S

Safety

▶ Crime Prevention, Dataveillance, and the Regulation of Information Communication Technologies

Sample Point

▶ Theory of Probability: Basics and Fundamentals

Sample Space

▶ Theory of Probability: Basics and Fundamentals

Sampling Effects in Social Network Analysis

Rick Grannis
Department of Sociology, UCLA, Los Angeles, CA, USA
Sociology, UCI, Irvine, CA, USA

Synonyms

Network sampling; Nonlinearity; Phase transition; Social networks

Glossary

Chain-referral sampling	A sampling technique that involves selecting initial respondents and then generating all future respondents by following along the contact network of those who have already sampled (also referred to as snowball, link-tracing, or random walk samples)
Clustering	The finding, common to social networks, that two social actors who share a common first neighbor have a disproportionately great probability of also being first neighbors of each other (also referred to as transitivity)
Dyad	A set of two social actors potentially or actually connected by a social relation
First neighbors	Social actors who directly connect to each other via a social relation
Giant component	A single component that connects the majority of network members
Mth neighbor	Social actors who connect to each other via a path consisting of m (but no fewer) social relations involving $m, -1$ (but no fewer) intermediaries
Phase transition	A sharp combinatorial transition point such that only a relatively small increase in network density (or some other individual-level

R. Alhajj, J. Rokne (eds.), *Encyclopedia of Social Network Analysis and Mining*,
https://doi.org/10.1007/978-1-4939-7131-2

	network characteristic) transitions the network from the situation in which virtually no network member connects via a path to any other network member to the situation in which most network members are connected via a path to most other network members (also referred to as a critical point or double jump threshold)
Second neighbors	Social actors who do not directly connect to each other via a social relation but whom both directly connect to the same intermediary
Social actor	An entity with the ability to create or nullify social relations
Social relation	Anything that exists, occurs, or flows between two social actors
Triad	A set of three social actors all of whom are potentially or actually connected to each other by social relations

Definition

While a great many useful statistical models and visualizations have been developed to explore large-scale complex networks, fewer have attempted to relate these models to data generated by samples. While, in many fields, complete (or near-complete) data is widely available, and while the Internet has made even more readily available, complete data about large-scale complex networks sufficient to answer many compelling social science questions does not exist and cannot be reasonably generated. In these cases, sampling theory must be used to connect data to models. This proves difficult for a variety of reasons such as: data collection methodologies which, while attempting to overcome nonresponse bias, deviate from standard sampling practices; and, the non-independence of both first neighbors and second neighbors. This entry reviews some of the different ways which have been created to faithfully translate data sampled from large-scale complex networks into useful statistics.

Introduction

One of the fundamental goals of social science is to understand how the interactions of individuals translate into the characteristics of the social systems they comprise (Schelling 1978). Network models have a potentially powerful role in this task.

To model the social world in network terms is to focus on social entities and the social relations among them, on the patterning of relations among social actors rather than the correlation among social actors' attributes. Often these entities forming the basis of social networks are individual people, but they can also be households, business firms, nations, or any social actor with the ability to create or nullify social relations. A social relation can mean anything that occurs or flows between two social actors such as information (advice, gossip, conversation, etc.), tangible resources (being exchanged, stolen, etc.), and emotional affect (Borgatti et al. 2009). Social relations never involve only a single social actor, and most, although not all, of them require at least the implied consent of both parties to exist. Social actors a and b determine the quality and quantity of their interaction with each other, including none at all, based not only on the motivations to interact with each other but also based on the social resources available to them through their other relations. Social relations are not independent.

Some research into large, complex networks ignores this nonindependence characteristic of social networks, and other research models its effects only with first-order neighbors and not second-order neighbors, threatening their theoretical reliability. Methodologies have been developed, however, to overcome these statistical obstacles in some cases and to fruitfully exploit them in others. This entry reviews these methods.

Key Points

Unlike many other types of network data, gathered from large databases or the Internet, social network data typically derives from samples and thus requires mathematical theory to reliably translate these sample statistics into estimates of population parameters.

When sampling from a social network, the non-independence of the sampled individuals can be both a resource to be exploited and an obstacle to be overcome.

This entry reviews three distinct ways in which social network researcher have dealt with nonindependence:

1. Well-defined statistical models faithfully deriving local-level network properties from sample data
2. Network sampling techniques which attempt to exploit the social network structure to yield more robust estimates of population parameters
3. Statistical models translating local-level information derived from traditional samples into reliable estimates of large-scale network realities

Deriving Local-Level Network Properties from Sample Data

Many useful sample estimators have been derived for local-level network properties (Granovetter 1976). As an example, Frank (1978) showed that a dyad count, C, a count of the number of distinct types of the $\binom{n}{2}$ possible dyads which can be induced from the network, has an unbiased estimator given by

$$\widehat{C} = \frac{N(N-1)Z}{n(n-1)}$$

where N is the total number of individuals in the population, n is the number of individuals in the sample, and Z is the dyad count within the sample and $n \geq 2$. If N is large, $n \geq 4$, and the sampling fraction n/N is a relatively small nonzero number p, the variance

$$\sigma^2 \approx \frac{C}{p^2}$$

has an unbiased estimator $\widehat{\sigma}^2$:

$$\widehat{\sigma}^2 \approx \frac{\sum_{i=1}^{N} Z_i^2 - Z}{p^4}$$

Frank (1978) further showed that a triad count C, a count of the number of distinct types of the $\binom{n}{3}$ possible dyads which can be induced from the network, has an unbiased estimator given by $\widehat{C} = Z/p^3$, where Z is the triad count within the sample, $p_r = n^{(r)}/N^{(r)}$, and $n \geq 3$.

If the sampling fraction p is a small number and N is large and $n \geq 6$, we have $\sigma^2 \approx \mathrm{C}/p^3$ and the variance has an unbiased estimator

$$\widehat{\sigma}^2 \approx \frac{Z - \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{Z_{ij.}^2}{2} + \sum_{i=1}^{N} \frac{Z_{ij}^2}{4}}{p^6}.$$

For decades, these types of statistics, among others, have proved foundational to the development of social network sampling theory. Other sample estimators have been derived which can be used to connect data sampled from a network to still other local-level network parameters (i.e., phenomena only depending on a single node or their immediate contacts). Deriving sample estimators for global-level network properties has proved more troublesome, however; the inherent non-independence of the sampled individuals and the relations interconnecting them inhibits analysis. Global-level properties, such as centrality and network centralization, for example, have proven to be highly sensitive to microlevel changes (Butts 2006, 2009).

Using Network Samples to Effectively Analyze Population Characteristics

While the nonindependence of sampled individuals, inherent to social networks, can potentially impede analysis when sampling from a network, under some circumstances, this same lack of independence can prove advantageous. One of these key circumstances involves estimating the characteristics of “hidden populations” such as drug users, the homeless, or artists (Heckathorn 1997), as well as the enormous set of nonresponders (both actual and potential) who would choose to opt out if a survey was requested and frustrate virtually all survey efforts (Grannis et al. 2011).

Standard sampling and estimation techniques, which require the researcher to select sample members with a known probability of selection, necessitate that researchers have a sampling frame listing all members in the population; however, for many populations of interest, such a list does not exist. Chain-referral (also known as snowball (Coleman 1958) or link-tracing) samples, which select initial respondents but then generate all future respondents by following along the contact network of those who have already been sampled, have been shown to be effective at penetrating hidden populations even without a sampling frame.

Such samples' use has been limited, however, due to the difficulty of making statistical inferences. Since members of the population to be sampled do not have the same probability of selection, those with many contacts are more likely to be included in the sample than social isolates. Also, biases in those "seeding" the sample, those first selected and from whom all subsequent respondents were indirectly recruited, may compound in unknown ways as the sampling process continued.

Because of this, chain-referral samples have often been considered to be nonprobability or convenience samples "which can only be assessed by subjective evaluation" (Kalton 1983) and "conventional wisdom among sociologists, public health researchers, and statisticians is that chain-referral sampling holds great promise for a number of problems, especially the study of hidden populations, but that it is so hopelessly biased that it cannot be used to make reliable estimates" (Salganik and Heckathorn 2004, p. 197).

Salganik and Heckathorn (2004), however, showed that, for any reciprocal relation, we can recover the proportion of any population belonging to a group A , PP_A, with knowledge only of the network structure connecting the population, the type of data which chain-referral samples most reliably generate:

$$PP_A = \frac{\widehat{D}_B \widehat{C}_{B,A}}{\widehat{D}_A \widehat{C}_{A,B} + \widehat{D}_B \widehat{C}_{B,A}}$$

where $\widehat{D}_A$ is the estimate of the average number of relations an individual of type A is involved with and $\widehat{C}_{AB}$ is the proportion of relations originating from an individual of type A which connect with an individual of type B.

They showed that the numerator and denominator of this statistic are both unbiased (Brewer and Hanif 1983) estimates of the population parameters and that the ratio of these two unbiased estimators is asymptotically unbiased with bias on the order of n^{-1} where n is the sample size (Cochran 1977).

The mean degree of a chain-referral sample would be higher than the mean degree of the population since these methods overrepresent people with high degree (Erickson 1979; Kalton and Anderson 1986; Eland-Goosensen et al. 1997); therefore Salganik and Heckathorn (2004) used two distinct mathematical approaches to show that, assuming only that nodes are drawn with probability proportional to their degree, the mean degree of the population can be estimated by

$$\widehat{D}_A = \frac{n_A}{\sum_{i=1}^{n_A} \frac{1}{d_i}}$$

where na is the number of individuals of type A in the sample, and d_i indexes over the number of individuals of type A in the sample.

Furthermore, Salganik and Heckathorn (2004) showed that since recruitments originating from a person of type A can be categorized into two sets, those that connect with another person in group A , r_{AA} and those which connect with a person in group B , r_{AB} and since the observed recruitments are a random sample from all edges, an unbiased estimate for C_{AB} is given by

$$\widehat{D}_{A,B} = \frac{r_{AB}}{r_{AA} + r_{AB}}$$

Thus, unlike a conventional probability sampling design, one can use data generated from a chain-referral sample not to directly estimate population parameters but rather to analyze the population *specifically as a network.* One uses the sample to make estimates about the network connecting the population and only then is this information about the network used to derive population proportions. By not attempting to estimate directly from the sample proportions to population proportions, one avoids many of the well-known problems with chain-referral samples (Heckathorn 2002).

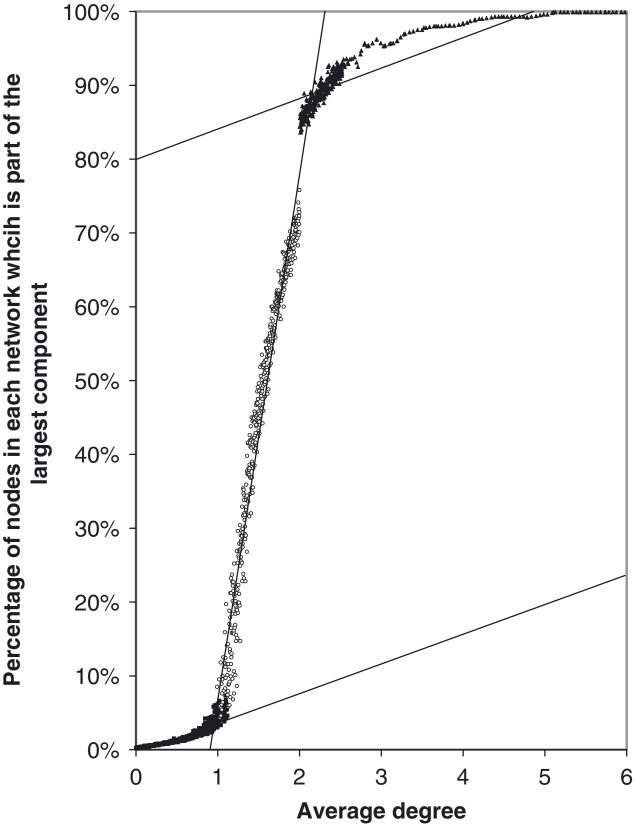
Modeling the Phase Transition from Sample Data

Arguably, the most fundamental global-level property of a large-scale, complex network concerns whether it contains a giant component, a single component that connects the majority of network members. If it does not, if the network is not connected, most other global network properties, like centralization, have little meaning. Furthermore, the only assumption about the structure of the network underlying chain-referral estimation procedures such as respondent-driven sampling is that the population basically consists of one connected component and that there exists a path between every person and every other person (Salganik and Heckathorn 2004).

While they depend to some extent on a network's individual-level properties, giant components, however, do not emerge as a linear response to individual-level changes but rather subtle changes in relations potentially produce extraordinarily different macro-level outcomes. As the average number of relations among individuals increases, the size of the components that they form does not grow relatively smoothly from small to large. Instead, in virtually all networks, a sharp threshold point exists combinatorially such that only a relatively small increase in the proportion of relations transitions the network from the situation in which virtually no network member is connected via a path to any other network member to the situation in which most network members are connected via a path to most other network members (Erdos and Rényi 1960; Janson et al. 2002). This has been referred to as the "critical point" or "the double-jump threshold" (Molloy and Reed 1995, 1998) or the "phase transition."

In thermodynamics, where the concept of a phase transition originated, it refers to an abrupt change in physical properties resulting from a relatively small change in temperature. Readers will be familiar with the phase transition occurring when water entirely transforms from a crystalline solid (ice) to a liquid over a relatively small temperature threshold. As temperature rises, water molecules remain organized in a crystalline structure until, during a very short interval of degrees, they completely transform into a liquid form. A relationship, similar to that between temperature and molecular structure, exists between relational density and social network structure. If one imagines a sparse network with nodes existing only in small, disconnected components, as the density of relations increases, there would be no large-scale effects until a threshold point was achieved, when the addition of a relatively few relations transforms the population from many small, disconnected, insular communities into a network composed primarily of one dominant, comprehensive community whose constituent members are mutually reachable via paths.

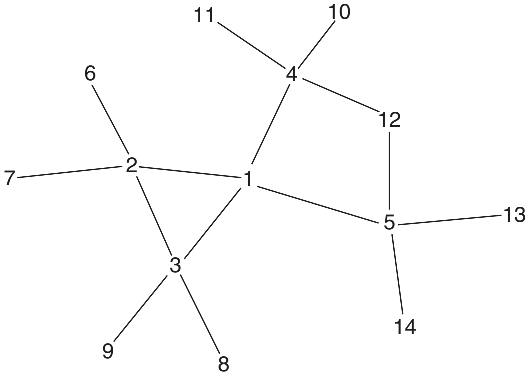
Figure 1 illustrates this. It represents 10,000 randomly generated networks, each with 1,000 nodes and each with an average degree varying between zero and six. It plots the proportion of nodes in each network that is part of the largest component as a function of the average degree of the network. The three lines drawn on the graph represent three distinct theoretical slopes clearly evidenced in this plot: before the phase transition, during the phase transition, and after the phase transition. Clearly, as the average degree rises, so does the size of the largest component in the network. While the increase is rather gradual as the average degree rises to one, the slope changes spectacularly, increasing almost 20-fold, as the average degree goes to two and then, just as dramatically, returns to a slope similar to its previous one. The size of the largest component, which is gradually increasing along most of the continuum of increasing degree, suddenly "jumps" to a new threshold, one that it would not have achieved until the network was 20 times as dense as it currently is, if the phase transition had not occurred.



Sampling Effects in Social Network Analysis, Fig. 1 The phase transition simulated by 10,000 randomly generated networks, each with 1,000 nodes. *Lines* represent theoretical slopes: before, during, and after the phase transition

Complications Unique to "Social" Networks

As a result of the extreme sensitivity of the phase transition, and other global network properties, to relatively trivial changes in local-network properties, great care must be exercised when using sample data to understand global-level network properties since virtually all samples are taken of local-network properties. For example, if we wanted to estimate whether the phase transition had occurred in a social network for which only sample data was available, there are known local-level properties, unique to social networks, which must be accounted for. Most important among these is clustering (illustrated in Fig. 2). In Fig. 2, node 1 connects to four others, and each of these also connects to four others (assume that the network continues on past the nodes labeled 6 and higher but that those edges simply are not illustrated here). While node 1 has four first neighbors, if we assumed that the number of second neighbors, neighbors of neighbors, would simply be a function of the number of first neighbors (i.e., that individuals only create or dissolve



Sampling Effects in Social Network Analysis, Fig. 2 An illustration of clustering. Node *1* has four first neighbors, labeled *2–5*, and nine second neighbors, neighbors of neighbors, labeled *6–14* network ties based on their immediate interactions), we would expect node 1 to have 12 (4 × 3) second neighbors. We would expect that each node connected to node 1 would have the same average number of neighbors as node l (four) but one less due to the fact that each has already spent one of their four relations connecting to node l. Instead, node 1 has only nine second neighbors. This occurs because nodes 2 and 3, both of which are first neighbors of node 1, each spend one of their relations connecting to each other and because nodes 4 and 5, both first neighbors of node 1, share a common second neighbor in node 12.

To account for the importance of clustering such as this on large-scale network properties, Grannis (2010) distinguished the number of first and second neighbors as two distinct variables, identifying the mean number of neighbors of a typical randomly chosen node as f_1 while letting f_2 represent the mean number of distinct second neighbors, regardless of how this number arises, whether influenced by transitivity or clustering or any other process that acts upon the distribution of second neighbors. The variable f_2 is measured independently of f_1. Because this variable, f_2, ignores those edges that do not contribute to unique second neighbors, it therefore explicitly accounts for clustering (as well as the necessary connection with the original node). Thus, the ratio $g = (f_2/f_1)$ equals the proportional increase in the number of new neighbors. Thus, in the network illustrated in Fig. 2, $f_1 = 4$, $f_2 = 9$, and $g = 2.25$.

Using this notation, we expect the average node has f first neighbors $f_2 = f_1 g^1$ second neighbors, $f_3 = f_1 g^2$ third neighbors, and $f_m = f_1 g^{m-1}$ mth neighbors. The total number of neighbors reached in l (or fewer) steps is given by the geometric series

$$\sum_{m=1}^{l} f_m = \sum_{m=1}^{l} f_1 g^{m-1} = f_1 \frac{g^l - 1}{g - 1}$$

The expected size of the connected component to which a typical node belongs equals one (itself) plus the number of neighbors it could reach after an infinite number of steps. Substituting ∞ for l into the formula above and adding one yields

$$f_1 \frac{g^\infty - 1}{g - 1} + 1$$

If $g < 1$, g^∞ asymptotically approaches zero and the expression reduces to

$$1 + \frac{f_1}{1 - g}$$

If $g > 1$, the first term (and thus the entire expression) approaches infinity; the average component size is infinite (i.e., a giant component has formed). If, however, $g = 1$, then the first term becomes indeterminate, the phase transition point. A giant component exists when $g > 1$ and does not exist when $g < 1$.

Intuitively, one can understand this as follows. Assume that Jacob connects to Sophia, Ben, and Hannah. We can consider these individuals as the starting points on branches originating from Jacob. Regardless of Sophia's, Ben's, or Hannah's initial degrees, they must use one tie connecting to Jacob (else they would not be Jacob's neighbor), and they may use some (perhaps none, perhaps all) of their other ties (if any) connecting to each other (i.e., clustering). Any remaining ties will ramify out into new branches. Assume that, after connecting to Jacob and perhaps to each other, Sophia, Ben, and Hannah have zero, two, and one remaining ties, respectively, available to connect to new nodes. By connecting to Sophia, the number of branches originating from Jacob has decreased; by connecting to Ben, the number of branches originating from Jacob has increased; and, by connecting to Hannah, the number of branches originating from Jacob has stayed the same. In general, if the neighbors which any typical node is likely to connect to will, after accounting for ties spent in clustering and the initial connection, on average yield more than one new branch each, this process will expand throughout the network and a giant component can be expected to form.

Illustrative Example: Sampling Core Discussion Networks from the General Social Survey

To illustrate, consider the General Social Survey (GSS) data on the confidants with whom Americans discuss important matters. Examination of trends in GSS discussion networks (which were collected in 1985, 1987, and 2004) at the individual level have reported important changes in the last generation. For example, McPherson et al. (2006, p. 353) noted that individual networks are a third smaller in 2004 than in 1985 (about two people instead of three) and that the number of people saying there is no one with whom they discuss important matters nearly tripled.

To estimate whether the phase transition has occurred and a giant component exists, we need to calculate the proportional increase in the number of new neighbors,

$$g = \frac{f_2}{f_1}(f_1 \neq 0)$$

If $g > 1$, then the phase transition has occurred. To perform this calculation, we need to know the number of a node's neighbors, f_1, as well as the number of distinct second neighbors that are not also first neighbors f_2. The GSS data provides information on the first variable f_1, the number of others each individual nominates as someone with whom they discuss important matters. Data concerning the second variable, f_2, however, is not readily apparent and requires some calculation.

The simplest, although probably not the most accurate, way to do this would be to assume that one's discussion partners do not discuss important matters with each other (i.e., that there is no clustering) but, rather, that they link to others randomly with the only stipulation being that each discussion partner has used one tie connecting to the respondent; all other ties extend outward. Table 1 displays results generated using this model. It shows that the value of g to be 2.975 in 1985 and 2.565 in 2004.

Alternatively, one could theorize that some of an individual's discussion partners also discuss important matters with each other. The GSS did not ask respondents which of the people they knew, whom they discussed important matters with, also discussed important matters with each other. The GIS did, however, ask respondents to characterize the relationship between each pair of the people they mentioned into three categories: as "especially close, as close or closer" than their relationship to the respondent; "total strangers"; or somewhere in-between. Which, if any, of these corresponds to discussing important matters is unknown.

If we theorize that all pairs of individuals identified as "especially close" are, in fact, discussion partners, then members of pairs so identified each spend one of their ties connecting to the other. Thus, fewer ties will extend outward to others. Table 1 shows that, under this model, the value of g is 1.894 in 1985 and 1.592 in 2004. For the "no clustering" model as well as the model of those who were "especially close" as discussion partners, $g > 1$ indicates that a giant component clearly unites most isolates.

Some might assume that only the more exclusive "especially close" relation represents those who would have in fact nominated each other as someone they discuss important matters with if the GSS had surveyed them. However, while it seems reasonable to assume that not all pairs of individuals whom respondents categorized as neither "especially close" nor "total strangers" would have nominated each other as someone they discuss important matters with, it is certainly arguable that some of them might have, given that this intermediate category implicitly included those "almost as close." Thus, if we further theorize that not only those who are "especially close" but also those in the intermediate category, neither "especially close" nor "strangers," are also discussion partners, then an even larger number of pairs of those directly connected to the respondent will spend ties connecting to each other. Table 1 shows that, using this model, the value of g is 0.8277 in 1985 and 0.6818 in 2004. If this model is correct, $g < 1$ tells us that in neither year has the phase transition occurred and all components are minuscule.

This definitional distinction has dramatic effects when one considers the network model it

Sampling Effects in Social Network Analysis, Table 1 Values for individual-level predictors of the phase transition

		1985		2004	
		Preferential attachment	Random	Preferential attachment	Random
Proportional increase in new neighbors (g)	No clustering	2.975 (0.06935)	2.975 (0.05055)	2.565 (0.09166)	2.565 (0.09771)
	Especially close	1.815 (0.04975)	1.894 (0.03968)	1.505 (0.06020)	1.592 (0.06161)
	Neither especially close nor strangers	0.6858 (0.02996)	0.8277 (0.03326)	0.5046 (0.03302)	0.6818 (0.03902)
Number of respondents		1,525		1,466	
Average number of first neighbors (f_1)		2.980 (0.4409)		1.987 (0.04409)	

generates. The difference between these two scenarios is not merely that one component is somewhat larger, but rather it is a difference in orders of magnitude. Theoretically, it signals the difference between a society that is primarily united into a single discussion network and a society that has utterly disintegrated.

In the first case, it is possible that the typical person is involved in an extended discussion network (e.g., she discusses important matters with someone who discusses important matters with someone who discusses important matters with someone, etc.) that ultimately includes multiplied millions of people. While it is unlikely that the specifics of one's discussions transmit over any distance, it is possible that general norms or values could diffuse and a general awareness, if not consensus, could form.

The second case is quite distinct. To understand just how tiny the nonphased components are, we can use the formula derived above for calculating average component size when $g < 1$

$$1 + \frac{f_1}{1 - g}$$

We find that, in this case, the size of the average discussion component is 18 in 1985 and 7 in 2004. Thus, in this case, most persons' discussion networks do not extend much beyond those they have direct discussions with. Instead of society consisting of an extended network diffusing norms and values, it would have been pulverized into tiny groups, perhaps no larger than a single individual's discussion network.

Summary

This entry has reviewed some of the many unique issues which arise when one uses sample data to model social networks. In some cases, such as with chain-referral sampling, social network conceptualizations may prove advantageous, as statistical methods have been created which allow researchers to translate otherwise questionable data into robust estimates of population parameters. In contrast, other cases demonstrate that when using conventional sample data to understand network processes, great care must be taken in how one theorizes, defines, and operationalizes local-level processes. Social actors, unlike nonsocial network nodes, are aware of and respond to the actions or inactions of both their first neighbors and their second neighbors (Friedkin 1983). Relatively trivial variations in the social responses by these individual actors may have dramatic effects on the theoretical understanding which results from analyzing the sampled data.

Cross-References

- ► Network Representations of Complex Data
- ► Probabilistic Analysis
- ► Probabilistic Graphical Models
- ► Research Designs for Social Network Analysis

References

Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. Science 323:892–895

Brewer KRW, Hanif M (1983) Sampling with unequal probability. Springer, New York

Butts CT (2006) Exact bounds for degree centralization. Soc Netw 28:283–296

Butts CT (2009) Revisiting the foundations of network analysis. Science 325:414–416

Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York

Coleman JS (1958) Relational analysis: the study of social organization with survey methods. Hum Organ 17:28–36

Eland-Goosensen M, Van De Goor L, Vollemans E, Hendriks V, Garretsen H (1997) Snowball sampling applied to opiate addicts outside the treatment system. Addict Res 5(4):317–330

Erdos P, Rényi A (1960) On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5:17–61

Erickson BH (1979) Some problems of inference from chain data. In: Karl FS (ed) Sociological methodology, vol 10. Jossey-Bass, San Francisco, pp 276–302

Frank O (1978) Sampling and estimation in large social networks. Soc Netw 1:91–101

Friedkin NE (1983) Horizons of observability and limits of informal control in organizations. Soc Forces 62:54–77

Grannis R (2010) Six degrees of 'who cares?'. Am J Sociol 115:991–1017

Grannis R, Freedy E, Freedy A (2011) Ultra-rapid social network sampling in cross-cultural environments. In: HSCB conference proceedings: integrating social science theory and analytic methods for operational use, Arlington

Granovetter M (1976) Network sampling: some first steps. Am J Sociol 81(6):1287–1303

Heckathorn DD (1997) Respondent-driven sampling: a new approach to the study of hidden populations. Soc Probl 44(2):174–199

Heckathorn DD (2002) Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. Soc Probl 49(1):11–34

Janson S, Luczak T, Rucinski A (2002) The phase transition, Chap. 5. In: Random graphs. Wiley, New York, pp 103–138

Kalton G (1983) Introduction to survey sampling. Sage, Beverly Hills

Kalton G, Anderson DW (1986) Sampling rare populations. J R Stat Soc Ser A 149:65–82

McPherson M, Smith-Lovin L, Brashears ME (2006) Social isolation in America: changes in core discussion networks over two decades. Am Sociol Rev 71:353–375

Molloy M, Reed B (1995) A critical point for random graphs with a given degree sequence. Random Struct Algorithms 6:161–179

Molloy M, Reed B (1998) The size of the giant component of a random graph with a given degree sequence. Comb Probab Comput 7:295–305

Salganik MJ, Heckathorn DD (2004) Sampling and estimation in hidden populations using respondent-driven sampling. Sociol Methodol 34:193–239

Schelling TC (1978) Micromotives and macrobehavior. Norton, New York

Recommended Reading

Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. Science 323:892–895

Doreian P, Woodard KL (1992) Fixed list versus snowball selection of social networks. Soc Sci Res 21:216–233

Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256

Newman MEJ (2010) Networks: an introduction. Oxford University Press, New York

Scalable Distributed Systems

► Cloud Computing

Scalable Graph Clustering and Its Applications

Hiroaki Shiokawa[1] and Makoto Onizuka[2]

[1]Center for Computational Sciences, University of Tsukuba, Ibaraki, Japan

[2]Graduate School of Information Science and Technology, Osaka University, Osaka, Japan

Synonyms

Cluster analysis; Community detection; Modularity

Glossary

Cluster	A group of densely interconnected nodes
Community Detection	A function of network analysis that identifies groups of densely connected nodes

Graph	A set of nodes and edges connecting the nodes
Hub	A special role of node that bridges multiple clusters
Network	A graph extended with semantics and interactions between nodes and edges, respectively
Outlier	A special role of node that is not hub and does not belong to any clusters. In many cases outliers are regarded as noises
Partition	Division of nonoverlapping subsets

Definition

Graph is one of the fundamental data structures and we can easily find graphs in many applications and services. Graph cluster analysis is a key technique to understand structures, characteristics, and interrelationships graphs. The problem of the graph cluster analysis is to find clusters inside of which nodes are densely connected and sparsely connected inter clusters, and this problem has been studies for some decades in many fields, particularity in computer science and physics.

Introduction

Graphs can represent data entities as well as the relationships among entities. They arise in a wide range of application domains from the Internet to biological networks and beyond. Graphs are becoming large year by year, and graphs of unprecedented size can be easily found. Therefore, there is no longer any doubt about the need for techniques that can analyze large graphs quickly. For these large-scale graphs, graph cluster analysis (a. k. a. community detection) is one of the most important techniques in various research areas such as data mining, social science, and computer networks. A cluster can be regarded as a group of nodes that are densely connected within a group and sparsely connected between different groups. By discovering the hidden cluster structures in large-scale graphs, we can understand the characteristics and interrelationships of the nodes forming the graph. This entry provides an overview of the vast literatures on efficient graph clustering algorithms that arise in different contexts.

Key Points

This entry is organized as follows: *Historical Background* briefly describes several areas of research studies relevant to this entry. *Scalable Clustering Algorithms for Large Graphs* introduces 15 graph clustering algorithms including several traditional and the state-of-the-art methods. *Key Applications* reviews several recent applications using the graph clustering algorithms what we introduced in this chapter. Finally, in *Future Directions*, we conclude this entry with a brief future work discussion.

Historical Background

The problem of finding clusters in a graph has been studied for decades in many fields, particularity in computer science and physics. Graph clustering algorithms that we investigate in this entry are formalized as follow:

Problem 1 (Graph clustering): *Given:* Graph $G = (V, E)$, where V and E are sets of nodes and edges included in the graph G, respectively. An objective function $f(C)$ that evaluates qualities of clustering results, where C is a set of clusters extracted from the graph G by using graph clustering algorithms.

Find: All disjoint clusters $C_i = \{V_i, E_i\}$ in the graph G that maximize $f(C)$, where $C_i \in C$, $V_i \subseteq V$, and $E_i \subseteq E$.

To simplify the representations, we assume graphs are undirected and unweighted, also $V = \sum V_i$ and $V_i \cap V_j = \emptyset$ for any $i \neq j$, without loss of generality.

Graph clustering problem is traditionally related to the *graph partitioning problem*. The goal of the graph partitioning problem is to find clusters that minimize the numbers of edges that bridge clusters. However, the graph partitioning has drawbacks in that there is no guarantee of the optimality of the partitioning result. Moreover, in order to find multiclusters from a graph, we have to previously determine the number of clusters that are to be extracted from the graph.

To overcome the above drawbacks, *modularity* has been proposed in 2004 as a measure of graph clustering quality. It measures the difference of a cluster structure from an expected random graph, and modularity-based clustering methods try to extract clusters that maximize the score of the modularity. Since modularity-based methods are simple and effective for complex graphs, they are widely used in many applications and further studies. However, it is difficult for modularity-based methods to find small clusters hidden in large-scale graphs; these methods fail to fully reproduce the ground-truth; this problem involves a new type of clustering algorithms, called *structural graph clustering*.

Structural graph clustering algorithms have been widely examined in many studies in the last few years. In this type of clustering, the algorithms extract densely connected subgraphs as clusters and they also find several special role nodes, hubs and outliers, which are not assigned to any clusters. By finding densely connected clusters, hubs, and outliers, they achieve more accurate clustering results than existing algorithms (e.g., edge-cut based methods and modularity based methods).

Scalable Clustering Algorithms for Large Graphs

Graph clustering algorithms that we introduce in this entry are roughly grouped into three classes: *edge-cut based algorithms, modularity-based algorithms,* and *structural similarity based algorithms*. Each class is also divided into two groups based on its clustering manner (i.e., top-down method or bottom-up method). The detail of each method will be discussed in the following subsections.

Edge-cut Based Algorithm

A number of algorithms for graph clustering have been studied for decades. In these studies, graph clustering problem is related to the minimum cut and the graph partitioning problem. This problem is desired to partition the graph so as to minimize the numbers of edges across the clusters. This problem is also referred to as the minimum-cut problems.

Top-down method

The simplest case is the two-way minimum-cut problem, in which we wish to partition the graph into two clusters that minimize the numbers of the edges across the clusters. Ding et al. proposed one of the most popular algorithms, called *min-max cut method* (Ding et al. 2001). It tries to partition a graph into two clusters C_u and C_v by using top-down manner. The main idea of min-max cut method is minimizing the number of edges between C_u and C_v and maximizing the number of edges within each of them. A cut is evaluated by the number of edges that connect the nodes in C_u and in C_v.

A major drawback of the method is that it falls into local optimum. That is to cut out a single node from other nodes is always the optimum. Thus, in practice, the min-max cut method must be accompanied with a constraint, such as C_u and C_v should be of equal or similar size (i.e., $|C_u| \approx |C_v|$). However, the constraint is not always appropriate for the real-world graphs. For example, in social networks, it is natural that some communities in the network are much larger than the others. To address this issue, *normalized cut method* (Shi and Malik 2000) was proposed by Shi and Malik. This method is an extension of min-max cut method that normalizes the total number edges between each cluster to the rest of the graph. As a result, the normalized cut method does not produce unnatural small clusters of the graph as an optimum cluster.

Both the min-max cut method and the normalized cut method are based on two-way partitioning that divides a graph into two clusters. In order to divide a graph into k clusters, we have to recursively apply the two-way partitioning by a top-down manner: splitting the graph into two clusters and then further splitting these clusters recursively, until k clusters are determined. However, there is no guarantee for the optimality of the recursive clustering results and there is no indicator to stop the bisection procedure.

Kerninghan-Lin algorithm, proposed by Kerninghan and Lin (1970), is one of the traditional techniques for multiway graph partitioning. This algorithm is based on hill climbing technique for optimizing graph partitioning. Initially, it starts from a randomly cut graph. After that, it incrementally swaps nodes among partitions until the number of overall edge-cuts is minimized (this procedure is called coarsening). The algorithm produces a local optimum result and it may not be the global optimum. Several improved variances of Kerninghan-Lin algorithm have been proposed, e.g., *METIS* (Karypis and Kumar 1998); however, it is difficult to apply these methods to large-scale graphs since they use coarsening techniques that require significantly large computational cost (Wang et al. 2014).

Bottom-up method

In order to achieve scalable multiway partitioning, Wang et al. proposed bottom-up method, called *MLP* (Wang et al. 2014). MLP uses multilevel label propagation to iteratively coarsen a graph until the coarsened graph is small enough. Then it uses high-quality off-the-shelf partitioning algorithms to generate the final partitions on the coarsened graph. In the literature (Wang et al. 2014), they used METIS (Karypis and Kumar 1998) for producing the final partitioning. MLP is designed easily pararellized and more effective than existing approaches. Hence, it successfully finds good partitions such that the clustering result has few edges across different partitions on billion-node graphs with acceptable memory space and running time.

Modularity-Based Algorithms

As we described in the previous subsection, the edge-cut-based methods have drawbacks that there is no guarantee of the optimality of the clustering result. Furthermore, so as to achieve multiway partitioning, we have to specify the number of clusters that we like to extract from graphs. To overcome the above drawbacks, *modularity* was proposed as a quality measure of graph clustering by Newman and Girvan (2004). It indicates the degree of difference of graph from its random graph. Specifically, the modularity is defined as follows:

Definition 1 (Modularity *Q*): Let e_{uv} be the number of edges between cluster C_u and C_v; a_u be the number of edges that are attaches to the nodes in cluster C_u; and $m = |E|$. The following equation gives the modularity score of the clustering result.

$$Q = \sum_{C_u \in C} \left\{ \frac{e_{uu}}{2m} - \left(\frac{a_u}{2m} \right)^2 \right\}$$

In Definition 1, $\frac{a_u}{2m}$ indicates the expected fraction of edges of cluster C_u that can be obtained when we assume the graph to be a random graph. Thus, well-clustered graphs have high modularity scores since the value of e_{uu} is larger than the random graph. The main task of the modularity-based method is to find groups of nodes that maximizes the modularity Q.

Top-down method

Girvan-Newman algorithm (Newman and Girvan 2004) is a divisive clustering algorithm based on *betweenness centrality.* The Girvan-Newman algorithm identifies the edge with high betweenness centrality and remove them so as to bisecting a graph into clusters. The formal definition of the betweenness centrality of an edge is defined as the proportion of shortest paths between any nodes that pass through the edge. Thus, for a given edge e, the betweenness centrality $\mathfrak{B}(e)$ is defined as follows:

Definition 2 (Betweenness centrality): Let $cp(e, i, j)$ be the number of shortest paths between node i and j that pass through edge e; and $sp(i, j)$ be the total number of shortest paths between node i and j. The betweenness centrality $\mathfrak{B}(e)$ is defined as follows:

$$\mathfrak{B}(e) = \frac{cp(e,i,j)}{sp(i,j)}$$

The algorithm sorts edges in E by $\mathfrak{B}(e)$, and removes the edges with the highest $\mathfrak{B}(e)$ score. In order to specify the number of clusters that achieves a good clustering result, Girvan-Newman algorithm employs the modularity Q as the objective function.

Bottom-up method

Although modularity-based algorithms are effective for many applications, finding the maximum modularity involves NP-hard complexity (Newman and Girvan 2004). This problem has led to the introduction of approximation approaches. Instead of performing an exhaustive search, a greedy modularity-based approach, named *Newman clustering*, was proposed by Newman (2004). It iteratively selects and merges a pair of nodes so as to maximize modularity. It produces reasonable clusters with hierarchical structure, which represents the history of merges. Despite the effectiveness in avoiding the NP-hard problem, it requires high computing cost $O(|V|^2)$, where $|V|$ is the number of nodes.

Various algorithms have been proposed to reduce the computational cost of Newman clustering. Clauset et al. proposed a greedy modularity-based algorithm, called *CNM* (Clauset et al. 2004), which is one of the most widely used methods recently. They used *modularity gain*, which is obtained after merging a pair of nodes, and nested heap structures of modularity gain for all pairs of nodes. It iteratively selects and merges the best pair of nodes so as to obtain the largest modularity gain. The pair is chosen from the heap until no pairs improve the modularity. The computational cost of CNM is $O(d|E| \log |V|)$, where d and $|E|$ are the depth of the hierarchical clustering result and the number of edges, respectively. However, Blondel et al. reported that CNM does not work in reasonable computational time for graphs with more than 500 thousand nodes (Blondel et al. 2008). Moreover, they also reported that CNM has a tendency to produce super-clusters, which contains a large fraction of nodes, with significant low modularity.

Blondel et al. proposed an efficient greedy algorithm *BGLL (a.k.a. Louvain method)* (Blondel et al. 2008). In contrast to CNM, it computes the modularity gain only for the adjoined nodes pairs for local optimization. Blondel et al. reported that BGLL requires almost 3 h to process graphs with 118 million nodes (Blondel et al. 2008). Although BGLL is effective for extracting high modularity clusters, it is difficult for BGLL to realize quick responses for graphs of unprecedented size, such as Web graphs with few billion edges. This is because it iteratively scans all nodes/edges as long as the modularity is incrementing. It is known that BGLL has near-linear time complexity for the size of edges (Shiokawa et al. 2013).

Recently, Shiokawa et al. proposed *incremental aggregation algorithm* (Shiokawa et al. 2013) for finding clusters from large-scale graphs. For the efficient clustering, they use three techniques; incremental aggregation, incremental pruning, and efficient nodes ordering. To the best of our knowledge, this is representative for the state of the art algorithm; it achieves fast clustering with high modularity scores. It is reported that this algorithm computes clusters within 3 min for a graph that have more than 100 million nodes and one billion edges. The time complexity of this algorithm is $O((1 - c)|E|)$ where c is an average clustering coefficient of the given graph.

As we described above, modularity-based method can handle significantly large-scale graphs and they are one of the most widely used in many applications. However, despite the efficiency of the modularity-based approach, these methods cannot identify hubs and outliers in graphs. Furthermore, a recent research pointed out that modularity is not a scale-invariant measure (Fortunato and Barthélemy 2007). Hence, it is difficult for modularity-based methods to find small clusters hidden in large-scale graphs; these methods fail to fully reproduce the ground-truth (Xu et al. 2007). This serious problem is famously known as the *resolution limit* of modularity (Fortunato and Barthélemy 2007).

Structural Similarity Based Algorithms

Due to resolution limit of modularity, structural similarity-based methods have been widely used in many research areas in the last few years. The main idea of the structural similarity-based methods are that if adjacent nodes are densely connected each other, they should be assigned to the same cluster. To find clusters based on this idea, these methods detect densely connected subgraphs as clusters. Furthermore, besides extracting clusters, they can find special role nodes, hubs, and outliers, which are not belong to any clusters. By finding densely connected clusters, hubs, and outliers, they achieve more accurate clustering results than existing algorithms (e.g., edge cut based methods and modularity-based methods).

In order to evaluate the density of adjacent nodes, Xu et al. proposed a measurement called *structural similarity* (Xu et al. 2007), formally:

Definition 3 (Structural Similarity): The following equation gives the structural similarity, denoted by $\sigma(u, v)$, between node u and v.

$$\sigma(u, v) = \frac{|N[u] \cap N[v]|}{\sqrt{|N[u]||N[v]|}}$$

where $\boldsymbol{N}[\boldsymbol{u}] = \{\boldsymbol{v} \in \boldsymbol{V} | (\boldsymbol{u}, \boldsymbol{v}) \in \boldsymbol{E}\} \cup \{\boldsymbol{u}\}$.

Structural similarity is a score ranging from zero to one according to the scale of matching degree of $N[u]$ and $N[v]$. When adjacent nodes share many members of their neighborhoods, their structural similarity gets large.

The main task of the structural similarity based methods is to find all clusters, hubs and outliers from the graph based on Definition 3 and user-specified parameters.

Top-down method

Yuruk et al. proposed *DHSCAN* (Yuruk et al. 2007), which is based on top-down clustering approach. As well as Girvan-Newman algorithm, DHSCAN iteratively removes edges between nodes in the ascending order of the structural similarity of the node given by Definition 3. The graph is divided into disconnected components by the removal of edges. This iterative and divisive procedure produces a dendrogram showing the hierarchical structure of the clusters. Additionally, this procedure stops when the structural similarity-based modularity (Feng et al. 2007), that is, a slightly modified version of Definition 1, is maximized. The formal definition of the structural similarity based modularity is as follows:

Definition 4 (Structural Similarity-Based Modularity Q_s): Let IS_u be the sum of the structural similarity of the nodes within cluster C_u, DS_u be the be the sum of the structural similarity between nodes in cluster C_u and any nodes in the graph, and TS be the total structural similarity of any adjacent nodes in the graph. The following equation gives the structural similarity based modularity score of the clustering result:

$$Q_s = \sum_{C_u \in C} \left\{ \frac{IS_u}{TS} - \left(\frac{DS_u}{TS} \right)^2 \right\}$$

Yuruk et al. reported that the quality of the clustering results obtained by Q_s is better than the one by using the modularity based algorithms (Yuruk et al. 2007). However, as well as Girvan-Newman method, DHSCAN requires high computational cost for clustering since it iteratively computes the structural similarity for all edges. Hence, it is difficult to apply DHSCAN to large-scale graphs.

Bottom-up method

To achieve more efficient clustering while keeping higher quality of clustering results than DHSCAN's, several bottom-up algorithms have been studied in recent few years.

SCAN , proposed by Xu et al. (2007), is the most popular method based on the structural similarity. It is an extension of a traditional density-based clustering method, *DBSCAN* (Ester et al. 1996). This algorithm can find clusters as well as hubs and outliers in a graph by specifying two parameters ϵ and μ. For finding clusters, SCAN first extracts seeds of clusters called *cores* from the graph. The core is formally defined as follows:

Definition 5 (Cores): Let $N_\epsilon[u] = \{v \in V | \sigma(u, v) \geq \epsilon\}$; ϵ and μ be user-specified parameters. A node u is a core if and only if $|N_\epsilon[u]| \geq \mu$.

Once SCAN determines node u as a core, it assigns all nodes in $N_\epsilon[u]$ to the same cluster of the core. It then selects a node from the cluster and recursively finds cores and expands clusters from the cores. After the termination of the cluster extraction, SCAN classifies the remainder nodes that are not belong to any clusters as hubs or outliers. If a remainder node bridges multiple clusters, it is regarded as a hub; otherwise it is an outlier. This algorithm requires $O\left(\frac{|E|^2}{|V|}\right)$ time on average. This is because SCAN evaluates the structural similarity for all adjacent nodes, and each similarity computation requires $O\left(\frac{|E|}{|V|}\right)$ time on average.

Huang et al. and Sun et al. proposed parameter-free methods named *SHRINK* (Huang et al. 2014) and *gSkeletonClu* (Sun et al. 2014), respectively. SHRINK is a parameter-free and hierarchical clustering algorithm that shares the advantages of structural similarity and modularity-based methods. It first computes structural similarities for node pairs of all adjacent nodes. Then it aggregates densely connected node pairs, called micro community, into same clusters if the aggregation improves Q_s score given by Definition 4. In this way, SHRINK achieves parameter-free and hierarchical clustering. In contrast, gSkeletonClu finds the best value of ϵ in the clustering process. As well as SHRINK, it first computes the structural similarities for node pairs of all adjacent nodes in a graph. After that, it extracts spanning trees from the graph by using the results of the structural similarities; and then it searches the effective values of ϵ in the trees. SHRINK and gSkeletonClu are user-friendly algorithms since they do not require the user-specified parameters. Moreover, Huang et al. reported that the clustering quality of them is better than that of SCAN (Huang et al. 2014; Sun et al. 2014). However, these methods require exhaustive similarity computations for node pairs of all adjacent nodes; hence the time complexities of SHRINK and gSkeletonClu are $O\left(\frac{|E|^2}{|V|}\log|V|\right)$ and $O\left(\frac{|E|^2}{|V|}+|V|\log|V|\right)$, respectively. Therefore, as well as SCAN, both SHRINK and gSkeletonClu require large computation time for large-scale graphs.

Lim et al. proposed LinkSCAN* (Lim et al. 2014), which uses SCAN to find overlapping communities from graphs. In order to detect overlapping communities very accurately, LinkSCAN* transforms graphs into link-space graphs that combines the advantages of graph and line graph. However, this transformation entails an increase of the size of graph for clustering. Hence, they introduced a graph sampling step. This approach is certainly efficient in reducing the computation time; to the best of our knowledge, LinkSCAN* is one of the most efficient methods for structural similarity-based clustering. However, it degrades the quality of clustering results compared to SCAN since sampling involves approximation.

Recently, a fast and exact structural clustering algorithm, called *SCAN++*, has been proposed by Shiokawa et al. (2015). The main idea behind SCAN++ is based on a property of real-world graphs: *if node u is two hops away from node v, their structural neighborhoods, $N[u]$ and $N[v]$, are likely to share large portion of nodes.* Based on this property, SCAN++ prunes the structural similarity computations for the nodes that are shared between a node and its two-hop-away node. Specifically, it employs the following techniques: (1) it uses a new data structure, directly two-hop-away reachable node set (DTAR), the set of nodes that are two hops away from a given node, (2) it reduces the cost of clustering by avoiding unnecessary density evaluations if the nodes are not included in DTARs, and (3) its density evaluation is efficient since DTAR allows the reuse of density evaluation results. As a result, SCAN++ achieves almost linear time complexity in terms of number of edges without sacrificing clustering quality compared with SCAN.

Key Applications

Here, we overview several related applications that employ graph clustering algorithms.

Information Networks

Graph clustering algorithms plays as an important tool for analyzing the function of information networks. In the context of the information networks, graph clustering algorithms have many applications such as market analysis and infrastructure management.

Graph clustering is used to identify and analyze connectivity and functions of substructures in computer networks. One of the most representative examples is the network topology design. In order to reduce the traffic load of a network, Grout and Cunningham applied a graph clustering algorithm to the optimization of a network topology (Grout et al. 2007). They made a graph from the network topology, which consists of network switches and nonswitches, and then they extracted clusters from the graph for specifying subnetwork topologies with high traffic loads.

Graph clustering is also employed in the structural design and operation of wireless ad hoc networks (Nguyen et al. 2011) and sensor networks (Furuta et al. 2006). Given that such networks exhibit dynamic topology changes; graph clustering is useful to find better network routing. For example, Nguyen et al. and Dinh et al. used graph clustering for mobile ad hoc networks (MANET) (Dinh et al. 2012). By extracting clusters in the MANET, they directly route or forward messages (i.e., network packets) to nodes in the same cluster as the destinations. By doing this way, they avoid unnecessary messages forwarded through nodes in different clusters, which reduces the number of duplicate messages as well as the overhead information.

Social Networks

One of the most popular applications of graph clustering is social network analysis. In the field of social network analysis, graph clustering serves as an essential tool to reveal social phenomena such as social trends, social behavior, and information diffusion processes. Zhou et al. employed several graph clustering algorithms to extract communities from social message networks such as emails, tweets on Twitter, and Facebook updates (Zhou et al. 2012). They applied graph clustering methods to the tweets collected from United States President Barak Obama's account from November 2009 to February 2010, and they revealed that the President roughly belongs to the following five communities. President: comments on his advocacy of the Presidency, Public Policy: related to domestic and international politics and policy, Holidays: mainly about wishing the best for American families and friends, Senate Vote: related to health bills and Blessing Haiti: a response to the tremendous earthquake.

Lee et al. also applied graph clustering algorithms to the online microblogging service Twitter to track the dynamically changing topics that are discussed on Twitter (Lee et al. 2014). Since such online microblogging services are noisy, informal, and surge quickly, they investigated an incremental density-based clustering approach. By finding clusters from Twitter and using Part-of-speech Tagger (Toutanova and Manning 2000) for topic identification, their method can effectively track, on the fly, events and topic dynamics even from large volumes of social network data. For example, their approach detected the emergence in Twitter of the "SOPA (Stop Online Piracy Act) protest" in January 2012.

Finding influential individuals from users of social networks and microblogging services is one of the most essential issues for applications such as viral marketing. This problem is formally defined as the problem to find the individuals that influence the greatest number of users. The approach proposed by Chen et al. offers an effective detection of influential individuals based on SCAN (Chen et al. 2014). In order to generate a small set of candidate individuals, this approach extracts clusters, hubs, and outliers from graphs by their SCAN-based algorithm. Then, it identifies the influential individuals from just the significantly large clusters and hubs that connect a lot of clusters. As a result, they revealed that information can spread easily from hubs to many adjacent clusters; their approach outperforms several previous approaches for real-world social networks.

In the current information society, the study of social networks trends to overlap the study of information networks, as the popularity and significance of electronic messaging has become overwhelming in the big data era.

Biological Networks

In recent years, large amounts of biological networks have been generated in the field of medical care and bioinfomatics. Biological networks are naturally decomposed into several functional components, which are commonly referred to modular networks such as transcriptional modules, protein complexes, gene functional groups, and signaling pathways.

Graph clustering algorithms are traditionally applied to classify the gene expression data such as gene-activation dependencies for genome analysis (Xu et al. 2002; Boyer et al. 2005). In the field of bioinformatics, gene-activation dependencies are represented by graphs, and thus, graph clustering algorithms are able to extract groups of the genes each which is strongly interacted with the others. For example, Xu et al. applied graph clustering algorithms to a gene expression data set of Arabidopsis and found a known cis-acting element of chitin responsive genes from the dataset (Xu et al. 2002).

As well as gene analysis, protein and protein interactions (PPI) network, which typically contains the key mechanisms determining disease and health, is a representative application of graph clustering in the field of biological networks (Janga and Tzakos 2009; Ding et al. 2012). For example, in order to effectively identify the key functional modules and biomarkers underlying the mechanisms of disease and toxicity, Ding et al. applied graph clustering to PPI networks (Ding et al. 2012). They reported that SCAN could find helpful biomarkers in real-world datasets. For example, Ding et al. applied SCAN to the gene interaction data of breast cancer patients. The clusters, hubs, and outliers obtained by graph clustering were better prognostic biomarkers than the traditional biomakers; the former were more helpful for identifying the patients not needing additional chemotherapy (Ding et al. 2012).

Another attractive application of the graph clustering is to analyze structural and functional connectivity in the human brain (Sporns et al. 2005). The human brain is structurally and functionally organized into a complex network enabling segmentation and integration of information processing. Recent studies have suggested that a combination of MRI techniques together with graph cluster analyses can help us to map structural and functional connectivity patterns of the human brain.

Future Directions

This entry provides an overview of the key techniques used for cluster analysis on large-scale graphs. We investigated 15 algorithms that are grouped in three types: edge-cut based method, modularity-based method, and structural similarity-based method. In the predawn of the graph clustering algorithms, existing algorithms tried to minimize the edges that across different partitions. Then, in order to capture more complicated cluster and community structures hidden in the graphs, modularity-based, and structural similarity-based algorithms were widely developed in the computer science and physics. Our brief survey indicates that the structural similarity-based method now requires high computational cost and they are still difficult to find clusters from billion-nodes graphs. We suggest that developing efficient clustering methods for structural similarity-based algorithms is one of the important future works.

Cross-References

- ▶ Benchmarking for Graph Clustering and Partitioning
- ▶ Community Detection: Current and Future Research Trends
- ▶ Dynamic Community Detection
- ▶ Modularity

References

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech: Theory Exp 2008(10):P10008

Boyer F, Morgat A, Labarre L, Pothier J, Viari A (2005) Syntons, metabolons and interactions: an exact graph-theoretical approach for exploring neighborhood between genomic and functional data. Bioinformatics 21(23):4209–4215

Chen YC, Zhu WY, Peng WC, Lee WC, Lee SY (2014) CIM: community-based influence maximization in social networks. ACM Trans Intell Syst Technol 5(2):25:1–25:31

Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70:066111

Ding CHQ, He X, Zha H, Gu M, Simon HD (2001) A min-max cut algorithm for graph partitioning and data clustering. In Proceedings of the 2001 I.E. international conference on data mining, San Jose, pp 107–114

Ding Y, Chen M, Liu Z, Ding D, Ye Y, Zhang M, Kelly R, Guo L, Su Z, Harris S, Qian F, Ge W, Fang H, Xu X, Tong W (2012) atBioNet – An integrated network analysis tool for genomics and biomarker discovery. BMC Genomics 13(1):1–12

Dinh T, Xuan Y, Thai M (2012) Towards social-aware routing in dynamic communication networks. In Proceedings IPCCC 2009, pp 161–168

Ester M, Kriegel HP, Sander J, Xu X, (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings SIGKDD 1996, pp 226–231

Feng Z, Xu X, Yuruk N, Schweiger TAJ (2007) A novel similarity-based modularity function for graph partitioning. In Proceedings DaWaK 2007, pp 385–396

Fortunato S, Barthélemy M (2007) Resolution limit in community detection. Proc Natl Acad Sci 104(1):36–41

Furuta T, Sasaki M, Ishizaki F, Suzuki A, Miyazawa H (2006) A new cluster formation method for sensor networks using facility location theory. Technical report NANZAN-TR-2006-01

Grout V, Cunningham S, Picking R (2007) Practical large-scale network design with variable costs for links and switches. Int J Comp Sci Netw Secur 7(7):113–125

Huang J, Deng H, Sun H, Sun Y, Han J, Liu Y (2014) SHRINK: a structural clustering algorithm for detecting hierarchical communities in networks. In Proceedings CIKM 2010, pp 219–228

Janga C, Tzakos A (2009) Structure and organization of drug-target networks: insights from genomic approaches for drug discovery. Mol BioSyst 5(12):1536–1548

Karypis G, Kumar V (1998) A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM J Sci Comput 20(1):359–392

Kernighan BW, Lin S (1970) An efficient heuristic procedure for partitioning graphs. Bell Syst Tech J 49(2):291–307

Lee P, Lakshmanan LVS, Milios EE (2014) Incremental cluster evolution tracking from highly dynamic network data. In Proceedings ICDE 2011, pp 3–14

Lim S, Ryu S, Kwon S, Jung K, Lee JG (2014) LinkSCAN∗: overlapping community detection using the link-space transformation. In Proceedings ICDE 2014, pp 292–303

Newman MEJ (2004) Fast algorithm for detecting community structure in networks. Phys Rev E 69:066133

Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113

Nguyen NP, Dinh TN, Xuan Y, Thai MT (2011) Adaptive algorithms for detecting community structure in dynamic social networks. In Proceedings INFOCOM 2011, pp 2282–2290

Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE TPAMI 22(8):888–905

Shiokawa H, Fujiwara Y, Onizuka M (2013) Fast algorithm for modularity-based graph clustering. In Proceedings of the 27th AAAI conference on artificial intelligence, Bellevue, pp 1170–1176

Shiokawa H, Fujiwara Y, Onizuka M (2015) SCAN++: efficient algorithm for finding clusters, hubs and outliers on large-scale graphs. Proc VLDB Endowment 8(11):1178–1189

Sporns O, Tononi G, Kötter R (2005) The Human Connectome: a structural description of the human brain. PLoS Comput Biol 1(4):e42

Sun H, Huang J, Han J, Deng H, Zhao P, Feng B (2014) gSkeletonClu: density-based network clustering via structure-connected tree division or agglomeration,3e4. In Proceedings ICDM 2010, pp 481–490

Toutanova K, Manning CD (2000) Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In Proceedings EMNLP 2000, pp 63–70

Wang L, Xiao Y, Shao B, Wang H (2014) How to partition a billion-node graph. In Proceedings ICDE2014, pp 568–579

Xu Y, Olman V, Xu D (2002) Clustering Gene expression data using a graph-theoretic approach: an application of minimum spanning trees. Bioinformatics 18(4):536–545

Xu X, Yuruk N, Geng Z, Schweiger TAJ (2007) SCAN: a structural clustering algorithm for networks. In Proceedings SIGKDD 2007, pp 824–833

Yuruk N, Mete M, Xu X, Shweiger TAJ (2007) A divisive hierarchical structural clustering algorithm for networks. In Proceedings ICDM Workshops 2007, pp 441–446

Zhou W, Jin H, Liu Y (2012) Community discovery and profiling with social messages. In Proceedings SIGKDD 2012, pp 388–396

Scale-Free Distributions

► Scale-Free Nature of Social Networks

Scale-Free Nature of Social Networks

Piotr Fronczak
Faculty of Physics, Warsaw University of
Technology, Warsaw, Poland

Synonyms

Complex networks; Network models; Scale-free distributions; Universal scaling

Glossary

Degree	The degree of a node in a network is the number of edges or connections to that node
ER graph	The network model in which edges are set between nodes with equal probabilities
Fat-tailed distributions	Have tails that decay more slowly than exponentially. All power-law distributions are fat tailed, but not all fat-tailed distributions are power laws (e.g., the lognormal distribution is fat tailed but is not a power-law distribution)
Node degree distribution	The distribution function $P(k)$ that gives the probability that a node selected at random has exactly k edges
Power-law distribution	Has a probability function of the form $P(x) \sim x^{-a}$
Scale-freeness	Feature of objects or laws that does not change if length scale is multiplied by a common factor; also known as scale invariance
SF network	The network with power-law distribution of node degrees

Definition

The notion of scale-freeness and its prevalence in both natural and artificial networks have recently attracted much attention. In physics and mathematics, scale-freeness (or more formally – scale invariance) is a feature of objects or laws that does not change if length scale is multiplied by a common factor. The term gained large popularity in 1999 when Barabasi and Albert used it as a descriptor of networks in which node degrees (vertex connectivity) follow a power-law distribution (Barabasi and Albert 1999). Since most large complex networks are characterized by the distributions which at least partially are reminiscent of power functions, the term "scale-free," applied to networks, lost its formal meaning and nowadays is widely used (albeit erroneously) to describe the network with fat-tailed node degree distribution.

The overwhelming number of studies conducted in the last decade made it clear that the scale-free network topology can have a strong impact on the dynamical processes taking place on these networks such as opinion formation (Aleksiejuk et al. 2002), diffusion of information (Cohen et al. 2000), and epidemic spreading (Pastor-Satorras and Vespignani 2001). Nowadays, the recently acquired knowledge about the network structure revolutionizes not only many fields of science, like biology, computer science, and economics, but also the society and its perception of the ubiquitous networks.

Introduction

Scale-freeness is the property which is fascinating especially for physicists, since most phenomena studied by physicists are not scale invariant. Among seminal exceptions are phase transitions in thermodynamic systems which are associated with the emergence of power-law distributions of certain quantities (Yeomans 2002). Similarly, the phenomenon known as self-organized criticality (a property of dynamical systems which have a critical point as an attractor) displays the spatial and/or temporal scale-free nature of the critical point of a phase transition, but without the need to tune control parameters to precise values (Bak 1996).

In mathematics, scale invariance is an exact form of self-similarity where at any magnification there is a smaller piece of the object that is similar to the whole. Self-similarity is a typical property of fractals.

A common aspect of both, phase transitions and self-similar fractals, is universality, i.e., the observation that there are properties for a large class of different systems that are independent of the dynamical details of the particular system.

These reasons (universality and criticality) explain the excitement of scientists, when the power-law character of node degree distribution has been observed in drastically increasing number of real networks. The promise of the discovery of the universal character of surrounding us social, technological, and natural networks made the notion of scale-freeness frequently misused. Nevertheless, it is a notion that has clearly taken root with today's society effectively guiding the communicative patterns of different scientific communities. In the following paragraphs, we will use this notion in its less formal meaning as a shortcut of the networks with fat-tailed (or almost power-law) distribution of node degrees.

Despite the pure mathematical differences, the properties of idealized (scale-free) and realistic (almost scale-free) networks have the same implications for real-world applications.

Key Points

To understand the scale-free architecture of the networks, it is useful to contrast it with the other network model which dominated the network research for decades, namely, the model of network developed by Erdos and Renyi in 1959 (ER graph) (Erdos and Renyi 1960). The importance of the ER graph for modeling real-world networks is currently diminished; however, it is still a fundamental model in the random network theory. In the following we will briefly introduce ER graphs and emphasize differences between them and SF networks. We will present the methods of detection of the scale-free character of the node degree distribution in networks. We will discuss the most popular model in which the growing network evolves into scale-free state. Finally, we will discuss the vulnerability of SF networks to epidemics and intentional attacks and their extreme tolerance on random failures.

Historical Background

Power-law distributions in nature and society were already known in the nineteenth century. Italian economist, Vilfredo Pareto, in 1897, was the first to discover that the distribution of income in society follows the power law (Barabasi 2002). In 1925, George Udny Yule proposed a stochastic process (later called the Yule process but is now better known as preferential attachment – see the next section) that leads to a distribution with a power-law tail – in this case, the distribution of species and genera (Yule 1925). In 1965, Derek John de Solla Price demonstrated a power-law distribution of links in a network of scientists linked by citation (Price 1965). Although D. Price was a physicist, his discovery was totally ignored in physical sciences. In physics, the lattices and random networks like ER graph were the main objects of study until late 1990s, when Barabasi and others rediscovered the importance of SF networks in technology, nature, and society.

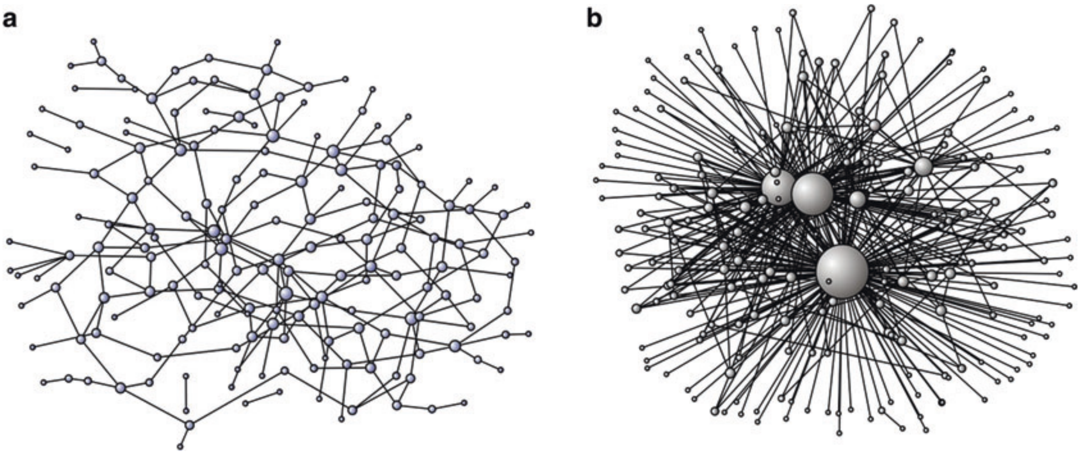
Properties of Scale-Free Networks

Two Opposing Models of Random Networks

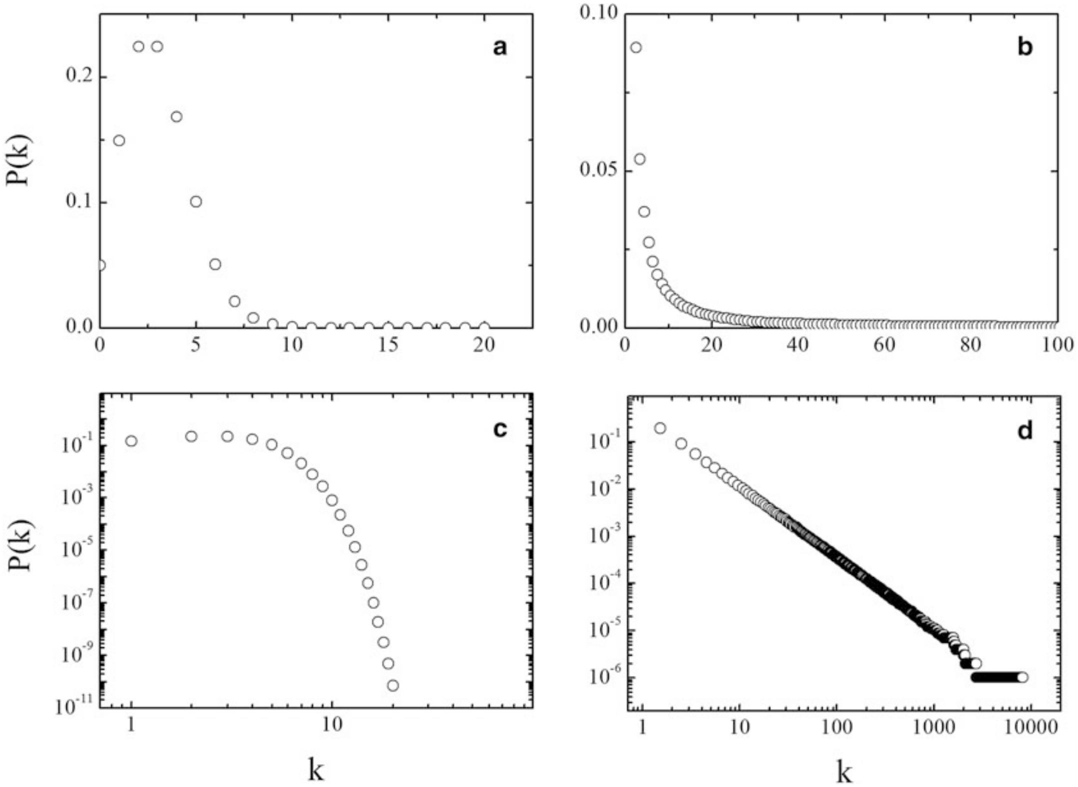
The definition of ER graph is simple: in a graph with N nodes, each possible pair of distinct nodes is connected with an edge with probability p. In that model the majority of nodes have a degree that is close to the average degree of the whole network, and this average has small variance (the number of nodes with a given degree decays exponentially fast away from the mean degree). In Fig. 1a, we show a typical representative of this model. As one can see, the sizes of all nodes (which reflect node degrees) are similar. For large N and infinitesimal p (i.e., for large and sparse networks), the node degree distribution follows a Poisson law

$$P(k) = e^{-pN} \frac{(pN)^k}{k!},$$

where k is a node degree and the average node degree $\langle k \rangle = pN$ (Newman et al. 2002). The characteristic bell shape of *P(k)* around the average node degree is visible in Fig. 2a.



Scale-Free Nature of Social Networks, Fig. 1 Two realizations of a ER graph (**a**) and SF network (**b**), both with the same number of nodes and edges. Size of the nodes is proportional to their degrees



Scale-Free Nature of Social Networks, Fig. 2 Node degree distributions of ER graph (*left column*) and SF network (*right column*) in normal (*top row*) and double logarithmic (*bottom row*) scale

As we stated previously, recent studies show that most large complex networks are characterized by a connectivity distribution different to a Poisson distribution (among the exceptions are train networks or electrical power grids). For example, the World Wide Web, Internet, e-mail, and collaboration networks have a degree distribution that follows (at least in some range) a power-law relationship defined by

$$P(k) \sim k^{-\gamma},$$

where γ, called *scale-free exponent*, ranges usually between 2 and 3 in real networks. Such networks have a very uneven distribution of connections. There are many nodes with only a few links and a few nodes with a large number of links. The difference between this type of network and a Poisson-like one is clearly visible in Fig. 1b, where some nodes act as "highly connected hubs" while the most of them have only one connection. The fat tail of this distribution, shown in Fig. 2b, is an evidence of an extreme heterogeneity of connections in the network.

Scale-Freeness of Networks with Power-Law Distribution of Node Degrees

Why the networks with the power-law distributions of node degrees are called scale-free? It is because a power-law distribution is scale invariant. If we rescale a measure of connectivity (e.g., counting how many tens of connections a node has, instead of counting all its connections), the connectivity distribution $P(10\ k)$ will be still proportional to the original *P(k)*. Mathematically, multiplying degree k by a constant c, the distribution remains the same and only scales the function: $P(ck) = c^{-\gamma}P(k)$, where $P(k) = ck^{-\gamma}$. To show that power-law distribution is the only one, which fulfills this condition, we take the logarithm of its both sides:

$$\ln\ P(ck) = -\gamma\ \ln\ c + \ln\ P(k)$$

Now, we introduce a new function $R(k)$ defined as $R(\ln\ k) = P(k)$. This gives

$$\ln\ R(\ln\ ck) = -\gamma\ \ln\ c + \ln\ R(\ln\ k)$$

and, after rearrangement,

$$\frac{\ln\ R(\ln\ c + \ln\ k) - \ln\ R(\ln\ k)}{\ln\ c} = -\gamma.$$

In the limit $\ln c \rightarrow 0$, the left side becomes a derivative

$$\frac{d\ln R(\ln k)}{d\ln k} = -\gamma.$$

Since the right side is constant, integrating the equation gives

$$\ln R(\ln k) = -\gamma \ln k + \text{const},$$

and finally

$$R(\ln k) = P(k) = \text{const} \cdot k^{-\gamma}.$$

An important difference between fat-tailed and Poisson-like distributions is that moments of the former (i.e., mean $\langle k \rangle$ and variance $\delta^2(k)$) poorly characterize the distribution (in fact, they are undefined for certain power-law distributions). The moments μ_{m} of order m are defined as follows:

$$\mu_m = \sum\nolimits_{k=0}^{k_{\max}} k^m P(k).$$

From the definition, in infinite networks (i.e., when $k_{\max} \rightarrow \infty$), all higher moments of order $m \geq \gamma - 1$ of the power-law distribution diverge. Since a mean and a variance are the moments of the first and the second order, respectively, a variance is infinite for γ in the range of typical real-world networks ($2 \leq \gamma \leq 3$):

$$\delta^2 = \mu_2 \sim \sum\nolimits_{k=0}^{\infty} k^2 k^{-\gamma} = \infty, \quad \text{for}\ \ \gamma \leq 3.$$

Although the real networks are finite, the variance can be still several orders larger than the mean. Since a variance describes the error of measured mean node degree, its enormously large value questions the quality of the measurement, and assigning the scale (related to the mean degree) to the network is a misuse.

Plotting Scale-Free Distributions

Since many fat-tailed distributions look similarly as in Fig. 2b (e.g., log-normal or stretched exponential distributions), to better expose the power-law nature of the node degree distribution, one usually plots the data on a double logarithmic scale. In that case the power law transforms into a straight line with a slope of $-\gamma$ (see Fig. 2d and compare it with Poisson distribution shown in Fig. 2c), as follows:

$$\begin{aligned} P(k) &= a \cdot k^{-\gamma} \\ \ln P(k) &= \ln\left(a \cdot k^{-\gamma}\right) \\ \ln P(k) &= \ln a + \ln\left(k^{-\gamma}\right) \\ \ln P(k) &= \ln a - \gamma \ln k \\ Y &= A - \gamma X \end{aligned}$$

where X and Y are transformed variables and A is a transformed constant.

In practice, measuring a slope directly from Fig. 2d is usually very erroneous, due to the poor statistics at the tail of the distribution. Direct histograms are almost always noisy in this region. The solution is to construct a histogram in which the bin sizes increase exponentially with degree. The number of samples in each bin is then divided by the width of the bin to normalize the measurement. Plotting histogram in a logarithmic degree scale, one obtains the even widths of the bins.

An even more discriminating method to verify potential power-law character of the node degree distribution is to plot the complementary cumulative distribution function

$$P_C(k) = \int_k^\infty P(k)dk \sim k^{-(\gamma+1)},$$

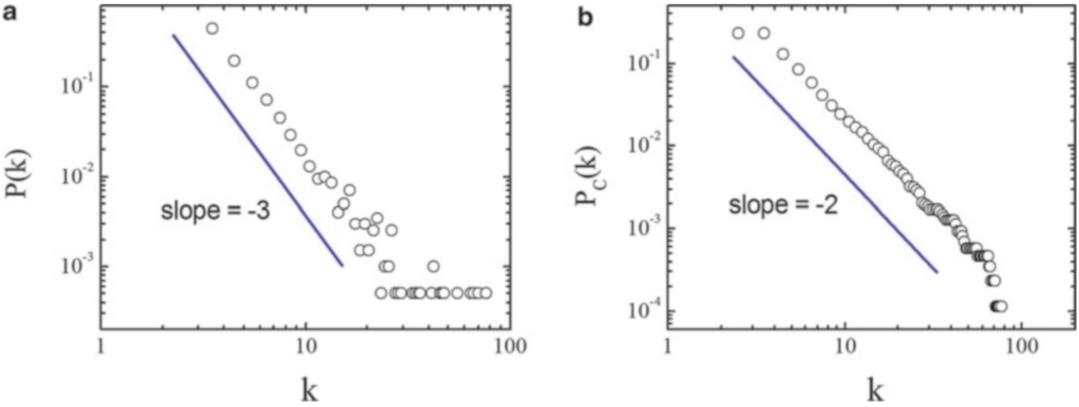
which is the probability that the degree of a randomly chosen node is greater than or equal to k. Such a plot has the advantage that all the original data are represented. When we make a conventional histogram by binning, any differences between the values of data points that fall in the same bin are lost. The cumulative distribution function does not suffer from this problem. The cumulative distribution also reduces the noise in the tail, which is clearly illustrated in Fig. 3.

Seminal Model of Preferential Attachment

Soon after the discovery of the scale-free structure of the World Wide Web, it has been realized that many other real networks also show power-law distribution of node degrees. This feature has been observed in the Internet, communication, and transportation networks (Albert et al. 1999; Guimera et al. 2005), as well as in many social networks, such as networks of scientific citations (Redner 1998), e-mail networks (Ebel et al. 2002), or even sexual contact networks (Liljeros et al. 2001). The initial surprise of omnipresence of SF networks quickly turned into a question: Why so many networks have the same scale-free character of connections? When a feature appears in many systems that do not have an obvious connection to each other, you should suspect that there is a common causal principle, which can be described in the most general terms, without reference to the details of this or any other system. Is a scale invariance of complex networks a result of some universal rules that govern the dynamics of these systems?

Although there are many different processes which can give rise to the same power-law structure of complex networks, the one deserves particular attention at least for the two reasons. Firstly, its universal character allows to adapt the process to many social but also technological and natural networks. Secondly, it has been independently rediscovered several times in different fields and ages. The process is currently known as Matthew effect, Yule process, Dulbecco's law, Rich gets richer, or preferential attachment (Barabasi and Albert 1999). Since it is used so widely across domains, the claim about its universality is reasonable.

The process, adopted to networks, comprises of two complementary mechanisms: network growth and preferential rules of joining nodes. Barabasi and Albert, who introduced the process to the modern science of complex networks, stated that real networks are not formed as a result



Scale-Free Nature of Social Networks, Fig. 3 Node degree distribution of SF network (**a**) and its cumulative distribution (**b**)

of purely random process, in which completely randomly selected nodes are connected by the edges. Most of the social and technological networks grow and change over time – they evolve. In networks, the newly added nodes prefer to create connections with such ones that already have a lot of other connections. The mechanisms underlying this preference can be different. For example, new actors are more likely to play supporting roles in films with established stars, than in those where there are only other unknown actors. Thus, the more famous you are, the more probably that you will attract new connections. The same principle seems to govern the structure of citation network. Preferential attachment corresponds to the feature that a publication with a large number of citations continues to be well cited in the future merely by virtue of being well-cited now. In the network of acquaintances, my friends introduce me to their friends. The more friends I have, the more recognized I am and the more chances to meet new people I have. In WWW, the more pages linked to a web page, the more Internet users visit that site and the greater the likelihood that they will place a link to this page on their own website.

The algorithm of the discussed process consists of two steps:

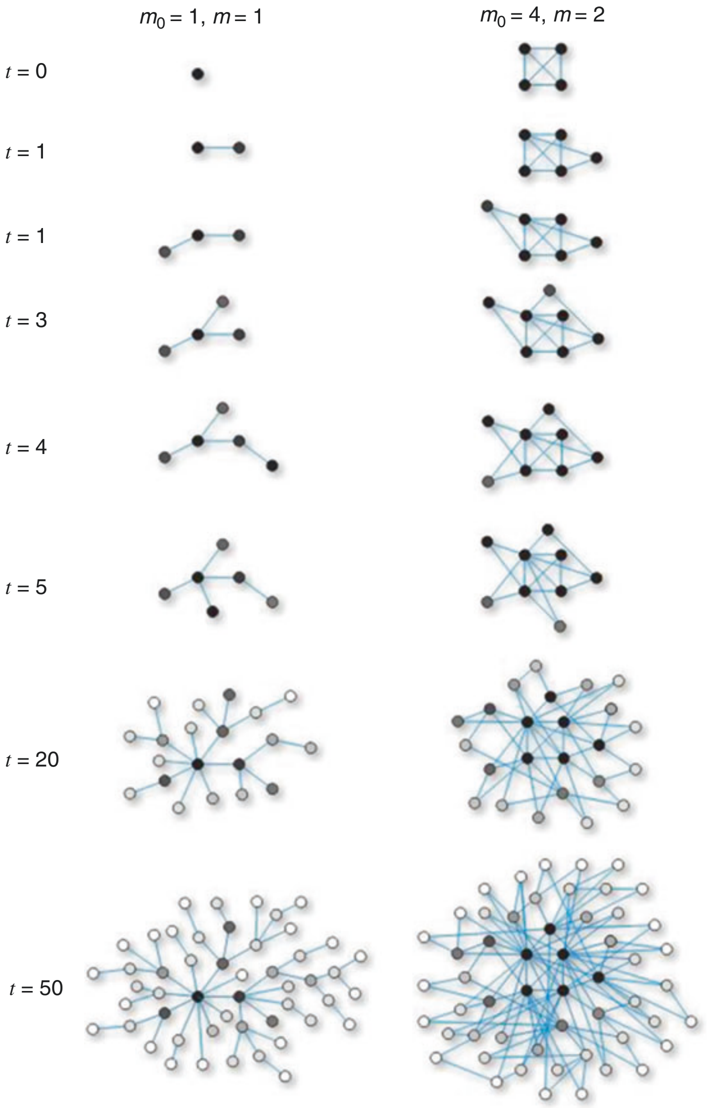
1. Starting with a small number m_0 of nodes, at every time step, we add a new node with $m \leq m_0$ edges that link the new node to m different nodes already present in the system.
2. When choosing the nodes to which the new node connects, we assume that the probability $\prod$ that a new node will be connected to node i depends on the degree k_i of node i, such that

$$(k_i) = \frac{k_i}{\sum_j k_j}.$$

After t time steps, this algorithm results in a network with $N = t + m_0$ nodes and mt edges. In Fig. 4 first steps of the network evolution have been shown. Already after several steps, the hubs in the network become clearly visible.

Mathematical derivations show that the node degree distribution of the network evolves into a scale-free one with the scale-free exponent $\gamma = 3$ independently of m, the only parameter in the model.

One has to keep in mind that the presented model does not share all properties observed in the real-world networks, e.g., it is less clustered. Soon, after this model was introduced, a large number of similar models, all based on some type of connecting preference, emerged, all leading to a power-law distribution of node degrees but also demonstrating a better agreement with real networks with reference to other network metrics.

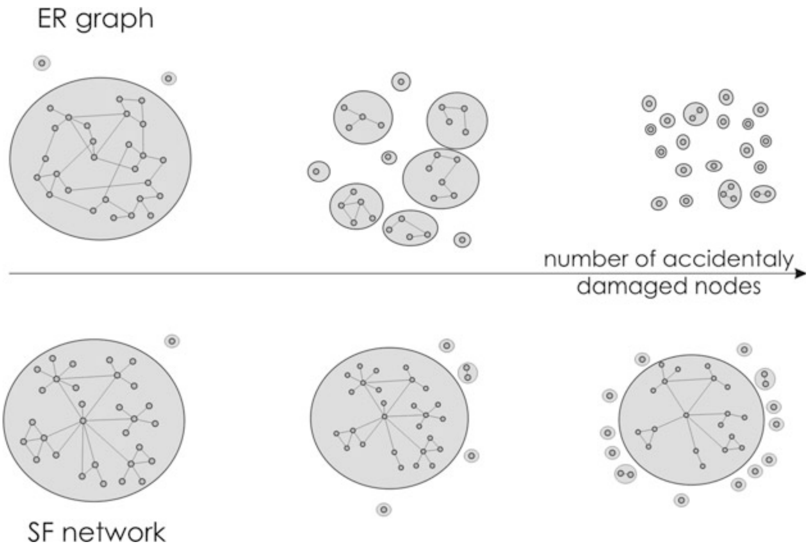


Scale-Free Nature of Social Networks, Fig. 4 Example of realization of two different growing networks in preferential attachment model. The *colors* of the nodes represent their age

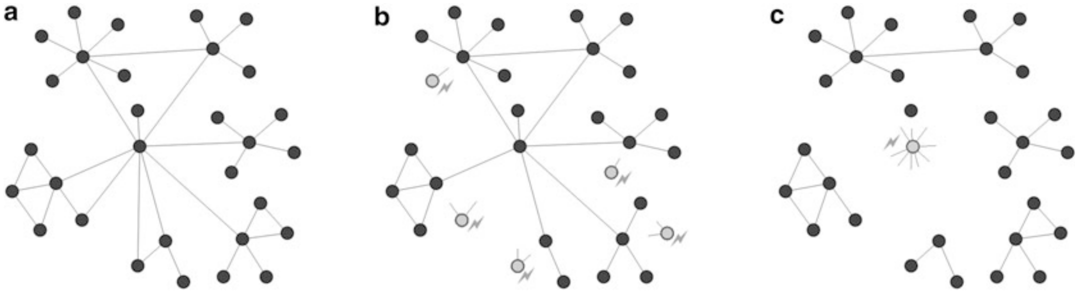
Preferential attachment is not the only possible explanation for the formation of scale-free structure of connections in complex networks. Among others there are rewiring processes (Aiello et al. 2002), optimization-based models (Valverde et al. 2002), and also static constructions (Park and Newman 2004; Goh et al. 2001).

Resilience and Vulnerability of SF Networks to Failures and Attacks

It has appeared recently that scale-free structure of complex networks has an important influence on their resilience to failures and attacks. In particular, SF networks seem much more robust than ER graphs in case of failures (modeled by a random removal of nodes or links) (Cohen et al.



Scale-Free Nature of Social Networks, Fig. 5 ER graphs break apart into small disconnected parts much faster than SF networks if the nodes are removed accidentally



Scale-Free Nature of Social Networks, Fig. 6 Random and targeted elimination of nodes. Original SF network (**a**), the network with randomly damaged nodes (**b**), and the network with the damaged hub (**c**)

2000), while they are more sensitive to attacks (modeled by the targeted removal of selected nodes or links) (Cohen et al. 2001). By the resilience we understood that despite removed nodes the main part of the network (so called giant component) is still interconnected (i.e., any two nodes in that part are connected to each other by paths). If the node elimination proceeds, then at some critical moment, the network breaks apart into small disconnected parts. The moment when this dramatic breakdown occurs strongly depends on the network structure as well as on the method of node's elimination (random or targeted). In the random removal case, the critical moment of destruction occurs much earlier in ER graphs, in opposite to SF networks (see Fig. 5). It means that SF network is much more resilient to accidental damages. However, in case of intentional attack, when the nodes of the network are removed in decreasing order of their degree, SF network appears to be much more vulnerable than ER graph (since the removal of the hubs results in the largest possible damage; see Fig. 6). This vulnerability of SF networks to intentional attacks has been described as their Achilles' heel.

Epidemic Spreading in SF Networks

Scale-free nature of social networks has a great implication for understanding the spread of information, diseases, opinions, and innovations in society. Standard epidemiological models usually consider networks with the well-defined average node degree, such as ER graphs. In those networks the models predict a critical threshold for the propagation of a contagion throughout a population. This epidemic threshold is determined by the virulence of the infection. In other words, if the spreading rate is larger than the threshold, the infection spreads and becomes persistent. Below the threshold, the infection dies out.

It turns out that in SF networks the above statement is no longer correct. In 2001, Pastor-Satorras and Vespignani found that in that case the threshold is zero (Pastor-Satorras and Vespignani 2001). It means that all viruses, even those that are weakly contagious, will spread and persist in the system. The main reason is that the presence of hub nodes can facilitate epidemic spreading due to the large numbers of neighbors. Infected hub passes the infection to numerous other nodes, faster than the typical node recovers.

Specifically, in SF network, the traditional random immunization could easily fail because nearly everyone would have to be treated to ensure that the hubs were not missed. New immunization strategies have to be developed to recover the epidemic threshold. It turns out that one of the most efficient approaches is to selectively immunize hub nodes. Such a strategy is known as targeted immunization (Pastor-Satorras and Vespignani 2002).

Key Applications

The paradigm of SF networks has numerous applications to problems in different areas including epidemiology (Colizza et al. 2007), human mobility (Gonzalez et al. 2008), social networks (Huberman et al. 2009), life sciences (Guimera and Nunes Amaral 2005), information flow (Helbing et al. 2006), and ecology (Montoya et al. 2006).

Future Directions

The structure, topological properties, and appropriate measures were the main research topics in complex network domain in recent years. Currently, dynamical processes taking place in the networks are quite intensively studied. It is believed that further understanding of dynamics on complex networks is the general direction of the field. There is a continuous shift from studies of networks in general and features that are common to most of them to more application-driven studies of increasingly narrow classes of networks. After a decade of mostly descriptive studies and just potential applications, there is a final need to transfer an acquired knowledge into concrete market applications. Complex networks research society should provide the manageable solutions to global challenges, like vaccination campaigns against serious viruses, risk reduction of financial crises, and preventing cascading bankruptcies among interlinked economies.

Cross-References

- ► Exponential Random Graph Models
- ► Network Models

References

Aiello W, Chung F, Lu L (2002) Random evolution of massive graphs. In: Pardalos PM, Abello J, Resende MGC (eds) Handbook of massive data sets. Kluwer, Dordrecht, pp 97–122

Albert R, Jeong H, Barabasi A-L (1999) Diameter of the World Wide Web. Nature 401:130–131

Aleksiejuk A, Holyst JA, Stauffer D (2002) Ferromagnetic phase transition in Barabasi-Albert networks. Phys A 310:260–266

Bak P (1996) How nature works: the science of self-organized criticality. Copernicus, New York

Barabasi A-L (2002) Linked: the new science of networks. Perseus Books Group, Reading

Barabasi A-L, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512

Cohen R, Erez K, ben Avraham D, Havlin S (2000) Resilience of the Internet to random breakdowns. Phys Rev Lett 85:4626

Cohen R, Erez K, ben Avraham D, Havlin S (2001) Breakdown of the Internet under intentional attack. Phys Rev Lett 86:3682
Colizza V, Barrat A, Barthelemy M, Valleron AJ, Vespignani A (2007) Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. PLoS Med 4:e13
Ebel H, Mielsch LI, Bornholdt S (2002) Scale-free topology of e-mail networks. Phys Rev E 66:035103
Erdos P, Renyi A (1960) On the evolution of random graphs. Pub Math Inst Hung Acad Sci 5:17–61
Goh K-I, Kahng B, Kim D (2001) Universal behaviour of load distribution in scale-free networks. Phys Rev Lett 87:278701
Gonzalez MC, Hidalgo CA, Barabasi AL (2008) Understanding individual human mobility patterns. Nature 453:779–782
Guimera R, Nunes Amaral LA (2005) Functional cartography of complex metabolic networks. Nature 433 (7028):895–900
Guimerà R, Mossa S, Turtschi A, Amaral LAN (2005) The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. Proc Natl Acad Sci 102:7794–7799
Helbing D, Armbruster D, Mikhailov AS, Lefeber E (2006) Information and material flows in complex networks. Physica A 363:11–16
Huberman BA, Romero DM, Wu F (2009) Social networks that matter: twitter under the microscope. First Monday 14:1–5
Liljeros F, Edling CR, Amaral LAN, Stanley HE, Aberg Y (2001) The web of human sexual contacts. Nature 411:907–908
Montoya J, Pimm SL, Solé RV (2006) Ecological networks and their fragility. Nature 442(7100):259–264
Newman ME, Watts DJ, Strogatz SH (2002) Random graph models of social networks. Proc Natl Acad Sci U S A 99(Suppl 1):2566–2572
Park J, Newman MEJ (2004) Statistical mechanics of networks. Phys Rev E 70:066117
Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. Phys Rev Lett 86:3200
Pastor-Satorras R, Vespignani A (2002) Immunization of complex networks. Phys Rev E 65:036104
Price DJ d S (1965) Networks of scientific papers. *Science* 149:510–515
Redner S (1998) How popular is your paper? An empirical study of the citation distribution. Eur Phys J B 4:131–134
Valverde S, Ferrer Cancho R, Solé RV (2002) Scale-free networks from optimal design. Europhys Lett 60:512
Yeomans JM (2002) Statistical mechanics of phase transitions. Oxford University Press, New York
Yule UG (1925) A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R. S. Philosophical transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character, vol 213. The Royal Society, pp 21–87

Scaling Subgraph Matching Queries in Huge Networks

Matthias Brücheler[1], Andrea Pugliese[2] and
V. S. Subrahmanian[1]
[1]Computer Science Department, University of Maryland, College Park, MD, USA
[2]DIMES Department, University of Calabria, Rende, Italy

Synonyms

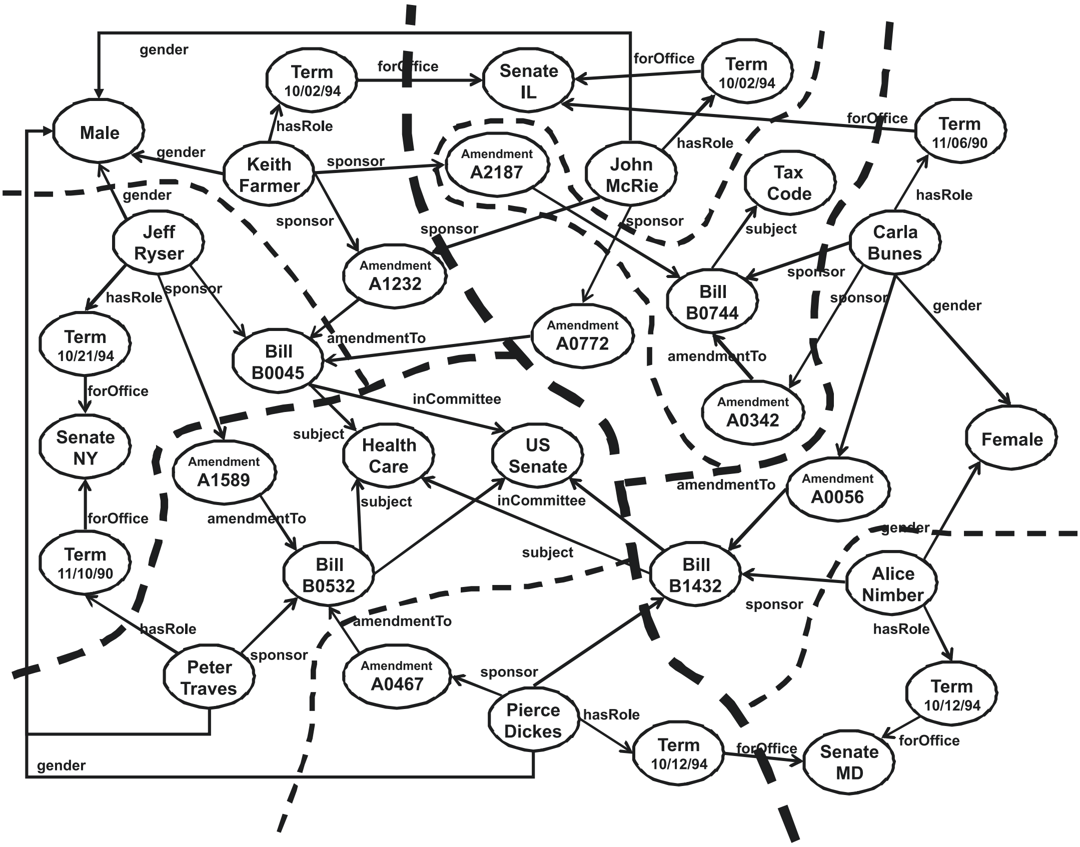
Graph matching; Subgraph identification; Subgraph isomorphic queries

Glossary

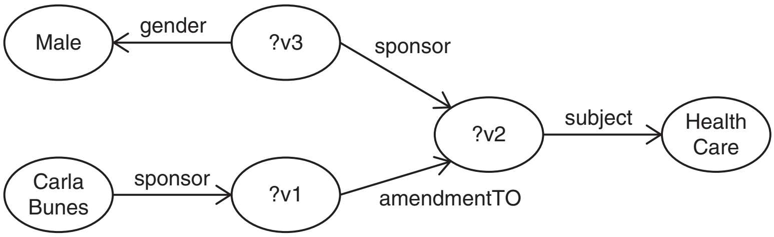
COSI	Cloud-Oriented Subgraph Identification
DOGMA	Disk-Oriented Graph Matching Algorithm
RDF	Resource Description Framework
SPARQL	SPARQL Protocol and RDF Query Language

Introduction

Both social network owners and social network users are interested in a variety of queries that involve subgraph matching. In addition, answering SPARQL queries in the Semantic Web's RDF framework largely involves subgraph matching. For example, the GovTrack dataset (2013) tracks events in the US Congress. In Fig. 1, we see that Jeff Ryster sponsored Bill B0045 whose subject is Health Care. A user who is using such a database might wish to ask queries such as that shown in Fig. 2. This query asks for all amendments (? $v1$) sponsored by Carla Bunes to bill (? $v2$) on the subject of health care that were originally sponsored by a male person (? $v3$). The reader can readily see that when answering this query, we want to find *all* matches for this query graph in the



Scaling Subgraph Matching Queries in Huge Networks, Fig. 1 Example graph



Scaling Subgraph Matching Queries in Huge Networks, Fig. 2 Example query

original graph. The reader who tries to answer this very simple query against this very tiny graph will see that it takes time to do so, even for a human being! In this entry, we show how to answer complex subgraph matching queries over huge graphs efficiently.

An important aspect of all past work is that it has focused on working in memory. Though memory prices are dropping while capacity is increasing, the increase in capacity is not even remotely large enough to store 1% of **Facebook** or **Twitter**. As a consequence, in order to efficiently answer queries to social network graphs, we are forced to store the data, as well as indexes for the data, on disk. In section "The DOGMA Index" we provide a description of the DOGMA (Disk-Oriented Graph Matching Algorithm) index (Brücheler et al. 2009) for building a disk-based index for huge networks. DOGMA is based on a simple observation: the size of any real-world social network graph is likely to be orders of magnitude larger than that of any subgraph matching query graph a user is likely to ask. This tells us that it should be possible to build an index for efficiently executing such queries that ensures that vertices in a social network graph that are "near" each other be stored together on a disk page.

Then, in section "Cloud-Oriented Subgraph Matching," we present the COSI (Cloud-Oriented Subgraph Identification) system (Brücheler et al. 2010). COSI distributes a graph across multiple compute machines and answers subgraph matching queries in parallel using an asynchronous query-answering algorithm that does not rely on central orchestration. Thus, computation is completely distributed and our goal is to minimize communication between compute machines so as to save time. Finally, section "Experimental Results" shows some experimental results assessing the performance of **DOGMA** and **COSI**, and section "Related Work and Conclusions" briefly discusses related work and outlines conclusions.

Basic Notation

Throughout this entry, we assume the existence of an arbitrary but fixed set V whose elements are called vertices. For example, V might consist of all strings that can form a valid user ID and/or the set of all valid identifiers for comments in a social network like **Facebook**. We also assume the existence of a finite set $\mathcal{P}$ of predicate symbols.

We model a social network graph S as a triple (V, E, λ) where V is the set of vertices, $E \subseteq V \times V$ is a *multiset* of edges from vertices to vertices, and $\lambda: E \rightarrow \mathcal{P}$ assigns a predicate symbol to each edge in E.

The *out neighborhood* of vertex v is the set out $(v) = \{u| \ (v, u) \in E\}$; the *in neighborhood* of node v is the set in $(v) = \{u| \ (u, v) \in E\}$. The neighborhood of v is the set ngh$(v) =$ out$(v) \cup$ in (v). Each of these neighborhoods can be restricted to a particular predicate symbol p: for example, $\text{out}_p(v) = \{u| \ (v, u) \in E \wedge \lambda(v, u) = p\}$.

When formulating queries, we assume the existence of a set **VAR** of variable symbols ranging over V. Each variable symbol starts with a?. A *query* Q is a triple $V_q\mathrm{E}_q$; q where $V\,q \subseteq V \cup$ **VAR**, $E\,q \subseteq V_{\,q} \times V\,Q$ is a multiset of edges, and λ Q: $E_{\,q} \rightarrow \mathcal{P}$. We use **VAR** Q to denote the set of variable vertices in query Q.

Suppose S is a social network graph and Q is a query. A *substitution* for query Q is a mapping **VAR** $Q \rightarrow V$. If θ is a substitution for query Q, then $Q\,\theta$ denotes the replacement of all variables? v in $V\,Q$ by $\theta(?\,v)$. Hence, the graph structure of $Q\theta$ is exactly like that of Q except that nodes labeled with variables are replaced by vertices in S. A substitution θ is an *answer* for query Q w.r.t. S iff $Q\theta$ is a subgraph of S. The *answer set* for query Q w.r.t. a S is the set $\{\theta|\, Q\theta \text{ is a subgraph of } \mathcal{S}\}$. For example, the substitution θ such that $\theta(?\,v1) =$ Amendment A0056, $\theta(?\,v2) =$ Bill B1432, and $\theta(?\,v3) =$ Pierce Dickes is the only answer for the query in Fig. 2.

The DOGMA Index

In this section, we define the **DOGMA** index and describe an algorithm to take an existing social network graph and create the **DOGMA** index for it. Then, we describe algorithms to answer subgraph matching queries.

Before we define the **DOGMA** index, we first define what it means to merge two graphs. Suppose $G = (V, E, \lambda)$ is a graph, and $G_1 = (V_1; E_1; \lambda_1)$ and $G_2 = (V_2; E_2; \lambda_2)$ are two graphs such that V_1, $V_2 \subseteq V$and $k >$ is an integer such that $k \leq \max(|V_1$, $|V|)$ Graph $G_{\,\mathrm{m}}(V_\mathrm{m}, E_\mathrm{m}, \lambda_\mathrm{m})$ is said to be a *k-merge of graphs* G_1, G_2 w.r.t. G iff: (i) $|V_\mathrm{m}| = k$; (ii) there is a *surjective* (i.e., onto) mapping $\mu: V_1 \cup V_2 \rightarrow V_\mathrm{m}$ called the *merge mapping* such that for all $v \in V_\mathrm{m}$, $\text{rep}(v) = \{v' \in V_1 \cup V_2|\ \mu(v') = v\}$, and $e_\mathrm{m} = (v_1, v_2) \in E_\mathrm{m}$ if there exist $v'_1 \in \text{rep}(v_1), v'_2 \in \text{rep}(v_2)$ such that $e = \left(v'_1, v'_2\right) \in E$. The basic idea tying k-merges to the **DOGMA** index is that we want **DOGMA** to be a binary tree, each of whose nodes occupies a disk page. Each node is labeled by a graph that "captures" its two children in some way. As each page has a fixed size, the number k limits the size of the graph so that it fits on one page. The idea is that if a node N has two children,

N_1 and N_2, then the graph labeling node N should be a k-merge of the graphs labeling its children.

A **DOGMA** index for a social network graph S is a generalization of the well-known binary tree specialized to represent social network graphs in the following manner.

Definition 1 *A DOGMA index of order k* ($k \geq 2$) is a binary tree **D** s with the following properties:

1. Each node in **D** s equals the size of a disk page and is labeled by a graph.
2. **D** s is balanced.
3. The labels of the set of leaf nodes of **D** s constitute a partition of S.
4. If node N is the parent of nodes N_1, N_2, then the graph G_N labeling node N is a k-merge of the graphs G_{N1}, G_{N2} labeling its children.

Note that a single social network database can have many **DOGMA** indexes.

Example 2 Suppose $k = 4$. A **DOGMA** index for the graph of Fig. 1 might split the graph into the eight components indicated by dashed lines in Fig. 1 that become the leaf nodes of the index (Fig. 3). Consider the two leftmost leaf nodes. They can be 4-merged together to form a parent node. Other leaf nodes can also be merged together (the results of k-merging are not shown in the inner nodes).

Building DOGMA Indexes

Even though many different **DOGMA** indexes can be constructed for the same social network graph, we want to find a **DOGMA** index with as few "cross" edges between subgraphs stored on different pages as possible. In other words, if node N is the parent of nodes N_1, N_2, then we would like relatively fewer edges in S between some node in GN_{N1} and some node in G_{N2}. The smaller this number of edges, the more "self-contained" nodes N_1, N_2 are and the less likely that a query will require looking at both nodes N_1 and N_2. **DOGMA** can employ any external graph partitioning algorithm (many of which have been proposed in the literature) that, given a weighted graph, partitions its vertex set in such a way that (i) the total weight of all edges crossing the components is minimized and (ii) the accumulated vertex weights are (approximately) equal for both components. In our implementation, we employ the *GGGP* graph partitioning algorithm proposed in Karypis and Kumar (1999).

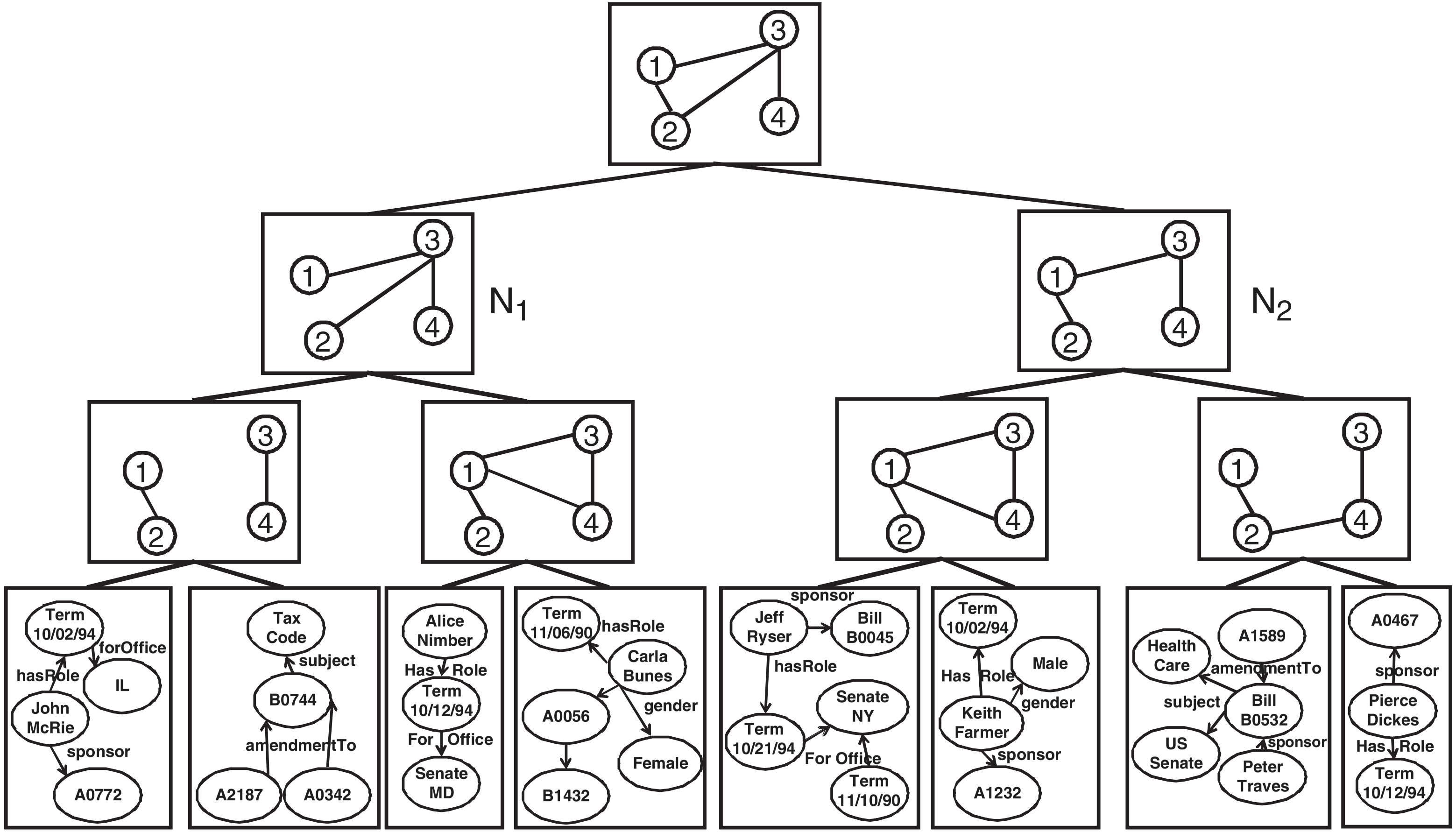
In order to generate a **DOGMA** index for a social network S, we can intuitively proceed through the two following phases (the fully detailed version of the algorithm can be found in Brücheler et al. (2009)).

Iterative Coarsening. Iteratively "coarsen" S by merging nodes in S. This generates a sequence of social network graphs $S = S_0, S_1, \ldots, S_k$ where S_{i+1} is obtained by randomly merging nodes (and corresponding edges) in S_i till the number of vertices in $S_{i\ +1}$ is less than or equal to half of those in S_i. We stop when we reach the smallest m such that the set V_m of vertices associated with S_m is small enough to fit on a disk page. Thus, m is proportional to $O(\log_2 (|V|))$. When constructing V_{i+1} (and the corresponding E_{i+1}) from V_i and E_i, respectively, we keep track of which vertices (resp. edges) in V_i (resp. E_i) were merged into which vertices (resp. edges) in V_{i+1} (resp. E_{i+1}). The "root" of the **DOGMA** index now corresponds to $S\ m$ which, implicitly, represents the entire S.

Hierarchical Decomposition. We now decompose S_m (the root) into two to get S_m 's two children, using any standard graph partitioning algorithm. Suppose this partitioning splits S_m into S_m^1 and S_m^2. We then go back and see which vertices in S_{m-1} got merged into the vertices in S_m^1 and replace the (merged vertices) in S_m^1 by the two vertices from which the merged vertex got created. We repeat this for S_m^2. This now gives us the two children of the root of the **DOGMA** index for social network database S. This process is applied iteratively till we reach the leaf level of the **DOGMA** index (we will know when to stop because the vertices in the index can no longer be "unfolded").

Processing Queries with DOGMA

The basic query answering for answering graph matching queries using the **DOGMA** index, called DOGMA_basic (Brücheler et al. 2009), is



Scaling Subgraph Matching Queries in Huge Networks, Fig. 3 A DOGMA index for the graph of Fig. 1

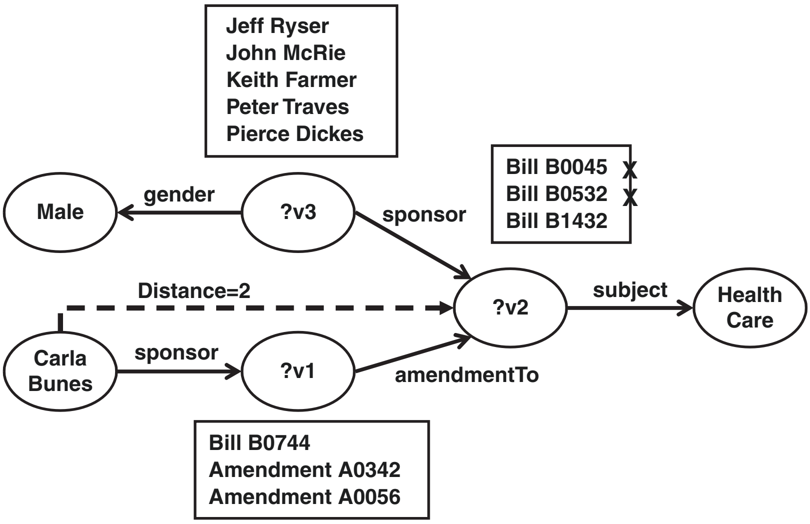
a recursive, depth-first algorithm which searches the space of all substitutions for the answer set to a given query Q w.r.t a graph S. For each variable vertex v in Q, the algorithm maintains a set of constant vertices $R\ v \subseteq V\ S$ (called result candidates) to prune the search space; for each answer substitution θ for Q, we have $\theta(v) \in R_v$. In other words, the result candidates must be a superset of the set of all matches for v. Hence, we can prune the search space by only considering those substitutions θ for which $\theta(v) \in R_v$ for all variable vertices v in Q. The algorithm initializes the result candidates for all variable vertices v in Q which are connected to a constant vertex c in Q through an edge having the label specified in Q. Here we employ the fact that any answer substitution θ must be such that θ(v) is a neighbor of c, and thus, the set of all neighbors of c in S reachable by an edge labeled l are result candidates for v. We use the **DOGMA** index to efficiently retrieve the neighborhood of c. If v is connected to multiple constant vertices, we take the intersection of the respective constraints on the result candidates.

At each recursive invocation, the algorithm extends the given substitution and narrows down the result candidates for all remaining variable vertices correspondingly. To extend the given substitution θ, we greedily choose the variable vertex w with the smallest set of result candidates. This yields a locally optimal branching factor of the search tree since it provides the smallest number of extensions to the current substitution. In fact, if the set of result candidates is empty, then we know that θ cannot be extended to an answer substitution, and we thus directly prune the search. Otherwise, we consider all the possible result candidates m for w by deriving extended substitutions θ′ from θ which assign m to w and then calling DOGMA_basic recursively on θ′. By assigning the constant vertex m to w, we can constrain the result candidates for all neighboring variable vertices as discussed above.

This basic query-answering algorithm only uses "short-range" dependencies, i.e., the immediate vertex neighborhood of variable vertices, to constrain their result candidates. While this suffices for most simple queries, considering "long-range" dependencies can yield additional constraints on the result candidates and thus improve query performance. For instance, the result candidates for $v\ _1$ in our example query not only must be immediate neighbors of "Carla Bunes": in addition, they must be at most at a distance of two from "Health Care." More formally, let $d_S\,(u, v)$ denote the length of the shortest path between two vertices $u, v \in V\ S$ in the undirected counterpart of a graph S, and let $d_q\ (u, v)$ denote the distance between two vertices in the undirected counterpart of a query Q. A long-range dependency on a variable vertex $v \in V_q$ is introduced by any constant vertex $c \in V_q$ with $d_q\ (v, c) > 1$.

We can exploit long-range dependencies to further constrain result candidates. Let v be a variable vertex in Q and c a constant vertex with a long range dependency on v. Then any answer substitution θ must satisfy $d_q\ (v, c) \geq d_s\ (\theta(v), c)$ which, in turn, means that $\{m|\, d_S(m,c) \leq d_Q(v,c)\}$ are result candidates for v. This is the core idea of the DOGMA_adv algorithm (Brücheler et al. 2009), which improves over and extends DOGMA_basic. In addition to the result candidates set R_v, the algorithm maintains sets of distance constraints C_v on them. As long as a result candidates set R_v remains uninitialized, we collect all distance constraints that arise from long-range dependencies on the variable vertex v in the constraints set C_v. After the result candidates are initialized, we ensure that all elements in R_v satisfy the distance constraints in Cv. Maintaining additional constraints therefore reduces the size of $R\ v$ and hence the number of extensions to θ we have to consider.

DOGMA_adv assumes the existence of a *distance index* to efficiently look up $d_S\ (u, v)$ for any pair of vertices $u, v \in V\ S$, since computing graph distances at query time is clearly inefficient. But how can we build such an index? Computing all-pairs shortest path has a worst-case time complexity $O(|V\ S|^3)$ and space complexity $O(|V\ S|^2)$, both of which are clearly infeasible for large social network graphs. However, we do not need to know the *exact* distance between two vertices for DOGMA_adv to be correct. Since all the distance constraints in DOGMA_adv are *upper bounds*, all we need is to



Scaling Subgraph Matching Queries in Huge Networks, Fig. 4 Using DOGMA_ipd for query answering

ensure that $\forall\, u, v \in V S$, the distance retrieved by the index is less than or equal to $d_s\,(u, v)$.

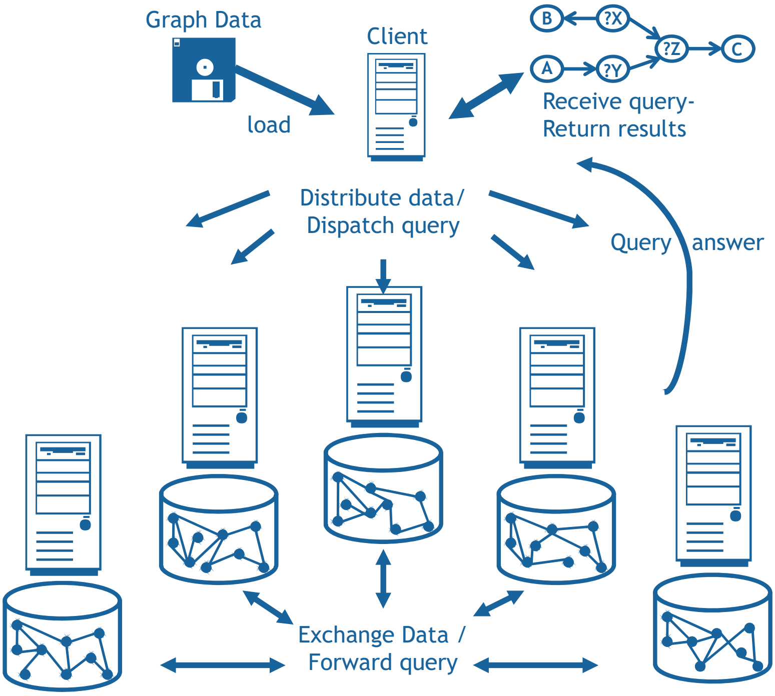
Thus, we can extend the **DOGMA** index to include distance information and build two "lower bound" distance indexes, DOGMA_ipd and DOGMA_epd, that use approximation techniques to achieve acceptable time and space complexity. As seen before, the leaf nodes of the **DOGMA** index **D** S are labeled by subgraphs which constitute a partition of S. For any node $N \in$ **D** S, let $P\ n$ denote the union of the graphs labeling all leaf nodes reachable from N. Hence, $P\ n$ is the union of all subgraphs in S that were eventually merged into the graph labeling N during index construction and therefore corresponds to a larger subset of S. For example, the dashed lines in Fig. 1 mark the subgraphs $P\ n$ for all index tree nodes N of the **DOGMA** index shown in Fig. 3, whereas bolder lines indicate boundaries corresponding to nodes of lower depth in the tree.

The **DOGMA** *internal partition distance* (DOGMA_ipd) index stores, for each index node N and vertex $v \in P_n$, the distance to the outside of the subgraph corresponding to P_n. We call this the *internal partition distance* of $v; N$, denoted ipd(v, N), which is thus defined as ipd (v, N) = min $u \in V S\ d\ S\,(v, u)$. We compute these distances during index construction. At query time, for any two vertices $v, u \in V S$ we first use the **DOGMA** tree index to identify those distinct nodes $N \neq M$ in **D** S such that $v \in P\ n$ and $u \in P\ m$, which are at the same level of the tree and closest to the root.

If such nodes do not exist (because v; u are associated with the same leaf node in **D** S), then we set $d_{\text{ipd}}(u, v) = 0$. Otherwise, we set $d_{\text{ipd}}(u, v) = \max(\text{ipd}(u, N), \text{ipd}(u, M))$. It is easy to see that d_{ipd} is an admissible lower bound distance, since $P\ n \cap P\ m = \varnothing$. By choosing those distinct nodes which are closest to the root, we ensure that the considered subgraphs are as large as possible, and hence, $d_{\text{ipd}}(u, v)$ is the closest approximation to the actual distance.

Example 3 Consider the example of Figs. 1 and 2. Fig. 4 shows the initial result candidates for each of the variable vertices in boxes. We can determine that there is a long-range dependency between "Carla Bunes" and variable vertex? v_2 at distance 2. The boldest dashed line in Fig. 1 marks the top-level partition and separates the sets $P\ N1$, $P\ N2$, where N_1, N_2 are the two nodes directly below the root in the **DOGMA** index in Fig. 3. We can determine that ipd.(Carla Bunes, N_2) = 3, and since Bill B0045 and B0532 lie in the other subgraph, it follows that d_{ipd}.Carla Bunes; B0045/ B0532/ = 3, and therefore, we can prune both result candidates.

The **DOGMA** *external partition distance* (DOGMA_epd) index also uses the partitions in the index tree to compute a lower bound distance.



Scaling Subgraph Matching Queries in Huge Networks, Fig. 5 Architecture of COSI

However, it considers the distance to *other* subgraphs rather than the distance within the same one. For some fixed level L, let N L denote the set of all nodes in **D** S at distance L from the root. As discussed above, $P = \{P_N\}_{N \in \mathcal{N}_L}$ is a partition of S. The idea behind DOGMA_epd is to assign a color from a fixed list of colors C to each subgraph P n ∊ P and to store, for each vertex v ∊ V S and color c ∊ C, the shortest distance from v to a subgraph colored by c. We call this the *external partition distance*, denoted epd(v, c), which is thus defined as $\mathrm{epd}(v,c) = \min_{u \in P_N,\, \phi(P_N)=c} d_S(v,u)$ where φ: $P \rightarrow C$ is the color assignment function. We store the color of P n with its index node N so that for a given pair of vertices u, v we can quickly retrieve the colors c u, c v of the subgraphs to which u and v belong. We then compute $d_{\mathrm{epd}}(v, u)$ = max.epd(v, c u), epd (v, c v) . It is easy to see that d_{epd} is an admissible lower bound distance.

Cloud-Oriented Subgraph Matching

This section presents the COSI system which distributes a social network graph across multiple compute machines and answers subgraph matching queries in parallel using an asynchronous query-answering algorithm that does not rely on central orchestration.

Figure 5 shows a schematic view of the architecture of COSI. We assume that a compute cloud consists of k compute nodes and one “client” node. Compute nodes communicate directly without going through the client node, thus preventing the client node from becoming a communication bottleneck. The client node takes a query and directs it or parts of it to one or more compute nodes that complete the computation of the answer. The complete answer is then shipped to the client node. In Fig. 5, $k = 5$ and the compute machines are shown in the lower half of the figure. Each of those machines stores a fragment of the graph in a local graph database and responds to query requests specifically addressed to it. The architecture and system we present in this entry does not depend on any particular local graph database.

Distributing a Social Network Graph

We now address the question: How do we distribute a social network graph across a cloud so that we can efficiently process subgraph matching queries? In partitioning the social network data, we follow two objectives: (i) all *k* compute machines should store roughly the same amount of data to balance the load across machines, and (ii) the partition should minimize the expected query execution time.

At a high level, we achieve these objectives as follows. First, we transform the social network graph *S* into a simple weighted graph **WG**.(*S*). Intuitively, the weight of an edge *e* = *(u, v)* in **WG**.(*S*) refers to the sum of the probability that *v* will be retrieved immediately after *u* and vice versa when an arbitrary query is processed. If this probability is (relatively) high, then the two vertices should be stored on the same compute node. Then, we use these to partition *S* across the *k* compute machines so that expected communication costs are minimized. In the remainder, we assume there is a probability distribution $\mathbb{P}$ over the space of all queries. Intuitively, $\mathbb{P}(Q)$ is the probability that a random subgraph matching query posed to a social network is *Q*. For any real-world online social network, $\mathbb{P}$ can be easily learned from frequency analysis of past query logs.

We start by introducing our formalization of *query plans and query traces*. A query plan *qp.(Q)* for a query *Q* is a sequence of two types of *operations*: the first type retrieves the neighborhood of vertex *v* (from whichever compute node it is on), and the second type performs some computation (e.g., check a selection condition or perform a join) on the results of previous operations. This definition is compatible with most existing definitions of query plans in the database literature. Now suppose *x* = *qp(Q)* is a query plan for a query *Q* on a social network graph *S*. The query trace of executing *x* on *S*, denoted *qt(x, S)*, consists of (i) all the vertices *v* in *S* whose neighborhood is retrieved during execution of query plan *x* on *S* and (ii) all *pairs* (*u, v*) of vertices where immediately after retrieving *u*'s neighborhood, the query plan retrieves *v*'s neighborhood (in the next operation of *x*). When processing a query, we make the reasonable assumption that index retrievals are cached so that repeated vertex neighborhood retrievals are read from memory and hence the query trace *qt(x, S)* can be defined as a set rather than as a multiset. Traces contain consecutive retrievals of vertex neighborhoods — this allows us to store neighborhoods of both *u* and *v* on the same compute node, avoiding unnecessary communication.

The probability distribution $\mathbb{P}$ on queries can be used to infer a probability distribution $\widetilde{\mathbb{P}}$ over the space of feasible query plans: $\widetilde{\mathbb{P}}(x) = \sum_{Q \in \mathcal{Q}: qp(Q)=x} \mathbb{P}(Q)$. This says that the probability of a query plan is the sum of the probabilities of all queries which use that query plan. In the rest of the entry, we will abuse notation and denote both PDFs by $\mathbb{P}$. We can now define the *probabilities of retrieval and coretrieval*. The probability, $\mathbb{P}(v)$, of retrieving *v* when executing a random query plan is $\sum_{x \in qp(\mathcal{Q}): v \in qt(x,\mathcal{S})} \mathbb{P}(x)$. Thus, the probability of retrieving *v* is the sum of the probabilities of all query plans that retrieve *v*. The probability $\mathbb{P}(v_1, v_2)$ of retrieving v_2 immediately after v_1 is $\sum_{x \in qp(\mathcal{Q}):(v_1, v_2) \in qt(x,\mathcal{S})} \mathbb{P}(x)$ This says that the probability of retrieving v_2 immediately after v_2 is the sum of the probabilities of all query plans that retrieve v_2 immediately after v $_1$.

We can associate a weighted graph WG(*S*) with the graph *S* = (*V, E,λ*). The weighted graph is the complete graph.(*V, V* × *V, w*) where $w(v_1, v_2) = \mathbb{P}(v_1, v_2) + \mathbb{P}(v_2, v_1)$. An *edge cut C* of a weighted graph is a partition of the vertices into components. An edge (*u, v*) in the graph is said to *cross the edge cut C* if *u* is in one component of the partition and *v* is in another. The size of an edge cut is the sum of the weights of the edges that cross the cut. *C* is said to be a *minimum cut* if there is no other cut *C′* such that the size of *C′* is less than the size of *C*. The important theorem we gave in Brücheler et al. (2010) shows that the minimal edge cut of WG(*S*) corresponds to the partition of *S* across *k* compute nodes that minimizes expected cost of executing a query.

Since computing minimal edge cuts is a well-known NP-hard problem, we develop heuristic techniques to partition the graph that allow us to obtain suboptimal partitions of high quality, without incurring in the expensive computational costs of obtaining the optimal ones. We start by defining the concept of *vertex force vector*. Let

$\mathcal{P} = \{P_1; \ldots, P_{\mathrm{k}}\}$ be a partition of S and consider any component P_i. The *vertex force vector*, denoted $|\vec{v}|$, of any vertex $v \in S$ is a k-dimensional vector where $|\vec{v}|\,[i] = f_{\mathcal{P}}\left(\sum_{x \in \mathrm{ngh}(v) \cap P_i} w((v,x))\right)$ and $f_{\mathcal{P}} : \mathbb{R}^{+} \to \mathbb{R}$ is a function called the *affinity measure*.

The vertex force vector intuitively specifies the "affinity" between a vertex and each component as measured by the affinity measure $f_{\mathcal{P}}$. An affinity measure takes the connectedness between a vertex v and the respective component as an argument. The vertex force vector captures the strength with which each component "pulls" on the vertex and is used as the basis for a vertex assignment decision: intuitively, if an inserted edge introduces a new vertex v, we first compute the vertex force vector $|\vec{v}|$ and then assign v to the component P_j where $j = \mathrm{argmax}_{1 \le i \le k} |\vec{v}|[i]$.

COSI uses an affinity measure that is a linear combination of three factors: *connectedness*, *imbalance*, and *size*. Obviously, evaluating the connectedness of a vertex v to a component $P\ i$ is crucial for edge cut minimization — we measure this as the number of edges that connect v to the vertices in $P\ i$. Moreover, balanced partitions lead to even workload distribution, thus enhancing parallelism. Let $|P_i|_E = \sum_{x \in P_i} \mathrm{deg}(x)$ be the number of edges in P_i and let T be an estimate (even a bad one) of the total number of edges that a given graph is expected to be. Then a reasonable measure of imbalance is the standard deviation of $\frac{\{|P_i|_E\}_{1 \le i \le k}}{T}$. Finally, we regulate the size of components by comparing the actual size of a component to its expected one. If a component grows beyond its expected size, we punish such growth more aggressively than imbalance does alone by reducing the affinity further according to the metric min $\left(-\frac{|P_i|_E - \frac{T}{k}}{T}, 0\right)$.

Consider now the case of a new set of edges to be inserted into a social network graph, given that a partition $\mathcal{P} = P\ _1, \ldots, P\ k$ of the graph already exists (this can be used to create a partition for the first time by assuming $S = \varnothing$). A naive GreedyInsert insertion algorithm would iterate over all new vertices v: for each vertex v it would compute the vertex force vector and assigns v to the component P_i such that $|\mathbf{v}|\ [i]$ is maximal — fortunately we can do better.

Our COSI_Partition algorithm (Brücheler et al. 2010) leverages graph *modularity* (Blondel et al. 2008) to identify a strongly connected subgraph that is loosely connected to the remaining graph. However, modularity cannot be used blindly as our balance requirement must also be met. The *modularity* of a partition $\mathcal{P}$ of an undirected graph $G = (V, E)$ with weight function w: $E \to \mathbb{R}$ is defined as

$$\mathrm{mod}(\mathcal{P}) = \sum_{P \in \mathcal{P}} \left(\frac{W(P,P)}{2|E|} - \frac{\deg_W(P)^2}{(2|E|)^2} \right)$$

where $\deg_w(v) = \sum_{x \in V} w((v, x))$ is the weighted degree of vertex v, $W(X, Y) = \sum_{x \in X, y \in Y} w((x, y))$ is the sum of edge weights connecting two sets of vertices $X, Y \subset V$, and $\deg_W(X) = \sum_{x \in X} \deg_w(x)$ is the weighted degree of a set of vertices $X \subset V$. Intuitively, components with high modularity are densely connected subgraphs which are isolated from the rest of the graph. Our algorithm iteratively builds high modularity components and then assigns all vertices in a component to one compute node based on the vertex force vector. Let $B \subset V$ be a set of vertices. We generalize the notion of a vertex force vector by defining $|\vec{B}|[i] = f_{\mathcal{P}}\left(\sum_{v \in B} \sum_{x \in \mathrm{ngh}(v) \cap P_i} w((v,x))\right)$ The intuition behind our partitioning algorithm is that assigning vertices at the aggregate level of isolated and densely connected components yields good partitions because (i) we respect the topology of the graph, (ii) most edges are within components and therefore cannot be cut, and (iii) force vectors of sets of vertices combine the connectedness information of many vertices leading to better assignment decisions.

Processing Queries with COSI

Our COSI_basic parallel query processing algorithm (Brücheler et al. 2010) operates asynchronously and in parallel across all compute nodes. A user issues query Q to the client node which "prepares" the query. In particular, it selects one constant vertex c from Q and determines the compute node that hosts c — the prepared query is then forwarded to this node.

The algorithm proceeds depth first, substituting vertices for variables in Q one at a time. We maintain a set of result candidates for each variable in Q. The algorithm assumes there is an index retrieval function that retrieves ngh l (v) from the local index (which could be implemented many ways — we used a DOGMA index in the experiments) on the compute node. The algorithm arbitrarily chooses the next vertex to be substituted. Incoming queries come with a selected variable to be instantiated with a vertex ID. The algorithm updates the candidate result sets by retrieving the neighborhood of the newly substituted vertex from the index. It then checks if any results have been found or whether the current substitution cannot yield a valid result. If neither condition holds, the algorithm selects the next variable v' to be substituted and forwards the query to those compute nodes that host potential substitution candidates for v'. All query results are sent to the client which returns them to the user.

COSI_basic does not rely on central orchestration — it uses depth-first search so the branches of the search tree are traversed in parallel while ensuring that no branch gets explored multiple times. After forwarding the prepared query to a compute node, the client waits for incoming results of that query and forwards those to the user. As we explore branches in parallel, the client node cannot be notified when the search for query results has completed. Keeping track of the current number of parallel executions for each query would introduce significant synchronization cost. Instead, the client node keeps track of the time t_{last} at which the last result of a running query has come in. If the difference between the current time and t_{last} exceeds a threshold, the client node asks all compute nodes for a list of query IDs of all currently running queries. The client node merges these lists and closes all queries whose IDs are not contained. To avoid the case where a query is being forwarded to another compute node at the very moment that the client node asks for all query IDs, each compute node keeps query IDs in their local list up to a certain grace period.

The choice of the next variable to be instantiated has profound implications on the running time of COSI_basic, as some substitutions yield larger branching factors in the search than others. Our COSI_heur algorithm (Brücheler et al. 2010) handles this by choosing the variable vertex v' which has the lowest cost according to a function h_{opt}. First, to reduce the branching factor, we could choose the variable vertex v' with the smallest number of result candidates. This heuristic only considers the branching factor of the immediate next iteration but is nevertheless an important metric to consider in the cost heuristic. Second, whenever we instantiate a vertex on a remote component, we have to send a message to the appropriate compute node which is expensive. Therefore, we consider the fraction of result candidates which are not stored locally as a cost metric. When we have to send a query to remote nodes for further processing, we would like to distribute the workload evenly across all nodes. Hence, we also analyze the distribution of result candidates by node via the cost metric

$$ds(v) = \sqrt{\sum_{1 \le i \le k} \left(|R_v^i| - \frac{|R_v|}{k} \right)^2}$$

where R_v^i is the set of result candidates for vertex v restricted to those which reside on compute node i. Finally, we define

$$h_{\text{opt}}(v) = |R_v| \times \left(1 - \frac{|R_v^l|}{\alpha \times |R_v|}\right) \times \left(1 + \beta \times \frac{ds(v)}{|R_v|}\right)$$

where l is the ID of the local compute node and α and β are constants that determine how much the model favors locality over parallelism.

Experimental Results

In this section, we present the results of the experimental assessment we performed of both the **DOGMA** index and the COSI system.

DOGMA

We tested the **DOGMA** index on RDF data and compared its performance with four RDF database systems developed in the Semantic Web community that are most widely used and have demonstrated superior performance in previous evaluations (Lee et al. 2008): Sesame2 (2013), *Jena2* (Wilkinson et al. 2003), JenaTDB (2013), and the internal memory version of *OWLIM* (Kiryakov et al. 2005). Moreover, we used three different RDF datasets. *GovTrack* (2013) consists of more than 14.5 million triples describing data about the US Congress. The *Lehigh University Benchmark* (LUBM) (2013) is frequently used within the Semantic Web community as the basis for evaluation of RDF and ontology storage systems — we generated a dataset of more than 13.5 million triples. Finally, we used a fragment of the Flickr social network (2013) collected by researchers of the MPI Saarbrücken to analyze online social networks (Mislove et al. 2007). The fragment was anonymized and contains approximately 16 million triples. The GovTrack and Flickr datasets are well connected (with the latter being denser than the former), whereas the dataset generated by the LUBM benchmark is a sparse and almost degenerate graph containing a set of small and loosely connected subgraphs.

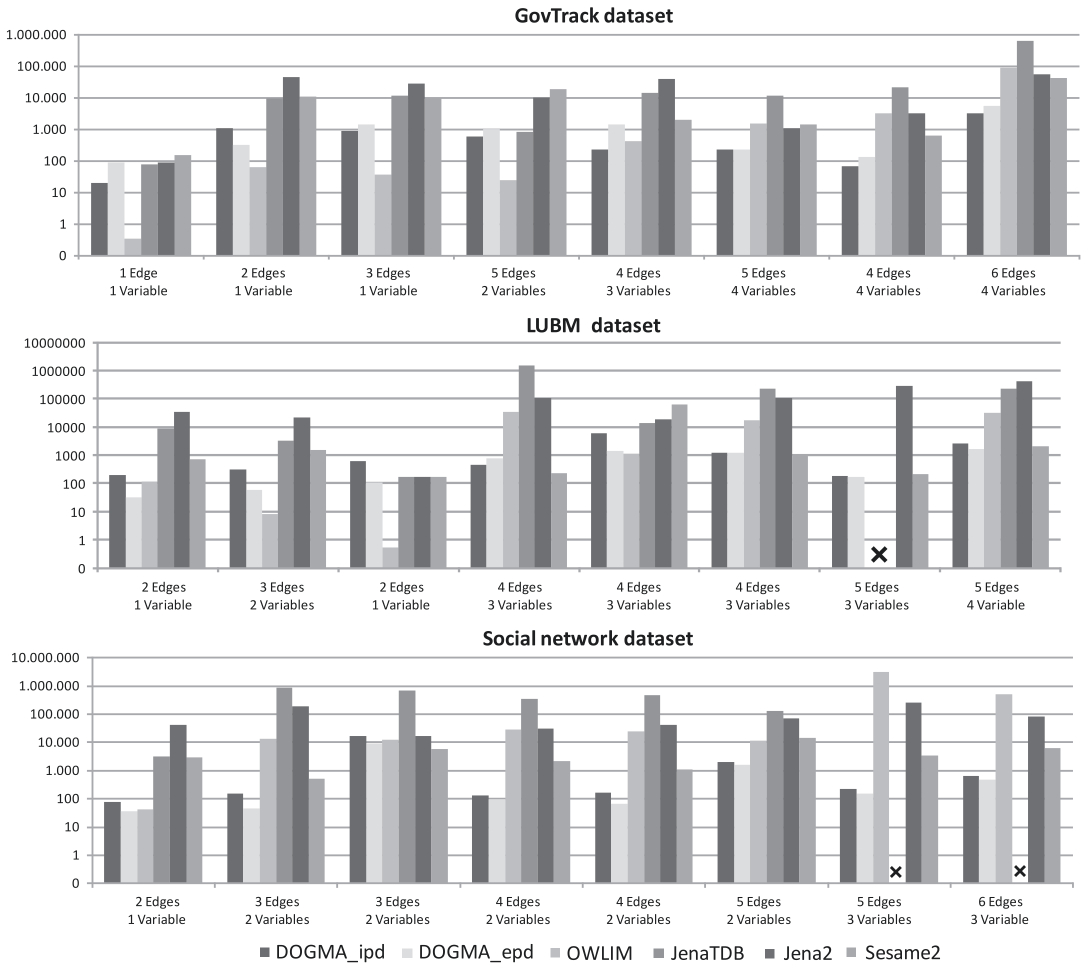
We designed a set of graph queries with varying complexity, where constant vertices were chosen randomly and queries with an empty result set were filtered out. Queries were grouped into classes based on the number of edges and variable vertices. We repeated the query time measurements multiple times for each query, eliminated outliers, and averaged the results. Finally, we averaged the query times of all queries in each class. All experiments were executed on a machine with a 2.4 GHz Intel Core 2 processor and 3 GB of RAM.

In first round of experiments, we designed several relatively simple graph queries for each dataset, containing no more than six edges, and grouped them into eight classes. The results of these experiments are shown in Fig. 6 which reports the query times for each query class on each of the three datasets. Missing values in the figure indicate that the system did not terminate on the query within a reasonable amount of time (around 20 min). Note that the query times are plotted in logarithmic scale to accommodate the large discrepancies between systems. The results show that OWLIM has low query times on low-complexity queries across all datasets. This result is not surprising, as OWLIM loads all data into main memory prior to query execution. The performance advantage of DOGMA_ipd and DOGMA_epd over the other systems increases with query complexity on the GovTrack and the Flickr dataset, where our proposed techniques are orders of magnitude faster on the most complex queries. On the LUBM dataset, however, Sesame2 performs almost equally for the more complex queries. Finally, DOGMA_epd is slightly faster on the LUBM and Flickr dataset, whereas DOGMA_ipd has better performance on the GovTrack dataset.

In second round of experiments, we significantly increased the complexity of the queries, which now contained up to 24 edges. Unfortunately, the OWLIM, JenaTDB, and Jena2 systems did not manage to complete the evaluation of these queries in reasonable time, so we exclusively compared with Sesame2. The results are shown in Fig. 7. On the GovTrack and Flickr dataset, DOGMA_ipd and DOGMA_epd continue to have a substantial performance advantage over Sesame2 on all complex graph queries of up to 40,000%. For the LUBM benchmark, the picture is less clear due to the particular structure of the generated dataset explained before.

Finally, Fig. 8 compares the storage requirements of the systems under comparison for all three datasets. The results show that DOGMA_ipd, DOGMA_epd, and Sesame2 are the most memory efficient.

In conclusion, we can observe that both DOGMA_ipd and DOGMA_epd are significantly faster than all other RDF database systems under comparison on complex graph queries over non-degenerate graph datasets. Moreover, they can efficiently answer complex queries on which most of the other systems do not terminate or take up to 400 times longer while maintaining a satisfactory storage footprint. DOGMA_ipd and DOGMA_epd have similar performance, yet



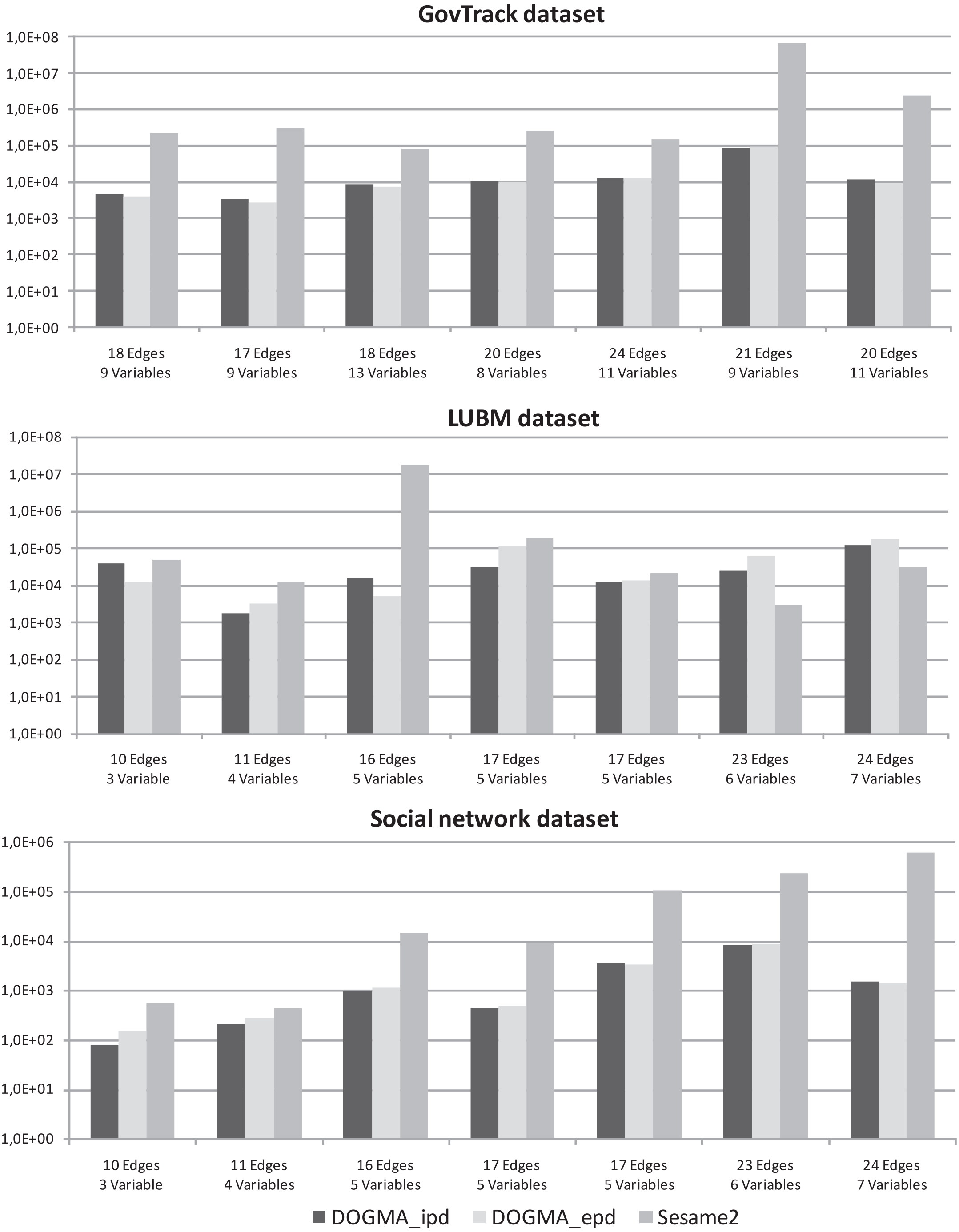
Scaling Subgraph Matching Queries in Huge Networks, Fig. 6 Query times (ms) for graph queries of low complexity

differences exist which suggest that each index has unique advantages for particular queries and graph structures.

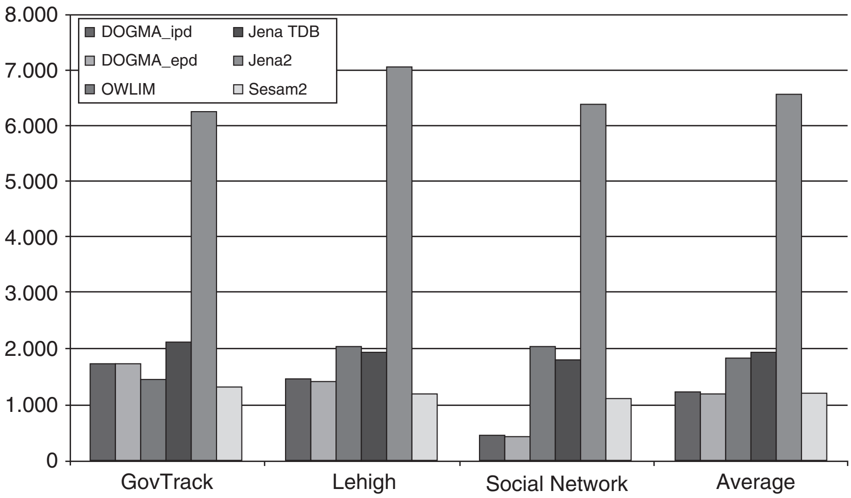
COSI

In the experiments with COSI, we used the social network graph studied in Mislove et al. (2007). This graph contains 778 M edges and describes personal relationships and group memberships crawled from Facebook, Orkut, Flickr, and LiveJournal. We fixed the coefficients for the affinity measure by hand. Both, the imbalance and excessive size metric, were given an equal weight of one. The connectedness measure was set relative to the number of edges we considered per batch. We experimented with different batch sizes and found best performance for half a million edges.

We developed a communication infrastructure for the compute nodes based on the Java NIO libraries which is used to send the graph data during the loading and the queries during the query-answering stages. The communication infrastructure handles contention at individual nodes and variations in network latency. It is optimized to ensure that the client node's requests for outstanding queries are answered quickly. In our experiments, we used a cluster of 16 compute nodes, out of which one served as a client node and the remaining 15 nodes served as compute

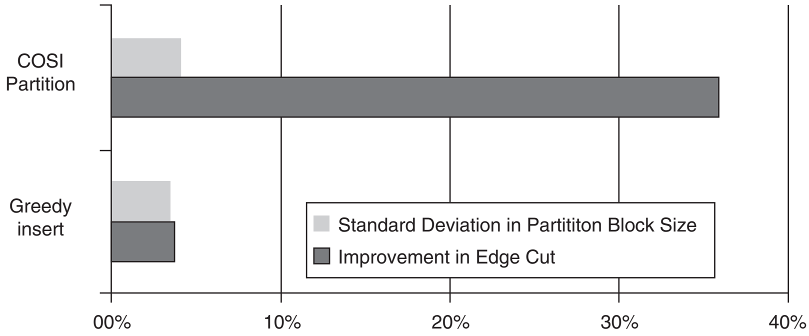


Scaling Subgraph Matching Queries in Huge Networks, Fig. 7 Query times (ms) for graph queries of high complexity



Scaling Subgraph Matching Queries in Huge Networks, Fig. 8 Index size (MB) for different datasets

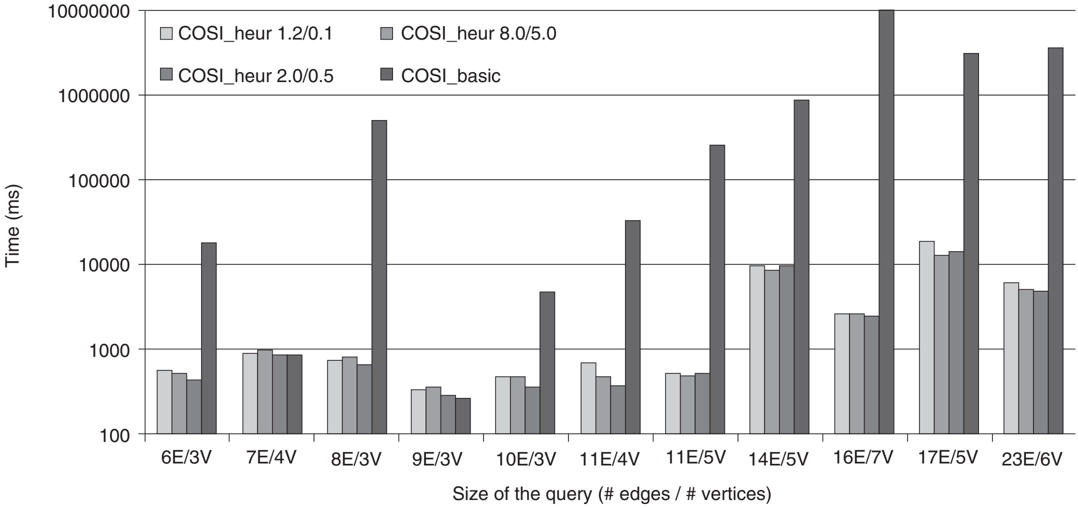
Scaling Subgraph Matching Queries in Huge Networks, Fig. 9 Comparison of partitioning methods



nodes. All compute nodes had an identical hardware configuration with a 4-core 2.16 GHz Intel CPU, 4 GB of RAM, and 80 GB IDE 7200 rpm hard drive. The client node's hardware differed slightly with an 8-core CPU and 8 GB of RAM.

Figure 9 compares COSI_Partition's performance with that of the GreedyInsert algorithm. To validate our experiments, we used a random partitioning scheme, which assigns vertices to compute nodes uniformly at random, as the naive baseline in our experiments and report all results in comparison to this baseline. COSI_Partition achieves a substantial 36% improvement in edge cut over the naive baseline at a total running time of 10.5 h for all 778 M edges. GreedyInsert only achieves a marginal improvement in edge cut. COSI_Partition significantly outperforms greedy batch insertion by 33% with only slightly higher imbalance as measured in the standard deviation in component size relative to average size of a component.

Figure 10 compares COSI_basic against COSI_heur for three different parameter settings of function h_{opt}: ($\alpha = 1{:}2$, $\beta = 0.1$) which strongly favors locality over parallelism, ($\alpha = 8.0$, $\beta = 5.0$), which strongly favors parallelism over locality, and ($\alpha = 2.0$, $\beta = 0.5$) which balances locality and parallelism. The queries have increasing complexity as measured by the number of edges (E) and variables (V) in the query graph. All query times were averaged across six independent runs with complete system restarts after each run to empty caches. Note that the graph is plotted



Scaling Subgraph Matching Queries in Huge Networks, Fig. 10 Query times by query-answering algorithm on the 778 M edge dataset

in logarithmic scale to accommodate the huge differences in query times.

COSI_heur drastically outperforms COSI_basic by up to four orders of magnitude on all but two queries, and the performance gap seems to grow exponentially with the query complexity. A close look at the difference in performance between the variants of COSI_heur reveals that the third configuration outperforms the first one on nine queries, with a tie on the remaining two, and outperforms the second configuration on eight queries, being slower only on three. These results suggest that a balanced choice of parameters leads to a better h_{opt}.

Related Work and Conclusions

The problem of efficiently evaluating subgraph matching queries over huge graphs/networks has been recently addressed in different scenarios, among which social network analysis and RDF database management play an important role (Martín and Gutierrez 2009). A wide variety of methods for social network analysis have been proposed (Borgatti et al. 2002; Nooy et al. 2005; Huisman and Duijn 2005). However, most algorithms operate solely in memory, loading the entire graph from disk and then executing the analysis. For social networks of the size of Facebook, Flickr, or Orkut, such an approach becomes infeasible. To handle social networks of such magnitude, one needs to store and query network data efficiently on disk. More importantly, complex queries involving even a few joins can quickly cause such approaches to run into trouble. Ronen and Shmueli (2009) introduce a social network-specific query language and show how such queries can be answered on moderately sized datasets. However, their query language is geared toward users of a social network in helping them communicate with friends.

Graph-structured RDF data has been studied in the Semantic Web community (Mahmoudi-Nasab and Sakr 2010). Initial approaches to RDF storage (Broekstra et al. 2003; Sintek and Kiesel 2006; Wilkinson et al. 2003) stored the graph in relational tables and then used a relational query engine to answer queries. Abadi et al. (2007) showed that storing RDF in a vertical database leads to significant query time improvements. Stocker et al. (2008) uses triple selectivity estimation techniques similar to those used in relational database systems. Pugliese et al. (2008) and Udrea et al. (2007) are the first to propose specific tree-structured indexes for RDF. All these approaches work on single machines. In response to the increasing need of scalability when facing extremely large RDF datasets, two approaches have essentially been proposed so far: *scale up and scale out*. In scaling up, existing RDF databases, such as RDF-3X (Neumann and Weikum 2008), Sesame (Broekstra et al. 2003), or YARS (Harth and Decker 2005), are simply run on more powerful machines. As such it requires no technological innovation but is very costly and limited by current hardware. In scaling out, multiple machines are utilized to store the data but all operations on the data are centrally executed. Parallel storage regimes, such as YARS2 (Harth et al. 2007), are cheaper but still limited in their scalability due to central execution. Our COSI system demonstrated efficient query answering across multiple machines without central orchestration.

Earlier work on database technologies for general graph data such as Lore (Goldman et al. 1999) considered much smaller graphs than the social networks we study here. More recent work (e.g., Cheng et al. 2009; Giugno and Shasha 2002; Ke et al. 2010; Sakr 2009; Zhang et al. 2010; Zhu et al. 2010; Brücheler et al. 2011a, b) focuses on heuristics to predict the cost of answering strategies based on statistics about the dataset and the current state of query processing and then choose a strategy to minimize cost. However, due to the highly heterogeneous nature of network data (Newman 2003), such predictions can become inaccurate. Zou et al. (2009) proposes to transform vertices into points in a vector space, thus converting queries into distance-based multiway joins over the vector space. In Cheng et al. (2008), the authors propose a two-step join optimization algorithm based on a cluster-based join index. *GADDI* is proposed in Zhang et al. (2009) that employs a structural distance-based approach and a dynamic matching scheme to minimize redundant calculations. GADDI can handle graphs with thousands of vertices, which are common in many biological applications. In Zhang et al. (2010), the authors propose *SUMMA*, which improves over GADDI and employs more advanced indices, becoming capable to handle graphs with up to tens of millions of vertices. The algorithm in Zhu et al. (2010) employs an aggressive pruning strategy based on an index storing label distributions. In Natale et al. (2010), the authors argue that existing indices over sets of data graphs do not support efficient pruning when they face graphs with tens of thousands of vertices. They propose an index that is specifically targeted at this scenario.

In this entry, we have described a disk-oriented index and a graph-partitioning technique that make the processing of complex subgraph matching queries on very large graph data feasible. The DOGMA index is based on the simple observation that the size of any real-world social network graph is likely to be orders of magnitude larger than that of any subgraph matching query graph a user is likely to ask. Thus, it is possible to build an index for efficiently executing such queries that ensures that vertices in a social network graph that are "near" each other be stored together on a disk page. On the other hand, the COSI system is able to effectively distribute a social network graph across multiple compute machines and answer subgraph matching queries asynchronously in parallel. The experimental results confirm the feasibility of both approaches.

Cross-References

- ▶ Extracting and Inferring Communities via Link Analysis
- ▶ Graph Matching
- ▶ RDF
- ▶ SPARQL
- ▶ Subgraph Extraction for Trust Inference in Social Networks

References

Abadi DJ, Marcus A, Madden S, Hollenbach KJ (2007) Scalable semantic web data management using vertical partitioning. In: VLDB, Vienna, pp 411–422

Blondel V, Guillaume J, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008:P10008

Borgatti SP, Everett MG, Freeman LC (2002) Ucinet for windows: software for social network analysis. Analytic Technologies, Harvard

Broekstra J, Kampman A, van Harmelen F (2003) Sesame: an architecture for storing and querying RDF data and schema information. In: Fensel D, Hendler JA, Lieberman H, Wahlster W (eds) Spinning the semantic web. MIT Press, Cambridge, pp 197–222

Brücheler M, Pugliese A, Subrahmanian VS (2009) DOGMA: a disk-oriented graph matching algorithm for RDF databases. In: ISWC, Chantilly, pp 97–113

Brücheler M, Pugliese A, Subrahmanian VS (2010) COSI: cloud oriented subgraph identification in massive social networks. In: Memon N, Alhajj R (eds) ASONAM, Odense. IEEE Computer Society, pp 248–255

Brücheler M, Pugliese A, Subrahmanian VS (2011a) Probabilistic subgraph matching on huge social networks. In: Alhajj R, Memon N, Ting I (eds) ASONAM, Kaohsiung. IEEE Computer Society, pp 271–278

Brücheler M, Pugliese A, Subrahmanian VS (2011b) A budget-based algorithm for efficient subgraph matching on huge networks. In: Abiteboul S, Bohm K, Koch C, Tan K (eds) ICDE workshop proceedings, Hannover. IEEE Computer Society, pp 94–99

Cheng J, Yu JX, Ding B, Yu PS, Wang H (2008) Fast graph pattern matching. In: ICDE conference, Cancun, pp 913–922

Cheng J, Ke Y, Ng W (2009) Efficient query processing on graph databases. ACM Trans Database Syst 2(1–2):48

Flickr (2013) http://www.flickr.com

Giugno R, Shasha D (2002) Graphgrep: a fast and universal method for querying graphs. In: ICPR conference, Québec City, pp 112–115

Goldman R, McHugh J, Widom J (1999) From semi-structured data to XML: migrating the Lore data model and query language. In: Proceedings of the 2nd international workshop on the web and databases (WebDB'99), Philadelphia, pp 25–30

GovTrack dataset (2013) http://www.govtrack.us

Harth A, Decker S (2005) Optimized index structures for querying RDF from the web. In: Proceedings of the 3rd Latin American web congress, Buenos Aires, pp 71–80

Harth A, Umbrich J, Hogan A, Decker S (2007) YARS2: a federated repository for querying graph structured data from the web. In: ISWC, Busan, pp 211–224

Huisman M, Duijn MAV (2005) Software for social network analysis. In: Carrington PJ, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, Cambridge/New York, pp 270–316

JenaTDB (2013) http://jena.apache.org

Karypis G, Kumar V (1999) A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM J Sci Comput 20:359–392

Ke Y, Cheng J, Yu JX (2010) Querying large graph databases. In: DASFAA conference, Tsukuba, pp 487–488

Kiryakov A, Ognyanov D, Manov D (2005) OWLIM – a pragmatic semantic repository for OWL. In: WISE workshops, New York, pp 182–192

Lee C, Park S, Lee D, Lee J, Jeong O, Lee S (2008) A comparison of ontology reasoning systems using query sequences. In: Proceedings of the 2nd international conference on ubiquitous information management and communication, Suwon. ACM, pp 543–546

MahmoudiNasab H, Sakr S (2010) An experimental evaluation of relational RDF storage and querying techniques. In: DASFAA workshops, Tsukuba, pp 215–226

Martín MS, Gutierrez C (2009) Representing, querying and transforming social networks with RDF/SPARQL. In: ESWC conference, Heraklion, pp 293–307

Mislove A, Marcon M, Gummadi PK, Druschel P, Bhattacharjee B (2007) Measurement and analysis of online social networks. In: Internet measurement conference, San Diego, pp 29–42

Natale RD, Ferro A, Giugno R, Mongiovì M, Pulvirenti A, Shasha D (2010) SING: subgraph search in non-homogeneous graphs. BMC Bioinform 11:96

Neumann T, Weikum G (2008) RDF-3X: a RISC-style engine for RDF. PVLDB 1(1):647–659

Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256

Nooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Structural analysis in the social sciences, vol 27. Cambridge University Press, New York

Pugliese A, Udrea O, Subrahmanian VS (2008) Scaling RDF with time. In: WWW, Beijing, pp 605–614

Ronen R, Shmueli O (2009) Evaluating very large datalog queries on social networks. In: EDBT, Saint-Petersburg, pp 577–587

Sakr S (2009) GraphREL: a decomposition-based and selectivity-aware relational framework for processing sub-graph queries. In: DASFAA conference, Brisbane, pp 123–137

Sesame2 (2013) http://www.openrdf.org

Sintek M, Kiesel M (2006) RDFBroker: a signature-based high-performance RDF store. In: ESWC, Budva, pp 363–377

Stocker M, Seaborne A, Bernstein A, Kiefer C, Reynolds D (2008) SPARQL basic graph pattern optimization using selectivity estimation. In: Proceeding of the 17th international conference on World Wide Web, ACM, Beijing, pp 595–604

The Lehigh University Benchmark (2013) http://swat.cse.lehigh.edu/projects/lubm

Udrea O, Pugliese A, Subrahmanian VS (2007) GRIN: a graph based RDF index. In: AAAI, Vancouver, pp 1465–1470

Wilkinson K, Sayers C, Kuno H, Reynolds D (2003) Efficient RDF storage and retrieval in Jena2. Proc SWDB 3:7–8

Zhang S, Li S, Yang J (2009) GADDI: distance index based subgraph matching in biological networks. In: EDBT conference, Saint-Petersburg, pp 192–203

Zhang S, Li S, Yang J (2010) SUMMA: subgraph matching in massive graphs. In: CIKM conference, Toronto, pp 1285–1288

Zhu K, Zhang Y, Lin X, Zhu G, Wang W (2010) NOVA: a novel and efficient framework for finding subgraph isomorphism mappings in large graphs. In: DASFAA conference, Tsukuba, pp 140–154

Zou L, Chen L, ßzsu MT (2009) Distancejoin: pattern match query in a large graph database. VLDB Conf 2 (1):886–897

Scamming

▶ Online Social Network Phishing Attack

Scan Statistics

▶ Disease Surveillance: Case Study

Schema Matching

▶ Ontology Matching

Scholarly Communication

▶ Scholarly Network Analysis

Scholarly Network Analysis

Erjia Yan[1] and Ying Ding[2]
[1]College of Computing and Informatics, Drexel University, Philadelphia, PA, USA
[2]School of Informatics and Computing, Indiana University, Bloomington, IN, USA

Synonyms

Disciplinary; Knowledge flow; Scholarly communication; Scholarly networks; Science maps; Scientific collaboration; Scientific evaluation; Topic identification

Glossary

Node (in scholarly networks)	Units such as words, papers, patents, authors, journals, institutions, fields, and countries
Edge (in scholarly networks)	Citation, co-citation, co-word, coauthor, bibliographic coupling, hybrid, or heterogeneous relations
Scholarly network	The combination of edge properties and node properties defines a scholarly network
Macro-level approach	Statistics that are used to identify the global structural features of networks, including component, bicomponent, shortest distance, clustering coefficient, degree distribution, and error and attack tolerance
Meso-level approach	Approaches that focus on the behaviors of a group of actors, including topic identification and community detection
Microlevel approach	Indicators that are useful to understand individual node's power, stratification, ranking, and inequality in social structures, including centrality measures and PageRank and its variants

Definition

Scholarly networks are a type of complex networks. Some scholarly networks belong to social networks, such as collaboration networks, while others belong to information networks, such as citation or co-citation networks. Scholarly networks can be constructed on several research units, including papers, authors, journals, institutions, and countries. Scholarly networks are useful to understand scholarly communication, science of science, and knowledge diffusion.

Introduction

In recent years, we have witnessed a growing trend of studying networks, such as social networks, information networks, technical networks, and biological networks (Newman 2003). These studies were informed by the social studies of human interactions, accelerated by the discovery of small-world and scale-free properties, and enriched by a variety of macro-level statistics, meso-level clustering techniques, and microlevel indicators.

Studying the characteristics of scholarly communication is crucial for understanding innovation, collaborations, and scientific activities in general. Scholars have used different types of networks to answer a wide spectrum of questions related to research interaction, scholarly communication, and science policy making. These efforts have advanced the scholarship of scientometrics and our understanding of science of science. For scholarly networks, papers are typically the basic research unit and can be aggregated into several higher levels, such as the author, journal, institution, and field levels. Network types define edge properties and aggregation levels define node properties. The combination of edge properties (i.e., citation, co-citation, co-word, coauthor, bibliographic coupling, or hybrid) and node properties (i.e., words, papers, patents, authors, journals, institutions, fields, or countries) precisely defines a network. Such networks are referred to as scholarly networks in this entry.

Scholarly networks provide an ideal research instrument to quantitatively study scholarly communication, in the areas of impact assessments (primarily through citation networks), scientific collaborations (primarily through collaboration networks), identifications of research specialties and topics (primarily through co-occurrence networks such as co-citation and bibliographic coupling networks), and examinations of knowledge flow patterns (primarily through citation networks).

Key Points

The key points in scholarly network analysis often involve the proper construction of networks given a set of research questions, the use of appropriate methods to analyze networks, and the accurate interpretations of obtained results. For instance, if one is interested in the global scientific collaboration, institution- or country-level collaboration networks work better than author-level networks given the research objectives. As will be introduced in detail in following paragraphs, there are three levels of approaches to analyze scholarly networks – each has its defined purposes. Last, accurate interpretations can also be critical. One should realize the limitations of the data and methods when interpreting results and should be aware of the unobserved factors that may contribute to the network structures.

Historical Background

The earliest well-defined network in scholarly communication is probably the paper bibliographic coupling network, proposed by Kessler in the 1960s (Kessler 1963). Paper co-citation was first proposed by Small (1973). Like bibliographic coupling networks, co-citation networks have the same goal of measuring the similarity between two scientific papers. Since then, various types of networks were proposed, for instance, citation networks, coauthorship networks, co-word networks, and hybrid networks.

Scholarly Networks as a Type of Networks

In an important review article on complex networks, Newman (2003) distinguished four kinds of real-world networks: social networks (e.g., collaboration networks), information networks (e.g., citation networks), technical networks (e.g., Internet router networks), and biological networks (e.g., protein-protein interaction networks). Based on this division, two types of scholarly networks can be distinguished: social networks and information networks. In social networks such as coauthorship networks, a node is a social actor (i.e., an author), while in information networks a node is typically an artifact, such as a paper, a journal, or an institution.

In addition to "social networks vs. information networks," another distinction can be made, which is "real connection-based networks vs. similarity-based networks." Coauthorship networks and citation networks are constructed based on real connections, whereas co-citation, bibliographic coupling, topical, and co-word networks are constructed based on similarity connections. Furthermore, these scholarly networks can also be viewed from their edge types: collaboration based, citation based, or word based. Citation-based scholarly networks include citation networks, co-citation networks, and bibliographic coupling networks; word-based scholarly networks include topical networks and co-word networks; collaboration-based networks include coauthorship networks. Those distinctions (social networks vs. information networks, real connection-based networks vs. similarity connection-based networks, and citation-based networks vs. non-citation-based networks) are helpful to understand how different types of scholarly networks relate to each other.

The Use of Scholarly Networks

Before network theories were introduced to scientometrics, accumulative citation counting was widely used in the area of impact assessment. In the same vein of research, several citation-based indicators were proposed, such as Journal Impact Factor and h-index (Hirsch 2005). The accumulative citation counting and citation-based indicators equate all citations to have the same weight, without considerations of the impact of citing papers, authors, or journals. This equal counting mechanism was questioned, as scholars argued that it is more reasonable to differentiate the weight of citations based on the source of endorsements (e.g., Pinski and Narin 1976; Bollen et al. 2006; Yan et al. 2011). This tension was largely alleviated by the construction of different types of scholarly networks and the invention of a variety of network-based bibliometric indicators. Compared to the traditional citation counting, scholarly networks have the advantage of considering the source of citation endorsements. In this way, scholarly networks can capture the complex research communication and interaction.

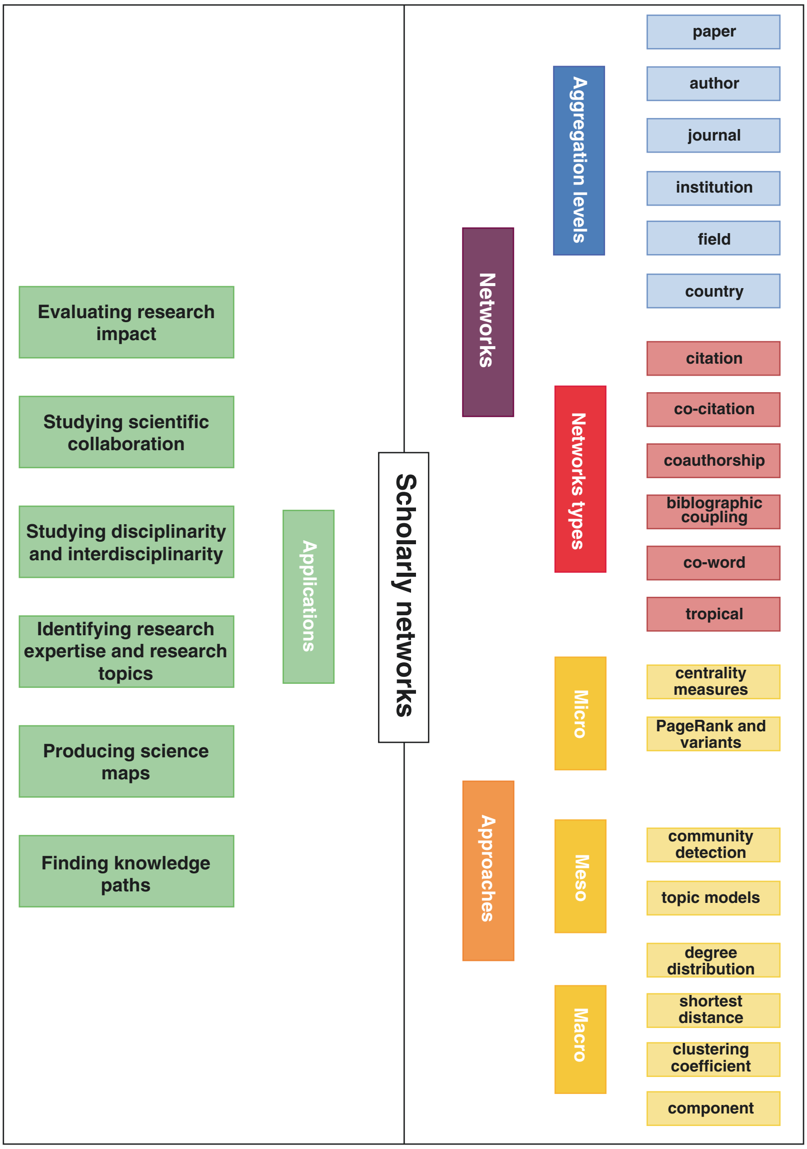
In addition to impact assessments, scholarly networks also contribute to other realms of scholarly communication and science policy making. For instance, coauthorship networks have been used to detect research communities and identify collaboration patterns (e.g., Newman and Girvan 2004); co-citation networks, bibliographic coupling networks, and co-word networks have been used to identify research specialties, examine interdisciplinary research, and map the backbone of science; and citation networks have been used to study knowledge flow and knowledge transfer in science and technology (e.g., Jaffe et al. 1993; Yan et al. 2013).

The Framework of Studying Scholarly Networks

Through scholarly network analysis, scientists and policy makers have gained unprecedented insights into the interaction of various research units. Studies of scholarly networks can be broadly presented in a framework (Fig. 1), including approaches, network-network types, network-aggregation levels, and applications.

Approaches

Given that we have established a scholarly network, we can describe its properties on three levels: macro-level metrics (global graph statistics), meso-level techniques (community characteristics), and microlevel metrics (individual actor properties). Macro-level metrics seek to describe the global characteristic and to capture the generic structural features of a network. Commonly used measures include diameter, density, mean distance, components, modularity, and degree distribution. Meso-level techniques focus on identifying research communities and studying how communities interact with each other. Micro-level metrics pertain to the analysis of the individual properties of network actors, for example, positions, status, and distances to others, which informs us about "the differential constraints and



Scholarly Network Analysis, Fig. 1 A framework of scholarly network studies

opportunities facing individual actors which shape their social behavior" (Yin et al. 2006, p. 1600). It zooms in to capture the features of the individual actors in a network with considerations of the topology of networks. Microlevel metrics typically include centrality measures, which dictate how central an actor is in a network.

Macro-level Macro-level metrics are useful to identify the global structural features of networks. There are many ways to characterize the structure of a network, such as component, bicomponent, *k*-core, mean distance, clustering coefficient, modularity, degree distribution, and error and attack tolerance of a network. In network analysis, connected graphs are called components.

- Component analysis can be used to learn about the macro-level structure of a network.
- In a bicomponent, no node can control the information flow between two other nodes

completely because there is always an alternative path that information may follow (Nooy et al. 2005).
- The *k*-core of a network is a substructure in which each node has ties to at least *k* other nodes (Seidman 1983).
- A geodesic is the shortest path between two nodes.
- The degree of a node is the number of other nodes connected with it. Degree distribution measures the character of a network. In a typical real-world network such as a coauthorship network, a few authors have many coauthors and majority have smaller numbers of coauthors.

Meso-level Meso-level scholarly network analyses focus on clustering various units to the same groups using certain clustering techniques. Clustering papers, authors, institutions, journals, and subject categories are usually referred to as community detection; and clustering words and research topics are usually referred to as topic identification. Broadly perceived, clustering techniques fall into two branches: one yields discrete results where a node in a scholarly network is grouped into one or multiple clusters, and the other branch yields fractional results where a node is grouped into clusters with certain probabilities. "Discrete" clustering techniques include graph partitioning (e.g., Kernighan-Lin algorithm), hierarchical clustering, partitional clustering (e.g., *k*-means), and spectral clustering (e.g., algorithms utilizing Laplacian matrices). In this decade, more clustering tasks have used modularity-based methods that use modules to measure the strength of communities. "Fractional" clustering techniques use probabilistic models to assign papers, journals, or authors to clusters. The outcomes of topic models are probability distributions of words, papers, journals, or authors for each topic (e.g., Blei et al. 2003).

Microlevel Freeman (1979) elaborated four concepts of centrality in social networks which have implications to scholarly networks as well. These concepts have since been further developed into degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality. Eigenvector is based on the principle that the importance of a node depends on the importance of its neighbors. PageRank can be seen as a type of eigenvector centrality derived from the influence weights proposed by Pinski and Narin (1976). PageRank is formally formulated by Brin and Page (1998), who developed a method of assigning a universal rank to Web pages based on a weight-propagation algorithm called PageRank. A page has a high rank if the sum of the ranks of its backlinks is high. Actors in the PageRank of Web information retrieval systems are Web pages, and actors in the PageRank of citation networks are papers. The underlying idea is that a citation from an influential publication, a prestigious journal, or a renowned author should be regarded as more valuable than a citation from an insignificant publication, an obscure journal, or an unknown author. It is sometimes argued that non-recursive indicators (such as h-index and Journal Impact Factor) measure popularity and recursive indicators (such as PageRank) measure prestige.

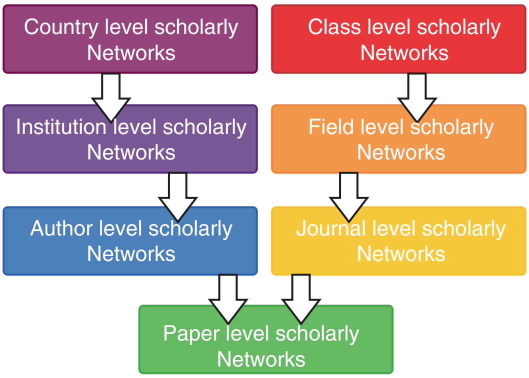
Network Types

In addition to the different levels of approaches, interactions of research units can be explored from different types of scholarly networks. Each type has its own use and can bring different perspectives to study scholarly communication. For example, social networks such as coauthorship networks focus on finding collaboration patterns of contacts or interactions among social actors. Similarity-based networks such as co-citation networks, bibliographic coupling networks, and co-word networks focus on identifying research topics or schools of thoughts. In citation networks, each node can be used to represent a piece of knowledge, and a link can be used to denote knowledge flow (Yan and Ding 2012).

Aggregation Levels

An article is a single research unit and can be aggregated into several higher levels. Figure 2 shows the different aggregation levels discussed in scholarly networks.

The right side of the cascade is connected through "journal-ship" affiliation: a paper is


Country level scholarly Networks
Class level scholarly Networks
Institution level scholarly Networks
Field level scholarly Networks
Author level scholarly Networks
Journal level scholarly Networks
Paper level scholarly Networks

Scholarly Network Analysis, Fig. 2 Aggregation levels of scholarly networks

published in a journal, a journal is classified into a subject category, and a subject category is further classified into a higher-level, more abstract class. The left side of the cascade is connected through authorship affiliation: a paper is written by an author, an author is affiliated to an institution, and an institution is located in a country. Through studies of different research units, we are provided with multiple focus lenses that allow us to zoom in and gain a concrete, detailed perspective on research interactions, while zooming out allows us to obtain holistic and integrated views of interacting institutions and disciplines.

Key Applications

Scholarly networks have rich applications in studies of scholarly communication. In this section, we show six major applications and give brief introductions.

Evaluating Research Impacts

Impact evaluation has become an important issue in the science community. Scientists as well as policy makers have a keen interest in evaluating scientific outputs. For scientists, evaluations of research impacts help them find potential collaborations, discover new research topics, and locate appropriate venues to publish their work. For science policy makers, evaluations of research impacts help inform how to allocate research funds and promote emerging research fields. Traditional citation-based bibliometric indicators do not consider the source of citation endorsements. However, in reality, being cited by a renowned author, a prestigious journal, and/or a highly influential paper differs from being cited by a remote author, a peripheral journal, and/or an obscure paper. Network-based bibliometric indicators are capable of considering the provenance of citation endorsements – specifically PageRank and its variants have gained popularity in evaluating research impacts. PageRank-like indicators denote a collection of algorithms based on Google's PageRank, such as Y-factor (Bollen et al. 2006), CiteRank (Walker et al. 2007), Eigenfactor (Bergstrom and West 2008), and SCImago Journal Rank (SCImago 2007). Among these network-based bibliometric indicators, citations are weighed differently depending on the status of the citing publication (e.g., Walker et al. 2007), the citing journal (e.g., Bollen et al. 2006; Pinski and Narin 1976), or the citing author (e.g., Radicchi et al. 2009). A detailed description of PageRank-related methods can be found at Waltman and Yan (2014).

Studying Scientific Collaborations

Scientific collaborations, a large-scale real-world social phenomenon, exhibit interests to scientists and scholars. Coauthorship networks provide an accurate and expedite medium, allowing scientists and scholars to explore various intriguing questions pertinent to this social phenomenon. Physicists and mathematicians have discovered the small-world and scale-free properties from coauthorship networks, for the first time providing a systematic inquiry into humans' social relationships. Later on, coauthorship networks have been used as a testing field for various clustering techniques (e.g., Newman and Girvan 2004). Such techniques are useful to examine scientific collaborations at a more granular level, providing insights to study the science of team science.

Studying Disciplinarity and Interdisciplinarity

Interdisciplinarity has long been a research focus for scientists. The quantitative study of interdisciplinarity has been enhanced by the availability of large-scale citation networks aggregated at the field level. Scholars usually choose representative journals, or all journals from a field based on certain journal classification schemes (e.g., the Thomson Reuter's classification), and then measure the extent to which the publications of the chosen field cited the publications of other fields. Network-based indicators have been employed to measure how interdisciplinary disciplines are, using measures such as entropy (Zhang et al. 2010), integration and specialization (Porter et al. 2006), diversity and coherence (Rafols and Meyer 2010), and relative openness (Rinia et al. 2002).

Identifying Research Expertise and Research Topics

Human knowledge, in the form of scholarly publications, increases at a fast pace. How to effectively organize the expanding knowledge has become an important issue. Hence, scholars proposed a variety of clustering techniques to group papers, authors, journals, institutions, and fields, to identify and organize research specialty in effective ways. For similarity-based scholarly networks such as co-citation and bibliographic coupling networks, the assumption is that if two research units co-occur frequently, they are more likely to have similar characteristics. Therefore, co-occurrence networks can successfully identify and organize scientific knowledge (e.g., White and McCain 1998; Boyack et al. 2005; Waltman et al. 2010).

Producing Science Maps

Clustering results can be presented in science maps, and these maps are able to deliver richer and more informative messages to a broader audience body. Science maps on author and journal communications (e.g., via collaboration and co-citation relations) typically are used to identify research topics (e.g., Boyack et al. 2005). As institutions are associated with geographical locations, science maps at the institution level are useful to illustrate the geographical distribution of scientific productivity and knowledge (e.g., Leydesdorff and Persson 2010). Science maps at the field level provide a unique view on the backbone of science (e.g., Boyack et al. 2005) or on the knowledge flow in scientific disciplines (e.g., Rosvall and Bergstrom 2008).

Finding Knowledge Paths

The production and creation of knowledge is not dependent on a single, isolated unit; instead, knowledge is diffused, exchanged, and circulated among various units. Knowledge flow, in the past 20 years, is becoming more inter-sectoral, interorganizational, interdisciplinary, and international. Thus, the following issues are becoming pertinent to understanding patterns of knowledge transfer and dissemination: how do scientific and technological knowledge, innovative ideas, management skills, or certain influences transfer within different sectors, between different organizations, and between different scientific disciplines. Citation networks serve as an ideal research instrument to uncover patterns of knowledge diffusion. In citation networks, a node is a research unit, and a link denotes a citation from the citing research unit to the cited unit. Using citation networks, scholars have examined disciplinary knowledge trading characteristics, dynamics, and diversity (Bettencourt et al. 2008; Yan 2015) and identified the most important knowledge hubs and knowledge paths (Yan and Yu 2015). A variety of approaches were designed and employed, such as knowledge trading metaphor (Yan et al. 2013), maximum spanning tree (Yan and Yu 2015), Shannon diversity (Yan 2015), epidemiological models (Bettencourt et al. 2008; Kiss et al. 2010), the clique percolation method (Herrera et al. 2010), small-world models (Cowan and Jonard 2004), and diffusion models (Zhuang et al. 2011).

Future Directions

Heterogeneous Scholarly Networks

Studies of scholarly networks usually choose one type of networks at one aggregation level. The choice of the type of networks can be inconsistent or even arbitrary, and the findings have been discrete and cannot be generalized to address a wider spectrum of research questions. We recommend that, in order to capture the varied aspects of research interactions, different types of networks can be integrated and thus form hybrid networks. Beyond the hybrid approach, scholars proposed heterogeneous scholarly networks to incorporate different units while keeping edge semantics (Sun et al. 2013). By adding multiple units (e.g., authors, journals, articles, words) in analyses, heterogeneous networks can capture more expansively the mutual engagement of various research units in complex academic environments.

Therefore, future research in this direction would benefit from (1) constructing hybrid and heterogeneous scholarly networks and (2) evaluating different approaches on hybrid or heterogeneous scholarly networks through possible "golden standards" (such as award lists or expert judgments) to determine which approaches can yield more precise clustering results and more useful information for studies of scholarly communication.

Entity-Based Scholarly Networks

Traditionally, papers are employed as the unit of analysis in scholarly network analytics, and this unit can be aggregated into several higher levels based on journal associations or authorship, as mentioned in the preceding paragraph. While these units of analysis delivered rich analyses of scholarly communication, knowledge is more effectively expressed through unstructured or semi-structured fields, such as titles, abstracts, keywords, or full texts. Thus, researchers and practitioners are demanding more fine-grained methods and tools to contextualize findings and make sense of bibliometric indicators and numbers. Therefore, another future research direction is to tackle the complexity of textual data and examine content-rich knowledge entities – expressions in the text that convey some discriminatory information about the research-relevant aspects of documents – such as theories, concepts, artifacts, algorithms, and methods. These entities, viewed as the essential cells of scientific literature and building blocks of knowledge, will drastically accelerate our understanding of science of science. Accordingly, examinations of entity-based scholarly networks (e.g., entity citation networks and entity co-occurrence networks) will reveal science of science at a new granular level and address questions on the provenance, diffusion, coevolution, trend, and impact of knowledge at a heretofore unexplored extent and depth. In addition, entity-based scholarly networks will enable us to merge heterogeneous data sources as well as disambiguate and integrate knowledge entities across different data sources, disciplines, organizations, and geographic locations.

Cross-References

- ▶ Analysis and Visualization of Dynamic Networks
- ▶ Classical Algorithms for Social Network Analysis: Future and Current Trends
- ▶ Components of the Network Around an Actor
- ▶ Detecting and Identifying Communities in Dynamic and Complex Networks: Definition and Survey
- ▶ Link Dynamics and Community Formation in Social Networks
- ▶ Node Centrality
- ▶ Similarity Metrics on Social Networks
- ▶ Social Interaction Analysis for Team Collaboration

References

Bergstrom CT, West JD (2008) Assessing citations with the Eigenfactor™ metrics. Neurology 71(23):1850–1851

Bettencourt LMA, Kaiser DI, Kaur J, Castillo-Chávez C, Wojick DE (2008) Population modeling of the emergence and development of scientific fields. Scientometrics 75(3):495–518

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3(4–5):993–1033

Bollen J, Rodriguez MA, Van de Sompel H (2006) Journal status. Scientometrics 69(3):669–687

Boyack KW, Klavans AR, Börner K (2005) Mapping the backbone of science. Scientometrics 64(3):351–374

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30(117):107

Cowan R, Jonard N (2004) Network structure and the diffusion of knowledge. J Econ Dyn Control 28(8):1557–1575

Freeman LC (1979) Centrality in social networks: conceptual clarification. Soc Netw 1(3):215–239

Herrera M, Roberts DC, Gulbahce N (2010) Mapping the evolution of scientific fields. PLoS ONE 5(5):6

Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci USA 102(46):16569–16572

Jaffe AB, Trajtenberg M, Henderson AD (1993) Geographical localization of knowledge spillovers by patent citations. Q J Econ 108(3):577–599

Kessler MM (1963) Bibliographic coupling between scientific papers. Am Doc 14(1):10–25

Kiss IZ, Broom M, Craze PG, Rafols I (2010) Can epidemic models describe the diffusion of topics across disciplines? J Informet 4(1):74–82

Leydesdorff L, Persson O (2010) Mapping the geography of science: distribution patterns and networks of relations among cities and institutes. J Am Soc Inf Sci Technol 61(8):1622–1634

Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256

Newman M, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026113

Nooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Cambridge University Press, Cambridge

Pinski G, Narin F (1976) Citation influence for journal aggregates of scientific publications: theory, with application to the literature of physics. Inf Process Manag 12 (5):297–312

Porter AL, Roessner JD, Cohen AS, Perreault M (2006) Interdisciplinary research: meaning, metrics and nurture. Res Eval 15(3):187–195

Radicchi F, Fortunato S, Markines B, Vespignani A (2009) Diffusion of scientific credits and the ranking of scientists. Phys Rev E 80(5):056103

Rafols I, Meyer M (2010) Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. Scientometrics 82(2): 263–287

Rinia EJ, Van Leeuwen TN, Bruins EEW, Van Vuren HG, Van Raan AFJ (2002) Measuring knowledge transfer between fields of science. Scientometrics 54(3): 347–362

Rosvall M, Bergstrom CT (2008) Maps of information flow reveal community structure in complex networks. Proc Natl Acad Sci USA 105(4):1118–1123

SCImago (2007) SJR: SCImago Journal & Country Rank. http://www.scimagojr.com. Retrieved 31 Aug 2009

Seidman SB (1983) Network structure and minimum degree. Soc Netw 5:269–287

Small H (1973) Co-citation in the scientific literature: A new measure of the relationship between two documents. J Am Soc Inf Sci 24(4):265–269

Sun Y, Norick B, Han J, Yan X, Yu PS, Yu X (2013) Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. ACM Trans Knowl Discov Data (TKDD) 7(3):11

Walker D, Xie H, Yan KK, Maslov S (2007) Ranking scientific publications using a simple model of network traffic. J Stat Mech: Theory Exp P06010. https://doi.org/10.1088/1742-5468/2007/06/P06010

Waltman L, Van Eck NJ, Noyons ECM (2010) A unified approach to mapping and clustering of bibliometric networks. J Informetr 4(4):629–635

Waltman L, Yan E (2014) PageRank-related methods for analyzing citation networks. In: Ding Y, Rousseau R, Wolfram D (eds) Measuring scholarly impact. Springer International Publishing, Cham, pp 83–100

White HD, McCain KW (1998) Visualizing a discipline: an author co-citation analysis of information science 1972–1995. J Am Soc Inf Sci 49(4):327–355

Yan E, Ding Y (2012) Scholarly network similarities: how bibliographic coupling networks, citation networks, co-citation networks, topical networks, coauthorship networks, and co-word networks relate to each other. J Am Soc Inf Sci Technol 63(7):1313–1326

Yan E, Ding Y, Sugimoto CR (2011) P-Rank: an indicator measuring prestige in heterogeneous scholarly networks. J Am Soc Inf Sci Technol 62(3): 467–477

Yan E, Ding Y, Cronin B, Leydesdorff L (2013) A bird's eye view of scientific trading: dependency relations among fields of science. J Informetr 7(2):249–264

Yan E (2015) Disciplinary knowledge production and diffusion in science. J Assoc Inf Sci Technol. https://doi.org/10.1002/asi.23541

Yan E, Yu Q (2015) Using path-based approaches to examine the dynamic structure of discipline-level citation networks: 1997–2011. J Assoc Inf Sci Technol. https://doi.org/10.1002/asi.23516

Yin L, Kretschmer H, Hanneman RA, Liu Z (2006) Connection and stratification in research collaboration: an analysis of the COLLNET network. Inf Process Manag 42(6):1599–1613

Zhang L, Liu X, Janssens F, Liang L, Glänzel W (2010) Subject clustering analysis based on ISI category classification. J Informetr 4(2):185–193

Zhuang E, Chen G, Feng G (2011) A network model of knowledge accumulation through diffusion and upgrade. Physica A: Statistical Mechanics and its Applications 390(13):2582–2592

Scholarly Networks

► Scholarly Network Analysis

Science Maps

▶ Scholarly Network Analysis

Science of Science (Sci2) Tool

▶ Plug-and-Play Macroscopes: Network Workbench (NWB), Science of Science Tool (Sci2), and Epidemiology Tool (EpiC)

Science of the Internet

▶ Web Science

Science of the Web

▶ Web Science

Scientific Collaboration

▶ Scholarly Network Analysis

Scientific Collaboration Network

▶ Link Dynamics and Community Formation in Social Networks

Scientific Communities

▶ Stability and Evolution of Scientific Networks

Scientific Evaluation

▶ Scholarly Network Analysis

Search and Querying Social Data

▶ Social Search and Querying

Search Engine

▶ Weblog Analysis

Search Missions

▶ Weblog Analysis

Second Life

▶ Social Media, Definition, and History

Security

▶ Crime Prevention, Dataveillance, and the Regulation of Information Communication Technologies
▶ Reconnaissance and Social Engineering Risks as Effects of Social Networking

Self-Confidence

▶ Self-Efficacy Versus Expertise

Self-Disclosure

▶ Privacy and Disclosure in a Social Networking Community
▶ Reconnaissance and Social Engineering Risks as Effects of Social Networking

Self-Efficacy Versus Expertise

Donghee Yvette Wohn and Chandan Sarkar
Department of Telecommunication, Information Studies and Media, Michigan State University, East Lansing, MI, USA

Synonyms

Self-confidence; Self-perception

Glossary

Expertise	Knowledge and actual skills
Self-efficacy	An individual's confidence about his or her skills

Definition

Self-efficacy (Bandura 1977a, b) is an individual's self-perception of his or her ability. By placing importance on the individual's perception as opposed to the individual's actual skill, this construct can explain why people have different behaviors even if they have a similar skill set. In much of social-psychological research, selfefficacy serves as a good proxy of predicting people's behaviors because it looks not only at perceived expertise (knowledge and actual skills) about a certain behavior but also perceived confidence. However, in the context of behaviors required to ensure privacy and security in an online environment, confidence in one's ability may not necessarily be the best factor that explains behavior. Because privacy behaviors, such as changing privacy settings and employing preventive security measures, require a certain degree of technical expertise, perceived expertise and perceived confidence can be false indicators. For example, one may have extremely strong confidence in one's ability but could very well be overestimating that ability. In the context of Internet privacy, this differentiation is important as privacy protection requires certain technical skills.

Efficacy beliefs are the product of a complex process of self-persuasion that rely on cognitive processing of diverse sources of information. Research has found that self-efficacy is an important construct that explains an individual's attitude about privacy, which ultimately affects their behavior (Rifon et al. 2005). A limitation for studies that examine self-efficacy, however, is that there was little consideration for actual expertise.

In semi-structured in-depth interviews with young adults aged 20–30, we found that the reality of how users process information and behave in relation to privacy and security issues online was sometimes inconsistent with their self-efficacy, especially among those with low levels of expertise. Expertise was determined by an individual's actual understanding of the technical aspects of privacy issues through a set of questions to participants answered explaining constructs such as phishing, Trojans, cookies, and browser privacy settings. Self-efficacy was measured asking participants if they consider themselves knowledgeable about the security risks and threats to privacy that exist online. We found that the reality of how users process information and behave in relation to privacy and security issues online was sometimes inconsistent with their self-efficacy, especially among those with low levels of expertise. In many cases, there was a difference between an individual's self-efficacy and expertise; novices were more likely to overestimate their self-efficacy while experts were more likely to underestimate it. It was expertise, not self-efficacy, that was a stronger predictor of their attitude and behavior. For example, participants with high self-efficacy but low expertise showed extreme caution and concern regarding the implication of privacy policies and were more upset about behavioral targeting than those with high knowledge, whereas participants with higher self-efficacy and high expertise showed accepting behavior towards behavioral targeting.

Individuals' abilities to change privacy settings, set up firewalls, and use security software among others were key indicators to how they perceived privacy issues and how they acted to deal with those issues. Novices were fearful of privacy and security threats and relied more on peripheral cues such as privacy seals and brand names to make their judgment. Experts were less concerned about such threats and interpreted peripheral cues differently from novices. Across different topics, we consistently saw differences between those who had expertise and those who did not.

Although this seems to challenge studies that show self-efficacy as a strong predictor of behavior, it may be that there are different dimensions of self-efficacy. For example, self-efficacy of behavior (how to run antivirus software) may be high, but self-efficacy of underlying concepts or mechanisms (how the software works) may be low. Thus, from a theoretical perspective, we suggest that in the context of privacy studies, researchers measure individuals' actual expertise in addition to self-efficacy, as expertise may play a moderating role in predicting behavior. From a practical perspective, the distinction between self-efficacy and expertise can inform how and what we should teach people about privacy and security issues. People using social networks face many challenges in terms of privacy and security threats. There are no regulations in terms of legal enforcement to what extent personal information can be gathered and used by the social network providers for purposes such as behavioral targeting. The average user has uncertainty about what kind of personal data is collected and for what purpose, let alone where that information is being sold or how they should protect themselves. Furthermore, most social network services collect network data, in which case information that a user reveals to another thinking it is private could still end up being collected and sold to third parties. Differentiating the true expertise from the self-perception of self-efficacy may enable us to identify individuals who are at high risk – those who think they know much but actually don't.

Cross-References

- ▶ Privacy in Social Networks, Current and Future Research Trends on
- ▶ User Behavior in Online Social Networks: Influencing Factors

References

Bandura A (1977a) Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev 84 (2):191–215

Bandura A (1977b) BibBook. Self-efficacy: the exercise of control. Freeman, New York

Rifon NJ, LaRose R, Choi SM (2005) Your privacy is sealed: effects of web privacy seals on trust and false assurances. J Consum Aff 39(2):337–360

Self-Perception

- ▶ Self-Efficacy Versus Expertise

Self-Presentation

- ▶ Privacy and Disclosure in a Social Networking Community

Semantic Networks

- ▶ NetMiner
- ▶ Semantic Social Networks Analysis

Semantic Perception

- ▶ Twitris: A System for Collective Social Intelligence

Semantic Sentiment Analysis of Twitter Data

Preslav Nakov
Qatar Computing Research Institute, HBKU,
Doha, Qatar

Synonyms

Microblog sentiment analysis; Twitter opinion mining

Glossary

Sentiment Analysis	This is text analysis aiming to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a piece of text

Definition

Sentiment analysis on Twitter is the use of natural language processing techniques to identify and categorize opinions expressed in a tweet, in order to determine the author's attitude toward a particular topic or in general. Typically, discrete labels such as *positive*, *negative*, *neutral*, and *objective* are used for this purpose, but it is also possible to use labels on an ordinal scale, or even continuous numerical values.

Introduction

Internet and the proliferation of smart mobile devices have changed the way information is created, shared, and spread, e.g., microblogs such as Twitter, weblogs such as LiveJournal, social networks such as Facebook, and instant messengers such as Skype and WhatsApp are now commonly used to share thoughts and opinions about anything in the surrounding world. This has resulted in the proliferation of social media content, thus creating new opportunities to study public opinion at a scale that was never possible before.

Naturally, this abundance of data has quickly attracted business and research interest from various fields including marketing, political science, and social studies, among many others, which are interested in questions like these: *Do people like the new Apple Watch? What do they hate about iPhone6? Do Americans support ObamaCare? What do Europeans think of Pope's visit to Palestine? How do we recognize the emergence of health problems such as depression? Do Germans like how Angela Merkel is handling the refugee crisis in Europe? What do republican voters in USA like/hate about Donald Trump? How do Scottish feel about the Brexit?*

Answering these questions requires studying the sentiment of opinions people express in social media, which has given rise to the fast growth of the field of sentiment analysis in social media, with Twitter being especially popular for research due to its scale, representativeness, variety of topics discussed, as well as ease of public access to its messages (Java et al. 2007; Kwak et al. 2010).

Despite all these opportunities, the rise of social media has also presented new challenges for natural language processing (NLP) applications, which had largely relied on NLP tools tuned for formal text genres such as newswire, and thus were not readily applicable to the informal language and style of social media. That language proved to be quite challenging with its use of creative spelling and punctuation, misspellings, slang, new words, URLs, and genre-specific terminology and abbreviations, e.g., RT for re-tweet and #hashtags. In addition to the genre difference, there is also a difference in length: social media messages are generally short, often length limited by design as in Twitter, i.e., a sentence or a headline rather than a full document. How to handle such challenges has only recently been the subject of thorough research (Barbosa and Feng 2010; Bifet et al. 2011; Davidov et al. 2010b; Jansen et al. 2009; Kouloumpis et al. 2011; O'Connor et al. 2010; Pak and Paroubek 2010; Tumasjan et al. 2010).

Key Points

Sentiment analysis has a wide number of applications in areas such as market research, political and social sciences, and for studying public opinion in general, and Twitter is one of the most commonly used platforms for this. This is due to its streaming nature, which allows for real-time analysis; to its social aspect, which encourages people to share opinions; and to the short size of the tweets, which simplifies linguistic analysis.

There are several formulations of the task of Sentiment Analysis on Twitter that look at different sizes of the target (e.g., at the level of words vs. phrases vs. tweets vs. sets of tweets), at different types of semantic targets (e.g., aspect vs. topic vs. overall tweet), at the explicitness of the target (e.g., sentiment vs. stance detection), at the scale of the expected label (2-point vs. 3-point vs. ordinal), etc. All these are explored at SemEval, the International Workshop on Semantic Evaluation, which has created a number of benchmark datasets and has enabled direct comparison between different systems and approaches, both as part of the competition and beyond.

Traditionally, the task has been addressed using supervised and semi-supervised methods, as well as using distant supervision, with the most important resource being sentiment polarity lexicons, and with feature-rich approaches as the dominant research direction for years. With the recent rise of deep learning, which in many cases eliminates the need for any explicit feature modeling, the importance of both lexicons and features diminishes, while at the same time attention is shifting toward learning from large unlabeled data, which is needed to train the high number of parameters of such complex models. Finally, as methods for sentiment analysis mature, more attention is also being paid to linguistic structure and to multi-linguality and cross-linguality.

Historical Background

Sentiment analysis emerged as a popular research direction in the early 2000s. Initially, it was regarded as standard document classification into topics such as business, sport, and politics (Pang et al. 2002). However, researchers soon realized that it was quite different from standard document classification (Sebastiani 2002), and that it crucially needed external knowledge in the form of sentiment polarity lexicons.

Around the same time, other researchers realized the importance of external sentiment lexicons, e.g., Turney (2002) proposed an unsupervised approach to learn the sentiment orientation of words/phrases: positive vs. negative. Later work studied the linguistic aspects of expressing opinions, evaluations, and speculations (Wiebe et al. 2004), the role of context in determining the sentiment orientation (Wilson et al. 2005), of deeper linguistic processing such as negation handling (Pang and Lee 2008), of finer-grained sentiment distinctions (Pang and Lee 2005), of positional information (Raychev and Nakov 2009), etc. Moreover, it was recognized that in many cases, it is crucial to know not just the polarity of the sentiment but also the topic toward which this sentiment is expressed (Stoyanov and Cardie 2008).

Until the rise of social media, research on opinion mining and sentiment analysis had focused primarily on learning about the language of sentiment in general, meaning that it was either genre-agnostic (Baccianella et al. 2010) or focused on newswire texts (Wiebe et al. 2005) and customer reviews (e.g., from web forums), most notably about movies (Pang et al. 2002) and restaurants (Pontiki et al. 2014) but also about hotels, digital cameras, cell phones, MP3 and DVD players (Hu and Liu 2004), laptops (Pontiki et al. 2014), etc. This has given rise to several resources, mostly word and phrase polarity lexicons, which have proven to be very valuable for their respective domains and types of texts, but less useful for short social media messages.

Later, with the emergence of social media, sentiment analysis in Twitter became a hot research topic. Unfortunately, research in that direction was hindered by the unavailability of suitable datasets and lexicons for system training, development, and testing. While some Twitter-specific resources were developed, initially they were either small and proprietary, such as the i-sieve corpus (Kouloumpis et al. 2011), were created only for Spanish like the TASS corpus (Villena-Roman et al. 2013), or relied on noisy labels obtained automatically, e.g., based on emoticons and hashtags (Mohammad 2012; Mohammad et al. 2013; Pang et al. 2002).

This situation changed with the shared task on *Sentiment Analysis on Twitter*, which was organized at SemEval, the International Workshop on Semantic Evaluation, a semantic evaluation forum previously known as SensEval. The task ran in 2013, 2014, 2015, and 2016, attracting over 40 participating teams in all four editions. While the focus was on general tweets, the task also featured out-of-domain testing on SMS messages, LiveJournal messages, as well as on sarcastic tweets.

SemEval-2013 Task 2 (Nakov et al. 2013) and SemEval-2014 Task 9 (Rosenthal et al. 2014) focused on expression-level and message-level polarity. SemEval-2015 Task 10 (Nakov et al. 2016b; Rosenthal et al. 2015) featured topic-based message polarity classification on detecting trends toward a topic and on determining the out-of-context (a priori) strength of association of Twitter terms with positive sentiment. SemEval-2016 Task 4 (Nakov et al. 2016a) introduced a 5-point scale, which is used for human review ratings on popular websites such as Amazon, TripAdvisor, Yelp, etc.; from a research perspective, this meant moving from classification to *ordinal regression*. Moreover, it focused on *quantification*, i.e., determining what proportion of a set of tweets on a given topic are positive/negative about it. It also featured a 5-point scale *ordinal quantification* subtask (Gao and Sebastiani 2015).

Other related tasks have explored aspect-based sentiment analysis (Pontiki et al. 2014, 2015, 2016), sentiment analysis of figurative language on Twitter (Ghosh et al. 2015), implicit event polarity (Russo et al. 2015), stance in tweets (Mohammad et al. 2016), out-of-context sentiment intensity of phrases (Kiritchenko et al. 2016), and emotion detection (Strapparava and Mihalcea 2007). Some of these tasks featured languages other than English.

Sentiment Analysis on Twitter: A SemEval Perspective

Variants of the Task at SemEval

Tweet-level sentiment. The simplest and also the most popular task of sentiment analysis on Twitter is to determine the overall sentiment expressed by the author of a tweet (Nakov et al. 2013, 2016a, b; Rosenthal et al. 2014, 2015). Typically, this means choosing one of the following three classes to describe the sentiment: POSITIVE, NEGATIVE, and NEUTRAL. Here are some examples:

(1). POSITIVE: @nokia lumia620 cute and small and pocket-size, and available in the brightest colours of day! #lumiacaption

(2). NEGATIVE: *I hate tweeting on my iPhone 5 it's so small :(*

(3). NEUTRAL: *If you work as a security in a samsung store . . . Does that make you guardian of the galaxy??*

Sentiment polarity lexicons. Naturally, the overall sentiment in a tweet can be determined based on the sentiment-bearing words and phrases it contains as well as based on emoticons such as;) and :(. For this purpose, researchers have been using lexicons of sentiment-bearing words. For example, *cute* is a positive word, while *hate* is a negative one, and the occurrence of these words in (1) and (2) can help determine the overall polarity of the respective tweet. We will discuss these lexicons in more detail below.

Prior sentiment polarity of multi-word phrases. Unfortunately, many sentimentbearing words are not universally good or universally bad. For example, the polarity of an adjective could depend on the noun it modifies, e.g., *hot coffee* and *unpredictable story* express positive sentiment, while *hot beer* and *unpredictable steering* are negative. Thus, determining the out-of-context (a priori) strength of association of Twitter terms, especially multi-word terms, with positive/negative sentiment is an active research direction (Nakov et al. 2016b; Rosenthal et al. 2015).

Phrase-level polarity in context. Even when the target noun is the same, the polarity of the modifying adjective could be different in different tweets, e.g., *small* is positive in (1) but negative in (2), even though they both refer to a phone. Thus, there has been research in determining the sentiment polarity of a term in the context of a tweet (Nakov et al. 2013; Rosenthal et al. 2015, 2014).

Sarcasm. Going back to tweet-level sentiment analysis, we should mention sarcastic tweets, which are particularly challenging as the sentiment they express is often the opposite of what the words they contain suggest (Davidov et al. 2010b; Rosenthal et al. 2014, 2015). For example,

(1) and (5) express a negative sentiment even though they contain positive words and phrases such as *thanks*, *love*, and*boosts my morale*.

(4). NEGATIVE: *Thanks manager for putting me on the schedule for Sunday.*

(5). NEGATIVE: *I just love missing my train every single day. Really boosts my morale.*

Sentiment toward a topic. Even though tweets are short, as they are limited to 140 characters by design (even though this was relaxed a bit as of September 19, 2016, and now media attachments such as images, videos, polls, etc., and quoted tweets no longer reduce the character count), they are still long enough to allow the tweet's author to mention several topics and to express potentially different sentiment toward each of them. A topic can be anything that people express opinions about, for example, a product (e.g., *iPhone6*), a political candidate (e.g., *Donald Trump*), a policy (e.g., *Obamacare*), an event (e.g., *Brexit*), etc. For example, in (6) the author is positive about Donald Trump but negative about Hillary Clinton. A political analyzer would not be interested so much in the overall sentiment expressed in the tweet (even though one could argue that here it is positive overall), but rather in the sentiment with respect to a topic of his/her interest of study.

(6). *As a democrat I couldn't ethically support Hillary no matter who was running against her. Just so glad that its Trump, just love the guy!*

(topic: *Hillary* → NEGATIVE).

(topic: *Trump* → POSITIVE).

Aspect-based sentiment analysis. Looking again at (1) and (2), we can say that the sentiment is not about the phone (*lumia620* and *iPhone 5*, respectively) but rather about some specific aspect thereof, namely, SIZE. Similarly, in (7) instead of sentiment toward the topic *lasagna*, we can see sentiment toward two aspects thereof: QUALITY (POSITIVE sentiment) and QUANTITY (NEGATIVE sentiment). Aspect-based sentiment analysis is an active research area (Pontiki et al. 2014, 2015, 2016).

(7). *The lasagna is delicious but do not come here on an empty stomach.*

Stance detection. A task related to, but arguably different in some respect from sentiment analysis, is that of *stance detection*. The goal here is to determine whether the author of a piece of text is in favor of, against, or neutral toward a proposition or a target (Mohammad et al. 2016). For example, in (8) the author has a negative stance toward the proposition *women have the right to abortion*, even though the target is not mentioned at all. Similarly, in (9) the author expresses a negative sentiment toward *Mitt Romney*, from which one can imply that she/he has a positive stance toward the target *Barack Obama*.

(8). *A foetus has rights too! Make your voice heard.*

(Target: *women have the right to abortion* → AGAINST)

(9). *All Mitt Romney cares about is making money for the rich.*

(Target: *Barack Obama* → INFAVOR).

Ordinal regression. The above tasks were offered in different granularities, e.g., 2-way (POSITIVE, NEGATIVE), 3-way (POSITIVE, NEUTRAL, NEGATIVE), 4-way (POSITIVE, NEUTRAL, NEGATIVE, OBJECTIVE), 5-way (HIGHLYPOSITIVE, POSITIVE, NEUTRAL, NEGATIVE, HIGHLYNEGATIVE), and sometimes even 11-way (Ghosh et al. 2015). It is important to note that the 5-way and the 11-way scales are ordinal, i.e., the classes can be associated with numbers, e.g., −2, −1, 0, 1, and 2 for the 5-point scale. This changes the machine learning task as not all mistakes are equal anymore (Pang and Lee 2005). For example, misclassifying a HIGHLYNEGATIVE example as HIGHLYPOSITIVE is a bigger mistake than misclassifying it as NEGATIVE or as NEUTRAL. From a machine learning perspective, this means moving from *classification* to *ordinal regression*. This also requires different evaluation measures (Nakov et al. 2016a).

Quantification. Practical applications are hardly ever interested in the sentiment expressed in a *specific tweet*. Rather, they look at estimating the *prevalence* of positive and negative tweets about a given topic in a set of tweets from some time interval. Most (if not all) tweet sentiment classification studies conducted within political science (Borge-Holthoefer et al. 2015; Kaya et al. 2013; Marchetti-Bowick and Chambers 2012), economics (Bollen et al. 2011; O'Connor et al. 2010), social science (Dodds et al. 2011), and market research (Burton and Soboleva 2011; Qureshi et al. 2013) use Twitter with an interest in aggregate data and *not* in individual classifications. Thus, some tasks, such as SemEval-2016 Task 4 (Nakov et al. 2016a), replace classification with class prevalence estimation, which is also known as *quantification* in data mining and related fields. Note that quantification is not a mere byproduct of classification, since a good classifier is not necessarily a good quantifier, and vice versa (Forman 2008). Finally, in case of multiple labels on an ordinal scale, we have yet another machine learning problem: *ordinal quantification*. Both versions of quantification require specific evaluation measures and machine learning algorithms.

Features and Learning

Pre-processing. Tweets are subject to standard preprocessing steps for text such as tokenization, stemming, lemmatization, stop-word removal, and part-of-speech tagging. Moreover, due to their noisy nature, they are also processed using some Twitter-specific techniques such as substitution/removal of URLs, of user mentions, of hashtags, and of emoticons, spelling correction, elongation normalization, abbreviation lookup, punctuation removal, detection of amplifiers and diminishers, negation scope detection, etc. For this, one typically uses Twitter-specific NLP tools such as part-of-speech and named entity taggers, syntactic parsers, etc. (Gimpel et al. 2011; Kong et al. 2014; Ritter et al. 2011).

Negation handling. Special handling is also done for negation. The most popular approach to negation handling is to transform any word that appeared in a negation context by adding a suffix *NEG* to it, e.g., *good* would become *good NEG* (Das and Chen 2007; Pang et al. 2002). A negated context is typically defined as a text span between a negation word, e.g., *no, not, shouldn't*, and a punctuation mark or the end of the message. Alternatively, one could flip the polarity of sentiment words, e.g., the positive word *good* would become negative when negated. It has also been argued (Zhu et al. 2014a) that negation affects different words differently, and thus it was also proposed to build and use special sentiment polarity lexicons for words in negation contexts (Kiritchenko et al. 2014).

Features. Traditionally, systems for Sentiment Analysis on Twitter have relied on handcrafted features derived from word-level (e.g., *great*, *freshly roasted coffee*, *becoming president*) and character-level *n*-grams (e.g., *bec*, *beco*, *comin*, *oming*), stems (e.g., *becom*), lemmata (e.g., *become*, *roast*), punctuation (e.g., exclamation and question marks), part-of-speech tags (e.g., adjectives, adverbs, verbs, nouns), word clusters (e.g., *probably*, *probly*, and *maybe* could be collapsed to the same word cluster), and Twitter-specific encodings such as emoticons (e.g.,*;)*,*: D*), hashtags (*#Brexit*), user tags (e.g., *@allenai org*), abbreviations (e.g., *RT*, *BTW*, *F2F*, *OMG*), elongated words (e.g., *soooo*, *yaayyy*), use of capitalization (e.g., proportion of ALL CAPS words), URLs, etc. Finally, the most important features are those based on the presence of words and phrases in sentiment polarity lexicons with positive/negative scores; examples of such features include number of positive terms, number of negative terms, ratio of the number of positive terms to the number of positive + negative terms, ratio of the number of negative terms to the number of positive + negative terms, sum of all positive scores, sum of all negative scores, sum of all scores, etc.

Supervised learning. Traditionally, the above features were fed into classifiers such as Maximum Entropy (MaxEnt) and Support Vector Machines (SVM) with various kernels. However, observation over the SemEval Twitter sentiment task in recent years shows growing interest in and by now clear dominance of methods based on deep learning. In particular, the best-performing systems at SemEval-2015 and SemEval-2016 used deep convolutional networks (Deriu et al. 2016; Severyn and Moschitti 2015b). Conversely, kernel machines seem to be less frequently used than in the past, and the use of learning methods other than the ones mentioned above is at this point scarce. All these models are examples of supervised learning as they need labeled training data.

Semi-supervised learning. We should note two things about the use of deep neural networks: first they can often do quite well without the need for explicit feature modeling, as they can learn the relevant features in their hidden layers starting from the raw text. Second, they have too many parameters, and thus they require a lot of training data, orders of magnitude more than it is realistic to have manually annotated. A popular way to solve this latter problem is to use self-training, a form of semi-supervised learning, where first a system is trained on the available training data only, then this system is applied to make predictions on a large unannotated set of tweets, and finally it is trained for a few more iterations on its own predictions. This works because parts of the network, e.g., with convolution or with LSTMs (dos Santos and Gatti 2014; Severyn and Moschitti 2015b; Wang et al. 2015), need to learn something like a language model, i.e., which word is likely to follow which one. Training these parts needs no labels. While these parts can be also pretrained, it is easier, and often better, to use self-training.

Distantly supervised learning. Another way to make use of large unannotated datasets is to rely on *distant supervision* (*Marchetti-Bowick and Chambers* 2012). For example, one can annotate tweets for sentiment polarity based on whether they contain a positive or a negative emoticon. This results in noisy labels, which can be used to train a system (Severyn and Moschitti 2015b), to induce sentiment-specific word embeddings (Tang et al. 2014), sentiment polarity lexicons (Mohammad et al. 2013), etc.

Unsupervised learning. Fully unsupervised learning is not a popular method for addressing sentiment analysis tasks. Yet, some features used in sentiment analysis have been learned in an unsupervised way, e.g., Brown clusters to generalize over words (Owoputi et al. 2012). Similarly, word embeddings are typically trained from raw tweets that have no annotation for sentiment (even though there is also work on sentimentspecific word embeddings (Tang et al. 2014), which uses distant supervision).

Sentiment Polarity Lexicons

Despite the wide variety of knowledge sources explored so far in the literature, sentiment polarity lexicons remain the most commonly used resource for the task of sentiment analysis.

Until recently, such sentiment polarity lexicons were manually crafted and were thus of small to moderate size, e.g., LIWC (Pennebaker et al. 2001) has 2300 words, the General Inquirer (Stone et al. 1966) contains 4206 words, Bing Liu's lexicon (Hu and Liu 2004) includes 6786 words, and MPQA (Wilson et al. 2005) has about 8000 words.

Early efforts toward building sentiment polarity lexicons automatically yielded lexicons of moderate sizes such as the SentiWordNet (Baccianella et al. 2010; Esuli and Sebastiani 2006). However, recent results have shown that automatically extracted large-scale lexicons (e.g., up to a million words and phrases) offer important performance advantages, as confirmed at shared tasks on Sentiment Analysis on Twitter at SemEval 2013–2016 (Nakov et al. 2016a, 2013;

Rosenthal et al. 2015, 2014). Using such large-scale lexicons was crucial for the performance of the top-ranked systems. Similar observations were made in the related Aspect-Based Sentiment Analysis task at SemEval 2014 (Pontiki et al. 2014). In both tasks, the winning systems benefitted from building and using massive sentiment polarity lexicons (Mohammad et al. 2013; Zhu et al. 2014b).

The two most popular large-scale lexicons were the Hashtag Sentiment Lexicon and the Sentiment140 lexicon, which were developed by the team of NRC Canada for their participation in the SemEval-2013 shared task on sentiment analysis on Twitter. Similar automatically induced lexicons proved useful for other SemEval tasks, e.g., for SemEval-2016 Task 3 on Community Question Answering (Balchev et al. 2016; Nakov et al. 2016a).

The importance of building sentiment polarity lexicons has resulted in a special subtask (Rosenthal et al. 2015) at SemEval-2015 (part of Task 4) and an entire task (Kiritchenko et al. 2016) at SemEval-2016 (namely, Task 7), on predicting the out-of-context sentiment intensity of words and phrases. Yet, we should note though that the utility of using sentiment polarity lexicons for sentiment analysis probably needs to be revisited, as the best system at SemEval-2016 Task 4 could win without using any lexicons (Deriu et al. 2016); it relied on semi-supervised learning using a deep neural network.

Various approaches have been proposed in the literature for bootstrapping sentiment polarity lexicons starting from a small set of seeds: positive and negative terms (words and phrases). The dominant approach is that of Turney (Turney 2002), who uses pointwise mutual information and bootstrapping to build a large lexicon and to estimate the semantic orientation of each word in that lexicon. He starts with a small set of seed positive (e.g., *excellent*) and negative words (e.g., *bad*) and then uses these words to induce sentiment polarity orientation for new words in a large unannotated set of texts (in his case, product reviews). The idea is that words that co-occur in the same text with positive seed words are likely to be positive, while those that tend to co-occur with negative words are likely to be negative. To quantify this intuition, Turney defines the notion of sentiment orientation (SO) for a term w as follows:

$$SO(w) = pmi(w, pos) - pmi(w, neg)$$

where PMI is the pointwise mutual information; *pos* and *neg* are placeholders standing for any of the seed positive and negative terms, respectively; and w is a target word/phrase from the large unannotated set of texts (here tweets).

A positive/negative value for $SO(w)$ indicates positive/negative polarity for the word w, and its magnitude shows the corresponding sentiment strength. In turn, $pmi(w, pos) = \frac{P(w,pos)}{P(w)P(pos)}$, where $P(w, pos)$ is the probability to see w with any of the seed positive words in the same tweet, $P(w)$ is the probability to see w in any tweet, and $P(pos)$ is the probability to see any of the seed positive words in a tweet; $pmi(w, neg)$ is defined similarly.

The pointwise mutual information is a notion from information theory: given two random variables A and B, the mutual information of A and B is the “amount of information” (in units such as bits) obtained about the random variable A, through the random variable B (Church and Hanks 1990).

Let a and b be two values from the sample space of A and B, respectively. The *pointwise* mutual information between a and b is defined as follows:

$$\begin{aligned} pmi(a; b) &= \log \frac{P(A = a, B = b)}{P(A = a) \cdot P(B = b)} \\ &= \log \frac{P(A = a | B = b)}{P(A = a)} \qquad (1) \end{aligned}$$

$pmi(a;\ b)$ takes values between $-\infty$, which happens when $P(A = a,\ B = b) = 0$, and $\min\{-\log P(A = a), -\log P(B = b)\}$ if $P(A = a| B = b) = P(B = b|A = a) = 1$.

In his experiments, Turney (Turney 2002) used five positive and five negative words as seeds. His PMI-based approach further served as the basis for the creation of the two abovementioned large-scale automatic lexicons for sentiment analysis in

Twitter for English, initially developed by NRC for their participation in SemEval-2013 (Mohammad et al. 2013). The *Hashtag Sentiment Lexicon* uses as seeds hashtags containing 32 positive and 36 negative words, e.g., #happy and #sad. Similarly, the *Sentiment140* lexicon uses smileys as seed indicators for positive and negative sentiment, e.g., :), :-), and :)) as positive seeds and :(and :-(as negative ones.

An alternative approach to lexicon induction has been proposed (Severyn and Moschitti 2015a), which, instead of using PMI, assigns positive/negative labels to the unlabeled tweets (based on the seeds) and then trains an SVM classifier on them, using word *n*-grams as features. These *n*-grams are then used as lexicon entries (words and phrases) with the learned classifier weights as polarity scores. Finally, it has been shown that sizable further performance gains can be obtained by starting with mid-sized seeds, i.e., hundreds of words and phrases (Jovanoski et al. 2016).

Key Applications

Sentiment analysis on Twitter has applications in a number of areas, including political science (Borge-Holthoefer et al. 2015; Kaya et al. 2013; Marchetti-Bowick and Chambers 2012), economics (Bollen et al. 2011; O'Connor et al. 2010), social science (Dodds et al. 2011), and market research (Burton and Soboleva 2011; Qureshi et al. 2013). It is used to study company reputation online (Qureshi et al. 2013), to measure customer satisfaction, to identify detractors and promoters, to forecast market growth (Bollen et al. 2011), to predict the future income from newly released movies, to forecast the outcome of upcoming elections (Marchetti-Bowick and Chambers 2012; O'Connor et al. 2010), to study political polarization (Borge-Holthoefer et al. 2015; Tumasjan et al. 2010), etc.

Future Directions

We expect the quest for more interesting formulations of the general sentiment analysis task to continue. We see competitions such as those at SemEval as the engine of this innovation, as they not only perform head-to-head comparisons but also create databases and tools that enable follow-up research for many years afterward.

In terms of methods, we believe that deep learning (dos Santos and Gatti 2014; Severyn and Moschitti 2015b; Wang et al. 2015), together with semi-supervised and distantly supervised methods (Davidov et al. 2010a; Tang et al. 2014), will be the main focus of future research. We also expect more attention to be paid to linguistic structure and sentiment compositionality (Socher et al. 2012, 2013). Moreover, we forecast more interest for languages other than English and for cross-lingual methods (Kaya et al. 2013; Mihalcea et al. 2007; Wan 2009), which will allow leveraging on the rich resources that are already available for English. Last, but not least, the increase in opinion spam on Twitter will make it important to study astroturfing (Ratkiewicz et al. 2011) and troll detection (Mihaylov et al. 2015a, b; Mihaylov and Nakov 2016).

Cross-References

- ▶ Microblog Sentiment Analysis
- ▶ Multi-Classifier System for Sentiment Analysis and Opinion Mining
- ▶ Sentiment Analysis in Social Media
- ▶ Sentiment Analysis of Microblogging Data
- ▶ Sentiment Analysis of Reviews
- ▶ Sentiment Analysis, Basic Tasks of
- ▶ Sentiment Quantification of User-Generated Content
- ▶ Social Media Analysis for Monitoring Political Sentiment
- ▶ Twitter Microblog Sentiment Analysis
- ▶ User Sentiment and Opinion Analysis

Recommended Reading

For general research on sentiment analysis, we recommend the following surveys: (Liu and Zhang 2012) and (Pang and Lee 2008). For sentiment analysis on Twitter, we recommend the overview article on *Sentiment Analysis on Twitter* about the SemEval task (Nakov et al. 2016b) as well as the task description papers for different editions of the task (Nakov et al. 2016a, 2013; Rosenthal et al. 2014, 2015).

References

Baccianella S, Esuli A, Sebastiani F (2010) SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the seventh international conference on language resources and evaluation, LREC '10 Valletta, Malta

Balchev D, Kiprov Y, Koychev I, Nakov P (2016) PMI-cool at SemEval-2016 task 3: experiments with PMI and goodness polarity lexicons for community question answering. In: Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval '16, San Diego, California, USA, pp 844–850

Barbosa L, Feng J (2010) Robust sentiment detection on Twitter from biased and noisy data. In: Proceedings of the 23rd international conference on computational linguistics, COLING '10, Beijing, China, pp 36–44

Bifet A, Holmes G, Pfahringer B, Gavalda R (2011) Detecting sentiment change in Twitter streaming data. J Mach Learn Res, Proceedings Track vol 17, pp 5–11

Bollen J, Mao H, Zeng XJ (2011) Twitter mood predicts the stock market. J Comput Sci 2(1):1–8

Borge-Holthoefer J, Magdy W, Darwish K, Weber I (2015) Content and network dynamics behind Egyptian political polarization on Twitter. In: Proceedings of the 18th ACM conference on computer supported cooperative work and social computing, CSCW '15, Vancouver, Canada, pp 700–711

Burton S, Soboleva A (2011) Interactive or reactive? Marketing with twitter. J Consum Mark 28(7):491–499

Church KW, Hanks P (1990) Word association norms, mutual information, and lexicography. Comput Linguist 16(1):22–29

Das SR, Chen MY (2007) Yahoo! For Amazon: sentiment extraction from small talk on the web. Manag Sci 53(9):1375–1388

Davidov D, Tsur O, Rappoport A (2010a) Enhanced sentiment learning using Twitter hashtags and smileys. In: Proceedings of the 23rd international conference on computational linguistics: posters, COLING '10, Beijing, China, pp 241–249

Davidov D, Tsur O, Rappoport A (2010b) Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In: Proceedings of the fourteenth conference on computational natural language learning, CoNLL '10, Uppsala, Sweden, pp 107–116

Deriu J, Gonzenbach M, Uzdilli F, Lucchi A, De Luca V, Jaggi M (2016) SwissCheese at SemEval-2016 task 4: sentiment classification using an ensemble of convolutional neural networks with distant supervision. In: Proceedings of the 10th international workshop on semantic evaluation, SemEval '16, San Diego, California, USA, pp 1124–1128

Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM (2011) Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. PLoS One 6(12)

Esuli A, Sebastiani F (2006) SENTIWORDNET: a publicly available lexical resource for opinion mining. In: Proceedings of the international conference on language resources and evaluation, LREC '06, Genoa, Italy, pp 417–422

Forman G (2008) Quantifying counts and costs via classification. Data Min Knowl Disc 17(2):164–206

Gao W, Sebastiani F (2015) Tweet sentiment: from classification to quantification. In: Proceedings of the 7th international conference on advances in social network analysis and mining, ASONAM '15, Paris, France, pp 97–104

Ghosh A, Li G, Veale T, Rosso P, Shutova E, Barnden J, Reyes A (2015) SemEval-2015 task 11: sentiment analysis of figurative language in Twitter. In: Proceedings of the 9th international workshop on semantic evaluation, SemEval '15, Denver, Colorado, USA, pp 470–478

Gimpel K, Schneider N, O'Connor B, Das D, Mills D, Eisenstein J, Heilman M, Yogatama D, Flanigan J, Smith NA (2011) Part-of-speech tagging for Twitter: annotation, features, and experiments. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, ACL-HLT '11, Portland, Oregon, USA, pp 42–47

Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '04, Seattle, Washington, USA, pp 168–177

Jansen B, Zhang M, Sobel K, Chowdury A (2009) Twitter power: tweets as electronic word of mouth. J Am Soc Inf Sci Technol 60(11):2169–2188

Java A, Song X, Finin T, Tseng B (2007) Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on web mining and social network analysis, pp 56–65

Jovanoski D, Pachovski V, Nakov P (2016) On the impact of seed words on sentiment polarity lexicon induction. In: Proceedings of the 26th international conference on computational linguistics, COLING '16 Osaka, Japan

Kaya M, Fidan G, Toroslu IH (2013) Transfer learning using Twitter data for improving sentiment classification of Turkish political news. In: Proceedings of the 28th international symposium on computer and information sciences, ISCIS '13, Paris, France, pp 139–148

Kiritchenko S, Zhu X, Mohammad SM (2014) Sentiment analysis of short informal texts. J Artif Intell Res 50:723–762

Kiritchenko S, Mohammad SM, Salameh M (2016) SemEval-2016 task 7: determining sentiment intensity of English and Arabic phrases. In: Proceedings of the 10th international workshop on semantic evaluation, SemEval ‘16 San Diego, California, USA

Kong L, Schneider N, Swayamdipta S, Bhatia A, Dyer C, Smith NA (2014) A dependency parser for tweets. In: Proceedings of the 2014 conference on empirical methods in natural language processing, EMNLP ‘14, Doha, Qatar, pp 1001–1012

Kouloumpis E, Wilson T, Moore J (2011) Twitter sentiment analysis: the good the bad and the OMG! In: Proceedings of the fifth international conference on weblogs and social media, ICWSM ‘11, Barcelona, Catalonia, Spain, pp 538–541

Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web, WWW ‘10, Raleigh, North Carolina, USA, pp 591–600

Liu B, Zhang L (2012) A survey of opinion mining and sentiment analysis. In: Aggarwal CC, Zhai C (eds) Mining text data. Springer, New York, pp 415–463

Marchetti-Bowick M, Chambers N (2012) Learning for microblogs with distant supervision: political forecasting with Twitter. In: Proceedings of the 13th conference of the European chapter of the association for computational linguistics, EACL ‘12, Avignon, France, pp 603–612

Mihalcea R, Banea C, Wiebe J (2007) Learning multilingual subjective language via crosslingual projections. In: Proceedings of the 45th annual meeting of the association of computational linguistics, ACL ‘07, Prague, Czech Republic, pp 976–983

Mihaylov T, Nakov P (2016) Hunting for troll comments in news community forums. In: Proceedings of the 54th annual meeting of the association for computational linguistics, ACL ‘16, Berlin, Germany, pp 399–405

Mihaylov T, Georgiev G, Nakov P (2015a) Finding opinion manipulation trolls in news community forums. In: Proceedings of the nineteenth conference on computational natural language learning, CoNLL ‘15, Beijing, China, pp 310–314

Mihaylov T, Koychev I, Georgiev G, Nakov P (2015b) Exposing paid opinion manipulation trolls. In: Proceedings of the international conference recent advances in natural language processing, RANLP ‘15, Hissar, Bulgaria, pp 443–450

Mohammad, S (2012) #Emotional tweets. In: Proceedings of *SEM 2012: the first joint conference on lexical and computational semantics – volume 1: proceedings of the main conference and the shared task, *SEM ‘12, Montreal, Canada, pp 246–255

Mohammad S, Kiritchenko S, Zhu X (2013) NRC-Canada: building the state-of-the-art in sentiment analysis of tweets. In: Proceedings of the second Joint conference on lexical and computational semantics (*SEM), volume 2: proceedings of the seventh international workshop on semantic evaluation, SemEval ‘13, Atlanta, Georgia, USA, pp 321–327

Mohammad SM, Kiritchenko S, Sobhani P, Zhu X, Cherry C (2016) SemEval-2016 task 6: detecting stance in tweets. In: Proceedings of the 10th international workshop on semantic evaluation, SemEval ‘16, San Diego, California, USA, pp 31–41

Nakov P, Rosenthal S., Kozareva Z, Stoyanov V, Ritter A, Wilson T (2013) SemEval-2013 task 2: sentiment analysis in Twitter. In: Proceedings of the second joint conference on lexical and computational semantics (*SEM), volume 2: proceedings of the seventh international workshop on semantic evaluation, SemEval ’13, Atlanta, Georgia, USA, pp 312–320

Nakov, P., Marquez L, Moschitti A, Magdy W, Mubarak H, Freihat AA, Glass J, Randeree B (2016a) SemEval-2016 task 3: community question answering. In: Proceedings of the 10th international workshop on semantic evaluation, SemEval ’16, San Diego, California, USA, pp 525–545

Nakov P, Rosenthal S, Kiritchenko S, Mohammad SM, Kozareva Z, Ritter A, Stoyanov V, Zhu X (2016b) Developing a successful SemEval task in sentiment analysis of twitter and other social media texts. Lang Resour Eval 50(1):35–65

O’Connor B, Balasubramanyan R, Routledge B, Smith N (2010) From tweets to polls: linking text sentiment to public opinion time series. In: Proceedings of the fourth international conference on weblogs and social media, ICWSM ‘10, Washington, DC, USA, pp 122–129

Owoputi O, Dyer C, Gimpel K, Schneider N (2012) Part-of-speech tagging for Twitter: word clusters and other advances. Tech. Rep. CMU-ML-12-107, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Pak A, Paroubek P (2010) Twitter based system: using Twitter for disambiguating sentiment ambiguous adjectives. In: Proceedings of the 5th international workshop on semantic evaluation, SemEval ’10, Uppsala, Sweden, pp 436–439

Pang B, Lee L (2005) Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the annual meeting of the association for computational linguistics, ACL ‘05, Ann Arbor, Michigan, USA, pp 115–124

Pang B, Lee L (2008) Opinion mining and sentiment analysis. Foundation Trend Inform Retriev 2(1–2):1–135

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the conference on empirical methods in natural language processing, EMNLP ‘02, Philadelphia, Pennsylvania, USA, pp 79–86

Pennebaker JW, Francis ME, Booth RJ (2001) Linguistic inquiry and word count. Lawerence Erlbaum Associates, Mahwah

Pontiki M, Papageorgiou H, Galanis D, Androutsopoulos I, Pavlopoulos J, Manandhar S (2014) SemEval-2014 task 4: aspect based sentiment analysis. In: Proceedings of the 8th international workshop on semantic evaluation, SemEval ’14, Dublin, Ireland, pp 27–35

Pontiki M, Galanis D, Papageorgiou H, Manandhar S, Androutsopoulos I (2015) SemEval2015 task 12: aspect based sentiment analysis. In: Proceedings of the 9th international workshop on semantic evaluation, SemEval ‘15, Denver, Colorado, USA, pp 486–495

Pontiki M, Galanis D, Papageorgiou H, Androutsopoulos I, Manandhar S, AL-Smadi M, Al-Ayyoub M, Zhao Y, Qin B, De Clercq O, Hoste V, Apidianaki M, Tannier X, Loukachevitch N, Kotelnikov E, Bel N, Jimenez-Zafra SM, Eryigit G (2016) SemEval-2016 task 5: aspect based sentiment analysis. In: Proceedings of the 10th international workshop on semantic evaluation, SemEval ’16, San Diego, California, USA, pp 19–30

Qureshi MA, O’Riordan C, Pasi G (2013) Clustering with error estimation for monitoring reputation of companies on Twitter. In: Proceedings of the 9th Asia information retrieval societies conference, AIRS ‘13, Singapore, pp 170–180

Ratkiewicz J, Conover M, Meiss M, Goncalves B, Patil S, Flammini A, Menczer F (2011) Truthy: mapping the spread of astroturf in microblog streams. In: Proceedings of the 20th international conference companion on World Wide Web, WWW ’11, Hyderabad, India, pp 249–252

Raychev V, Nakov P (2009) Language-independent sentiment analysis using subjectivity and positional information. In: Proceedings of the international conference on recent advances in natural language processing, RANLP ’09, Borovets, Bulgaria, pp 360–364

Ritter A, Clark S, Mausam EO (2011) Named entity recognition in tweets: an experimental study. In: Proceedings of the conference on empirical methods in natural language processing, EMNLP ’11, Edinburgh, Scotland, UK, pp 1524–1534

Rosenthal S, Ritter A, Nakov P, Stoyanov V (2014) SemEval-2014 task 9: sentiment analysis in Twitter. In: Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval ‘14, Dublin, Ireland, pp 73–80

Rosenthal S, Nakov P, Kiritchenko S, Mohammad S, Ritter A, Stoyanov V (2015) SemEval2015 task 10: sentiment analysis in Twitter. In: Proceedings of the 9th international workshop on semantic evaluation, SemEval ‘15, Denver, Colorado, USA, pp 450–462

Russo I, Caselli T, Strapparava C (2015) SemEval-2015 task 9: CLIPEval implicit polarity of events. In: Proceedings of the 9th international workshop on semantic evaluation, SemEval ’15, Denver, Colorado, USA, pp 442–449

dos Santos C, Gatti M (2014) Deep convolutional neural networks for sentiment analysis of short texts. In: Proceedings of the 25th international conference on computational linguistics, COLING ’14, Dublin, Ireland, pp 69–78

Sebastiani F (2002) Machine learning in automated text categorization. ACM Comput Surv 34(1):1–47

Severyn A, Moschitti A (2015a) On the automatic learning of sentiment lexicons. In: Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: human language technologies, NAACL-HLT ‘15, Denver, Colorado, USA, pp 1397–1402

Severyn A, Moschitti A (2015b) Twitter sentiment analysis with deep convolutional neural networks. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, SIGIR ’15, Santiago, Chile, pp 959–962

Socher R, Huval B, Manning CD, Ng AY (2012) Semantic compositionality through recursive matrix-vector spaces. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, EMNLPCoNLL ’12, Jeju Island, Korea, pp 1201–1211

Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng A, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 conference on empirical methods in natural language processing, EMNLP ’13. Seattle, Washington, USA, pp 1631–1642

Stone PJ, Dunphy DC, Smith MS, Ogilvie DM (1966) The general inquirer: a computer approach to content analysis. MIT Press, Cambridge, MA

Stoyanov V, Cardie C (2008) Topic identification for fine-grained opinion analysis. In: Proceedings of the 22nd international conference on computational linguistics, COLING ’08, Manchester, United Kingdom, pp. 817–824

Strapparava C, Mihalcea R (2007) SemEval-2007 task 14: affective text. In: Proceedings of the international workshop on semantic evaluation, SemEval ’07, Prague, Czech Republic, pp 70–74

Tang D, Wei F, Yang N, Zhou M, Liu T, Qin B (2014) Learning sentiment-specific word embedding for Twitter sentiment classification. In: Proceedings of the 52nd annual meeting of the association for computational linguistics, ACL ’14, Baltimore, Maryland, USA, pp 1555–1565

Tumasjan A, Sprenger T, Sandner P, Welpe I (2010) Predicting elections with Twitter: what 140 characters reveal about political sentiment. In: Proceedings of the fourth international conference on weblogs and social media, ICWSM ’10, Washington, DC, USA, pp 178–185

Turney PD (2002) Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the annual meeting of the association for computational linguistics, ACL ’02, Philadelphia, Pennsylvania, USA, pp 417–424

Villena-Roman J, Lana-Serrano S, Martınez-Camara E, Cristobal JCG (2013) TASS workshop on sentiment analysis at SEPLN. Procesamiento del Lenguaje Natural 50:37–44

Wan X (2009) Co-training for cross-lingual sentiment classification. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP, ACL-AFNLP ’09, Singapore, pp 235–243

Wang X, Liu Y, SUN C, Wang B, Wang X (2015) Predicting polarities of tweets by composing word embeddings with long short-term memory. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing, ACL-IJCNLP '15, Beijing, China, pp 1343–1353

Wiebe J, Wilson T, Bruce R, Bell M, Martin M (2004) Learning subjective language. Comput Linguist 30(3):277–308

Wiebe J, Wilson T, Cardie C (2005) Annotating expressions of opinions and emotions in language. Lang Resour Eval 39(2–3):165–210

Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on human language technology and empirical methods in natural language processing, HLT-EMNLP '05, Vancouver, Canada, pp 347–354

Zhu X, Guo H, Mohammad SM, Kiritchenko S (2014a) An empirical study on the effect of negation words on sentiment. In: Proceedings of the annual meeting of the association for computational linguistics, ACL '14, Baltimore, MD, USA, pp 304–313

Zhu X, Kiritchenko S, Mohammad SM (2014b) NRC-Canada-2014: detecting aspects and sentiment in customer reviews. In: Proceedings of the international workshop on semantic evaluation, SemEval '14, Dublin, Ireland, pp 437–442

Semantic Social Networks

Peter Gloor[1] and Jana Diesner[2]
[1]Center for Collective Intelligence, MIT, Cambridge, MA, USA
[2]School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL, USA

Synonyms

Context networks; Ontologies; Taxonomies; Text networks; Topic networks

Glossary

Semantic network	Structured representations of knowledge that are used for reasoning and inference
RDF	Resource Description Framework
FOAF	Friend of a friend
NLP	Natural language processing

Definition

Semantic social network analysis combines the collection and examination of connections between social agents and the content produced or shared in the network. Representations of the content are either attached to the nodes (e.g., if the nodes in the network are Web sites) or to the edges (e.g., if the edges in the network are e-mails). Network structure is then derived from similarity in the content exchanged or from similarity in content between the nodes. Analysis happens mostly on the word level through natural language processing techniques, and statistical approaches to comparing combinations of characters and words are also popular.

Introduction

Semantic networks represent the relationships between pieces of knowledge or information. They were originally designed to be used for performing inference and reasoning on the meaning of these data (Sowa 1992; Woods 1975). Social networks represent the interactions between social agents, typically people and/or organizations (Freeman 2004). The analyses of semantic networks and social networks are both active areas of research and innovation, while work at their intersection is less prevalent. However, it has been long recognized that combining both types of network data enables researchers and practitioners alike to ask more advanced yet highly practically relevant questions such as:

Who is talking to whom (social network level) about what (semantic network level) (Danowski 1993; McCallum et al. 2007)?

Do shared knowledge or interests in similar topics (semantic network) increase the likelihood of becoming acquaintances (social networks) and vice versa (Crandall et al. 2008; Wenger 1999)?

How do information, opinions, and rumors (semantic network) emerge, spread, and vanish in society and on social networking sites (social network) (Adar and Adamic 2005; Leskovec et al. 2009)?

Can social network data be made more accurate, complete, and useful by exploiting background information on social agents (Berners-Lee et al. 2001; Van Atteveldt 2008)?

The last one of these questions originates from a more specific use of the concept of "semantic social networks": in addition to referring to the combination of semantic and social networks, this term also describes the enhancement of relational data with background information on any type of node or entity. Again, these enhanced graphs can then be used for conducting inference and reasoning. One prominent example for this approach is the semantic Web (Berners-Lee et al. 2001). The key idea with the semantic Web is to mark up data objects on the Web, such as words and relations, as they occur on Web pages, by using a standardized annotation language, also called Resource Description Framework or RDF. These enhanced data structures are then used to generate machine-readable definitions of data that can be interpreted by computers. An example for a stream of work that originates from the semantic Web philosophy is the "friend-of-a-friend" (FOAF) framework. FOAF combines social network data with semantic network data, where both types of data structures are denoted according to a predefined, machine-readable description language, allowing for automated inference on the data. On a general level, such data can be used to recommend social ties between people who share some interests or are involved in the same events or to recommend activities and pieces of information that are new to a person whose friends have endorsed these items.

Key Points

With the rise of the social Web and social media, combining the analysis of network structure with the analysis of network content has become increasingly important. Relationships between nodes are influenced through similarities and differences in node and edge content. In combination with sentiment analysis, semantic social network analysis greatly extends the predictive capabilities of traditional social network analysis.

Historical Background

Efforts to combine social and semantic networks trace back their roots to well before the advent of computers and the Internet. Vannevar Bush (1945), advisor to US president Roosevelt during the Second World War, envisioned Memex, a device for organizing all knowledge of mankind in a structured way. In the 1960s, Ted Nelson coined the term "hypertext." In the 1970s and 1980s, a vibrant field around the concept of hypertext emerged in disciplines such as artificial intelligence and computer science. In the 1980s and 1990s, the ACM annual hypertext conference regularly drew between 500 and 1000 researchers. Early on these researchers started combining hypertext with semantic networks (Brachman 1979). In 1991 Tim Berners-Lee, together with Robert Cailliau, presented the Web at the ACM hypertext conference in San Antonio, and just a few years later, he broadened his concept to the semantic Web (Berners-Lee et al. 2001). Shortly after, Hermann Maurer and his colleagues envisioned a hypermedia system called Hyper-G, which combines semantics with social networking (Andrews et al. 1995).

Combining Social and Semantic Networks

Combining social networks and semantic networks opens up novel ways for extracting meaning from social interactions. One strategy for bringing together social networks with semantics about the network data is to enhance a given social network with additional information about agents and their connections. This can be done by exploiting external data sources such as the Web, news and information archives, and domain-specific databases (Van Atteveldt 2008). For example, for the individuals being co-mentioned in a news article, one could search knowledge bases such as Wikipedia or the Web for further information on those people's roles and locations. Adding this information to the social network contextualizes the data, which allows for richer and more fine-grained indexing and retrieval. This approach basically adds semantics to a social network. It can also help to disambiguate social agents, e.g., people who have the same name, but differ in their job, location, or date of birth as indicated on knowledge bases such as Wikipedia, and to specify the types of relationships between agents. The inverse of this principle, i.e., utilizing, searching, or suggesting connections between people who share some knowledge, interests, or likings, is also an active area of research and development. Examples for this approach include recommender systems, dating services, and social networking platforms.

Another common strategy for bringing together social networks with semantic information is to enhance a social network with the information produced, processed, or shared by members from within or outside the network (Diesner and Carley 2011). This information typically represents salient information from natural language text data, such as people's interests as indicated on their social networking profile or key terms and themes that are explicitly or implicitly contained in documents that people authored. For communication data, for instance, a social network can be built from the explicit information about communication partners (who talks to whom). Then, agents can be linked to nodes representing words and short phrases that occur with a high (weighted) frequency in the underlying text data. Suitable data sources used for this procedure include transcripts of conversations and meetings and online discussion forums. An early example for a tool that jointly analyzed the social network of e-mail senders and receivers *and* the content of e-mail bodies is TeCFlow, now called Condor (Gloor and Zhao 2004). Today, a variety of methods and tools has been developed that support the joint collection, visualization, and analysis of relational social and semantic information from communication archives. Remaining challenges in this domain include the partitioning of people into meaningful groups prior to associating social clusters with content and selecting pieces of information to link to agents in a scalable yet nonarbitrary fashion (Diesner 2013).

Sometimes, the information about social networks is encoded in sources that are typically used for conducting semantic analysis, namely, unstructured, natural language text data. In these cases, social network data can be extracted from the text data (Diesner and Carley 2011; Roth and Yih 2002). Typical data sources include newswire data, interviews, communication data, and social media data, such as microblogging services and social networking sites. The main steps involved in this task are the identification of entities, i.e., nodes, and the relations between the nodes. These entities are sometimes further categorized into different classes, such as people, organizations, and locations, and can entail one-word units as well as multi-word units (Diesner and Carley 2008). The types of relations can be defined over entity types, such as social networks between people or a membership network between people and organizations. Alternatively, applicable types of relations can be specified in an ontology or a taxonomy, which can be predefined or extracted from the data (Brin 1999; Roth and Yih 2002). An example would be to classify social network ties as representing friendship or antagonism. Highly accurate, automated, and scalable methods for relation extraction typically exploit a combination of lexical (words and their structure), semantic (meaning of words), syntactic (relationships between words and grammar), and statistical information from text data (Diesner and Carley 2008; Mihalcea and Radev 2011). These methods, which have been developed in the fields of natural language processing and computational linguistics, typically combine routines from statistics and machine learning and sometimes also consider models and methods from sociolinguistics and sociology (Corman et al. 2002; Diesner and Carley 2008). Once such network data have been extracted from a text corpus, they can serve as input to regular network analysis (Carley et al. 2007; Corman et al. 2002). Used this way, relation extraction can serve as a complementary or alternative method for collecting data about social networks. These social networks can be combined with semantic networks that are also extracted from text data. In fact, some models and methods consider “knowledge” or “information” as a node classes for relation extraction (Diesner and Carley 2008). Combining social networks and semantic networks extracted from text data or built from other sources can be useful for addressing the following types of questions (Barthelemy et al. 2005; Carley et al. 2007; Gloor et al. 2009):

Which social agents are associated with what ideas, beliefs, or pieces of knowledge?

Which agents are prominent with respect to their association with information? These people might function as information brokers or gatekeepers if they have a high betweenness centrality or are somewhere between well informed and overloaded with information if they feature a high degree centrality (for details on these metrics, see the section on centrality measures).

Which agents are linked to too many knowledge items and thus might suffer from task overload?

Which agents have exclusive access to some information? In an organizational context, such people might represent a vulnerability, which can be mitigated by converting tacit knowledge residing in people to information being documented in written form.

One caveat with distilling network data from text data is that research on resembling ground truth data for social networks by exploiting the substance of text data has shown that the overlap between text-based social networks, e.g., those extracted from e-mail bodies, and social networks constructed from associated metadata, e.g., e-mail headers, shows only minimal agreement (Diesner 2013).

Key Applications

One prominent example for employing semantic social networks in practice is expert finder systems (Dooley et al. 2002). A specific instance is SmallBlue, a tool developed and deployed at IBM (Ehrlich et al. 2007). This system augments social network data with information on who knows what, allowing people to search for the shortest social path to knowledge through their wider and potentially remote network of coworkers. In such systems, the information on people’s expertise can be pulled from internal data sources, such as organizational databases, as well as from public sources, such as blog posts and tags. Ehrlich and colleagues evaluated the SmallBlue system to be particularly useful for locating experts on very particular pieces of knowledge, which complements the general understanding about broadly regarded experts on certain topics. Generalizing from this idea, professional social network sites such as LinkedIn are based on the same premise: they allow people to search for experts on certain topics within the professional network of their immediate acquaintances. If a match has been found but the identified individual is not a contact of the person executing the search, one could mobilize their social capital to be introduced to a person via the shortest social path of mutual acquaintances.

Another real-world example for semantic social networks is Wikipedia (Brandes et al. 2009; Crandall et al. 2008). This knowledge base not only provides a vast amount of socially vetted information but also entails metadata about the authors and detailed information about every single contribution. The metadata surrounding the content pages contains information about the *what* (the edits of pages), *when* (edits over time), *who* (which authors edited the pages), and *how* (which links to other pages inside and outside of Wikipedia of information). These types of data can be fused into dynamic, multimodal network data. Moreover, these data open new opportunities for investigating the processes that lie behind the life cycle of the creation of content and contributing knowledge to the public domain. Visualizations built on this information can provide maps of concepts, knowledge, and trends, which can be displayed by content domain, geophysical region, cultural background, etc. Analyzing and comparing these maps and semantic networks across time, space, and languages can contribute toward a better understanding of societies and cultures. In addition, constructing coauthorship networks, where links between Wikipedia articles are drawn based on the same person editing different articles, enables the identification of domain experts as well as trusted arbitrators. Furthermore, Wikipedia has an implicit social organization of its own, composed of networks of contributors. Analyzing this social network can help to understand if active Wikipedians operate under an implicit set of rules that has evolved within the Wikipedia community and might generalize to other open-source production systems or to traditional organizations. Studying collaboration among Wikipedians also gives indications of the role of social capital for teams in organizations where members are collaborating virtually without much face-to-face contact. In the same way that social network surveys made visible the importance of the informal organization within large corporations, the analysis of Wikipedia editor networks enables the measuring of the role of social capital in voluntary online collaboration: social capital indeed seems to increase efficiency in this emerging organizational setting as well (Nemoto et al. 2011).

Future Directions

An example for ongoing research on semantic social networks is the measurement of team performance over time. In this work the performance and creativity of organizations are analyzed by correlating social network data, content-based semantic analyses, and creative performance of work teams (Grippa et al. 2012). This approach is a first step toward articulating a systematic theory of social networks coming from the dynamic and causal dimensions of relationships. This work represents the general idea behind semantic social networks, namely, enabling the investigation of ties among community members not only under the quantitative aspect related to SNA metrics but also under the qualitative aspect related to the content of the ties. Such an emergent theory will give new meaning to the “relational” and “cognitive” dimensions of social capital (Stinchcombe 1990).

Cross-References

- ▶ Analysis and Mining of Tags, (Micro)Blogs, and Virtual Communities
- ▶ Automatic Document Topic Identification Using Social Knowledge Network
- ▶ Clustering Algorithms
- ▶ Collecting Qualitative Data to Enhance Social Network Analysis and Data Mining
- ▶ Combining Link and Content for Community Detection
- ▶ Combining Online Social Networks with Text Analysis
- ▶ Data Mining
- ▶ Flickr and Twitter Data Analysis
- ▶ Inferring Social Ties
- ▶ Mapping Online Social Media Networks
- ▶ Multi-classifier System for Sentiment Analysis and Opinion Mining
- ▶ Network Data Collected via Web
- ▶ Network Text Analysis
- ▶ Ontology Matching
- ▶ Recommender Systems, Semantic-Based
- ▶ Semantic Sentiment Analysis of Twitter Data
- ▶ Semantic Social Networks Analysis
- ▶ Sentiment Analysis
- ▶ Sentiment Analysis in Social Media
- ▶ Sentiment Analysis of Microblogging Data
- ▶ Sentiment Analysis of Reviews
- ▶ Sentiment Analysis, Basic Tasks of
- ▶ Social Web Search
- ▶ Sources of Network Data

References

Adar E, Adamic L (2005) Tracking information epidemics in blogspace. Paper presented at the 2005 IEEE/WIC/ACM international conference on web intelligence, Compiegne, Sept 2005

Andrews K, Kappe F, Maurer H (1995) Serving information to the web with Hyper-G. Paper presented at the third international world-wide web conference, computer networks and ISDN systems

Barthelemy M, Chow E, Eliassi-Rad T (2005) Knowledge representation. Issues in semantic graphs for relationship detection. Paper presented at the AAAI spring symposium on AI Technologies for Homeland Security, Stanford, http://pages.cs.wisc.edu/~eliassi/chow-aaai-ss2005.pdf

Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Sci Am 284(5):34–43

Brachman RJ (1979) On the epistemological status of semantic networks. In: Findler NV (ed) Associative networks: representation and use of knowledge by computers. Academic, New York, pp 3–50

Brandes U, Kenis P, Lerner J, Van Raaij D (2009) Network analysis of collaboration structure in Wikipedia. Paper presented at the 18th international conference on world wide web, Madrid

Brin S (1999) Extracting patterns and relations from the world wide web. Paper presented at the the world wide web and databases, Valencia, 27–28 Mar 1998

Bush V (1945) As we may think. Atl Mon 176(1):101–108

Carley KM, Diesner J, Reminga J, Tsvetovat M (2007) Toward an interoperable dynamic network analysis toolkit. Decis Support Syst 43(4):1324–1347. Special Issue Cyberinfrastructure for Homeland Security

Corman SR, Kuhn T, Mchee RD, Dooley KJ (2002) Studying complex discursive systems: centering resonance analysis of communication. Hum Commun Res 28(2):157–206

Crandall D, Cosley D, Huttenlocher D, Kleinberg J, Suri S (2008) Feedback effects between similarity and social influence in online communities. Paper presented at the proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, Las Vegas

Danowski JA (1993) Network analysis of message content. Prog Commun Sci 12:198–221

Diesner J (2013) From texts to networks: detecting and managing the impact of methodological choices for extracting network data from text data. Künstliche Intelligenz/Artificial Intelligence 27(1):75–78. https://doi.org/10.1007/s13218-012-0225-0

Diesner J, Carley KM (2008) Conditional random fields for entity extraction and ontological text coding. J Comput Math Org Theory 14:248–262. https://doi.org/10.1007/s10588-008-9029-z

Diesner J, Carley KM (2011) Words and networks. In: Barnett G, Golson JG (eds) Encyclopedia of social networking. Sage, Thousand Oaks, pp 958–961

Dooley KJ, Corman SR, Mchee RD (2002) A knowledge directory for identifying experts and areas of expertise. Hum Syst Manag 21(4):217–228

Ehrlich K, Lin C, Griffiths-Fisher V (2007) Searching for experts in the enterprise: combining text and social network analysis. Paper presented at the 2007 international ACM conference on supporting group work, Sanibel Island

Freeman LC (2004) The development of social network analysis. Empirical Press, Vancouver

Gloor P, Zhao Y (2004) TeCFlow – a temporal communication flow visualizer for social networks analysis. Paper presented at the ACM CSCW conference, workshop on social networks, Chicago

Gloor P, Krauss J, Nann S, Fischbach K, Schoder D, Switzerland B (2009) Web science 2.0: identifying trends through semantic social network analysis. Sani-bel Island Paper presented at the IEEE conference on social computing (SocialCom-09), Vancouver

Grippa F, Palazzolo M, Bucuvalas J, Gloor P (2012) Monitoring changes in the social network structure of clinical care teams resulting from team development efforts. Int J Org Des Eng 2(2):149–166

Leskovec J, Backstrom L, Kleinberg J (2009) Meme-tracking and the dynamics of the news cycle. Paper presented at the 15th ACM SIGKDD international conference on knowledge discovery and data mining

McCallum A, Wang X, Corrada-Emmanuel A (2007) Topic and role discovery in social networks with experiments on Enron and academic email. J Artif Intell Res 30:249–272

Mihalcea RF, Radev DR (2011) Graph-based natural language processing and information retrieval. Cambridge University Press, Cambridge

Nemoto K, Gloor P, Laubacher R (2011) Social capital increases efficiency of collaboration among Wikipedia editors. Paper presented at the HT'11 22nd ACM conference on hypertext and hypermedia

Roth D, Yih W (2002) Probabilistic reasoning for entity and relation recognition. Paper presented at the international conference on computational linguistics (COLING), Taipei

Sowa J (1992) Semantic networks. In: Shapiro SC (ed) Encyclopedia of artificial intelligence, 2nd edn. Wiley, New York, pp 1493–1511

Stinchcombe AL (1990) Information and organizations. University of California Press, Berkley

Van Atteveldt W (2008) Semantic network analysis: techniques for extracting, representing, and querying media content. BookSurge Publishers, Charleston

Wenger E (1999) Communities of practice: learning, meaning, and identity. Cambridge University Press, New York

Woods W (1975) What's in a link: foundations for semantic networks. In: Bobrow D, Collins A (eds) Representation and understanding: studies in cognitive science. Academic, New York, pp 35–82

Semantic Social Networks Analysis

Christophe Thovex[1], Bénédicte LeGrand[2], Ofelia Cervantes[1], J. Alfredo Sánchez[1] and Francky Trichet[3]
[1]French-Mexican Laboratory of Informatics and Automatic Control (LAFMIA, UMI CNRS 3175), Lyon, France
[2]Centre de Recherche en Informatique, Paris, France
[3]Laboratory of Computer Sciences (LINA, UMR CNRS 6241), University of Nantes, Nantes, France

Synonyms

Detection of communities; Graph mining; Knowledge engineering; Networks dynamics analysis; Semantic networks; Social trends discovery; Text mining

Glossary

Data Mining	Extracting implicit information and knowledge from numeric data
Graph Mining	Extracting implicit information and knowledge from graphs
Knowledge Engineering	Discipline studying, extracting, and managing knowledge implicitly defined within digital data structures
SNA	Social network analysis (see Definition section)
Social Capital	Knowledge and skills owned by employees (human capital) when shared in a collaborative context and defining a network of professional interactions
SW	Semantic Web (see Historical background section)
Text Mining	Extracting implicit information and knowledge from text corpora

Definition

Social networks analysis (SNA) enables to figure out the position of people and communities within social networks, represented as social graphs. It defines a set of methods and measures, such as graph clustering for community detection or closeness centrality and betweenness centrality, which identify and rank members or communities based on the semantic analysis of the connections found in these social graphs. When these kinds of methods and measures also take into account the semantics of the digital content shared within social networks or semantic information about people, SNA turns into **Semantic Social Networks Analysis** (SSNA).

Introduction

Standard SNA measures mostly consider ties and relationships within social networks and thus remain blind to the semantics of the digital content shared by their members and/or implicitly expressed by their profiles. Therefore, searching opinion leaders within a planetary network such as Facebook or MSN using SNA measures generally returns the most mediatized people, whether they are journalists, politicians, or international artists.

Indeed, most SNA methods and measures are based on the statistical analysis of social graphs topology (Freeman et al. 1989). A graph *G(V; E)* is a set of vertices and a set of edges. Each element of *G(V; E)* is possibly weighted and/or labeled with one or more values. As a result, we find structure-based measures, such as the stress centrality defined in (Shimbel 1953), and flow-based measures (Newman 2005) for undirected or directed graphs, integrating various metrics such as information flows or virality (Brandes and Fleischer 2005; Miramontes and Luque 2002).

Semantic SNA is mostly based on interdisciplinary models merging SNA and knowledge engineering (KE). On the one hand, it refines SNA measures and metrics in order to enhance the processing of data, text, and knowledge tied to the members of social networks. On the other hand, it refines KE principles, techniques, and methods such as linguistic statistics and ontologies, so as to provide KE capabilities adaptable to SNA models. Therefore, semantic SNA measures make possible to retrieve opinion leaders within a large network, for specific topics or keywords. For instance, the semantic betweenness centrality defined in (Thovex and Trichet 2012) enables to retrieve polyvalent experts in specific domains of professional activity defined by seized keywords, such as "database administration and website management," even if managers are much more connected and relay more communications than technical experts within the enterprise social network.

Sense and semantics are extracted from social data and contents such as text, relation types and quantitative or qualitative data related to social content. As a consequence, Semantic Social Networks Analysis defines new hybrid approaches and tends to provide openings in artificial intelligence.

Key Points

This essay proposes an insight of theoretical aspects of semantic social networks analysis, of its epistemic extents, and of the applications it enables to develop. Based on the state of the art in the domain, we study the theoretical foundations of SSNA from their graphic aspects such as topology and flows or static and dynamic behavior within social networks, to their KE aspects such as data mining, text mining, graph mining, or ontologies and the semantic Web. Some theoretical aspects are illustrated with an application of SSNA for enterprises and with examples of applications impacting our social life, economic life, professional life or private life. Lastly, the future and theoretical directions of SSNA are presented under the epistemic aspects of the way paved by SSNA foundations, and future applicative directions are explored in terms of social, economic, and strategic outcomes.

Historical Background

Some premises of SNA have appeared in (Moreno 1934) with the notions of sociogram and sociometry, then in (Freeman et al. 1960) concerning the study of social relationships and leadership in communities. Introduced as a sociological discipline, SNA started to have recourse to mathematics and statistics to develop new measures adapted to large-scale analysis, mostly centralities and modularity measures (Shimbel 1953; Freeman 1977). These measures and metrics are now considered as standard in SNA, and while sociologists carry on studying socialization and group behaviors (Tajfel et al. 1971), the theoretical foundations of standard measures continually inspire new refinements (Brandes 2001; Miramontes and Luque 2002; Pearson and West 2003), so as to face the new challenges raised by the planetary networks of the social Web – e.g., Twitter, Facebook, MSN, Instagram, Wechat. As the Web started to be semantic before being social, semantic SNA is currently becoming a main stream in SNA (Erétéo 2011; Thovex and Trichet 2012).

Increasing interest in Web information retrieval led to the semantic Web initiative (Berners-Lee et al. 2001) from the World Wide Web Consortium in 2001. Semantic standards have been widely used since then, even outside the scope of the Web. However, the main limit in the use of such techniques is the need for explicit semantics from users, as fully automatic semantic annotation systems are not yet reliable. The semantic Web is well fitted to merge with the social Web as both content and actors are generally strongly tied, through semantics-to-content relationships and members-to-members relationships (*e.g.*, finding appropriate files/endpoints within peer-to-peer networks). Defining enhanced capabilities for SNA and taking advantage of the semantic Web, semantic social networks analysis tends now to provide new decisional models and openings in artificial intelligence.

Semantic SNA Is Interdisciplinary

Sociology, mathematics, knowledge engineering, these three disciplines summarize the interdisciplinary aspect of semantic social networks analysis. In this section, we present SSNA from the standpoint of computer sciences, focusing on the theoretical aspects of standard SNA models then on knowledge engineering techniques, before summing up with an epistemic overview of the main conceptual bridges discovered in the presented domain.

Theoretical Aspects in SNA

Standard SNA models are mainly based on the study of topology and flows within social graphs. We differentiate static models and dynamic models.

Topology and Flows

A graph is identified by its vertices and its edges (connections in the case of social networks). When connections are distributed depending on a Gaussian law, the graph structure is named random graph (Erdos and Rényi 1959), and when they depend on a power law (i.e., the probability for a given node to be connected to k other nodes is proportional to $k^{-\gamma}$, where γ is a parameter generally comprised between 2 and 3), the structure is called scale-free network (Barabasi and Albert 1999); scale-free networks contain many nodes with a very low number of connections, and a few “hubs” connected to many other nodes (“rich get richer”). The Web and social networks are identified as scale-free networks depending on preferential attachments (Barabasi and Albert 1999). Standard SNA measures are sensitive to topology because they generally follow geodesic paths – i.e., shortest paths connecting pairs of nodes (i, j) within a graph – so as to proceed to pairwise comparisons of nodes such as in the betweenness centrality defined in (Freeman 1977) as follows:

$$C_{B(P_k)} = \sum\nolimits_{i>j}^{n} \sum b_{ij(P_k)}$$

The definition above adds $b_{ij(P_k)} = 1$ to the betweenness centrality of a point P_k for each geodesic path between the pair of nodes (i, j) comprising P_k, and so on for each pair (i, j) of a social graph. It has been successfully implemented and experimented in (Eréteo 2011), in the context of a project deploying semantic Web languages and tools (i.e., RDF, SPARQL) on a professional dataset based on semantic annotations and collaborative documents sharing – cf. *Ontologies and the Semantic Web* section.

Integrating flows values enhances the results of SNA models, because it fosters the discrimination of representative positions such as leaders or eccentric influencers within social networks. It enables to differentiate hubs regarding the information they share, and to take into account various flows metrics such as read/written textual content, social media viewing/listening, shared knowledge, positive/negative opinions, or friendliness (Chen and Qi 2011; Zhuhadar et al. 2011).

In order to produce relevant flows values, knowledge engineering techniques enable to define semantic flows metrics based on the content shared within social networks. For instance, the study of professional skills and activities in enterprises and/or institutions social networks introduces metrics of semantic intensity (*SemI*) and semantic resistance *SemR* based on linguistic analysis techniques (Thovex and Trichet 2012), such as in the following definitions:

$$SemI_{U,T,D} = TF(T, D); SemR_{U,T,C} = IDF(T, C.)$$

In these definitions, U, T, D, C represent respectively a node, a term, a document, and the corpus of text documents tied to the studied social networks. *TF* and *IDF* are well-known measures in the domain of linguistic statistics, which are trivially defined as follows:

$$\text{Term Frequency}_{(\text{term})} = |\text{term occurences}| \times / |\text{terms} \in \text{document}|$$

$$\text{Inverse Document Frequency}_{(\text{term})} = \log(|\text{documents} \in \text{corpus} | / |\text{document} \ni \text{term}|)$$

They are introduced in SSNA by Robertson and (Sparck Jones 1976) before to be coupled with semantic indexation and research services, so as to produce semantic metrics that enable to value the ties between people and terms within a social graph, depending on the endogenous content – i.e., the content generated and shared within the studied social network. With such a graph, when i represent an individual and j represents content, $b_{ij(P_k)} = 1$ is easily weighted using the semantic edge metrics SemI and/or SemR as factors, in order to define a new semantic betweenness centrality and new semantic centralities based on standard SNA measures and paths walks (Newman 2005). As *SemI* and/or *SemR* are calculated for a subset of edges, we have defined a dynamic model propagating the metrics in a coherent way within the whole graph. Therefore, all nodes and arcs of social and semantic networks are endowed with semantic values respecting an overall and coherent heuristic.

Static Models and Dynamic Models

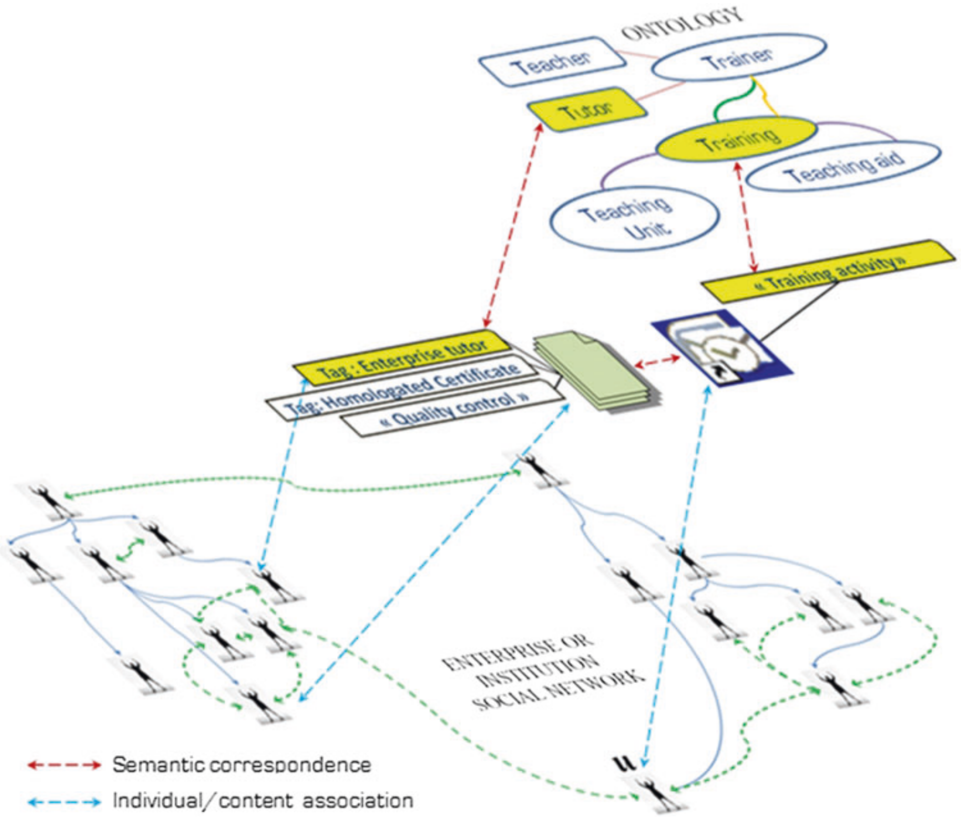
It seems essential to differentiate static SNA models, in which the values found within social graphs are not dependent on each other (i.e., static values), from dynamic SNA models in which the values are (temporally or not) dependent on each other, like in electric circuits where the current of a part depends on the other parts. This metaphor is significant, regarding the main contributions to dynamic SNA based on the analogy between information and electronic flows (Newman 2005; Brandes and Fleischer 2005). It is also developed in physics, introducing SNA measures so as to prevent failures in electric power grids (Wang et al. 2010).

Static models are powerful when social graphs under study are fully weighted before being analyzed. For instance, applying the metrics *SemI* and *SemR* to a network representing the relationships between the members of an enterprise and the terms found in the mails they exchange, it is possible to define a social graph in which all edges are weighted by semantic values compared to current flows (Newman 2005). In such a context, a semantic referential such as an ontology representing the terms found in the network should increase the weighting relevance of each term, using semantic metrics such as defined in (Aimé et al. 2010). Merging semantic networks and social networks fosters the development of relevant SSNA models.

The example of hybrid model illustrated in Fig. 1 is a case in point. It represents a multilayered view of semantic network and social network merged in a single structure. At the bottom of the picture, the dotted lines represent collaborative social relationships, and the full lines represent organizational relationships, within an enterprise social network. The individual u shares the content of an email comprising the expression “training activity.” Through the term “training,” represented in the semantic layer at the center of the picture, and thanks to the ontological relationships about this term, the email associated to the individual u is associated to the documentary resources comprising terms or annotations similar or close to “training” – e.g., “tutor.” Individuals associated to these documentary resources are then more tightly associated to u, thanks to shared knowledge. So, the three individuals pointed by the individual/content associations (blue arrows) become prominent nodes of a same semantic and social sub-graph related to training and tutoring.

Moreover, the other expressions indexing the documentary resources of the socio-semantic sub-graph detected (i.e., “homologated certificate,” “quality control”) can help in the automatic classification of endogenous resources, having recourse to semantic indexing and natural language processing techniques.

Dynamic SNA models enable to introduce heuristics based on natural behaviors, such as encountered in physics or biology (Galam 2008; Giugliano 2009). For instance, with the enterprise social graph that we previously took as an example, applying *SemI* and *SemR* only produces weighs on the edges, not on the nodes representing people or terms. Furthermore, the metrics are not always coherent within the whole graph and could transgress simple rules such as “for each node, the sum of incoming flows equals the sum of outgoing flows.” The issue is solved by a dynamic SSNA model implementing Kirchoff’s and Ohm’s laws (Thovex and Trichet 2012). This model enables to weigh the whole graph by ensuring the coherence of all weights, according to the natural balance of electronic flows in solid state circuits. This naturally coherent heuristic still does not take electromagnetic losses and interactions



Semantic Social Networks Analysis, Fig. 1 Semantic network and social network merged in a single structure

into account but improves previous epistemic approaches (Newman 2005; Brandes and Fleischer 2005), which introduce Kirchhoff's point law in dynamic SNA without integrating the Ohm's law, although it is a prerequisite in physics. The dynamic method of flow propagation defined in (Thovex and Trichet 2012) owns two temporal aspects. On the one hand, it produces a coherent distribution through the whole graph, of the semantic values coming out from *SemI* and *SemR* on the edges connecting people to content. This phase enables to weight nodes and people-to-people and/or content-to-content edges, in a coherent way. On the other hand, temporal changes occurring within the input dataset might be processed, so as to compare the states of studied socio-semantic networks together and to produce temporal analysis of socio-semantic networks following a timeline.

Theoretical Aspects in Knowledge Engineering

Knowledge engineering aims at integrating knowledge in computer systems in order to lower the need for human intervention. Two main issues need to be addressed: knowledge discovery (i.e., mining techniques such as linguistic statistics) and knowledge representation and exchange – i.e., semantic formalisms. These issues are developed in the following sections.

Data Mining, Text Mining, and Graph Mining

Data mining aims at automatically finding patterns (such as rules or outliers) in large datasets.

The underlying motivation comes from the data explosion that has taken place since several decades. Daily data generated by social networks contribute to the apparition of Big Data techniques. Moreover, a significant part of these data needs to be processed in real time, which represents an additional challenge to the one related to scalability (off-line vs. online analysis).

Many scientific areas contribute to data mining techniques, e.g., statistics, artificial intelligence, machine learning, and optimization. Data mining solutions include data classification or clustering, association rules generation, and weak signals or outlier detection.

Data mining is applied to many sectors, among which text analysis, Web mining, marketing, financial or biological data analysis, or fraud detection. Moreover, the development of networked data such as computer, biological, or social networks has created new challenges for data mining and graph mining in particular. Indeed, these – often large and heterogeneous – real networks also called *complex networks* may be represented as graphs. Complex networks analysis has raised interest in the scientific community, and various graph mining techniques have therefore been developed in order to describe these real graphs and design models for generating realistic networks. Another trend in graph mining consists in identifying clusters of strongly connected nodes in the network, called communities (Fortunato 2010). Finally, very little is known about complex networks dynamics, and much remains to be done – e.g., study of communities evolution over time, ties, and interactions between data types.

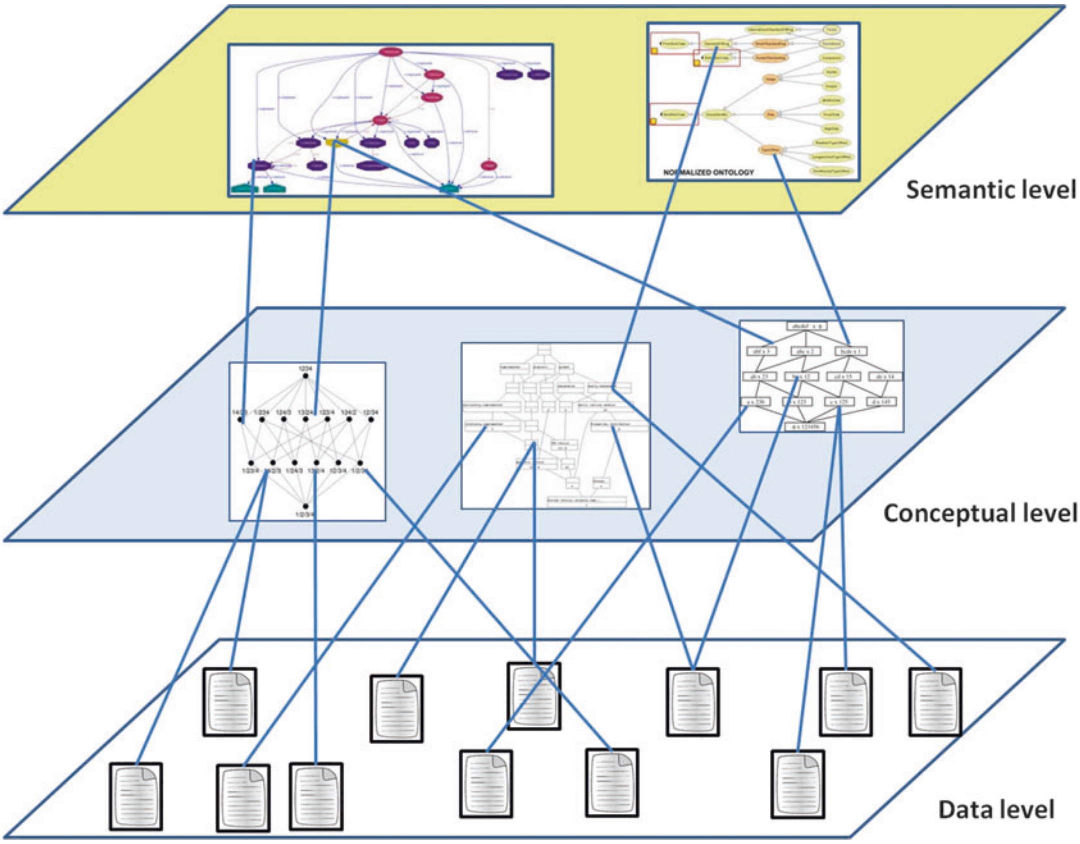
Ontologies and the Semantic Web

Various formalisms exist within the semantic Web framework, with different levels of complexity and expressiveness, from simple annotation syntaxes to sophisticated reasoning capabilities. The eXtensible Markup Language (XML), the Resource Description Framework (RDF) (Lassila and Swick 1998), ontologies (Gruber 1993), rules, and logic all belong to the semantic Web picture. Many definitions of ontologies may be found in the literature; among them, Tom Gruber's (1993) is frequently referred to: "*An ontology is a formal specification of a shared conceptualization*." An ontology basically describes concepts and the relationships among these concepts. A thesaurus may be seen as a light ontology as it also describes concepts, but the relationships among them are not specified as formally as in ontologies.

Conceptual graphs (Sowa 1976) constitute a way to represent and organize knowledge. Such graphs may be built from structured or unstructured data, for example, through the computation of Galois lattice based on formal concept analysis – FCA (Ganter and Wille 1998). From a set of elements (called *objects* in the FCA terminology) described by their properties (called *attributes*), a Galois lattice builds a partially ordered set of concepts, consisting each in objects sharing common attributes. Based on semantic attributes and relationships, it defines semantic networks and clusters that can be compared to the notion of community encountered in SNA. Semantic networks such as Galois lattices and ontologies represent the topology of semantic relationships between concepts enriched with various qualitative information. Hence, intrinsic features of members of a social network such as age, education, address, or hobbies (i.e., profiles) may be used for the identification of communities or for the recommendation of new contacts.

Conceptual graphs, ontologies, and thesauri are excellent candidates for supporting the convergence of semantic analysis and social network analysis, since both disciplines are based on graphic representations and heuristics. Recent research has shown that graph theory and natural language understanding are intimately connected. It brings together topics as diverse as lexical semantics, text summarization, text mining, ontology construction, text classification, and text retrieval (Rada and Dragomir 2011).

Conceptual graphs and semantic networks provide an intermediate layer between analyzed data and semantics, as shown on Fig. 2. Indeed, they have been successfully used for the analysis of social networks extracted from Myspace, Flickr, Dailymotion (Riadh et al. 2009), and Twitter (Melo et al. 2012). A node of the conceptual layer may be linked to several nodes of the semantic layer, creating a bridge among various ontologies. New similarity metrics for ontology matching may also be derived from graph-based metrics. Conversely, a semantic node may be related to distinct concepts, allowing the



Semantic Social Networks Analysis, Fig. 2 Three-tier architecture: data, conceptual, and semantic levels

navigation from a conceptual graph to another via the semantic layer.

Figure 2 illustrates how, in KE, a semantic network maps onto data. It represents the conceptual bases illustrated in Fig. 1 – which shows an ontology snippet mapped onto a social network *via* the content shared by its members. Similarity between both figures reveals how semantic networks are propitious to define SSNA models, but not only. The conceptual bridge it entails can be crossed from KE to SNA/SSNA, but also from SNA/SSNA to KE, in order to research new methods to build and/or populate semantic networks using SNA/SSNA models. In such a context, the possibility of discovering virtuous and self-learning models appear as a new opening in artificial intelligence.

As explained earlier, the interest of semantics in social network analysis has been acknowledged; conversely, SNA results may help maintain and enrich ontologies. For instance, communities identified through topological links within a social network may correspond to emerging concepts to be added in ontologies. Intra/inter-communitarian ties between members may help with ontology building in the interdisciplinary context of semantic and social networks analysis. This raises challenging research questions in order (1) to identify, within social networks and social content, useful and relevant ties for ontology building and/or matching and (2) to define convolutional self-learning processes in SSNA, or based on SSNA.

Semantic SNA: An Interdisciplinary Approach

In the first part of the current section, we have explored the bases of social networks analysis and shown how semantic SNA enhances standard SNA – integrating semantics related to endogenous content into SNA measures – thanks to knowledge engineering methods such as linguistic statistics. In the second part, we have discovered a singular analogy between social networks and semantic networks, through a presentation of conceptual graphs, Galois lattices, ontologies, and thesauri – i.e., semantic representations based on graphs. We state that it opens an epistemic track paving the way for future and interdisciplinary SSNA models. We define SSNA as an interdisciplinary approach based on SNA and KE. SSNA introduces a generation of models that adapt the results of standard measures and metrics depending on the semantics found in the content shared within social networks. Current experimentations show a significant improvement of SNA results, thanks to SSNA models. We can imagine future extensions like merging opinion analysis into SSNA.

Unfortunately, current SSNA models are mostly dependent on the existence of text within the endogenous content, while social networks include more and more pictures, audio and/or video streams, bookmarks, or geographical locations. In (Sánchez et al. 2007, 2008; Sánchez and Valdiviezo 2011), teams of experts in specific domains and in the needs of users communities, generate tags that are meaningful for multiple contexts of use and dynamic ontologies. They carry out this task as a side effect for their responsibilities of assisting users in exploring complex information networks represented by inter-linked digital documents.

Due to the diversity, volume and permanent updating of the social Web, human annotations are cost-effective only for local applications. Hybrid approaches coupling human participation with automatic means for semantic network analysis are likely to generate effective results for medium scale applications.

Therefore, before providing social and semantic recommendations, the social Web requires a lot of various techniques for processing signal in visual and/or audio streams and for knowledge extraction in bookmarks and locations – e.g., face recognition, multilingual speech to text, musical pattern recognition, Web crawling, linguistic analysis, association rules, and fuzzy logic. Hence, without heavy preprocessing frameworks extracting textual representations and semantics from social media, SSNA omits a large part of the social knowledge it is supposed to process. Though the explicit relationships within social networks and social media provide a turnaround, this seems to be one of the biggest obstacles to the deployment of SSNA in the socio-semantic Web, with more general problems such as processing very large datasets, mining the hidden/deep Web, or subtle psychosocial knowledge regarding human behaviors, hidden intentions, and subconscious opinions.

Key Applications

As social networks touch our social life, private life, economic life, and professional life, the application domain of SSNA is potentially vast and linked to SNA applications. One of the first major trends we have seen developed is criminal networks analysis for counterterrorism. Obviously, SNA/SSNA is an important decisional leverage for marketing agencies and strategies. In 2008, two founders of Facebook declared having to leave the enterprise to conceive a new kind of products that "will become to your work life what Facebook.com is to your social life" – cf. ASANA and http:/en.wikipedia.org/wiki/Dustin_ Moskovitz. In the same time, enterprise social networks became as usual as mail exchanges in certain professional branches, and when they are consensually accepted, they are considered as tools fostering collaboration and productivity. They could be also precious for human resources management and social capital management.

Experimenting the semantic metrics and measures defined in (Thovex and Trichet 2012) on collaborative enterprise dataset, we have identified and ranked significant terms and teams within skills networks – i.e., socio-semantic networks representing professional collaborations. As a result, SSNA of skills networks provided a set of relevant indications helping in (1) self-managed collaboration and teams organization; (2) detection of critical topics, in terms of stress at work;

and (3) redeployment of human resources, according to dynamic requirements in terms of competencies and workload. Evaluating our predictive and epistemic model with the experts involved in experimental phase, the produced recommendation enabled to retrieve a group of users sharing an anonymous account, though there was no explicit data allowing to identify these users in the studied dataset.

In the past decade, new forms of communication, such as micro-blogging and text messaging have emerged and become ubiquitous. While there is no limit to the range of information conveyed by micro-blogging and text messaging, often these short messages are used to share opinions and sentiments that people have about what is going on in the world around them. Those texts are typically short and the knowledge representation of their content for automatically mining and understanding the opinions and sentiments that people communicate has become an attractive research topic such as in (Pang and Lee 2008). Promising results have been obtained in (Castillo et al. 2015a) using centrality measures to extract the most representative words for common sentiments. These words are successfully used in a supervised learning algorithm as features to obtain the polarity of unknown documents. Further analysis on the use of centrality measures and on the methodology for constructing the network of words will allow to find more accurate semantic features that could be used in new supervised learning methods as presented in (Castillo et al. 2015b).

Recommender systems are changing the way people find products, points of interest, services or even new friends. The models of recommendations now processes user qualitative preferences and social influences, such as in SSNA-based recommender engine for tourism and smart cities presented in (Cervantes et al. 2015a). It is based on linguistic statistics, social networks analysis and hybrid graphs structures that represent interactions in-between points of interests, territorial knowledge and ubiquitous mobile uses. The model enables personalized recommendations fostering social and territorial uses, thanks to centrality measures, semantic predominance and similarity metrics applied within social and semantic graphs.

We should not be surprised to find various uses of SSNA by the GAFA (Google, Apple, Facebook, Amazon), as it is a key domain for viral and/or predictive marketing and other strategic analytics. Other social networks platforms such as Instagram or FourSquare are obviously concerned but not only. Indeed, SSNA can benefit to applications based on social interactions tracks created by digital humanities, in professional or personal information systems such as workflows and mails databases.

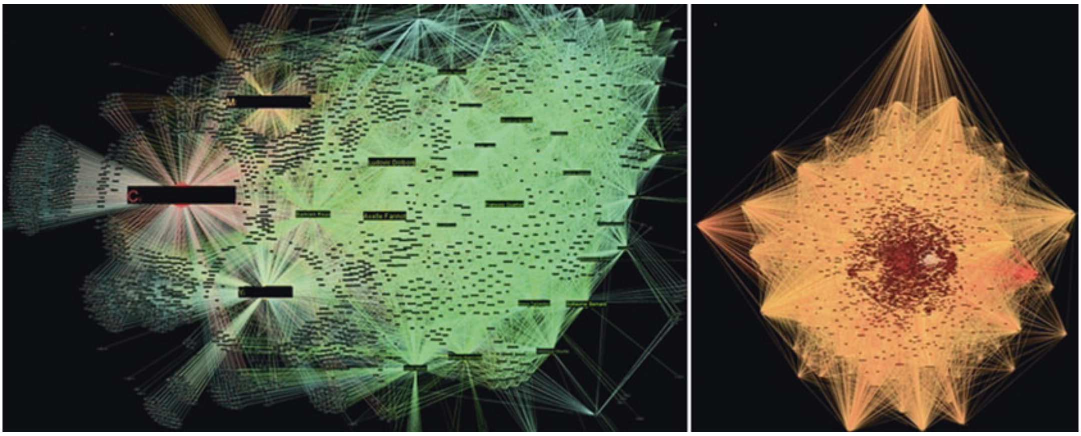
Future Directions

It has been shown that the Semantic Social Network Analysis field is growing and expanding its horizons to involve more disciplines and to support new decision making approaches applied in various fields. The SSNA is a research area that continues opening challenges and opportunities for proposing new conceptual proposals and effective solutions.

Semantic SNA: An Interdisciplinary Way to Be Paved

We have presented earlier the interdisciplinary dimension of SSNA. According to our knowledge of the domain, this dimension raises an unsuspected set of epistemic issues far beyond the analogy between electric current flows and information flows in social networks developed in (Newman 2005); (Brandes and Fleischer 2005); (Thovex and Trichet 2012). The interdisciplinary dimension of SSNA can be explored in depth as well as transversally.

Pursuing the in-depth exploration of epistemic equivalences between electrophysics and SSNA, we could intend to merge electromagnetic and thermodynamic principles into our current model (e.g., Maxwell’s equations, Joule effect), so as to detect probable and implicit information flows not visible within the graph, semantic ties or risks of psychological burnout within social networks.



Semantic Social Networks Analysis, Fig. 3 Visualization of SSNA results – samples

Such an epistemic metaphor might lead to introduce the Schrödinger's and/or Dirac's equations into SSNA, researching a future generation of quantum analytics for digital humanities.

Figure 3 represents a sample of our experimental results, studying collaboration relationships within an enterprise. At the left on the picture, weighted degree centrality (i.e., the sum of all weights from edges connected to a node) based on *SemI*SemR* (named "semantic tension") defines the size and color of nodes and the color of edges, from light blue for weak values to red for high values. It enables to identify the most important collaborators (largest hubs at the left) in terms of skills and knowledge (small nodes/terms around the hubs). The same dataset is represented at the right on the picture, based on semantic closeness centrality values such as defined in (Thovex and Trichet 2012). Orange color shows how semantic closeness is concentrated on average values, in this case. The most common terms are tied with most of the collaborators, represented as the core of the network. At the periphery, we find eccentric collaborators working on rare terms, sometimes with high semantic tension (red edges), which represent rare but important knowledge/skills within the enterprise.

Transversally exploring the interdisciplinary dimension of SSNA, we might discover epistemic connections between biological similarities and socio-semantic networks formulations, between knowledge networks and collective behaviors or geography of social networks, for instance.

From Social Outcomes to Strategic Outcomes

While social networks thoroughly describe our social, private, economic, and professional lives, SSNA outcomes are gradually turning into strategic outcomes. The sum of indications and recommendations they provide quickly becomes strategic for economy, politics, education, and information sharing all around the world. Observing trends such as wearable devices and quantified-self, we can guess that the diversification and multiplication of sensors within the Internet of Things (IoT), will turn knowledge graphs and digital social networks into an augmented paradigm of digital humanities. A large range of our societal uses should be impacted by the future analytics adapted to such a data deluge.

Privacy is an important stake while social data become signals for economy, education, politics and strategic governance. Empowering social ties within society, current and future SSNA applications might play a major role in future social networks, regarding security *versus* individual freedom, and evolution of societal organization such as participative democracy, for instance. Based on current facts and trends, citizens and governments might take care of the benefits of SSNA, keeping them predominant while facing the perverse effects possibly resulting from the digitalization of humanity.

Cross-References

- ▶ Analysis and Mining of Tags, (Micro)Blogs, and Virtual Communities
- ▶ Classical Algorithms for Social Network Analysis: Future and Current Trends
- ▶ Community Detection: Current and Future Research Trends
- ▶ Dynamic Community Detection
- ▶ Managerial Networks
- ▶ Modeling Social Preferences Based on Social Interactions
- ▶ Ontology Matching
- ▶ Path-Based and Whole-Network Measures
- ▶ Recommender Systems, Semantic-Based
- ▶ Role Identification of Social Networkers
- ▶ Semantic Social Networks
- ▶ Social Capital
- ▶ Social Interaction Analysis for Team Collaboration
- ▶ Social Network Analysis in Organizational Structures Evaluation
- ▶ Social Recommendation in Dynamic Networks
- ▶ Virtual Team

References

Aimé X, Furst F, Kuntz P, Trichet F (2010) Prototypicality gradient and similarity measure: a semiotic-based approach dedicated to ontology personalization. J Intell Inf Manag 2(2):65–158. 2150–8194

Barabasi A-L, Albert R (1999) Emergence of scaling in random networks. Sci Mag 286(5439):509–512

Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Sci Am Mag 284(5):34–43

Brandes U (2001) A faster algorithm for betweenness centrality. J Math Sociol 25:163–177

Brandes U, Fleischer D (2005) Centrality measures based on current flow. In: 22nd symposium theoretical aspects of computer science (STACS 05), Stuttgart, LNCS, vol 3404, Springer, pp 533–544

Castillo E, Cervantes O, Vilariño D, Báez D, Sánchez A (2015a) UDLAP: Sentiment analysis using a graph based representation. In: Proceedings of the 9th international workshop on semantic evaluation (Sem Eval 2015), pages 556–560, Denver, 4–5 June 2015. 2015 Association for Computational Linguistics

Castillo E, Cervantes O, Vilariño D, Báez D (2015b) Author verification using a graph-based representation. Int J Comput Appl (0975–8887) 123(15):1–8

Cervantes O, Gutiérrez F, Gutiérrez E, Sánchez A, Rizwan M, Wan W (2015). A recommendation engine based on social metrics. In: The 6th workshop on semantics for smarter cities collocated with the 14th international semantic web conference (ISWC2015) 11–12 Oct, Bethlehem (S4SC 2015)

Chen L, Qi L (2011) Social opinion mining for supporting Buyer's complex decision making: exploratory user study and algorithm comparison. Soc Netw Anal Min J 1(4):301–320. Journal by Springer MathSciNet

Erdos P, Rényi A (1959) On random graphs. Publ Math 6:290–297

Erétéo G (2011) Semantic social network analysis. PhD thesis, Laboratoire d'Informatique, Signaux et Systémes de Sophia-Antipolis (L3S, UMR6070 CNRS), Université de Nice Sophia-Antipolis

Fortunato S (2010) Community detection in graphs. Phys Rep 486:75–174

Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40:35–41

Freeman LC, Bloomberg W, Koff SP, Sunshine MH, Fararo TJ (1960) Local community leadership. University College of Syracuse University, Syracuse

Freeman LC, White DR, Romney AK (1989) Research methods in social network analysis. George Mason University Press, Fairfax

Galam S (2008) Sociophysics: a review of galam models. Int J Mod Phys C 19(3):409–440

Ganter B, Wille R (1998) Formal concept analysis: mathematical foundations, 1st edn. Springer, Berlin

Giugliano M (2009) Calcium waves in astrocyte networks: theory and experiments. Front Neurosci 3(2):160–161

Gruber TR (1993) A translation approach to portable ontology specifications. Knowl Acquis 5(2):199–220

Lassila O, Swick RR (1998) Resource description framework (RDF) model and syntax specification. Technical report, World Wide Web Consortium, Cambridge

Melo C, Le Grand B, Aufaure M-A (2012) A conceptual approach to characterize dynamic communities in social networks: application to business process management. In: BPMS2 2012: 5th international workshop on business process management and social software, Tallin

Miramontes O, Luque B (2002) Dynamical small-world behavior in an epidemical model of mobile individuals. Physica 168–169:379–385

Moreno J (1934) Who shall survive? – (Trad. fr) Fondements de la sociométrie. PUF, Washington, DC

Newman MEJ (2005) A measure of betweenness centrality based on random walks. Soc Netw 27(1):39–54

Pang B, Lee L (2008) Analysis mining opinion sentiment. J Found Trends Inf Retr 2:1–135

Pearson M, West P (2003) Drifting smoke rings: social network analysis and Markov processes in a longitudinal study of friendship groups and risk taking. Connect Bull Int Netw Soc Netw Anal 25(2):59–76

Rada M, Dragomir R (2011) Graph based natural language processing and information Retrieval. Cambridge University Press, Cambridge

Riadh TM, Le Grand B, Aufaure M-A, Soto M (2009) Conceptual and statistical footprints for social networks' characterization. In: SNA-KDD '09: proceedings of the

3rd workshop on social network mining and analysis. ACM, Paris, pp 1–8

Robertson SE, Sparck Jones K (1976) Relevance weighting of search terms. J Am Soc Inf Sci 27(3):129–146

Sánchez JA, Valdiviezo O (2011) Enhancing productivity through social computing. Book chapter. In: Papadopoulou P, Kanellis P, Martakos D (eds) Social computing theory and practice: interdisciplinary approaches. Information Science Reference, pp 133–153

Sánchez JA, Arzamendi-Pétriz A., Valdiviezo O (2007) Induced tagging: promoting resource discovery and recommendation in digital libraries. In: Proceedings of the joint conference on digital libraries (JCDL 2007, Vancouver), pp 396–397

Sánchez JA, Valdiviezo O, Aquino E, Paredes R (2008) REC: improving the utilization of digital collections by using induced tagging. Res Comput Sci 39:83–93

Shimbel A (1953) Structural parameters of communication networks. Bull Math Biophys 15:501–507. Stress rate

Sowa JF (1976) Conceptual graphs for a data base interface. IBM J Res Dev 20(4):336–357

Tajfel H, Billig M, Bundy R, Flament C (1971) Social categorization and intergroup behavior. Eur J Soc Psychol 1:149–178

Thovex C, Trichet F (2012) Semantic social networks analysis: towards a sociophysical knowledge analysis. Soc Netw Anal Min J 2(1):1–15. Journal by Springer

Wang Z, Scaglione A, Thomas RJ (2010) Electrical centrality measures for electric power grid vulnerability analysis. In: IEEE (ed) Proceedings of the 49th IEEE conference on decision and control, CDC 2010, Atlanta, 15–17 Dec 2010, pp 5792–5797

Zhuhadar L, Nasraoui O, Wyatt R, Yang R (2011) Visual knowledge representation of conceptual semantic networks. Soc Netw Anal Min J 3:219–299. Journal by Springer

Recommended Reading

Memon N, Alhajj R (eds) (2010) From sociology to computing in social networks theory, foundations and applications. Lecture notes in social networks, vol 1. Springer, Wien

Semantic Social Web

▶ Twitris: A System for Collective Social Intelligence

Semantic Web

▶ Recommender Systems Based on Linked Open Data

▶ Sources of Network Data

Semantic Web Modeling Language

▶ Web Ontology Language (OWL)

Semantic Web Search

▶ Query Answering in the Semantic Social Web: An Argumentation-Based Approach

Semantic Web Service Composition

▶ Web Service Composition

Semantic Web Services

▶ Service Discovery

Semi-Discrete Decomposition

Cherukuri Aswani Kumar
School of Information Technology and
Engineering, VIT University, Vellore, India

Synonyms

Bump hunting technique; Matrix decomposition; Singular value decomposition

Glossary

Clustering
An unsupervised data mining technique that places the data objects into different groups (clusters) such that the objects in a cluster are more similar to each other and dissimilar to objects in other clusters

Data Matrix — A rectangular array with m rows and n columns, the rows representing different observations or individuals or objects and the columns representing different attributes or variables or items

Data Mining — Nontrivial extraction of previously unknown, potentially useful, and reliable patterns from a set of data

Dimensionality Reduction — The process of embedding a set of n points in a d-dimensional space into a k-dimensional space, where d is sufficiently large and k is much smaller than d

High-Dimensional Data — Data in which the objects are described by a large number of features, where each feature factor corresponds to a dimension. While analyzing a data matrix of size m by n, we refer the matrix as n -dimensional since we consider the view of m points in an n -dimensional space

Matrix Decomposition — Transformation of original data matrix into a given canonical form, as a product of new matrices. This transformation is aimed at revealing the latent structures or relations in the original data matrix. This transformation is also known as matrix factorization

Matrix Rank Reduction — Given a data matrix A having rank r, the process of finding a matrix $\hat{A}$ having rank k where $k < r$ and minimizes $\| A - \hat{A}\|$

Definition

Semi-discrete decomposition (SDD) is a matrix decomposition technique that produces low-rank approximation of original matrix as a weighted sum of outer products. With its approximation, SDD defines new axes that capture the variance in the data. Though this approximation is similar to that of singular value decomposition (SVD), the axes of the SDD transformed space are not orthonormal, and coordinates of the points in the transformed space are restricted to the set of values $\{-1,0,1\}$. With such a restriction, SDD achieves the storage economization than other decomposition techniques like SVD. Hence, the higher-rank approximations can be stored for smaller amount of storage. With an iterative procedure, SDD aims to find and extract the locations in the given dataset having extremely large magnitude values which are both positive and negative. SDD represents the data matrix as the sum of bumps and arrange the bumps such that the most significant bump appears first. Hence, SDD is generally treated as a bump hunting technique and proved to be effective in finding the outlier clusters in the data. Also SDD produces an unsupervised, hierarchical, and ternary classification by partitioning the data items having similar attribute values. Hence, SDD is applied in classification and clustering problems. SDD has a unique property of discovering more latent factors than the available features in the dataset. In addition to its primary motivation in digital image processing, SDD has successful applications in finding outliers in the data, semantic indexing, etc.

Introduction

Most of the engineering, scientific, and computer applications result in high-dimensional datasets containing large number of variables associated with each observation. Also such data is often a combination of several underlying processes coupled with noise. The dimension of a dataset is defined by the number of variables that are measured on each observation. However, all these variables are not necessary to understand the latent structure of the data. Such highdimensional data coupled with noise poses several computational challenges. In addition to the complexity prevailing in analyzing the highdimensional datasets, the similarities between the objects in the high-dimensional space diminish with regard to the Euclidean distance. This would negatively influence the accuracy of the analysis. This problem is referred to as curse of dimensionality (Cunningham 2007). The solution to this curse is to apply dimensionality reduction techniques as a preprocessing step. This process requires identification of a suitable low-dimensional representation of original high-dimensional data. Also this preprocessing step improves the accuracy of the data analysis (Dobsa et al. 2012). Several dimension reduction techniques are available in the literature. Interested readers can refer to few authoritative references (Kumar 2009; Cunningham 2007; Fodor 2002). An optimal technique can efficiently map the original data to suitable lower dimension while preserving the properties of original data. By representing the data in the form of a matrix, we get a convenient way to store and analyze data. If a data matrix A of size m × n, containing m objects and n attributes, each object can be considered as a point in the n-dimensional space spanned by the attributes.

Matrix rank reduction techniques from linear algebra are popular in data analysis and mining problems for finding low-dimensional representation of data (Elden 2006). Rank of the matrix is defined as the number of linearly independent rows or columns of the matrix and helps to measure the contents of the matrix. Redundancy in the matrix arises due to the dependent row or column vectors. This redundancy can be removed by mapping or replacing the dependent vectors with linear combination of other linearly independent vectors.

Generally decomposition of the matrix refers to the decomposition to some approximation. Decomposition of a matrix produces two or more factor matrices. The original matrix can be represented as a product of these factor matrices. The main motivation behind the matrix decomposition lies in the fact that the inner dimension value of k is much smaller than the original dimensions (m;n) of the data matrix. The matrix decomposition techniques are mainly intended to segregate the different processes that are captured by the dataset and to cluster the similar objects of the dataset in some standard understandable way. These techniques can be applied as a stand alone or in combination of other techniques.

The notions like dimensionality reduction, matrix rank reduction, matrix factorization, and data compression are closely related and are based on Wedderburn rank reduction theorem (Miettinen 2009; Park and Elden 2003; Elden 2007; Skillicorn 2007).

Several matrix rank reduction techniques are available that include singular value decomposition (SVD), semi-discrete decomposition (SDD), and nonnegative matrix factorization (NMF) (Miettinen 2009). Each of these techniques differs in the way they decompose the matrix, constraints that they impose on the elements, relationship among the rows and columns, etc. Recently heuristic techniques like clustering and random projections are also used in the literature for matrix rank reduction (Kumar and Srinivas 2010; Kumar 2011).

Key Points

Semi-discrete decomposition (SDD) was originally introduced by O'Leary and Peleg (1983) for the purpose of digital image compression. Later it is extended as a storage efficient variant of SVD in latent semantic indexing (LSI)-based IR application. Based on vector space representation, LSI finds low-rank approximation of term document collection using SVD (Kumar and Srinivas 2006; Berry et al. 1999; Deerwester et al. 1990). However, if the original matrix is sparse, the low-rank approximation achieved through SVD requires more storage than the original matrix. To overcome this difficulty Kolda and O'Leary (1998) have proposed to use SDD for LSI. An analogy can be brought between SDD and SVD, Boolean matrix decompositions. Both the factor matrices and the matrix multiplication in Boolean decompositions are binary (Miettinen 2009). Similar to SVD, SDD obtains three matrices, but elements of outer product vectors are restricted to -1,0, and 1.

Given a data matrix A of size.(m,n), with m objects and n attributes, SDD finds the approximation of A to a lower dimension k as follows:

$$A_k = X_k D_k Y_k^{\mathrm{T}}$$

where the matrix X_k is of size m × k; D_k is a diagonal matrix of size k × k, and Y_K^T is a matrix of size k × n. The entries of D_k matrix are nonnegative real numbers. Each D_i value ($1 \leq i \leq k$) indicates significance of the ith factor. The rows of the matrix X_k are considered as the coordinates of an object in the space defined by the new axes that are described by the rows of Y_k^T. The variation in original data is captured and is concentrated along the earlier axes defined by Y_k^T. The lower axes in which the lesser variance in the data is concentrated can be removed to achieve the approximation. The axes of the transformed space are not orthonormal. The coordinates can have different interpretations depending on the application.

The elements of the matrices X_k, Y_k^T are from the set $\{-1,0,1\}$. This transformation of original n-dimensional space into new k-dimensional space results in dimensionality reduction of original matrix. Generally in SVD applications the value of k will be chosen as $m \leq k \leq n$. However in SDD, the value of k can even be higher than n. Hence, SDD can identify the more latent factors than the existing features in the dataset. SDD tries to define new axis that captures larger variation in the original data. The algorithm starts by identifying the values of first column of the matrix X_k, the first axis vector of Y_k^{T} and multiplier in D_k that gives the least amount of error between the approximation matrix, A_k and the original data matrix A. The iterative process continues by selecting successive fields of these matrices in such a manner that reduces the error in the approximation.

The SDD decomposition can have three types of interpretations, namely, factor interpretation, geometric interpretation and component interpretation, (Skillicorn 2007). By considering the rows of the matrix Y_k^{T} as factors that are mixed by the rows of X_k and diagonal entries of D, we can obtain the factor interpretation. This representation is useful in image processing. In geometric interpretation, the rows of the matrix Y_k^{T} define the generalized quadrants, and the values of the matrix X_k can then identify whether a given object is placed in the given quadrant or not. Component interpretation can be obtained by expressing the original matrix A as sum of the outer product matrices, i.e., the ith column of matrix X, the ith entry on the diagonal of the matrix D, and the ith row of the matrix Y^{T}. Though graph interpretation for SDD can be obtained, it provides no new insights about the data.

The approximation of a matrix using SDD is achieved through an iterative and greedy algorithm, which computes a new column, a diagonal element, and a row in each step. Let A be the data matrix of size m × n. Choose a value k that represents the maximum number of terms in the approximation. Let A_0 be the zero matrix of size m × n, x_i be the ith column of the matrix X_k , d_i be the diagonal element of the matrix D_k , and y_i be the ith row of Y_k^{T}. Let R_i be the residual matrix obtained at ith step, i.e., $R_i = A - A_{i\,-\,1}$. Consider $R_1 = A$. In the following we present the algorithm:

1. Outer iteration, for each step of i = 1 to k.
2. Choose an initial y vector such that $R_i\, y \neq 0$.
3. Inner iteration:
   i. Fix y and let x solve $\max\limits_{x \in \mathrm{I}^m} \dfrac{(x^T R_i y)^2}{\|x\|_2^2}$.
   ii. Fix x and let y solve $\max\limits_{y \in \mathrm{I}^n} \dfrac{(y^T R_i x)^2}{\|y\|_2^2}$.
   iii. Repeat the inner iterationuntil some heuristic convergence criterion is satisfied.
4. Let x $x_i = x, y_i = y, d_i = \frac{x_i^T R_i y_i}{\|x_i\|_2^2 \|y_i\|_2^2}$.
5. Calculate the ith term approximation $A_i = A_{i-1} + d_i x_i y_i^T$.
6. Calculate the residual matrix, $R_{i+1} = R_i - d_i x_i y_i^T$.
7. Repeat the outer iteration until i = k.

The convergence criterion for stopping the inner loop is to verify whether the residual improvement is further possible. O'Leary and Peleg (1983) have proposed a method to determine the condition for stopping the inner iterations. Computing SDD on the data matrix of size (m;n) to approximate it to a rank k, under the assumption of fixed number of inner iterations, the above heuristic algorithm has a complexity of O.(k^2.(m + n) + mlog m + nlog n). Generally it is observed that the number of inner iterations required is averaged near 10. Kolda and O'Leary (2000) have shown that the SDD algorithm converges linearly to the original matrix. Also they have discussed strategies that can be used to initialize the y vector in outer iteration shown in step 2. Since the algorithm is a heuristic variant, the parameters need changes depending on the dataset. Implementation of this algorithm in MATLAB and C is available from http://www.cs.umd.edu/~oleary/SDDPACK.

In the SDD basic setting, the heuristic component is the selection of initial y_i . This selection does not always identify and remove the largest possible bump from the data matrix. Hence, a rearrangement of these bumps is required (Skillicorn 2007). Once the X_k, Y_K^T, and D_k matrices are computed, the product of d_i^s with corresponding nonzero column entries of Y_K^T is formed. The columns of X_k, elements of D_k, and rows of Y_K^T are sorted into decreasing order of the products of d_i^s with Y_K^T. This reordering ensures that the axes with largest weight or the axes that capture large variation appear first in the ordering, and hence the strongest outlier will be placed closest to the top of the decision tree (Knight and Carosielli 2003).

Since elements of the outer product matrices X_k and Y_K^T obtained from SDD contain the values {−1, 0, 1}, higher-rank approximations can be stored at less amount of space. For rank k approximation of matrix of size (m,n), SDD requires the storage of k(m + n) values from the set {−1,0,1} for the matrices X_k, Y_k^T and k scalar values for the matrix D_k. To store the values from the set {−1,0,1} requires log $_2$3 bits. The scalar values for the matrix D_k need to be only single precision values. However, the SVD is computed with double precision values and hence requires nearly 32 times more space than SDD (Kolda and O'Leary 1998).

Unlike SVD, even for value k = n, the SDD does not produce the approximation matrix that is equal to the original matrix, i.e., $A_k \neq A$ for k = n (Snasel et al. 2008). When the data is organized naturally in many small and well-separated clusters, SDD and SVD tend to agree and hence produce similar results. This is the main reason for usage of SDD as a replacement of SVD in LSI, since term-document matrices usually contain several natural small clusters (Kolda and O'Leary 1998, 2000). However, SVD and SDD do not produce similar results on the datasets that are organized in the form of large clusters. The basic problem with SDD is that the approximation takes five times more time than computing SVD. However, SDD updating is much easier than the SVD updating (Kolda and O'Leary 2000). SDD can be extended as weighted SDD and tensor SDD. These extensions along with their convergence issues are discussed in Kolda and O'Leary (2000).

Objects of the data matrix A can be hierarchically classified using the columns of the matrix X_k. The analysis start, with the first column of the matrix X_k. Objects (i.e., rows) of the matrix A are divided into three classes according to the value {−1,0,1} that appears in the first column of the matrix X_k. The objects whose value is +1 in the first column of X_k are in one class, the objects whose value is −1 are in one class and the objects whose value is 0 are in the third class so that the classification forms a ternary decision tree structure.

From each class, the objects are further divided into three subclasses depending on the value {−1,0,1}, corresponding to each object of matrix A, in the second column of the matrix X_k and so on. The process can be stopped when a set of objects cannot be separated by the next levels or when each object is alone at a particular level. The analysis generates a ternary, hierarchical decision tree structure of depth k. In contrast to the conventional decision trees, the decision tree induced by the SDD is an unsupervised structure. By following the same procedure on the Y_k^T matrix, we can obtain the hierarchical classification structure of attributes. The general notion is that the classes −1 and 1 represent the data objects that have attributes significantly different from the normal data objects represented by the class 0.

In another perspective, by treating each class as a partition, we can consider that SDD performs partitional clustering. The division of data objects into three groups using the first column of matrix X_k and further subdivision of each group based on the subsequent columns of X_k results in a hierarchical clustering of the objects of A. The clustering contains k levels. The partitions at each level are independent. Unlike standard hierarchical clustering, the result of SDD-based hierarchical clustering is a ternary tree. The branches with the groups −1 and +1 are equal and opposite, but not different. Similarity among the objects or attributes can be computed using a distance measurement metric in the ternary representation structure.

Another important perspective of SDD is as a bump hunting technique (McConnell and Skillicorn 2002). Let us consider the original data matrix A as a grid of entries. Each positive entry of A is considered as a bump/tower at that position in the grid, with a height proportional to the values of the entry. Similarly each negative entry of A is treated as negative bump/hole with the depth proportional to the value of the entry. SDD searches for the regions of similar height and depth. One particular component of the decomposition is identified, once such a region is found. The average height or depth of the region is computed and subtracted from all the bumps and holes involved. Then the process continues for searching such similar regions and identifying the components of the decomposition. At each iteration, the position of the region is identified using the product of x_i and Y_k^T, and the height of the bump is identified using d_i If the original matrix A is represented as sum of a set of A_i^s, then each A_i,- represents a bump. The bumps are discovered based on their volume. Since the SDD selects the bump/hole based on the height and region, it is not scale independent. However, SVD is a scale-independent technique since the scaling process does not change the decomposition result. Scaling the magnitudes by squaring, SDD first selects the smaller regions of large magnitudes. Similarly if the magnitudes are replaced by their signed square roots, then SDD first selects the larger regions of smaller magnitude. There are other bump hunting techniques like PRIM and rule-based techniques. Methods based on SVD are available that result in decision tree classification like Principle Direction Division Partitioning (PDDP) (Skillicorn 2007).

Key Applications

SDD has found several applications in the literature. SDD for outlier detection was used by McConnell and Skillicorn (2002). Based on this application, SDD is further used for counter terrorism, social network analysis, detecting deceptive communications in the e-mails, etc. (Divya et al. 2011; Keila and Skillicorn 2005; Knight and Carosielli 2003; Skillicorn 2004; Snasel et al. 2010).

In collaborative filtering applications and recommender systems, SDD can be applied to identify the groups of objects that are rated highly by the individuals (Skillicorn 2007). SDD is successfully applied for image and video compression. Pattern matches and motion vectors in video coding can be computed using SDD (Zyto et al. 2002). For compressing the large images, truncated SDD of the image matrix can be considered as approximation to the original image.

With its features, SDD is well suited for hierarchical clustering and decision tree classification problems (Skillicorn 2007). For information retrieval and text mining applications, SDD is used as an alternate method for SVD (Kolda and O'Leary 1998). In addition to the IR applications, LSI technique can be augmented with SDD in automated text categorization application (Pilato et al. 2005). SDD can also be applied for obtaining the subsymbolic representation of words (Qiang et al. 2004), topic identification (Snasel et al. 2008). In social network and link analysis, the matrix decompositions can be useful to derive the higher-order information about the relationships among the individuals in the network. Based on the relationship, the members in the network can be ranked (Skillicorn 2007).

Based on the application, SDD can be directly applied on the data matrix or on the correlation matrix of the original data matrix, or on the approximated correlation matrix. SVD and SDD can be combined so as to complement each strength. On a dataset by applying SVD, we can visualize latent clusters within the data. But SVD cannot provide a way to label these clusters.

On the other hand, SDD provides the clusters within the data and label them. However, SDD cannot produce the visualization of these clusters. While performing this combination, first SVD can be computed on the data matrix A and decompose it to an appropriate rank A matrix A_k matrix using SDD. The advantage of this computation is that the SVD performs denoising of the data, and SDD identifies and labels the clusters within the data (Kumar and Palanisamy 2010). Also this computation can effectively be applied for classification of protein sequences and exploration of minerals, galaxies, etc. (Skillicorn 2007). Also SDD can be applied on the correlation matrix obtained from truncated SVD matrix A_k. In this case, the SDD is used to find the correlation structure within the denoised data.

Illustrative Examples

From the above discussion, we can understand that the SDD can be used for bump selection, hierarchical clustering, LSI-based information retrieval, etc. In the following we see some of the examples illustrating these applications. To better understand SDD as a bump hunting, let us consider the following matrix:

$$A = \begin{bmatrix} 1 & 1 & 5 & 5 & 5 \\ 1 & 9 & 9 & 1 & 1 \\ 1 & 9 & 9 & 1 & 1 \\ 1 & 9 & 9 & 1 & 1 \\ 1 & 1 & 5 & 5 & 5 \end{bmatrix}$$

The SDD on this data matrix produces the following factorization matrices:

$$X_k = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}, \ D_k = \begin{bmatrix} 9 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and $Y_k^T = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{bmatrix}$.

Now consider the first outer product i = 1, then $X_i^* Y_i^{\mathrm{T}}$ is

$$\begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} * [01100] = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

We can understand that the resultant matrix is a stencil, representing the region of the array elements having the value 9 (which is the value d_1). Similarly for i = 2, the second outer product produces $Xi^* Y_i^{\mathrm{T}}$ as

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} * [00111] = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

It is clear that the second outer product result is a stencil representing the region of the array elements having the value 5 (which is the value d_2). Similarly we can obtain the other stencil regions for the values of d using corresponding outer products.

SDD derives a hierarchical clustering of the objects producing ternary tree structure. To illustrate this process, let us consider a well-explained example from Skillicorn et al. (2003). The following is the data matrix of size 9 × 8.

$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 3 & 4 & 4 & 5 & 5 & 6 & 7 & 9 \\ 1 & 8 & 2 & 7 & 3 & 6 & 4 & 5 \\ 9 & 8 & 7 & 6 & 5 & 4 & 3 & 2 \\ 9 & 4 & 8 & 3 & 7 & 2 & 6 & 1 \\ 2 & 3 & 2 & 4 & 2 & 5 & 2 & 6 \\ 3 & 4 & 3 & 4 & 4 & 3 & 4 & 3 \\ 3 & 2 & 4 & 3 & 2 & 4 & 3 & 2 \\ 5 & 5 & 4 & 4 & 6 & 6 & 2 & 2 \end{bmatrix}$$

We apply SDD on this data matrix for rank k = 8. After rearrangement of the bumps as discussed in the above section, the following are the X_k, Y_k^{T} and $\mathrm{D_k}$ matrices of sizes 9 × 8;8 × 8 and 8 × 8, respectively.

$$X_k = \begin{bmatrix} -1 & -1 & 1 & -1 & 0 & 1 & 0 & 1 \\ -1 & 0 & 1 & -1 & 0 & 1 & 1 & 1 \\ -1 & 1 & 0 & 1 & 0 & -1 & 0 & 0 \\ 1 & 1 & -1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & -1 & 0 & 1 & 1 & -1 \\ -1 & -1 & -1 & -1 & 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & -1 & 0 & -1 & 0 & -1 \\ 0 & -1 & -1 & -1 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 1 & -1 & 1 & -1 & -1 \end{bmatrix} \text{ and } Y_k^{\mathrm{T}} = \begin{bmatrix} 1 & 1 & 0 & -1 & 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ -1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 & -1 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 & 1 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$D_k = \begin{bmatrix} 4.6134 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2.685 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.905 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.834 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.5527 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.271 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1.1556 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.1239 \end{bmatrix}$$

The hierarchical clustering structure can be obtained from the data objects present in the matrix $\mathrm{X_k}$ as shown in Fig. 1.

Starting from the first column, the objects are grouped or clustered based on their entries −1, 0, and 1. Members of each group are further divided into subgroups based on their entries in subsequent columns. For example, in the tree structure shown in Fig. 1, we can understand that based on the entries from the first column, the objects {1,2,3, 6} are grouped under label −1; the objects {7,8,9} are grouped under the label 0, and the objects {4,5} are grouped under the label +1. In the next level, each of these groups is subdivided into three groups based on the entries in their second column. Similarly we can obtain hierarchical clustering on the attributes of the data represented in the matrix Y_k^{T}, by following the procedure illustrated above.

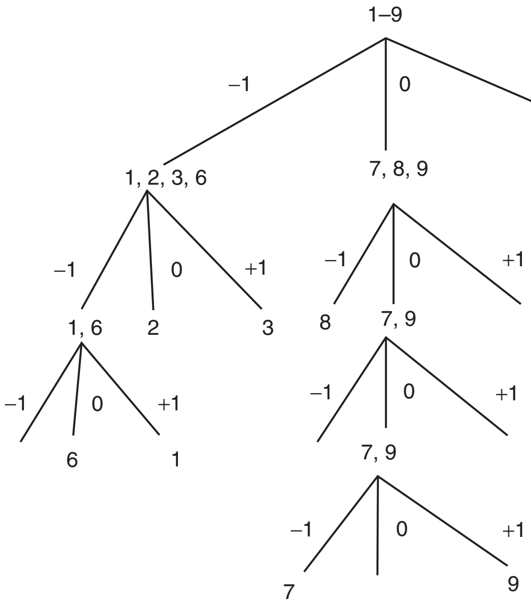
In information retrieval applications, SDD can be used by augmenting with LSI model (Kolda and O'Leary 1998). Let us consider A is a term-document matrix of size m × n, having m terms and n documents with rank r. Let q be the query vector of length m used to probe on the document collection. Column normalization will be performed on the term-document matrix. After applying SDD on the data matrix A, we obtain the factor matrices $\mathrm{X_k}$, $\mathrm{D_k}$, and Y_k^{T} having the sizes m × k; k × k, and k × n, respectively, as discussed above. We apply the query on this reduced dimensional space to compute the similarity of the documents. However before processing, the query vector should be projected onto the lowerdimensional space obtained by SDD as:

$$q_k = qX_kD_k$$

Now the similarity between the document and query vectors in the reduced dimensional space is calculated as

$$\mathrm{sim} = q_kY_k^{\mathrm{T}}$$

Based on the similarity documents can be ranked and returned to the user. Consider a term-document matrix of size 9 × 7 having 9 index



Semi-Discrete Decomposition, Fig. 1 Hierarchical clustering structure obtained from data objects

terms (T) and 7 documents (D) (Berry et al. 1999). The following are the terms and documents:

- T1: Bab(y,ies's), T2: Child(ren's), T3: Guide, T4: Health, T5: Home, T6: Infant, T7: Proofing, T8: Safety, and T9: Toddler
- D1: Infant & Toddler First Aid, D2: Babies & Children's Room, D3: Child Safety at Home,
- D4: Your Baby's Health & Safety: From Infant Babies Collector's Guide. Now the termto Toddler,
- D5: Baby Proofing Basics,
- D6: Your document for this collection is Guide to Easy Rust Proofing, and
- D7: Beanie Babies Collector's Guide. Now the termdocument for this collection is

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Now by making the unit columns, the normalized term-document matrix is

$$A = \begin{bmatrix} 0 & 0.5774 & 0 & 0.4472 & 0.7071 & 0 & 0.7071 \\ 0 & 0.5774 & 0.5774 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.7071 & 0.7071 \\ 0 & 0 & 0 & 0.4472 & 0 & 0 & 0 \\ 0 & 0.5774 & 0.5774 & 0 & 0 & 0 & 0 \\ 0.7071 & 0 & 0 & 0.4472 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.7071 & 0.7071 & 0 \\ 0 & 0 & 0.5774 & 0.4472 & 0 & 0 & 0 \\ 0.7071 & 0 & 0 & 0.4472 & 0 & 0 & 0 \end{bmatrix}$$

By applying SDD on the term-document matrix to approximate to rank k = 4, we obtain the following X_4, Y_4^{T}, and D$_4$ matrices:

$$X_4 = \begin{bmatrix} 1 & 1 & 1 & -1 \\ 1 & 0 & -1 & 0 \\ 1 & -1 & 1 & 1 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 0 & 1 & 0 & 1 \end{bmatrix} Y_4^{\mathrm{T}} = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} D_4 = \begin{bmatrix} 0.2389 & 0 & 0 & 0 \\ 0 & 0.2412 & 0 & 0 \\ 0 & 0 & 0.2289 & 0 \\ 0 & 0 & 0 & 0.3054 \end{bmatrix}$$

The following is the rank k (k = 4) approximation matrix of A obtained using SDD:

$$A_4 = \begin{bmatrix} -0.0642 & 0.2389 & 0.0101 & 0.4801 & 0.4678 & 0.2266 & 0.4678 \\ 0 & 0.2389 & 0.4678 & 0.2389 & 0.0101 & 0.0101 & 0.0101 \\ 0.0642 & 0.2389 & 0.0101 & -0.0022 & 0.4678 & 0.7089 & 0.4678 \\ -0.0642 & 0 & 0 & 0.2412 & 0 & -0.2412 & 0 \\ 0 & 0.2389 & 0.4678 & 0.2389 & 0.0101 & 0.0101 & 0.0101 \\ 0.5465 & 0 & 0 & 0.2412 & 0 & -0.2412 & 0 \\ 0.0624 & 0.2389 & 0.0101 & -0.0022 & 0.4678 & 0.7089 & 0.4678 \\ -0.0642 & 0.2389 & 0.4678 & 0.4801 & 0.0101 & -0.2311 & 0.0101 \\ 0.5465 & 0 & 0 & 0.2412 & 0 & -0.2412 & 0 \end{bmatrix}$$

We note from this approximated termdocument matrix that the elements have negative values. These values represent the linear combination of elements of original termdocument matrix (Berry and Browne 2005). However, the individual term component of document vectors does not define the semantic content. The approximation automatically provides an association with relevant terms in each document. For example, along with the original terms T6 and T9 in the approximation space, the document D1 is associated with T3 and T7 also. Consider that the user wants to find the books on Child Home Safety from the document collection listed above. The corresponding query vector constituted from these terms is

$$q = [0 \;\; 1 \;\; 0 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \;\; 1 \;\; 0]$$

Now before we process the query, we project it on the reduced dimensional space and obtain the following representation:

$$q_4 = [0.7168 \quad 0.2412 \quad -0.6866 \quad -0.3054]$$

Now we process the reduced dimensional query in the approximated space and obtain the similarity as

$$\mathrm{sim} = \;\; [-0.0642 \quad 0.7168 \quad 1.4033 \quad 0.9579 \quad 0.0302 \quad -0.2109 \quad 0.0302]$$

Generally the documents whose similarity values higher than some threshold value are considered as relevant to the given query. Considering a threshold limit of 0.5, we understand from the document similarity vector that the documents D2, D3, and D4 are relevant to the query and hence returned to the user.

In order to better understand the relation between SDD and SVD, discussed above, we verify this retrieval process using SVD. The SVD on the data matrix A produces rank k approximation as shown below:

$$A_k = U_k S_k V_k^T$$

where the unitary projection matrices U_k and V_4^T represent truncated left and right singular vectors of the original matrix, respectively. The matrix S_k holds the first k number of singular values of the original matrix. SVD provides the best approximation of the original matrix with regard to Frobenius norm. Generally SVD is regarded as one of the powerful decomposition technique since it provides all the fundamental spaces of the original matrix A, i.e., the orthogonal basis for Range space and Null space of the matrices A and U^T (Park and Elden 2003).

By applying SVD, we obtain rank 4 approximation of the column normalized term-document matrix. After processing the query in the reduced dimensional space, we obtain the document similarity vector as shown below:

$$\text{sim} = [0.0705 \quad 1.2360 \quad 1.6855 \quad 0.3747 - 0.0117 \quad 0.0128 \quad - 0.0117]$$

From this vector documents D2 and D3 are relevant to the query, and hence they will be returned to the user. From this result we can understand that the SDD and SVD have commonly identified the documents D2 and D3. Now let us consider another query aimed at retrieving the documents on Child Proofing from this collection. The corresponding query vector would be

$$q = [0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0]$$

After projecting the query in the reduced dimensional space obtained using SDD, we compute the similarity of the document vectors. The following is the similarity vector of all the documents for the given query:

$$\text{sim} = [0.0642 \ \ 0.4779 \ \ 0.4779 \ \ 0.2367 \ \ 0.47790.7190 \ \ 0.4779]$$

With the chosen threshold limit of 0.5, we understand that the document D6 is the only relevant document for this query. However, let us consider posing this query in the reduced dimensional space obtained using SVD. After processing, we get the following similarity:

$$\text{sim} = [-0.0721 \ \ 0.4872 \ \ 0.6307 \ \ 0.07300.3690 \ \ 0.6903 \ \ 0.3690]$$

With the threshold value of 0.5, we get the documents D3 and D6 as relevant to the given query. In this case, the SDD and SVD have commonality in D6. For equal rank values of approximation, SDD requires significantly less number of floating point operations than SVD to process the query. On standard document collections, it is proved that SDD-based LSI retrieves documents similar to SVD-based LSI with a lesser time to process the query and lesser storage. These illustrative examples provide an understanding about the usage of SDD in some of the applications. Interested readers can explore the literature cited in section "Key Applications" for more details on the applications of SDD.

Cross-References

- ► Eigenvalues: Singular Value Decomposition
- ► Matrix Algebra, Basics of
- ► Matrix Decomposition
- ► Principal Component Analysis
- ► Spectral Analysis

Acknowledgments Author sincerely acknowledges the support from National Board of Higher Mathematics, Dept. of Atomic Energy, Govt. of India under the grant 2/48(11)/2010-R&D II/10806.

References

Berry MW, Browne M (2005) Understanding search engines: mathematical modeling and text retrieval. SIAM, Philadelphia

Berry MW, Drmac Z, Jessup ER (1999) Matrices, vector spaces and information retrieval. SIAM Rev 41(2):335–362

Cunningham P (2007) Dimension reduction. Technical report UCD-CSI-2007-7, School of computer science and informatics, University College, Dublin

Deerwester S, Dumais S, Furnas G, Landauer T, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):391–407

Divya R, Kumar CA, Saijanani S, Priyadharshini M (2011) Deceiving communication links on an organization email corpus. Malays J Comput Sci 24(1):17–33

Dobsa J, Praus P, Kumar CA, Praks P (2012) Classification of hydrochemical data in reduced dimensional space. J Inf Org Sci 36(1):27–37

Elden L (2006) Numerical linear algebra in data mining. Acta Numer 15:327–384

Elden L (2007) Matrix methods in data mining and pattern recognition. SIAM, Philadelphia

Fodor IK (2002) A survey of dimension reduction techniques. Technical report UCRL-148494, University of California

Keila PS, Skillicorn DB (2005) Structure in the Enron email dataset. J Comput Math Organ Theory 11(3):183–199

Knight GS, Carosielli L (2003) Detecting malicious use with unlabelled data using clustering and outlier analysis. In: Proceedings of 18th IFIP international conference on information security, Athens, pp 205–216

Kolda TG, O'Leary DP (1998) A semi-discrete matrix decomposition for latent semantic indexing in information retrieval. ACM Trans Inf Syst 16(4):322–346

Kolda TG, O'Leary DP (2000) Computation and uses of the semidiscrete matrix decomposition. ACM Trans Math Softw 26(3):415–435

Kumar CA (2009) Analysis of unsupervised dimensionality reduction techniques. Comput Sci Inf Syst 6(2):217–227

Kumar CA (2011) Reducing data dimensionality using random projections and fuzzy k-means clustering. Int J Intell Comput Cybern 4(3):353–365

Kumar CA, Palanisamy R (2010) Comparison of matrix dimensionality reduction methods in uncovering latent structures in the data. J Inf Knowl Manag 9(1):81–92

Kumar CA, Srinivas S (2006) Latent semantic indexing using eigenvalue analysis for efficient information retrieval. Int J Appl Math Comput Sci 16(4):551–558

Kumar CA, Srinivas S (2010) A note on weighted fuzzy k-means clustering for concept decomposition. Cybern Syst 41:455–467

McConnell S, Skillicorn DB (2002) Semidiscrete decomposition: a bump hunting technique. In: Proceedings of the Australian data mining workshop, Canberra, pp 75–82

Miettinen P (2009) Matrix decomposition methods for data mining: computational complexity and algorithms. Academic dissertation A-2009-04, Department of Computer Science, University of Helsinki, Finland

O'Leary D, Peleg S (1983) Digital image compression by outer product expansion. IEEE Trans Commun 31:441–444

Park H, Elden L (2003) Matrix rank reduction for data analysis and feature extraction. Technical report TR 03–015, University of Minnesota

Pilato G, Vassallo G, Gaglio S (2005) Wordnet and semi-discrete decomposition for sub-symbolic representation of words. In: Apolloni B et al (eds) Biological and artificial intelligence environments. Springer, Dordrecht, pp 191–198

Qiang W, XiaoLong L, Yi G (2004) A study of semi-discrete matrix decomposition for LSI in automated text categorization. In: Proceedings of 1st International joint conference on natural language processing. Springer, Berlin/Heidelberg, pp 606–615

Skillicorn DB (2004) Applying matrix decompositions to counterterrorism. Technical report, Queen's University, Kingston

Skillicorn DB (2007) Understanding complex datasets data mining with matrix decompositions. Chapman & Hall/CRC, New York

Skillicorn DB, McConnell SM, Soong EY (2003) Handbook of data mining using matrix decompositions. School of Computing, Queen's University, Kingston

Snasel V, Moravec P, Pokorny J (2008) Using semi-discrete decomposition for topic identification. In: 8th

International conference on intelligent systems design and applications, Kaohsiung, pp 415–420

Snasel V, Horak Z, Abraham A (2010) Link suggestions in terrorist networks using semi discrete decomposition. In: 6th International conference on information assurance and security. Atlanta, pp 23–25

Zyto S, Grama A, Szpankowski W (2002) Semi-discrete matrix transforms for image and video compression. In: Proceedings of the data compression conference. Snowbird, Utah, p 484

Semiring

▶ Semirings and Matrix Analysis of Networks

Semirings and Matrix Analysis of Networks

Monika Cerinšek[1] and Vladimir Batagelj[2,3,4]
[1]Abelium d.o.o, Research and Development, Ljubljana, Slovenia
[2]Faculty of Mathematics and Physics, Department of Mathematics, University of Ljubljana, Ljubljana, Slovenia
[3]Department of Theoretical Computer Science, Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia
[4]University of Primorska, Andrej Marušič Institute, Koper, Slovenia

Synonyms

Algebraic path problem; Matrix; Multiplication of vector and matrix; Network; Network multiplication; Semiring; Simple walk; Value matrix; Walk

Glossary

Algebraic structure	A set with one or more operations defined on it and rules that hold for them
Network analysis	A study of networks as representations of relations between discrete objects
Sparse matrix	A matrix with most of entries equal to zero
Large network	A network with several thousands or millions of nodes
Complete graph	K_n – A graph in which every pair of nodes is linked

Definition

A network can be represented also with a corresponding matrix. Using matrix operations (addition and multiplication) over an appropriate semiring a unified approach to several network analysis problems can be developed. Matrix multiplication is about traveling on network.

Introduction

Semirings are algebraic structures with two operations that provide the basic conditions for studying matrix addition and multiplication and path problems in networks. Several results and algorithms from different fields of application turn out to be special cases over the corresponding semirings.

Semirings

Let $\mathbb{K}$ be a set and a, b, c elements from $\mathbb{K}$. A semiring (Abdali and Saunders 1985; Baras and Theodorakopoulos 2010; Batagelj 1994; Carré 1979) is an algebraic structure $(\mathbb{K}, \oplus, \odot, 0, 1)$ with two binary operations (addition $\oplus$ and multiplication $\odot$) where:

- $(\mathbb{K}, \oplus, 0)$ is an abelian monoid with the neutral element 0 (zero):

$$\begin{array}{ll} a \oplus b = b \oplus a & \text{commutativity} \\ (a \oplus b) \oplus c = a \oplus (b \oplus c) & \text{associativity} \\ a \oplus 0 = a & \text{existence of zero} \end{array}$$

- $(\mathbb{K}, \odot, 1)$ is a monoid with the neutral element 1 (unit):

$$\begin{array}{ll} (a \odot b) \odot c = a \odot (b \odot c) & \text{associativity} \\ a \odot 1 = 1 \odot a = a & \text{existence of a unit} \end{array}$$

- Multiplication $\odot$ distributes over addition $\oplus$:

S

$$a \odot (b \oplus c) = a \odot b \oplus a \odot c$$
$$(b \oplus c) \odot a = b \odot a \oplus c \odot a$$

In formulas we assume precedence of multiplication over addition.

A semiring $(\mathbb{K}, \oplus, \odot, 0, 1)$ is *complete* if the addition is well defined for countable sets of elements and the commutativity, associativity, and distributivity hold in the case of countable sets. These properties are generalized in this case; for example, the distributivity takes the form

$$\left(\bigoplus_i a_i\right) \odot \left(\bigoplus_j b_j\right) = \bigoplus_i \left(\bigoplus_j (a_i \odot b_j)\right)$$
$$= \bigoplus_{i,j} (a_i \odot b_j).$$

The addition is *idempotent* if $a \oplus a = a$ for all $a \in \mathbb{K}$. In this case the semiring over a finite set $\mathbb{K}$ is complete.

A semiring $(\mathbb{K}, \oplus, \odot, 0, 1)$ is *closed* if for the additional (unary) *closure* operation * it holds for all $a \in \mathbb{K}$:

$$a^* = 1 \oplus a \odot a^* = 1 \oplus a^* \odot a.$$

The *power* a^n, $n \in \mathbb{N}$ of an element $a \in \mathbb{K}$ is defined by $a^0 = 1$ and $a^{n+1} = a^n \odot a$ for $n \geq 0$.

Different closures over the same semiring can exist. A complete semiring is always closed for the closure

$$a^* = \bigoplus_{k \in \mathbb{N}} a^k.$$

In a closed semiring we can also define a *strict closure* $\overline{a}$ as

$$\overline{a} = a \odot a^*.$$

In a semiring $(\mathbb{K}, \oplus, \odot, 0, 1)$ the *absorption law* holds if for all $a, b, c \in \mathbb{K}$:

$$a \odot b \oplus a \odot c \odot b = a \odot b.$$

Because of the distributivity, it is sufficient for the absorption law to check the property $1 \oplus c = 1$ for all $c \in \mathbb{K}$.

Combinatorial Semiring $(\mathbb{N}, +, \cdot, 0, 1)$

This is the most commonly used semiring. Also some other sets are used: $\mathbb{R}$, $\mathbb{R}_0^+$, $\mathbb{Q}$. For $\overline{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$, the semiring is closed for $a^* = \sum_{k \in \overline{\mathbb{N}}} a^k$ because it is a complete semiring. An example of a closure for $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ is $a^* = 1/(1 - a)$ for $a \neq 1, \infty$ and $0^* = 1$, $1^* = \infty$, and $\infty^* = \infty$. This semiring is commutative because it holds $a \odot b = b \odot a$ for all a and b in the set. Combinatorial semiring is not an idempotent semiring.

Reachability Semiring $(\{0, 1\}, \vee, \wedge, 0, 1)$

The logical (boolean, reachability) semiring is suitable for solving the connectivity questions in networks. The multiplication is commutative and the absorption law holds. The reachability semiring is closed for $a^* = 1 \vee a \wedge a^* = 1$.

Shortest Paths Semiring $(\overline{\mathbb{R}}_0^+, \min, +, \infty, 0)$

The commutativity of multiplication holds in this semiring. The semiring is closed for $a^* = \min(0, a + a^*) = 0$ (0 is the smallest element in the set $\mathbb{R}_0^+$). Since $\min(0, a) = 0$, the absorption law also holds. For the set $\overline{\mathbb{N}}$, the semiring is called a tropical semiring. Another set is $\overline{\mathbb{R}}$ and in this case the semiring is isomorphic ($x \to -x$) to max-plus semiring $(\mathbb{R} \cup \{-\infty\}, \max, +, -\infty, 0)$.

Pathfinder Semiring $(\overline{\mathbb{R}}_0^+, \min, \boxed{r}, \infty, 0)$

The Pathfinder semiring (Schvaneveldt et al. 1988) is a special case from the family of semirings obtained as follows. Let $B \subseteq \overline{\mathbb{R}}$ be such that $(B, +, \cdot, 0, 1)$ or $(B, \min, +, U, 0)$ is a semiring ($U = \max(B)$). Therefore $0 \in B$ and $1 \in B$. Let $A \subseteq \mathbb{R}$ be such that $g : A \to B$ is a bijection. Let us define operations $\oplus$, ∇, $\odot$ so that g is a homomorphism:

$$\begin{aligned} g(a \oplus b) &= g(a) + g(b), \\ g(a \nabla b) &= \min(g(a),\ \ g(b)), \\ g(a \odot b) &= g(a) \cdot g(b). \end{aligned}$$

This is equivalent to

$$\begin{aligned} a \oplus b &= g^{-1}(g(a) + g(b)), \\ a \nabla b &= g^{-1}(\min(g(a),\ \ g(b))), \\ a \odot b &= g^{-1}(g(a) \cdot g(b)). \end{aligned}$$

The function g^{-1} is also a homomorphism. If g is strictly increasing function, then

$$a \nabla b = g^{-1}(\min(g(a),\ \ g(b))) = \min(a, b).$$

Since the homomorphisms preserve the algebraic properties, also the structures

$$(A,\ \oplus\ ,\ \odot\ , g^{-1}(0), g^{-1}(1))$$

and

$$(A, \nabla,\ \oplus\ , g^{-1}(U), g^{-1}(0))$$

are semirings.

For g(:r) $= x^r$, $g^{-1}(y) = \sqrt[r]{y}$, we get the *Pathfinder semiring* ($\overline{\mathbb{R}}_0^+$, min, $\boxed{r}$, ∞, 0). The multiplicative operation is the *Minkowski operation* $a \boxed{r} b = \sqrt[r]{a^r + b^r}$. This semiring is closed for $a^* = 0$ and the absorption law holds in it.

In Pathfinder algorithm the value r for the Minkowski operation is selected according to a dissimilarity measure. For a value $r = 1$, the semiring is the shortest path semiring, and for a value $r = \infty$, the semiring is the min-max semiring.

More about semirings and several other examples can be found in (Baras and Theodorakopoulos 2010; Batagelj and Praprotnik 2016; Burkard et al. 1984; Carré 1979; Glazek 2002; Golan 1999; Gondran and Minoux 2008; Kepner and Gilbert 2011).

Matrices

An $m \times n$ matrix $\mathbf{A}$ over a set $\mathbb{K}$ is a rectangular array of elements from the set $\mathbb{K}$ that consists of m rows and n columns. The entry in the *i-th* row and j-th column is denoted by a_{ij}. If $m = n$ the matrix $\mathbf{A}$ is called a *square* matrix. The matrix with all entry values equal to 0 is called the *zero* matrix and is denoted by $\mathbf{0}_{mn}$.

The *transpose* of a matrix $\mathbf{A}$ is a matrix $\mathbf{A}^T$ in which the rows of $\mathbf{A}$ are written as the columns of $\mathbf{A}^T$: $a_{ij}^T = a_{ji}$. A square matrix $\mathbf{A}$ is *symmetric* if $\mathbf{A} = \mathbf{A}^T$.

A *diagonal matrix* is a square matrix $\mathbf{A}$ such that only diagonal elements are nonzero: $a_{ij} = 0$, for $i \neq j$. If $a_{ii} = 1$, $i = 1, \ldots, n$, a diagonal matrix is called the *identity* matrix I_n of order n. A square matrix $\mathbf{A}$ is *upper triangular* if $a_{ij} = 0,\ i > j$, and its transpose is a *lower triangular* matrix.

Let $\mathcal{M}_{mn}(\mathbb{K})$ be a set of matrices of order $m \times n$ over the semiring ($\mathbb{K}$, $\oplus$, $\odot$, 0, 1) in which we additionally require

$$\forall a \in \mathbb{K} : a \odot 0 = 0 \odot a = 0,$$

and let $\mathcal{M}(\mathbb{K})$ be a set of all matrices over the $\mathbb{K}$. The operations $\oplus$ and $\odot$ can be extended to the $\mathcal{M}(\mathbb{K})$:

$$\begin{aligned} &\mathbf{A}, \mathbf{B} \in \mathcal{M}_{mn}(\mathbb{K}) : \mathbf{A} \oplus \mathbf{B} = [a_{uv} \oplus b_{uv}] \in \mathcal{M}_{mn}(\mathbb{K}) \\ &\mathbf{A} \in \mathcal{M}_{mk}(\mathbb{K}), \\ &\mathbf{B} \in \mathcal{M}_{kn}(\mathbb{K}) : \mathbf{A} \odot \mathbf{B} = \left[\oplus_{t=1}^{k} a_{ut} \odot b_{tv}\right] \in \mathcal{M}_{mn}(\mathbb{K}). \end{aligned}$$

Then:

- $(\mathcal{M}_{mn}(\mathbb{K}),\ \oplus\ , \mathbf{0}_{mn})$ is an abelian monoid.
- $(\mathcal{M}_{n^2}(\mathbb{K}),\ \odot,\ \mathbf{I}_n)$ is a monoid.
- $(\mathcal{M}_{n^2}(\mathbb{K}),\ \oplus,\ \odot,\ 0_n,\ \mathbf{I}_n)$ is a semiring.

For matrices $\mathbf{A}$ and $\mathbf{B}$, it holds

$$(\mathbf{A} \odot \mathbf{B})^T = \mathbf{B}^T \odot \mathbf{A}^T.$$

Network Multiplication

A (simple directed) network $\mathcal{N}$ is an ordered pair of sets ($\mathcal{V}$, $\mathcal{A}$) where $\mathcal{V}$ is the set of nodes and $\mathcal{A} \subseteq\ \mathcal{V} \times\ \mathcal{V}$ is the set of arcs (directed links). We assume that the set of nodes is finite $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$. Let $\mathcal{N} = ((\mathcal{I}, \mathcal{J}), \mathcal{A}, w)$ be a *simple two-mode network*, where $\mathcal{I}$ and $\mathcal{J}$ are disjoint (sub)sets of nodes ($\mathcal{V} = \mathcal{I} \cup\ \mathcal{J}$, $\mathcal{I} \cap$

$\mathcal{J} = \emptyset$), $\mathcal{A}$ is a set of arcs linking $\mathcal{I}$ and $\mathcal{J}$, and the mapping w: $\mathcal{A} \rightarrow \mathbb{K}$ is the *arcs value function* also called a *weight*. We can assign to a network its *value matrix* $\mathbf{W} = [w_{ij}]$ with elements

$$w_{ij} = \begin{cases} w((i,j)) & (i,j) \in \mathcal{A} \\ 0 & \text{otherwise.} \end{cases}$$

The problem with value matrices in computer applications is their size. The value matrices of large networks are sparse. There is no need to store the zero values in a matrix, and different data structures can be used for saving and working with value matrices: special dictionaries and lists.

Let $\mathcal{N}_\mathbf{A} = ((\mathcal{I}, \mathcal{K}), \mathcal{A}_\mathbf{A}, w_\mathbf{A})$ and $\mathcal{N}_\mathbf{B} = ((\mathcal{K}, \mathcal{J}), \mathcal{A}_\mathbf{B}, w_\mathbf{B})$ be a pair of networks with corresponding matrices $\mathbf{A}_{\mathcal{I}\times\mathcal{K}}$ and $\mathbf{B}_{\mathcal{K}\times\mathcal{J}}$, respectively. Assume also that w_A: $\boldsymbol{\mathcal{A}}_\mathbf{A} \rightarrow \mathbb{K}$, $w_\mathbf{B}$: $\boldsymbol{\mathcal{A}}_\mathbf{B} \rightarrow \mathbb{K}$, and ($\mathbb{K}$, $\oplus$, $\odot$, 0, 1) is a semiring. We say that such networks/matrices are *compatible.* The *product* $\boldsymbol{\mathcal{N}}_\mathbf{A} * \boldsymbol{\mathcal{N}}_\mathbf{B}$ of networks $\boldsymbol{\mathcal{N}}_\mathbf{A}$ and $\boldsymbol{\mathcal{N}}_\mathbf{B}$ is a network $\boldsymbol{\mathcal{N}}_\mathbf{C} = ((\mathcal{I}, \mathcal{J}), \boldsymbol{\mathcal{A}}_\mathbf{C}, w_\mathbf{C})$ for $\boldsymbol{\mathcal{A}}_\mathbf{C} = \{(i, j); i \in \mathcal{I}, j \in \mathcal{J}, c_{ij} \neq 0\}$ and $w_\mathbf{C}((i, j)) = c_{ij}$ for $(i, j) \in \boldsymbol{\mathcal{A}}_\mathbf{C}$, where $\mathbf{C} = [c_{ij}] = \mathbf{A} \odot \mathbf{B}$. If all three sets of nodes are the same ($\mathcal{I} = \mathcal{K} = \mathcal{J}$), we are dealing with ordinary one-mode networks (square matrices).

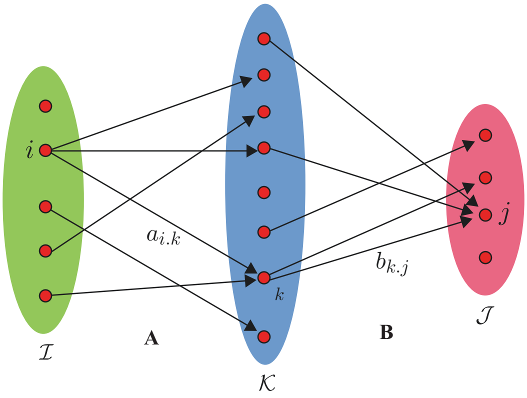
When do we get an arc in the product network? Let's look at the definition of the matrix product

$$c_{ij} = \bigoplus_{k \in \mathcal{K}} a_{ik} \odot b_{kj}.$$

There is an arc $(i, j) \in \boldsymbol{\mathcal{A}}_\mathbf{C}$ if c_{ij} is nonzero. Therefore at least one term $a_{ik} \odot b_{kj}$ is nonzero, but this means that both a_{ik} and b_{kj} should be nonzero, and thus $(i,k) \in \boldsymbol{\mathcal{A}}_\mathbf{A}$ and $(k,j) \in \mathcal{A}_\mathbf{B}$ (see Fig. 1):

$$c_{ij} = \bigoplus_{k \in N_\mathbf{A}(i) \cap N_\mathbf{B}^-(j)} a_{ik} \odot b_{kj},$$

where $N_\mathbf{A}(i)$ are the *successors* of node i in the network $\boldsymbol{\mathcal{N}}_\mathbf{A}$ and $N_\mathbf{B}^-(j)$ are the *predecessors* of node j in the network $\boldsymbol{\mathcal{N}}_\mathbf{B}$. The value of the entry c_{ij} equals to the value of all paths (of length 2) from $i \in \mathcal{I}$ to $j \in \mathcal{J}$ passing through some node $k \in \mathcal{K}$.



Semirings and Matrix Analysis of Networks, Fig. 1 Multiplication of networks

The standard procedure to compute the product of matrices $\mathbf{A}_{\mathcal{I}\times\mathcal{K}}$ and $\mathbf{B}_{\mathcal{K}\times\mathcal{J}}$ has the complexity $O(|\mathcal{I}| \cdot |\mathcal{K}| \cdot |\mathcal{J}|)$ and is therefore too slow to be used for large networks. Since the matrices of large networks are usually sparse, we can compute the product of two networks much faster considering only nonzero entries (Batagelj and Cerinšek 2013; Batagelj and Mrvar 2008):

```
for  $k \in \mathrm{K}$  do
    for  $i \in N_\mathbf{A}^-(k)$  do
        for  $j \in N_\mathbf{B}(k)$  do
            If  $\exists c_{ij}$  then
                $c_{ij} = c_{ij} \oplus a_{ik} \odot b_{kj}$
            else
                $c_{ij} = a_{ik} \odot b_{kj}.$
```

In general the multiplication of large sparse networks is a "dangerous" operation since the result can "explode" – it is not sparse.

From the network multiplication algorithm, we see that each intermediate node $k \in \mathcal{K}$ adds to a product network a complete two-mode subnetwork $K_{N_\mathbf{A}^-(k), N_\mathbf{B}(k)}$ (or, in the case $\mathbf{A} = \mathbf{B}$, a complete subnetwork $\mathrm{K}_{N(k)}$). If both degrees $\deg_\mathbf{A}(k) = \left|N_\mathbf{A}^-(k)\right|$ and $\deg_\mathbf{B}(k) = |N_\mathbf{B}(k)|$ are large, then already the computation of this complete subnetwork has a quadratic (time and space) complexity – the result "explodes."

If for the sparse networks $\boldsymbol{\mathcal{N}}_\mathbf{A}$ and $\boldsymbol{\mathcal{N}}_\mathbf{B}$, there are in $\mathcal{K}$ only few nodes with large degree and no one among them with large degree in both networks, then also the resulting product network $\mathcal{N}_C$ is sparse.

The Algebraic Path Problem

The use of a special semiring and a multiplication of networks can lead us to the essence of the shortest path problem (Baras and Theodorakopoulos 2010). Many other network problems can be solved by replacing the usual addition and multiplication with the corresponding operations from an appropriate semiring.

Let $\mathcal{N} = (\mathcal{V}, \mathcal{A}, w)$ be a network where $w: \mathcal{A} \to \mathbb{K}$ is the value (weight) of arcs such that $(\mathbb{K}, \oplus, \odot, 0, 1)$ is a semiring. We denote the number of nodes as $n = |\mathcal{V}|$ and the number of arcs as $m = |\mathcal{A}|$.

A finite sequence of nodes $\sigma = (u_0, u_1, u_2, \ldots, u_{p-1}, u_p)$ is a *walk* of *length p* on $\mathcal{N}$ if every pair of neighboring nodes is linked: $(u_{i-1}, u_i) \in \mathcal{A}$, $i = 1,\ldots,p$. Finite sequence σ is a *semiwalk* or chain on $\mathcal{N}$ if every pair of neighboring nodes is linked neglecting the direction of an arc: $(u_{i-1}, u_j) \in \mathcal{A} \vee (u_i, u_{i-1}) \in \mathcal{A}, i = 1, \ldots, p$. A (semi)walk is closed if its end nodes coincide: $u_0 = u_p$. A walk is *simple* or a *path* if no node repeats in it. A closed walk with different nodes, except first and last, is called a *cycle.*

We can extend the weight w to walks and sets of walks on $\mathcal{N}$ by the following rules (see Fig. 2):

- Let $\sigma_v = (v)$ be a null walk in the node $v \in \mathcal{V}$; then $w(\sigma_v) = 1$.
- Let $\sigma = (u_0, u_1, u_2,\ldots, u_{p-1}, u_p)$ be a walk of length $p \geq 1$ on $\mathcal{N}$; then

$$w(\sigma) = \bigodot_{i=1}^{k} w(u_{i-1},\ u_i).$$

- For empty set of walks $\varnothing$ it holds $w(\varnothing) = 0$.
- Let $S = \{\sigma_1, \sigma_2, \ldots\}$ be a set of walks in $\mathcal{N}$; then

$$w(\mathcal{S}) = \bigoplus_{\sigma \in \mathcal{S}} w(\sigma).$$

Let σ_1 and σ_2 be compatible walks on $\mathcal{N}$: the end node of the walk σ_1 is equal to the start node of the walk σ_2. Such walks can be concatenated in a new walk $\sigma_1 \circ \sigma_2$ for which holds

$$w(\sigma_1 \circ \sigma_2) = \begin{cases} w(\sigma_1) \odot w(\sigma_2) & \sigma_1 \text{ and } \sigma_2 \text{ are compatible} \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be finite sets of walks; then

$$w(\mathcal{S}_1 \cup \mathcal{S}_2) \oplus w(\mathcal{S}_1 \cap \mathcal{S}_2) = w(\mathcal{S}_1) \oplus w(\mathcal{S}_2).$$

In the special case when $\mathcal{S}_1 \cap \mathcal{S}_2 = \varnothing$, it holds $w(\mathcal{S}_1 \cup \mathcal{S}_2) = w(\mathcal{S}_1) \oplus w(\mathcal{S}_2)$. Also the concatenation of walks can be generalized to sets of walks:

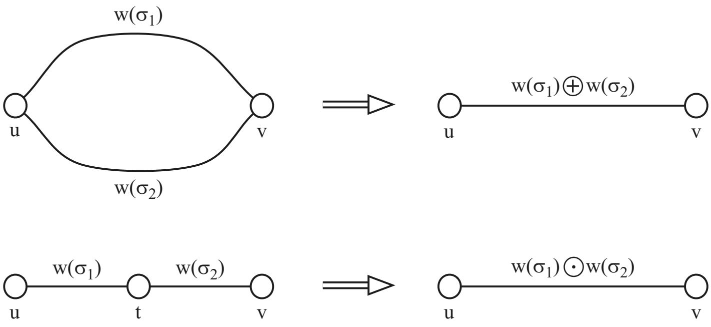
$$\mathcal{S}_1 \circ \mathcal{S}_2 = \{\sigma_1 \circ \sigma_2 : \sigma_1 \in \mathcal{S}_1, \sigma_2 \in \mathcal{S}_2, \sigma_1 \text{ and } \sigma_2 \text{ are compatible}\}.$$

It also holds $\mathcal{S} \circ \varnothing = \varnothing \circ \mathcal{S} = \varnothing$.

We denote by:

- $\mathcal{S}_{uv}^k$ the set of all walks of length k from node u to node v

Semirings and Matrix Analysis of Networks, Fig. 2 Semiring operations and values of walks



- $\boldsymbol{\mathcal{S}}_{uv}^{(k)}$ the set of all walks of length at most k from node u to node v
- $\boldsymbol{\mathcal{S}}_{uv}^{*}$ the set of all walks from node u to node v
- $\overline{\boldsymbol{\mathcal{S}}}_{uv}$ the set of all nontrivial walks from node u to node v
- $\mathcal{E}_{uv}$ the set of all simple walks (paths) from node u to node v

The following relations hold among these sets:

$$\boldsymbol{\mathcal{S}}_{uv}^{k} \subseteq \boldsymbol{\mathcal{S}}_{uv}^{(k)} \subseteq \boldsymbol{\mathcal{S}}_{uv}^{*}$$

$$k \neq l \Leftrightarrow \boldsymbol{\mathcal{S}}_{uv}^{k} \cap \boldsymbol{\mathcal{S}}_{uv}^{l} = \varnothing$$

$$\boldsymbol{\mathcal{S}}_{uv}^{(k)} = \bigcup_{i=0}^{k} \boldsymbol{\mathcal{S}}_{uv}^{i}, \quad \boldsymbol{\mathcal{S}}_{uv}^{*} = \bigcup_{k=0}^{\infty} \boldsymbol{\mathcal{S}}_{uv}^{k}$$

$$k \geq n-1 : \mathcal{E}_{uv} \subseteq \boldsymbol{\mathcal{S}}_{uv}^{(k)}$$

$$w\left(\boldsymbol{\mathcal{S}}_{uv}^{(k)}\right) = \sum_{i=0}^{k} w\left(\boldsymbol{\mathcal{S}}_{uv}^{i}\right).$$

A set of walks $\boldsymbol{\mathcal{S}}$ is *uniquely factorizable* to sets of walks $\boldsymbol{\mathcal{S}}_1$ and $\boldsymbol{\mathcal{S}}_2$ if $\boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{S}}_1 \circ \boldsymbol{\mathcal{S}}_2$, and for all walks $\sigma_1, \sigma_1' \in \boldsymbol{\mathcal{S}}_1$, $\sigma_2, \sigma_2' \in \boldsymbol{\mathcal{S}}_2, \sigma_1 \neq \sigma_1', \sigma_2 \neq \sigma_2'$, it holds $\sigma_1 \circ \sigma_2 \neq \sigma_1' \circ \sigma_2'$.

For example, for s, $0 < s < k$, a nonempty set $\boldsymbol{\mathcal{S}}_{uv}^{k}$ is uniquely factorizable to sets $\boldsymbol{\mathcal{S}}_{u\bullet}^{s}$ and $\boldsymbol{\mathcal{S}}_{\bullet v}^{k-s}$, where $\boldsymbol{\mathcal{S}}_{u\bullet}^{s} = \cup_{t \in \mathcal{V}} \boldsymbol{\mathcal{S}}_{ut}^{s}$, etc.

Theorem 1 *Let the finite set* $\boldsymbol{\mathcal{S}}$ *be uniquely factorizable to* $\boldsymbol{\mathcal{S}}_1$ *and* $\boldsymbol{\mathcal{S}}_2$ *or a semiring is idempotent. Then it holds*

$$w(\boldsymbol{\mathcal{S}}_1 \circ \boldsymbol{\mathcal{S}}_2) = w(\boldsymbol{\mathcal{S}}_1) \odot w(\boldsymbol{\mathcal{S}}_2).$$

The k-th power $\mathbf{W}^k$ of a square matrix $\mathbf{W}$ over $\mathbb{K}$ is unique because of associativity.

Theorem 2 *The entry* w_{uv}^{k} *of k-th power* $\mathbf{W}^k$ *of a value matrix* $\mathbf{W}$ *is equal to the value of all walks of length k from node u to node v:*

$$w\left(\boldsymbol{\mathcal{S}}_{uv}^{k}\right) = \mathbf{W}^{k}[u,v] = w_{uv}^{k}.$$

Therefore if a network $\boldsymbol{\mathcal{N}}$ is acyclic, then it holds for a value matrix $\mathbf{W}$:

$$\exists k_0 : \forall k > k_0 : \mathbf{W}^k = 0,$$

where k_0 is the length of the longest walk in the network.

If $\mathbf{W}$ is the network adjacency matrix over the combinatorial semiring, the entry w_{uv}^{k} counts the number of different walks of length k from u to v.

Let us denote

$$\mathbf{W}^{(k)} = \bigoplus_{i=0}^{k} \mathbf{W}^{i}.$$

In an idempotent semiring, it holds $\mathbf{W}^{(k)} = (1 \oplus \mathbf{W})^k$.

Theorem 3

$$w\left(\boldsymbol{\mathcal{S}}_{uv}^{k}\right) = \mathbf{W}^{(k)}[u,v] = w_{uv}^{(k)}.$$

For the combinatorial semiring and the network adjacency matrix $\mathbf{W}$, the entry $w_{uv}^{(k)}$ counts the number of different walks of length at most k from u to v.

The matrix semiring over a complete semiring is also complete and therefore closed for $\mathbf{W}^* = \oplus_{k=0}^{\infty} \mathbf{W}^k$.

Theorem 4 *For a value matrix* $\mathbf{W}$ *over a complete semiring with closure* $\mathbf{W}^*$ *and strict closure* $\overline{\mathbf{W}}$ *hold:*

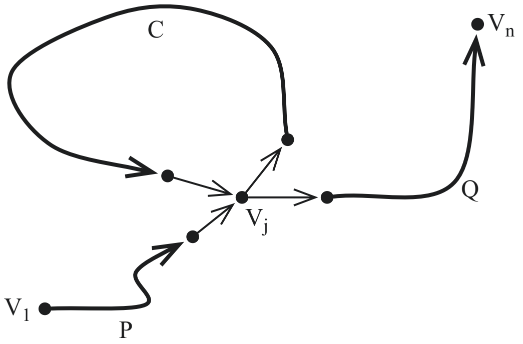
$$w\left(\boldsymbol{\mathcal{S}}_{uv}^{*}\right) = \mathbf{W}^{*}[u,v] = w_{uv}^{*} \quad \text{and}$$
$$w\left(\overline{\boldsymbol{\mathcal{S}}}_{uv}\right) = \overline{\mathbf{W}}[u,v] = \overline{w}_{uv}.$$

For the reachability semiring and the network adjacency matrix $\mathbf{W}$, the matrix $\overline{\mathbf{W}}$ is its transitive closure.

For the shortest paths semiring and the network value matrix $\mathbf{W}$, the entry w_{uv}^{*} is the value of the shortest path from u to v.

The paper (Quirin et al. 2008) could be essentially reduced to the observation that the structure $\left(\mathbb{R}_0^+, \min, \boxed{\text{r}}, \infty, 0\right)$ is a (Pathfinder) complete semiring.

Let $(\mathbb{K}, \oplus, \odot, 0, 1)$ be an absorptive semiring and σ be a nonsimple walk from a set $\boldsymbol{\mathcal{S}}_{uv}^{*}$.



Semirings and Matrix Analysis of Networks, Fig. 3 Example of a walk that is not a path

Therefore at least one node $\boldsymbol{v}_j$ appears more than once in σ. The part of a walk between its first and last appearance is a closed walk C (see Fig. 3). The whole walk can be written as $\sigma = P \circ C \circ Q$ where P is the initial segment of σ from u to the first appearance of $\boldsymbol{v}_j$, and Q is the terminal segment of σ from the last appearance of $\boldsymbol{v}_j$ to v. Note that $P \circ Q$ is also a walk. The value of both walks together is $w(\{P \circ Q, P \circ C \circ Q\}) = w(P \circ Q)$. We see that the walks that are not paths do not contribute to the value of walks. Therefore $w(\mathcal{S}^*_{uv}) = w(\mathcal{E}^*_{uv})$. This equality holds also for $\mathcal{S}^*_{uv} = \emptyset$.

Since the node set $\mathcal{V}$ is finite, also the set $\mathcal{E}_{uv}$ is finite which allows us to compute the value $w(\mathcal{S}^*_{uv})$. We already know that $\mathbf{W}^* = \mathbf{W}^{(k)} = (\mathbf{1} \oplus \mathbf{W})^k$ for k large enough.

To compute the closure matrix $\mathbf{W}^*$ of a given matrix over a complete semiring $(\mathbb{K}, \oplus, \odot, 0, 1)$, we can use the Fletcher's algorithm (Fletcher 1980):

```
$\mathbf{C}_0 = \mathbf{W}$
for $k = 1, \ldots, n$ do
    for $i = 1, \ldots, n$ do
        for $j = 1, \ldots, n$ do
            $c_k[i,j] = c_{k-1}[i,j] \oplus c_{k-1}[i,k] \odot$
                $(c_{k-1}[k,k])^* \odot c_{k-1}[k,j]$
    $c_k[k,k] = 1 \oplus c_k[k,k]$
$\mathbf{W}^* = \mathbf{C}_n$
```

If we delete the statement $c_k[k, k] = 1 \oplus c_k[k, k]$, we obtain the algorithm for computing the strict closure $\overline{\mathbf{W}}$. If the addition $\oplus$ is idempotent, we can compute the closure matrix in place – we omit the subscripts in matrices $\mathbf{C}_k$.

The Fletcher's algorithm is a generalization of a sequence of algorithms (Kleene, Warshall, Floyd, Roy) for computing closures on specific semirings.

Multiplication of Matrix and Vector

Let e_i be a unit vector of length n – the only nonzero element is at the i-th position and it is equal to 1. It is essentially a $1 \times n$ matrix. The product of a unit vector and a value matrix of a network can be used to calculate the values of walks from a node i to all the other nodes.

Let us denote $q_1^T = e_i^T \odot \mathbf{W}$. The values of elements of the vector q_1 are equal to the values of walks of the length 1 from a node i to all other nodes: $q_1[j] = w\left(\mathcal{S}^1_{ij}\right)$. We can calculate iteratively the values of all walks of the length s, $s = 2, 3, \ldots, k$ that start in the node

$$i\colon q_s^T = q_{s-1}^T \odot \mathbf{W}$$
$$\text{or } q_s^T = e_i^T \odot \mathbf{W}^s \text{ and } q_s[j] = w\left(\mathcal{S}^s_{ij}\right).$$

Similarly we get

$$q^{(k)^T} = e_i^T \odot \mathbf{W}^{(k)}, q^{(k)}[j] = w\left(\mathcal{S}^{(k)}_{ij}\right)$$
$$\text{and } q^{*^T} = e_i^T \odot \mathbf{W}^*, q^*[j] = w\left(S^*_{ij}\right).$$

This can be generalized as follows. Let $\mathcal{I} \subseteq \mathcal{V}$ and $e_{\mathcal{I}}$ is the characteristic vector of the set $\mathcal{I}$ – it has value 1 for elements of $\mathcal{I}$ and is 0 elsewhere. Then, for example, for $q_k^T = e_{\mathcal{I}}^T \odot \mathbf{W}^k$, it holds $q_k[j] = w\left(\cup_{i \in \mathcal{I}} \mathcal{S}^k_{ij}\right)$.

Future Directions

New network analysis problems are emerging all the time. For some of them a semiring-based approach can prove to be useful. Recently we proposed a longitudinal approach to analysis of temporal networks based on semirings of temporal quantities (Batagelj and Praprotnik 2016).

Cross-References

- ▶ Eigenvalues: Singular Value Decomposition
- ▶ Iterative Methods for Eigenvalues/Eigenvectors
- ▶ Markov Chain Monte Carlo Model
- ▶ Matrix Algebra, Basics of
- ▶ Spectral Analysis

Acknowledgments The first author was financed in part by the European Union, European Social Fund. The work was partially supported by the EUROCORES Programme EUROGIGA of the European Science Foundation, project GReGAS and by Slovenian Research Agency (ARRS) projects J5-5537, J1-6720, and Z7-7614 (B), and grant P1-0294.

References

Abdali SK, Saunders BD (1985) Transitive closure and related semiring properties via eliminants. Theor Comput Sci 40:257–274

Baras JS, Theodorakopoulos G (2010) Path problems in networks. Morgan & Claypool, Berkeley

Batagelj V (1994) Semirings for social networks analysis. J Math Soc 19(1):53–68

Batagelj V, Cerinšek M (2013) On bibliographic networks. Scientometrics 96(3):845–864

Batagelj V, Mrvar A (2008) Analysis of kinship relations with Pajek. Soc Sci Comput Rev 26(2):224–246

Batagelj V, Praprotnik S (2016) An algebraic approach to temporal network analysis based on temporal quantities. Soc Netw Anal Min 6(1):1–22

Burkard RE, Cuninghame-Greene RA, Zimmermann U (eds) (1984) Algebraic and combinatorial methods in operations research, Annals of discrete mathematics, vol 19. North Holland, Amsterdam/New York

Carré B (1979) Graphs and networks. Clarendon, Oxford

Fletcher JG (1980) A more general algorithm for computing closed semiring costs between vertices of a directed graph. Commun ACM 23(6):350–351

Glazek K (2002) A guide to the literature on semirings and their applications in mathematics and information sciences. Kluwer Academic Press, Dordrecht

Golan JS (1999) Semirings and their applications. Springer, Dordrecht

Gondran M, Minoux M (2008) Graphs, dioids and semirings: new models and algorithms. Springer, New York

Kepner J, Gilbert J (2011) Graph algorithms in the language of linear algebra. SIAM, Philadelphia

Quirin A, Cordón O, Santamaria J, Vargas-Quesada B, Moya-Anegón F (2008) A new variant of the Pathfinder algorithm to generate large visual science maps in cubic time. Inf Process Manag 44(4):1611–1623

Schvaneveldt RW, Dearholt DW, Durso FT (1988) Graph theoretic foundations of Pathfinder networks. Comput Math Appl 15(4):337–345

Sentiment Analysis

▶ Multi-classifier System for Sentiment Analysis and Opinion Mining

▶ Sentiment Analysis, Basic Tasks of

▶ Social Media Analysis for Monitoring Political Sentiment

▶ User Sentiment and Opinion Analysis

Sentiment Analysis in Social Media

Noor Fazilla Abd Yusof[1], Chenghua Lin[1] and Yulan He[2]
[1]Department of Computing Science, University of Aberdeen, Aberdeen, UK
[2]School of Engineering and Applied Science, Aston University, Birmingham, UK

Synonyms

Data mining; Knowledge discovery; Opinion mining; Sentiment classification; Social media analysis

Glossary

NB	Naive Bayes classifier
SVM	Support vector machines
MaxEnt	Maximum entropy classifier
PMI	Point-wise mutual information
POS	Part-of-speech
SO	Sentiment orientation

Definition

Sentiment analysis aims to understand subjective information such as opinions, attitudes, and feelings expressed in text. Sentiment analysis tasks include but not limited to the following:

- **Sentiment classification** which classifies a given piece of text as positive, negative, or neutral.
- **Opinion retrieval** which retrieves opinions in relevance to a specific topic or query.
- **Opinion summarization** which summarizes opinions over multiple text sources towards a certain topic.
- **Opinion holder identification** which identifies who express a specific opinion.
- **Topic/sentiment dynamics tracking** which aims to track sentiment and topic changes over time.
- **Opinion spam detection** which identifies fake/untruthful opinions.
- **Prediction** which predicts people's behaviors, market trends, political election outcomes, etc., based on opinions or sentiments expressed in online contents.

Introduction

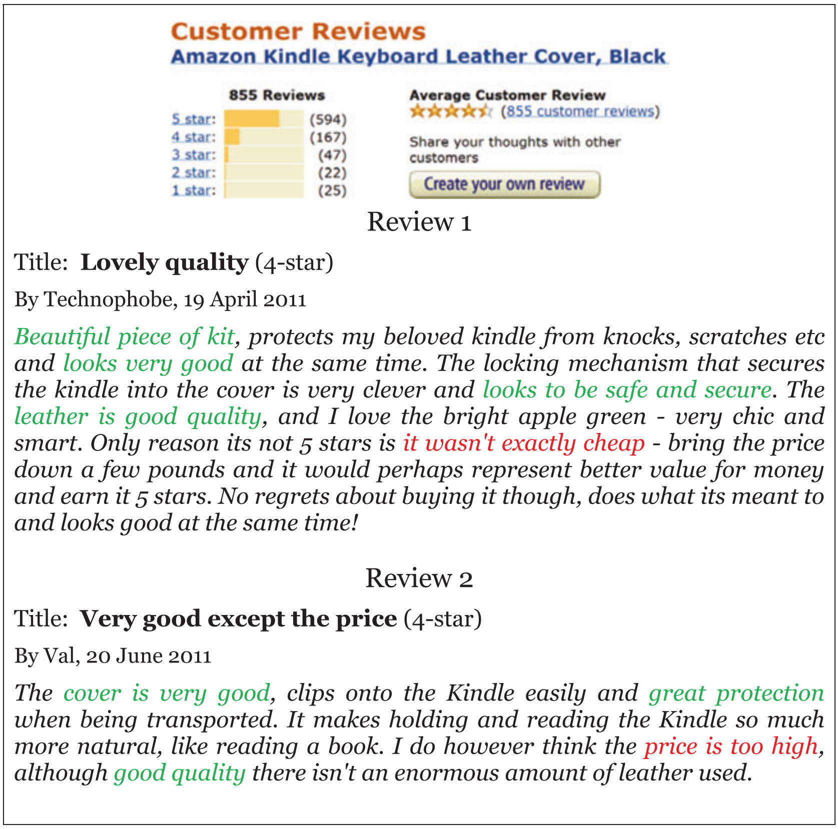
With the explosion of people's attitudes and opinions expressed in social media such as blogs, discussion forums, and tweets, detecting sentiment or opinion from the Web is becoming an increasingly popular way of interpreting data. Sentiment analysis in social media allows business organizations to monitor their reputations, find public opinions about their products or services and those of their competitors, and provide them with insight into emerging trends and potential changes in market opinion, etc.

Customers also rely on online reviews to make more informed purchase decisions. Taking the Amazon Kindle cover reviews shown in Fig. 1 as an example, this Kindle cover receives a very high average rating of 4.5 stars from a total of 855 reviews. Nevertheless, some reviews with high star ratings might still contain negative comments. Two example 4-star reviews shown in Fig. 1 reveal that although people think the design and quality of the cover are very good, it is overpriced. With such information, the cover would still be a good buy for price-insensitive customers, whereas other customers may choose a less expensive alternative. With the sheer volume of social media data published every day on the web and driven by the demand of gleaning insights into such great amounts user-generated data, there have been a large number of sentiment analysis software tools developed for alleviating users' information seeking burden.

There has been research (Liu et al. 2016; Moe and Schweidel 2012) on analyzing rating and reviews which enables brands to capture customer's opinions about their services and products. Lightweight tools, such as Tweetfeel (Barbosa and Feng 2010), scours Twitter for tweets and shows how positively or negatively Twitter users feel about a particular topic. The *Financial Times* introduced Newssift (Wan and Paris 2014; Denecke 2009), a search tool that matches business topics to users' queries, sorts articles into positive and negative sentiment, and identifies people, companies, places, and connections across all stories allowing for further refined search. There has also been increasing use of social media in developing virtual communities that facilitate professional networking, knowledge sharing, and evidence-informed practice in the healthcare domain. For instance, PatientsLikeMe and CureTogether (Grajales III et al. 2014) provide user-driven platforms for patients to share information on symptoms, causes, treatment efficacy, as well as documenting the progression of their disease.

Key Points

This entry primarily focuses on sentiment classification from social media data. It describes some of the prominent approaches to sentiment classification including corpus-based approaches, lexicon-based approaches, and the incorporation of social networks into sentiment classification.



Customer Reviews

Amazon Kindle Keyboard Leather Cover, Black

855 Reviews

5 star:	(594)
4 star:	(167)
3 star:	(47)
2 star:	(22)
1 star:	(25)

Average Customer Review

☆☆☆☆☆ (855 customer reviews)

Share your thoughts with other customers

Create your own review

Review 1

Title: **Lovely quality** (4-star)

By Technophobe, 19 April 2011

Beautiful piece of kit, protects my beloved kindle from knocks, scratches etc and looks very good at the same time. The locking mechanism that secures the kindle into the cover is very clever and looks to be safe and secure. The leather is good quality, and I love the bright apple green - very chic and smart. Only reason its not 5 stars is it wasn't exactly cheap - bring the price down a few pounds and it would perhaps represent better value for money and earn it 5 stars. No regrets about buying it though, does what its meant to and looks good at the same time!

Review 2

Title: **Very good except the price** (4-star)

By Val, 20 June 2011

The cover is very good, clips onto the Kindle easily and great protection when being transported. It makes holding and reading the Kindle so much more natural, like reading a book. I do however think the price is too high, although good quality there isn't an enormous amount of leather used.

Sentiment Analysis in Social Media, Fig. 1 Amazon Kindle cover reviews. Text highlighted in *green* and *red* indicate the pros and cons of the product respectively

Historical Background

In the past, the majority of work in text information processing focused on mining and retrieving factual information, such as classifying documents according to their subject matter (e.g., politics vs. religion and sports vs. arts). In recent years, there has been a rapid growth of research interests in natural language processing that seeks to better understand sentiment or opinion expressed in text. One reason is that with the rise of various types of social media, communicating on the web has become increasingly popular, where millions of people broadcast their thoughts and opinions on a great variety of topics, such as feedbacks on products and services, opinions on political development and events, and information sharing on global disasters. Therefore, new computational tools are needed to help organize, summarize, and understand this vast amount of information. Additionally, the discovery of opinions reflecting people's attitudes towards various topics enables many useful applications, which is another motivation of sentiment analysis.

Sentiment analysis can be considered as computational treatments of subjective information such as opinions and emotions expressed in text. In the simplest setting, sentiment analysis aims to automatically identify whether a given piece of text expresses positive or negative opinion. Early approaches (Pang et al. 2002; Matsumoto et al. 2005) view sentiment classification as a text classification problem where a corpus with sentiment orientation annotated is required for classifiers training. Supervised sentiment classification approaches usually perform well when the training set is large enough, where the state-of-the-art approach (Matsumoto et al. 2005) can achieve more than 90% accuracy on the movie review data. However, there are some noticeable issues. One is that supervised classifier trained on one domain often fails to produce satisfactory performance when tested on other domains, and secondly, online content varies widely in domains and evolves rapidly over time, making corpora annotation for each domain unrealistic.

In response to the domain transfer and labeling cost problems faced by supervised approaches, there has been rising interest in exploring semi-supervised methods leveraging a large amount of unlabeled data and a small amount of labeled data for classifier training (Aue and Gamon 2005; Blitzer et al. 2007; He et al. 2011). Some representative work in this line are that of Aue and Gamon (2005) which explored various strategies for training SVM classifiers for the target domain lacking sufficient labeled data, and the work of Blitzer et al. (2007) which addressed domain transfer problem with structural correspondence learning (SCL).

Unsupervised or weakly supervised approaches are mostly lexicon-based which do not require labeled document for training. Instead, they assume that the sentiment orientation of a document is an averaged sum of the sentiment orientations of its words and phrases. Given the difficulties of supervised and semi-supervised sentiment analysis, it is conceivable that unsupervised or weakly supervised approaches to sentiment classification are even more challenging. Nevertheless, solutions to unsupervised or weakly supervised sentiment classification are of practical significance owing to its domain-independent nature.

The pioneer work is the pointwise mutual information approach proposed in (Turney 2002), who calculated the sentiment orientations of phrases in documents that contain adjectives or adverb as the pointwise mutual information (PMI) with a positive prototype "excellent" minus the PMI with a negative prototype "poor." The proposed approach achieved an accuracy of 84% for automobile reviews and 66% for movie reviews. Also work such as Read and Carroll (2009) are good examples of this lexicon-based approach.

Weakly supervised sentiment classification approaches are similar to unsupervised approaches in that they do not require labeled documents for training. Instead, they typically incorporate supervision information either from sentiment lexicons containing a list of words marked as positive or negative (usually much larger in size than the sentiment seed words used in unsupervised approaches), or from user feedbacks. Lin and He (2009) proposed a joint sentiment-topic (JST) model to detect document-level sentiment and extract sentiment bearing topics simultaneously from text. By incorporating a small set of domain-independent sentiment words as prior knowledge for model learning, the weakly supervised JST model is able to achieve comparable performance to semi-supervised approaches with 40% labeled data (Lin et al. 2011).

Compared to the vast majority of work in sentiment analysis mainly focuses on the domains of product reviews and blogs, Twitter sentiment analysis is considered as a much harder problem than sentiment analysis on conventional text, mainly due to the short length of tweet messages, the frequent use of informal and irregular words, and the rapid evolution of language in Twitter. Annotated tweets data are impractical to obtain. Previous work on twitter sentiment analysis (Go et al. 2009; Pak and Paroubek 2010; Barbosa and Feng 2010) rely on noisy labels or distant supervision, for example, by taking emoticons as the indication of tweet sentiment to train supervised classifiers. Other works explore feature engineering in combination of machine learning methods to improve sentiment classification accuracy on tweets (Agarwal et al. 2011; Kouloumpis et al. 2011).

Prominent Methodologies

Research on sentiment classification has attracted a great deal of attention, where different classification tasks focus on various levels of granularity, e.g., from the document level (Pang et al. 2002) to the finer-grained sentence and word/phrase level (Turney and Littman 2002). In this section, we investigate the work which deals with computational treatments of sentiment using corpus-based and lexicon-based approaches, with a focus on document-level sentiment classification.

Corpus-Based Approaches

Corpus-based approaches (Pang et al. 2002; Pang and Lee 2004; Boiy et al. 2007) rely on annotated corpora where each document is annotated with a polarity label such as positive, negative, and neutral. Standard classifiers such as naïve Bayes (NB), maximum entropy (MaxEnt), and support vector machines (SVMs) can then be trained from such annotated corpora to detect the sentiment of text. In Twitter sentiment analysis where annotated data are impractical to obtain, noisy labels such as emoticons (":−)", ":D", ":(", etc.) appeared in tweets are used to label tweets as positive or negative (Go et al. 2009).

Pioneering work on document-level sentiment classification is by Pang et al. (2002) who employed machine learning techniques including SVMs, NB, and MaxEnt to determine whether the sentiment expressed in a movie review was "*thumbs up*" or "*thumbs down.*" They achieved the best classification accuracy with SVMs using binary features coding whether a unigram was present or not. In subsequent work, Pang and Lee (2004) further improved sentiment classification accuracy on the movie review dataset using a cascaded approach. Instead of training a classifier on the original feature space, they first filtered out the objective sentences from the dataset using a global min-cut inference algorithm, and then used the remaining subjective sentences as input for sentiment classifier training. The classification improvement achieved by the cascaded approach suggests that the subjective sentences contain features which are more discriminative and informative than the full dataset for sentiment classification. The movie review dataset (also known as the polarity dataset, http://www.cs.cornell.edu/people/pabo/movie-review-data/) used in Pang et al. (2002) and Pang and Lee (2004) has later on become a benchmark for many sentiment classification studies (Whitelaw et al. 2005; Matsumoto et al. 2005). Whitelaw et al. (2005) used fine-grained semantic distinctions in features for sentiment classification, namely the appraisal groups. Specifically, an *appraisal group* is defined as coherent groups of words that express together a particular attitude, such as "*extremely boring*" and "*not terribly funny.*" By training a SVM classifier on the combination of different types of appraisal group features and bag-of-word features, they achieved the best accuracy of 90.2% on the movie review dataset. Matsumoto et al. (2005) proposed a method using the extracted word sub-sequences and dependency sub-trees as features for SVMs training and attained the state-of-the-art accuracy of 93.7%.

A common assumption made by the aforementioned line of work (Pang et al. 2002; Pang and Lee 2004; Whitelaw et al. 2005) is that the entire document is represented as a flat feature vector (i.e., a bag-of-words format), which limits their ability to exploit sentiment or subjectivity information at a finer-grained level. In this regard, there has been work on incorporating sentence or sub-sentence level sentiment label information for document-level sentiment classification (McDonald et al. 2007; Zaidan et al. 2007).

McDonald et al. (2007) proposed a fully supervised structured model for joint sentence- and document-level sentiment classification based on sequence classification techniques using constrained Viterbi inference. The joint model leverages both document-level and sentence-level label information and allows classification decisions from one level (e.g., the document level) to influence decisions at another level (e.g., the sentence level). It was reported that the joint model significantly outperformed both the document- and sentence-classifier that predict a single level label only. Zaidan et al. (2007) used human annotators to mark the sub-sentence level text spans known as "*annotator rationales*," which support the document's sentiment label. These annotator rationales were used as additional constraints for SVMs training, which ensure that the resulting classifier will be less confident in classifying the documents that do not contain the rationales. By exploiting the rationales during the classifier training, the proposed approach achieved 92.2% accuracy on the movie review dataset, and significantly outperformed the baseline SVM which only used the full text of the original documents for training.

Apart from exploiting structured information for sentiment classification, there are also work on exploring various features such as unigrams, bigrams, and part-of-speech (POS) tags, for building sentiment classifiers. Agarwal et al. (2011) studied using the feature-based model and the tree kernel–based model for sentiment classification. They explored a total of 50 different feature types and showed that both the feature-based and tree kernel–based models perform similarly and they outperform the unigram baseline. Kouloumpis et al. (2011) compared various features including n-gram features, lexicon features based on the existence of polarity words from the MPQA subjectivity lexicon (http://www.cs.pitt.edu/mpqa/), POS features, and microblogging features capturing the presence of emoticons, abbreviations, and intensifiers. They found that micoblogging features are most useful in sentiment classification.

Example: Sentiment Classification Based on Supervised Learning

In this section, we illustrate an example of how to train a binary naive Bayes classifier for document-level sentiment classification, i.e., to determine the sentiment orientation of a document as positive or negative. The procedures of classifier training involves three steps as depicted in Fig. 2.

Step 1 Prepare a training set: Given a set of opinionated documents $\mathcal{D} = \{d_1, d_2, \cdots, d_D\}$, each document $d \in \mathcal{D}$ needs to be annotated with a sentiment label $c \in \mathcal{C}$ as positive or negative prior to classifier training. Thus, training examples can be represented as pairs of documents and the corresponding sentiment labels as $\{\mathcal{D}, \mathcal{C}\} = \{(d_1, c_1), \ldots, (d_D, c_D)\}$. Also, using V to denote the number of distinct terms in the training set, each document can then be represented as a V-dimensional binary vector with each dimension t corresponding to term w_t. By employing the binary naive Bayes model which only encodes the presence of words, the feature presence indicator λ_{it} (i.e., the tth dimension of document d_i) can only take two possible values, i.e., 0 indicating w_t does not appear in d_i or 1 indicating w_t has occurred in d_i at least once.

Step 2 Train a sentiment model: Given a training set $\{\mathcal{D}, \mathcal{C}\}$, the goal of model training is to calculate the optimal parameter estimates of a naive Bayes model $\mathcal{M}$. Specifically, for each term w_t in the vocabulary and each class label c_j, we need to calculate $P(w_t|c_j)$, i.e., the probability of generating w_t given class label c_j. Using the independence assumptions of NB that all attributes of data examples are independent of each other given a class label Lewis (1998), $P(w_t|c_j)$ can be approximated from training data as

$$P(w_t|\, c_j) = \frac{\#\text{documents with label } c_j \text{ that contain } w_t}{\#\text{documents with label } c_j}. \tag{1}$$

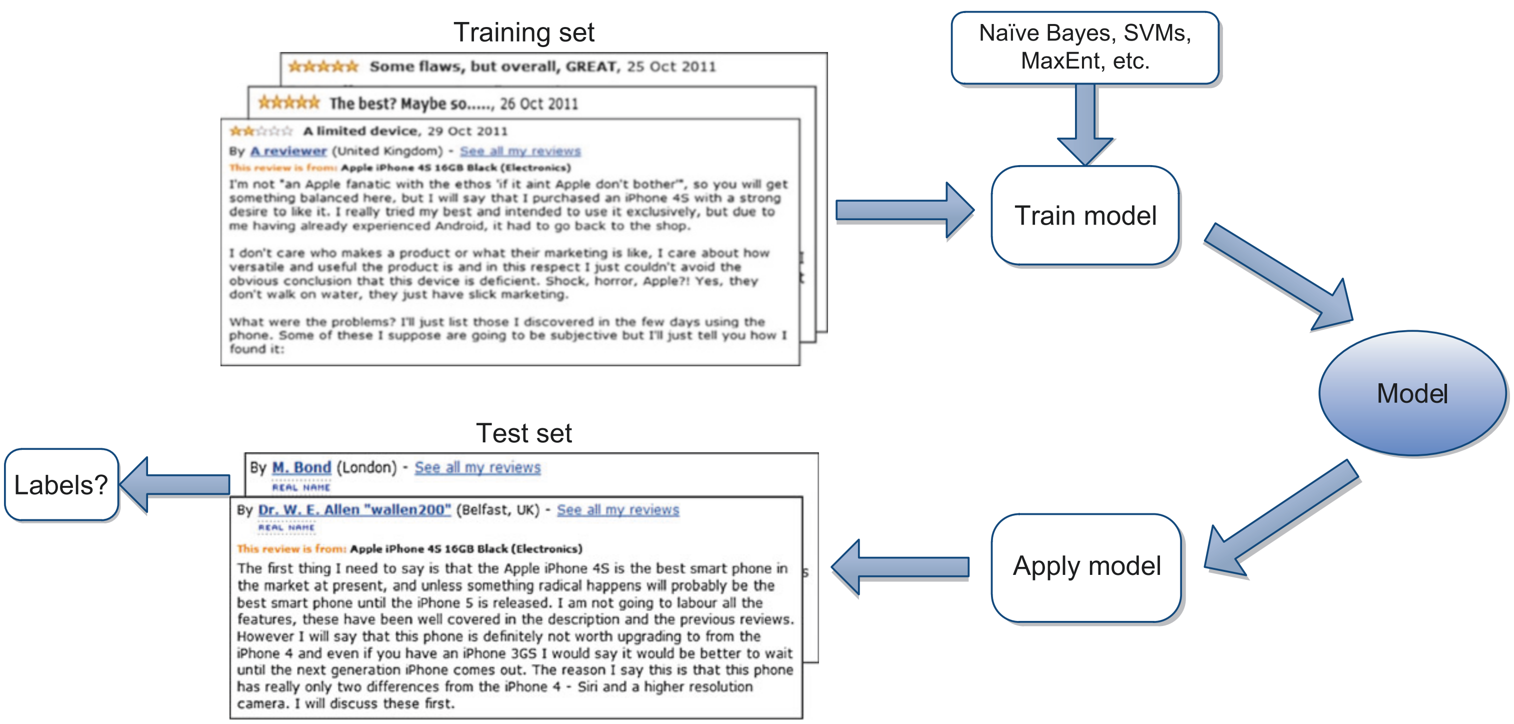
We also need to compute the sentiment class probability, $P(c_j)$, which can be estimated as the proportion of documents labeled as class c_j in the training data:

$$P(w_t|\, c_j) = \frac{\#\text{documents with label } c_j}{\#\text{documents in the training data}}. \tag{2}$$

Step 3 Predict sentiment label for unseen documents: Given a set of unseen documents $\mathcal{D}_u$, the final step is to predict the most probable sentiment class label $\tilde{c}$ for each unseen document $d_u \in \mathcal{D}_u$. By applying the previously trained model $\mathcal{M}$, the posterior $p(c_j|d_u)$, i.e., the probability that unseen document d_u belongs to class c_j, can be calculated as

$$\begin{aligned} P(c_j|\, d_u) &= \frac{P(c_j)P(d_u|\, c_j)}{P(d_u)} \\ &= \frac{P(c_j)\prod_{t=1}^{V} P(w_t|\, c_j)}{P(d_u)}, \end{aligned} \tag{3}$$

where $P(d_u)$ is a normalization constant which plays no role in classification, and $P(c_j)$ and



Some flaws, but overall, GREAT, 25 Oct 2011

The best? Maybe so....., 26 Oct 2011

A limited device, 29 Oct 2011

By A reviewer (United Kingdom) - See all my reviews

This review is from: Apple iPhone 4S 16GB Black (Electronics)

I'm not "an Apple fanatic with the ethos 'if it aint Apple don't bother", so you will get something balanced here, but I will say that I purchased an iPhone 4S with a strong desire to like it. I really tried my best and intended to use it exclusively, but due to me having already experienced Android, it had to go back to the shop.

I don't care who makes a product or what their marketing is like, I care about how versatile and useful the product is and in this respect I just couldn't avoid the obvious conclusion that this device is deficient. Shock, horror, Apple?! Yes, they don't walk on water, they just have slick marketing.

What were the problems? I'll just list those I discovered in the few days using the phone. Some of these I suppose are going to be subjective but I'll just tell you how I found it:

By M. Bond (London) - See all my reviews

REAL NAME

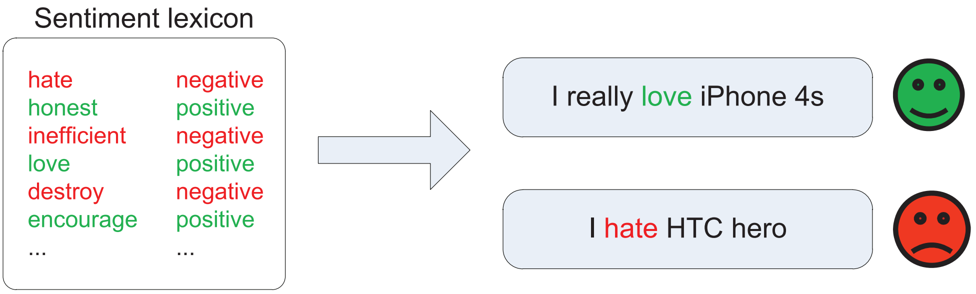
By Dr. W. E. Allen "wallen200" (Belfast, UK) - See all my reviews

REAL NAME

This review is from: Apple iPhone 4S 16GB Black (Electronics)

The first thing I need to say is that the Apple iPhone 4S is the best smart phone in the market at present, and unless something radical happens will probably be the best smart phone until the iPhone 5 is released. I am not going to labour all the features, these have been well covered in the description and the previous reviews. However I will say that this phone is definitely not worth upgrading to from the iPhone 4 and even if you have an iPhone 3GS I would say it would be better to wait until the next generation iPhone comes out. The reason I say this is that this phone has really only two differences from the iPhone 4 - Siri and a higher resolution camera. I will discuss these first.

Sentiment Analysis in Social Media, Fig. 2 Illustration of corpus-based approaches



Sentiment Analysis in Social Media, Fig. 3 Illustration of lexicon-based approaches

$P\ (w_t|c_j)$ are the probabilities estimated from training data in Step 2. Finally, the class label of d_u is determined as $\widehat{c_j} = \text{argmax}_{c_j} P\left(c_j|\, d_u\right)$.

Lexicon-Based Approaches

Lexicon-based approaches for sentiment classification are mostly unsupervised or weakly supervised. As unsupervised classifiers are usually not able to identify which features are relevant to polarity classification in the absence of annotated data, they normally resort to sentiment seed words or lexicons as a form of prior polarity knowledge for model learning as illustrated in Fig. 3. Such domain-independent sentiment lexicons can be acquired automatically or semiautomatically with much less effort compared to labeling a large training dataset.

The pioneering work in this line is that of Turney and Littman (2002), which classified a document as positive or negative by the average sentiment orientation of the phrases containing adjectives or adverbs in the document. The sentiment orientation of a phrase is calculated as the pointwise mutual information (PMI) with a positive word "*excellent*" minus the PMI with a negative word "*poor.*" The proposed approach achieved an accuracy of 84% for automobile reviews and 66% for movie reviews. In the Sentiment lexicon same vein, Read and Carroll (2009) measured the similarity between words and polarity prototypes such as "*excellent*" and "*good*" with three different methods, namely, lexical association (using PMI), semantic spaces, and distributional similarity. While Turney and Littman (2002) only used one polarity prototype for each sentiment class, Read and Carroll experimented with seven polarity prototypes obtained from Roget's Thesaurus and WordNet (http://wordnet.princeton.edu/) through a selection process based on their frequency in the Gigaword corpus. The best result was achieved using PMI with 69.1% accuracy obtained on the movie review data.

While a fixed number of sentiment seed words have been used in the aforementioned work (Turney and Littman 2002; Read and Carroll 2009), there have been attempts to incrementally enlarge the unlabeled examples with self-training based on the original seed word input (Zagibalov and Carroll 2008a, b). Starting with a single Chinese sentiment seed word meaning "*good*," Zagibalov and Carroll (2008b) used iterative retraining to gradually enlarge the seed vocabulary. Those enlarged sentiment-bearing words are selected based on their relative frequency in both the positive and negative parts of the current training data. The sentiment orientation of a document is then determined by the sum of the sentiment scores of all the sentiment-bearing lexical items found in the document. Problems with this approach are that there is no principled mechanism for determining the optimal iteration number for training as well as for selecting the initial seed word, where inappropriate seed word selection may result in very poor accuracy. As such, in subsequent work, Zagibalov and Carroll (2008a) introduced a way for automatic seed word selection based on some heuristic knowledge, and an iteration control method was proposed so that iterative training stops when there is no change to the classification of any document over the previous two iterations.

Weakly supervised sentiment classification approaches are mostly lexicon-based, some of which integrate with corpus-based methods as a hybrid model (Qiu et al. 2009). Compared to the seed words used in unsupervised methods, the sentiment lexicon, consisting of a list of positive and negative sentiment bearing words, is usually much larger in size and is used as reference features for sentiment classification. Analogous to the unsupervised approach that uses iterative retraining (Zagibalov and Carroll 2008b), Qiu et al. (2009) also used a lexicon-based iterative process to iteratively enlarge an initial sentiment dictionary from the first phrase. But instead of using a single seed word as Zagibalov and Carroll (2008b), they started with a much larger Chinese sentiment dictionary *HowNet* (http://www.keenage.com/download/sentiment.rar) as the initial lexicon. Documents classified from the first phase were taken as a training set to train SVMs, which were subsequently used to revise the results produced from the first phase. This self-supervised approach was tested on reviews from ten different domains and outperformed the best results of the approach by Zagibalov and Carroll (2008a) on the same data over 6% in F-measure. The weakly supervised joint sentiment topic (JST) model (Lin and He 2009) can detect sentiment and topic simultaneously from text by incorporating a small set of domain independent sentiment lexicon (http://mpqa.cs.pitt.edu/). Unlike supervised approaches to sentiment classification which often fail to produce satisfactory performance when applied to other domains, the weakly supervised nature of JST makes it highly portable to other domains, and it is able to achieve comparable performance to the semi-supervised approaches using 40% labeled data for training.

Example: Sentiment Classification Based on Unsupervised Learning

In this section, we show how to perform sentiment classification using pointwise mutual information (PMI) (Turney and Littman 2002) as this is one of the pioneering work of lexicon-based approach for sentiment classification. The PMI algorithm can be boiled down into three steps.

Step 1 Extract phrases containing adjectives or adverbs: The first step of the PMI algorithm is to extract two-word phrases from the document where one member of the phrase is an adjective or an adverb and the second provides context. The rationale behind is that although adjectives are generally considered good indicators for subjectivity detection from text, using an isolated adjective alone may be insufficient to determine sentiment orientation as sentiment is context dependent. For instance, the adjective "complicated" may have negative sentiment orientation as "complicated setting" in an electronic product review, and conveys positive sentiment as "complicated plot" in a movie review. This phrase extraction process consists of two steps by firstly applying POS tagger to documents and then discarding the phrases with POS tags that do not conform to some predefined syntactic patterns. Readers may refer to the original paper (Turney and Littman 2002) for a full list of POS tag patterns.

Step 2 Estimate phrase sentiment orientation: In order to calculate the sentiment orientation (SO) of each extracted phrases, two sentiment polarity reference words are used, with word "excellent" indicating positive sentiment and "poor" indicating negative sentiment. So the SO of a phrase is measured by the difference of its PMI with positive word "excellent" and negative word "poor" as follows

$$\begin{aligned}\mathrm{SO}(phrase) = \ & \mathrm{PMI}\left(phrase, \text{“excellent”}\right) \\ & -\mathrm{PMI}\left(phrase, \text{“poor”}\right).\end{aligned} \tag{4}$$

Formally, the PMI of words w_1 and w_2 is given by.

$$\mathrm{PMI}\left(w_1, w_2\right) = \log_2\left(\frac{p(w_1 \wedge w_2)}{p(w_1)p(w_2)}\right), \tag{5}$$

where $p(w_1 \wedge w_2)$ is the joint probability of how likely that word w_1 and w_2 co-occur. If w_1 and w_2 are statistically independent, this joint probability is equivalent to $p(w_1)p(w_2)$. Thus, the ratio between $p(w_1 \wedge w_2)$ and $p(w_1)p(w_2)$ essentially measures the degree of statistical dependence between the words. In practice, the probabilities required for calculating PMI can be acquired by issuing quires to a public search engine (http://www.altavista.com/sites/search/adv), and then based on the results returned, we can approximate $p(w_1)$ with the number of hits that documents contain w_1 and approximate $p(w_1 \wedge w_2)$ with the number of hits that documents contain both w_1 and w_2 within a range of ten words. Thus, Eq. 4 can be rewritten as

$$\mathrm{SO}(\mathit{phrase}) = \log_2\left(\frac{\mathrm{hits}(\mathit{phrase}\ \mathrm{NEAR}\ \text{“excellent”})\ \mathrm{hits}(\text{“poor”})}{\mathrm{hits}(\mathrm{phrase}\ \mathrm{NEAR}\ \text{“poor”})\ \mathrm{hits}(\text{“excellent”})}\right). \tag{6}$$

Step 3 Calculate document sentiment: The final step is to calculate the average SO of all extracted phrases in the document, and then classify the document as positive if the average SO is positive and as negative otherwise.

Explore Social Networks for Sentiment Analysis

Recently, there have been increasing interests in employing social relations for both document-level and user-level sentiment analysis. It is based on a hypothesis that users connected with each other are likely to express similar opinions. In Twitter, social relations can be established by the following links, through retweeting using RTusername or via username, or by referring to other users in one’s messages using “@” mentions.

Speriosu et al. (2011) argued that using noisy sentiment labels may hinder the performance of sentiment classifiers. They proposed exploiting the Twitter follower graph to improve sentiment classification and constructed a graph that has users, tweets, word unigrams, word bigrams, hashtags, and emotions as its nodes which are connected based on the link existence among them (e.g., users are connected to tweets they created; tweets are connected to word unigrams that they contain). They then applied a label propagation method where sentiment labels were propagated from a small set of nodes seeded with some initial label information throughout the graph. They claimed that their label propagation method outperforms MaxEnt trained from noisy labels and obtained an accuracy of 84.7% on the subset of the twitter sentiment test set from Go et al. (2009).

Tan et al. (2011) incorporated both textual and social relations revealed by the following links and “@” mentions in a single heterogeneous graph on a certain topic such as “Obama,” where nodes correspond to either users or tweets. Started from some seed user nodes labeled as positive or negative, they proposed a transductive learning method to propagate sentiment label to all the users in the graph. In a similar vein, Calais Guerra et al. (2011) also proposed modeling the user opinion prediction problem as a relational learning problem over a network of users connected by endorsement (e.g., retweets in Twitter) where the goal is to classify the nodes of a partially labeled network.

In the psychology domain, evidences show that analyzing the sentiment of social media data may present innovative opportunities for mental health studies, complementing the traditional methods used in clinical settings (De Choudhury et al. 2013b; Park et al. 2013). Moreno et al. (2011) employed negative binomial regression analysis to ascertain the pervasiveness of depression symptoms among college students. They assessed 1-year Facebook status updates of students and explored whether the depression symptoms reported on Facebook met the symptom criteria for Major Depressive Episode (MDE). It was found that depression symptoms disclosed on Facebook indeed correlate with the DSM diagnosis under a clinical setting. In a similar study, De Choudhury et al. (2013a) identified the level of depression in populations by analyzing different features of Twitter posts, including one’s social activity level, emotion words usage, and the language patterns of posts.

Key Applications

Social media such as Twitter and Facebook has become an increasingly popular communication channel, which have enabled many useful applications by discovering opinions reflecting people's attitudes towards various topics or events from the massive user-generated data. These social media centric applications are particularly proliferous in the domains of financial marketing, brand and consumer perception, as well as mental healthcare.

Financial Marketing

Sentiment analysis has shown great impact on financial markets, where financial organizations are embracing new tools and techniques to help make sense of the massive amounts of unstructured data available on social media for making more informed decisions and maximizing the performance of their trading strategies. For instance, Thomson Reuters launched a sentiment analytics service for Internet news and social media, which is capable to mine expansive wealth of social media and blog content to deliver digestible analytics for algorithmic trading systems as well as risk management and human decision support processes (Uhl 2014). There are also research which used Twitter to predict the up or down movement of stock price (Si et al. 2014; Nguyen et al. 2015; Si et al. 2013). For instance, Nguyen et al. (2015) and Si et al. (2013) used topic-sentiment from social media to predict stock price movement, with a focus on identifying the sentiment of the specific topics of that particular company such as product, services, or dividend.

Brand and Consumer Perception

Engaging with consumers and gaining perceptions of brands is another active domain of applying social media analytics, where commercial products preserve similar visions such as to supporting brands to better understand costumer segments, what consumers value about the brands, and how consumers perceive their products and services and those of their competitors.

IBM has developed an internally used system called Banter for monitoring and analyzing the contents of blogs and social network conversation (Thom and Millen 2012). It answers key questions for marketers such as *How do I identify relevant blogs?*, *Who are the key influencers?*, and *What is the sentiment about these relevant topics?*, etc. In terms of commercial products, one of the leading companies is Bazaarvoice (Sarner et al. 2011) which provides a comprehensive social media analytics platform covering a range of services, such as gathering consumer generated opinions from customer conversations on social networks as well as capturing and responding to consumer questions about products and services, etc. Another major player in this market is PowerReviews (Freedman 2011). In contrast to Bazaarvoice which targets enterprises, PowerReviews is more focused on small and medium-sized business (SMB) solutions.

Mental Healthcare

Applying social media for mental healthcare related studies has increased significantly in the past few years (Barak and Grohol 2011; Grajales III et al. 2014). Notwithstanding, Facebook has led an initiative for suicide prevention by allowing Facebook users to flag posts with suicidal intents. Users being flagged will receive a private message from the Facebook safety team with appropriate supports (Ruder et al. 2011). EmotionDiary (Park et al. 2013) is a mobile application which helps users to recognize if they have any potential symptoms of depression by tracking users' emotional state over time. If depressive mood pattern is detected, users will be given suggestions such as seeking help from mental health professionals is needed. Social media has also been used for health education and promotion (Barak and Grohol 2011). For instance, Second Life (Boulos et al. 2007), one of the most popular virtual 3D social network platforms, provides a wide range of medical and healthcare education to its virtual citizens, e.g., educating people about schizophrenia (Yellowlees and Cook 2006), and skills for dealing with mental health issues, etc.

Future Directions

This chapter gave an introduction to sentiment analysis in social media. Despite the recent successes, the field of sentiment analysis is still relatively new and many challenges remained to be tackled:

1. Topic-dependent sentiment analysis. Sentiment is domain dependent, where sentiment expressions in different domains can be quite different. Besides, even for data from the same domain, sentiment distributions may vary over time, especially for collections that span years or decades, and the fast evolving social media data such as Twitter data. Therefore, topic-sensitive sentiment analysis, as well as detecting and tracking the dynamics in both topic and sentiment over time in time-variant datasets, are promising areas for research.
2. Multilingual analysis. Most of the sentiment analysis systems are monolingual which typically process English only. However, a sentiment system with multilingual capability is important as users such as international companies often need to gain insights into markets for more than one country, e.g., for both the USA and China.
3. In Twitter and other social media sites such as Facebook and YouTube, short, ungrammatical utterances are commonplace. Finding an effective way of correcting these spelling mistakes are important for improving sentiment analysis system performance.
4. Sarcasm and slang. Sentiment is often embodied in subtle linguistic mechanisms such as the use of sarcasm and slang, which poses great challenges for automated sentiment analysis. For instance, without taking context into account, sarcasms expressing negative sentiment could be wrongly interpreted as extremely positive sentiment. On the other hand, understanding slang is also very difficult as it changes by geographical location. Therefore, addressing this challenge would require deeper linguistic understanding and incorporating richer background knowledge for model learning.

Cross-References

▶ Data Mining
▶ Machine Learning
▶ Multi-Classifier System for Sentiment Analysis and Opinion Mining
▶ Twitter Microblog Sentiment Analysis
▶ User Sentiment and Opinion Analysis

Acknowledgments This work is supported by the awards made by the UK Engineering and Physical Sciences Research Council (Grant number: EP/P005810/1) and the UK Economic & Social Research Council (Grant number: ES/P011004/1).

References

Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R (2011) Sentiment analysis of twitter data. In: Proceedings of the workshop on languages in social media. Association for Computational Linguistics, Portland, pp 30–38

Aue A, Gamon M (2005) Customizing sentiment classifiers to new domains: a case study. In: Proceedings of recent advances in natural language processing (RANLP), Borovets

Barak A, Grohol JM (2011) Current and future trends in internet-supported mental health interventions. J Technol Hum Serv 29(3):155–196

Barbosa L, Feng J (2010) Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd international conference on computational linguistics: posters. Association for Computational Linguistics, Beijing, pp 36–44

Blitzer J, Dredze M, Pereira F (2007) Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification. In: Proceedings of the Association for Computational Linguistics (ACL), Prague, pp 440–447

Boiy E, Hens P, Deschacht K, Moens MF (2007) Automatic sentiment analysis of on-line text. In Proceedings of the 11th International Conference on Electronic Publishing, pages 349–360, Vienna

Boulos MNK, Hetherington L, Wheeler S (2007) Second life: an overview of the potential of 3-d virtual worlds in medical and health education. Health Inf Libr J 24 (4):233–245

Calais Guerra P, Veloso A, Meira Jr W, Almeida V (2011) From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, San Diego, pp 150–158

De Choudhury M, Counts S, Horvitz E (2013a) Social media as a measurement tool of depression in populations. In: Proceedings of the 5th annual ACM web science conference. ACM, New York, pp 47–56

De Choudhury M, Gamon M, Counts S, Horvitz E (2013b) Predicting depression via social media. In: Proceedings of ICWSM, p 2

Denecke K (2009) Assessing content diversity in medical weblogs. In: Proceedings of the first international workshop on living web at the 8th international semantic web conference (ISWC)

Freedman L (2011) The 2011 Social Shopping Study [Online]. Available: http://www.powerreviews.com/assets/download/Social_Shopping_2011_Brief1.pdf [Accessed 2017- 01-13]

Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford 1(2009): 12

Grajales FJ III, Sheps S, Ho K, Novak-Lauscher H, Eysenbach G (2014) Social media: a review and tutorial of applications in medicine and health care. J Med Internet Res 16(2):e13

He Y, Lin C, Alani H (2011) Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In: Proceedings of the 49th annual meeting of the Association for Computational Linguistics: human language technologies, vol vol 1. Association for Computational Linguistics, Portland, pp 123–131

Kouloumpis E, Wilson T, Moore JD (2011) Twitter sentiment analysis: the good the bad and the omg! Icwsm 11 (538–541):164

Lewis D (1998) Naive (bayes) at forty: the independence assumption in information retrieval. In: Machine learning: ECML-98. Springer, Chemnitz, pp 4–15

Lin C, He Y (2009) Joint sentiment/topic model for sentiment analysis. In: The 18th ACM conference on information and knowledge management (CIKM), Hong Kong

Lin C, He Y, Everson R, Rüger S (2011) Weakly-supervised joint sentiment-topic detection from text. IEEE Trans Knowl Data Eng (TKDE) 24(6):1134–1145

Liu Q, Liu B, Zhang Y, Kim DS, Gao Z (2016) Improving opinion aspect extraction using semantic similarity and aspect associations. In: Proceedings of the thirtieth AAAI conference on artificial intelligence. AAAI Press, pp 2986–2992

Matsumoto S, Takamura H, Okumura M (2005) Sentiment classification using word sub-sequences and dependency sub-trees. In: Proceedings of the Pacific-Asia conference on knowledge discovery and data mining (PAKDD). Springer, Hanoi, pp 301–310

McDonald R, Hannan K, Neylon T, Wells M, Reynar J (2007) Structured models for fine-to-coarse sentiment analysis. In: Proceedings of the annual meeting of the Association of Computational Linguistics (ACL), Prague, pp 432–439

Moe WW, Schweidel DA (2012) Online product opinions: incidence, evaluation, and evolution. Mark Sci 31(3):372–386

Moreno MA, Jelenchick LA, Egan KG, Cox E, Young H, Gannon KE, Becker T (2011) Feeling bad on facebook: depression disclosures by college students on a social networking site. Depress Anxiety 28(6): 447–455

Nguyen TH, Shirai K, Velcin J (2015) Sentiment analysis on social media for stock movement prediction. Expert Syst Appl 42(24):9603–9611

Pak A, Paroubek P (2010) Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of LREc, vol 10

Pang B, Lee L (2004) A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, p 271

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on empirical methods in natural language processing, vol 10. Association for Computational Linguistics, pp 79–86

Park S, Lee SW, Kwak J, Cha M, Jeong B (2013) Activities on facebook reveal the depressive state of users. J Med Internet Res 15(10):e217

Qiu L, Zhang W, Hu C, Zhao K (2009) SELC: a self-supervised model for sentiment classification. In: Proceeding of the 18th ACM conference on information and knowledge management (CIKM), Hong Kong, pp 929–936

Read J, Carroll J (2009) Weakly supervised techniques for domain-independent sentiment classification. In: Proceeding of the 1st international CIKM workshop on topic-sentiment analysis for mass opinion, Hong Kong, pp 45–52

Ruder TD, Hatch GM, Ampanozi G, Thali MJ, Fischer N (2011) Suicide announcement on facebook. Crisis 32(5):280–282

Sarner A, Thompson E, Drakos N, Fletcher C, Mann J, Maoz M (2011) Magic quadrant for social crm. Gartner, Stamford

Si J, Mukherjee A, Liu B, Li Q, Li H, Deng X (2013) Exploiting topic based twitter sentiment for stock prediction. ACL 51(2):24–29

Si J, Mukherjee A, Liu B, Pan SJ, Li Q, Li H (2014) Exploiting social relations and sentiment for stock prediction. In: Proccedings of EMNLP, vol 14, pp 1139–1145

Speriosu M, Sudan N, Upadhyay S, Baldridge J (2011) Twitter polarity classification with label propagation over lexical links and the follower graph. In: Proceedings of the first workshop on unsupervised learning in NLP. Association for Computational Linguistics, Edinburgh, pp 53–63

Tan C, Lee L, Tang J, Jiang L, Zhou M, Li P (2011) User-level sentiment analysis incorporating social networks. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, San Diego, 21–24 Aug 2011, pp 1397–1405

Thom J, Millen DR (2012) Stuff ibmers say: Microblogs as an expression of organizational culture. In: ICWSM

Turney PD (2002) Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, Philadelphia, pp 417–424

Turney PD, Littman ML (2002) Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical Report EGB-1094, National Research Council Canada

Uhl MW (2014) Reuters sentiment and stock returns. J Behav Financ 15(4):287–298

Wan S, Paris C (2014) Improving government services with social media feedback. In: Proceedings of the 19th international conference on intelligent user interfaces. ACM, Haifa, pp 27–36

Whitelaw C, Garg N, Argamon S (2005) Using appraisal groups for sentiment analysis. In: Proceedings of the ACM international conference on information and knowledge management (CIKM), Bremen, pp 625–631. https://doi.org/10.1145/1099554.1099714

Yellowlees PM, Cook JN (2006) Education about hallucinations using an internet virtual reality system: a qualitative survey. Acad Psychiatry 30(6):534–539

Zagibalov T, Carroll J (2008a) Automatic seed word selection for unsupervised sentiment classification of Chinese text. In: Proceedings of the 22nd international conference on computational linguistics (COLING), Manchester, pp 1073–1080

Zagibalov T, Carroll J (2008b) Unsupervised classification of sentiment and objectivity in chinese text. In: Proceedings of the third international joint conference on natural language processing, Hyderabad, pp 304–311

Zaidan O, Eisner J, Piatko C (2007) Using annotator rationales to improve machine learning for text categorization. In: Proceedings of NAACL-HLT, Rochester, pp 260–267

Sentiment Analysis in Social Media, Aspect Extraction for

Lei Zhang[1] and Bing Liu[2]
[1]LinkedIn, Sunnyvale, CA, USA
[2]Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA

Synonyms

Entity extraction; Feature extraction; Target extraction

Glossary

Aspect-based sentiment analysis	Extract and summarize opinions on entities and aspects of entities from text
Sentiment analysis or opinion mining	Computational study of people’s opinions, sentiments, appraisals, attitudes, and emotions from text
Sentiment words or opinion words	Words bearing positive or negative sentiment
User-generated content	Contents created by users of social media, such as product reviews, forum discussions, blogs, and tweets

Definition

Aspect-based sentiment analysis is the computational study of people’s opinions, sentiments, appraisals, attitudes, and emotions toward entities and their aspects expressed in text.

Introduction

With the rapid growth of social media, sentiment analysis, also called opinion mining, has become one of the most active research areas in natural language processing, which also has widespread applications from business analytics to social study. Researchers have studied sentiment analysis at the document, sentence, and aspect levels. Because mining opinions at aspect level can provide more useful knowledge, aspect-based sentiment analysis is often desired and utilized in practical applications. Aspect extraction is one of the core tasks of aspect-based sentiment analysis. In this chapter, we provide a broad overview of aspect extraction for sentiment analysis and introduce related extraction techniques.

Key Points

In a nutshell, aspect extraction is to identify entities and aspects of entities as opinion targets from opinion documents. The task belongs to both the fields of sentiment analysis and information extraction. Due to special characteristics of opinion documents, many domain-specific approaches to aspect extraction have proposed in research and utilized in practical applications. Note that some researchers use the term *feature* to mean aspect and the term *object* to mean entity. Some others do not distinguish aspects and entities and call both of them opinion targets, topics, or simply attributes.

Historical Background

Sentiment and opinion and their related concepts, such as evaluation, appraisal, attitude, affect, emotion, and mood, are about people's subjective beliefs and feelings. They are key influencers of human behaviors. Whenever people need to make a decision, he or she often seeks out others' opinions. Likewise, organizations need to analyze public's opinions to take proper actions.

The development of sentiment analysis coincides with the growth of social media. For the first time in human history, we now possess a huge volume of opinion data recorded in digital forms, such as product reviews, forum discussions, blogs and tweets. These *user-generated contents* (UGC) are full of people's opinions. Mining useful knowledge from these corpora leads to the task of sentiment analysis. Since early 2000, sentiment analysis has been one of the most active research areas in natural language processing (NLP) (Pang and Lee 2008; Liu 2015). The research and applications have also spread from computer science to management science and social sciences because of its importance to business and society as well.

Generally, researchers have studied sentiment analysis at three levels of granularity: document level, sentence level, and aspect level. Document level sentiment classification (Pang et al. 2002; Turney 2002) classifies an opinionated document (e.g., a product review) as expressing an overall positive or negative opinion, assuming the document is a basic information unit of opinion. While at the sentence level, sentiment classification is applied to individual sentences in a document (Wiebe et al. 2004; Wilson et al. 2005). Each sentence cannot be assumed to be opinionated. Therefore, one often first classifies a sentence as opinionated or not opinionated, which is called subjectivity classification. The resulting opinionated sentences are then classified as expressing positive or negative opinions.

Although sentiment analysis at the document level and the sentence level is useful in many cases, it lacks the capacity for fine-grained analysis. A positive evaluative text on a particular entity does not mean that the author has positive opinions on every aspect of the entity. Likewise, a negative evaluative text for an entity does not mean that the author dislikes everything about the entity. To obtain in-depth sentiment analysis, one needs to delve into the aspect level, which gives aspect-based sentiment analysis (first called the feature-based sentiment analysis in (Hu and Liu 2004)).

Aspect-Based Sentiment Analysis

Technically, aspect-based sentiment analysis is to extract and summarize sentiment quadruple (s, g, h, t) from opinion document, which covers four components: sentiment orientation s, sentiment target g, opinion holder h, and time t. *Sentiment* is the underlying feeling, attitude, evaluation, or emotion associated with an opinion. Sentiment orientation can be *positive*, *negative*, or *neutral*. *Sentiment target*, also known as the *opinion target*, is an *entity* or an *aspect* of the *entity* that the sentiment has been expressed upon. *Opinion holder* is an individual or organization that holds an opinion. *Time* is when the opinion is expressed. In some applications, we can decompose opinion target g to entity e and aspect a for more fine-grained analysis. Thus, the quadruple becomes quintuple (s, e, a, h, t), in which e and a represents entity and aspect, respectively.

We use the following camera review as an example (an ID number is associated with each sentence for easy reference):

> Posted by John Smith Date: September 10, 2011
>
> (1) *I bought a Canon G12 camera six months ago.* (2) *I simply love it.* (3) *The picture quality is amazing.* (4) *The battery life is also long.* (5) *However, my wife thinks it is too heavy for her.*

Given the review, the task of aspect-based sentiment analysis aims to extract the following sentiment quintuples from sentences 2, 3, 4, and 5, respectively:

(*positive*, *Canon G12 camera*, *GENERAL*, John Smith, *2011/09/10*)
(*positive*, *Canon G12 camera*, *picture quality,* John Smith, *2011/09/10*)
(*positive*, *Canon G12 camera*, *battery life,* John Smith, *2011/09/10*)
(*negative*, *Canon G12 camera, weight,* John Smith*'s wife*, *2011/09/10*)

The entity is "Canon G12 camera" and the aspects are "picture quality," "battery life," and "weight" of "Canon G12 camera". The GENERAL represents the entity itself. An aspect can be *explicit* (e.g., "battery life*"*) or *implicit* (e.g., "weight" is indicated by "heavy").

An opinion from a single opinion holder is usually not actionable in opinion mining application. The user often needs opinions from a large number of opinion holders, which leads to opinion summary. A summary of opinions is normally constructed based on statistics of extracted sentiment quintuples. Aspect-based opinion summary has been employed in many scenarios, especially in E-commerce.

Aspect Extraction Approaches

Aspect extraction is one of the core tasks of aspect-based sentiment analysis, which needs to identify and extract opinion targets (aspect or entity) from opinion documents. We do not differentiate aspect extraction and entity extraction here, because aspect extraction and entity extraction are closely related tasks and ideas and methods proposed for aspect extractions can be applied to the entity extraction task as well. Currently, aspect extraction methods can be roughly grouped into four categories: frequency, syntactic rules, sequence learning models, and topic modeling. Recently, a new machine-learning framework, *lifelong machine learning*, is proposed to improve the performance of syntactic rules and topic modeling approaches.

Frequency

Frequency-based approach is simple but effective, which mines frequent nouns or noun phrases as aspects. In product reviews, it is observed that people often use the same words when they comment on the same product aspects. Hu and Liu (2004) made use of this observation to mine aspects using frequent itemset mining (Agrawal and Srikant 1994). Naturally, more frequent noun phrases are also more likely to be important aspects, since people usually comment on those important aspects many times. Blair-Goldensohn et al. (2008) refined the approach by considering mainly those noun phrases that are in sentiment-bearing sentences or in some syntactic patterns which indicate sentiments. Several filters were applied to remove unlikely aspects, for example, dropping aspects that do not have sufficient mentions alongside known sentiment words. The frequency-based idea was also utilized in Popescu and Etzioni (2005), Ku et al. (2006), Moghaddam and Ester (2010), and Zhu et al. (2009).

Syntactic Rules

This approach is to exploit syntactic relations of sentiment words and their aspects to identify aspects, which assumes opinion always has target, and there are often syntactic relations that connect sentiment words and targets in a sentence. Such idea was first embodied by Hu and Liu (2004). They found that adjective sentiment words often modify noun aspects in product reviews and proposed to use relations between aspects and opinion words to identify infrequent aspects. That is, product aspects identified by frequency-based method are used to find opinion words first, which are then applied to identify infrequent aspects.

The using of modifying relationships of opinion words and aspects to extract aspects can be generalized to using dependency relations. Zhuang et al. (2006) employed the dependency grammar to generate language templates, which extract aspect and sentiment word pairs from movie reviews by template matching. In Wu et al. (2009), a phrase dependency parser was used for extracting noun phrases and verb phrases as aspect candidates. Unlike a normal dependency parser that identifies dependency of individual words only, a phrase dependency parser identifies dependency of phrases. Dependency relations have also been exploited by Kessler and Nicolov (2009).

Wang and Wang (2008) found that by utilizing the relation between product aspect and opinion word, they can identify both of them simultaneously. Given a list of seed opinion words, a bootstrapping method is employed to identify product aspects and opinion words in an alternating fashion. Mutual information is utilized to measure association between potential aspects and opinion words and vice versa. In addition, linguistic rules are extracted to identify infrequent aspects and opinion words. The similar bootstrapping idea is also utilized in (Hai et al. 2012).

Qiu et al. (2011) further developed above ideas and proposed an algorithm called *double propagation* (DP). DP uses a set of manually compiled dependency rules derived from some dependency relations to identify both aspects and opinion words simultaneously through a bootstrapping process. That is, sentiment words can be recognized by identified aspects and aspects can be identified by known sentiment words. The extracted sentiment words and aspects are utilized to identify new sentiment words and new aspects, which are used again to extract more sentiment words and aspects. DP is pretty effective and scales well for different domains. However, DP may have low precision for large corpora and low recall for small corpora. The reason is that the patterns based on direct dependencies have a large chance of introducing noises for large corpora and the patterns are limited for small corpora. Zhang et al. (2010) proposed an approach to alleviate the issues, which consists of two steps: aspect extraction and aspect ranking. For aspect extraction, it still adopts DP to populate aspect candidates. However, some new linguistic patterns are introduced to increase recall. After extraction, it ranks aspect candidates by aspect importance. High-rank aspect candidate are more likely to be important and genuine aspects. Liu et al. (2015) proposed an automated rule set selection/learning method to improve DP as well.

Liu et al. (2012) also utilized the relation between opinion word and aspect to perform extraction. However, they formulated the opinion relation identification between aspects and opinion words as a word alignment task. They employed the word-based translation model (Brown et al. 1993) to perform monolingual word alignment. Basically, the associations between aspects and opinion words are measured by translation probabilities, which can capture opinion relations between opinion words and aspects more precisely and effectively than linguistic rules or patterns.

Popescu and Etzioni (2005) proposed a method to extract product aspects by utilizing a discriminator relation in context, i.e., the relation between aspects and product class. They first extract noun phrases with high frequency from reviews as candidate product aspects. Then they evaluate each candidate by computing a pointwise mutual information (PMI) score between the candidate and some meronymy discriminators associated with the product class. For example, for “scanner”, the meronymy discriminators for the scanner class are patterns such as “of scanner,” “scanner has,” “scanner comes with,” etc.

Kobayashi et al. (2007) proposed an approach to extract aspect evaluation (aspect-opinion expression) and aspect of relations from blogs, which also makes use of association between aspect, opinion expression, and product class. For example, in aspect-evaluation pair extraction, evaluation expression is first determined by a dictionary lookup. Then, syntactic patterns are employed to find its corresponding aspect to form the candidate pair. The candidate pairs are tested and validated by a classifier, which is trained by incorporating two kinds of information: contextual and statistical clues in corpus. The contextual clues are syntactic relations between words in a sentence, which can be determined by the dependency grammar, and the statistical clues are normal co-occurrences between aspects and evaluations.

Sequence Learning Models

Sequence learning models such as hidden Markov models (HMM) and conditional random fields (CRF) are widely used in information extraction. Aspect extraction can be regarded as a sequence labeling task since entity, aspect, and opinion expressions are often interdependent and occur in a sentence sequence. Jin and Ho (2009) utilized lexicalized HMM to extract product aspects and opinion expressions from reviews. Different from traditional HMM, they integrated linguistic features such as part-of-speech and lexical patterns into HMM.

One limitation of HMM is that its assumptions may not be adequate for real-life problems, which leads to reduced performance. To address the limitation, linear-chain conditional random fields (CRF) are proposed as an undirected sequence model, which allows relaxation of the strong independence assumptions made by HMM. Jakob and Gurevych (2010) utilized CRF to extract opinion aspects from opinion sentences. They employed features such as token, part-of-speech, short dependency path, and word distance as input for the CRF-based approach. They adopted the inside-outside-begin (IOB) labeling schema: B-Target, identifying the beginning of an opinion target; I-Target, identifying the continuation of a target, and O for other (nontarget) tokens.

Similar work has been done in Li et al. (2010). In order to model the long-distance dependency with conjunctions (e.g., “and,” “or,” “but”) at the sentence level and deep syntactic dependencies for aspects, positive opinions and negative opinions, they used the skip-tree CRF models to detect product aspects and opinions.

Topic Modeling

Topic models are based on the idea that documents are mixtures of topics, and each topic is a probability distribution over words. Naturally, topic models can be applied to aspect extraction. We can deem that each aspect is a unigram language model, i.e., a multinomial distribution over words. Although such a representation is not as easy to interpret as aspects, its advantage is that different words expressing the same or related aspects (more precisely aspect expressions) can be automatically grouped together under the same aspect. Currently, a great deal of research has been done on aspect extraction using topic models. They basically adapted and extended the probabilistic latent semantic analysis (pLSA) model and the latent Dirichlet allocation (LDA) model.

In the context of aspect extraction, aspects are basically topics in topic modeling. Mei et al. (2007) proposed a model for extracting both aspects and sentiment words. Titov and McDonald (2008) pointed out that global topic models such as PLSA and LDA might not be suitable for detecting aspects from reviews. To tackle this problem, they proposed some multi-grain topic models to discover aspects, which models two distinct types of topics: global topics and local topics. Lin and He (2009) proposed a joint topic-sentiment model, which extended LDA by adding a sentiment layer. It detects sentiment and aspect simultaneously from the corpus. Further works along a similar line have been done in Brody and Elhadad (2010), Wang et al. (2010), Zhao et al. (2010), Jo and Oh (2011), and Diao et al. (2014).

Improvements by Lifelong Machine Learning

Although effective and scalable for different domains, syntactic rules and topic modeling still have room for major improvements. For example, for aforementioned syntactic rule-based method DP, it has low precision and low recall issues due to the flexibility of natural languages. One good way to improve is to use the prior knowledge in a new machine-learning framework, namely, *lifelong machine learning*.

Lifelong machine learning retains its past experiences and learned results as knowledge and uses it to help new learning to extraction. In other words, if the system already knows a lot before extraction, it clearly can do much better. The prior knowledge is mined automatically by exploiting the abundance of reviews for all kinds of products on the Web. This idea is workable because many products actually share aspects, for example, many electronic products have aspects *screen* and *battery*.

Such idea can be utilized to improve results of topic modeling. Mukherjee and Liu (2012) proposed *knowledge-based models* to exploit prior domain knowledge to produce better results. Chen and Liu (2014) proposed *lifelong topic models*, which exploit the big data to automatically mine prior knowledge to be used in the modeling process.

Lifelong learning can also be employed for syntactic rule-based methods. Liu et al. (2016) proposed a recommendation-based approach to improve the DP method based on the framework of lifelong learning. Two forms of recommendation were presented, which exploited two types of interesting knowledge about aspects: aspect similarity and association.

Miscellaneous Methods

Yi et al. (2003) proposed a method for aspect extraction based on the likelihood-ratio test. Bloom et al. (2007) manually built a taxonomy of aspects, which indicates aspect types. They also constructed an aspect list by starting with a sample of reviews that the list would apply to. They examined the seed list manually and used WordNet to suggest additional terms to add to the list. Lu et al. (2010) exploited the online ontology Freebase to obtain aspects of a topic and used them to organize scattered opinions and to generate structured opinion summaries. Ma and Wan (2010) exploited Centering theory to extract opinion targets from news comments. The approach uses global information in news articles as well as contextual information in adjacent sentences of comments. Ghani et al. (2006) formulated aspect extraction as a classification problem and used both traditional supervised learning and semi-supervised learning methods to extract product aspects. Yu et al. (2011a) used a partially supervised learning method called one-class SVM to extract aspects. Using one-class SVM, one only needs to label some positive examples, which are aspects. In their case, they only extracted aspects from pros and cons of reviews. Li et al. (2012) formulated aspect extraction as a shallow semantic parsing problem. A parse tree is built for each sentence, and structured syntactic information within the tree is used to identify aspects.

Aspect Grouping and Hierarchy

It is common that people use different words and expressions to describe the same aspect. For example, photo and picture refer to the same aspect in digital camera reviews. Although topic models can identify and group aspects to some extent, the results are not fine-grained because such models are based on word co-occurrences rather than word semantic meanings. As a result, a topic is often a list of related words about a general topic rather than a set of words referring to the same aspect. For example, a topic about battery may contain words like “life,” “battery,” “charger,” “long,” and “short.” We can clearly see that these words do not mean the same thing, although they may co-occur frequently. Alternatively, we can extract aspect expressions first and then group them into different aspect categories.

Grouping aspect expressions indicating the same aspect are essential for opinion mining applications. Although WordNet and other thesaurus dictionaries can help, they are far from sufficient due to the fact that many synonyms are domain dependent. For example, “picture” and “movie” are synonyms in movie reviews, but they are not synonyms in digital camera reviews as picture is more related to photo while movie refers to video. It is also important to note that although most aspect expressions of an aspect are domain synonyms, they are not always synonyms. For example, “expensive” and “cheap” can both indicate the aspect price, but they are not synonyms of price.

Liu et al. (2005) attempted to solve the problem by using the WordNet synonym sets, but the results were not satisfactory because WordNet is not sufficient for dealing with domain dependent synonyms. Carenini et al. (2005) also proposed a method to solve this problem in the context of opinion mining. Their method is based on several similarity metrics defined using string similarity, synonyms, and distances measured using WordNet. However, it requires taxonomy of aspects to be given beforehand for a particular domain. The algorithm merges each discovered aspect expression to an aspect node in the taxonomy. Guo et al. (2009) proposed a multilevel latent semantic association technique (called mLSA) to group product aspect expressions. At the first level, all the words in product aspect expressions are grouped into a set of concepts/ topics using LDA. The results are used to build some latent topic structures for product aspect expressions. At the second level, aspect expressions are grouped by LDA again according to their latent topic structures produced from level 1 and context snippets in reviews.

Zhai et al. (2010) proposed a semi-supervised learning method to group aspect expressions into the user-specified aspect groups or categories. Each group represents a specific aspect. To reflect the user needs, they first manually label a small number of seeds for each group. The system then assigns the rest of the discovered aspect expressions to suitable groups using semi-supervised learning based on labeled seeds and unlabeled examples. The method used the expectation-maximization (EM) algorithm. Two pieces of prior knowledge were used to provide a better initialization for EM, i.e., (1) aspect expressions sharing some common words are likely to belong to the same group and (2) aspect expressions that are synonyms in a dictionary are likely to belong to the same group. Zhai et al. (2011) further proposed an unsupervised method, which does not need any pre-labeled examples. Additionally, it is enhanced with lexical (or WordNet) similarities. The algorithm also exploited a piece of natural language knowledge to extract more discriminative distributional context to help grouping.

Mauge et al. (2012) used a maximum entropy-based clustering algorithm to group aspects in a product category. It first trains a maximum-entropy classifier to determine the probability p that two aspects are synonyms. Then, an undirected weighted graph is constructed. Each vertex represents an aspect. Each edge weight is proportional to the probability p between two vertices. Finally, approximate graph partitioning methods are employed to group product aspects.

Zhao et al. (2014) proposed a concept called *sentiment distribution consistency* for aspect clustering, which states that different words or phrases (e.g., “price,” “money,” “worth,” etc.) of the same aspect tend to have consistent sentiment distribution. Then, they formalize the concept as a soft constraint, integrate the constraint into a context-based probabilistic model, and then solve the problem in the posterior regularization framework. An EM-like two-stage optimization algorithm is utilized.

Closely related to aspect grouping, aspect hierarchy is to present product aspects as a tree or hierarchy. The root of the tree is the name of the entity. Each non-root node is a component or subcomponent of the entity. Each link is a part of relation. Each node is associated with a set of product aspects. Yu et al. (2011b) proposed a method to create aspect hierarchy. The method starts from an initial hierarchy and inserts the aspects into it one by one until all the aspects are assigned. Each aspect is inserted into the optimal position by semantic distance learning. Wei and Gulla (2010) studied sentiment analysis based on aspect hierarchy trees.

Identifying Implicit Aspects

Sometimes, an aspect is presented implicitly. For example, the sentence “The show ticket is cheap” implies the aspect *price*. There are many types of implicit aspect expressions. Adjectives are perhaps the most common type. Many adjectives modify or describe some specific attributes or properties of entities. For example, the adjective “heavy” usually describes the aspect weight of an entity. “Beautiful” is normally used to describe (positively) the aspect look or appearance of an entity. By no means, however, does this say that these adjectives only describe such aspects. Their exact meanings can be domain dependent. For example, “heavy” in the sentence “The traffic is heavy” does not describe the weight of the traffic. Note that some implicit aspect expressions are very difficult to extract and to map, for example, "*fit in pockets*" in the sentence "This phone will not easily fit in pockets."

Limited research has been done on identifying implicit aspects. People often try to map implicit aspects to their explicit aspects. In Su et al. (2008), a clustering method was proposed to map implicit aspect expressions, which were assumed to be sentiment words, to their corresponding explicit aspects. The method exploits the mutual reinforcement relationship between an explicit aspect and a sentiment word forming a co-occurring pair in a sentence. Such a pair may indicate that the sentiment word describes the aspect, or the aspect is associated with the sentiment word. The algorithm finds the mapping by iteratively clustering the set of explicit aspects and the set of sentiment words separately. In each iteration, before clustering one set, the clustering results of the other set is used to update the pairwise similarity of the set. The pairwise similarity in a set is determined by a linear combination of intra-set similarity and interset similarity. The intra-set similarity of two items is the traditional similarity. The interset similarity of two items is computed based on the degree of association between aspects and sentiment words. The association (or mutual reinforcement relationship) is modeled using a bipartite graph. An aspect and an opinion word are linked if they have co-occurred in a sentence. The links are also weighted based on the co-occurrence frequency. After iterative clustering, strong links between aspects and sentiment word groups form the mapping.

In Hai et al. (2011), a two-phase co-occurrence association rule mining approach was proposed to match implicit aspects (which are also assumed to be sentiment words) with explicit aspects. In the first phase, the approach generates association rules involving each sentiment word as the condition and an explicit aspect as the consequence, which co-occur frequently in sentences of a corpus. In the second phase, it clusters the rule consequents (explicit aspects) to generate more robust rules for each sentiment word mentioned above. For application or testing, given a sentiment word with no explicit aspect, it finds the best rule cluster and then assigns the representative word of the cluster as the final identified aspect. Fei et al. (2012) focused on finding implicit aspects (mainly nouns) indicated by opinion adjectives, for example, to identify price, cost, etc. for adjective expensive. A dictionary-based method was proposed, which tries to identify attribute nouns from the dictionary gloss of the adjective. They formulated the problem as a collective classification problem, which can exploit lexical relations of words (e.g., synonyms, antonyms, hyponym, and hypernym) for classification. Some other related work for implicit aspect identification includes those in Wang and Wang (2008), Yu et al. (2011b), and Zhang and Liu (2011).

Key Applications

Aspect-based sentiment analysis, which aims to obtain detailed information about opinions, has attracted a great deal of attention from both the research community and the industry. It has widespread applications in many domains, especially in social media analytics and E-commerce. The former is to listen and analyze public opinions on specific entities mentioned in social media channels. The results can help organizations make informed decisions. For the latter, aspect-based sentiment analysis can provide better shopping experiences for consumers. As there are a huge number of reviews for almost any kind of products at E-commerce websites, consumers can easily understand pros and cons of almost any products, which enables them to make the right purchase decisions. Besides, businesses can also benefit from aspect-based sentiment analysis as it can help them to better monitor product quality and understand the needs of the market. In recent years, applications have also spread to social, political, health, and medical domains. For all these applications, aspect extraction is critical. Without knowing entities and aspects of entities as opinion targets, the mined opinions or sentiments are of little use.

Future Directions

In this chapter, we reviewed some representative works for aspect extraction from opinion documents. Although significant progresses have been made, the problem remains to be challenging. Semi-supervised and unsupervised methods with the help of lifelong machine learning are promising. We expect that future work will improve the performance significantly.

Cross-References

- ▶ Sentiment Analysis in Social Media
- ▶ Sentiment Analysis of Microblogging Data
- ▶ Sentiment Analysis of Reviews
- ▶ Sentiment Analysis, Basic Tasks of

Acknowledgments Bing Liu's work was partially supported by the National Science Foundation under the grants IIS-1407927 and IIS-1650900.

References

Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the international conference on very large databases (VLDB-1994)

Blair-Goldensohn S, Hannan K, McDonald R, Neylon T, Reis GA, Reyna J (2008) Building a sentiment summarizer for local service reviews. In: Proceedings of international conference on world wide web workshop of NLPIX (WWW-NLPIX-2008)

Bloom K, Grag N, Argamon S (2007) Extracting appraisal expressions. In: Proceedings of the 2007 annual conference of the North American chapter of the ACL (NAACL-2007)

Brody S, Elhadad S (2010) An unsupervised aspect-sentiment model for online reviews. In: Proceedings of the annual conference of the North American chapter of the ACL(NAACL-2010)

Brown FP, Della Pietra SA, Della Pietra VJ, Mercer RL (1993) The mathematics of statistical machine translation: parameter estimation. Computational Linguist 19 (2):263–311

Carenini G, Ng R, Zwart E (2005) Extracting knowledge from evaluative text. In: Proceedings of third international conference on knowledge capture (K-CAP-2005)

Chen Z, Liu B (2014) Topic modeling using topics from many domains, lifelong learning and big data. In: Proceedings of the international conference on machine learning (ICML-2014)

Diao Q, Qiu M, Wu C, Smola A, Jiang J, Wang C (2014) Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS). In: Proceedings of ACM SIGKDD international conference on knowledge discovery and data mining (KDD-2014)

Fei G, Liu B, Hsu M, Castellanos M, Ghosh R (2012) A dictionary-based approach to identifying aspects implied by adjectives for opinion mining. In: Proceedings of international conference on computational linguistics (COLING-2012)

Ghani R, Probst K, Liu Y, Krema M, Fano A (2006) Text mining for product attribute extraction. SIGKDD Explorations 1(8):41–48

Guo H, Zhu H, Guo Z, Zhang X, Su Z (2009) Product feature categorization with multilevel latent semantic association. In: Proceedings of ACM international conference on information and knowledge management (CIKM-2009)

Hai Z, Chang K, Kim J (2011) Implicit feature identification via co-occurrence association rule mining. In: Computational linguistic and intelligent text processing

Hai Z, Chang K, Cong G (2012) One seed to find them all: mining opinion features via association. In: Proceedings of ACM international conference on information and knowledge management (CIKM-2012)

Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of ACM SIGKDD international conference on knowledge discovery and data mining (KDD-2004)

Jakob N, Gurevych I (2010) Extracting opinion targets in a single- and cross-domain setting with conditional random fields. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP-2010)

Jin W, Ho HH (2009) A novel lexicalized HMM-Based learning framework for web opinion mining. In: Proceedings of international conference on machine learning (ICML-2009)

Jo Y, Oh A (2011) Aspect and sentiment unification model for online review analysis. In: Proceedings of the conference on web search and web data mining (WSDM-2011)

Kessler J, Nicolov N (2009) Targeting sentiment expressions through supervised ranking of linguistic configurations. In: Proceedings of the inter-national AAAI conference on weblogs and social media (ICWSM-2009)

Kobayashi N, Inui K, Matsumoto Y (2007) Extracting aspect-evaluation and aspect-of relations in opinion mining. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-2007)

Ku L, Liang Y, Chen H (2006) Opinion extraction, summarization and tracking in news and blog corpora. In: Proceedings of AAAI-CAAW-2006

Li F, Han C, Huang M, Zhu X, Xia Y, Zhang S, Yu H (2010) Structure-aware review mining and summarization. In: Proceedings of international conference on computational linguistics (COLING-2010)

Li S, Wang R, Zhou G (2012) Opinion target extraction using a shallow semantic parsing framework. In: Proceedings of national conference on artificial intelligence (AAAI-2012)

Lin C, He Y (2009) Joint sentiment/topic model for sentiment analysis. In: Proceedings of ACM international conference on information and knowledge management (CIKM-2009)

Liu B (2015) Sentiemnt analysis: mining opinions, sentiments, and emotions. Cambridge University Press, Cambridge

Liu B, Hu M, Cheng J (2005) Opinion observer: analyzing and comparing opinions on the web. In: Proceedings of international conference on world wide web (WWW-2005)

Liu K, Xu L, Zhao J (2012) Opinion target extraction using word-based translation model. In: Proceeding of conference on empirical methods in natural language processing (EMNLP-2012)

Liu K, Xu L, Zhao J (2013) Syntactic patterns versus word alignment: extracting opinion targets from online reviews. In: Proceedings of the annual meeting of the association for computational linguistics (ACL-2013)

Liu Q, Gao Z, Liu B, Zhang Y (2015) Automated rule selection for aspect extraction in opinion mining. In: Proceedings of international joint conference on artificial intelligence (IJCAI-2015)

Liu Q, Liu B, Zhang Y, Kim D, Gao Z (2016) Improving opinion aspect extraction using semantic similarity and aspect associations. In: Proceedings of the thirtieth AAAI conference on artificial intelligence (AAAI-2016)

Lu Y, Duan H, Wang H, Zhai C (2010) Exploiting structured ontology to organize scattered online opinions. In: Proceedings of international conference on computational linguistics (COLING-2010)

Mauge K, Rohanimanesh K, Ruvini JD (2012) Structuring e-commerce inventory. In: Proceedings of annual meeting of the association for computational linguistics (ACL-2012)

Mei Q, Ling X, Wondra M, Su H, Zhai C (2007) Topic sentiment mixture: modeling facets and opinions in weblogs. In: Proceedings of international conference on world wide web (WWW-2007)

Moghaddam, S., M. Ester (2010) Opinion digger: an unsupervised opinion miner from unstructured product reviews. In: Proceedings of ACM international conference on information and knowledge management (CIKM-2010)

Mukherjee A, Liu B (2012) Aspect extraction through semi-supervised modeling. In: Proceedings of the annual meeting of association for computational linguistics (ACL-2012)

Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of conference on empirical methods in natural language processing (EMNLP-2002)

Popescu A, Etzioni O (2005) Extracting product features and opinions from reviews. In: Proceedings of conference on empirical methods in natural language processing (EMNLP-2005)

Qiu G, Liu B, Bu J, Chen C (2011) Sentiment word expansion and target extraction through double propagation. Comput Linguist

Su Q, Xu X, Guo H, Guo Z, Wu X, Zhang X, Swen B, Su Z (2008) Hidden sentiment association in Chinese web opinion mining. In: Proceedings of international conference on world wide web (WWW-2008)

Titov I, McDonald R (2008) Modeling online reviews with multi-grain topic models. In: Proceedings of international conference on world wide web (WWW-2008)

Turney PD (2002) Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the annual meeting of the association for computational linguistics (ACL-2002)

Wang B, Wang H (2008) Bootstrapping both product features and opinion words from Chinese customer reviews with cross-inducing. In: Proceedings of the international joint conference on natural nanguage processing (IJCNLP-2008)

Wang H, Lu Y, Zhai C (2010) Latent aspect rating analysis on review text data: a rating regression approach. In: Proceedings of ACM SIGKDD international conference on knowledge discovery and data mining (KDD-2010)

Wei W, Gulla JA (2010) Sentiment learning on product reviews via sentiment ontology tree. In: Proceedings of annual meeting of the association for computational linguistics (ACL-2010)

Wiebe J, Wilson T, Bruce R, Bell M, Martin M (2004) Learning subjective language. Comput Linguist 30(3):277–308

Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the human language technology conference and the conference on empirical methods in natural language processing (HLT/EMNLP-2005)

Wu Y, Zhang Q, Huang X, Wu L (2009) Phrase dependency parsing for opinion mining. In: Proceedings of conference on empirical methods in natural language processing (EMNLP-2009)

Yi J, Nasukawa T, Bunescu R, Niblack W (2003) Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In: Proceedings of international conference on data mining (ICDM-2003)

Yu J, Zha Z, Wang M, Chua T (2011a) Aspect ranking: identifying important product aspects from online consumer reviews. In: Proceedings of annual meeting of the association for computational linguistics (ACL-2011)

Yu J, Zha Z, Wang M, Wang K, Chua T (2011b) Domain-Assisted product aspect hierarchy generation: towards hierarchical organization of unstructured consumer reviews. In: Proceedings of conference on empirical methods in natural language processing (EMNLP-2011)

Zhai Z, Liu B, Xu H, Jia P (2010) Grouping product features using semi-supervised learning with soft-constraints. In: Proceedings of international conference on computational linguistics (COLING-2010)

Zhai Z, Liu B, Xu H, Jia P (2011) Clustering product features for opinion mining. In: Proceedings of ACM international conference on web search and data mining (WSDM-2011)

Zhang L, Liu B (2011) Identifying noun product features that imply opinions. In: Proceedings of the annual meeting of the association for computational linguistics (ACL-2011)

Zhang L, Liu B, Lim S, O'Brien-Strain E (2010) Extracting and ranking product features in opinion documents. In: Proceedings of international conference on computational linguistics (COLING-2010)

Zhao W, Jiang J, Yan H, Li X (2010) Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In: Proceedings of conference on empirical methods in natural language processing (EMNLP-2010)

Zhao L, Huang M, Chen H, Cheng J, Zhu X (2014) Clustering aspect-related phrases by leveraging sentiment distribution consistency. In: Proceedings of conference on empirical methods in natural language processing (EMNLP-2014)

Zhu J, Wang H, Tsou BK, Zhu M (2009) Multi-aspect opinion polling from textual reviews. In: Proceedings of ACM international conference on information and knowledge management (CIKM-2009)

Zhuang L, Jing F, Zhu X (2006) Movie review mining and summarization. In: Proceedings of ACM international conference on information and knowledge management (CIKM-2006)

Sentiment Analysis of Microblogging Data

Pierpaolo Basile[1], Valerio Basile[2],
Malvina Nissim[3], Nicole Novielli[1] and
Viviana Patti[4]
[1]Department of Computer Science, University of Bari Aldo Moro, Bari, Italy
[2]Université Côte d'Azur, Inria, CNRS, Sophia Antipolis, France
[3]University of Groningen, Groningen, The Netherlands
[4]Department of Computer Science, University of Turin, Turin, Italy

Synonyms

Opinion mining of microblogging data

Glossary

Microblogging	Broadcast messaging where posts are constrained to a specific size, e.g., Twitter (140 characters per message)
NLP	Natural language processing, the area of computer science that studies natural language by computational means
Polarity	The characteristic of a subjective message of conveying a positive or negative sentiment. Polarity is typically represented either by discrete classes (e.g., *positive*, *negative*, *neutral*) or on a continuous scale of sentiment ranging from negative to positive
Sentiment analysis	The study of opinion and emotions expressed in natural language

Definition

Sentiment analysis is the task of identifying the subjectivity (neutral vs. emotionally loaded) and the polarity (positive vs. negative semantic orientation) of a text, by exploiting natural language processing, text analysis, and computational linguistics. Sentiment analysis is typically adopted to mine and classify customers' reviews and user-generated contents in social media in several application domains, ranging from customer service to marketing. This contribution focuses on microblogging data. A microblog differs from traditional blogs and other crowd-generated contents in the size of its content, which is limited by the length constraints posed in microblogging platforms. Furthermore, microblogs allow users to exchange small elements of text enriched by images or video. For these reasons, microblogs are interesting to analyze, yet their format makes it a challenge beyond traditional sentiment analysis. Also, the sheer size of data available makes microblog data attractive for the industry.

Introduction

Flourished in the last decade, sentiment analysis is the study of the subjectivity and polarity (positive vs. negative) of a text (Pang and Lee 2008). Traditionally, sentiment analysis techniques have been successfully exploited for opinionated corpora, such as news (Wiebe et al. 2005) or reviews (Hu and Liu 2004). With the worldwide diffusion of social media, sentiment analysis on microblogging (Pak and Paroubek 2010) is now regarded as a powerful tool for modeling socioeconomic phenomena (Jansen et al. 2009; O'Connor et al. 2010).

The success of the tasks of sentiment analysis on Twitter at NLP evaluation campaigns since 2013 (Nakov et al. 2013; Rosenthal et al. 2014, 2015) attests this growing trend (Basile et al. 2014). In a world where e-commerce is part of our everyday life and social media platforms are regarded as new channels for marketing and for fostering trust of potential customers, such great interest in opinion mining from microblogging is not surprising. In fact, one fifth of all tweets refer to opinions about products or brands (Jansen et al. 2009).

In the following, we define the main problems addressed by research on sentiment analysis on microblogging data. We provide overview on the open challenges of sentiment analysis on microblogging data. Finally, we discuss investigation, with particular focus on sentiment analysis of microblogging data containing figurative language and entity-based sentiment analysis of microposts.

Key Points

In the following we will discuss the main aspects, open challenges, and opportunities of sentiment analysis in microblogging data.

First, we will discuss the problems posed to traditional sentiment analysis approaches by microblogging data. Microblog differs from traditional blogs and other crowd-generated contents, mainly in the content size that forces users to adapt their writing style to the limited amount of characters (e.g., using abbreviations, emoji, etc.). Furthermore, microblogs are usually enriched by images, videos, hashtags, and user mentions. In this sense, microblogs are interesting to analyze and represent a challenge beyond classical sentiment analysis approaches developed on longer documents, such as reviews, emails, news, etc.

Second, different tasks can be performed on microblogging data. Traditional tasks involve detecting the subjectivity and polarity of microblogs at the message level, i.e., determining whether a tweet is subjective or not and, if subjective, determining its positive or negative semantic orientation. However, comments and opinions are usually directed toward a specific target or aspect of interest, and in this perspective finer-grained tasks can be envisioned. For instance, *aspect-based sentiment analysis* addresses the problem of identifying the aspects of given target entities and the sentiment expressed for each aspect, while the *stance detection* emerging task focuses on detecting what particular stance a user takes with respect to some specific target, which is particularly interesting for studying political debates in social media. Furthermore, figurative language, including irony and sarcasm, is widely adopted in social media. As a consequence, the accuracy of sentiment analysis can be affected by the use of such devices, which can invert the literal polarity of a text (e.g., irony may involve using positive language to convey a negative opinion). Irony and figurative language detection is becoming paramount in sentiment analysis research, toward the development of irony-aware sentiment analysis systems.

Finally, the sheer size of data available makes microblog data attractive for a huge range of application domains, from industry to politics and from social sciences to software engineering. We will provide an overview of the key applications of sentiment analysis for business, social utility, and research.

Historical Background

Nowadays, affective computing (Picard 1997) is an established research field, and sensing emotions and opinion from text is now regarded as a crucial task for several application domains. Mining emotions and opinions from text involves, first of all, choosing the most appropriate model to operationalize sentiment (Carofiglio et al. 2009). Bidimensional models, such as the circumplex model of affect proposed by Russel (1980), usually represent the sentiment polarity on the *x* axis and its level of activation or intensity on the *y* axis. On the other hand, other theoretical frameworks explicitly deal with discrete emotion labeling of text, by looking for linguistic cues of specific emotions in a limited set. It is the case, for example, of Ekman's model, which defines a set of six basic emotional states *{happy*, *sad*, *angry*, *fearful*, *disgusted*, *and surprised}* or Lazarus' framework (Lazarus 1991), which describes nine negative {*anger*, *anxiety*, *guilt*, *shame*, *sadness*, *envy*, *jealousy*, *disgust*} and six positive {*joy*, *pride*, *love*, *relief*, *hope*, *compassion*} emotions.

Among the various possibilities of modeling and recognizing the sentiment conveyed by a text, sentiment analysis research is mainly concerned with the classification of text according to the expressed polarity, i.e., along the valence dimension (i.e., positive vs. negative).

Originally, sentiment analysis had been mainly concerned with the classification of movie reviews and longer documents in general (Wiebe et al. 2005; Wilson et al. 2005). The first shared task on sentiment analysis specifically on Twitter data was organized within the SemEval campaign in 2013 (Nakov et al. 2013), followed by another one in 2014 (Rosenthal et al. 2014). Interest boomed at SemEval 2015 when sentiment analysis became a whole track, and that year involved four different tasks. Three of these concerned tweets: a general classic task on *sentiment analysis in Twitter* (Task 10, (Rosenthal et al. 2015), with four subtasks); a task focused on figurative language, *sentiment analysis of figurative language in Twitter* (Task 11, Ghosh et al. 2015a); and an *aspect-based sentiment analysis* task (Task 12, Pontiki et al. 2015), where systems had to identify aspects of entities and the sentiment expressed for each aspect. We explain such tasks in more details in sections "Traditional Task: Classifying the Overall Polarity of a Text," "Irony and Figurative Language," and "Aspect-Based Sentiment Analysis," respectively.

At the most recent SemEval-2016, in addition to the by now classic sentiment analysis task (Nakov et al. 2016) and the already established aspect-based sentiment analysis task (AL-Smadi et al. 2016), two additional tasks have been organized, targeting emerging challenges: one on detecting *stance* in tweets (Task 6, Mohammad et al. 2016b), i.e., detecting the position of the author with respect to a given target (against/in, favor/neutral), which we review in section "Emerging Task: Stance Detection," and one on determining sentiment intensity (Task 7, Kiritchenko et al. 2016). Interest is not confined to English only, as indeed some of these tasks provide datasets in more than one language. Additionally, dedicated tasks for specific languages have also been organized, for example, for Italian (Basile et al. 2014) and Spanish (Villena-Roman et al. 2015).

Sentiment Analysis of Microblogging Data

Sentiment analysis is the area of natural language processing that studies the sentiments and how they are expressed in written or spoken language. This field is sometimes referred to as *opinion mining* (Pang and Lee 2008), highlighting how it mainly deals with subjective language conveying personal opinions, intended as views, judgments, or appraisals about a particular matter (Fersini et al. 2017). With the worldwide diffusion of social media, a huge amount of textual data has been made available, and sentiment analysis on microblogging is now regarded as a powerful tool for modeling socioeconomic phenomena (Jansen et al. 2009; O'Connor et al. 2010). Dealing with such informal text poses new challenges due to the presence of slang, misspelled words, and microblogging features such as hashtags or links. In the following we provide an overview on the main tasks and approaches.

Data

Microblogging Data

The most prominent source of raw microblogging material is undoubtedly Twitter (https://twitter.com). Twitter (the name is an intentional misspelling of "tweeter," someone who tweets) is an online messaging platform launched in 2006 that stood out from the crowd of alternatives at the time for its characteristic of limiting the length of messages to 140 characters. Users of Twitter broadcast their short messages (*tweets*), which are public by default unless the author's account is explicitly set to private. Users of Twitter (sometimes called *tweeps*) can "follow" each other in a unilateral or bilateral way, thus controlling the stream of tweets that is shown to them when accessing the website or the mobile app. However, all tweets that are not sent by a private account, and that have not been deleted, are accessible through the search facility.

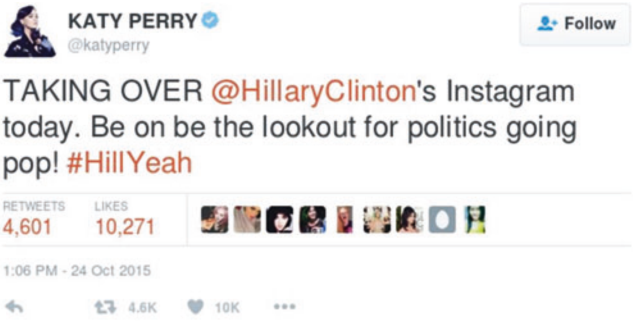
Besides the short text, a tweet contains a certain amount of metadata: information about the user and its social network induced by the following relation, a fine-grain timestamp; geographical coordinates (if the option is activated by the user); what is the language used; images, videos, and web links included in the text; and more. A tweet can also be simply a *retweet*, that is, a propagation of a tweet to one's followers, possibly including an additional comment. Other elements that make Twitter text peculiar are *mentions* (screen names of other Twitter users) and *hashtags*, short strings preceded by the symbol # that are supposed to mark a topic specific to the message. Figure 1 shows an example of a tweet from the account *@KatyPerry*, the most popular on the service at the time of this writing, exhibiting both a mention *@HillaryClinton* and a hashtag *#HillYeah*.

Twitter's Web page and the mobile application are not the only way of accessing its content. Since its inception, the service is founded on a public HTTP API that allows registered users to perform a great deal of operations concerning messages and accounts of Twitter. Several methods are provided to retrieve large quantities of messages, based on searching for keywords, intercepting the continuous stream of messages, or accessing the users' profiles. Each method is more or less effective depending on the task (Basile and Nissim 2013), for instance, use the streaming API to retrieve a fraction of all the messages containing keywords from a fixed list, in real time, in order to build a corpus of Italian tweets.

At the time of this writing, Twitter counts roughly 300 million active users, sending about 350,000 tweets per minute (500 million daily). Statistics are taken from http://www.internetlivestats.com/twitter-statistics/.

Lexical Resources for Sentiment Analysis

Sentiment-related information is often encoded in lexical resources, such as affective lists and corpora. Both sentiment and emotion lexicons, and psycholinguistic resources available for English, refer to various affective models (see section "Historical Background") and capture different nuances of affect, such as sentiment polarity, emotional categories, and emotional dimensions. Such lexicons are usually lists of words to which a positive or negative or emotion-related label



KATY PERRY
@katyperry

TAKING OVER @HillaryClinton's Instagram today. Be on be the lookout for politics going pop! #HillYeah

Sentiment Analysis of Microblogging Data, Fig. 1 Example of message from Twitter (or score) is associated. Classic and widely used lexicons for sentiment are the General Inquirer, the MPQA (Wilson et al. 2005), Bing Liu's Opinion Lexicon (Hu and Liu 2004), AFINN (Nielsen 2011), and a wide library or sentiment and emotion resources developed by the NRC group and available at http://saifmohammad.com/WebPages/lexicons.html, more specifically for emotions: EmoLex and psycholinguistically (LIWC) (Pennebaker et al. 2001), the Dictionary of Affect in Language (DAL, ftp://perceptmx.com/wdalman.pdf) developed by Whissell (2009).

Besides flat lists of affective words, lexical taxonomies have also been enriched with sentiment and/or emotion information. Both SentiWordNet (Esuli and Sebastiani 2006) and WordNet-Affect (Strapparava and Valitutti 2004) are extensions of WordNet (Fellbaum 1998): the former including sentiment information in the form of scores associated to positive, negative, and neutral values associated to each synset and the latter assigning labels representing affective concepts to synsets (Strapparava and Valitutti 2004). A reduced version of WN-Affect also exists (Strapparava and Mihalcea 2007), where a subset of the synsets are labeled with the six emotions from Ekman (1992): joy, fear, anger, sadness, disgust, and surprise. Another sense-level affective lexicon is SentiSense (http://nlp.uned.es/jcalbornoz/SentiSense.html), which attaches emotional meanings to concepts from WordNet (Carrillo de Albornoz et al. 2012). It has been developed semiautomatically using several semantic relations between synsets in WordNet. It is composed by 2,190 synsets tagged with 14 labels from a set of emotional categories: *joy*, *fear*, *surprise*, *anger*, *disgust*, *love*, *anticipation*, *hope*, *despair*, *sadness*, *calmness*, *like*, *hate*, *and ambiguous*. Such categories are also related via an antonym relationship and refer to a merge of models by Magda Arnold (1960), Robert Plutchik (1980), and W. Gerrod Parrot (2001).

In order to match interest in sentiment analysis for languages other than English, and exploiting existing mappings from English WordNet to WordNets in other languages, clones of SentiWordNet have been produced multilingually (e.g., Sentix for Italian (Basile and Nissim 2013)).

Additionally, other resources include and model semantic, conceptual, and affective information associated with multi-word natural language expressions, by enabling concept-level analysis of sentiment and emotions conveyed in texts (SenticNet family).

All such resources represent a rich and varied lexical knowledge about affect, under different perspectives, and virtually all sentiment analysis systems that detect polarity of microtexts incorporate lexical information derived from them. For a comprehensive description and an evaluation of the different ways lexicons are employed in sentiment analysis systems, see Nissim and Patti (2017) and Taboada et al. (2011).

Tasks

In this section we review the main tasks associated with sentiment analysis, ranging from a general polarity detection task to more fine-grained challenges. Although numerous and diverse datasets are being developed to train and test models, the tasks we describe here are closely related to those that are organized within the SemEval campaign and that we have mentioned at the end of section "Historical Background." It is worth mentioning that in addition to the datasets produced to support specific tasks, and thus tailored to encode specific information (e.g., irony), there have recently been increasing efforts toward producing *layered* datasets, where information useful for the various subtasks is all encoded, via different layers, on exactly the same data (Basile et al. 2014, 2015a; Mohammad et al. 2016b) (see also section "Future Directions").

Traditional Task: Classifying the Overall Polarity of a Text

In the 2013 edition of SemEval, a popular campaign of evaluation of NLP methods and technologies, the task #2 ("sentiment analysis on Twitter") proposed two subtasks to its participants: contextual and message-level polarity classification (Nakov et al. 2013). Since then, the classification of the polarity (or lack thereof) of a given message is considered the central task of sentiment analysis. A text is classified as *positive* (or *negative*) if its overall message expresses a positive (or negative) sentiment. In the absence of any detected polarity, a message is *neutral*. Some evaluation frameworks are open to the possibility of both polarities existing at the same time, thus yielding a *mixed* message (Basile et al. 2014).

Another alternative, to traditional *binary* classification task with *positive* and *negative* as possible class values, is a *multi-label* classification task including also *neutral* as a possible value for polarity to denote those cases for which no opinion is conveyed by the text. Such an evaluation framework is usually related to the task of subjectivity classification. The classification of the degree of subjectivity of a message can be cast an independent task, yet usually it is considered somewhat depending on the polarity detection, i.e., only subjective utterances can be polarized.

Recently, we are also assisting to a shift of focus from these more traditional frameworks addressing also the sentiment *strength*, that is, the intensity of the opinion or emotion being expressed. In such a scenario, we deal with an *ordinal classification* task aiming at assessing sentiment strength on an ordered scale, such as the one defined by *VeryPositive*, *Positive*, *Neutral*, *Negative*, and *VeryNegative* (Pontiki et al. 2016) or by numeric values in $[N, -N]$ (Thelwall et al. 2012), where 0 is usually adopted to denote neutral texts.

Looking at the recent studies on sentiment analysis, with particular focus to the systems submitted to the evaluation campaigns since 2013 (Basile et al. 2014; Nakov et al. 2013), the majority of approaches to polarity classification relies on supervised learning. More in detail, the most popular machine learning included SVM, maximum entropy, CRFs, and linear regression.

More recently a clear trend emerged to prefer supervised methods based on deep learning, including convolutional neural networks and recurrent neural networks. Another widely adopted technique exploits word embeddings. Distributional semantic models are usually generated via word2vec (Mikolov et al. 2013) and are exploited to derive semantic features for classifiers (Basile and Novielli 2014). In this vein, some authors use general-purpose, pretrained embeddings, while others build their own vector space either on gold standard dataset developed on purpose for sentiment analysis tasks (see (Nakov et al. 2016) for an updated overview on recent studies).

The main features used are usually derived from sentiment lexicons, the most popular ones being MPQA (Wilson et al. 2005), SentiWordNet (Esuli and Sebastiani 2006), and Bing Liu's Opinion Lexicon (Hu and Liu 2004). Some researchers derive their own lexicons based on manually labeled resources (Mohammad and Turney 2010) or by automatically adapting existing English lexicons to other languages (Basile and Novielli 2014) (for a wider overview of lexical resources and how they are employed in sentiment analysis, see Nissim and Patti (2017)). Along with traditional linguistic features such as negations, which acts as a reverser of the polarity of words in their scope, researchers also need to take into account character repetitions and use of upper case, which are usually used as intensifiers in microposts. Other traditional linguistic features such as word or characters n-grams are also adopted, in combination with new emerging ones that derive from typical features of microblogging platforms, such as the user mentions.

Even if they currently are the most widely adopted by researchers, supervised approaches present the main drawback of being highly domain dependent. This means that systems are very likely to perform poorly outside the domain they were trained on (Gamon et al. 2005). Different approaches were also investigated by empirical researchers in this fields that led to the development of lexicon- (Thelwall et al. 2012) and rule-based systems (Tromp and Pechenizkiy 2014).

Aspect-Based Sentiment Analysis

Sentiment analysis approaches generally attempt to detect the overall polarity of a sentence, paragraph, or text span, irrespective of the entities (e.g., restaurants) and their aspects (e.g., food, service) mentioned. However, quite often comments and subjective opinions are expressed on single aspects of interest.

A more fine-grained sentiment detection is usually performed in aspect-based sentiment analysis (ABSA) whose aim is to identify the aspects of given target entities and the sentiment expressed for each aspect. For example, given the comment "The food was lousy – too sweet or too salty and the portions tiny," we can identify two distinct aspects: *food* and *portions*. The sentiment with respect to these two aspects is negative. Traditionally, ABSA aimed at summarizing the content of users' reviews in several commercial domains like consumer electronics (Ganu et al. 2013; Hu and Liu 2004; Thet et al. 2010) providing a general polarity score for each aspect.

ABSA is gaining increasing importance in recent years and the SemEval series dedicated a task to this problem since 2014 (Pontiki et al. 2014). The last edition of this task (AL-Smadi et al. 2016) attracts 245 submissions from 29 teams confirming the interest in this task. Generally, the ABSA task requires to work on semantic and syntactic features. In Brun et al. (2016), the authors combine several machine learning approaches (CRF and ensemble models) relying on bag of words, lemmas, bigrams, punctuation, sentiment lexicon, and syntactic features.

In the first attempts to solve the ABSA task, most systems are based on state-of-the-art machine learning algorithms such as SVMs (Brun et al. 2014; Brychcın et al. 2014; Kiritchenko et al. 2014; Wagner et al. 2014) or CRFs (Hamdan et al. 2015; Toh and Wang 2014), with lexical information, bigrams, and POS as features. In SemEval 2014, Kiritchenko et al. (2014) achieved good results on aspect category and aspect polarity detection, using SVMs combined with rich linguistic features including dependency parsing. However, in the last edition of SemEval, promising results have been achieved by systems based on deep learning. In Su and Toh (2016), the authors propose a combination of a binary classifiers trained using single layer feed-forward network for aspect and a sequential labeling classifiers for opinion target extraction. In Ruder et al. (2016), a convolutional neural network (CNN) for both aspect extraction and aspect-based sentiment analysis is exploited in a multilingual setting. Both the previous systems provide best results in SemEval-2016. In Khalil and El-Beltagy (2016), an ensemble classifier which combines convolutional neural network (CNN) and a support vector machine (SVM) classifier based on the bag of words model achieved promising results.

Future directions involving the ABSA task are moving toward the evaluation of cross-lingual or language-agnostic approaches as already experimented in SemEval-2016 and the analysis of irony and figurative language.

Irony and Figurative Language

Communications in social media platforms include a high percentage of linguistic devices for figurative language, such as irony and sarcasm (Davidov et al. 2010; Gonzalez-Iban~ez et al. 2011; Reyes and Rosso 2014; Reyes et al. 2013), and systems for detecting the sentiment in microblogging data experienced the phenomenon of wrong polarity classification of ironic messages (Bosco et al. 2013; Ghosh et al. 2015a). Indeed, the presence in a text of ironic devices can flip the polarity of an opinion expressed with positive words to the intended negative meaning (one says something "good" to mean something "bad") – or vice versa – working as an unexpected polarity reverser. This can undermine systems' accuracy. The automatic detection of irony is, therefore, crucial for the development of irony-aware sentiment analysis systems, and a special task SemEval-2015 was recently dedicated to the sentiment analysis of figurative language in Twitter (Ghosh et al. 2015a). At the same time, it is also an interesting conceptual challenge from a cognitive point of view and can help to shed some light on how human beings use irony as a communicative tool.

Different approaches to the task of recognizing verbal irony in texts have been developed. The majority of them take advantage only of the textual content itself, since in textual messages other paralinguistic cues, like, for instance, the tone or corporal movements, are not available. Twitter is the most widely used source of information to experiment with irony detection. This is mainly due to availability of a large set of samples of ironic texts, which are easy to be collected relying on the behavior of Twitter users, who often explicitly mark their ironic messages by using hashtags such as “#irony” or “#sarcasm.” The pretty good reliability of the user-generated hashtags as golden labels for irony has been experimentally confirmed by Kunneman et al. (2015). Moreover, it seems that, due to the interaction model underlying the microblogging platform, irony expressed here could be somehow easier to analyze. Indeed, Twitter users have to be sharp and short, having only 140 characters for expressing their comments, and most of the times the ironic posts do not require knowledge about the conversational context to be understood. Several works have been carried out using tweets for experimental purposes (Bamman and Smith 2015; Barbieri et al. 2014; Davidov et al. 2010; Gonzalez-Iban~ez et al. 2011; Farias et al. 2015; Joshi et al. 2015; Karoui et al. 2015b; Ptacek et al. 2014; Rajadesingan et al. 2015; Reyes et al. 2013; Riloff et al. 2013; Wang 2013). Furthermore, there are some efforts in other social media such as customer reviews from Amazon (http://www.amazon.com/) (Buschmeier et al. 2014; Filatova 2012); comments from the online debate sites such as 4forums.com (http://www.4forums.com/political/) (Abbott et al. 2011; Lukin and Walker 2013) and, recently, Reddit (http://www.reddit.com) (Wallace et al. 2015).

The majority of the research in irony detection has been addressed in English, although there is some research in other languages, such as Dutch (Kunneman et al. 2015), Italian (Bosco et al. 2013), Czech (Ptacek et al. 2014), French (Karoui et al. 2015b), Portuguese (Carvalho et al. 2009), and Chinese (Tang and Chen 2014). A shared task for English on sentiment analysis of figurative language in Twitter has been organized at SemEval-2015 for the first time (Ghosh et al. 2015a), and a pilot shared task for Italian on irony detection has been proposed in Sentipolc-2014 within the periodic evaluation campaign EVALITA (Basile et al. 2014). This confirms the growing interest for this task in the research community, especially for understanding the impact of the ironic devices on sentiment analysis.

Irony detection has been modeled as a binary classification problem, where mostly tweets labeled with certain hashtags (i.e., #irony, #sarcasm, #sarcastic, #not) have been considered as ironic utterances. Following this framework, different approaches have been proposed (Barbieri et al. 2014; Davidov et al. 2010; Gonzalez-Iban~ez et al. 2011; Farias et al. 2015; Ptacek et al. 2014; Reyes et al. 2013; Riloff et al. 2013). The authors proposed models that exploit mainly features related to textual content such as punctuation marks, emoticons, part-of-speech labels, discursive terms, and specific patterns (e.g., according to Riloff et al. (2013), a common form of sarcasm in Twitter consists of a positive sentiment contrasting with a negative situation), among others.

Another key characteristic for irony is *unexpectedness* (Attardo 2000). According to many theoretical accounts, people infer irony when they recognize an incongruity between an utterance and what is known (or expected) about the speaker and/or the environment. This is something that can be referred to as the pragmatic context. Recent approaches started to address such issue, taking into account information about context (Bamman and Smith 2015; Rajadesingan et al. 2015; Wallace et al. 2015).

For what concerns the affective information, some approaches already used in their models some kind of sentiment and emotional information. Reyes et al. (2013) included in their model some features to characterize irony in terms of elements related to sentiments, attitudes, feelings, and moods exploiting the Dictionary of Affect in Language proposed by Whissell (2009). Barbieri et al. (2014) considered the amount of positive and negative words by using SentiWordNet. Hernandez et al. (2015) exploited two widely applied sentiment lexicons (Hu&Liu and AFINN) as features in their model. Recent works focused specifically on studying the role of affective information in a comprehensive manner, by exploring the use of a wide range of lexical resources available for English, reflecting different aspects of a multifaceted phenomenon (Hernandez Farıas et al. 2016).

Few preliminary studies address the task to investigate the differences between irony and sarcasm, which can also be interesting in order to reason on the possibility to observe different polarity reversal patterns behind the two figurative devices. A contribution on this line is given in Sulis et al. (2016), where authors analyze messages explicitly tagged by users with #irony, #sarcasm, and #not in order to test the hypothesis to deal with different linguistic phenomena, with a special focus on the role of features related to the multifaceted affective information expressed in such texts.

There are also other figurative devices, like metaphor, where we can observe that the sentiment polarity of the literal meaning differs from that of the intended figurative meaning (Ghosh et al. 2015a). Metaphorical expressions represent a great variety, ranging from conventional metaphors to poetic and largely novel ones (Shutova 2010). The use of metaphor can sometimes also contribute to give an ironic twist to a sentence and, being pervasive also in social media, represents a further serious challenge for sentiment analysis systems working on this kind of texts.

Emerging Task: Stance Detection

In addition to knowing whether a microblogger expresses a positive or negative opinion in a given text, interest has recently focused on detecting what particular position, or *stance*, a user has or takes with respect to some specific target. Note that this closely relates to aspect-based sentiment analysis (see section "Aspect-Based Sentiment Analysis"). While interest in stance detection originated mainly in relation to online debates and essays (Faulkner 2014; Somasundaran and Wiebe 2010; Walker et al. 2012), and thus longer documents, it has recently shown also on Twitter. For example, Rajadesingan and Liu (2014) propose a semi-supervised framework to detect Twitter users' stance on hotly contested issues such as abortion or gun reforms, and Djemili et al. (2014) develop a set of linguistically motivated rules to detect ideology in tweets produced by French politicians.

To match the growing interest in stance detection on microblogs, a stance-annotated Twitter dataset has recently been developed (Mohammad et al. 2016a) and used as benchmark in one of the tasks at SemEval-2016 (Mohammad et al. 2016b). At this competition, systems were asked to detect users' positions as being in *favor*, *against*, or *neutral* with respect to pre-chosen targets, in two different settings. For the first one, annotated training data was provided, so that supervised learning could be employed. For the second, a weakly supervised framework was provided, where systems were given a large amount of topic-specific tweets, without, however, any manual annotation.

Almost 20 teams participated in the 2016 edition, using both standard text classification features, such as n-grams, and standard sentiment analysis features, such as lexical semantic information derived from existing resources. Some teams attempted to exploit additional data using distant supervision techniques, or employed deep neural networks, but none of them managed to beat the system used as baseline by the organizers (Mohammad et al. 2016b, c), which achieved an f-score of 0.703 on the test data. This is a linear SVM trained with basic features such as word and character n-grams as well as features derived from external lexica and word-embedding features derived from unannotated data.

Key Applications

From an applicative perspective, microposts comprise an invaluable wealth of data, ready to be mined for training predictive models. We review here three areas of interest from an application perspective: business, general social well-being, and research.

Business – Analyzing the sentiment conveyed by microposts can yield a competitive advantage for businesses (Jansen et al. 2009), and mining opinions about specific aspects of entities being discussed is of paramount importance in this sense. Due to the pervasiveness of mobile devices and the ubiquitous diffusion of social media platforms, information analysis and sentiment analysis of microposts are now being exploited to forecast real-world market outcomes (Asur and Huberman 2010). Many companies are interested in using results of sentiment analysis in order to develop marketing strategies. In fact, user-generated contents are a precious mine for grasping opinions of people about a specific topic or product; thus, they can constitute a valuable asset for firms to directly tap into the customer's preferences. But the leveraging of social media for the purpose of tracking product image requires sentiment-related technologies, and in response to these needs, NLP-related companies that offer monitoring and analysis of social media to learn more about consumer behavior toward brands, products, and services are getting more and more popular (Bhatt et al. 2010). Fora such as the American *Sentiment Analysis Symposium* (http://sentimentsymposium.com/) are annually organized with the explicit aim to bridge business- and sentiment-related technology for mining and exploiting opinions, emotions, and intents in online, social, and enterprise contents.

Social utility – Beyond the pure commercial application domain, analysis of microposts can serve to gain crucial insights in the field of security, politics, and social well-being in general. The availability of a constant flow of information in the form of short texts makes it in theory possible to collect real-time information about events, which in turn can provide crucial information in times of emergencies. TopicSketch, a system for automatically determining bursting events on Twitter (Xie et al. 2013), and ReDites, an event detection and visualization system developed in order to help analysts to identify security-related events (Osborne et al. 2014), are prime examples of this line of applicative research. Specific sentiment-related research in this sense has also emerged in the last years, with efforts toward estimating emotional response to terrorism-related events from social media (Colbaugh and Glass 2010) and more in general toward understanding public emotional reactions from Twitter (Sykora et al. 2013; Wan and Paris 2015). The automatic analysis of microblogs has also been exploited to support humanitarian aid and disaster relief (Kumar et al. 2011; Tapia et al. 2011).

Political sentiment and election results (Tumasjan et al. 2010), as well as political movements (Starbird and Palen 2012), have also been the object of sentiment analysis-based investigations on microposts. Health wise, mining emotions on microblogging platform is now regarded as a powerful tool for monitoring public health issues (Michael and Paul 2011). In particular, researchers in this field have leveraged sentiment analysis on Twitter to predict postpartum changes in mood of new mothers (De Choudhury et al. 2013) and to assess how emotion sharing in social media may indicate desensitization to violence experienced in communities plagued by political conflicts or protracted violence (Choudhury et al. 2015). Similarly, emotion expressions of employees on enterprise microblogging tools may be seen as an indicator toward the understanding of the "affective climate" of a workplace (De Choudhury and Counts 2013).

Research – Besides the goals of several industry branches, sentiment analysis techniques are often also interesting for their application to other open research problems. One such field is *argumentation mining*, the area of NLP and Artificial Intelligence that models arguments between several participants about different topics and develops techniques to extract such models from raw data such as natural language text. Recent works such as Grosse et al. (2015) employ sentiment analysis in order to extract useful features from text for the ultimate purpose of identifying the opinions and stances of the participants in a debate. Similarly, Basile et al. (2016) directly compare cognitive and sentiment analysis approaches to study online debates. Finally, research in empirical software engineering is now devoting increasing attention to sentiment analysis of tweets reporting about software products (Guzman et al. 2016).

Future Directions

Exploring Unsupervised Approaches

Despite being in principle agnostic with respect to language and domain, supervised approaches are in practice highly domain dependent, as systems are very likely to perform poorly outside the domain they are trained on (Gamon et al. 2005). In fact, when training classification models, it is very likely to include consideration of terms that associate with sentiment because of the context of use. It is the case, for example, of political debates, where names of countries afflicted by wars might be associated to negative sentiments; analogous problems might be observed for the technology domain, where killer features of devices referred in positive reviews by customers usually become obsolete in relatively short periods of time (Thelwall et al. 2012). Such terms are usually referred by researchers as indirect affective words to distinguish them from direct ones (Valitutti and Stock 2006). Indeed, according to the emotion theory defined by Clore et al. (1987), it is possible to distinguish between words that directly refer to sentiment (e.g., *fear*, *joy*, *cheerful*, *sad*) and those having only an indirect reference to an emotional state, depending on the context (e.g., the words which indicates emotional causes such as *killer* or *monster* or emotional responses to an event such as *cry* or *laugh*). While representing a promising answer to the cross domain generalizability issue of sentiment classifiers in social web (Thelwall et al. 2012), unsupervised approaches have not been exhaustively investigated and represent an interesting direction for future research.

Sentiment Analysis of Figurative Language

Another main concern is the detection of irony and the correct polarity classification of tweets containing figurative language (Ghosh et al. 2015b; Stranisci et al. 2016; Reyes et al. 2013). In fact, in spite of the limitations due to the text length limit imposed by microblogging platforms (i.e., the 140 characters limit of Twitter), users adopt creative linguistic devices to convey their opinions, feelings, and emotions. It is the case of metaphor, irony, sarcasm, and, more in general, figurative language, which poses peculiar challenges to polarity classifiers. Irony has been explicitly addressed so far in both the Italian (Basile et al. 2014) and the English evaluation campaigns (Karoui et al. 2015a) that recently invited participants to deal with different forms of figurative language, and the goal of the task was to detect polarity of tweets using it. In both cases, participant submitted systems obtaining promising performance. Still, the complex relation between sentiment and figurative use of language needs to be further investigated. While, in fact, irony seems to mainly act as a polarity reverser, other linguistic devices might impact sentiment in different ways.

Entity-Based Sentiment Analysis

Mining information from a microblogging platform involves reliably identifying entities in tweets. As a consequence, entity linking on Twitter is gaining increasing attention (Guo et al. 2013). The overall goal of entity linking is to automatically extract entities from text and link them to the corresponding taxonomies and/or knowledge base as DBpedia or Freebase. As for sentiment, also entity linking in Twitter is now becoming a popular task for evaluation campaigns. Such tasks typically focus on entity extraction from tweets and linking to the corresponding entry in English DBpedia (Rizzo et al. 2015) resources, if the linkage exists.

By including explicit reference to entities, ABSA could broaden its impact beyond its traditional application in the commercial domain. While classical ABSA focuses on the sentiment/opinion with respect to a particular aspect, entity-based sentiment analysis (Batra and Rao 2010) tackles the problem of identifying the sentiment about an entity, for example, a politician, a celebrity, or a location. An extension to the sentiment analysis challenge at EVALITA that explicitly incorporates entity linking and sentiment analysis has been proposed in Basile et al. (2015b).

Cross-References

- ▶ Analysis and Mining of Tags, (Micro)Blogs, and Virtual Communities
- ▶ Behavior Analysis in Social Networks
- ▶ Behavior Modeling in Social Networks
- ▶ Emotions and Personality in Recommender Systems
- ▶ Microblog
- ▶ Microblog Sentiment Analysis
- ▶ Multi-classifier System for Sentiment Analysis and Opinion Mining
- ▶ Opinion Diffusion and Analysis on Social Networks
- ▶ Sentiment Analysis
- ▶ Sentiment Analysis, Basic Tasks of

- ▶ Sentiment Analysis in Social Media
- ▶ Sentiment Analysis of Reviews
- ▶ Sentiment Quantification of User-Generated Content
- ▶ Semantic Sentiment Analysis of Twitter Data
- ▶ Social Media Analysis for Monitoring Political Sentiment
- ▶ Twitter Microblog Sentiment Analysis
- ▶ User Sentiment and Opinion Analysis

Acknowledgments This work is partially funded by the project "EmoQuest – Investigating the Role of Emotions in Online Question & Answer Sites," funded by MIUR (Ministero dellUniversita' e della Ricerca) under the program "Scientific Independence of young Researchers" (SIR 2014) and the project "Multilingual Entity Liking" funded by the Apulia Region under the program FutureInResearch.

References

Abbott R, Walker M, Anand P, Fox Tree JE, Bowmani R, King J (2011) How can you say such things?!?: recognizing disagreement in informal political argument. In: Proceedings of the workshop on languages in social media, LSM '11, pp 2–11. Association for Computational Linguistics, Stroudsburg

Arnold MB (1960) Emotion and personality, vol 1. Columbia University Press, New York

Asur S, Huberman BA (2010) Predicting the future with social media. In: Proceedings of the 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology – vol 01, WI-IAT '10, pp 492–499. IEEE Computer Society, Washington, DC. https://doi.org/10.1109/WI-IAT.2010.63

Attardo S (2000) Irony as relevant inappropriateness. J Pragmat 32(6):793–826

Bamman D, Smith NA (2015) Contextualized sarcasm detection on twitter. In: Proceedings of the ninth international conference on web and social media, ICWSM 2015, pp 574–577. AAAI, Oxford

Barbieri F, Saggion H, Ronzano F (2014) Modelling sarcasm in twitter, a novel approach. In: Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis, pp 50–58. Association for Computational Linguistics, Baltimore

Basile V, Nissim M (2013) Sentiment analysis on Italian tweets. In: Proceedings of WASSA 2013, Atlanta. pp 100–107

Basile P, Novielli N (2014) UNIBA at EVALITA 2014-SENTIPOLC task: predicting tweet sentiment polarity combining micro-blogging, lexicon and semantic features. In: Proceedings of EVALITA 2014, Pisa. pp 58–63

Basile V, Bolioli A, Nissim M, Patti V, Rosso P (2014) Overview of the Evalita 2014 SENTIment POLarity classification task. In: Proceedings of EVALITA 2014, pp 50–57. Pisa University Press, Pisa

Basile P, Basile V, Nissim M, Novielli N (2015a) Deep tweets: from entity linking to sentiment analysis. In: Proceedings of CLiC-it 2015, Trento, Italy. p 41

Basile P, Basile V, Nissim M, Novielli N (2015b) Deep tweets: from entity linking to sentiment analysis. CLiC-it, Trento, Italy. p 41

Basile V, Cabrio E, Villata S, Frasson C, Gandon F (2016) A pragma-semantic analysis of the emotion/sentiment relation in debates. In: 4th international workshop on artificial intelligence and cognition. New York

Batra S, Rao D (2010) Entity based sentiment analysis on twitter. Science 9(4):1–12

Bhatt R, Chaoji V, Parekh R (2010) Predicting product adoption in large-scale social networks. In: Proceedings of the 19th ACM international conference on information and knowledge management, CIKM '10, pp 1039–1048. ACM, New York. https://doi.org/10.1145/1871437.1871569

Bosco C, Patti V, Bolioli A (2013) Developing corpora for sentiment analysis: the case of irony and senti-tut. IEEE Intell Syst 28(2):55–63

Brun C, Popa ND, Roux C (2014) Xrce: hybrid classification for aspect-based sentiment analysis. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), pp 838–842. Association for Computational Linguistics. Dublin, Ireland. http://aclweb.org/anthology/S14-2149

Brun C, Perez J, Roux C (2016) Xrce at semeval-2016 task 5: feedbacked ensemble modeling on syntactico-semantic knowledge for aspect based sentiment analysis. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016), pp 277–281. Association for Computational Linguistics. San Diego, California

Brychcın T, Konkol M, Steinberger J (2014) Uwb: machine learning approach to aspect-based sentiment analysis. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), pp 817–822. Association for Computational Linguistics. Dublin, Ireland. http://aclweb.org/anthology/S14-2145

Buschmeier K, Cimiano P, Klinger R (2014) An impact analysis of features in a classification approach to irony detection in product reviews. In: Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis, pp 42–49. Association for Computational Linguistics, Baltimore

Carofiglio V, Rosis Fd, Novielli N (2009) Cognitive emotion modeling in natural language communication. In: Tao J, Tan T (eds) Affective information processing. Springer, London, pp 23–44

Carrillo de Albornoz J, Plaza L, Gervas P (2012) Sentisense: an easily scalable concept-based affective lexicon for sentiment analysis. In: Proceedings of the eight international conference on language resources and evaluation (LREC'12), pp 3562–3567. European Language Resources Association (ELRA), Istanbul

Carvalho P, Sarmento L, Silva MJ, de Oliveira E (2009) Clues for detecting irony in user-generated contents: Oh...!! it's "so easy" ;-). In: Proceedings of the 1st international CIKM workshop on topic-sentiment analysis for mass opinion, TSA '09, pp 53–56. ACM, New York

Choudhury MD, Monroy-Hernandez A, Mark G (2015) "Narco" emotions: affect and desensitization in social media during the Mexican drug war. CoRR abs/1507.01287. http://arxiv.org/abs/1507.01287

Clore GL, Ortony A, Foss MA (1987) The psychological foundations of the affective lexicon. J Pers Soc Psychol 53(4):751–766

Colbaugh R, Glass K (2010) Estimating sentiment orientation in social media for intelligence monitoring and analysis. In: Intelligence and Security Informatics (ISI), 2010 I.E. international conference on, Canada. pp 135–137. IEEE

Davidov D, Tsur O, Rappoport A (2010) Semi-supervised recognition of sarcastic sentences in twitter and amazon. In: Proceedings of the fourteenth conference on computational natural language learning, CoNLL '10, pp 107–116. Association for Computational Linguistics, Uppsala

De Choudhury M, Counts S (2013) Understanding affect in the workplace via social media. In: Proceedings of the 2013 conference on computer supported cooperative work, CSCW '13, pp 303–316. ACM, New York. https://doi.org/10.1145/2441776.2441812

De Choudhury M, Counts S, Horvitz E (2013) Predicting postpartum changes in emotion and behavior via social media. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI '13, pp 3267–3276. ACM, New York. https://doi.org/10.1145/2470654.2466447

Djemili S, Longhi J, Marinica C, Kotzinos D, Sarfati GE (2014) What does twitter have to say about ideology? In: NLP 4 CMC: natural language processing for computer-mediated communication/social media-pre-conference workshop at Konvens 2014, vol 1. Universitätsverlag, Hildesheim

Ekman P (1992) An argument for basic emotions. Cognit Emot 6(3–4):169–200

Esuli A, Sebastiani F (2006) Sentiwordnet: a publicly available lexical resource for opinion mining. In: Proceedings of LREC, Italy. pp 417–422

Faulkner AR (2014) Automated classification of argument stance in student essays: a linguistically motivated approach with an application for supporting argument summarization. In: Proceedings of the twenty-seventh international Florida artificial intelligence research society conference. Florida, USA

Fellbaum C (1998) WordNet: an electronic lexical database. Bradford Books. Cambridge, Massachusetts

Filatova E (2012) Irony and Sarcasm: corpus generation and analysis using crowdsourcing. In: Proceedings of the eighth international conference on language resources and evaluation (LREC-2012), pp 392–398. European Language Resources Association (ELRA), Istanbul

Gamon M, Aue A, Corston-Oliver S, Ringger E (2005) Pulse: mining customer opinions from free text. In: Proceedings of the 6th international conference on advances in intelligent data analysis, IDA'05, pp 121–132. Springer, Berlin. https://doi.org/10.1007/11552253 12

Ganu G, Kakodkar Y, Marian A (2013) Improving the quality of predictions using textual information in online user reviews. Inf Syst 38(1):1–15. https://doi.org/10.1016/j.is.2012.03.001

Ghosh A, Li G, Veale T, Rosso P, Shutova E, Barnden J, Reyes A (2015a) Semeval-2015 task 11: sentiment analysis of figurative language in twitter. In: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), pp 470–478. Association for Computational Linguistics, Denver

Ghosh A, Li G, Veale T, Rosso P, Shutova E, Reyes A, Barnden J (2015b) Semeval-2015 task 11: sentiment analysis of figurative language in twitter. In: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), pp 470–475. Association for Computational Linguistics, Denver. http://www.aclweb.org/anthology/S14-2004

Gonzalez-Iban~ez R, Muresan S, Wacholder N (2011) Identifying sarcasm in twitter: a closer look. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, HLT '11, pp 581–586. Association for Computational Linguistics, Portland

Grosse K, Gonzalez MP, Chesne~var CI, Maguitman AG (2015) Integrating argumentation and sentiment analysis for mining opinions from twitter. AI Commun 28(3):387–401. https://doi.org/10.3233/AIC-140627

Guo S, Chang MW, Kiciman E (2013) To link or not to link? A study on end-to-end tweet entity linking. In: Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 1020–1030. Association for Computational Linguistics, Atlanta

Guzman E, Alkadhi R, Seyff N (2016) A needle in a haystack: what do twitter users say about software? In: 24nd IEEE international requirements engineering conference (RE'16). Beijing, China

Hamdan H, Bellot P, Bechet F (2015) Lsislif: Crf and logistic regression for opinion target extraction and sentiment polarity analysis. In: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), pp 753–758. Association for Computational Linguistics. Denver, Colorado. http://aclweb.org/anthology/S15-2128

Hernandez Farias I, Benedi JM, Rosso P (2015) Applying basic features from sentiment analysis for automatic irony detection. In: Pattern recognition and image analysis. Lecture notes in computer science, vol 9117. Springer International Publishing, Santiago de Compostela, pp 337–344

Hernandez Farıas I, Patti V, Rosso P (2016) Irony detection in twitter: the role of affective content. ACM Trans Internet Technol 16(3):19:1–19:24

Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, Seattle, WA, USA. pp 168–177

Jansen BJ, Zhang M, Sobel K, Chowdury A (2009) Twitter power: tweets as electronic word of mouth. J Am Soc Inf Sci Technol 60(11):2169–2188

Joshi A, Sharma V, Bhattacharyya P (2015) Harnessing context incongruity for sarcasm detection. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (vol 2: Short papers), pp 757–762. Association for Computational Linguistics, Beijing

Karoui J, Benamara F, Moriceau V, Aussenac-Gilles N, Belguith LH (2015a) Towards a contextual pragmatic model to detect irony in tweets. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing of the Asian Federation of natural language processing, ACL 2015, July 26–31, 2015, Beijing, vol 2: Short papers, pp 644–650. http://aclweb.org/anthology/P/P15/P15-2106.pdf

Karoui J, Benamara F, Moriceau V, Aussenac-Gilles N, Hadrich-Belguith L (2015b) Towards a contextual pragmatic model to detect irony in tweets. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (vol 2: Short papers), pp 644–650. Association for Computational Linguistics, Beijing

Khalil T, El-Beltagy SR (2016) Niletmrg at semeval-2016 task 5: deep convolutional neural networks for aspect category and sentiment extraction. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016), pp 271–276. Association for Computational Linguistics. San Diego, California

Kiritchenko S, Zhu X, Cherry C, Mohammad S (2014) Nrc-canada-2014: detecting aspects and sentiment in customer reviews. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), pp 437–442. Association for Computational Linguistics. Dublin, Ireland. http://aclweb.org/anthology/S14-2076

Kiritchenko S, Mohammad S, Salameh M (2016) Semeval-2016 task 7: determining sentiment intensity of English and Arabic phrases. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016), pp 42–51. Association for Computational Linguistics, San Diego. http://www.aclweb.org/anthology/S16-1004

Kumar S, Barbier G, Abbasi MA, Liu H (2011) Tweettracker: an analysis tool for humanitarian and disaster relief. In: ICWSM. Barcelona

Kunneman F, Liebrecht C, van Mulken M, van den Bosch A (2015) Signaling sarcasm: from hyperbole to hashtag. Inf Process Manag 51(4):500–509

Lazarus RS (1991) Emotion and adaptation. Oxford University Press, New York

Lukin S, Walker M (2013) Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. In: Proceedings of the workshop on language analysis in social media, pp 30–40. Association for Computational Linguistics, Atlanta

Michael J, Paul MD (2011) You are what you tweet: analyzing twitter for public health. In: Proceedings of the fifth international AAAI conference on weblogs and social media, Barcelona. pp 265–272

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781

Mohammad SM, Turney PD (2010) Emotions evoked by common words and phrases: using mechanical turk to create an emotion Lexicon. In: Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text, CAAGET '10, Los Angels CA. pp 26–34

Mohammad S, Kiritchenko S, Sobhani P, Zhu X, Cherry C (2016a) A dataset for detecting stance in tweets. In: N.C.C. (Chair), Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard+ B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the tenth international conference on language resources and evaluation (LREC 2016). European Language Resources Association (ELRA), Paris

Mohammad SM, Kiritchenko S, Sobhani P, Zhu X, Cherry C (2016b) Semeval-2016 task 6: detecting stance in tweets. In: Proceedings of the international workshop on semantic evaluation, SemEval '16. San Diego

Mohammad SM, Sobhani P, Kiritchenko S (2016c) Stance and sentiment in tweets. arXiv preprint arXiv:1605.01655

Nakov P, Rosenthal S, Kozareva Z, Stoyanov V, Ritter A, Wilson T (2013) Semeval-2013 task 2: sentiment analysis in twitter. In: Second joint conference on lexical and computational semantics (*SEM), vol 2: Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013), pp 312–320. Association for Computational Linguistics, Atlanta. http://www.aclweb.org/anthology/S13-2052

Nakov P, Ritter A, Rosenthal S, Sebastiani F, Stoyanov V (2016) Semeval-2016 task 4: sentiment analysis in twitter. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016), pp 1–18. Association for Computational Linguistics, San Diego. http://www.aclweb.org/anthology/W/W05/W05-0201

Nielsen FA: Afinn (2011). http://www2.imm.dtu.dk/pubdb/p.php?6010

Nissim M, Patti V (2017) Semantic aspects in sentiment analysis, chap. 3. In: Pozzi FA, Fersini E, Messina E, Liu B (eds) Sentiment analysis in social networks. Morgan Kaufmann, pp 31–48. https://doi.org/

10.1016/B978-0-12-804412-400003-6. http://www.sciencedirect.com/science/article/pii/B9780128044124000036

O'Connor B, Balasubramanyan R, Routledge B, Smith N (2010) From tweets to polls: linking text sentiment to public opinion time series. In: International AAAI conference on weblogs and social media (ICWSM), vol 11, Washington, DC. pp 122–129

Osborne M, Moran S, McCreadie R, Von Lunen A, Sykora MD, Cano E, Ireson N, Macdonald C, Ounis I, He Y et al (2014) Real-time detection, tracking, and monitoring of automatically discovered events in social media. In: Proceedings of ACL 2014: system demonstrations, pp 37–42. Association for Computational Linguistics. Baltimore

Pak A, Paroubek P (2010) Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of the seventh international conference on language resources and evaluation (LREC'10). Malta

Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1–2):1–135

Parrot WG (2001) Emotions in social psychology: essential readings. Psychology Press, Philadelphia

Pennebaker JW, Francis ME, Booth RJ (2001) Linguistic inquiry and word count: LIWC 2001. Lawrence Erlbaum Associates, Mahwah

Picard RW (1997) Affective computing. MIT Press, Cambridge, MA

Plutchik R (1980) A general psychoevolutionary theory of emotion. In: Plutchik R, Kellerman H (eds) Emotion: theory, research, and experience: vol 1. Theories of emotion. Academic, New York, pp 3–33

Pontiki M, Galanis D, Pavlopoulos J, Papageorgiou H, Androutsopoulos I, Manandhar S (2014) Semeval-2014 task 4: aspect based sentiment analysis. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), pp 27–35. Association for Computational Linguistics. Dublin, Ireland. http://aclweb.org/anthology/S14-2004

Pontiki M, Galanis D, Papageorgiou H, Manandhar S (2015) I.: Semeval-2015 task 12: Aspect based sentiment analysis. In: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), pp 486–495. Association for Computational Linguistics, Denver. http://www.aclweb.org/anthology/S15-2082

Pontiki M, Galanis D, Papageorgiou H, Androutsopoulos I, Manandhar S, AL-Smadi M, Al-Ayyoub M, Zhao Y, Qin B, De Clercq O, Hoste V, Apidianaki M, Tannier X, Loukachevitch N, Kotelnikov E, Bel N, Jimenez-Zafra SM, Eryiit G (2016) Semeval-2016 task 5: aspect based sentiment analysis. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016), pp 19–30. Association for Computational Linguistics San Diego, California

Pozzi F, Fersini E, Messina E, Liu B (2017) Chapter 1 – challenges of sentiment analysis in social networks: an overview. In: Pozzi FA, Fersini E, Messina E, Liu B (eds) Sentiment analysis in social networks. Morgan Kaufmann, Boston, pp 1–11. https://doi.org/10.1016/B978-0-12-804412-4.00001-2. http://www.sciencedirect.com/science/article/pii/B9780128044124000012

Ptacek T, Habernal I, Hong J (2014) Sarcasm detection on Czech and English twitter. In: Proceedings of COLING 2014, the 25th international conference on computational linguistics, pp 213–223. Dublin City University and Association for Computational Linguistics, Dublin

Rajadesingan A, Liu H (2014) Identifying users with opposing opinions in twitter debates. In: International conference on social computing, behavioral-cultural modeling, and prediction, pp 153–160. Springer International Publishing. Washington DC, USA

Rajadesingan A, Zafarani R, Liu H (2015) Sarcasm detection on twitter: a behavioral modeling approach. In: Proceedings of the eighth ACM international conference on web search and data mining, WSDM '15, pp 97–106. ACM, Shanghai

Reyes A, Rosso P (2014) On the difficulty of automatically detecting irony: beyond a simple case of negation. Knowl Inf Syst 40(3):595–614

Reyes A, Rosso P, Veale T (2013) A multidimensional approach for detecting irony in twitter. Lang Resour Eval 47(1):239–268

Riloff E, Qadir A, Surve P, Silva LD, Gilbert N, Huang R (2013) Sarcasm as contrast between a positive sentiment and negative situation. In: Proceedings of the 2013 conference on empirical methods in natural language processing, (EMNLP 2013), pp 704–714. Association for Computational Linguistics, Seattle

Rizzo G, Cano Basave AE, Pereira B, Varga A, Rowe M, Stankovic M, Dadzie A (2015) Making sense of Microposts Named Entity rEcognition and Linking (NEEL) challenge. In: Proceedings of the 5th workshop on making sense of Microposts co-located with the 24th international World Wide Web conference (WWW 2015), vol 1395. CEUR. Aachen, Germany

Rosenthal S, Ritter A, Nakov P, Stoyanov V (2014) SemEval-2014 Task 9: sentiment analysis in Twitter. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), pp 73–80. Dublin. http://www.aclweb.org/anthology/S14-2009

Rosenthal S, Nakov P, Kiritchenko S, Mohammad SM, Ritter A, Stoyanov V (2015) SemEval-2015 Task 10: Sentiment analysis in Twitter. In: Proceedings of the 9th international workshop on semantic evaluation, SemEval '2015, Denver

Ruder S, Ghaffari P, Breslin JG (2016) Insight-1 at semeval-2016 task 5: deep learning for multilingual aspect-based sentiment analysis. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016), pp 330–336. Association for Computational Linguistics. San Diego, California

Russell JA (1980) A circumplex model of affect. J Pers Soc Psychol 39:1161–1178

Shutova E (2010) Models of metaphor in nlp. In: Proceedings of the 48th Annual meeting of the association for computational linguistics, ACL '10, pp 688–697. Association for Computational Linguistics, Stroudsburg

Somasundaran S, Wiebe J (2010) Recognizing stances in ideological on-line debates. In: Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text, pp 116–124. Association for Computational Linguistics. Los Angels, CA

Starbird K, Palen L (2012) (how) will the revolution be retweeted?: Information diffusion and the 2011 Egyptian uprising. In: Proceedings of the ACM 2012 conference on computer supported cooperative work, CSCW '12, pp 7–16. ACM, New York

Stranisci M, Bosco C, Faras DIH, Patti V (2016) Annotating sentiment and irony in the online italian political debate on #labuonascuola. In: N.C.C. (Chair), Choukri K, Declerck T, Grobelnik M, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the tenth international conference on language resources and evaluation (LREC 2016). European Language Resources Association (ELRA), Paris

Strapparava C, Mihalcea R (2007) SemEval-2007 task 14: affective text. In: Proceedings of the 4th international workshop on semantic evaluations, SemEval '07, pp 70–74. Association for Computational Linguistics, Stroudsburg

Strapparava C, Valitutti A (2004) WordNet-Affect: an affective extension of WordNet. In: Proceedings of the International conference on language resources and evaluation (LREC'04), vol 4, pp 1083–1086. European Language Resources Association (ELRA). Reykjavik, Iceland

Su J, Toh Z (2016) Nlangp at semeval-2016 task 5: improving aspect based sentiment analysis using neural network features. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016), pp 282–288. Association for Computational Linguistics

Sulis E, Hernandez Farıas I, Rosso P, Patti V, Ruffo G (2016) Figurative messages and affect in twitter: differences between #irony, #sarcasm and #not. Knowl-Based Syst 108:132–143. 10.1016/j.knosys.2016.05.035. New avenues in knowledge bases for natural language processing

Sykora MD, Jackson TW, OBrien A, Elayan S (2013) National security and social media monitoring: a presentation of the EMOTIVE and related systems. In: Intelligence and security informatics conference (EISIC), 2013 European, pp 172–175. IEEE. Uppsala, Sweden

Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. Comput Linguist 37(2):267–307

Tang YJ, Chen HH (2014) Chinese irony corpus construction and ironic structure analysis. In: Proceedings of COLING 2014, the 25th international conference on computational linguistics, pp 1269–1278. Association for Computational Linguistics, Dublin

Tapia AH, Bajpai K, Jansen BJ, Yen J, Giles L (2011) Seeking the trustworthy tweet: can microblogged data fit the information needs of disaster response and humanitarian relief organizations. In: Proceedings of the 8th international ISCRAM conference, Lisbon, Portugal. pp 1–10

Thelwall M, Buckley K, Paltoglou G (2012) Sentiment strength detection for the social web. J Am Soc Inf Sci Technol 63(1):163–173. https://doi.org/10.1002/asi.21662.URL

Thet TT, Na JC, Khoo CS (2010) Aspect-based sentiment analysis of movie reviews on discussion boards. J Inf Sci 36(6):823–848. https://doi.org/10.1177/0165551510388123

Toh Z, Wang W (2014) Dlirec: aspect term extraction and term polarity classification system. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), pp 235–240. Association for Computational Linguistics. http://aclweb.org/anthology/S14-2038

Tromp E, Pechenizkiy M (2014) Rule-based emotion detection on social media: putting tweets on Plutchik's wheel. CoRR abs/1412.4682. http://arxiv.org/abs/1412.4682

Tumasjan A, Sprenger T, Sandner P, Welpe I (2010) Predicting elections with twitter: what 140 characters reveal about political sentiment. In: International AAAI conference on web and social media. Canada

Valitutti R, Stock O (2006) The affective weight of lexicon. In: Proceedings of the fifth international conference on language resources and evaluation. Italy

Villena-Roman J, Garcıa-Morera J, Cumbreras MA G, Martınez-Camara E, MartınValdivia MT, Lopez LAU (2015) Overview of TASS 2015. In: Villena-Roman J, Garcıa-Morera J, Cumbreras MAG, Martınez-Camara E, Martın-Valdivia MT, Lopez LAU (eds) Proceedings of TASS 2015: workshop on sentiment analysis at SEPLN co-located with 31st SEPLN conference (SEPLN 2015), Alicante, Sept 15, 2015. CEUR workshop proceedings, vol 1397, pp 13–21. CEUR-WS.org. http://ceur-ws.org/Vol-1397/overview.pdf

Wagner J, Arora P, Cortes S, Barman U, Bogdanova D, Foster J, Tounsi L (2014) Dcu: aspect-based polarity classification for Semeval task 4. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), pp 223–229. Association for Computational Linguistics. Dublin, Ireland. http://aclweb.org/anthology/S14-2036

Walker MA, Anand P, Abbott R, Grant R (2012) Stance classification using dialogic properties of persuasion. In: Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: human language technologies, NAACL HLT '12, pp 592–596. Association for Computational Linguistics, Stroudsburg. http://dl.acm.org/citation.cfm?id=2382029.2382124

Wallace BC, Choe DK, Charniak E (2015) Sparse, contextually informed models for irony detection: exploiting user communities, entities and sentiment. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (vol 1: Long papers), pp 1035–1044. Association for Computational Linguistics, Beijing

Wan S, Paris C (2015) Understanding public emotional reactions on twitter. In: Proceedings of the 2015 AAAI international conference on weblogs and social media (ICWSM), Oxford, UK. pp 715–716

Wang AP (2013) #irony or #sarcasm – a quantitative and qualitative study based on twitter. In: Proceedings of the PACLIC: the 27th Pacific Asia conference on language, information, and computation, pp 349–356. Department of English, National Chengchi University, Taipei

Whissell C (2009) Using the revised Dictionary of affect in language to quantify the emotional undertones of samples of natural languages. Psychol Rep 2(105):509–521

Wiebe J, Wilson T, Cardie C (2005) Annotating expressions of opinions and emotions in language. Lang Resour Eval 1(2). http://www.cs.pitt.edu/wiebe/pubs/papers/lre05withappendix.pdf

Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on human language technology and empirical methods in natural language processing, HLT '05, pp 347–354. Association for Computational Linguistics, Stroudsburg. 10.3115/1220575.1220619

Xie W, Zhu F, Jiang J, Lim EP, Wang K (2013) Topicsketch: real-time bursty topic detection from twitter. In: 2013 I.E. 13th international conference on data mining, USA. pp 837–846. IEEE

Sentiment Analysis of Reviews

Subhabrata Mukherjee
Department of Databases and Information Systems, Max-Planck-Institut für Informatik, Saarbrücken, Germany

Synonyms

Aspect or facet mining; Data mining; Knowledge discovery; Opinion mining; Sentiment classification; Social media analysis

Glossary

JAST	Joint author sentiment topic model
JST	Joint sentiment topic model
KB	Knowledge bases
MaxEnt	Maximum entropy classifier
NB	Naive Bayes classifier
PMI	Point-wise mutual information
POS	Part of speech
SA	Sentiment analysis
SO	Semantic orientation
SVM	Support vector machines

Definition

Sentiment analysis (SA) of reviews refers to the task of analyzing natural language text in forums like Amazon, TripAdvisor, Yelp, IMDB, etc. to obtain the writer's feelings, attitudes, and emotions expressed therein toward a particular topic, product, or entity. It involves overlapping approaches in several domains like natural language processing (NLP), computational linguistics (CL), information extraction (IE), text mining, and machine learning (ML).

Introduction

In recent years, the explosion of social networking sites (e.g., Facebook, Twitter), blogs (e.g., Mashable, TechCrunch), and online review portals (e.g., Amazon, TripAdvisor, IMDB) provides overwhelming amount of information about products and services. Millions of people express uninhibited opinions about various product (and service) features and their nuances. As consumers cannot test the functionality of a product (or service) prior to consumption, these reviews help them make an informed decision to buy the product (or service) or not. Sentiment analysis aims to tap this gold mine of information by analyzing the vast repository of largely unstructured text – in the form of reviews, comments, questions, and requests, etc. – to retrieve users' opinions about featured products and services. It can be performed at different levels of *granularity* as follows. Consider the following movie review:

Example 1 "This film is based on a true-life incident. It sounds like a great plot and the director makes a decent attempt in narrating a powerful story. However, the film does not quite make the mark due to sloppy acting."

- *Document-level* SA (Dave et al. 2003; Pang et al. 2002) aims to find the *overall* polarity of the review as *positive* or *negative*, depending on whether its author liked it or not (for instance, it is *negative* in Example 1).
- *Sentence-level* SA (Mukherjee and Bhattacharyya 2012a; Yu and Hatzivassiloglou 2003) analyzes individual sentences in the review to find its polarity. For example, the second sentence in Example 1 depicts a *positive* sentiment about the movie.
- *Phrase-level* SA (Mukherjee and Bhattacharyya 2012a; Wu et al. 2009) analyzes individual phrases to determine its polarity. For example, "does not quite make the mark" depicts a *negative* sentiment. This is particularly interesting, in case of *negation* operators in the scope, which flips the polarity of the phrase, for instance, contrast "not only good..." (positive) with "not good" (negative).
- *Facet-level* SA (Lakkaraju et al. 2011; Lin and He 2009; Subhabrata Mukherjee and Joshi 2014) finds the polarity of a review with respect to the underlying *facets* (attributes, features, or aspects) of the item in the review under consideration. For instance, the sentiment in Example 1 is *positive* w.r.t. the movie facets "plot" and "director"; however, it is *negative* w.r.t. "acting."
- *Author-level* SA (Mukherjee et al. 2013; Subhabrata Mukherjee and Joshi 2014) aims to find the polarity of a review w.r.t. the *person* who wrote the review. Since reviews are subjective, the same review may attain different polarities depending on the preferences of the person who authored it.

Considering all of the different aspects outlined above, sentiment (or opinion) can be defined as:

Definition 1 "Sentiment (or, opinion) is a quintuple: $\langle o_j , f_{jk} , so_{ijkl} , h_i, t_l \rangle$, where o_j is a target object (entity or item), f_{jk} is a feature (facet/aspect) of the object $o\ j$, $soi\ jkl$ is the sentiment on feature f_{jk} of object o_j, h_i is the opinion holder, and t_l is the time when the opinion is expressed by h_i. The sentiment so_{ijkl} is +ve, -ve, or neutral, or expressed with different strength/intensity levels, e.g., 1 to 5 stars as used by most review sites on the Web" (Pang and Lee 2008).

Key Points

This work gives a broad overview of the various approaches and state-of-the-art systems for sentiment analysis of reviews. It focuses on three key areas: (i) commonly used lexical resources for SA, (ii) a broad overview of the various features and aspects of SA, and (ii) some prominent machine learning approaches to SA with a brief overview of some of the commonly used classifiers and techniques.

Historical Background

Early works in sentiment analysis starting with the seminal work of Dave et al. (2003), Pang and Lee (2004b), Pang et al. (2002), Turney (2002), and Yu and Hatzivassiloglou (2003) considered reviews as bag of words and focused on classifying them as positive, negative, or neutral using classifiers. Later works developed more sophisticated features based on phrasal and dependency relations, narratives, perspectives, lexical resources, etc. The flurry of activity in this domain, in recent times, can be attributed to the availability of large-scale datasets with the explosion of social networking sites and online review portals and significant advances in machine learning algorithms for natural language processing, text mining, and information extraction tasks.

Lexical Resources

The simplest approach to sentiment analysis is to consider *word-level* features. Given a review document with a sequence of words, and access to a lexical resource containing the *annotated* polarity (positive, negative, or objective) of opinion words and phrases, a count-based approach assigns the majority polarity of the opinion words in the document as the polarity of the review. The following are some commonly used lexical resources for SA:

- SentiWordNet (Esuli and Sebastiani 2006) is a lexical resource, where each WordNet synset (i.e., sense) *s* is associated with three numerical scores (each ranging from [0 to 1] on a simplex) – Obj(*s*), Pos(*s*), and Neg(*s*) – denoting the corresponding polarity scores of the synset. This corresponds to a graded evaluation of opinion, as opposed to a hard one. For instance, the synset "estimable," with the sense "may be computed or estimated," has an Obj score of 1.0 and corresponding Pos and Neg scores as 0.0, whereas its Obj, Pos, and Neg scores corresponding to the sense "deserving of high respect or high regard" are 0.25, 0.75, and 0 respectively.
- Subjectivity lexicon (Wiebe et al. 2004) is a resource that annotates words with tags like parts of speech, prior polarity, magnitude of prior polarity (weak/strong), etc. The prior polarity can be positive, negative, or neutral.
- Inquirer (Stone et al. 1966) is a list of words marked as positive, negative, and neutral.
- Taboada (Taboada and Grieve 2004) is a word list that gives a count of collocations with positive and negative seed words. A word closer to a positive seed word is predicted to be positive and vice versa.
- Bing Liu sentiment lexicon (Hu and Liu 2004) contains a list of manually annotated positive and negative opinion words.

Features and Aspects of Sentiment Analysis

Feature engineering is a basic and essential task for most machine learning-based approaches to sentiment analysis. Converting a piece of review text to a feature vector is the basic step in any data-driven approach to SA. In the following section, we will see different aspects and features for sentiment analysis.

Term Presence Versus Term Frequency

Term (or word) frequency has always been considered essential in traditional information retrieval (IR) and text classification tasks. Pang et al. (2002) empirically found term presence to be more important to sentiment analysis than term frequency. This involves binary-valued feature vectors, with entries indicating the presence or absence of words. Unlike in traditional text classification tasks, the presence of a strong sentiment-bearing word can change the overall polarity of a sentence in SA. It has also been seen that the occurrence of rare words contains more information than frequently occurring words – a phenomenon called hapax legomenon.

Term Position

Words appearing in certain positions in the text carry more sentiment or weight than words appearing elsewhere. This is similar to IR where words appearing in topic titles, subtitles, abstracts, etc. are given more weight than those appearing in the body. For instance, the review text in Example 1 contains more positive words than negative; however, the presence of negative sentiment in the *concluding* sentence makes the overall sentiment of the review *negative*. This is a typical example of *thwarting* which has been investigated in some recent works (Mukherjee and Joshi 2014; Ramteke et al. 2013).

N-Grams

N-grams are capable of capturing context in texts and are widely used in natural language processing tasks. Whether higher-order N-grams are increasingly useful is a matter of debate. Pang et al. (2002) reported unigrams to outperform bigrams when classifying movie reviews by sentiment polarity; however, Dave et al. (2003) found that in some settings bigrams and trigrams perform better.

Part of Speech

Part-of-speech information is most commonly exploited in all NLP tasks. One of the important reasons is that they provide a crude form of word sense disambiguation (WSD). Adjectives have been used most frequently as features among different parts of speech. There is a strong correlation between *adjectives* and subjectivity in SA. Although many part-of-speech tags are important, most of the common sentiment-bearing words are adjectives. Pang et al. (2002) achieved an accuracy of around 82.8% in movie review domains using only adjectives as features.

Apart from using only adjectives, the a*djective-adverb* combination is also informative. Most of the adverbs have no prior polarity. But when they occur with sentiment-bearing adjectives, they can play a major role in determining the sentiment of a sentence. Benamara et al. (2007) demonstrated how adverbs can alter the sentiment value of the adjectives that they are used with. For example, the sentiment orientation of "immensely good" is more positive than that of using only the adjective "good." Similarly "barely good" is more negative than the positive sentiment-bearing adjective "good." This fine-grained analysis can be used to assign graded polarity scores to sentiment-bearing words and phrases, rather than assigning a hard polarity label as positive or negative.

Semantic Orientation with PMI

Semantic orientation (SO) (Turney 2002) refers to a real number measure of the positive or negative sentiment expressed by a word or phrase. The SO of a phrase is determined based on the phrase's point-wise mutual information (PMI) with the words "excellent" and "poor." PMI is defined as follows: $PMI(w_1, w_2) = log_2(p(w_1\&w_2)/p(w_1)p(w_2))$.

The SO for a phrase is the difference between its PMI with the word "excellent" and its PMI with the word "poor." This yields values above zero for positive sentiment-bearing phrases and below zero for negative ones. An SO value of zero would indicate a neutral semantic orientation.

Discourse and Modalities

"An important component of language comprehension in most natural language contexts involves connecting clauses and phrases together in order to establish a coherent discourse" (Wolf and Gibson 2004). The presence of linguistic constructs like connectives, modals, conditionals, and negation can alter sentiment at the sentence level as well as the clausal or phrasal level.

Example 2 "I'm quite excited about Tintin, despite not really liking original comics. Probably because Joe Cornish had a hand in."

The overall sentiment of Example 2 is *positive*, although there are an equal number of positive and negative sentiment-bearing words. This is due to the connective "despite" which gives more *emphasis* on the previous discourse segment. Any bag-of-words model would be unable to classify this sentence without considering the discourse marker "despite." Mukherjee and Bhattacharyya (2012b) probe the influence of these discourse markers and semantic operators (like modals and negations) for fine-grained sentiment analysis of sentences at the *discourse level*. They outline a lightweight approach to incorporate these insights as rules and features in a simple classifier that is effective for SA, even in noisy mediums (e.g., Twitter with short and noisy texts), where heavyweight approaches based on *parsing* typically fail.

Parsing

In dependency grammar, structure is determined by the relation between a head and its dependents. The dependent is a modifier or complement, and the head plays a more important role in determining the behaviors of the pair. Dependency parsing captures short-range and long-range dependencies between words in a sentence.

Example 3 "I have an iPod and it is a great buy but I'm probably the only person that dislikes the iTunes software."

The review in Example 3 is *positive* w.r.t. "iPod" but *negative* w.r.t. "iTunes." Feature-specific SA finds the polarity of a review with respect to the target facets or features in the review. Mukherjee and Bhattacharyya (2012a) use dependency parsing (e.g., Stanford dependency parser) for *facet-level* SA to capture the relations between the facets (aspects or features of items in a review) and their associated opinions. The idea is to capture the association between any specific facet and the expressions of opinion that come together to describe that feature. However, not all dependency relations are important for SA. The authors show that dependency relations like “nsubj, dobj, advmod, amod, neg, prep of, acomp, xcomp, conj and ccomp, iobj,” etc. are more important for SA than other relations.

Wu et al. (2009) use *phrase dependency parsing* for opinion mining. This approach trades off the information loss of the word-level dependency in dependency parsing, as it does not explicitly provide local structures and syntactic categories of phrases and the information gain in extracting long-distance relations. Hence, they extend the dependency tree node with phrases.

Chen and Yao (2010) use dependency parsing and shallow semantic analysis for Chinese opinion-related expression extraction. They categorize relations as topic and sentiment located in the same subsentence and quite close to each other (like the rule “an adjective plus a noun” is mostly a potential opinion-element relation), topic and sentiment located in adjacent subsentences and the two subsentences which are parallel in structure (that is to say, the two adjacent subsentences are connected by some coherent word, like although/but, and, etc.), topic and sentiment located in different sub-sentences, either being adjacent or not, etc.

World Knowledge: Encyclopedic, Semantic, and Ontological Resources

Sentiment analysis often requires access to external or background knowledge (e.g., knowledge bases (KB) like the Knowledge Graph (KG) (Dong et al. 2014)) to understand entity-specific concepts in the text as in Example 4. This *objective* synopsis of the movie will be misclassified as *negative* using any lexical resource for word-level SA due to the presence of the word “dangerous” – which is a part of the movie plot.

Example 4 “L.I.E. stands for Long Island Expressway, which slices through the strip malls and middle-class homes of suburbia. Filmmaker Michael Cuesta uses it as a metaphor of dangerous escape for his 15-year old protagonist, Howie (Paul Franklin Dano).”

Mukherjee and Bhattacharyya (2012c) use Wikipedia to understand concepts specific to movies like the movie plot, fictional characters, objective facts about the crew and cast, their past performance, and characteristics of the movie genre. This is used for extractive summarization to derive relevant subjective (opinionated) sentences significant for sentiment analysis of the movie review and not objective facts about the movie.

Yu and Hatzivassiloglou (2003) propose to find subjective sentences using lexical resources where the authors hypothesize that subjective sentences will be more similar to opinionated sentences than to factual sentences. As a measure of similarity between two sentences, they used different measures including shared words, phrases, and the WordNet. Potthast and Becker (2010) focus on extracting top sentiment keywords based on point-wise mutual information (PMI) measure (Turney 2002).

The pioneering work for subjectivity detection is done in Pang and Lee (2004a), where the authors use *min-cut* to leverage coherency between the sentences. The fundamental assumption is that local proximity preserves the objectivity or subjectivity relation in the review. Agarwal (2005) integrates graph cut with linguistic knowledge in the form of WordNet to exploit similarity in the set of documents to be classified.

Example 5 “I bought a Canon EOS 7D (DSLR). It’s very small, sturdy, and constructed well. The handling is quite nice with a powder-coated metal frame. It powers on quickly and the menus are fairly easy to navigate. The video modes are nice, too. It works great with my 8GB Eye-Fi SD card. A new camera isn’t worth it if it doesn’t exceed the picture quality of my old 5Mpixel SD400 and this one doesn’t. The auto white balance is poor. I’d need to properly balance every picture taken so far with the ELPH 300. With 12 Mpixels, you’d expect pretty good images, but the problem is that the ELPH 300 compression is turned up so high that the sensor’s acuity gets lost (softened) in compression.”

Review Example 5 depicts the complexity involved in analyzing product reviews. The review has a mix of good and bad comments about various features of the product. A flat classification model which considers all features to be equally important will fail to capture the proper polarity of the review. The reviewer seems happy with the "camera size, structure, easy use, video modes, SDHC support," etc. However, the "auto-white balance" and "high compression" leading to "sensor acuity" seem to disappoint him. Now, the primary function of a camera is to take good pictures and videos. Thus "picture, video quality, resolution, color balance," etc. are of primary importance, whereas "size, video mode, easy use," etc. are secondary in nature. The overall review polarity, in this example, should be *negative* as the reviewer shows concerns about the most important features of the camera.

In order to analyze this review, we not only need to understand the different underlying *facets* but also their *relations*. This can be captured using *ontological* information – where the concepts are nodes and edges between them capture relations. Typically, the concepts higher up in the ontology are more important than the concepts further down the hierarchy.

Wei and Gulla (2010) propose a hierarchical learning method to label a product's attributes and their associated sentiments in product reviews using a Sentiment Ontology Tree (HLSOT). The HLSOT approach is supervised, requiring the reviews to be annotated with product attribute relations, as well as feature-specific opinion expressions. Mukherjee and Joshi (2013) use ConceptNet (Liu and Singh 2004) as a knowledge resource to automatically construct a domain-specific ontology tree for product reviews, without requiring any labeled training data. They present a novel sentiment aggregation approach to combine the feature-specific polarities with ontological information to find the overall polarity of the review. ConceptNet relations have an inherent structure which helps in the construction of an ontology tree from the resource.

Machine Learning Approaches

In the previous section, we reviewed various aspects of sentiment analysis and different classes of features that are commonly used, like n-grams, part of speech, parsing (e.g., dependency relations like "nsubj, dobj, advmod," and "amod"), sentiment lexicons for prior polarity of words, discourse and modality features, etc. Now given access to all of these features, we can create a feature vector $f = \{f_1, f_2, \ldots, f_m\}$ (considering m number of features) for each review text and use a machine learning classifier to classify it as *positive* or *negative*. We now review some commonly used classifiers for sentiment analysis. Starting with the seminal work of Pang and Lee (2004a) and Pang et al. (2002), the most commonly used machine learning classifiers for SA have been Naive Bayes, maximum entropy, and support vector machines.

Naive Bayes (NB)

The simplest approach is to assign a sentiment label l to a document d that maximizes the conditional probability: $l* = argmax_l \ P(l|d)$. Assuming *conditional independence* between the features: $P(l|d) = \frac{P(l)\prod_{i=1}^{m} P(f_i|l)^{n_i(d)}}{P(d)}$, where $n_i(d)$ denotes the number of times feature f_i appears in document d and $P(f_i|l)$ is computed based on the number of times f_i is observed in documents with the sentiment label l. Despite the conditional independence assumptions, this approach works fairly well with a good *smoothing* technique (e.g., Laplace smoothing).

Maximum Entropy (MaxEnt)

Unlike Naive Bayes, maximum entropy does not make any assumptions regarding the relationships between the features and therefore may perform better when conditional independence assumptions are not met. Its estimate of $P(l|d)$ has the following exponential form: $P(l|d) = \frac{1}{Z(d)} \exp\left(\sum_i \lambda_{i,l} f_{i,l}(d,l)\right)$, where $Z(d)$ is a normalization function. $F_{i,l}$ is a feature/class function for feature f_i and class l, defined as a binary indicator function that assumes the value 1 or 0 corresponding to whether the feature f_i is present for class l or not. The weight λ_i for f_i learned from data depicts the importance of that feature for the classification task. The weights are learned to maximize the entropy of the distribution subject to the constraint that the expected values of the feature/class functions with respect to the model are equal to their expected values with respect to the training data. The underlying philosophy is to choose the model that makes the fewest assumptions about the data, while remaining consistent with it.

A more generalized model of this flavor that has been recently shown to outperform other classifiers is the *conditional random field (CRF)*. It is typically used for sequence classification tasks, where the objective is to predict a *sequence* of labels instead of a single one – where the model leverages interaction between the features of different label classes for more accurate predictions. It also does not suffer from the label bias problem due to global normalization, unlike the (locally normalized) maximum-entropy Markov model (MEMM) classifiers.

Support Vector Machines (SVM)

Support vector machines (SVM) have been shown to be extremely effective for text classification tasks. SVM maps the features (using kernels) to a high-dimensional space and constructs a hyperplane to separate the two categories. Although there can be an infinite number of such hyperplanes possible, SVM constructs the one with the largest functional margin given by the distance of the nearest point to the hyperplane on each side of it.

The solution can be written as $\overline{w} = \sum_j \alpha_j \cdot c_j \cdot \overline{d_j}$, where the α_j 's are obtained by solving a dual optimization problem. The $\overline{d_j}$ corresponding to $\alpha_j > 0$ are called support vectors, since they are the only document vectors contributing to $\overline{w}$. New points are mapped to the same space and classified to a category based on which side of the hyperplane they lie.

Facet or Aspect-Specific Sentiment Analysis

A review may have multiple facets or topics, with a different opinion about each facet. Consider the review in Example 1. The review is positive with respect to the topics "direction" and "story," but negative with respect to "acting." In order to analyze the sentiment of this review, it is necessary to identify the different facets or aspects in the review and then analyze the sentiment about those facets. The overall sentiment of the review can then be computed as a weighted aggregation of the facet or aspect-specific sentiments.

Latent Aspect Rating Analysis Model (LARAM) (Wang et al. 2010, 2011) jointly identifies latent aspects, aspect ratings, and weights placed on the aspects in a review. A shallow dependency parser is used to learn product aspects and aspect-specific opinions in Yu et al. (2011) by jointly considering the aspect frequency and the consumers' opinions about each aspect. A rated aspect summary of short comments is done in Lu et al. (2009). A topic model is used in Titov and McDonald (2008) to assign words to a set of induced topics. The model is extended through a set of maximum entropy classifiers, one per each rated aspect, which are used to predict aspect-specific ratings.

A joint sentiment topic model (JST) is described in Lin and He 2009) which detects sentiment and topic simultaneously from text. In JST, each document has a sentiment label distribution. Topics are associated with sentiment labels, and words are associated to both topics and sentiment labels. In contrast to Titov and McDonald (2008) and some other similar works (Lu et al. 2009; Snyder and Barzilay 2007; Wang et al. 2010, 2011; Yu et al. 2011) which require some kind of a supervised setting like aspect ratings or overall review rating (Mukherjee et al. 2013), JST is fully unsupervised. The CFACTS model (Lakkaraju et al. 2011) extends the JST model to capture facet coherences in a review using hidden Markov model. All of these generative models have their root in latent Dirichlet allocation model (Blei et al. 2003). LDA assumes a document to have a probability distribution over a mixture of topics and topics to have a probability distribution over words.

Author-Specific Sentiment Analysis

However, the models described so far do not consider any authorship information to incorporate author preferences for the facets or author style information for maintaining coherence in reviews. The overall sentiment for review in

Example 1 will differ for different authors depending on their topic preferences; if a reviewer watches a movie for a good story and narration, then his (overall) sentiment for the movie will be different than that if he watches it only for the acting skills of the protagonists in this example.

An approach to capture author-specific topic preferences is described in Mukherjee et al. (2013). The work considers *predefined* seed facets for restaurants like “food, ambience, service,” etc. and uses dependency parsing with a sentiment lexicon to find the sentiment about each facet. A WordNet similarity metric is used to assign each facet to a seed facet. Thereafter, they use linear regression to learn author preference for the seed facets from review ratings.

Joint author sentiment topic model (JAST) (Subhabrata Mukherjee and Joshi 2014) describes a generative process of writing a review by an author, without any supervision, building an author layer over the joint sentiment topic model (Lin and He 2009). Authors have different topic preferences, “emotional” attachment to topics, writing style based on the distribution of semantic (topic), and syntactic (background) words and their tendency to switch topics. JAST uses latent Dirichlet allocation to learn the distribution of author-specific topic preferences and emotional attachment to topics. It uses a hidden Markov model to capture short-range syntactic and long-range semantic dependencies in reviews to capture coherence in author writing style. JAST jointly discovers the topics in a review, author preferences for the topics, topic sentiment, as well as the overall review sentiment from the point of view of the author of the review.

Key Applications

Word of mouth (WOM) is the process of conveying information from person to person and plays a major role in customer buying decisions. In commercial situations, WOM involves consumers sharing attitudes, opinions, or reactions about businesses, products, or services with other people. WOM communication functions based on social networking and trust. People rely on families, friends, and others in their social network. Research also indicates that people appear to trust seemingly disinterested opinions from people outside their immediate social network, such as online reviews. This is where sentiment analysis comes into play. The growing availability of opinion-rich resources like online review sites, blogs, and social networking sites has made this “decision-making process” easier for us. With explosion of Web 2.0 platforms, consumers have a soapbox of unprecedented reach and power by which they can share opinions. Major companies have realized that these consumer voices shape voices of other consumers. Sentiment analysis thus finds its use in:

- Consumer market for product reviews for knowing consumer attitudes and trends for marketing-related activities
- Social media for finding general opinion about recent hot topics in town
- Product domains (like movies and electronics) to find whether a recently launched item is gaining popularity

Pang and Lee (2008) broadly classify the applications into the following categories:

- Applications to review related websites like movie and product reviews
- Applications as a subcomponent technology detecting antagonistic, heated language in mails, spam detection, context-sensitive information detection, etc.
- Applications in business and government intelligence to know consumer attitudes and trends
- Applications across different domains to know public opinions for political leaders, their notions about rules and regulations in place, etc.

Future Directions

Sentiment analysis has largely been restricted to textual analysis. An interesting extension can be to perform a *multimodal* analysis by considering speech, images, and videos to capture human behavior, interactions, and sentiment in real time which can have applications in the gaming industry and can improve security and intelligence.

Opinions are *dynamic* in nature, evolving with user experience, maturity, and social interactions. For instance, a review on a given item (e.g., movie) can have different polarities at different points in time – depending on when or at what level of maturity the user authored it in his life cycle in the community (Mukherjee et al. 2015, 2016). The dynamics on how users update their opinions on topics over time is affected by their social interactions as well. These can be used for forecasting users' sentiment, emotion, and attitudes over time (De et al. 2015).

With the recent advance of deep learning and representation learning in natural language processing tasks, efforts have been underway to understand the compositionality of forming sentiment expressions in text at fine-grained granularity and develop end-to-end systems for more robust sentiment predictions using *neural networks* (Socher et al. 2013). This area is likely to attract a lot of attention in coming times.

Cross-References

- ▶ Microblog Sentiment Analysis
- ▶ Multi-classifier System for Sentiment Analysis and Opinion Mining
- ▶ Semantic Sentiment Analysis of Twitter Data
- ▶ Sentiment Analysis
- ▶ Sentiment Analysis in Social Media
- ▶ Sentiment Analysis of Microblogging Data
- ▶ Sentiment Analysis, Basic Tasks of
- ▶ Sentiment Quantification of User-Generated Content
- ▶ Social Media Analysis for Monitoring Political Sentiment
- ▶ Twitter Microblog Sentiment Analysis
- ▶ User Sentiment and Opinion Analysis

References

Agarwal A (2005) Sentiment analysis: a new approach for effective use of linguistic knowledge and exploiting similarities in a set of documents to be classified. In: Proceedings of the international conference on natural language processing (ICON)

Benamara F, Cesarano C, Picariello A, Reforgiato D, Subrahmanian V (2007) Sentiment analysis: adjectives and adverbs are better than adjectives alone. In: Proceedings of the international conference on weblogs and social media (ICWSM)

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

Chen V, Yao T (2010) Combining dependency parsing with shallow semantic analysis for Chinese opinion-element relation identification. In: 4th international universal communication symposium, IUCS 2010, Beijing, 18–19 Oct 2010, pp 299–305

Dave K, Lawrence S, Pennock DM (2003) Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: Proceedings of the 12th international conference on world wide web, WWW '03. ACM, New York, pp 519–528

De A, Valera I, Ganguly N, Bhattacharya S, Gomez-Rodriguez M (2015) Modeling opinion dynamics in diffusion networks. CoRR, abs/1506.05474

Dong X, Gabrilovich E, Heitz G, Horn W, Lao N, Murphy K, Strohmann T, Sun S, Zhang W (2014) Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '14. ACM, New York, pp 601–610

Esuli A, Sebastiani F (2006) Sentiwordnet: a publicly available lexical resource for opinion mining. In: Proceedings of the 5th conference on language resources and evaluation (LREC'06), pp 417–422

Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD international conference on knowledge discovery and data mining, KDD '04. ACM, New York, pp 168–177

Lakkaraju H et al. (2011) Exploiting coherence in reviews for discovering latent facets and associated sentiments. In: SDM '11, pp 28–30

Lin C, He Y (2009) Joint sentiment/topic model for sentiment analysis. CIKM '09, pp 375–384

Liu H, Singh P (2004) Conceptnet: a practical commonsense reasoning tool-kit. BT Technol J 22(4):211–226

Lu Y, Zhai C, Sundaresan N (2009) Rated aspect summarization of short comments. In: WWW, pp 131–140

Mukherjee S, Bhattacharyya P (2012a) Feature specific sentiment analysis for product reviews. Lecture notes in computer science, vol 7181. Springer, Berlin/Heidelberg, pp 475–487

Mukherjee S, Bhattacharyya P (2012b) Sentiment analysis in twitter with lightweight discourse analysis. In: COLING, pp 1847–1864

Mukherjee S, Bhattacharyya P (2012c) Wikisent: weakly supervised sentiment analysis through extractive summarization with wikipedia. ECML PKDD'12, pp 774–793

Mukherjee S, Joshi S (2013) Sentiment aggregation using conceptnet ontology. In: Sixth international joint conference on natural language processing, IJCNLP 2013, Nagoya, 14–18 Oct 2013, pp 570–578

Mukherjee S, Joshi S (2014). Author-specific sentiment aggregation for polarity prediction of reviews. In: Proceedings of the ninth international conference on language resources and evaluation, LREC 2014, Reykjavik, 26–31 May 2014, pp 3092–3099

Mukherjee S, Basu G, Joshi S (2013) Incorporating author preference in sentiment rating prediction of reviews. In: WWW

Mukherjee S, Lamba H, Weikum G (2015) Experience-aware item recommendation in evolving review communities. In: 2015 I.E. international conference on data mining, ICDM 2015, Atlantic City, 14–17 Nov 2015, pp 925–930

Mukherjee S, Gu¨nnemann S, Weikum G (2016) Continuous experience-aware language model. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, 13–17 Aug 2016, pp 1075–1084

Pang B, Lee L (2004a). A sentimental education: sentiment analysis using subjectivity. In: Proceedings of ACL, pp 271–278

Pang B, Lee L (2004b) A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. ACL '04

Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1–2):1–135

Pang B, Lee VS, Lillian (2002) Thumbs up?: sentiment classification using machine learning techniques. EMNLP '02.

Potthast M, Becker S (2010) Opinion summarization of web comments. Springer, Berlin, pp 668–669

Ramteke A, Malu A, Bhattacharyya P, Nath JS (2013) Detecting turnarounds in sentiment analysis: thwarting. In: ACL (2). The Association for Computer Linguistics, pp 860–865

Snyder B, Barzilay R (2007) Multiple aspect ranking using the good grief algorithm. In: HLT 2007. ACL, Apr 2007, pp 300–307

Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng AY, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 conference on empirical methods in natural language processing. Association for Computational Linguistics, Stroudsburg, Oct 2013, pp 1631–1642

Stone PJ, Dunphy DC, Smith MS, Ogilvie DM (1966) The general inquirer: a computer approach to content analysis. MIT Press, Cambridge

Subhabrata Mukherjee GB, Joshi S (2014) Joint author sentiment topic model. In: Proceedings of the 2014 SIAM international conference on data mining (SDM'14)

Taboada M, Grieve J (2004) Analyzing appraisal automatically. In: In Proceedings of the AAAI spring symposium on exploring attitude and affect in text: theories and applications, pp 158–161

Titov I, McDonald RT (2008) A joint model of text and aspect ratings for sentiment summarization. ACL, Columbus, Ohio, pp 308–316

Turney PD (2002) Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting on Association for Computational Linguistics, ACL '02. Association for Computational Linguistics, Stroudsburg, pp 417–424

Wang H, Lu Y, Zhai C (2010) Latent aspect rating analysis on review text data: a rating regression approach. In: KDD, pp 783–792

Wang H et al (2011) Latent aspect rating analysis without aspect keyword supervision. In: KDD '11

Wei W, Gulla JA (2010) Sentiment learning on product reviews via sentiment ontology tree. In: Proceedings of the 48th annual meeting of the Association for Computational Linguistics, ACL '10. Association for Computational Linguistics, Stroudsburg, pp 404–413

Wiebe J, Wilson T, Bruce R, Bell M, Martin M (2004) Learning subjective language. Comput Linguist 30(3):277–308

Wolf F, Gibson E (2004) Representing discourse coherence: a corpus-based analysis. In: COLING 2004, 20th international conference on computational linguistics, proceedings of the conference, Geneva, 23–27 Aug 2004

Wu Y, Zhang Q, Huang X, Wu L (2009) Phrase dependency parsing for opinion mining. In: Proceedings of the 2009 conference on empirical methods in natural language processing: volume 3 volume 3, EMNLP '09. Association for Computational Linguistics, Stroudsburg, pp 1533–1541

Yu H, Hatzivassiloglou V (2003) Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In: Proceedings of the 2003 conference on empirical methods in natural language processing, EMNLP '03. Association for Computational Linguistics, Stroudsburg, pp 129–136

Yu J et al (2011) Aspect ranking: identifying important product aspects from online consumer reviews. ACL, pp 1496–1505

Sentiment Analysis, Basic Tasks of

Iti Chaturvedi, Soujanya Poria and Erik Cambria
School of Computer Science and Engineering,
Nanyang Technological University, Singapore,
Singapore

Synonyms

Convolutional neural network; Deep learning aspect extraction; Polarity distribution; Sentiment analysis; Subjectivity detection

Glossary

Aspect	Feature related to an opinion target
BOW	Bag of words
CNN	Convolutional neural network
Convolution	features made of consecutive words
LDA	Latent Dirichlet allocation
NLP	Natural language processing

Definition

Subjectivity detection is the task of identifying objective and subjective sentences. Objective sentences are those which do not exhibit any sentiment. So, it is desired for a sentiment analysis engine to find and separate the objective sentences for further analysis, e.g., polarity detection. In subjective sentences, opinions can often be expressed on one or multiple topics. Aspect extraction is a subtask of sentiment analysis that consists in identifying opinion targets in opinionated text, i.e., in detecting the specific aspects of a product or service the opinion holder is either praising or complaining about.

Introduction

While sentiment analysis research has become very popular in the past 10 years, most companies and researchers still approach it simply as a polarity detection problem. In reality, sentiment analysis is a "suitcase problem" that requires tackling many natural language processing (NLP) subtasks, including microtext analysis, sarcasm detection, anaphora resolution, subjectivity detection, and aspect extraction. In this chapter, we focus on the last two subtasks as they are key for ensuring a minimum level of accuracy in the detection of polarity from social media.

The two basic issues associated with sentiment analysis on the Web, in fact, are that (1) a lot of factual or non-opinionated information needs to be filtered out and (2) opinions are most times about different aspects of the same product or service rather than on the whole item and reviewers tend to praise some and criticize others. Subjectivity detection, hence, ensures that factual information is filtered out, and only opinionated information is passed on to the polarity classifier, and aspect extraction enables the correct distribution of polarity among the different features of the opinion target (instead of having one unique, averaged polarity assigned to it). In this chapter, we offer some insights about each task and apply an ensemble of deep learning and linguistics to tackle both.

The opportunity to capture the opinion of the general public about social events, political movements, company strategies, marketing campaigns, and product preferences has raised increasing interest of both the scientific community (because of the exciting open challenges) and the business world (because of the remarkable benefits for marketing and financial market prediction). Today, sentiment analysis research has its applications in several different scenarios. There are a good number of companies, both large and small scale, that focus on the analysis of opinions and sentiments as part of their mission (Cambria 2016). Opinion mining techniques can be used for the creation and automated upkeep of review and opinion aggregation websites, in which opinions are continuously gathered from the Web and not restricted to just product reviews but also to broader topics such as political issues and brand perception. Sentiment analysis also has a great potential as a sub-component technology for other systems. It can enhance the capabilities of customer relationship management and recommendation systems; for example, allowing users to find out which features customers are particularly interested in or to exclude items that have received overtly negative feedback from recommendation lists. Similarly, it can be used in social communication for troll filtering and to enhance anti-spam systems. Business intelligence is also one of the main factors behind corporate interest in the field of sentiment analysis (Cambria et al. 2014).

Sentiment analysis is a "suitcase" research problem that requires tackling many NLP subtasks, including semantic parsing (Rajagopal et al. 2013), named entity recognition (Ma et al.

2016), sarcasm detection (Poria et al. 2016b), subjectivity detection, and aspect extraction. In opinion mining, different levels of analysis granularity have been proposed, each one having its own advantages and drawbacks (Cambria 2013; Cambria et al. 2013b). Aspect-based opinion mining (Ding et al. 2008; Hu and Liu 2004) focuses on the relations between aspects and document polarity. An aspect, also known as an opinion target, is a concept in which the opinion is expressed in the given document.

For example, in the sentence, "The screen of my phone is really nice and its resolution is superb" for a phone review contains positive polarity, i.e., the author likes the phone. However, more specifically, the positive opinion is about its *screen* and *resolution*; these concepts are thus called opinion targets, or aspects, of this opinion. The task of identifying the aspects in a given opinionated text is called aspect extraction. There are two types of aspects defined in aspect-based opinion mining: explicit aspects and implicit aspects. Explicit aspects are words in the opinionated document that explicitly denote the opinion target. For instance, in the above example, the opinion targets *screen* and *resolution* are explicitly mentioned in the text. In contrast, an implicit aspect is a concept that represents the opinion target of an opinionated document but which is not specified explicitly in the text. One can infer that the sentence, "This camera is sleek and very affordable" implicitly contains a positive opinion of the aspects *appearance* and *price* of the entity *camera*. These same aspects would be explicit in an equivalent sentence: "The appearance of this camera is sleek and its price is very affordable."

Most of the previous works in aspect term extraction have either used conditional random fields (CRFs) (Jakob and Gurevych 2010; Zhiqiang and Wenting 2014) or linguistic patterns (Hu and Liu 2004; Poria et al. 2015b). Both of these approaches have their own limitations: CRF is a linear model, so it needs a large number of features to work well; linguistic patterns need to be crafted by hand, and they crucially depend on the grammatical accuracy of the sentences. In this chapter, we apply an ensemble of deep learning and linguistics to tackle both the problem of aspect extraction and subjectivity detection.

The remainder of this chapter is organized as follows: Section "Subjectivity Detection" and "Aspect-Based Sentiment Analysis" propose some introductory explanation and some literature for the tasks of subjectivity detection and aspect extraction, respectively; section "Preliminaries" illustrates the basic concepts of deep learning adopted in this work; section "Deep Learning Algorithm" describes in detail the proposed algorithm; section "Evaluation" shows evaluation results; and finally, section "Conclusion" concludes the chapter.

Key Points

- We consider deep convolutional neural networks where each layer is learned independent of the others resulting in low complexity.
- We model temporal dynamics in product reviews by pre-training the deep CNN using dynamic Gaussian Bayesian networks.
- We combine linguistic aspect mining with CNN features for effective sentiment detection.

Historical Background

Traditional methods prior to 2001 used handcrafted templates to identify subjectivity and did not generalize well for resource-deficient languages such as Spanish. Later works published between 2002 and 2009 proposed the use of deep neural networks to automatically learn a dictionary of features (in the form of convolution kernels) that is portable to new languages. Recently, recurrent deep neural networks are being used to model alternating subjective and objective sentences within a single review. Such networks are difficult to train for a large vocabulary of words due to the problem of vanishing gradients. Hence, in this chapter we consider use of heuristics to learn dynamic Gaussian networks to select significant word dependencies between sentences in a single review.

Further, in order to relation between opinion targets and the corresponding polarity in a review, aspect-based opinion mining is used. Explicit aspects were models by several authors using statistical observations such mutual information between noun phrase and the product class. However this method was unable to detect implicit aspects due to high level of noise in the data. Hence, topic modeling was widely used to extract and group aspects, where the latent variable "topic" is introduced between the observed variables "document" and "word." In this chapter, we demonstrate the use of "commonsense reasoning" when computing word distributions that enable shifting from a syntactic word model to a semantic concept model.

Subjectivity Detection

Subjectivity detection is an important subtask of sentiment analysis that can prevent a sentiment classifier from considering irrelevant or potentially misleading text in online social platforms such as Twitter and Facebook. Subjective extraction can reduce the amount of review data to only 60% and still produce the same polarity results as full text classification (Bonzanini et al. 2012). This allows analysts in government, commercial, and political domains who need to determine the response of people to different crisis events (Bonzanini et al. 2012; Murray and Carenini 2011; Tang et al. 2014). Similarly, online reviews need to be summarized in a manner that allows comparison of opinions, so that a user can clearly see the advantages and weaknesses of each product merely with a single glance, both in unimodal (Tang et al. 2009) and multimodal (Cambria et al. 2013a; Poria et al. 2016d) contexts. Furthermore, we can do in-depth opinion assessment, such as finding reasons or aspects (Poria et al. 2016a) in opinion-bearing texts, for example, "poor acting," which makes the film "awful." Several works have explored sentiment composition through careful engineering of features or polarity shifting rules on syntactic structures. However, sentiment accuracies for classifying a sentence as positive/negative/neutral have not exceeded 60%.

Early attempts used general subjectivity clues to generate training data from unannotated text (Riloff and Wiebe 2003). Next, bag-of-words (BOW) classifiers were introduced that represent a document as a multi-set of its words disregarding grammar and word order. These methods did not work well on short tweets. Co-occurrence matrices also were unable to capture difference in antonyms such as "good/bad" that have similar distributions. Subjectivity detection hence progressed from syntactic to semantic methods in Riloff and Wiebe (2003), where the authors used extraction pattern to represent subjective expressions. For example, the pattern "hijacking" of <x> looks for the noun "hijacking" and the object of the preposition <x>. Extracted features are used to train machine-learning classifiers such as SVM (Wiebe and Riloff 2005) and ELM (Oneto et al. 2016). Subjectivity detection is also useful for constructing and maintaining sentiment lexicons, as objective words or concepts need to be omitted from them (Cambria et al. 2016).

Since subjective sentences tend to be longer than neutral sentences, recursive neural networks were proposed where the sentiment class at each node in the parse tree was captured using matrix multiplication of parent nodes (Glorot et al. 2011; Kalchbrenner et al. 2014). However, the number of possible parent composition functions is exponential; hence in Socher et al. (2013) recursive neural tensor network was introduced that use a single tensor composition function to define multiple bilinear dependencies between words. In Maas et al. (2011), the authors used logistic regression predictor that defines a hyperplane in the word vector space where a word vectors positive sentiment probability depends on where it lies with respect to this hyperplane. However, it was found that while incorporating words that are more subjective can generally yield better results, the performance gain by employing extra neutral words is less significant (Lin et al. 2011). Another class of probabilistic models called latent Dirichlet allocation assumes each document is a mixture of latent topics. Lastly, sentence-level subjectivity detection was integrated into document-level sentiment detection using graphs where each node is a sentence. The contextual constraints between sentences in a graph led to significant improvement in polarity classification (Pang and Lee 2004).

Similarly, in Suzuki and Isozaki (2006) the authors take advantage of the sequence encoding method for trees and treat them as sequence kernels for sentences. Templates are not suitable for semantic role labeling, because relevant context might be very far away. Hence, deep neural networks have become popular to process text. In word2vec, for example, a word's meaning is simply a signal that helps to classify larger entities such as documents. Every word is mapped to a unique vector, represented by a column in a weight matrix. The concatenation or sum of the vectors is then used as features for prediction of the next word in a sentence (Collobert et al. 2011a). Related words appear next to each other in a *d* dimensional vector space. Vectorizing them allows us to measure their similarities and cluster them. For semantic role labeling, we need to know the relative position of verbs; hence the features can include prefix, suffix, distance from verbs in the sentence, etc. However, each feature has a corresponding vector representation in *d* dimensional space learned from the training data.

Recently, convolutional neural network (CNN) is being used for subjectivity detection. In particular, (Kalchbrenner and Blunsom 2013) used recurrent CNNs. These show high accuracy on certain datasets such as Twitter; we are also concerned with a specific sentence within the context of the previous discussion. The order of the sentences preceding the one at hand results in a sequence of sentences also known as a time series of sentences (Kalchbrenner and Blunsom 2013). However, their model suffers from overfitting; hence in this work we consider deep convolutional neural networks, where temporal information is modeled via dynamic Gaussian Bayesian networks.

Aspect-Based Sentiment Analysis

Aspect extraction from opinions was first studied by Hu and Liu (2004). They introduced the distinction between explicit and implicit aspects. However, the authors only dealt with explicit aspects and used a set of rules based on statistical observations. Hu and Liu's method was later improved by Popescu and Etzioni (2005) and by Blair-Goldensohn et al. (2008). Popescu and Etzioni (2005) assumed the product class is known in advance. Their algorithm detects whether a noun or noun phrase is a product feature by computing the point-wise mutual information between the noun phrase and the product class.

Scaffidi et al. (2007) presented a method that uses language model to identify product features. They assumed that product features are more frequent in product reviews than in a general natural language text. However, their method seems to have low precision since retrieved aspects are affected by noise. Some methods treated the aspect term extraction as sequence labeling and used CRF for that. Such methods have performed very well on the datasets even in cross-domain experiments (Jakob and Gurevych 2010; Zhiqiang and Wenting 2014).

Topic modeling has been widely used as a basis to perform extraction and grouping of aspects (Chen and Liu 2014; Hu et al. 2014). Two models were considered: pLSA (Hofmann 1999) and LDA (Blei et al. 2003). Both models introduce a latent variable "topic" between the observable variables "document" and "word" to analyze the semantic topic distribution of documents. In topic models, each document is represented as a random mixture over latent topics, where each topic is characterized by a distribution over words.

Such methods have been gaining popularity in social media analysis like emerging political topic detection in Twitter (Rill et al. 2014). The LDA model defines a Dirichlet probabilistic generative process for document-topic distribution; in each document, a latent aspect is chosen according to a multinomial distribution, controlled by a Dirichlet prior α. Then, given an aspect, a word is extracted according to another multinomial distribution, controlled by another Dirichlet prior β. Among existing works employing these models are the extraction of global aspects (such as the brand of a product) and local aspects (such as the property of a product (Titov and McDonald 2008)), the extraction of key phrases (Branavan et al. 2009), the rating of multi-aspects (Wang et al. 2010), and the summarization of aspects and sentiments (Lu et al. 2009). Zhao et al. (2010) employed the maximum entropy method to train a switch variable based on POS tags of words and used it to separate aspect and sentiment words.

Mcauliffe and Blei (2008) added user feedback to LDA as a response variable related to each document. Lu and Zhai (2008) proposed a semi-supervised model. DF-LDA (Andrzejewski et al. 2009) also represents a semi-supervised model, which allows the user to set must-link and cannot-link constraints. A must-link constraint means that two terms must be in the same topic, while a cannot-link constraint means that two terms cannot be in the same topic. Poria et al. (2016c) integrated commonsense in the calculation of word distributions in the LDA algorithm, thus enabling the shift from syntax to semantics in aspect-based sentiment analysis. Wang et al. (2014) proposed two semi-supervised models for product aspect extraction based on the use of seeding aspects. In the category of supervised methods, (Jagarlamudi et al. 2012) employed seed words to guide topic models to learn topics of specific interest to a user, while (Wang et al. 2010) and (Mukherjee and Liu 2012) employed seeding words to extract related product aspects from product reviews. On the other hand, recent approaches using deep CNNs (Collobert et al. 2011b; Poria et al. 2015a) showed significant performance improvement over the state-of-the-art methods on a range of NLP tasks. Collobert et al. (2011b) fed word embeddings to a CNN to solve standard NLP problems such as named entity recognition (NER), part-of-speech (POS) tagging, and semantic role labeling.

Preliminaries

In this section, we briefly review the theoretical concepts necessary to comprehend the present work. We begin with a description of maximum likelihood estimation of edges in dynamic Gaussian Bayesian networks where each node is a word in a sentence. Next, we show that weights in the CNN can be learned by minimizing a global error function that corresponds to an exponential distribution over a linear combination of input sequence of word features.

Notations: Consider a Gaussian network (GN) with time delays which comprises a set of N nodes and observations gathered over T instances for all the nodes. Nodes can take real values from a multivariate distribution determined by the parent set. Let the dataset of samples be $X = \{x_i(t)\}_{\mathrm{N}\times\mathrm{T}}$, where $x_i(t)$ represents the sample value of the ith random variable in instance t. Lastly, let a be the set of parent variables regulating variable i.

Gaussian Bayesian Networks

In tasks where one is concerned with a specific sentence within the context of the previous discourse, capturing the order of the sequences preceding the one at hand may be particularly crucial.

We take as given a sequence of sentences $s(1)$, $s(2)$, ..., $s(T)$, each in turn being a sequence of words so that $s(t) = (x_1(t), x_2(t), ..., x_{\mathrm{L}}(t))$, where L is the length of sentence $s(t)$. Thus, the probability of a word $p(x_i(t))$ follows the distribution:

$$p(x_i(t)) = P\big(x_i(t) \mid \big(x_1(t), x_2(t), \ldots, x_{i-1}(t)\big), \big(s(1), s(2), \ldots, s(t-1)\big) \qquad (1)$$

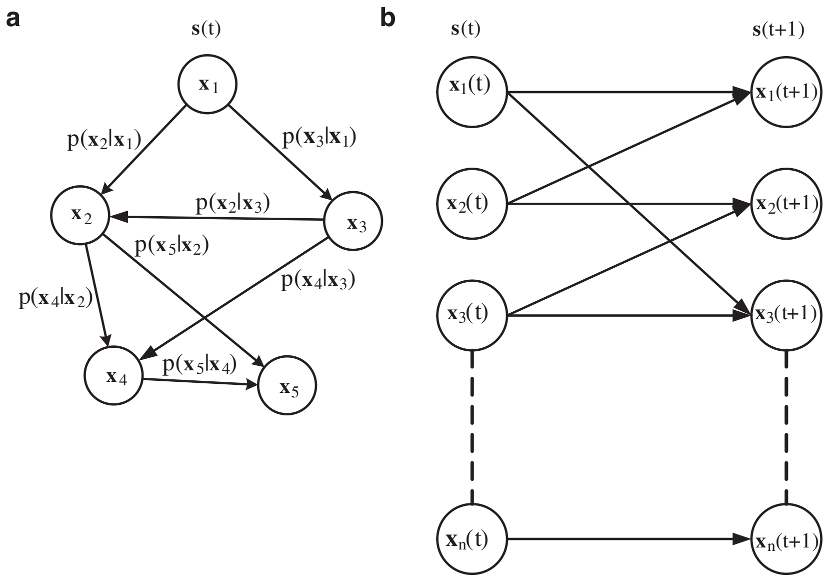
A Bayesian network is a graphical model that represents a joint multivariate probability distribution for a set of random variables (Prinzie and Van den Poel 2009). It is a directed acyclic graph S with a set of parameters θ that represents the strengths of connections by conditional probabilities.

The BN decomposes the likelihood of node expressions into a product of conditional probabilities by assuming independence of non-descendant nodes, given their parents.

$$p(X \mid S, \theta) = \prod_{i=1}^{\mathrm{N}} p\big(\mathrm{x}_i \mid a_i, \theta_{i,a_i}\big), \qquad (2)$$

where $p(\mathrm{x}_i|a_i,\theta_{i,ai})$ denotes the conditional probability of node expression x_i given its parent node expressions a_i and $\theta_{i,ai}$ denotes the maximum likelihood(ML) estimate of the conditional probabilities.

Figure 1a illustrates the state space of a Gaussian Bayesian network (GBN) at time instant t where each node $x_i(t)$ is a word in the sentence $s(t)$. The connections represent causal dependencies over one or more time instants. The observed state vector of variable i is denoted as x_i and the



Sentiment Analysis, Basic Tasks of, Fig. 1 State space of different Bayesian models

conditional probability of variable *i* given variable *j* is $p(\mathrm{x}_i|\mathrm{x}_j)$. The optimal Gaussian network S^* is obtained by maximizing the posterior probability of S given the data X. From Bayes theorem, the optimal Gaussian network S^* is given by

$$S^* = \arg\max_S p(S|X) = \arg\max_S p(S)p(X|S), \tag{3}$$

where $p(S)$ is the probability of the Gaussian network and $p(X|S)$ is the likelihood of the expression data given the Gaussian network.

Given the set of conditional distributions with parameters $\theta = \{\theta_{i,a_i}\}_{i=1}^{\mathrm{N}}$ the likelihood of the data is given by

$$p(X|S) = \int p(X|S,\theta)p(\theta|S)d\theta, \tag{4}$$

To find the likelihood in Eq. 4 and to obtain the optimal Gaussian network as in Eq. 3, Gaussian BN assumes that the nodes are multivariate Gaussian. That is, expression of node *i* can be described with mean μ_i and covariance matrix $\sum_i$ of size N × N. The joint probability of the network can be the product of a set of conditional probability distributions given by

$$p(\mathrm{x}_i|a_i) = \theta_{i,a_i} \sim \mathcal{N}\left(\mu_i + \sum_{j\in a_i}(\mathrm{x}_j - \mu_j)\beta, \Sigma_i'\right), \tag{5}$$

where $\sum_i' = \sum_i - \sum_{i,a_i}\sum_{a_i}^{-1}\sum_{i,a_i}^{T}$ and β denotes the regression coefficient matrix, $\sum_i'$ is the conditional variance of x_i given its parent set a_i, $\sum_{i,a_i}$ is the covariance between observations of x, and the variables in a_i, and $\sum_{a_i}$ is the covariance matrix of a_i. The acyclic condition of BN does not allow feedback among nodes, and feedback is an essential characteristic of real-world GN.

Therefore, dynamic Bayesian networks have recently become popular in building GN with time delays mainly due to their ability to model causal interactions as well as feedback regulations (Friedman et al. 1998). A first-order dynamic BN is defined by a transition network of interactions between a pair of Gaussian networks connecting nodes at time instants τ and τ + 1. In time instant τ + 1, the parents of nodes are those specified in the time instant τ. Similarly, the Gaussian network of a R-order dynamic system is represented by a Gaussian network comprising (R + 1) consecutive time points and N nodes or a graph of (R + 1) × N nodes. In practice, the sentence data is transformed to a BOW model where each sentence is a vector of frequencies for each word in the vocabulary. Figure 1b illustrates the state space of a first-order Dynamic GBN models transition networks among words in sentences $s(t)$ and $s(t + 1)$ in consecutive time points, the lines correspond to first-order edges among the words learned using BOW.

Hence, a sequence of sentences results in a time series of word frequencies. It can be seen that such a discourse model produces compelling discourse vector representations that are sensitive to the structure of the discourse and promise to capture subtle aspects of discourse comprehension, especially when coupled to further semantic data and unsupervised pre-training.

Convolutional Neural Networks

The idea behind convolution is to take the dot product of a vector of k weights w_k also known as kernel vector with each k-gram in the sentence $s(t)$ to obtain another sequence of features $c(t) = (c_1(t), c_2(t), \ldots, c_L(t))$.

$$c_j = {w_k}^T \cdot \mathrm{x}_{i:i+k-1} \tag{6}$$

We then apply a max-pooling operation over the feature map and take the maximum value $\hat{c}(t) = \max\{c(t)\}$ as the feature corresponding to this particular kernel vector. Similarly, varying kernel vectors and window sizes are used to obtain multiple features (Kalchbrenner et al. 2014).

For each word $x_i(t)$ in the vocabulary, an d dimensional vector representation is given in a lookup table that is learned from the data (Collobert et al. 2011a). The vector representation of a sentence is hence a concatenation of vectors for individual words. Similarly, we can have lookup tables for other features. One might want to provide features other than words if these features are suspected to be helpful. Now, the convolution kernels are applied to word vectors instead of individual words.

We use these features to train higher layers of the CNN that can represent bigger groups of words in sentences. We denote the feature learned at hidden neuron h in layer l as F_h^l. Multiple features may be learned in parallel in the same CNN layer. The features learned in each layer are used to train the next layer

$$F^l = \sum\nolimits_{h=1}^{n_h} w_l^h * F^{l-1} \tag{7}$$

where * indicates convolution and W_k is a weight kernel for hidden neuron h and n_h is the total number of hidden neurons. Training a CNN becomes difficult as the number of layers increases, as the Hessian matrix of second-order derivatives often does not exist. Recently, deep learning has been used to improve the scalability of a model that has inherent parallel computation. This is because hierarchies of modules can provide a compact representation in the form of input-output pairs. Each layer tries to minimize the error between the original state of the input nodes and the state of the input nodes predicted by the hidden neurons.

This results in a downward coupling between modules. The more abstract representation at the output of a higher layer module is combined with the less abstract representation at the internal nodes from the module in the layer below. In the next section, we describe deep CNN that can have arbitrary number of layers.

Convolution Deep Belief Network

A deep belief network (DBN) is a type of deep neural network that can be viewed as a composite of simple, unsupervised models such as restricted Boltzmann machines (RBMs) where each RBMs hidden layer serves as the visible layer for the next RBM (Chaturvedi et al. 2016). RBM is a bipartite graph comprising two layers of neurons: a visible and a hidden layer; it is restricted such that the connections among neurons in the same layer are not allowed. To compute the weights W of an RBM, we assume that the probability distribution over the input vector x is given as

$$p(x|\,W) = \frac{1}{Z(W)}\exp^{-\mathrm{E}(x;\,\mathrm{W})} \tag{8}$$

where $Z(W) = \sum_{\mathrm{x}} \exp^{-\mathrm{E}}(\mathrm{x}; W)$ is a normalization constant. Computing the maximum likelihood is difficult as it involves solving the normalization constant, which is a sum of an exponential number of terms. The standard approach is to approximate the average over the distribution with an average over a sample from $p(x|W)$, obtained by Markov chain Monte Carlo until convergence.

To train such a multilayer system, we must compute the gradient of the total energy function E with respect to weights in all the layers. To learn these weights and maximize the global energy function, the approximate maximum likelihood contrastive divergence (CD) approach can be used. This method employs each training sample to initialize the visible layer. Next, it uses the Gibbs sampling algorithm to update the hidden layer and then reconstruct the visible layer consecutively, until convergence (Hinton 2002). As an example, here we use a logistic regression model to learn the binary hidden neurons, and each visible unit is assumed to be a sample from a normal distribution (Taylor et al. 2007).

The continuous state $\widehat{h}_j$ of the hidden neuron j, with bias b_j, is a weighted sum over all continuous visible nodes v and is given by

$$\widehat{h}_j = b_j + \sum_i v_i w_{ij}, \qquad (9)$$

where W_{ij} is the connection weight to hidden neuron j from visible node v_i. The binary state h_j of the hidden neuron can be defined by a sigmoid activation function:

$$h_j = \frac{1}{1 + e^{-\widehat{h}_j}}. \qquad (10)$$

Similarly, in the next iteration, the binary state of each visible node is reconstructed and labeled as v_{recon}. Here, we determine the value to the visible node i, with bias c_i, as a random sample from the normal distribution where the mean is a weighted sum over all binary hidden neurons and is given by

$$\widehat{v}_i = c_i + \sum_j h_i w_{ij}, \qquad (11)$$

where W_{ij} is the connection weight to hidden neuron j from visible node v_i. The continuous state v_i is a random sample from $\mathcal{N}(\widehat{v}_i, \sigma)$, where σ is the variance of all visible nodes. Lastly, the weights are updated as the difference between the original and reconstructed visible layer using

$$\Delta w_{ij} = \alpha\left(< v_i h_j >_{\text{data}} - < v_i h_j >_{recon}\right), \qquad (12)$$

where α is the learning rate and $<V_i h_j>$ is the expected frequency with which visible unit i and hidden unit j are active together when the visible vectors are sampled from the training set and the hidden units are determined by Eq. 9. Finally, the energy of a DNN can be determined in the final layer using $\mathrm{E} = -\sum_{i,j} v_i h_j w_{ij}$.

To extend the deep belief networks to convolution deep belief network (CDBN), we simply partition the hidden layer into Z groups. Each of the Z groups is associated with a $k \times d$ filter where k is the width of the kernel and d is the number of dimensions in the word vector. Let us assume that the input layer has dimension $\mathrm{L} \times d$ where L is the length of the sentence. Then the convolution operation given by Eq. 6 will result in a hidden layer of Z groups each of dimension $(\mathrm{L}-k + 1)\times(d-d + 1)$. These learned kernel weights are shared among all hidden units in a particular group. The energy function is now a sum over the energy of individual blocks given by

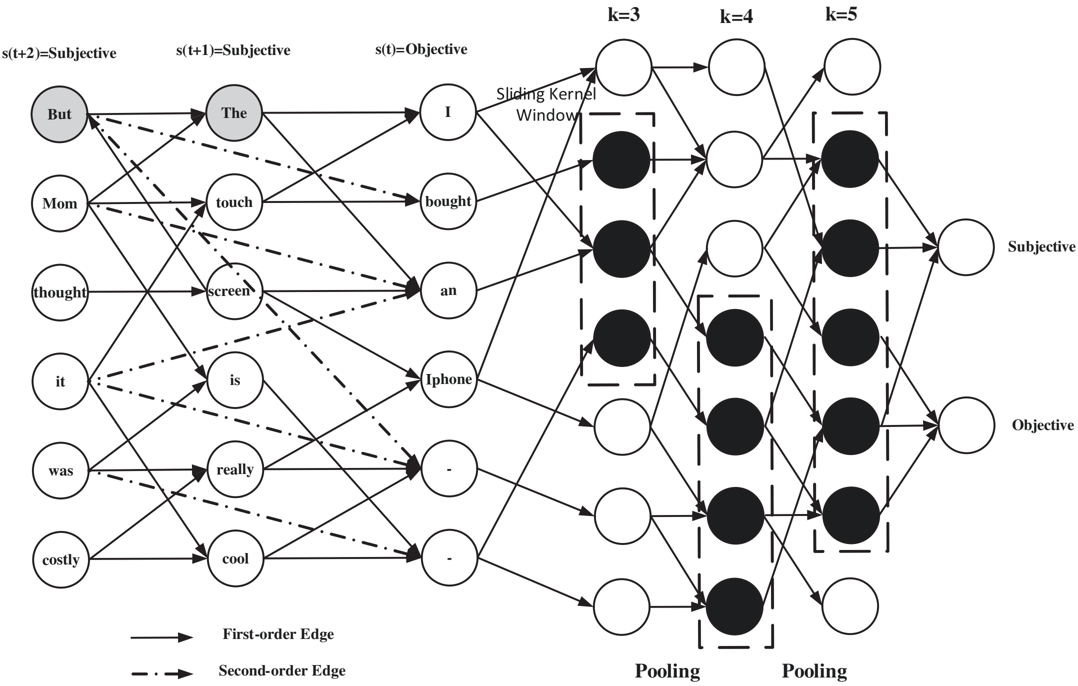
$$\mathrm{E} = -\sum_{z=1}^{Z} \sum_{i,j}^{\mathrm{L}-k+1,1} \sum_{r,s}^{k,d} v_{i+r-1,j+s-1} h_{ij}^{z} w_{rs}^{k} \qquad (13)$$

The CNN sentence model preserve the order of words by adopting convolution kernels of gradually increasing sizes that span an increasing number of words and ultimately the entire sentence (Kalchbrenner and Blunsom 2013). However, several word dependencies may occur across sentences; hence, in this work we propose a Bayesian CNN model that uses dynamic Bayesian networks to model a sequence of sentences.

Deep Learning Algorithm

Subjectivity Detection

In this work, we integrate a higher-order GBN for sentences into the first layer of the CNN. The



Sentiment Analysis, Basic Tasks of, Fig. 2 State space of Bayesian CNN where the input layer is pre-trained using a dynamic GBN

GBN layer of connections β is learned using maximum likelihood approach on the BOW model of the training data. The input sequence of sentences $s(t : t - 2)$ are parsed through this layer prior to training the CNN. Only sentences or groups of sentences containing high ML motifs are then used to train the CNN. Hence, motifs are convolved with the input sentences to generate a new set of sentences for pre-training.

$$F^0 = \sum_{h=1}^{M} \beta^h * s \tag{14}$$

where M is the number of high ML motifs and s is the training set of sentences in a particular class.

Figure 2 illustrates the state space of Bayesian CNN where the input layer is pre-trained using a dynamic GBN with up to two time point delays shown for three sentences in a review on iPhone. The dashed lines correspond to second-order edges among the words learned using BOW. Each hidden layer does convolution followed by pooling across the length of the sentence. To preserve the order of words, we adopt kernels of increasing sizes.

Since the number of possible words in the vocabulary is very large, we consider only the top subjectivity clue words to learn the GBN layer. Lastly, in order to preserve the context of words in conceptual phrases such as “touchscreen,” we consider additional nodes in the Bayesian network for phrases with subjectivity clues. Further, the word embeddings in the CNN are initialized using the log-bilinear language model (LBL) where the d dimensional vector representation of each word $x_i(t)$ in Eq. 2 is given by

$$x_i(t) = \sum_{k=1}^{i-1} C_k x_k(t) \tag{15}$$

where C_k are the $d \times d$ co-occurrence or context matrices computed from the data.

The time series of sentences is used to generate a subset of sentences containing high ML motifs using Eq. 14. The frequency of a sentence in the new dataset will also correspond to the corresponding number of high ML motifs in the sentence. In this way, we are able to increase the weights of the corresponding causal features among words and concepts extracted using Gaussian Bayesian networks.

The new set of sentences is used to pre-train the deep neural network prior to training with the complete dataset. Each sentence can be divided into chunks or phrases using POS taggers. The phrases have hierarchical structures and combine in distinct ways to form sentences. The *k*-gram kernels learned in the first layer hence correspond to a chunk in the sentence.

Aspect Extraction

In order to train the CNN for aspect extraction, instead, we used a special training algorithm suitable for sequential data, proposed by Collobert et al. (2011b). We will summarize it here, mainly following (Fonseca and Rosa 2013). The algorithm trains the neural network by back-propagation in order to maximize the likelihood over training sentences. Consider the network parameter θ. We say that h_y is the output score for the likelihood of an input x to have the tag y. Then, the probability to assign the label y to x is calculated as

$$p(y|\, x,\theta) = \frac{\exp(h_y)}{\sum_j \exp(h_j)}. \qquad (16)$$

Define the log-add operation as

$$\underset{i}{\operatorname{logadd}}\, h_i = \log \sum_i \exp\, h_i, \qquad (17)$$

then for a training example, the log-likelihood becomes

$$\log p(y|\, x,\theta) = h_y - \underset{i}{\operatorname{logadd}}\, h_i. \qquad (18)$$

In aspect term extraction, the terms can be organized as chunks and are also often surrounded by opinion terms. Hence, it is important to consider sentence structure on a whole in order to obtain additional clues. Let it be given that there are T tokens in a sentence and y is the tag sequence while $h_{t,\ i}$ is the network score for the f-th tag having z-th tag. We introduce $A_{i,\ j}$ transition score from moving tag i to tag j. Then, the score tag for the sentence s to have the tag path y is defined by

$$s(x,y,\theta) = \sum_{t=1}^{T} \left(h_{t,y_t} + A_{y_{t-1},y_t}\right). \qquad (19)$$

This formula represents the tag path probability over all possible paths. Now, from Eq. 17 we can write the log-likelihood

$$\log p(y|\, x,\theta) = s(x,y,\theta) - \underset{\forall j}{\operatorname{logadd}}\, s(x,j,\theta). \qquad (20)$$

The number of tag paths has exponential growth. However, using dynamic programming techniques, one can compute in polynomial time the score for all paths that end in a given tag (Collobert et al. 2011b). Let y_t^k denote all paths that end with the tag k at the token t. Then, using recursion, we obtain

$$\begin{aligned}\delta_t(k) = \underset{\forall y_t^k}{\operatorname{logadd}}\, s\left(x,y_t^k,\theta\right) = h_{t,k} \\ + \underset{j}{\operatorname{logadd}}\, \delta_{t-1}(j) + A_{j,k}.\end{aligned} \qquad (21)$$

For the sake of brevity, we shall not delve into details of the recursive procedure, which can be found in Collobert et al. (2011b). The next equation gives the log-add for all the paths to the token T:

$$\underset{\forall y}{\operatorname{logadd}}\, s(x,y,\theta) = \underset{i}{\operatorname{logadd}}\, \delta_T(i). \qquad (22)$$

Using these equations, we can maximize the likelihood of Eq. 20 over all training pairs. For inference, we need to find the best tag path using the Viterbi algorithm, e.g., we need to find the best tag path that minimizes the sentence score Eq. 19.

The features of an aspect term depend on its surrounding words. Thus, we used a window of 5 words around each word in a sentence, i.e., ±2 words. We formed the local features of that window and considered them to be features of the middle word. Then, the feature vector was fed to a CNN.

The network contained one input layer, two convolution layers, two max-pool layers, and a fully connected layer with softmax output. The first convolution layer consisted of 100 feature maps with filter size 2. The second convolution layer had 50 feature maps with filter size 3. The stride in each convolution layer is 1 as we wanted to tag each word. A max-pooling layer followed each convolution layer. The pool size we use in the max-pool layers was 2. We used regularization with dropout on the penultimate layer with a constraint on L2-norms of the weight vectors, with 30 epochs. The output of each convolution layer was computed using a nonlinear function; in our case we used tanh.

As features, we used word embeddings trained on two different corpora. We also used some additional features and rules to boost the accuracy; see section "Features and Rules Used." The CNN produces local features around each word in a sentence and then combines these features into a global feature vector. Since the kernel size for the two convolution layers was different, the dimensionality $L_x \times L_y$ mentioned in section "Convolutional Deep Belief Network" was 3×300 and 2×300, respectively. The input layer was 65×300, where 65 was the maximum number of words in a sentence and 300 the dimensionality of the word embeddings used, per each word.

The process was performed for each word in a sentence. Unlike traditional max-likelihood leaning scheme, we trained the system using propagation after convolving all tokens in the sentence. Namely, we stored the weights, biases, and features for each token after convolution and only back-propagated the error in order to correct them once all tokens were processed using the training scheme described in this section on "Aspect Extraction."

If a training instance s had n words, then we represented the input vector for that instance as $s_{1:n} = s_1 \bigoplus s_2 \bigoplus \ \ldots \ \bigoplus s_n$. Here, $s_i \in \Re^k$ is a k-dimensional feature vector for the word *Si*. We found that this network architecture produced good results on both of our benchmark datasets. Adding extra layers or changing the pooling size and window size did not contribute to the accuracy much and, instead, only served to increase computational cost.

Evaluation

Subjectivity Detection

Datasets Used

We use the MPQA corpus (Wiebe and Riloff 2005), a collection of 535 English news articles from a variety of sources manually annotated with subjectivity flag. From the total of 9700 sentences in this corpus, 55% of the sentences are labeled as subjective, while the rest are objective. We also compare with the Movie Review (MR) benchmark dataset (Pang and Lee 2004) that contains 5000 subjective movie review snippets from Rotten Tomatoes website and another 5000 objective sentences from plot summaries available from the Internet Movies Database. All sentences are at least ten words long and drawn from reviews or plot summaries of movies released post 2001.

The data pre-processing included removing top 50 stop-words and punctuation marks from the sentences. Next, we used a POS tagger to determine the part-of-speech for each word in a sentence. Subjectivity clues dataset (Riloff and Wiebe 2003) contains a list of over 8000 clues identified manually as well as automatically using both annotated and unannotated data. Each clue is a word and the corresponding part of speech.

The frequency of each clue was computed in both subjective and objective sentences of the MPQA corpus. Here we consider the top 50 clue words with highest frequency of occurrence in the subjective sentences. We also extracted 25 top concepts containing the top clue words using the method described in Poria et al. (2015b). The CNN is collectively pre-trained with both subjective and objective sentences that contain high ML word and concept motifs. The word vectors are initialized using the LBL model and a context window of size 5 and 30 features. Each sentence is wrapped to a window of 50 words to reduce the number of parameters and hence the overfitting of the model. A CNN with three hidden layers of 100 neurons and kernels of size {3,4,5} is used. The output layer corresponds to two neurons for each class of sentiments.

Sentiment Analysis, Basic Tasks of, Table 1 F-measure by different models for classifying sentences in a document as subjective and objective in MPQA and MR dataset

Dataset	NBSVM	CNN-MC	SWSD	UWSD	BCDBN
MPQA	86.3	89.4	80.35	60	**93.2**
MR	93.2	93.6	–	55	**96.4**

Experimental Results

We used tenfold cross-validation to determine the accuracy of classifying new sentences using the trained CNN classifier. A comparison is done with classifying the time series data using baseline classifiers such as Naive Bayes SVM (NBSVM) (Wang and Manning 2013), Multichannel CNN (CNN-MC) (Kim 2014), Subjectivity Word Sense Disambiguation (SWSD) (Ortega et al. 2013), and Unsupervised-WSD (UWSD) (Akkaya et al. 2009). Table 1 shows that BCDBN outperforms previous methods by 5–10% in accuracy on both datasets. Almost 10% improvement is observed over NBSVM on the movie review dataset. In addition, we only consider word vectors of 30 features instead of the 300 features used by CNN-MC and hence are 10 times faster.

Aspect Extraction

Datasets Used

In this subsection, we present the data used in our experiments.

Google Embeddings

Mikolov et al. (2013) presented two different neural network models for creating word embeddings. The models were log-linear in nature, trained on large corpora. One of them is a bag-of-word-based model called CBOW; it uses word context in order to obtain the embeddings. The other one is called skip-gram model; it predicts the word embeddings of surrounding words given the current word. Those authors made a dataset called word2vec publicly available. These 300-dimensional vectors were trained on a 100-billion-word corpus from Google News using the CBOW architecture.

Sentiment Analysis, Basic Tasks of, Table 2 SemEval Data used for evaluation

Domain	Training	Test	Total
Laptop	3041	800	3841
Restaurant	3045	800	3845
Total	6086	1600	7686

Our Amazon Embeddings

We trained the CBOW architecture proposed by Mikolov et al. (2013) on a large Amazon product review dataset developed by McAuley and Leskovec (2013). This dataset consists of 34,686,770 reviews (4.7 billion words) of 2,441,053 Amazon products from June 1995 to March 2013. We kept the word embeddings 300-dimensional (http://sentic.net/AmazonWE.zip). Due to the nature of the text used to train this model, this includes opinionated/affective information, which is not present in ordinary texts such as the Google News corpus.

Evaluation Corpora

For training and evaluation of the proposed approach, we used two corpora:

- Aspect-based sentiment analysis dataset developed by Qiu et al. (2011)
- SemEval 2014 dataset. The dataset consists of training and test sets from two domains, Laptop and Restaurant; see Table 2.

The annotations in both corpora were encoded according to IOB2, a widely used coding scheme for representing sequences. In this encoding, the first word of each chunk starts with a “B-Type” tag, “I-Type” is the continuation of the chunk, and “O” is used to tag a word which is out of the chunk. In our case, we are interested to determine whether a word or chunk is an aspect, so we only have “B–A,” “I–A,” and “O” tags for the words.

Here is an example of IOB2 tags:

also/O excellent/O operating/B-A system/I-A ,/O size/B-A and/O weight/B-A for/O optimal/O mobility/B-A excellent/O durability/B-A of/O *the/O battery/B-A the/O functions/O provided/ O by/O the/O trackpad/B-A is/O unmatched/O by/O any/O other/O brand/O*

Features and Rules Used

In this section, we present the features, the representation of the text, and linguistic rules used in our experiments.

We used the following the features:

- **Word embeddings.** We used the word embeddings described earlier as features for the network. This way, each word was encoded as 300-dimensional vector, which was fed to the network.
- **Part of speech tags.** Most of the aspect terms are either nouns or noun chunk. This justifies the importance of POS features. We used the POS tag of the word as its additional feature. We used six basic parts of speech (noun, verb, adjective, adverb, preposition, conjunction) encoded as a 6-dimensional binary vector. We used Stanford Tagger as a POS tagger.

These two features vectors were concatenated and fed to CNN. So, for each word the final feature vector is 306 dimensional.

In some of our experiments, we used a set of linguistic patterns (LPs) derived from sentic patterns (LP) (Poria et al. 2015b), a linguistic framework based on SenticNet (Cambria et al. 2016). SenticNet is a concept-level knowledge base for sentiment analysis built by means of sentic computing (Cambria and Hussain 2015), a multidisciplinary approach to natural language processing and understanding at the crossroads between affective computing, information extraction, and commonsense reasoning, which exploits both computer and human sciences to better interpret and process social information on the Web. In particular, we used the following linguistic rules:

Rule 1 Let a noun *h* be a subject of a word *t*, which has an adverbial or adjective modifier present in a large sentiment lexicon, SenticNet. Then mark *h* as an aspect.

Rule 2 Except when the sentence has an auxiliary verb, such as *is*, *was*, *would*, *should*, *could*, etc., we apply:

Rule 2.1 If the verb *t* is modified by an adjective or adverb or is in adverbial clause modifier relation with another token, then mark *h* as an aspect. For example, in “The battery lasts little,” *battery* is the subject of *lasts*, which is modified by an adjective modifier *little*, so *battery* is marked as an aspect.

Rule 2.2 If *t* has a direct object, a noun *n*, not found in SenticNet, then mark *n* an aspect, as, e.g., in “I like the lens of this camera.”

Rule 3 If a noun *h* is a complement of a copular verb, then mark *h* as an explicit aspect. For example, in “The camera is nice,” *camera* is marked as an aspect.

Rule 4 If a term marked as an aspect by the CNN or the other rules is in a noun-noun compound relationship with another word, then instead form one aspect term composed of both of them. For example, if in “battery life,” “battery” or “life” is marked as an aspect, then the whole expression is marked as an aspect.

Rule 5 The above rules 1–4 improve recall by discovering more aspect terms. However, to improve precision, we apply some heuristics: e.g., we remove stop-words such as *of*, *the*, *a*, etc., even if they were marked as aspect terms by the CNN or the other rules.

We used the Stanford parser to determine syntactic relations in the sentences.

We combined LPs with the CNN as follows: both LPs and CNN-based classifier are run on the text; then all terms marked by any of the two classifiers are reported as aspect terms, except for those unmarked by the last rule.

Experimental Results

Table 4 shows the accuracy of our aspect term extraction framework in laptop and restaurant domains. The framework gave better accuracy on restaurant domain reviews, because of the lower variety of aspect available terms than in laptop domain. However, in both cases recall was lower than precision.

Sentiment Analysis, Basic Tasks of, Table 3 Random features vs. Google embeddings vs. Amazon embeddings on the SemEval 2014 dataset

Domain	Feature	F-score (%)
Laptop	Random	71.21
Laptop	Google embeddings	77.32
Laptop	Amazon embeddings	**80.68**
Restaurant	Random	77.05
Restaurant	Google embeddings	83.50
Restaurant	Amazon embeddings	**85.70**

Table 4 shows improvement in terms of both precision and recall when the POS feature is used. Pre-trained word embeddings performed better than randomized features (each word's vector initialized randomly); see Table 3. Amazon embeddings performed better than Google word2vec embeddings. This supports our claim that the former contains opinion-specific information which helped it to outperform the accuracy of Google embeddings trained on more formal text – the Google news corpus. Because of this, in the sequel we only show the performance using Amazon embeddings, which we denote simply as WE (word embeddings).

In both domains, CNN suffered from low recall, i.e., it missed some valid aspect terms. Linguistic analysis of the syntactic structure of the sentences substantially helped to overcome some drawbacks of machine-learning-based analysis. Our experiments showed good improvement in both precision and recall when LPs were used together with CNN; see Table 5.

As to the LPs, the removal of stop-words, Rule 1, and Rule 3 were most beneficial. Figure 4 shows a visualization for the Table 5. Table 6 and Figure 3 show the comparison between the proposed method and the state of the art on the Semeval dataset. It is noted that about 36.55% aspect terms present in the laptop domain corpus are phrase, and restaurant corpus consists of 24.56% aspect terms. The performance of detecting aspect phrases is lower than single word aspect tokens in both domains. This shows that the sequential tagging is indeed a tough task to do. Lack of sufficient training data for aspect phrases is also one of the reasons to get lower accuracy in this case (Table 7).

Sentiment Analysis, Basic Tasks of, Table 4 Feature analysis for the CNN classifier

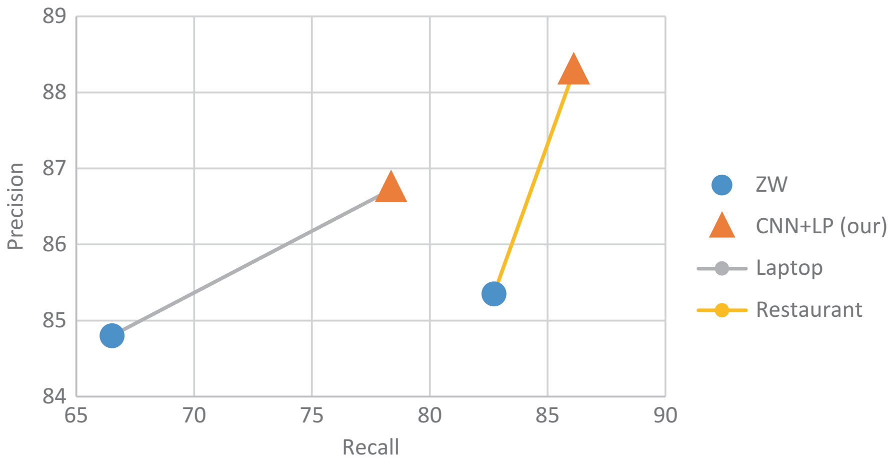
Domain	Features	Recall (%)	Precision (%)	F-score (%)
Laptop	WE	75.20	86.05	80.68
Laptop	WE+POS	**76.31**	**86.46**	**81.06**
Restaurant	WE	84.11	87.35	85.70
Restaurant	WE+POS	**85.01**	**87.42**	**86.20**

Sentiment Analysis, Basic Tasks of, Table 5 Impact of Sentic Patterns on the SemEval 2014 dataset

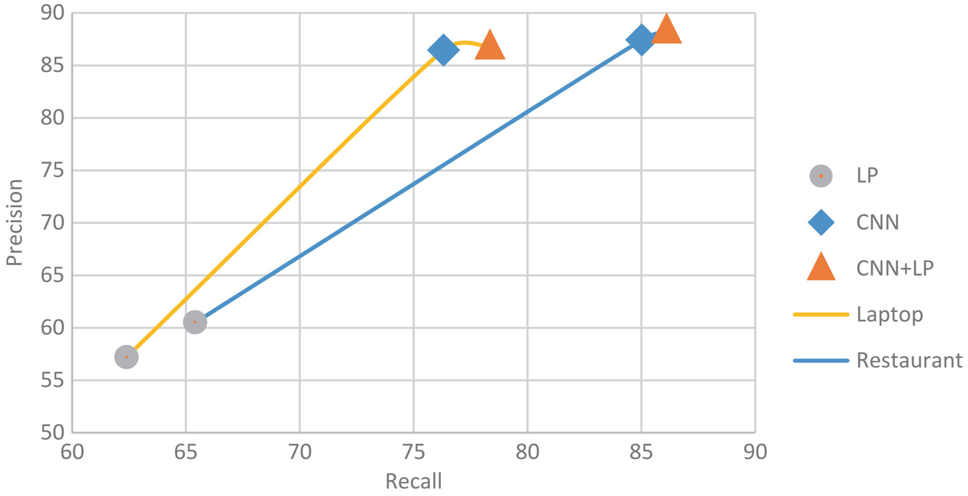
Domain	Classifiers	Recall (%)	Precision (%)	F-score (%)
Laptop	LP	62.39	57.20	59.68
Laptop	CNN	76.31	86.46	81.06
Laptop	CNN+LP	**78.35**	**86.72**	**82.32**
Restaurant	LP	65.41	60.50	62.86
Restaurant	CNN	85.01	87.42	86.20
Restaurant	CNN+LP	**86.10**	**88.27**	**87.17**

In particular, we got 79.20% and 83.55% F-score to detect aspect phrases in laptop and restaurant domain, respectively. We observed some cases where only 1 term in an aspect phrase is detected as aspect term. In those cases Rule 4 of the LPs helped to correctly detect the aspect phrases. We also carried out experiments on the aspect dataset originally developed by Qiu et al. (2011). This is to date the largest comprehensive aspect-based sentiment analysis dataset. The best accuracy on this dataset was obtained when word embedding features were used together with the POS features. This shows that while the word embedding features are most useful, the POS feature also plays a major role in aspect extraction.

As on the SemEval dataset, LPs together with CNN increased the overall accuracy. However, LPs have performed much better on this dataset than on the SemEval dataset. This supports the observation made previously (Qiu et al. 2011) that on this dataset LPs are more useful. One of the possible reasons for this is that most of the sentences in this dataset are grammatically correct and contain only one aspect term. Here we combined LPs and a CNN to achieve even better results than the approach of by Qiu et al. (2011) based only on LPs. Our experimental results showed that this



Sentiment Analysis, Basic Tasks of, Fig. 3 Comparison of the performance with the state of the art



Sentiment Analysis, Basic Tasks of, Fig. 4 Comparison between the performance of CNN, CNN-LP, and LP on Bing Liu dataset

Sentiment Analysis, Basic Tasks of, Table 6 Comparison with the state of the art. ZW stands for (Zhiqiang and Wenting 2014); LP stands for sentic patterns

Domain	Framework	Recall (%)	Precision (%)	F-score (%)
Laptop	ZW	66.51	84.80	74.55
Laptop	CNN+LP	**78.35**	**86.72**	**82.32**
Restaurant	ZW	82.72	85.35	84.01
Restaurant	CNN+LP	**86.10**	**88.27**	**87.17**

ensemble algorithm (CNN+LP) can better understand the semantics of the text than (Qiu et al. 2011)'s pure LP-based algorithm and thus extracts more salient aspect terms. Table 8 and Figure 5 show the performance and comparisons of different frameworks.

Figure 6 compares the proposed method with the state of the art. We believe that there are two key reasons for our framework to outperform

Sentiment Analysis, Basic Tasks of, Table 7 Impact of the POS feature on the dataset by Qiu et al. (2011)

Domain	Classifiers	Precision (%)	Recall (%)	F-score (%)
Canon	WE	82.74	75.15	78.76
Canon	WE+POS	**85.42**	**77.21**	**81.10**
Nikon	WE	73.19	79.27	76.10
Nikon	WE+POS	**77.65**	**82.30**	**79.90**
DVD	WE	84.41	77.26	80.67
DVD	WE+POS	**85.48**	**79.25**	**82.24**
Mp3	WE	87.35	81.23	84.17
Mp3	WE+POS	**89.40**	**83.77**	**86.49**
Cellphone	WE	86.01	81.32	83.59
Cellphone	WE+POS	**90.15**	**83.47**	**86.68**

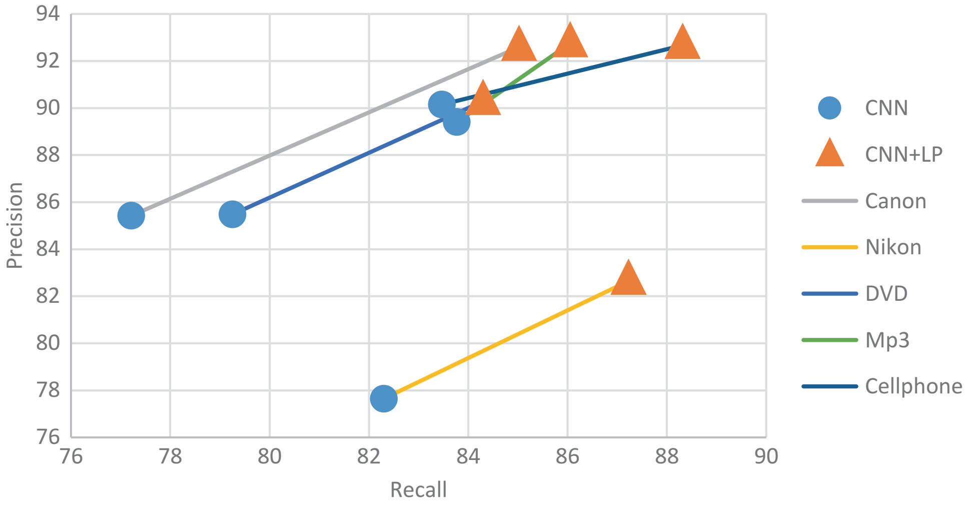
Sentiment Analysis, Basic Tasks of, Table 8 Impact of Sentic Patterns on the dataset by Qiu et al. (2011)

Domain	Classifiers	Precision (%)	Recall (%)	F-score (%)
Canon	CNN	85.42	77.21	81.10
Canon	CNN+LP	**92.59**	**85.02**	**88.64**
Nikon	CNN	77.65	82.30	79.90
Nikon	CNN+LP	**82.65**	**87.23**	**84.87**
DVD	CNN	85.48	79.25	82.24
DVD	CNN+LP	**90.29**	**84.30**	**87.19**
Mp3	CNN	89.40	83.77	86.49
Mp3	CNN+LP	**92.75**	**86.05**	**89.27**
Cellphone	CNN	90.15	83.47	86.68
Cellphone	CNN+LP	**92.67**	**88.32**	**90.44**

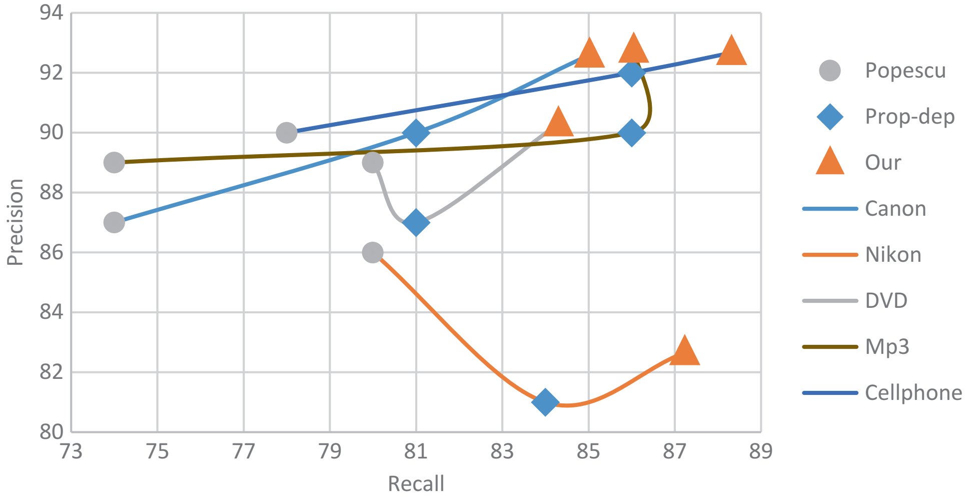
state-of-the-art approaches. First, a deep CNN, which is nonlinear in nature, better fits the data than linear models such as CRF. Second, the pre-trained word embedding features help our framework to outperform state-of-the-art methods that do not use word embeddings. The main advantage of our framework is that it does not need any feature engineering. This minimizes development cost and time.

Key Applications

Subjectivity detection can prevent the sentiment classifier from considering irrelevant or potentially misleading text. This is particularly useful in multi-perspective question answering summarization systems that need to summarize different opinions and perspectives and present multiple answers to the user based on opinions derived from different sources. It is also useful to analysts in government, commercial, and political domains who need to determine the response of the people to different crisis events. After filtering of subjective sentences, aspect mining can be used to provide clearer visibility into the emotions of people by connecting different polarities to the corresponding target attribute.



Sentiment Analysis, Basic Tasks of, Fig. 5 Comparison between the performance of CNN, CNN-LP, and LP on Bing Liu dataset



Sentiment Analysis, Basic Tasks of, Fig. 6 Comparison of the performance with the state of the art on Bing Liu dataset

Conclusion

In this chapter, we tackled the two basic tasks of sentiment analysis in social media: subjectivity detection and aspect extraction. We used an ensemble of deep learning and linguistics to collect opinionated information and, hence, perform fine-grained (aspect-based) sentiment analysis. In particular, we proposed a Bayesian deep convolutional belief network to classify a sequence of sentences as either subjective or objective and used a convolutional neural network for aspect extraction. Coupled with some linguistic rules, this ensemble approach gave a significant improvement in performance over state-of-the-art techniques and paved the way for a more multifaceted (i.e., covering more NLP subtasks) and multidisciplinary (i.e., integrating techniques from linguistics and other disciplines) approach to the complex problem of sentiment analysis.

Future Directions

In the future we will try to visualize the hierarchies of features learned via deep learning. We can also consider fusion with other modalities such as YouTube videos.

Cross-References

- ▶ Semantic Sentiment Analysis of Twitter Data
- ▶ Sentiment Quantification of User-Generated Content
- ▶ Twitter Microblog Sentiment Analysis

Acknowledgement This work was funded by Complexity Institute, Nanyang Technological University.

References

Akkaya C, Wiebe J, Mihalcea R (2009) Subjectivity word sense disambiguation. In: Proceedings of the 2009 conference on empirical methods in natural language processing, Singapore, pp 190–199

Andrzejewski D, Zhu X, Craven M (2009) Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In: Proceedings of the 26th annual international conference on machine learning, ACM, Montreal, pp 25–32

Blair-Goldensohn S, Hannan K, McDonald R, Neylon T, Reis GA, Reynar J (2008) Building a sentiment summarizer for local service reviews. In: Proceedings of WWW-2008 workshop on NLP in the information explosion era, Beijing, pp 14–23

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Bonzanini M, Martinez-Alvarez M, Roelleke T (2012) Opinion summarisation through sentence extraction: an investigation with movie reviews. In: Proceedings of the 35th international ACM SIGIR conference on

Research and development in information retrieval, SIGIR '12, Oregon, pp 1121–1122

Branavan S, Chen H, Eisenstein J, Barzilay R (2009) Learning document-level semantic properties from free-text annotations. J Artif Intell Res 34(2):569

Cambria E (2013) An introduction to concept-level sentiment analysis. In: Castro F, Gelbukh A, González M (eds) Advances in soft computing and its applications, Lecture notes in computer science, vol 8266. Springer, Berlin, pp 478–483

Cambria E (2016) Affective computing and sentiment analysis. IEEE Intell Syst 31(2):102–107

Cambria E, Hussain A (2015) Sentic computing: a common-sense-based framework for concept-level sentiment analysis. Springer, Cham

Cambria E, Howard N, Hsu J, Hussain A (2013a) Sentic blending: scalable multimodal fusion for continuous interpretation of semantics and sentics. In: IEEE SSCI, Singapore, pp 108–117

Cambria E, Mazzocco T, Hussain A (2013b) Application of multi-dimensional scaling and artificial neural networks for biologically inspired opinion mining. Biol Inspired Cogn Archit 4:41–53

Cambria E, Wang H, White B (2014) Guest editorial: big social data analysis. Knowl Based Syst 69:1–2

Cambria E, Poria S, Bajpai R, Schuller B (2016) SenticNet 4: a semantic resource for sentiment analysis based on conceptual primitives. In: COLING, Osaka, pp 2666–2677

Chaturvedi I, Ong Y-S, Tsang I, Welsch R, Cambria E (2016) Learning word dependencies in text by means of a deep recurrent belief network. Knowl Based Syst 108:144–154

Chen Z, Liu B (2014) Mining topics in documents: standing on the shoulders of big data. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, Newyork, pp 1116–1125

Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P (2011a) Natural language processing (almost) from scratch. J Mach Learn Res 12:2493–2537

Ding X, Liu B, Yu PS (2008) A holistic lexicon-based approach to opinion mining. In: Proceedings of first ACM international conference on web search and data mining (WSDM-2008). Stanford University, Stanford, pp 231–240

Fonseca ER, Rosa JLG (2013) A two-step convolutional neural network approach for semantic role labeling. In: The 2013 international joint conference on neural networks (IJCNN), IEEE, Dallas, 4–9 Aug 2013, pp 1–7

Friedman N, Murphy K, Russell S (1998) Learning the structure of dynamic probabilistic networks. In: Proceedings of the 14th annual conference on uncertainty in artificial intelligence, Madison, pp 139–14

Glorot X, Bordes A, Bengio Y (2011) Domain adaptation for large-scale sentiment classification: a deep learning approach. In: Proceedings of the twenty-eight international conference on machine learning, ICML, Washington

Hinton GE (2002) Training products of experts by minimizing contrastive divergence. Neural Comput 14 (8):1771–1800

Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of 22nd ACM SIGIR conference on Research and development in information retrieval, ACM, Berkeley, pp 50–57

Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of ACM SIGKDD conference on knowledge discovery & data mining, Seattle, pp 168–177

Hu Y, Boyd-Graber J, Satinoff B, Smith A (2014) Interactive topic modeling. Mach Learn 95(3):423–469

Jagarlamudi J, Daumé III H, Udupa R (2012) Incorporating lexical priors into topic models. In: Proceedings of the 13th EACL conference, Association for Computational Linguistics, France, pp 204–213

Jakob N, Gurevych I (2010) Extracting opinion targets in a single- and cross-domain setting with conditional random fields. In: Proceedings of EMNLP-2010, ACL, Sweden, pp 1035–1045

Kalchbrenner N, Blunsom P (2013) Recurrent convolutional neural networks for discourse compositionality. CoRR, abs/1306.3584

Kalchbrenner N, Grefenstette E, Blunsom P (2014) A convolutional neural network for modelling sentences. In: Proceedings of the 52nd annual meeting of the Association for Computational Linguistics, ACL 2014, Baltimore, 22–27 June 2014, vol 1: Long papers, pp 655–665

Kim Y (2014) Convolutional neural networks for sentence classification. CoRR, abs/1408.5882

Lin C, He Y, Everson R (2011) Sentence subjectivity detection with weakly-supervised learning. In: The 5th international joint conference on natural language processing (IJCNLP), Thailand

Lu Y, Zhai C (2008) Opinion integration through semi-supervised topic modeling. In: Proceedings of the 17th international conference on World Wide Web, ACM, Beijing, pp 121–130

Lu Y, Zhai C, Sundaresan N (2009) Rated aspect summarization of short comments. In: Proceedings of 18th World Wide Web conference, ACM, Madrid, pp 131–140

Ma Y, Cambria E, Gao S (2016) Label embedding for zero-shot fine-grained named entity typing. In: COLING, Osaka, pp 171–180

Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C (2011) Learning word vectors for sentiment analysis. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies – volume 1, Oregon, pp 142–150

McAuley J, Leskovec J (2013) Hidden factors and hidden topics: understanding rating dimensions with review text. In: Proceedings of RecSys'13, Hong Kong, 12–16 Oct 2013

Mcauliffe JD, Blei DM (2008) Supervised topic models. In: Advances in neural information processing systems, Vancouver, pp 121–128

Mikolov T, Yih W-T, Zweig G (2013) Linguistic regularities in continuous space word representations. In: HLT-NAACL, Atlanta, pp 746–751

Mukherjee A, Liu B (2012) Aspect extraction through semi-supervised modeling. In: Proceedings of 50th annual meeting of the ACL: long papers, vol 1, ACL, Korea, pp 339–348

Murray G, Carenini G (2011) Subjectivity detection in spoken and written conversations. Nat Lang Eng 17:397–418

Oneto L, Bisio F, Cambria E, Anguita D (2016) Statistical learning theory and ELM for big social data analysis. IEEE Comput Intell Mag 11(3):45–55

Ortega R, Fonseca A, Gutiérrez Y, Montoyo A (2013) Improving subjectivity detection using unsupervised subjectivity word sense disambiguation. Procesamiento del Lenguaje Natural 51:179–186

Pang B, Lee L (2004) A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd annual meeting of the Association for Computational Linguistics (ACL-04), Barcelona

Popescu A-M, Etzioni O (2005) Extracting product features and opinions from reviews. In: Proceedings of EMNLP-2005, Vancouver, pp 3–28

Poria S, Cambria E, Gelbukh A (2015a) Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: EMNLP, Lisbon, pp 2539–2544

Poria S, Cambria E, Gelbukh A, Bisio F, Hussain A (2015b) Sentiment data flow analysis by means of dynamic linguistic patterns. IEEE Comput Intell Mag 10(4):26–36

Poria S, Cambria E, Gelbukh A (2016a) Aspect extraction for opinion mining with a deep convolutional neural network. Knowl Based Syst 108

Poria S, Cambria E, Hazarika D, Vij P (2016b) A deeper look into sarcastic tweets using deep convolutional neural networks. In: COLING, pp 1601–1612

Poria S, Chaturvedi I, Cambria E, Bisio F (2016c) Sentic LDA: improving on LDA with semantic similarity for aspect-based sentiment analysis. In: IJCNN, pp 4465–4473

Poria S, Chaturvedi I, Cambria E, Hussain A (2016d) Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: ICDM, Barcelona, pp 439–448

Prinzie A, Van den Poel D (2009) Dynamic Bayesian Networks for acquisition pattern analysis: a financial-services cross-sell application. In: New frontiers in applied data mining, vol 5433, Lecture notes in computer science. Springer, Berlin/Heidelberg, pp 123–133

Qiu G, Liu B, Bu J, Chen C (2011) Opinion word expansion and target extraction through double propagation. Comput Linguist 37(1):9–27

Rajagopal D, Cambria E, Olsher D, Kwok K (2013) A graph-based approach to common-sense concept extraction and semantic similarity detection. In: WWW, Rio De Janeiro, pp 565–570

Rill S, Reinel D, Scheidt J, Zicari R (2014) Politwi: early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. Knowl Based Syst 69:14–23

Riloff E, Wiebe J (2003) Learning extraction patterns for subjective expressions. In: Proceedings of the 2003 conference on empirical methods in natural language processing, pp 105–112

Scaffidi C, Bierhoff K, Chang E, Felker M, Ng H, Jin C (2007) Red opal: Product-feature scoring from reviews. In: Proceedings of the 8th ACM conference on electronic commerce, ACM, pp 182–191

Socher R, Perelygin A, Wu JY, Chuang J, Manning CD, Ng AY, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank 1631. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), 2013, pp 1642

Suzuki J, Isozaki H (2006) Sequence and tree kernels with statistical feature mining. In: Advances in neural information processing systems 18, Vancouver, pp 1321–1328

Tang H, Tan S, Cheng X (2009) A survey on sentiment detection of reviews. Expert Syst Appl 36(7): 10760–10773

Tang D, Wei F, Yang N, Zhou M, Liu T, Qin B (2014) Learning sentiment-specific word embedding for Twitter sentiment classification. In: Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Baltimore, Maryland, pp 1555–1565

Taylor GW, Hinton GE, Roweis ST (2007) Modeling human motion using binary latent variables. In: Schölkopf B, Platt J, Hoffman T (eds) Advances in neural information processing systems 19. MIT Press, Cambridge, MA, pp 1345–1352

Titov I, McDonald R (2008) Modeling online reviews with multi-grain topic models. In: Proceedings of 17th conference on World Wide Web, ACM, Beijing, pp 111–120

Wang S, Manning C (2013) Fast dropout training. In: Proceedings of the 30th international conference on machine learning (ICML-13), Atlanta, pp 118–126

Wang H, Lu Y, Zhai C (2010) Latent aspect rating analysis on review text data: a rating regression approach. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, Washington, pp 783–792

Wang T, Cai Y, Leung H-f, Lau RY, Li Q, Min H (2014) Product aspect extraction supervised with online domain knowledge. Knowl Based Syst 71:86–100

Wiebe J, Riloff E (2005) Creating subjective and objective sentence classifiers from unannotated texts. In: Proceedings of the 6th international conference on computational linguistics and intelligent text processing, Mexico, pp 486–497

Zhao WX, Jiang J, Yan H, Li X (2010) Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In: Proceedings of the 2010 conference on empirical methods in natural language processing, Association for Computational Linguistics, Sweden, pp 56–65

Zhiqiang T, Wenting W (2014) DLIREC: aspect term extraction and term polarity classification system. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), Ireland, pp 235–240

Sentiment Classification

▶ Sentiment Analysis in Social Media
▶ Sentiment Analysis of Reviews

Sentiment Detection and Analysis

▶ Opinion Diffusion and Analysis on Social Networks

Sentiment Quantification of User-Generated Content

Fabrizio Sebastiani
Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy

Synonyms

Estimating prevalence of sentiment classes in user-generated content

Glossary

Prevalence of c in set $\mathcal{S}$ — Percentage of items in $\mathcal{S}$ that belong to class c and also known as the "relative frequency" of c or the "prior probability" (or simply "prior") of c

Quantification — Estimation of the prevalence of each class $c \in \mathcal{C}$ in a set $\mathcal{S}$ of unlabeled items (or estimation of the distribution of $\mathcal{S}$ across the classes in $\mathcal{C}$), synonym of "supervised prevalence estimation" and "class prior estimation," and also previously referred to as "counting."

Sentiment classification — A classification task whereby items (e.g., tweets, product reviews, comments, answers to open-ended questions) are classified based on the sentiment they convey (or opinion they express) about a certain entity or topic. It may take the form of binary classification (when the available classes are $\mathcal{C} =$ {Positive, Negative}) or ternary classification (when $\mathcal{C}$ also contains a Neutral class) or ordinal classification (when there are more than two classes, and these classes are ordered according to a total order, e.g., Excellent, Good, Fair, Poor, Disastrous)

Definition

User-generated content (UGC) is defined as content (usually in the form of text, spoken audio, imagery, video, etc.) authored by casual users (as opposed to professional users) of a digital content delivery platform. Examples of user-generated textual content are tweets, blog posts, product reviews, and as are all of the comments that other users upload as a reaction to them; examples of user-generated non-textual content are images uploaded on Instagram (or similar social networking services) or videos uploaded on platforms such as YouTube. *Sentiment quantification* of UGC is defined as the activity (carried out via supervised learning) of estimating the prevalence (aka percentage or relative frequency) of sentiment-related classes (e.g., Positive, Negative, Neutral) in a set of unlabeled UGC items.

Introduction

User-generated content (UGC) has turned into a goldmine for market researchers, social scientists, political scientists, and professionals involved in reputation management, since it gives near-instant access to a potentially enormous quantity of data from which the collective sentiment about products, companies, policies, and political candidates can be gauged.

Possibly the most important task underlying attempts to tap into this goldmine is *sentiment classification*, the task of classifying an item of UGC (e.g., a tweet, a product review, a post on a social networking service) according to the sentiment it conveys (or opinion it expresses) about a certain entity or topic. From a niche, esoteric topic that only a handful of NLP researchers were investigating, in the last 10 years, sentiment analysis (and sentiment classification, which is its most prominent incarnation) has blossomed into a research field with thousands of active researchers and into a multimillion industry too (almost all providers of textual content analysis tools nowadays boast a sentiment analysis solution as part of their commercial offer).

However, it turns out that in many applications, the final goal of sentiment classification is not that of determining the class of individual UGC items but that of estimating the prevalence of UGC items that belong to a certain class; the latter is a more specific task than the former, since a solution for the former is also a solution for the latter, but not vice versa. When prevalence estimation is tackled via supervised learning, it is known as *quantification*. Quantification has recently been investigated as a task of its own (i.e., as something which is not a mere by-product of classification), following experimental evidence that using quantification-specific algorithms, rather than standard classification-oriented ones, delivers superior quantification accuracy.

We here give an introduction to the task of quantifying UGC by sentiment, to the methods that have been proposed in the literature, and to the measures that have been used for evaluating the accuracy of different quantification algorithms.

Key Points

The obvious method for dealing with quantification is "classify and count," i.e., classifying each unlabeled object via a standard classifier and estimating class prevalence by counting the objects that have been labeled with the class. However, this strategy is suboptimal, since a good classifier is not necessarily a good "quantifier" (i.e., prevalence estimator). To see this, consider that a binary classifier h_1 for which $FP = 20$ and $FN = 20$ (FP and FN standing for the "false positives" and "false negatives," respectively, which it has generated on a given dataset) is worse, in terms of classification accuracy, than a classifier h_2 for which, on the same dataset, $FP = 18$ and $FN = 20$. However, h_1 is intuitively a better binary quantifier than h_2; indeed, h_1 is a perfect quantifier, since FP and FN are equal and thus, when it comes to class frequency estimation, compensate each other, so that the distribution of the test items across the class and its complement is estimated perfectly. In other words, a good quantifier needs to have small *bias* (i.e., needs to distribute its errors as evenly as possible across FP and FN).

Recent research (e.g., Barranquero et al. 2015; Bella et al. 2010; Esuli and Sebastiani 2015; Forman 2008) has convincingly shown that, since classification and quantification pursue different goals, quantification should be tackled as a task of its own, using different evaluation measures and, as a result, different learning algorithms. One reason why it seems sensible to pursue quantification directly, instead of tackling it via classification, is that classification is a more general task than quantification (since quantification can be framed in terms of classification, while the opposite is not true). A training set might thus contain information sufficient to generate a good quantifier but not a good classifier, which means that performing quantification via "classify and count" might be a suboptimal way of performing quantification. In other words, performing quantification via "classify and count" looks like a violation of "Vapnik's principle" (Vapnik 1998), which asserts that:

> If you possess a restricted amount of information for solving some problem, try to solve the problem directly and never solve a more general problem as an intermediate step. It is possible that the available information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem.

In the rest of this work, we will thus look at approaches that have tackled quantification as a task of its own rather than as a by-product of classification.

Historical Background

Sentiment quantification for UGC is a task at the crossroads of two main streams or research, namely, (a) sentiment analysis of UGC and (b) quantification.

Sentiment analysis (see Feldman 2013; Liu 2012; Pang and Lee 2008) is a fairly recent task, since until 15 years ago, it had only resulted in sporadic efforts, mainly originating from the NLP area. It was actually the rising importance of UGC, caused by the birth of new modes of expression (e.g., blogs, user-contributed product reviews) and new platforms for hosting them (e.g., electronic commerce portals, social networking services) that contributed to the growing importance of sentiment analysis. Indeed, while the Web of the 1990s was primarily a repository of factual content generated by professional authors, the Web of the new millennium has progressively become rife with opinion-laden content generated by casual users, and it is the sentiment-laden nature of this content that has prompted the explosion of sentiment analysis.

Earlier efforts at sentiment analysis were extremely primitive, some of them consisting of algorithms that counted the number p of occurrences of "positive words" (i.e., words that conveyed a sense of positivity, such as "truthful," "exceptional," and the like) and the number n of occurrences of "negative words" (i.e., words that conveyed a sense of negativity, such as "inaccurate," "pathetic," and the like) within a textual UGC item and decreed the item a Positive one if $p > n$ and a Negative one otherwise. This phase could be preceded by a check that $p + n$ exceeded a certain threshold, failing which the item was decreed a Neutral one. These primitive efforts met with some success, but clashed against the lack of high-coverage, manually crafted *sentiment lexicons*, i.e., dictionaries where each word of a given language is classified as a positive or a negative or a neutral word. This encouraged many researchers to devote their attention to devising methods for the automatic (or semiautomatic) extraction of such lexicons (also for languages other than English) from textual data.

The last 10 years have seen sentiment analysis researchers adopt increasingly sophisticated tools for linguistic analysis and text mining, thus leaving behind the "counting positive and negative words" phase; some of these tools will be discussed in section "Representing Sentiment for User-Generated Content."

Quantification has instead a more complex history, and it is fair to say that different strands in the quantification literature evolved almost independently (and somehow unknown to each other) within (a) statistics (e.g., Hopkins and King 2010), (b) machine learning (e.g., du Plessis and Sugiyama 2012; Saerens et al. 2002), and (c) data mining (e.g., Forman 2008). One interesting fact is that in strand (b), prevalence estimation is not a goal in itself but is functional to improving the accuracy of a classifier in situations characterized by *distribution drift* (i.e., by class prevalences that are likely to change substantially from the training set to the test set); instead, in strand (c) prevalence estimation is a goal in itself (the case of UGC that we are analyzing here belongs to this latter case).

Different quantification algorithms have been proposed over the years (on this, see section "Learning to Quantify" for more details). Most of them are batch learning methods, which require all the training examples to be loaded in memory at the same time, while incremental "online" methods, which relax this requirement and thus start learning after the first training examples are loaded in memory, are a rarity (see Kar et al. 2016 for an example).

The first work where sentiment analysis meets quantification is Esuli and Sebastiani (2010b), which observes that many applications of sentiment classification really have quantification, and not classification, as their goal, and thus argues for the importance of developing learning algorithms that explicitly target quantification and not classification. See section "Key Applications" for other pointers to the literature on sentiment quantification for UGC.

Main Approaches to Sentiment Quantification

Setting up a system that performs sentiment quantification of UGC essentially involves setting up two software modules:

1. A module that generates vectorial representations of UGC items (e.g., blog posts, tweets) which can be fed to an algorithm which learns a quantifier from training data
2. A module that learns a quantifier from the vectorial representations of labeled UGC items.

After discussing in section "Evaluating Quantification" the main measures that are used in the literature for evaluating quantification, in sections "Representing Sentiment for User-Generated Content" and "Learning to Quantify," we discuss the main techniques for building modules 1 and 2, respectively.

Evaluating Quantification

Different measures have been proposed in the literature for evaluating quantification error. We here concentrate on the measures that have been proposed for tackling *single-label multi-class* (SLMC) quantification. Note that a measure for SLMC quantification is also a measure for binary quantification, since the latter task is a special case of the former.

Notation-wise, by $\Lambda(p, \hat{p}, \mathcal{S}, \mathcal{C})$ we will indicate a *quantification loss*, i.e., a measure Λ of the error made in estimating a distribution p defined on set $\mathcal{S}$ and classes $\mathcal{C}$ by another distribution $\hat{p}$; we will often simply write $\Lambda(p, \hat{p})$ when $\mathcal{S}$ and $\mathcal{C}$ are clear from the context. (Consistently with most mathematical literature, we use the caret symbol (ˆ) to indicate estimation.)

The simplest measure for SLMC quantification is *absolute error* (AE), which corresponds to the average (across the classes $c \in \mathcal{C}$) absolute difference between the predicted class prevalence $\hat{p}(c)$ and the true class prevalence $p(c)$; i.e.,

$$\mathrm{AE}(p,\hat{p}) = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} |\hat{p}(c) - p(c)| \tag{1}$$

It is easy to show that AE ranges between 0 (best) and

$$\frac{2\left(1 - \min_{c\in\mathcal{C}} p(c)\right)}{|\,\mathcal{C}\,|}$$

(worst). The main advantage of AE is that it is intuitive and easy to understand to non-initiates too.

However, AE does not address the fact that the same absolute difference between predicted class prevalence and true class prevalence should count as a more serious mistake when the true class prevalence is small. For instance, predicting $\hat{p}(c) = 0.10$ when $p(c) = 0.01$ and predicting $\hat{p}(c) = 0.50$ when $p(c) = 0.41$ are equivalent errors according to AE, but the former is intuitively a more serious error than the latter. *Relative absolute error* (RAE) addresses this problem by relativizing the value $|\hat{p}(c) - p(c)|$ in Eq. 1 to the true class prevalence, i.e.,

$$\mathrm{RAE}(p,\hat{p}) = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \frac{|\hat{p}(c) - p(c)|}{p(c)} \tag{2}$$

RAE may be undefined in some cases, due to the presence of zero denominators. To solve this problem, in computing RAE, we can smooth both $p(c)$ and $\hat{p}(c)$ via additive smoothing, i.e.,

$$p_s(c) = \frac{\varepsilon + p(c)}{\varepsilon|\mathcal{C}| + \sum_{c\in\mathcal{C}} p(c)} \tag{3}$$

where $p_s(c)$ denotes the smoothed version of $p(c)$, the denominator is just a normalizing factor, and where the quantity $\varepsilon = \frac{1}{2|\mathcal{S}|}$ is typically used as a smoothing factor (Esuli and Sebastiani 2015; Forman 2008; Gao and Sebastiani 2015); $\hat{p}_s(c)$, the smoothed version of the predicted class prevalence, is defined analogously. The smoothed versions of $p(c)$ and $\hat{p}(c)$ are then used in place of their original versions in Eq. 2; as a result, RAE is always defined and still returns a value of 0 when p and $\hat{p}$ coincide. It is easy to show that RAE ranges between 0 (best) and

$$\frac{|\mathcal{C}| - 1 + \dfrac{1 - \min\limits_{c \in \mathcal{C}} p_s(c)}{\min\limits_{c \in \mathcal{C}} p_s(c)}}{|\mathcal{C}|}$$

(worst).

A third measure, and the one that has become somehow standard in the evaluation of SLMC quantification, is *normalized cross-entropy*, better known as *Kullback-Leibler Divergence* (KLD – see, e.g., Cover and Thomas 1991). KLD was proposed as a SLMC quantification measure in Forman (2005) and is defined as

$$\mathrm{KLD}(p, \hat{p}) = \sum_{c \in \mathcal{C}} p(c) \log_e \frac{p(c)}{\hat{p}(c)} \qquad (4)$$

KLD was originally devised as a measure of the inefficiency incurred when estimating a true distribution p over a set $\mathcal{C}$ of classes by means of a predicted distribution $\hat{p}$. KLD is thus suitable for evaluating quantification, since quantifying exactly means predicting how the items in set $\mathcal{S}$ are distributed across the classes in $\mathcal{C}$. KLD ranges between 0 (best) and $+\infty$ (worst). Note that, unlike AE and RAE, the upper bound of KLD is not finite since Eq. 4 has predicted prevalences, and not true prevalences, at the denominator: that is, by making a predicted prevalence $\hat{p}(c)$ infinitely small, we can make KLD be infinitely large.

Also KLD may be undefined in some cases. While the case in which $p(c) = 0$ is not problematic (since continuity arguments indicate that $0 \log \frac{0}{a}$ should be taken to be 0 for any $a \geq 0$), the case in which $\hat{p}(c) = 0$ and $p(c) > 0$ is indeed problematic, since $a \log \frac{a}{0}$ is undefined for $a > 0$. To solve this problem, also in computing KLD, we use the smoothed prevalences of Eq. 3; as a result, KLD is always defined and still returns a value of zero when p and $\hat{p}$ coincide.

While KLD is less easy to understand to non-initiates than AE or RAE, its advantage is that it is a very well-known measure, having been the subject of intense study within information theory (Csiszar and Shields 2004) and, although from a more applicative angle, within the language modeling approach to information retrieval and to speech processing. As a consequence, it has emerged as the *de facto* standard in the SLMC quantification literature.

Concerning ordinal quantification, the only known measure is the *Earth Mover's Distance* (EMD), a measure well-known in the field of computer vision. It is defined for the general case in which a distance $d(c', c'')$ is defined for each $c', c'' \in \mathcal{C}$; when there is a total order on the classes in $\mathcal{C}$, if we assume that $d(c_i, c_{i+1}) = 1$ for all $i \in \{1, ..., (\mathcal{C} - 1)\}$, the Earth Mover's Distance is defined as

$$\mathrm{EMD}(p, \hat{p}) = \sum_{j=1}^{|\mathcal{C}|-1} \left| \sum_{i=1}^{j} \hat{p}(c_i) - \sum_{i=1}^{j} p(c_i) \right| \qquad (5)$$

and can be computed in $|\mathcal{C}|$ steps from the estimated and true class prevalences. EMD ranges between 0 (best) and $(|\mathcal{C}| - 1)$ (worst).

Representing Sentiment for User-Generated Content

For quantification, sentiment analysis comes into play when generating vectorial representations for the labeled examples that are used to generate a sentiment quantifier and for the unlabeled examples that are the object of quantification. For generating these representations, no technique specific to sentiment *quantification* has emerged, which means that the same techniques used for sentiment *classification* are also used for quantification.

The traditional "bag of words" (BoW) approach to representing textual content in classification by topic cannot be used for classifying by sentiment: to see why, simply consider the fact that two sentences such as "A horrible hotel in a beautiful town" and "A beautiful hotel in a horrible town" would be assigned the same class if relying on a BoW representation. As a result, classification by sentiment relies on more sophisticated linguistic tools than classification by topic; these tools include:

- Part-of-speech taggers (e.g., to detect the difference between "good" as an adjective and "good" as a noun)
- Valence shifter detectors (e.g., to detect the presence of negated contexts, since negation usually inverts the polarity of the sentiment expressed in the scope of the negation)
- Detectors of amplifiers (e.g., "very") and diminishers (e.g., "scarcely"), since they change the intensity of any sentiment expressed within their scope
- Parsers and named entity recognizers (e.g., to detect which entity or entity sentiments are about)

One linguistic tool of particular importance in sentiment analysis is a sentiment lexicon, i.e., a dictionary (or thesaurus) where lexical entries are tagged according to whether they carry positive sentiment (e.g., the adjective "phenomenal"), negative sentiment (e.g., the adjective "disappointing"), or no sentiment at all (e.g., the adjective "electronic"). The availability of a sentiment lexicon is crucially important, e.g., to distinguish sentences such as "The room had a comfortable bed" (which indeed carries positive sentiment toward the hotel being reviewed) from sentences such as "The room had a rectangular bed" (which carries no sentiment).

Sentiment analysis, like many other NLP tasks, has not been indifferent to the "deep learning revolution" that has swept the field of machine learning in the last 5 years. In sentiment analysis, the most visible effect of this revolution has been the adoption of "word embeddings," which allow the vectorial representations of UGC items to leverage the distributional semantics of the words occurring in them (Tang et al. 2014).

An important fact to be noted is that UGC, because of its informal nature, is often fraught with syntactic inaccuracies, abbreviations, typos, slang expressions, etc.; these features make the linguistic analysis (and, as a consequence, the analysis by sentiment) of UGC more difficult than the analysis of more formal, polished language. It is thus often beneficial to use linguistic analysis tools (including sentiment lexicons) that are UGC specific and often even specific to the particular medium to be analyzed (e.g., Twitter).

Learning to Quantify

In the last 10 years several supervised learning approaches to prevalence estimation have been proposed, the two main classes being the *aggregative* and the *non-aggregative* methods. While the former requires the classification of each individual item as an intermediate step, the latter does not and estimates class prevalences holistically. Most methods fall in the former class, while the latter has few representatives (e.g., Gonzalez-Castro et al. 2013; King and Lu 2008).

We here introduce in some detail a few representative approaches of the aggregative type. Let us fix some notation. We assume a domain $\mathcal{X}$ of UGC items; a generic item will be indicated by $\mathbf{x}$. We assume the availability of a set *Tr* of training items and of a set *Te* of test items in which the accuracy of a quantifier is evaluated. A classifier (or *hypothesis*) trained on *Tr* will be denoted by $h\mathcal{X} \rightarrow \mathcal{C}$, where $\mathcal{X}$ is our domain of interest (e.g., the set of tweets) and $\mathcal{C}$ is the set of classes (e.g., Positive, Negative, Neutral). By $\widehat{p}_{\mathcal{S}}^{M}(c)$ we denote the true prevalence of class $c \in \mathcal{C}$ in set $\mathcal{S}$, while by $\widehat{p}_{\mathcal{S}}^{M}(c)$ we denote the prevalence of class $c \in \mathcal{C}$ in set $\mathcal{S}$ as estimated via method *M*.

Classify and Count (CC). An obvious method for quantification consists of training a classifier from *Tr* via a standard learning algorithm, classifying the items in *Te*, and estimating p_{Te} by simply counting the fraction of items in *Te* that are predicted to belong to the class. If by $\widehat{c}$ we denote the event "class *c* has been assigned by the classifier," so that $p_{Te}(\widehat{c})$ represents the fraction of test documents that have been assigned to c by the classifier, this corresponds to computing

$$\hat{p}^{CC}_{Te}(c) = p_{Te}(\hat{c}) = \frac{|\{\mathbf{x} \in Te | h(\mathbf{x}) = c\}|}{|Te|} \quad (6)$$

Forman (2008) calls this the *classify and count* (CC) method. This is the classification-oriented method that often uses as a baseline in quantification experiments.

Probabilistic Classify and Count (PCC). A variant of CC consists in generating a classifier from *Tr*, classifying the items in *Te*, and computing $p_{Te}(c)$ as the *expected* fraction of items predicted to belong to c. If by $p(c|\mathbf{x})$ we indicate the posterior probability, i.e., the probability of membership in c of test item $\mathbf{x}$ as estimated by the classifier, and by $E[x]$ we indicate the expected value of x; this corresponds to computing

$$\hat{p}^{PCC}_{Te}(c) = E[p_{Te}(\hat{c})] = \frac{1}{|Te|} \sum_{\mathbf{x} \in Te} p(c|\mathbf{x}) \quad (7)$$

The rationale of PCC is that posterior probabilities contain richer information than binary decisions, which are usually obtained from posterior probabilities by thresholding.

The PCC method is dismissed as unsuitable in Forman (2005, 2008), on the grounds that, when the training set distribution p_{Tr} and the test set distribution p_{Te} are different (as they should be assumed to be in any application of quantification), probabilities calibrated on *Tr* (*Tr* being the only available set where calibration may be carried out) cannot be, by definition, calibrated for *Te* at the same time. Experimental evidence on PCC is not conclusive, since PCC performed better than CC in the experiments of Bella et al. (2010) (where it is called "Probability Average") and Tang et al. (2010) but underperformed CC in the (much more extensive) experiments of Esuli and Sebastiani (2015). Interestingly enough, in the experiments of Gao and Sebastiani (2016), PCC proves the best performer, outperforming several among the methods that we discuss in this section.

Adjusted Classify and Count (ACC). Forman (2005, 2008) uses a further method which he calls "Adjusted Count" and which we will call (consistently with Esuli and Sebastiani 2015) *Adjusted Classify and Count* (ACC) so as to make its relation with CC more explicit.

ACC is based on the observation that, thanks to the law of total probability, it holds that

$$p_{Te}(\hat{c}_j) = \sum_{c_i, c_j \in \mathcal{C}} p_{Te}(\hat{c}_j | c_i) \cdot p_{Te}(c_i) \quad (8)$$

where $p_{Te}(\hat{c}_j | c_i)$ represents the fraction of test documents belonging to c_i that have been instead assigned to c_j by the classifier. Note that, once the classifier has been trained and applied to *Te*, the quantity $p_{Te}(\hat{c}_j)$ can be observed, and the quantity $p_{Te}(\hat{c}_j | c_i)$ can be estimated from *Tr* via *k*-fold cross-validation; the quantity $p_{Te}(c_i)$ is instead unknown and is indeed the quantity we want to estimate. Since there are $|\mathcal{C}|$ equations of the type described in Eq. 8 (one for each possible $\hat{c}_j$), and since there are $|\mathcal{C}|$ quantities of type $p_{Te}(c_i)$ to estimate (one for each choice of c_i), we are in the presence of a system of $|\mathcal{C}|$ linear equations in $|\mathcal{C}|$ unknowns. This system can be solved via standard techniques, thus yielding the required $\hat{p}_{Te}(c_i)$ estimates.

One problem with ACC is that it is not guaranteed to return a value in [0,1], due to the fact that the estimates of $p_{Te}(\hat{c}_j | c_i)$ may be imperfect. This has led most authors (see, e.g., Forman 2008) to (i) "clip" the $\hat{p}_{Te}(c_i)$ estimates (i.e., equate to 1 every value higher than 1 and to 0 every value lower than 0) and (ii) rescale them so that they sum up to 1.

Probabilistic Adjusted Classify and Count (PACC). The PACC method (proposed in Bella et al. (2010), where it is called "Scaled Probability Average") is a probabilistic variant of ACC, i.e., it stands to ACC like PCC stands to CC. Its underlying idea is to replace, in Eq. 8, $p_{Te}(\hat{c}_j)$ and $p_{Te}(\hat{c}_j | c_i)$ with their expected values. Equation 8 is thus transformed into

$$E[p_{Te}(\hat{c}_j)] = \sum_{c_i, c_j \in \mathcal{C}} E[p_{Te}(\hat{c}_j | c_i)] \cdot p_{Te}(c_i) \quad (9)$$

where

$$\begin{aligned} E\left[p_{Te}\left(\widehat{c}_j\right)\right] &= \frac{1}{|Te|}\sum_{\mathbf{x}\in Te} p\left(c_j|\mathbf{x}\right) \\ E\left[p_{Te}\left(\widehat{c}_j|c_i\right)\right] &= \frac{1}{|Te_i|}\sum_{\mathbf{x}\in Te_i} p\left(c_j|\mathbf{x}\right) \end{aligned} \tag{10}$$

and Te_i indicates the set of items in Te whose true class is c_i. Like for ACC, once the classifier has been trained and applied to Te, the quantity $E\left[p_{Te}\left(\widehat{c}_j\right)\right]$ can be observed, and the quantity $E\left[p_{Te}\left(\widehat{c}_j|c_i\right)\right]$ can be estimated from Tr via k-fold cross-validation, which means that we are again in the presence of a system of $|\mathcal{C}|$ linear equations in $|\mathcal{C}|$ unknowns that we can solve by standard techniques. Like ACC, also PACC can return values of $\widehat{p}_{Te}(c_i)$ that fall off the [0,1] range; again, clipping and rescaling is the only solution in these cases.

Like PCC, also PACC is dismissed as unsuitable in Forman (2005, 2008), for the same reasons for which PCC was also dismissed. Some experimental evidence seems instead in favor of PACC, since the experimental results published in Bella et al. (2010), Esuli and Sebastiani (2015), and Tang et al. (2010) indicate PACC to outperform all of CC, PCC, and ACC.

Expectation Maximization for Quantification (EMQ). EMQ, proposed by Saerens et al. (2002), is an instance of Expectation Maximization (Dempster et al. 1977), a well-known iterative algorithm for finding maximum-likelihood estimates of parameters (in our case, the class prevalences) for models that depend on unobserved variables (in our case, the class labels). Essentially, EMQ (see Algorithm 1) incrementally updates (Line 12) the posterior probabilities by using the class prevalences computed in the last step of the iteration and updates (Line 14) the class prevalences by using the posterior probabilities computed in the last step of the iteration, in a mutually recursive fashion.

All of the above methods require an underlying classifier that, given an item, predicts whether it belongs to class c or not (CC, ACC) or outputs a posterior probability of membership in c (PCC, PACC); any learning method for generating this classifier can be used. If the classifier only returns confidence scores that are not probabilities (as is the case, e.g., when the scores do no range on [0,1]), for PCC and PACC, these scores must be converted into true probabilities. If the score is a monotonically increasing function of the classifier's confidence in the fact that the item belongs to the class, this may be obtained by applying a logistic function. *Well-calibrated* probabilities (defined as the probabilities such that the prevalence $p_{\mathcal{S}}(c)$ of a class c in a set $\mathcal{S}$ is equal to $\sum_{\mathbf{x}\in\mathcal{S}} p(c|\mathbf{x})$) may be obtained by using a *generalized* logistic function.

Within the class of aggregative methods, a further distinction can be made between methods that use general-purpose learning algorithms, sometimes tweaking them or post-processing their prevalence estimates to account for their estimated bias, and methods that instead make use of learning algorithms explicitly devised for quantification. All of the methods described so far belong to the former class; let us now look at two methods of the latter type.

SVMs Optimized for KLD **(SVM(KLD)).** SVM(KLD), proposed in Esuli and Sebastiani (2010b, 2015), is an instantiation of SVM-perf (Joachims 2005) that uses KLD as the loss to optimize. (In Joachims (2005), SVM-perf is actually called SVM-multi, but the author has released its implementation under the name SVM-perf; we will thus use this latter name.) SVM-perf is a "structured output prediction" algorithm in the support vector machines (SVMs) family. Unlike traditional SVMs, SVM-perf is capable of optimizing any nonlinear, multivariate loss function that can be computed from a contingency table (as all the measures presented in section "Evaluating Quantification" are). Instead of handling hypotheses $h: \mathcal{X} \rightarrow \mathcal{Y}$ that map an individual item (e.g., a tweet) $\mathbf{x}_i$ into an individual label $y_i \in \mathcal{Y}$, SVM-perf considers hypotheses $\overline{h}: \overline{\mathcal{X}} \rightarrow \overline{\mathcal{Y}}$ that map entire tuples of items (e.g., entire sets of tweets) $\overline{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ into tuples of labels $\overline{\mathbf{y}} = (y_1, \ldots, y_n)$. Instead of learning the traditional hypotheses of type

$$h(\mathbf{x}) = \operatorname{sign}(\mathbf{w}\cdot\mathbf{x} + b) \tag{11}$$

SVM-perf thus learns hypotheses of type

```
Input  : Class prevalences $p_{Tr}(c)$ on $Tr$, for all $c \in \mathcal{C}$;
         Posterior probabilities $p(c|\mathbf{x})$, for all $c \in \mathcal{C}$ and for all $\mathbf{x} \in Te$;
Output: Estimates $\hat{p}_{Te}(c)$ of class prevalences on $Te$;
/* Initialization                                                      */
$s \leftarrow 0$;
for $c \in \mathcal{C}$ do
    $\hat{p}^{(s)}_{Te}(c) \leftarrow p_{Tr}(c)$;
    for $\mathbf{x} \in Te$ do
        $p^{(s)}(c|\mathbf{x}) \leftarrow p(c|\mathbf{x})$;
    end
end
/* Main Iteration Cycle                                                */
while stopping condition = false do
    $s \leftarrow s + 1$;
    for $c \in \mathcal{C}$ do
        for $\mathbf{x} \in Te$ do
            $p^{(s)}(c|\mathbf{x}) \leftarrow \dfrac{\dfrac{\hat{p}^{(s)}_{Te}(c)}{\hat{p}^{(0)}_{Te}(c)} \cdot p^{(0)}(c|\mathbf{x})}{\sum_{c \in \mathcal{C}} \dfrac{\hat{p}^{(s)}_{Te}(c)}{\hat{p}^{(0)}_{Te}(c)} \cdot p^{(0)}(c|\mathbf{x})}$
        end
        $\hat{p}^{(s)}_{Te}(c) \leftarrow \frac{1}{|Te|} \sum_{\mathbf{x} \in Te} p^{(s-1)}(c|\mathbf{x})$
    end
end
/* Generate output                                                     */
for $c \in \mathcal{C}$ do
    $\hat{p}_{Te}(c) \leftarrow \hat{p}^{(s)}_{Te}(c)$
end
```

Sentiment Quantification of User-Generated Content, Algorithm 1 The EMQ algorithm (Saerens et al. 2002)

$$\bar{h}(\bar{\mathbf{x}}) = \arg \max_{\bar{y}\in\bar{\mathcal{Y}}} (\mathbf{w} \cdot \Psi(\bar{\mathbf{x}},\bar{y})) \qquad (12)$$

where $\mathbf{w}$ is the vector of parameters to be learnt during training and

$$\Psi(\bar{\mathbf{x}},\bar{y}) = \sum_{i=1}^{n} \mathbf{x}_i y_i \qquad (13)$$

(the *joint feature map*) is a function that scores the pair of tuples $(\bar{\mathbf{x}},\bar{\mathbf{y}})$ according to how "compatible" $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are. In other words, while classifiers trained via traditional SVMs classify individual instances $\mathbf{x}$ one at a time, models trained via SVM-perf classify entire sets $\bar{\mathbf{x}}$ of instances in one shot and can thus make the labels assigned to the individual items mutually depend on each other. This is of fundamental importance in quantification, where, say, an additional false positive may even be *beneficial* when the rest of the data is expected to contain more false negatives than false positives.

While the optimization problem of classic soft-margin SVMs consists of finding

$$\begin{aligned} &\arg \min_{\mathbf{w},\,\xi_i \geq 0} \quad \frac{1}{2}\mathbf{w}\cdot\mathbf{w} + C\sum_{i=1}^{|Tr|}\xi_i \\ &\text{such that} \quad y'_i\left[\mathbf{w}\cdot\mathbf{x}'_i + b\right] \geq (1-\xi_i) \\ &\qquad\qquad\qquad \text{for all } i \in \{1,\ldots,|Tr|\} \end{aligned} \qquad (14)$$

(where the $\left(\mathbf{x}'_i, y'_i\right)$ denote the training examples), the corresponding problem of SVM-perf consists instead of finding

$$\begin{aligned} &\arg \min_{\mathbf{w},\, \xi_i \geq 0} \quad \frac{1}{2}\mathbf{w} \cdot \mathbf{w} + C\xi \\ &\text{such that} \quad \mathbf{w} \cdot \left[\Psi(\overline{\mathbf{x}}', \overline{y}') - \Psi(\overline{\mathbf{x}}', \overline{y}) + b\right] \\ &\qquad \geq \Lambda(\overline{y}', \overline{y}) - \xi \ \text{ for all } \ \overline{y} \in \overline{\mathcal{Y}}/\overline{y}' \end{aligned} \tag{15}$$

where $(\overline{\mathbf{x}}', \overline{y}')$ indicates a sequence of training examples and the corresponding sequence of their true labels. Here, the relevant fact to observe is that the multivariate loss Λ explicitly appears in the optimization problem.

Note that the set of possible labels $\mathcal{Y}$ is equal to $\{c_i, \overline{c}_i\}$, where c_i is any of the classes we are concerned with and $\overline{c}_i$ is its complement; that is, SVM-perf can only deal with binary decisions, which makes SVM(KLD) apt for binary quantification only. If we want to tackle SLMC quantification with $|\mathcal{C}| > 2$ classes, we need to independently generate $|\mathcal{C}|$ predicted prevalences $\widehat{p}(c)$ for each $c \in \mathcal{C}$ via $|\mathcal{C}|$ instances of SVM (KLD) and then rescale these predicted prevalences so that they sum up to 1.

SVMs Optimized for Q (SVM(Q)). SVM(Q), originally proposed in Barranquero et al. (2015), is, like SVM(KLD), an instantiation of SVM-perf. The authors optimize a "multi-objective" measure (which they call *Q-measure*) that combines classification accuracy and quantification accuracy; the rationale is that by maximizing both measures at the same time, one tends to obtain quantifiers that are not just effective (thanks to the high quantification accuracy) but also reliable (thanks to the high classification accuracy). The authors' Q-measure is

$$Q_\beta(p, \widehat{p}) = \frac{(\beta^2 + 1)\Gamma_c(p, \widehat{p}) \cdot \Gamma_q(p, \widehat{p})}{\beta^2 \Gamma_c(p, \widehat{p}) + \Gamma_q(p, \widehat{p})} \tag{16}$$

where Γ_c and Γ_q are a measure of classification "gain" (the opposite of loss) and a measure of quantification gain, respectively, and $0 \leq \beta \leq +\infty$ is a parameter that controls the relative importance of the two; for $\beta = 0$, the Q_β measure coincides with Γ_c, while when β tends to $+\infty$, Q_β asymptotically tends to Γ_q. As a measure of classification gain, the authors use recall, while as a measure of quantification gain, they use (1 − RAE), where RAE is as defined in Eq. 2. The authors motivate the (apparently strange) decision to use recall as a measure of classification gain with the fact that, while recall by itself is not a suitable measure of classification gain (since it is always possible to arbitrarily increase recall at the expense of precision or specificity), to include precision or specificity in Q_β is unnecessary, since the presence of Γ_q in Q_β has the effect of ruling out anyway those hypotheses characterized by high recall and low precision/specificity (since these hypotheses are indeed penalized by Γ_q).

Key Applications

One of the first applications of quantification to the field of user-generated content was described in Hopkins and King (2010), where the authors estimate the prevalence of support for different political candidates from blog posts, using the quantification algorithm pioneered in King and Lu (2008).

Another such application is described in Gao and Sebastiani (2015), which discusses sentiment quantification of tweets; the authors show experimentally that SVM(KLD) outperforms a "classify and count" approach implemented via linear SVMs on several tweet sentiment datasets. Sentiment quantification of tweets is also the topic of Da San Martino et al. (2016), whose authors tackle quantification at the ordinal level (using a totally ordered set of five degrees of sentiment strength) via a hierarchical quantification approach. Tweet quantification is also one of the subjects of the recent SemEval Task 4 "Sentiment Analysis in Twitter" shared task (Nakov et al. 2016), where tweets are labeled according to the sentiment they convey toward a certain topic; Subtask D consists of a binary quantification task, and Subtask E consists of an ordinal quantification task (with tweets labeled according to a five-point scale).

More in general, fields such as the social sciences, political science, reputation management, and market research are obvious application fields for sentiment quantification of UGC. This derives from the fact that these fields tend to be *inherently* interested in aggregate (rather than individual) views of people's attitudes. For instance, social scientists study the distribution of a given phenomenon across a population of interest (sometimes breaking up the population according to age or geographical location or religion or others) and are hardly interested in whether a single individual is affected by the phenomenon. Broadly speaking, we might say that researchers in these fields are usually less interested in finding the needle in the haystack than in characterizing the haystack itself.

A large number of works in the disciplines mentioned in the previous paragraph use quantification "without knowingly doing so"; that is, unaware of the existence of methods specifically optimized for quantification, they use classification with the only goal of estimating class prevalences. In other words, these works use plain "classify and count," but the application they are looking at is an obvious candidate for "real" quantification techniques. Among them, Mandel et al. (2012) use tweet quantification in order to estimate, from a quantitative point of view, the emotional responses of the population (broken down according to location and gender) to a natural disaster; Esuli and Sebastiani (2010a) quantify responses to open-ended surveys for market research applications; O'Connor et al. (2010) analyze the correlation between public opinion as measured via tweet sentiment quantification and via traditional opinion polls; and Dodds et al. (2011) use tweet sentiment quantification in order to infer spatiotemporal happiness patterns.

Future Directions

Research in sentiment quantification of user-generated content is still at an early stage; what can we expect for the near future?

First of all, one aspect that would be worth investigating is how to generate sentiment-laden vectorial representations of UGC that are specific to quantification; up to now, the same representations used for sentiment classification have been used, and this may be suboptimal.

Second, areas that will likely see new developments are those of single-label multi-class quantification and ordinal quantification. Up to now, most research on quantification has focused on the binary case, but going beyond binary is important for sentiment analysis, where classes other than Positive and Negative are often important.

More in general, as awareness of quantification as a task on its own grows, we may expect fewer and fewer applied works to use simple "classify and count" and more and more of them to apply methods that have proven more accurate in the quantification literature. This in turn may encourage the investigation of new learning methods specific to quantification.

Cross-References

- ▶ Microblog Sentiment Analysis
- ▶ Semantic Sentiment Analysis of Twitter Data
- ▶ Sentiment Analysis in Social Media
- ▶ Sentiment Analysis of Microblogging Data
- ▶ Sentiment Analysis of Reviews
- ▶ Sentiment Analysis, Basic Tasks of
- ▶ Social Media Analysis for Monitoring Political Sentiment
- ▶ Twitter Microblog Sentiment Analysis
- ▶ User Sentiment and Opinion Analysis

References

Barranquero J, Diez J, del Coz JJ (2015) Quantification-oriented learning based on reliable classifiers. Pattern Recogn 48(2):591–604

Bella A, Ferri C, Hernandez-Orallo J, Ramirez-Quintana MJ (2010) Quantification via probability estimators. In: Proceedings of the 11th IEEE international conference on data mining (ICDM 2010), Sydney, pp 737–742

Cover TM, Thomas JA (1991) Elements of information theory. Wiley, New York

Csiszar I, Shields PC (2004) Information theory and statistics: a tutorial. Found Trends Commun Inf Theory 1(4):417–528

Da San Martino G, Gao W, Sebastiani F (2016) Ordinal text quantification. In: Proceedings of the 39th ACM conference on research and development in information retrieval (SIGIR 2016), Pisa, pp 937–940. https://doi.org/10.1145/2911451.2914749

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39(1):1–38

Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM (2011) Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. PLoS One 6(12). https://doi.org/10.1371/journal.pone.0026752

du Plessis MC, Sugiyama M (2012) Semi-supervised learning of class balance under class-prior change by distribution matching. In: Proceedings of the 29th international conference on machine learning (ICML 2012), Edinburgh

Esuli A, Sebastiani F (2010a) Machines that learn how to code open-ended survey data. Int J Mark Res 52(6):775–800

Esuli A, Sebastiani F (2010b) Sentiment quantification. IEEE Intell Syst 25(4):72–75

Esuli A, Sebastiani F (2015) Optimizing text quantifiers for multivariate loss functions. ACM Trans Knowl Discov Data 9(4), Article 27

Feldman R (2013) Techniques and applications for sentiment analysis. Commun ACM 56(4):82–89

Forman, G (2005) Counting positives accurately despite inaccurate classification. In: Proceedings of the 16th European conference on machine learning (ECML), Porto, pp 564–575

Forman G (2008) Quantifying counts and costs via classification. Data Min Knowl Disc 17(2):164–206

Gao W, Sebastiani F (2015) Tweet sentiment: from classification to quantification. In: Proceedings of the 7th international conference on advances in social network analysis and mining (ASONAM 2015), Paris, pp 97–104

Gao W, Sebastiani F (2016) From classification to quantification in tweet sentiment analysis. Soc Netw Anal Min 6(19):1–22

Gonzalez-Castro V, Alaiz-Rodriguez R, Alegre E (2013) Class distribution estimation based on the Hellinger distance. Inf Sci 218:146–164

Hopkins DJ, King G (2010) A method of automated nonparametric content analysis for social science. Am J Polit Sci 54(1):229–247

Joachims T. (2005) A support vector method for multivariate performance measures. In: Proceedings of the 22nd international conference on machine learning (ICML 2005), Bonn, pp 377–384

Kar P, Li S, Narasimhan H, Chawla S, Sebastiani F (2016) Online optimization methods for the quantification problem. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2016), San Francisco, pp 1625–1634. https://doi.org/10.1145/2939672.2939832

King G, Lu Y (2008) Verbal autopsy methods with multiple causes of death. Stat Sci 23(1):78–91

Liu B (2012) Sentiment analysis and opinion mining. Morgan and Claypool Publishers, San Rafael

Mandel B, Culotta A, Boulahanis J, Stark D, Lewis B, Rodrigue J (2012) A demographic analysis of online sentiment during hurricane Irene. In: Proceedings of the NAACL/HLT workshop on language in social media, Montreal, pp 27–36

Nakov P, Ritter A, Rosenthal S, Sebastiani F, Stoyanov V (2016) SemEval-2016 Task 4: sentiment analysis in Twitter. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016), San Diego, pp 1–18

O'Connor B, Balasubramanyan R, Routledge BR, Smith NA (2010) From tweets to polls: linking text sentiment to public opinion time series. In: Proceedings of the 4th AAAI conference on weblogs and social media (ICWSM 2010), Washington, DC

Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1/2):1–135

Saerens M, Latinne P, Decaestecker C (2002) Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. Neural Comput 14(1):21–41

Tang L, Gao H, Liu H (2010) Network quantification despite biased labels. In: Proceedings of the 8th workshop on mining and learning with graphs (MLG 2010), Washington, DC, pp 147–154

Tang D, Wei F, Yang N, Zhou M, Liu T, Qin B (2014) Learning sentiment-specific word embedding for Twitter sentiment classification. In: Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (ACL 2014), Baltimore, pp 1555–1565

Vapnik V (1998) Statistical learning theory. Wiley, New York

Sentiment-Emotion-Intent Analysis

► Twitris: A System for Collective Social Intelligence

Sequence

► Paths in Complex Networks

Server-Side Scripting Languages

Ludger Martin
Faculty of Design – Computer Science – Media,
Hochschule RheinMain, Wiesbaden, Germany

Synonyms

ASP; ASP.NET; CGI; JSF; JSP; Perl; PHP; Ruby on Rails

Glossary

AJAX	Asynchronous JavaScript and XML
CGI	Common Gateway Interface
HTML	Hypertext Markup Language
JSF	Java Server Faces
JSON	Java Script Object Notation
JSP	Java Server Pages

Definition

Server-side scripting languages are programming languages developed especially for creating HTML pages (or Web pages) on the server side. These languages usually provide special libraries that facilitate creating HTML pages. In times of Web 2.0 and AJAX, these scripting languages can also serve as data sources (services) for AJAX.

There are two different types of scripting languages. The first variant can be embedded in HTML. The language can be embedded, for example, in places where a particular functionality is needed. The second variant is languages which can be used to create HTML tags. They provide an interface for creating HTML tags.

On the server side, a special interpreter is necessary for each scripting language. This interpreter is introduced to the Web server so that the server will be able to use it for the script execution, when required.

Introduction

There is a large number of server-side scripting languages. It is their task to dynamically build HTML pages (Web pages) on the server side. To achieve this, a Web server that is to distribute the HTML pages must be told where to find an interpreter for a particular script. Most of the server-side scripts are interpreted. A small number can also be compiled.

Without server-side scripting languages, you can only create static Web pages. Then it is not possible to customize anything for single users. A customization can be something as simple as the display of search results. Because of these languages, HTML pages can be created dynamically, i.e., on request.

If we look at today's Web sites, we will find that most of them were created using server-side scripting languages, among them Web sites which are run by a content management system. The content management system itself has been developed using a server-side scripting language. There will be very few exceptions which do not use such a language.

Web 2.0 pages that use JavaScript for controlling their content also need a server-side scripting language, e.g., AJAX was used to send requests for database access to a server. This can only be done using a server-side scripting language.

Key Points

There is a large number of server-side scripting languages. It is their task to dynamically build Web pages on the server side. A Web server needs an interpreter for a particular scripting language. A small number can also be compiled. Despite the development of Web 2.0 and the relocation of functionality to the client side, i.e., the browser, server-side scripting languages will still remain beneficial. Using AJAX, for example, a data source must be provided on the server side.

Historical Background

It is hard to say which server-side scripting language was first. It is a fact though that Perl was one of the first languages. The first version of Perl as a universal scripting language was presented in 1987 by Larry Wall. Only later, in the 1990s, did it become useful for Web pages because CGI was introduced.

In 1995, Rasmus Lerdorf developed PHP. At the beginning, PHP was based on Perl. In 1997, with version 2, the first parser for PHP was delivered. From then on PHP has been particularly suited for Web pages. PHP is a scripting language embedded in HTML. From the very start the evaluation of form variables has been important. By now, PHP has become one of the most widely used scripting languages for Web pages. But PHP has also become a universal scripting language which can be used anywhere.

Python is another universal scripting language, which was developed by Guido van Rossum in 1991. Today it is also commonly used for Web applications.

At the end of the 1990s, Sun Microsystems presented the language JavaServer Pages (JSP). JSP is based on the language Java, but it is embedded in HTML. Just as with Java, the JSP pages must be compiled before the byte-code that was created can be executed in a virtual machine. Nowadays, JSP is considered outdated. It was superseded by JavaServer Faces (JSF) in 2004. Particularly for Web pages, JSF, as opposed to JSP, is component oriented. JSF consequently focuses on the model–view–controller pattern.

Microsoft developed the Active Server Pages (ASP) particularly for the Internet Information Service (IIS). The technology, which was presented in 1998, can be programmed among others with VBScript or JScript. The relevant programming language is also embedded in HTML. In 2002 it was superseded by ASP.NET. That is the Web-based technology which is embedded in the . NET framework. Programming languages for ASP.NET are VBScript and C#.

Another popular language for Web pages is Ruby on Rails. The programming language Ruby was presented by Yukihiro Matsumoto in 1995. At the beginning it was only known in Japan. Ruby on Rails, which was developed in 2005, is a specific library for Web applications.

A more recent language is node.js. It is introduced in 2009 by Ryan Dahl. It is a server-side system with a special event-based implementation. The programming language for node.js is JavaScript.

Server-Side Scripting Languages

In the following, a sample server-side scripting language will be described. We chose PHP because it is one of the most widely used languages. After that, we will take a glance at Perl, which is unlike PHP a language that is not embedded in HTML.

PHP

From version 2 on, PHP has been developed for the dynamic creation and evaluation of Web pages. At first it was a procedural programming language. Version 4.0 (2000) introduced objects, which were revised in version 5.0 (2004). PHP provides a comprehensive procedural and object-oriented library.

PHP (Lerdorf et al. 2006) is a programming language that is embedded in HTML (Kessin 2011). It is always interpreted on the server. Output of a PHP script is usually an HTML page. But it is also possible to create different text formats and binary formats such as JSON, PDF, or PNG. Because the scripts are executed on the server, the user cannot see the source code. Users only get to see the output. This way, stealing the source code is not possible. If a Web server is very busy, the PHP scripts can be compiled beforehand, and then only the byte-code can be executed. Without parsing and compiling, execution performance is strongly improved.

Figure 1 shows a very small PHP file whose browser output is hello world. It shows clearly that the file begins with HTML source code. PHP is embedded in HTML; the actual PHP part starts only in line eight. This is marked by the

```
1  <DOCTYPE html>
2  <html>
3      <head>
4          <title>Hello World </title>
5          <meta charset="UTF-8"/>
6      <head>
7      <body>
8          <?php
9              echo "<p>Hello World!</p>";
10         ?>
11     </body>
12 </html>
```

Server-Side Scripting Languages, Fig. 1 Hello World PHP page

string <?php. The command echo makes the browser display <p>Hello World!</p>. HTML tags can also be output using echo. The PHP part ends in line ten with the string?>. You can include PHP parts anywhere and any number of times. It is not necessary to include HTML source code in a PHP file, which is often the case with classes. In this scenario, the PHP file starts directly with <?php. If the file ends with a PHP source text, you can leave out the closing? >. Formerly, <? and ?> were used, but they had caused problems with XHTML. A PHP filename must always finish with .php for the Web server to know that it is a PHP file.

You can also include comments in the PHP sections. Introduce single-line comments using // or #. Multiline comments should be enclosed in /* and */.

Variables and Operators

Variables are a central feature in a programming language. PHP is an untyped programming language. This means that usually you do not need to specify types. There are two exceptions, which will be explained later. As a result, you do not need to define variables; you can simply use them.

Variables always begin with $. Then, an arbitrary sequence of characters and numbers may follow; the first character after $ must be a letter. PHP distinguishes between capital letters and small letters. As it is not necessary to define the variables, there is a certain danger. If you access an undefined variable, in the best case you will receive a warning. In the worst case you will only notice that the program does not work as expected.

You can assign a type to a variable by giving it a value. Figure 2 shows several examples of this. In lines 1–4, numbers are assigned. The lines five and six treat booleans, where *true* is the same as 1 and *false* equals an empty string. Lines seven and eight demonstrate how character strings are assigned. It is important to realize that there is a great difference between the opening and closing "and."

Only if " is specified, variables and escape sequences (e.g., \t for tabulator) in character strings are resolved. That means that as the lines nine and ten show $x is replaced by the numeric value 17. if ′ is used, $x will remain a character string.

You can also create arrays. Line 11 shows an array with three elements. Line 12, however, shows an associative array. Using = > you can separate the keys from the values. Because PHP is an untyped language, the values and the keys of the types may vary within an array.

PHP changes the variable type according to the situation, if necessary. If, for example, two variables are added as numbers and one of them is a character string, then this string will automatically be changed to a number. Sometimes, not often, an explicit-type conversion may be necessary. This can be done placing the required type in round brackets in front of the variable: (int)$x.

You can check variables using the functions in Fig. 3. In the following, functions and methods will always be specified including the expected and returned types. Because PHP is an untyped language, the types will only be checked at runtime, and only afterward an error message will be displayed if the types do not match.

Server-Side Scripting Languages, Fig. 2 Variables

```
$x = 42;
$x = 0xFF;
$x = 4.2;
$x = 4.2e6;
$x = TRUE; // or 1
$x = FALSE; // or ''
$x = "abc";
$x = 'abc';
$x = 8;
$y = "it's $x o'clock"; // value: it's 8 o'clock
$a = array('one', 'two', 'three');
$a = array(1 => 'one', 2 => 'two', 3 => 'three');
```

```
1   int isset (mixed var);
2   int unset (mixed var);
3   $var   =  NULLL;
4   string gettype(mixed var);
```

Server-Side Scripting Languages, Fig. 3 Checking of variables

Specification of mixed means that different types may be used. isset() checks if a variable has been specified. With lines two and three you can specify a variable as undefined. gettype() determines the current type of a variable value. This value may change during the execution of a program. The type itself will be returned as a string.

There are very few specifics for operators. As we have already seen, = is an assignment. +, −, *, /, and % are mathematical operators, whereby division and multiplication come before addition and subtraction. To concatenate character strings, use .. For comparisons, ==,! =, <>, <, >, <=, and >.= are available. Variable types can be compared using === and! ==. For grouping operators you can use round brackets. You can use AND, &&, OR, ||, XOR and ! as logical operators, e.g., for conditions. AND and && are equivalent and so are OR and ||.

```
if ($i > 0){
      echo '$i is grater than 0';
} elseif ($i = 0) {
      echo '$i is equal to 0';
} else {
      echo '$i is less than 0';
}
```

Server-Side Scripting Languages, Fig. 4 If condition

Program Control

PHP offers the usual options for program control. Figure 4 shows an if condition. After the keyword if, you must specify a condition in round brackets. Unlike in other languages, in addition to the else

```
$a = array(1 => 'one', 2 => 'two', 3 => 'three');
foreach($a as $key => $value) {
    echo "$value has key $key";
}
```

Server-Side Scripting Languages, Fig. 5 Foreach loop

branch, there are one or more alternative conditions elseif. Besides the if condition there is additionally a switch statement.

Loops are also an option. A do while (. . .) {. . .} loop checks the condition after every loop run, while the (. . .){. . .} loop checks the condition prior to every loop run.

The for loop corresponds to the C syntax in that you can specify separately first an initialization statement, followed by a condition and then an incrementation using. Additionally, there is a foreach loop which has been developed specially for arrays. This loop executes the following statements once for each value in the array. Figure 5 gives an example. Inside the brackets the array is specified first and then follows the keyword. After that, one or two variables are specified, which will be used to store the value and, optionally, the key of the array. The variable for the key and the characters = > can be left out if only the values are of interest.

Classes, Objects, Error Handling

PHP too allows you to define custom functions and objects. With version 5, object orientation has been thoroughly revised.

Using the keyword class you can define classes. Classes' attributes and methods can be defined as public, protected, and private to ensure access protection. PHP supports only single-class inheritance, which can be specified using the keyword extends. Alternatively, there are interfaces, which a new class can implement. Classes can also be abstract. This is the case as soon as at least one method of a class has been marked as abstract. These methods are not yet implemented. The class which inherits must, similar to the interfaces, implement the abstract methods.

```
$aDateTime = new DateTime();
echo $aDateTime->format(’Y-m-d H:i:s’);
```

Server-Side Scripting Languages, Fig. 6 Instantiate a class

```
try {
     $date = new DateTime(’2012-08-01’);
} catch (Exception $e) {
      echo $e-> getMessage();
}
```

Server-Side Scripting Languages, Fig. 7 Error handling

Constructors must be named __construct (with two _). For downward compatibility with PHP 4, the constructor may have the same name as the class. A destructor must be named __destruct (with two _).

As Fig. 6 shows, you can create instances of classes using the keyword new. An object of the class DateTime is instantiated. You can access the methods or attributes of objects using the –> operator, as demonstrated by the method format().

PHP also allows creating static attributes and methods. Polymorphic methods are not supported, as PHP is an untyped language. It is only possible to specify default values for single parameters, so that the values can be omitted.

Error handling is also possible. Figure 7 shows how a try-catch is specified. After try you specify all the statements which might cause errors. One or more catch statements control the error handling. Only in the catch statement, it is necessary

```
<form action="form.php" method="post">
  <p>
     username <input type="text" name="username"/><br/>
     password <input type="password" name="password"/><br/>
     <input type="submit" name="submit" value="submit"/>
  </p>
</form>
```

Server-Side Scripting Languages, Fig. 8 Small HTML-form

to specify a type within the brackets. This type determines the class of the exception that is to be treated. This enables you to react appropriately to different exceptions. Using the keyword throw, you can throw a new exception.

Interaction with HTML Forms

The interaction with HTML forms is a central point with server-side scripting languages. Here, you need to be particularly careful because these places are popular goals for attacks on a Web page or an application. For this very reason, PHP as well has undergone a variety of improvements in the course of time.

In Fig. 8 you can see a small HTML form which consists of two input fields for the username and the password and a send button. In the form, the send method post has been specified. You have two options. get is the simplest method. Here, data is coded and committed with the URL. An advantage is that the data is visible in the URL. The disadvantages are that the URL and with it the data amount is limited in size and anyone can see and modify the committed data. Especially for passwords, this method is not recommended. Here, the post method can help. Data is sent to the URL separately, and there is no limit in the amount of data. If you want to transfer entire files with a form, the method post is obligatory. It is more difficult to modify this data but not impossible.

```
1  if ($_POST['submit'] == 'submit') {
2      if (($_POST['username'] == ...)&&
3          ($_POST['password'] == ...)) {
4          // secure action
5      }
6  }
```

Server-Side Scripting Languages, Fig. 9 Evaluate form

As destination for the form, the PHP file form.php is specified. The <input> tags each have an attribute name, which determines the name of the variable as it shall be available in the destination script. In the past, these names could be used in PHP directly. Today, for security reasons they are stored in the two arrays $_GET and $_POST. The name of each <input> tag is its array index. Line one in Fig. 9 shows how to check whether the button submit was pressed. Because the method post was used, in the array $_POST the index "submit" will be searched for, and a check will be done to see if the array contains the value submit. If so, a similar procedure can check the username and password.

You can use cookies if a Web application is to store data on a client (or browser) permanently. Cookies may contain any data, but they have a

```
1 session_start(void);
2 $_SESSION['text'] = 'Hello world!';
3 echo $_SESSION['text'];
```

Server-Side Scripting Languages, Fig. 10 Session

maximum length of 4 KB. But a Web application can send up to 20 cookies to the browser. If you want to define how long data shall be kept, you can specify an individual expiry date for a cookie. Cookies can only be placed in a Web page's header. This is possible only if no character of the actual Web page has been output. It is recommended to specify a cookie for a script as early as possible. As soon as a cookie was set, the browser will send it automatically to the server with every query. In PHP cookies can be read using the array $_COOKIE. The problem with cookies is that users can decide whether or not their browser shall accept cookies. A Web application can only find out about this decision if it checks whether cookies are transferred back.

Web pages are generally independent of a context. When developing Web applications, this can be annoying. You can solve this problem by using sessions. Sessions provide a separate storage area on the server for each user. A session is assigned to a user through a unique identifier. This ID must always be transferred between the browser and the server. To achieve this, three variants are available, which have been presented earlier: get, post, cookie. Cookie is the most popular variant because it causes the least work. It is, however, the most problematic as the developer does not know whether the browser accepts cookies. If cookies are used, again the cookie must be transferred in the header.

To use a session in PHP, each file which uses the session must call the command session_start() (see Fig. 10). This command checks the session ID if one was transferred. If the ID is correct, an existing session will continue to be used. If it is not correct or not available, a new session will be created. The array $_SESSION is available for storing data (lines two and three). The data remains stored on the server for some time, and the various PHP scripts can access it. This time limit is important because sometimes it may not be clear whether a user is still active. If a user is inactive over a longer period of time, the session will be deleted. Sessions can also be deleted by PHP.

```
1 header('Content-type: application/json');
2 $object = new ...;
3 $json = json_encode($object);
4 echo $json;
```

Server-Side Scripting Languages, Fig. 11 Modify header

Any kind of data can be stored in a session. You can also store objects. But there is one condition: the object's class must be known in each PHP file before the session can be started. With respect to security (Hope and Walther 2008), sessions must be particularly protected. The usual attacks are session hijacking and session riding.

Other Data Formats

Besides HTML, PHP (Loudon 2010) allows you to create any other data format. These formats could be XML, JSON, or images. JSON, for example, is often used if it is a Web 2.0 application and you want to send back queries to a server using AJAX. For the browser to be able to recognize what kind of data it is, the HTTP header that specifies the data format must be modified. Figure 11 shows how to set the type for JSON data. For the function header(), it is important that in the body no data has yet been sent. It is the same as for cookies.

JSON is a frequently used data format. Therefore, special functions are available for creating JSON automatically from PHP objects. In line two a sample object is instantiated. Using the function json_decode(), it can be changed to a JSON character string, which can then be output with echo. Analogously, the function json_decode() can change a JSON character string into a PHP object.

Embedded Files and Debugging

If you develop classes, usually each class is stored in a separate file. Then it is also necessary that these files can be embedded in PHP scripts. To do so, in PHP two functions are available. require() embeds the specified file and displays an error message if this was not successful. Processing is stopped. include(), however, will only output a warning and continue the script. For each one there is an alternative method require_once() and include_once(). These functions ensure that a script will be embedded only once, even if it was specified more often. This is particularly useful for, e.g., classes.

Troubleshooting is also very important for server-side applications. To enable debugging of an application, there are currently two modules for the Web server Apache, Xdebug and Zend Debugger, which allow remote debugging. Using these modules in a compatible development environment, you can execute the PHP source code step by step and examine the variables as well.

Perl

Perl (Guelich et al. 1999; Wall et al. 2000) is a universal programming language that can be used for developing server-side applications, too. The module CGI provides an interface that can be used to create HTML elements very easily. Figure 12 shows an example of the hello world program in Perl.

Line two provides the CGI module. Then, an object whose class is CGI is instantiated. Then some methods are used to create the individual HTML elements. The methods header and start_html create the entire header. For each HTML element in the body, there is one method for creating that particular element. You can see this in line six for the <p> —tag. Each method returns a string, which is output using print.

```
1  #!/usr/local/bin/perl -w
2  use CGI;
3  $q = CGI ->new;
4  print $q ->header,
5        $q ->start_html('hello world'),
6        $q ->p('hello world'),
7        $q ->end_html;
```

Server-Side Scripting Languages, Fig. 12 Hello World Perl page

Key Applications

A large area of application for server-side scripting languages is a content management system. An e-commerce platform is another common application. The development of Web 2.0 applications needs server-side scripting languages for data sources, those must be provided on the server side. Single-page applications are not able to access data bases from the client side.

Future Directions

Despite the development of Web 2.0 and the relocation of functionality to the client side, i.e., the browser, server-side scripting languages will still remain beneficial. Using AJAX, for example, a data source must be provided on the server side. Concerning the various libraries, a trend can be seen that the server-side scripting languages are more and more used for the automatic creation of JavaScript source code. So it will remain exciting to watch how the fast-moving world of the Web will develop.

Cross-References

▶ HTML

References

Guelich S, Gundavaram S, Birznieks G (1999) CGI programming with Perl. O'Reilly Media, Sebastopol

Hope P, Walther B (2008) Web security testing cookbook: systematic techniques to find problems fast. O'Reilly Media, Sebastopol

Kessin Z (2011) Programming HTML5 applications: building powerful cross-platform environments in javascript. O'Reilly Media, Sebastopol

Lerdorf R, Tatroe K, MacIntyre P (2006) Programming PHP, 2nd edn. O'Reilly Media, Sebastopol

Loudon K (2010) Developing large web applications. O'Reilly Media, Sebastopol

Wall L, Christiansen T, Orwant J (2000) Programming Perl: there's more than one way to do it. O'Reilly Media, Sebastopol

Service Delivery Networks

▶ Interorganizational Networks

Service Discovery

Matthias Klusch
German Research Center for Artificial,
Intelligence (DFKI), Saarbruecken, Germany

Synonyms

Semantic web services; Service-oriented architectures; Service search and selection; Web services

Glossary

CAN	Content-Addressable Network
DHT	Distributed Hash Table
JSON	JavaScript Object Notation
NAICS	North American Industrial Classification System
OWL-S	Ontology Web Language for Services
REST	Representational State Transfer
SA-REST	Semantic Annotation of Web Resources
SA-WSDL	Semantically Annotated WSDL
SIC	Standard Industrial Classification
SML	Service Modeling Language
SOA	Service-Oriented Architecture
SOAP	Simple Object Access Protocol
UDDI	Universal Description, Discovery, Integration
UNSPSC	United Nations Standard Products and Services Code
USDL	Unified Service Description Language
UUID	Unique Universal Identifiers
WADL	Web Application Description Language
WSDL	Web Service Description Language
WSML	Web Service Modeling Language

Definition

Service discovery is the process of locating existing services that are relevant for a given request based on the description of their functional and non-functional semantics. Approaches to service discovery differ in their support of service description language(s), the organization of the search, and the utilized means of service selection.

Introduction

The continuous proliferation of web services which encapsulate business software and hardware assets, e-business, or social software applications in the web 2.0 holds promise to further revolutionize the way of interaction within today's society and economy. A service can be defined as a kind of action, performance, or promise that is exchanged for value between provider and client. In other words, it is a provider-client interaction that creates and captures value for all parties involved. At present, there are tens of thousands of web services for a huge variety of applications and in many heterogeneous formats available for the common user of the web. One main challenge of web service technology is to provide scalable and effective means for an automated discovery of relevant services with minimal human intervention in any user and application context. This paper provides an overview of service discovery in a nutshell. For a more comprehensive survey on the subject, the interested reader is referred to, for example, Crasso et al. (2011) and Klusch (2008b, 2012).

Preliminaries

Service discovery can be performed in different ways depending on how the services of the considered search space are described, how the search process is organized, and which means of service selection are used for the search.

Service Description In general, a web service can be described in terms of what it does and how it actually works. These aspects of its functional semantics (aka capability) are described in a service profile and a service process model, respectively.

A service profile describes the signature of a service in terms of its input and output (I/O) parameters and the service specification, i.e., the preconditions and effects (P/E) of the service execution. The profile also describes nonfunctional service semantics such as information about its provenance, name, business category, pricing, delivery constraints, and quality. Prominent approaches to represent such profiles are the XML-based web service description languages WSDL (Chinnici et al. 2007), SML (Pandit et al. 2009), USDL (Oberle et al. 2013), and WADL (Hadley 2009) and the HTML micro-format hREST (Kopecky et al. 2008). Other examples are the textual documentations of RESTful services (Fielding and Taylor 2002) and the ontology-based service description languages OWL-S (Martin et al. 2004), WSML (De Bruijn and Lausen 2005), SAWSDL (Farrell and Lausen 2007), SA-REST (Gomadam et al. 2010), and Linked USDL (Pedrinaci and Leidig 2011).

A service process model describes the operational behavior of a service in terms of its internal control and data flow. Such models are described, for example, in OWL-S, WSML, and USDL by use of standard workflow operators like sequence, split + join, and choice, while other representation approaches are adopting process algebraic languages like the pi-calculus and Petri nets for this purpose.

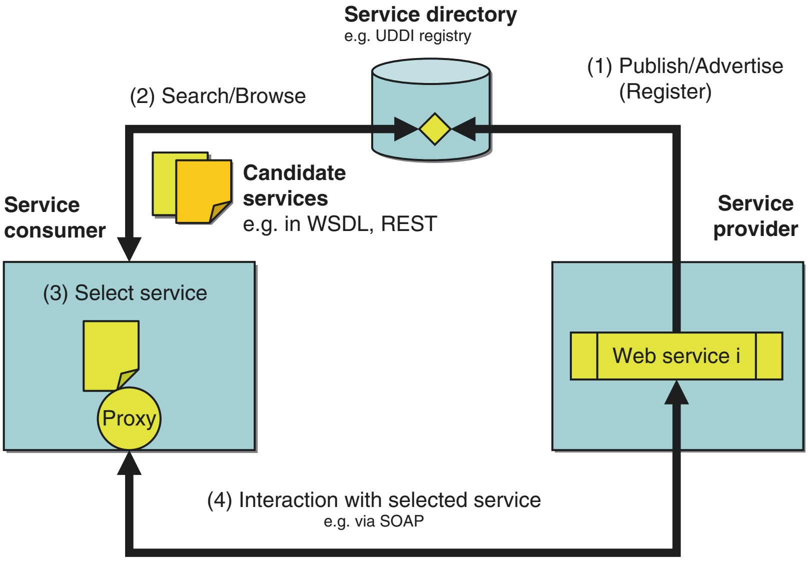
Discovery Architectures Approaches to organize the service search can be classified as either directory-based (aka structured) or directory-less (aka unstructured), or hybrid peer-to-peer (P2P). In the scenario of a directory-based search, service providers register their services with either one central and possibly replicated directory or multiple distributed (federated) service directories at distinguished nodes of the underlying network. Service consumers are informed about available services in the network only through these directory nodes.

Centralized directory-based service discovery can be performed by using either a contemporary web search engine or a specialized web service search engine or a dedicated and authoritative web service directory with query interface. In any case, the W3C web service interaction lifecycle for service-oriented architectures (SOA) expects a central service directory to act as an intermediary between provider and consumer (cf. Fig. 1), though it represents a potential single point of failure and performance bottleneck for dependent applications.

Decentralized directory-based service discovery relies on a structured P2P network overlay and a respective query routing protocol. In this case, services are placed and discovered by all peer nodes according to the global distribution or replication scheme and the location mechanism of the network. Classic examples of structured P2P overlays are the DHT-based Chord ring, Pastry, Tapestry, CAN, P-Grid, or a compound routing index, and a hierarchically structured federation of service directories with super-peers. In general, this type of service discovery provides a search guarantee in the sense of total recall and logarithmic complexity in the size of the network for finding popular, i.e., highly replicated, as well as rare services. On the other hand, it comes at the cost of high communication overhead for publishing and maintaining the structured overlay when peers are joining or leaving the network or the set of services which they provide changes.

Directory-less service discovery is performed in an unstructured P2P network without any given overlay structure. Each peer initially knows only about services provided by its own or its direct

Service Discovery, Fig. 1 W3C web service interaction life cycle



neighbor peers. Prominent examples of service location or query routing schemes in such networks are query flooding and k-random walks with replication and caching strategies, as well as informed probabilistic adaptive search. This type of service discovery is effective for finding popular but not rare services and provides only probabilistic search guarantees, i.e., incomplete recall.

Hybrid P2P service discovery is performed in networks with structured and unstructured overlay parts. For example, service requests can be routed to super-peers in the structured overlay part in order to find relevant rare services or processed with restricted flooding or broadcasting to peers of the unstructured network part to find relevant popular services.

Service Selection The performance of service discovery depends, in particular, on the used service selection method. The process of service selection (aka service matchmaking) encompasses (a) the pairwise semantic matching of a given service request with each service that is registered with the matchmaker and (b) the semantic relevance ranking of these services. In contrast to service brokers, a matchmaker only returns a rank list of relevant services and related provenance information to its human user or application but does not handle the interaction with selected services. In principle, a matchmaker can be used for any organizational approach to service discovery. For example, matchmakers can be part of either the query interface of one central directory or federated directories or local directories owned by peers in an unstructured P2P network (Klusch and Sycara 2001).

Types of Service Selection Current approaches to the semantic matching of web services can be classified as non logic-based, logic-based, or hybrid, depending on the nature of reasoning means used for this purpose. Non logic-based semantic matching exploits, for example, means of graph matching, schema matching, data mining, and text similarity measurement, while logic-based semantic matching performs logical reasoning on service descriptions. Hybrid semantic matching is a combination of both types of matching, while adaptive selection means learn how to best aggregate different matching filters off or on line. In any case, it is commonly assumed that service requests and offers are given in the same format or are appropriately transformed by the considered service matchmaker.

Benchmarking Systems and tools for service discovery, in particular service matchmakers, can be evaluated according to the following five criteria: (1) the support of different service description formats and languages; (2) the usability of the tool and required amount of effort for its configuration; (3) the support of service composition planning through, for example, context-aware pruning of the search space or interactive recommendations for a step-wise forward or backward chaining of services by the user; (4) the policy to preserve user data privacy; and (5) the service retrieval performance in terms of correctness and average query response time over given service test collections. Correctness is commonly evaluated with classical information retrieval measures such as average precision and macro-averaged precision at standard recall levels for binary relevance, as well as the normalized discounted cumulative gain or Q measure for graded relevance. Current evaluation initiatives include the WS Challenge and the SWS Challenge for (semantic) web service composition and the S3 Contest for semantic web service selection (Klusch 2012; Küster et al. 2009).

Web Service Discovery

Most web services are described in the standard WSDL, USDL, or according to the REST paradigm of the web. Some service providers also publish the functional description of their services in multiple formats and languages. The number and variety of web services which are available in the public web appears tremendously high, though there are still no common and comprehensive statistics on the subject available. However, the portal seekda.com reported about 30 k web services in November 2011, and the public directory programmableweb.com alone already offered about 16 k single or composite RESTful web services in March 2013. In this section, we focus on the discovery of WSDL and RESTful services.

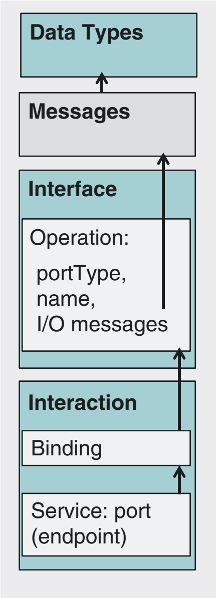
WSDL Services The W3C web services framework offers a set of technical specifications including WSDL and SOAP SOAP 2007 that codify mechanisms for XML-based interoperability between business services that are accessible in the web over stateless HTTP. A web service which profile is described in WSDL (in short WSDL service) exposes one or multiple operations which consume inputs and produce outputs both encoded in XML. Applications or other services can interact with these operations by means of XML-SOAP messaging.

Description The XML-based W3C standard language WSDL describes the functionality of a service by the set of signatures of its service operations and the set of network endpoints or ports (URIs) at which these operations can be invoked and how this can be achieved (Fig. 2). In particular, each port is associated with a respective interface which binds the operation to a given protocol for transport and messaging. The definitions of the I/O messages of each service operation include references to their data types which are defined in common XMLS namespaces. Several non functional service parameters can be added to such a WSDL service profile on demand. The description of service profiles in WSDL remains stateless, since the specification of service preconditions and effects is not part of the standard. Besides, a WSDL service description does not include any process model. In this respect, WSDL is commonly considered as weak in describing what the service actually does.

Discovery and Selection Most approaches to directory-based or directory-less discovery of WSDL services utilize means of non logic-based semantic selection, in particular, structural XML and text similarity-based matching.

Central directory-based discovery of WSDL services is the most popular. One classic example is the instantiation of the W3C service interaction lifecycle (cf. Fig. 1) with some UDDI-compliant (Bellwood et al. 2004) registry of WSDL services and using SOAP (Mitra and Lafon 2007) for service interaction. In such an XML-based UDDI business registry (UBR), the services and their providers are categorized with standard taxonomies such as NAICS, SIC, and UNSPSC. Registration of WSDL services and their retrieval from a UBR is through its APIs PublishSOAP and InquireSOAP. In general, a UBR may provide information on the business entities of services (aka white pages), service categories (aka yellow pages), and the technical model (tModel) of services (aka green pages). Search queries to a UBR are regular expressions with identifiers and keywords for service tModels, names, and categories.

Service Discovery, Fig. 2 Example of web service description in WSDL



```
<description xmlns=“http://www.w3.org/ns/wsdl”>
<types> <xs:schema … xmlns:pdc=“http://www.parts-
            depot.com/schemas/pdc“ … >
   <xs:element name = “cid" type = "xs:string"/>…</types>
<interface name = “PartsListInterface">
  <operation name = “GetPartsList"
            pattern = “http://www.w3.org/ns/wsdl/in-out”…>
   <input messageLabel = "In" element = “xs:cid" />
   <output messageLabel="Out” element = “pdc:parts_list“/>
  </operation> </interface>
<binding name = “PListHTTPBinding”
  type = “http://www.w3.org/ns/wsdl/http”
  interface = "tns:PartsListInterface“
  <operation ref = "tns:GetPartsList" whttp:method="GET”/>
</binding>
<service name = “PartsDepot“
           interface = "tns:PartsListInterface">
 <endpoint name =“PListHTTPEndpoint
            binding = "tns:PListHTTPBinding"
            address = “http://www.parts-depot.com/parts/”>
</endpoint> </description>
```

Accordingly, service selection by a UBR is, in principle, based on string matching without any logical reasoning on service relationships or non functional service parameters. Thus, it requires a rather cumbersome browsing of the registry by the user to find relevant services. Since 2005, UDDI is not supported by its originally main supporters IBM and Microsoft.

Examples of non-UDDI compliant WSDL service directories are RemoteMethods.com, Xmethods.net, WebserviceX.net, webservicelist.com, service-repository.com, and wsindex.org. Most of them rely on keyword search, and service category or simple list browsing. An example of a specialized web service search engine is Woogle (Dong et al. 2004) which retrieves and indexes WSDL services from a given set of UBRs. The WSDL service selection tool WSDLAnalyzer (Zinnikus et al. 2006) returns a rank list of similar WSDL services for a given WSDL service and produces a mapping between their I/O messages. In particular, it recursively computes the XML-tree similarity of a given pair of WSDL files with integrated text matching of tree node names, using WordNet-distance and string matching, and a binary compatibility check of XMLS data types. Other approaches to WSDL service selection exploit techniques for matching software components, graphs, or schemas (Stroulia and Wang 2005), or perform a full-text matching of service names or the content of WSDL files as a whole. In addition, there are approaches to preference-, trust- or reputation-based matching of non functional parameters including quality of service, pricing, and service policies (Crasso et al. 2011; Garofalakis et al. 2006).

Decentralized directory-based discovery of WSDL services in structured P2P networks still appears in its infancies. One example is the PWSD system (Li et al. 2004) in which WSDL files and requests are distributed and located in a Chord ring of service peers. The DUDE system (Banerjee et al. 2005) enables WSDL service discovery in a hierarchical DHT-based overlay for multiple local UDDI registries. There is no approach to directory-less discovery of WSDL services in unstructured P2P networks available yet.

REST Services Web service interaction is not restricted to XML-SOAP messaging. A RESTful web service (in short REST service) represents resources which states shall be accessed only over the stateless HTTP according to the REST paradigm of the web Fielding and Taylor 2002.

The call of a REST service with given input values may return output values in XML or in the text-based JSON or RSS formats. For example, the call of some REST service "books" hosted at a portal www.bookstore. com with input parameter "subject" for books on the topic Eclipse is of the form http://www.bookstore.com/books/?subject=computers/eclipse and may return book list entries like <booklist:book url = http://www.bookstore.com/books/0321288157title="Eclipse Distilled"/> in XML.

Description At present, there is no standard for describing the functionality of REST services. Most REST service APIs are documented by their developers on dedicated, public HTML pages in more or less plain text and tables; some APIs are described in XML-based WADL files or the HTML micro-format hRESTS. This heterogeneity is a major barrier for the automated discovery of REST service APIs in the web to date.

Discovery and Selection Centralized directory-based discovery of REST services can be performed with the prominent directory programmableweb.com. It offers about 9 k REST service APIs and 7 k REST service mash-ups (as of April 2013). Another open source REST API directory in the web is APIS.io. The selection of relevant REST services through their query interfaces is done by keyword search which relies on the textual description of the registered service APIs or other meta-information provided by their developers. The web services search engine seekda! identifies relevant REST service APIs based on adaptive text classification and feature extraction. An approach to automated extraction of information from REST service APIs like service operation name, description, and URI is proposed in Ly et al. (2012). It integrates means of DOM processing, information extraction, and natural language processing. An approach to structural and textual matching of REST services is proposed in Khorasgani et al. (2011). In this case, a given pair of REST service APIs is first semi automatically converted into WADL descriptions. The REST service matching score is then computed as the maximum flow in the graph of WADL service description elements.

Approaches to directory-less discovery of REST services in mobile ad hoc networks mostly rely on simple lookup methods based on the matching of service classes, UUID, or service attribute names (Schiele et al. 2004).

Semantic Web Service Discovery

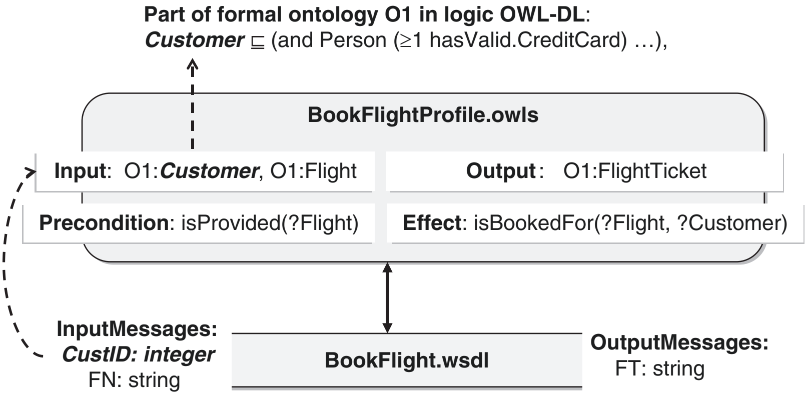
One major challenge of automated service discovery is to make service-based applications or intelligent software agents actually "understand" the semantics of service requests and offers. From the perspective of strong AI, this requires some well-founded logic specification of service profile and process model. However, contemporary web service descriptions are lacking such formal semantics. It is well known that this problem can be addressed by exploiting semantic web technologies (Hitzler et al. 2011).

Description The key idea of encoding web service semantics not only in a machine-readable but machine- understandable way is as follows: The semantics of web service interface elements are described by references to appropriate concepts and rules which are formally defined in a shared ontology in some W3C standard ontology language like RDFS or OWL2. Such semantically annotated web services are called semantic web services (in short semantic services). Current frameworks for semantic service description include OWL-S (Martin et al. 2004), WSML (De Bruijn and Lausen 2005), the W3C standard SAWSDL (Farrell and Lausen 2007), and Linked USDL (Pedrinaci and Leidig 2011) which is USDL modeled in RDFS. These ontology-based semantic service description languages mainly differ in their formal logic-based foundation and the possible extent of annotating services.

OWL-S In OWL-S the service I/O parameters are annotated with concepts which are exclusively defined in the formal logic-based W3C standard ontology language OWL2 (cf. Fig. 3). Service preconditions and effects may be specified in the formal semantic web rule language SWRL.

WSML The description of service profile semantics in one of five variants of WSML is formally grounded in the respective variant of the logic programming language F-Logic (Fensel et al. 2010). Both, WSML and OWL-S, are also

Service Discovery, Fig. 3 Example of semantic service profile in OWL-S


Part of formal ontology O1 in logic OWL-DL:
Customer ⊑ (and Person (≥1 hasValid.CreditCard) …),
BookFlightProfile.owls
Input: O1:Customer, O1:Flight
Output: O1:FlightTicket
Precondition: isProvided(?Flight)
Effect: isBookedFor(?Flight, ?Customer)
InputMessages:
CustID: integer
FN: string
BookFlight.wsdl
OutputMessages:
FT: string

providing the developer with a set of workflow operators like sequence, iterate, choice, and split + join for specifying the operational semantics of a single or composite service in its process model. The process model can be mapped to service orchestrations in BPEL as the semantic service can be grounded with a WSDL service.

SAWSDL and SA-REST The W3C standard SAWSDL allows the annotation of WSDL service elements with references to web resources of any media type such as plain text, video, picture, audio podcast, and concepts in a formal ontology. The same approach is taken in the SA-REST framework for semantically annotating REST service APIs (Gomadam et al. 2010). Both SAWSDL and SA-REST do not allow the specification of preconditions and effects, and the handling of semantic annotations is completely outside these frameworks. In this sense, unlike OWL-S and WSML, neither of both has unique formal semantics. For more details on semantic service description, the reader is referred to, for example, Klusch (2008a) and the above cited relevant technical specifications.

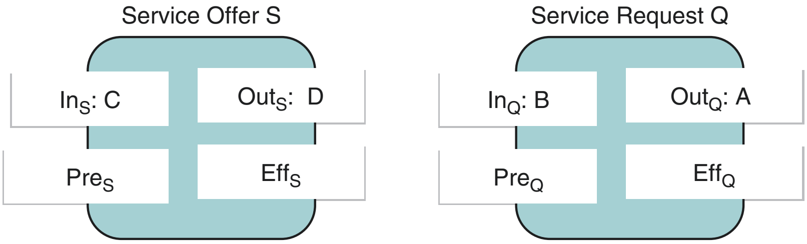
At present, there are no public statistics about semantic web services available. A survey conducted with the semantic service search engine Sousuo (Klusch and Xing 2008) in April 2013 reported about 3500 semantic services in OWL-S, WSML, WSDL-S, and SAWSDL in the public web, though most of them are available only in distinguished test collections.

Discovery and Selection In the past decade, the semantic web research community has developed a wide range of solutions for the automated discovery and selection of semantic services. The degree of semantic correspondence between a pair of semantic web services particularly relies on the matching of the semantic annotations of their service profile and/or process model.

Types of Selection The types of semantic service selection are logic-based, non logic-based, and hybrid semantic. Classical examples of logic-based semantic matching filters are the logical I/O concept subsumption-based plugin match of service signatures and the logical specification plugin match of preconditions and effects (cf. Fig. 4). Logical and full functional (IOPE) profile matching combines the scores of logical signature (IO) and specification (PE) matching. Non logic-based semantic matching of annotated service signatures is mostly based on the textual similarity of the concept names or the text of their logical unfolding in the referenced ontology. Additional examples include the structural similarity-based matching of I/O concepts in terms of the shortest path or upward co-topic distances between them in the shared ontology.

Currently, most approaches to semantic service selection are hybrid, i.e., they combine non logic-based with logic-based semantic service matching. Besides, the majority of them support either OWL-S or SAWSDL, but only a few are devoted to WSML, or other description formats, and hardly any matchmaker is even language-agnostic (Klusch 2012). In the following, we focus on approaches to the discovery and selection of services in OWL-S and SAWSDL. More

Service Discovery, Fig. 4 Logic-based semantic service plugin matching



Logical Signature Plug-In Match of S with Q:

$$(\forall C \in In_S \ \exists B \in In_Q: \ B \sqsubseteq C) \wedge$$
$$(\forall A \in Out_Q \ \exists D \in Out_S: \ D \sqsubseteq A)$$

Logical Specification Plug-In Match of S with Q:

$$KB \models (Pre_Q \Rightarrow Pre_S) \wedge (Eff_S \Rightarrow Eff_Q)$$

information on the subject is provided, for example, in Klusch (2008b, 2012).

Centralized Discovery and Selection There are quite a few tools and systems for central directory-based discovery of semantic services available.

Matchmakers. For example, the matchmaker iSeM (Klusch and Kapahnke 2012) performs an adaptive and hybrid semantic selection of OWL-S services. Its logic-based semantic matching of services relies on the computation of strict and approximated logical I/O concept subsumption relations and the logical specification plugin relation. Like its predecessor OWLS-MX2 (Klusch et al. 2009), it also performs non logic-based semantic matching with different classical token-based text similarity measures, as well as ontology-based structural matching of signature annotation concepts. Finally, it learns how to best aggregate the results of its matching filters by use of a binary SVM relevance classifier with an evidential coherence-based weighting scheme.

An example of a hybrid semantic and adaptive matchmaker for SAWSDL services is LOG4SWS (Schulte et al. 2010). Like iSeM it performs a logical service signature matching which is complemented with ontology-based structural matching based on the shortest path lengths between concepts. In case there are no semantic annotations of WSDL service signature elements, it exploits the WordNet distance between the element names. LOG4SWS does not consider service preconditions and effects, but learns off line how to best aggregate the matching results by use of an ordinary least square-based classifier.

The logic-based semantic service matchmaker SPARQLent (Sbodio et al. 2010) considers the full functional profile of OWL-S services. It performs an RDF entailment rule-based matching of I/O concepts, preconditions, and effects described in SPARQL.

According to the results of the international S3 contest (Klusch 2012), iSeM and LOG4SWS are currently the best performing matchmakers for OWL-S and SAWSDL services, respectively. In fact, they provide the best trade-off between average precision and response time.

An example of a hybrid semantic matchmaker for WSML services is WSMO-MX (Klusch and Kaufer 2009): It recursively determines service matching degrees based on ontology-based signature parameter type matching, logical constraint (PE) matching, and syntactic matching with text similarity measurements.

Specialized search engines. Examples of search engines for semantic services are S3E (Giantsiou et al. 2009) and Sousuo (Klusch and Xing 2008). The latter performs a meta-search through the public web search engines Google and A9 and complements it by crawling the web with its own focused topic crawler. It also utilizes the semantic web search engine Swoogle for an inverse ontology-based search and performs a full-text search of the public scientific archive citeseer in the web. Service selection through

Sousuo's query interface relies on full-text or keyword search in its XML-encoded service index.

Alternatively, the S3E engine is encoding the profiles of crawled semantic services in RDF. The selection of services from an internal RDF store with SPARQL relies, in particular, on textual matching of profile parameters. Another search engine which is restricted to a QoS-based discovery of semantic services is presented in Vu et al. (2006).

Registries. At present, there are no central and authoritative registries of semantic services available in the public web. Public collections of semantic services are, for example, the prominent OWLS-TC for OWL-S services, the SAWSDL-TC for SAWSDL services, and hREST-TC for annotated REST services; each of these collections is available at the portal semwebcentral.org. iServe (Pedrinaci et al. 2010) is a software platform that can be used to build and maintain a registry of semantic services described in SAWSDL, OWL-S, MicroWSMO, and WSMO-Lite. The services are internally represented in iServe according to a minimal service model and then exposed in HTML and RDF as linked services with a unique and resolvable HTTP URI. Any iServe registry can be queried through an SPARQL endpoint. For service selection, iServe provides means of keyword search, functional classification, and service I/O parameter matching based on RDFS reasoning.

Centralized P2P search. An example for the discovery of WSDL-S (a predecessor of SAWSDL) services in a structured P2P system is the METEOR-S system (Verma et al. 2005). It consists of a set of service-providing and service-consuming peers which may form groups on given domains or topics and one central super-peer which serves as a central service matchmaker for all peers. For this purpose, the super-peer maintains and utilizes a global registry ontology which covers the concept taxonomies of all local service registries of peers in the network. The super-peer also provides the peers with mappings between the message types and signature annotation concepts of registered services. The non logic-based semantic selection of services by the super-peer relies on structural XMLS matching and the computation of NGram-based text similarities and taxonomic relations. The super-peer can be replicated for reasons of scalability.

Decentralized Discovery and Selection A directory-based discovery of OWL-S services in structured P2P systems can be performed, for example, with the AGORA-P2P system (Küngas and Matskin 2006). It relies on a Chord ring for distributed storage and location of services. In particular, the service signature concept labels are hashed as literals to unique integer keys such that peers holding the same key are offering services with equal literals in the circular key space. Service selection for multi-key queries relies on exact key matching.

Directory-less discovery of semantic services can be performed with, for example, the RS2D system (Basters and Klusch 2006). It is a solution for informed and adaptive probabilistic service search in unstructured P2P networks. In particular, each peer dynamically builds and maintains its local view of the semantic overlay of the network and uses the OWLS-MX matchmaker for hybrid semantic service selection. A peer also learns the average query-answering behavior of its direct neighbors in the network. The peer's decision to whom to forward a semantic service request is then driven by its estimated probabilistic risk of routing failure in terms of semantic loss and communication costs. Other examples are discussed, for example, in Klusch (2008b) and Staab and Stuckenschmidt (2006).

Future Directions

Despite the progress made in the field in the past decade, a major open problem is the scalable and dynamic interleaving of discovery of services with their composition, negotiation, and execution in the converging Internet of Things and Internet of Services. Examples of potential applications of solutions are intelligent condition monitoring based on large-scale, wireless, and semantic sensor service networks; the intelligent collaborative design of products in shared 3D spaces; and mobile ad hoc and context-aware business travel planning or product recommendation services.

Cross-References

- ▶ RDF
- ▶ Web Ontology Language (OWL)
- ▶ Web Service Composition
- ▶ WSDL

References

Banerjee S, Basu S, Garg Sh, Garg S, Lee SJ, Mullan P, Sharma R (2005) Scalable grid service discovery based on UDDI. In: Proceedings of the 3rd international workshop on middleware for grid computing, Grenoble. ACM

Basters U, Klusch M (2006) RS2D: fast adaptive search for semantic web services in unstructured P2P networks. In: Proceedings of the 5th international semantic web conference, Athens. Springer

Bellwood P, Capell S, Clement L, Colgrave J, Dovey MJ, Feygin D, Hately A, Kochman R, Macias P, Novotny M, Paolucci M, von Riegen C, Rogers T, Sycara K, Wenzel P, Wu Z (2004) UDDI 3.02. http://uddi.org/pubs/uddi_v3.htm

Chinnici R, Moreau JJ, Ryman A, Weerawarana S (2007) Web services description languages 2.0. W3C recommendation. http://www.w3.org/TR/wsdl20/WSDL1. 1 (2001): www.w3.org/TR/wsdl

Crasso M, Zunino A, Campo M (2011) A survey of approaches to web service discovery in service-oriented architectures. J Database Manag 22(1):102–132. IGI Global

De Bruijn J, Lausen H (eds) (2005) Web service modeling language (WSML). http://www.w3.org/Submission/WSML/

Dong X, Halevy AY, Madhavan J, Nemes E, Zhang J (2004) Simlarity search for web services. In: Proceedings of the 30th conference on very large databases, Toronto

Farrell J, Lausen H (2007) Semantic annotations for WSDL and XML schema. www.w3.org/TR/sawsdl/

Fensel D, Lausen H, Polleres A (2010) Enabling semantic web services. Springer, Heidelberg

Fielding RT, Taylor RN (2002) Principled design of the modern web architecture. J Trans Internet Technol 2(2):115–150. ACM

Garofalakis J, Panagis Y, Sakkopoulos E, Tsakalidis A (2006) Contemporary web service discovery mechanisms. J Web Eng 5(3):265–289. Rinton Press

Giantsiou L, Loutas N, Peristeras V, Tarabanis K (2009) Semantic service search engine (S3E): an approach for finding services on the web. In: Proceedings of the 2nd world summit on the knowledge society, Crete. Springer

Gomadam K, Ranabahu A, Sheth A (2010) SA-REST: semantic annotation of web resources. www.w3.org/Submission/2010/SUBM-SA-REST-20100405/

Hadley M (2009) Web application description language. www.w3.org/Submission/wadl/

Hitzler P, Krötzsch M, Rudolph S (2011) Foundations of semantic web technologies. CRC, Boca Raton/New York/London

Khorasgani RR, Stroulia E, Zaiane OR (2011) Web service matching for RESTful web services. In: Proceedings of the 13th IEEE international symposium on web systems evalution. IEEE, Williamsburg

Klusch M (2008a) Semantic web service description. In: Schumacher M, Helin H, Schuldt H (eds) CASCOM-intelligent service coordination in the semantic web, chapter 3. Birkhäuser, Basel

Klusch M (2008b) Semantic web service coordination. In: Schumacher M, Helin H, Schuldt H (eds) CASCOM – intelligent service coordination in the semantic web, chapter 4. Birkhäuser, Basel

Klusch M (2012) Overview of the S3 contest: performance evaluation of semantic service matchmakers. In: Blake MB, Cabral L, König-Ries B, Küster U, Martin D (eds) Semantic web services, chapter 2. Springer, Berlin/London

Klusch M, Kapahnke P (2012) The iSeM matchmaker: a flexible approach for adaptive hybrid semantic service selection. J Web Semant 15:1–14. Elsevier

Klusch M, Kaufer F (2009) WSMO-MX: a hybrid semantic web service matchmaker. J Web Intell Agent Syst 7(2):1–14. IOS Press

Klusch M, Sycara K (2001) Brokering and matchmaking for coordination of agent societies: a survey. In: Omicini A et al (eds) Coordination of internet agents, chapter 8. Springer, Berlin/New York

Klusch M, Xing Z (2008) Deployed semantic services for the common user of the web: a reality check. In: Proceedings of the 2nd IEEE international conference on semantic computing (ICSC), Santa Clara. IEEE

Klusch M, Fries B, Sycara K (2009) OWLS-MX: a hybrid semantic web service matchmaker for OWL-S services. J Web Semant 7(2):121–133. Elsevier

Kopecky J, Gomadam K, Vitvar T (2008) HTML microformat for describing RESTful web services and APIs. In: Proceedings of the international conference on web intelligence, Sydney. IEEE

Küngas P, Matskin M (2006) Semantic web service composition through a P2P-based multi-agent environment. In: Proceedings of the 4th international workshop on agents and peer-to-peer computing, Utrecht. LNCS 4118. Springer

Küster U, Koenig-Ries B, Klusch M (2009) Evaluating semantic web service technologies: criteria, approaches and challenges. In: Lytras MD, Sheth A (eds) Progressive concepts for semantic web evolution: applications and developments. IGI Global, Hershey

Li Y, Zou F, Wu Z, Ma F (2004) PWSD: a scalable web service discovery architecture based on peer-to-peer overlay network. In: Proceedings of the APWeb04, Hangzhou. LNCS 3007. Springer
Ly PA, Pedrinaci C, Domingue J (2012) Automated information extraction from web APIs documentation. In: Proceedings of the 13th international conference on web information system engineering, Paphos. Springer
Martin D, Burstein M, Hobbs J, Lassila O, McDermott D, McIlraith S, Narayanan S, Paolucci M, Parsia B, Payne T, Sirin E, Srinivasan N, Sycara K (2004) OWL-S.: semantic markup for web services. www.w3.org/Submission/OWL-S/
Mitra N, Lafon Y (2007) SOAP 1.2 Primer. www.w3.org/TR/2007/REC-soap12-part0-20070427/
Oberle D, Barros A, Kylau U, Heinzl S (2013) A unified description language for human to automated services. J Inf Syst 38(1):155–181. Elsevier
Pandit B, Popescu V, Smith V (2009) Service modeling language 1.1. www.w3.org/TR/sml/
Pedrinaci C, Leidig T (2011) Linked USDL Core. www.linked-usdl.org/ns/usdl-core
Pedrinaci C, Liu D, Maleshkova M, Lambert D, Kopecky J, Domingue J (2010) iServe: a linked services publishing platform. In: Proceedings of the 7th extended semantic web conference workshop on ontology repositories and editors for the semantic web, Heraklion. LNCS. Springer
Sbodio ML, Martin D, Moulin C (2010) Discovering semantic web services using SPARQL and intelligent agents. J Web Semant 8(4):310–328. Elsevier
Schiele G, Becker C, Rothermel K (2004) Energy-efficient cluster-based service discovery for ubiquitous computing. In: Proceedings of the 11th ACM SIGOPS European workshop, Belgium
Schulte S, Lampe U, Eckert J, Steinmetz R (2010) LOG4SWS.KOM: self-adapting semantic web service discovery for SAWSDL. In: Proceedings of the 6th world congress of services, Miami. IEEE
SOAP (2007) W3C recommendation SOAP 1.2. www.w3.org/TR/2007/REC-soap12-part0-20070427/
Staab S, Stuckenschmidt H (eds) (2006) Semantic web and peer-to-peer. Springer, Berlin
Stroulia E, Wang Y (2005) Structural and semantic matching for assessing web-service similarity. J Coop Inf Syst 14:407–437. World Scientific
Verma K, Sivashanmugam K, Sheth A, Patil A, Oundhakar S, Miller J (2005) METEOR–S WSDI: a scalable P2P infrastructure of registries for semantic publication and discovery of web services. J Inf Technol Manag 6(1):17–39. Kluwer
Vu L-H, Hauswirth M, Porto F, Aberer K (2006) A search engine for QoS-enabled discovery of semantic web services. J Bus Process Integr Manag 1(4):244–255
Zinnikus I, Rupp H-J, Fischer K (2006) Detecting similarities between web service interfaces: the WSDL analyzer. In: Proceedings of the 2nd international workshop on web services and interoperability, Bordeaux. Wiley

Service Search and Selection

▶ Service Discovery

Service Systems

▶ Queueing Theory

Service-Oriented Architectures

▶ Service Discovery

Sex Industry

▶ Pornography Online

Sharding

▶ Weblog Analysis

Siena: Statistical Modeling of Longitudinal Network Data

Tom A. B. Snijders
Department of Sociology, University of Groningen, Groningen, The Netherlands
Department of Statistics and Nuffield College, University of Oxford, Oxford, UK

Synonyms

Coevolution of networks and behavior; Network dynamics; Network panel data; Peer influence; Statistical modeling

Glossary

Network panel data	longitudinal data consisting of two or more repeated observations of a network on a given set of nodes
Panel waves	the data observed for one given observation moment in a panel study
Social actors	individuals, companies, etc., represented by the nodes in the network
Stochastic actor-oriented model	a probability model for network dynamics where changes may take place at arbitrary moments in continuous time, and where these changes are regarded as consequences of choices made by the actors
RSiena	R package implementing statistical inference according to a stochastic actor-oriented model given network panel data
Effects	model components defining the probabilities of tie changes in the stochastic actor-oriented model
Method of Moments	one of the traditional methods in statistics for parameter estimation
Dependent variable	the variable defining the outcome space in a statistical model

Definition

The name "Siena" stands for *Simulation Investigation for Empirical Network Analysis*. It is a method for the statistical analysis of longitudinal network data, observed in two or more panel waves. This method was implemented in the standalone program *Siena*, first released in 1997, going through many versions, and superseded by the R package *RSiena* in 2009. *Siena* was programmed by Tom Snijders in Delphi, with contributions by Christian Steglich, Mark Huisman, and Michael Schweinberger. *RSiena* was originally programmed by Ruth Ripley and Krists Boitmanis, under the direction of Tom Snijders. Since 2012 it is maintained by Tom Snijders, in collaboration with Christian Steglich and Johan Koskinen; other contributors have been Josh Lospinoso, Charlotte Greenan, Paulina Preciado, and Felix Schönenberger. Current versions of *RSiena* and its development version *RSienaTest* are available from R-Forge, http://r-forge.r-project.org/R/?group_id=461.

RSiena is a contributed package of the statistical software system R, and as such is free, distributed under the GNU Public License, running under Unix-like, Windows, and Mac families of operating systems. The methods used are based on Monte Carlo simulation, and therefore can be time-consuming for larger data sets. The package is programmed in a combination of R and C++, the latter for the computationally intensive parts.

The orientation of the *Siena* method is primarily to the social sciences, but this of course is not exclusive.

Introduction

Statistical modeling is based on model assumptions, mostly assumptions about independence or conditional independence, but one of the main characteristics of networks is the strong and complex dependence between network ties. If the assumptions in statistical modeling are not good representations of the data structures or the mechanisms that may have led to the observed data, then results of statistical inference can be grossly misleading. This leads to difficulties in proposing plausible statistical models for network data.

Modeling longitudinal network data can be simpler than modeling single observations of networks, because the time structure poses a constraint on the dependence structure: the present depends on the past, not on the future. The *Siena* method is based on a probability model that represents network dynamics as a Markov chain running in continuous time, called the *stochastic actor-oriented model*. The basic state space is the set of all digraphs on a given node set. The model has been expanded to allow multiple (i.e., multivariate) digraphs and also actor-based variables as components of the state space. The actor-based variables are usually referred to as "behavior," thus allowing the modeling of the coevolution, or interdependent dynamics, of networks and behavior. Thus, in the basic type of stochastic actor-oriented model, there is one dependent variable, viz., a directed network; in the extended models, there can be several dependent networks and also one or more dependent actor-based variables. Two-mode networks can also be included as dependent networks.

Key Points

The *Siena* method is defined in the usual paradigm of statistical modeling. It presupposes the availability to the user of *network panel data or network and behavior panel data*. This means that for a given node set, at a finite number (two or more) of observation points (also called *panel waves*), a network on this node set was observed, possibly complemented with a behavioral variable. In the ideal case the node set is constant and the data are complete; some changes in the node set (nodes entering or exiting) and some fraction of data being missing is allowed. Up to 10% of missing tie variables is in practice not a problem, more than 20% is not advisable.

The nodes are supposed to represent *social actors*, and the model is said to be *actor-oriented*, meaning that tie changes are regarded as the consequence of choices made by the senders of the ties. The user specifies a model by defining a set of effects (see below), which are model components defining the probabilities of tie changes. Given the model specification, the *RSiena* package can estimate parameters (which are coefficients indicating the strength of the effects) and test hypotheses about the parameters. With a given specification and given parameters, *RSiena* can also be used to simulate the dynamics of a network.

Historical Background

A historical overview of early work on probability models for network dynamics is given in Snijders (1995), which also was the first paper about stochastic actor-oriented models. Some important papers that are part of the general background preceding the work on this methodology are Holland and Leinhardt (1977), Wasserman (1979, 1980), Zeggelink (1994), and Leenders (1995).

The development of the stochastic actor-oriented model for digraphs was stimulated by the empirical work in van de Bunt (1999). After two precursor papers (Snijders 1996; Snijders and van Duijn 1997), the main presentation of this model was given in Snijders (2001). Methods for the coevolution of networks and behavior were developed in Snijders et al. (2007) and elaborated in Steglich et al. (2010); the extension to the coevolution of multiple (multivariate) networks was presented in Snijders et al. (2013). All these papers use the Method of Moments (one could also say, the method of Estimating Equations) to estimate the parameters. This is the main method implemented in *RSiena*. In addition, Bayesian methods (Koskinen and Snijders 2007) and an algorithm for Maximum Likelihood estimation (Snijders et al. 2010a) were developed and implemented, but these are much more time-consuming and therefore are less used.

Statistical Model

This section outlines the basic probability model implemented in *RSiena*, for the basic case of one dependent variable, assuming this is a one-mode network. Further elaboration and details can be found in the publications mentioned above and in Snijders (2009), on which much of the explanation below is based.

The network is represented by the node set $\{1,\cdots,n\}$ with tie variables x_{ij}, where $x_{ij} = 1$ or 0 indicates whether the tie $i \rightarrow j$ is present or absent. The tie variables are collected in the $n \times n$ adjacency matrix $\mathrm{x} = (x_{ij})$. Self-ties are excluded, so that $x_{ii} = 0$ for all i. The concepts of network (directed graph) and matrix (its adjacency matrix) will be used interchangeably. Random variables will be indicated by capitals, and observations, or other nonrandom variables, by lower case. The ties are assumed to be outcomes of time-dependent random variables, denoted by $X_{ij}(t)$ and collected in the time-dependent random matrix $\mathrm{X}(t)$.

In addition to the network X(t), which can be regarded as the dependent variable of the model, there can be other variables, the so-called *covariates*, regarded as independent or explanatory variables in the sense that their values are not modeled but accepted as given, and which may influence the network. Examples are the gender of actors (actor variable) and their spatial proximity (dyadic variable). For conciseness, these are disregarded in this brief overview; in practice, they are included in most data sets and of great practical importance.

Basic Model Definition

The following basic assumptions are made.

1. Time, denoted by t, is a continuous variable.

   This assumption separates time as observed (two or more moments of observation) from time that determines network dynamics (continuous).
2. X(t) is a Markov process.

   This means that the conditional distribution of future states depends on the past only as a function of the present. This assumption corresponds to the network ties being regarded as states rather than events.
3. At any given moment t, no more than one tie variable $X_{ij}(t)$ can change.

   This set of assumptions was first proposed by Holland and Leinhardt (1977), and is very helpful because it allows representing network dynamics as a feedback process, where the actors create the network as the endogenously changing environment for themselves and each other (Zeggelink 1994); while requiring only to specify the probabilities of changes of single tie variables.

In the further model elaboration, two aspects are distinguished: the *change opportunity process* and the *change determination model.*

Opportunity for change. For each actor i, opportunities to establish one new outgoing tie $i \rightarrow j$, or dissolve one existing tie $i \rightarrow j$, occur according to a Poisson process with rate λ_i. This means that the probability that an opportunity for change occurs for actor i in the time interval from t to $t + \varepsilon$, where ε is a small positive number, is approximated (in the limit for ε tending to 0) by $\lambda_i \in$.

Determination of change. When actor i has an opportunity for change, s/he is permitted to choose one of the outgoing tie variables X_{ij} and change this into its opposite value, changing 0 to 1 (creating a new tie) or changing 1 to 0 (terminating an existing tie). The probabilities depend on the so-called *objective function* $f_i(\mathrm{x}^0, \mathrm{x})$, indicating how “attractive” it is to go to state x given the current state x^0. The set of potential new network states, denoted by $C(\mathrm{x}^0)$, is the set composed of x^0 itself together with the $n - 1$ matrices which are equal to x^0 except for exactly one nondiagonal element in line i which is replaced by its opposite, $x_{ij} = 1 - x_{ij}^0$. The probability that the new state is x is given by

$$\begin{aligned} &\mathrm{P}\{X(t) \text{ changes to } x \mid X(t) = x^0, \\ &\qquad\qquad i \text{ has a change opportunity at time } t\} \\ &= p_i(\mathrm{x}^0, \mathrm{x}) = \frac{\exp(f_i(\mathrm{x}^0, \mathrm{x}))}{\sum_{\mathrm{x}' \in C(\mathrm{x}^0)} \exp(f_i(\mathrm{x}^0, \mathrm{x}'))}. \end{aligned} \tag{1}$$

The two model components can be put together by giving the transition rate matrix, also called Q-matrix, of which the nondiagonal elements are defined by

$$\begin{aligned} &q_{\mathrm{x}^0, \mathrm{x}} \\ &= \lim_{dt \downarrow 0} \frac{\mathrm{P}\{\mathrm{X}(t + dt) = \mathrm{x} | \mathrm{X}(t) = \mathrm{x}^0\}}{dt} \; (\mathrm{x} \neq \mathrm{x}^0) \end{aligned}$$

(see textbooks on continuous-time Markov chains, such as Norris 1997). Note that the assumptions imply that $q_{\mathrm{x}^0, \mathrm{x}} = 0$ whenever $x_{ij} \neq x_{ij}^0$ for more than one element (i, j).

For digraphs x and x^0 which differ from each other only in one element in row i, the transition rate is

$$q_{\mathrm{x}^0, \mathrm{x}} = \lambda_i(\mathrm{x}^0) p_i(\mathrm{x}^0, \mathrm{x}). \tag{2}$$

S

Model Specification

The model specification consists of defining the network X (and other dependent variables, if any; see Steglich et al. 2010, and Snijders et al. 2013), covariates, the rate function λ_i, and the objective function f_i. If there are more than one dependent variable, each has its own rate function and objective function. The rate function may be constant between waves, or depend on actor-based variables through an exponential link function. The focus of model specification is on the objective function, specified as a linear combination

$$f_i\left(\mathrm{x}^0, \mathrm{x}\right) = \sum_k \beta_k s_{ki}\left(\mathrm{x}^0, \mathrm{x}\right) \tag{3}$$

where the functions s_{ki} are the so-called *effects* driving the network dynamics, while the weights β_k are parameters indicating the strength of these effects and which can be estimated from the data. The effects represent internal network dependencies as well as dependence on covariates, and are discussed in the mentioned literature. The manual Ripley et al. (2017) contains the long list of implemented effects, and this list is frequently added to because of requests from applied researchers.

Parameter Estimation

The main estimation method implemented in *RSiena* is an application of the Method of Moments (or Estimating Equations). It makes good use of the Markov property by conditioning on the preceding observation. This enables computer simulation of the process in a straightforward way, and does away with the need for an assumption of stationary marginal distributions. The moment equations, or estimating equations, define the parameter estimate θ as a function of the data $\mathrm{x} = x(t_1), \cdots, x(t_M)$ (assuming there are M waves), and are given by

$$\begin{aligned}\sum_{m=1}^{M-1} \mathrm{E}_\theta\{U(\mathrm{X}(t_m), \mathrm{X}(t_{m+1}))|\mathrm{X}(t_m) = \mathrm{x}(t_m)\}\\ = \sum_{m=1}^{M-1} U(\mathrm{x}(t_m), \mathrm{x}(t_{m+1}))\end{aligned} \tag{4}$$

for suitable functions $U(x(t_m), x(t_{m+1}))$ chosen in correspondence with the estimated parameter θ. The choice of the statistics U is discussed in Snijders (2001) and Snijders et al. (2007). The latter publication also specifies the estimating equations for the case of more than one independent variable, which are slightly more involved.

To solve the estimating Eq. (4), in the absence of ways to calculate analytically the expected values, stochastic approximation methods are used. Variants of the Robbins–Monro (1951) algorithm (see, e.g., Chen 2002, for a more up to date treatment) have been used with good success. This is a stochastic iteration method which produces a sequence of estimates $\theta^{(N)}$ which is intended to converge to the solution of (4), and which works here as follows. For a given provisional estimate $\theta^{(N)}$, the model is simulated so that for each $m = 1, \cdots, M-1$, a random draw is obtained from the conditional distribution of $\mathrm{X}(t_{m+1})$ given that $\mathrm{X}(t_m) = \mathrm{x}(t_m)$. This simulated network is denoted $\mathrm{X}^{(N)}(t_{m+1})$. Denote $U^{(N)} = \sum_{m=1}^{M-1} U\left(\mathrm{x}(t_m), \mathrm{X}^{(N)}(t_{m+1})\right)$, and let u^{obs} be the right-hand side of (4). Then the iteration step in the Robbins–Monro algorithm for obtaining the method of moments estimate is given by

$$\theta^{(N+1)} = \theta^{(N)} - a_N D^{-1}\left(U^{(N)} - u^{obs}\right), \tag{5}$$

where D is a suitable matrix and a_N sequence of positive constants tending to 0. This equation is reminiscent of the iteration step in the Newton-Raphson algorithm, but in this case the function for which the root is sought is not directly computable, and instead we simulate random variables having this function as their expected value. Tuning details of the algorithm, including the choices of D and a_N, are given in Snijders (2001). The Bayesian estimators for these models presented in Koskinen and Snijders (2007) and the Maximum Likelihood estimators of Snijders et al. (2010a) are also implemented in *RSiena.* Since Maximum Likelihood estimates can also be defined by an equation of type (4), where now U is the score function (and therefore also depends on the parameter θ), also for this purpose the Robbins–Monro algorithm is used.

A general issue for Monte-Carlo based estimation is to assess the convergence of a given run of the estimation algorithm. The output resulting from the Method of Moments as well as Maximum Likelihood estimation algorithms contains simple indications for convergence, the so-called *t-ratios for convergence*, which indicate the extent to which the estimates found indeed satisfy approximately the Eq. (4), based on independent simulations with the value of θ resulting from the estimation algorithm. A more thorough convergence indicator is the *overall maximum convergence ratio*, defined as the maximum *t*-ratios for convergence for any linear combination of the statistics *U*. For relatively simple models, it is quite usual that the first run of the algorithm produces good estimates. For more complex models or data sets it may be necessary to iterate the algorithm, using the estimates obtained as starting values for the next run of the algorithm.

Elements of the Package

The *RSiena* package operates as all R packages by a collection of functions, and the user can mix the use of *RSiena* with using all other functions in R and its contributed packages. Also in line with the R environment, the package is totally object-oriented: data sets, model specifications, estimation results, etc., all are defined as objects on which the user can operate, and about which information can be requested.

Without going into the specifics of the model, it nevertheless may be helpful to indicated briefly the main types of functions that are available.

1. Functions to specify data objects for a specific use (as covariates, dependent variables, etc.) in the model.
2. Functions to specify the model. These create the "**sienaEffects"** objects containing the model specification, and further modify such objects.
3. Functions for estimating parameters. The main workhorse here is called **siena07** – the name was given for historical reasons, because in the original *Siena* version 1 suite this was meant to be the seventh in a sequence of executable programs. Function **siena07** can be used for estimation according to the Method of Moments as well as Maximum Likelihood estimation. It can also be used for simulating the model without parameter estimation. In combination with the estimation it is also possible to test a hypothesized value for some of the parameters without estimating them by so-called *score-type tests* (Schweinberger 2012) for the Method of Moments, or by regular score tests for Maximum Likelihood estimation.

   Further there are a function **sienaBayes** for Bayesian estimation, and a function **siena08** for the meta-analytic combination of the evidence produced by estimating the same model for a number of independent data sets (Snijders and Baerveldt 2003).
4. Functions for assessing the fit of the model. The main functions of this kind currently are **sienaTimeTest** for testing time homogeneity across multiple waves (Lospinoso et al. 2011), and **sienaGOF** ("goodness of fit") for assessing the adequacy of the model in reproducing a number of features of the data (Lospinoso 2012).
5. A variety of functions for summarizing results obtained.

Key Applications

The *Siena* method as implemented in the *RSiena* package has been applied in a variety of studies in sociology, psychology, political science, and other disciplines. Two special issues of the journal *Social Networks* on network dynamics, published in January 2010 and July 2012, contain a couple of examples. Many applications are listed at the Siena website, http://www.stats.ox.ac.uk/siena/. The following is a very small and somewhat arbitrary selection.

Applications to dynamics of one social network (i.e., without the inclusion of dependent behavior variables), started with van de Bunt et al. (1999), a study of friendship in a group of freshman students. This was the first publication on network dynamics that found statistical evidence for transitivity and for homophily (on gender, age, and smoking behavior) using a method that allows each tested effect to be controlled for all other effects – this being a basic purpose of the *Siena* method. Another example study on friendship development is Selfhout-Van Zalk et al. (2010), concentrating on effects of personality characteristics (the “Big Five”), and finding evidence for homophily with respect to agreeableness, extraversion, and openness to experience; and further concluding (less surprising) that individuals high on extraversion tended to select more friends and individuals high on agreeableness tended to be selected more as friends.

A large group of applications is about *peer effects or social influence*, i.e., the question whether individuals are being influenced in their behavior, performance, or attitudes by those to whom they have network ties. It has long been debated whether the similarity between friends with respect to smoking behavior is a consequence of homophilous selection of friends or of social influence. Mercken et al. (2009) applied *Siena* to a data set of 7704 adolescents (aged 12–15 years) in seventy schools from six European countries (Denmark, Finland, the Netherlands, Portugal, the UK, and Spain). They found evidence for homophilous selection in all countries, and for peer influence with respect to smoking only in Finland and the Netherlands. Obesity is another health-related variable for which the question of peer influence, and how to assess it, has recently received attention in scholarly journals. De la Haye et al. (2011) found, in a data set of two cohorts in the initial two years in high school, that similarities between friends with respect to their body mass index (BMI) were due mainly to processes of friend selection, and not to peer influence. Since the extent of peer influence may well depend on age, family background, cultural and contextual aspects, etc., it is quite plausible that peer influence may differ between countries and social settings. One study can therefore not give a definitive answer about questions of peer selection with respect to variables such as smoking or obesity, and further research is necessary and ongoing.

An example application in political science is Berardo and Scholz (2010), studying governance processes between organizations in 10 US estuaries, and how partner selection for collaboration depended on general trust in the institutional environment as expressed by representants of the organizations. They found that partner selection is not directly dependent on trust, but trust is influenced by the trust expressed by collaboration partners.

The method can also be used for diffusion of innovations (event history analysis, survival analysis) coevolving with a social network, as elaborated by Greenan (2015).

The *Siena* method has also been applied to network data collected by other methods than self-report surveys. An example is Lewis et al. (2012), a study of Facebook friendships and cultural tastes which concluded that friendship formation is influenced by similarity in taste for music and movies, but not for books; and that there is little influence for diffusion of tastes through Facebook ties, with the exception of a taste for classical/jazz music.

Future Directions

The stochastic actor-oriented methodology and the *RSiena* package are areas of active ongoing research and development, as can be seen from the website. Current work includes the development of models for continuous dependent behavior variables (Niezink), models with unobserved heterogeneity between actors (Koskinen), random effect models for multiple groups (Koskinen and Snijders), and the *settings model* (Preciado and Snijders) which is meant to make the actor-oriented approach applicable also to larger networks (with a few hundred to a few thousand nodes); such data sets are not well-suited for the current software because the basic model (like other models for network dynamics) makes assumptions of homogeneity, and of accessibility of actors to each other, that are less plausible for such large networks because the time taken by the computer simulations becomes prohibitive.

Cross-References

- ▶ Actor-Based Models for Longitudinal Networks
- ▶ Analysis and Visualization of Dynamic Networks
- ▶ Exponential Random Graph Models
- ▶ Human Behavior and Social Networks
- ▶ R Packages for Social Network Analysis
- ▶ Social Influence Analysis
- ▶ Statistical Research in Networks: Looking Forward
- ▶ Temporal Networks
- ▶ Theory of Statistics: Basics and Fundamentals

Acknowledgements Work on the *Siena* method and software has taken place at the University of Groningen (Department of Sociology) and the University of Oxford (Department of Statistics; Department of Politics and International Relations; Nuffield College). NWO (the Netherlands Organisation for Scientific Research) has supported development of this methodology and software through the project *Statistical Methods for the Joint Development of Individual Behavior and Peer Networks* (project number 575-28-012, researcher Mark Huisman), the integrated research program *The Dynamics of Networks and Behavior* (project number 401-01-550, methodological researchers Christian Steglich and Michael Schweinberger), and the ECRP-Eurocores program *Models for the Evolution of Networks and Behavior* (project number 461-05-690, methodological researcher Christian Steglich). The US National Institutes of Health have provided funding for methodological developments (Johan Koskinen) and software development (Ruth Ripley and Krists Boitmanis) as part of the project *Adolescent Peer Social Network Dynamics and Problem Behavior* (grant number 1R01HD052887-01A2).

References

Berardo R, Scholz JT (2010) Self-organizing policy networks: risk, partner selection and cooperation in estuaries. Am J Polit Sci 54:632–649

Chen HF (2002) Stochastic approximation and its applications. Kluwer Academic, Dordrecht

de la Haye K, Robins G, Mohr P, Wilson C (2011) Homophily and contagion as explanations for weight similarities among adolescent friends. J Adolesc Health 49:421–427

Greenan C (2015) Diffusion of innovations in dynamic networks. J R Stat Soc Ser A 178:147–166

Holland PW, Leinhardt S (1977) A dynamic model for social networks. J Math Sociol 5:5–20

Koskinen JH, Snijders TAB (2007) Bayesian inference for dynamic social network data. Journal of Statistical Planning and Inference 13:3930–3938

Leenders R (1995) Models for network dynamics: a Markovian framework. J Math Sociol 20:1–21

Lewis K, Gonzalez M, Kaufman J (2012) Social selection and peer influence in an online social network. Proceedings of the National Academy of Sciences, USA 109(1):68–72

Lospinoso JA (2012) Statistical models for social network Dynamics. DPhil Thesis, University of Oxford, Oxford

Lospinoso JA, Schweinberger M, Snijders TAB, Ripley RM (2011) Assessing and accounting for time heterogeneity in stochastic actor oriented models. Advances in Data Analysis and Computation 5:147–176

Mercken L, Snijders TAB, Steglich CEG, de Vries H (2009) Dynamics of adolescent friendship networks and smoking behavior: social network analyses in six European countries. Soc Sci Med 69:1506–1514

Norris JR (1997) Markov Chains. Cambridge University Press, Cambridge

Ripley RM, Snijders TAB, Preciado P (2017) Manual for Siena version 4.0. Technical report, Department of Statistics, University of Oxford, Oxford/Nuffield College. http://www.stats.ox.ac.uk/siena/. Accessed 28 May 2013

Robbins H, Monro S (1951) A stochastic approximation method. Ann Math Stat 22(3):400–407

Schweinberger M (2012) Statistical modeling of network panel data: goodness-of-fit. British Journal of Statistical and Mathematical Psychology 65:263–281

Selfhout-Van Zalk MHW, Burk W, Branje SJT, Denissen J, van Aken M, Meeus WHJ (2010) Emerging late adolescent friendship networks and big five personality traits: a social network approach. J Pers 78:509–538

Snijders TAB (1995) Methods for longitudinal social network data: Review and Markov process models. In: Tiit E, Kollo T, Niemi H (eds) New Trends in Probability and Statistics Vol. 3: Multivariate Statistics and Matrices in Statistics. Proceedings of the 5th Tartu Conference, TEV Vilnius, Lithuania, pp 211–227

Snijders TAB (1996) Stochastic actor-oriented dynamic network analysis. J Math Sociol 21:149–172

Snijders TAB (2001) The statistical evaluation of social network dynamics. In: Sobel ME, Becker MP (eds) Sociological methodology – 2001, vol 31. Basil Blackwell, Boston and London, pp 361–395

Snijders TAB (2005) Chap. 11, Models for longitudinal network data. In: Carrington P, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, New York, pp 215–247

Snijders TAB (2009) Longitudinal methods of network analysis. In: Meyers B (ed) Encyclopedia of complexity and system science. Springer, Berlin, pp 5998–6013

Snijders TAB, Baerveldt C (2003) A multilevel network study of the effects of delinquent behavior on friendship evolution. J Math Sociol 27:123–151

Snijders TAB, van Duijn MAJ (1997) Simulation for statistical inference in dynamic network models. In:

Conte R, Hegselmann R, Terna P (eds) Simulating social phenomena. Springer, Berlin, pp 493–512

Snijders TAB, Steglich CEG, Schweinberger M (2007) Modeling the co-evolution of networks and behavior. In: van Montfort K, Oud H, Satorra A (eds) Longitudinal models in the behavioral and related sciences. Lawrence Erlbaum, Mahwah, pp 41–71

Snijders TAB, Koskinen JH, Schweinberger M (2010a) Maximum likelihood estimation for social network dynamics. Ann Appl Stat 4:567–588

Snijders TAB, van de Bunt GG, Steglich C (2010b) Introduction to actor-based models for network dynamics. Soc Networks 32:44–60

Snijders TAB, Lomi A, Torl'o V (2013) A model for the multiplex dynamics of two-mode and one-mode networks, with an application to employment preference, friendship, and advice. Soc Networks 35:265–276

Steglich CEG, Snijders TAB, Pearson MA (2010) Dynamic networks and behavior: separating selection from influence. Sociol Methodol 40:329–393

van de Bunt GG (1999) Friends by choice; an actor-oriented statistical network model for friendship networks through time. Thesis Publishers, Amsterdam

van de Bunt GG, van Duijn MAJ, Snijders TAB (1999) Friendship networks through time: an actor-oriented statistical network model. Computational and Mathematical Organization Theory 5:167–192

Wasserman S (1979) A stochastic model for directed graphs with transition rates determined by reciprocity. In: Schuessler KF (ed) Sociological methodology 1980. Jossey-Bass, San Francisco

Wasserman S (1980) Analyzing social networks as stochastic processes. J Am Stat Assoc 75:280–294

Zeggelink EP (1994) Dynamics of structure: an individual oriented approach. Soc Networks 16:295–333

Recommended Reading

In addition to the help pages that are available as for all R packages, there is an extensive manual (Ripley et al. 2017) and a tutorial paper (Snijders et al. 2010b). A textbook about the Siena method and an edited volume with example applications are in preparation. The website http://www.stats.ox.ac.uk/siena/ is actively maintained and contains references to the basic methodology, references to applications, R scripts, example data sets, workshop announcements, and more

For those who wish to read more about the mathematical and methodological background, a recommended sequence of readings could be Snijders (1996) as an introduction to the idea of stochastic actor-oriented models, Snijders (2001) or Snijders (2005) for the basic definition of the model for one dependent network defined as a changing digraph, and Steglich et al. (2010) for models for the dynamics of networks and behavior, which might be followed by Snijders et al. (2010a) for Maximum Likelihood estimation or Snijders et al. (2013) for models with multiple dependent networks.

Signatures

► Telecommunications Fraud Detection, Using Social Networks for

Signed Graphs

Krzysztof Stefaniak and Mikołaj Morzy
Institute of Computing Science, Poznan University of Technology, Poznań, Poland

Synonyms

Biased graph; Gain graph; Signed network

Glossary

Arc	An ordered pair of nodes adjacent in the graph
Cycle	A loop of at least three nodes in which the first node and the last node are the same
Digraph	A graph in which all relations are directed
Dyad	A pair of nodes and the incidence relation between them
Edge	A pair of nodes adjacent in the graph
Graph	A data structure consisting of a set of nodes, and a set of pairs of nodes, called edges or arcs
Loop	A walk in the graph in which all edges are distinct
Path	A walk in the graph in which all edges and nodes are distinct
Sociomatrix	Representation of the incidence relation as a two-dimensional matrix in which rows and columns represent nodes, and cells represent relation values
Triad	A triple of nodes and all incidence relations between them
Valence	Semantic orientation of an edge in a signed graph

Definition

Given a set of nodes $N = \{n_1,..., n_m\}$ and a set of edges $E = \{e_1,..., e_n\}$, where each edge is a pair of nodes, $e_k = (n_i, n_j)$. A *signed graph* is a triple $G_{\pm} = <N, E, S>$ consisting of a set of nodes N, a set of edges E, and a mapping S which is a function $S : E \rightarrow \{+, -\}$, i.e., the mapping S associates with every edge $e_k \in E$ either a positive valence, typically denoted by (+), or a negative valence, denoted by (−). Positive valence of an edge usually denotes the fact that the relationship modeled by the edge (the type of association between nodes) has some positive quality, such as kindness, friendship, or trust. Likewise, the negative valence represents antagonizing feelings between nodes, such as enmity, dislike, or distrust. Edges can be lacking directional information, in such case the relationship is considered symmetrical. If edges are directional, such graph is called a *signed digraph*. Some formulations also allow for the existence of multiedges, as well as halfedges (which are edges with only one endpoint) and loose edges (which are edges without any endpoints), but halfedges and loose edges are not signed. A *complete signed graph* is a signed graph in which each pair of nodes belongs to the set of edges.

Introduction

Signed graphs have been used for a long time in social network analysis to model opposite relationships. In a signed graph, each edge is assigned either a positive or negative sign, referred to as *valence*. For instance, in a social network representing acquaintance between people, positive edges can represent friendship, while negative edges can represent animosity. If the signed graph is modeling diplomatic relations between countries, a positive edge can represent cooperation and a negative edge can represent some kind of political tension.

In general, edges can be attributed with more values, leading to the so-called *valued graphs*. Signed graphs are a special case of valued graphs in which edges are allowed only two opposing values, and the aggregation of values along loops is performed by multiplication rather than by addition. It should be stressed that a negative edge between nodes is different from the lack of an edge between nodes. While the lack of an edge suggests the lack of interaction between nodes, a negative edge is a clear mark of an inimical relationship. Another frequent misunderstanding is that signed graphs are simply graphs with edges weighted by either +1 or −1 numerical values. Such a graph would be a regular graph with a constraint imposed on the set of possible values for edges. It would be very different from a signed graph because in regular valued graphs edge values are added, and not multiplied. Another example where similar graphs are being used is the knot theory, where color is used to mark edges. Again, methods and algorithms are very different because the color of an edge does not convey the intrinsic opposition of positive and negative valence.

At the core of a signed graph lies a *signed relation*. It is the relation that can easily convey both positive and negative sentiments. Examples of signed relations include esteem/disrespect, like/dislike, praise/blame, and influence/negative influence, as presented by Sampson (1968). It is possible to treat these opposing sentiments as two independent relations, but in reality the two sentiments are clearly associated, because one sentiment is usually the antonym of the other. Some graph theorists also require that the relation in question should satisfy the *principle of antithetical duality*, which is to say that the dual (the antonym) of a signed graph simply changes the signs of the loops. Computing the dual of the changed graph brings back the original graph. Without this property a graph cannot be used in the light of signed graph theory and balance theory. Therefore, traditional social networks, where relations usually represent some kind of social interaction, e.g., communication or interaction, cannot be modeled as signed graphs and cannot be studied using the balance theory.

Key Points

Signed graphs have several features that make them a useful tool for sociological and psychological research, but signed graphs can be also used outside of social sciences, e.g., in the field of physics or chemistry. To understand the benefit and utility of the signed graph model, we must first observe the key points that differentiate signed graphs from more general valued graphs. One of the most important methods of network analysis developed within the domain of signed graphs is the *triad analysis* that aims at capturing the dynamics of relations between very small groups of nodes. Triad analysis is described in detail in section "Triad Analysis". Triad analysis has been refined and extended in the field of social psychology under the moniker of *P-O-X triples analysis*, which we scrutinize in section "*P-O-X* Triples". Probably, the most famous concept originating from the signed graph theory is the idea of *structural balance*. In section "Structural Balance" we define the notion of structural balance and we introduce the fundamental Harary's theorem, along with its proof. We discuss the implications of a signed graph being balanced and we show a simple method for testing whether a graph is balanced. We also present several measures for the amount of imbalance in section "Balance Index". The last concept pertaining to signed graphs is the notion of *frustration*, discussed in section "Frustration".

Historical Background

Signed graphs have been studied since 1950s. They were first introduced by Harary (Cartwright and Harary 1956; Harary 1953) in a structural balance theory, which was the generalization of the Heider's theory (Heider 1946) from sociology. Heider's theory of social balance focuses on the balance of sentiments between people. According to this theory, in small subgroups of people certain relationships tend to be more socially plausible. For example, in a group of two individuals (a dyad) there is only one relationship, positive or negative, but when we look at a complete graph of relationships between three people (a triad), we can distinguish four different principles: "a friend of my friend is my friend," "an enemy of my enemy is my friend," "a friend of my friend is my enemy," and "an enemy of my enemy is my enemy," with the two latter ones clearly causing cognitive dissonance and thus making the whole graph unbalanced.

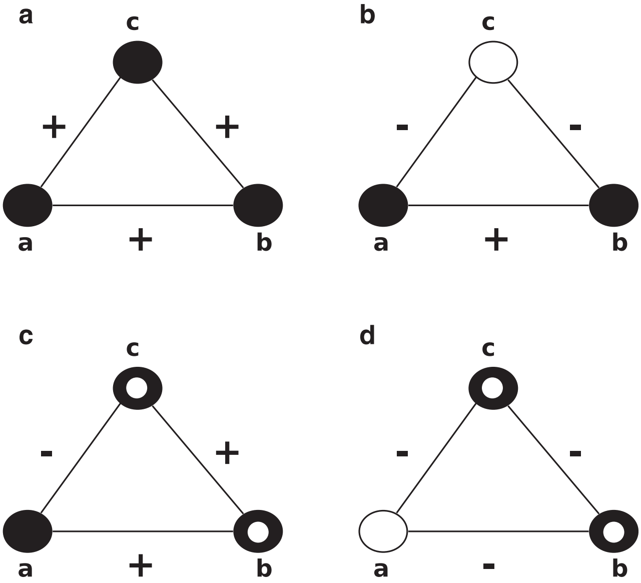
There is a tendency to avoid unbalanced structures and increase balance of the graph even if it makes shifts of relationship signs necessary. Sign changes include enemies becoming friends (positive edge) or friends becoming enemies (negative edge). Davis questioned the significance of the last principle (Davis 1967), arguing that it is rather difficult to make any three mutual enemies friendly towards each other. Consequently, he proposed *weakly balanced graphs*, which rule out only the structure with one negative edge reflecting the principle "a friend of my friend is my enemy". In consonance with the original structural balance theory, a balanced graph can be divided into two groups (bipartite graph) (Harary 1953), but in case of weakly balanced graphs, it is possible to have multiple clusters with positive edges inside the group and negative edges between the subgroups.

Signed Graphs

Triad Analysis

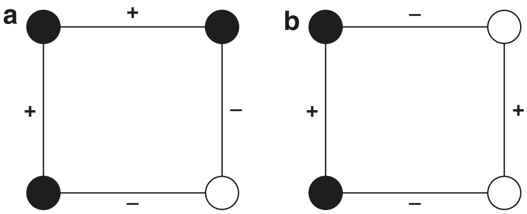
Most of analysis of signed graphs is concerned with the analysis of dyads and triads. Each dyad can be in one of three states: a positive relationship, a negative relationship, and no relationship between the nodes in the dyad. For complete signed graphs, each dyad is either positive or negative. For triads (for the sake of brevity we are considering only complete triads) each triad can be in one of four states depending on the number of negative relationships between the nodes in the triad (zero, one, two, or three).

Consider possible triad configurations shown in Fig. 1. Nodes that share a positive relationship are marked with the same color. The configuration (a) is the simplest and most obvious, all nodes have positive feelings about all other nodes in the triad, so there is no room for a conflict. Similarly, the configuration (b) is stable, since nodes



Signed Graphs, Fig. 1 Possible signed triad configurations

a and *b* like each other and share the same negative feeling towards node *c*. This configuration is stable in the sense that it is coherent and no node has to choose between any other node. Now compare previous configurations with the configuration presented in (c). This configuration is unstable, because node *b* is torn in its allegiance to nodes *a* and *c*, who dislike each other. In order to maintain social ties node *b* has to choose between nodes *a* and *c*, and the remaining relationship will probably become broken. Finally, the configuration presented in (d) is also considered unstable. What is characteristic about this configuration is that the "*enemy of my enemy is my friend*" rule of thumb does not apply here. This fundamental difference in triad configurations can be very easily expressed by the number of negative signs along the loop. Triads with an even number of negative edges tend to be stable, whereas triads with an odd number of negative edges tend to be unstable and eventually break down. This observation can be extended to loops of the length greater than 3, as depicted in Fig. 2. Experimental research suggests that this type of stability is quite often encountered in real networks because unstable configurations appear far less often in real networks than stable configurations.

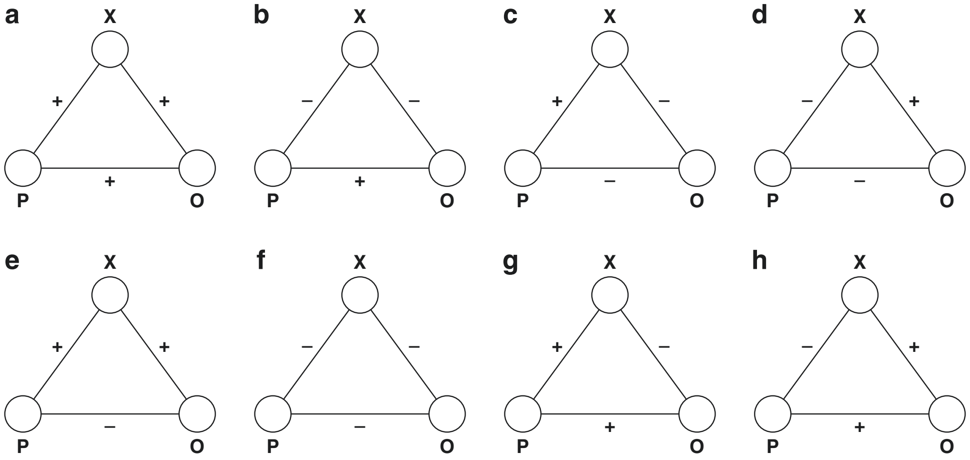


Signed Graphs, Fig. 2 Stable configurations of loops of the length four

P-O-X Triples

Another usage for signed graphs comes from the field of social psychology, where signed graphs are used to model the cognition of social relationships. A well-known example of this line of research is the analysis of the so-called *P-O-X* triples. According to this model, *P* denotes a person, *O* denotes another individual (the other), and *X* denotes an entity or object. The task is to find how the positive or negative attitude of the primary person *P* towards the object *X* is consistent with the attitude of the other *O*. This analysis is fairly similar to the discussion of the basic triad model presented above, but with some slight differences which we will underline next.

To make our discussion as general as possible, we will assume that the relationships depicted in



Signed Graphs, Fig. 3 *P-O-X* triples

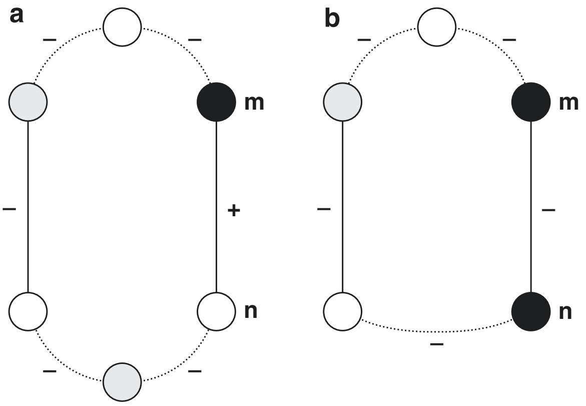
Fig. 3 represent the attitudes of liking (positive valence denoted with the (+) sign) and disliking (negative valence denoted with the (−) sign). In accordance with the tradition which is well established in sociology, we will use the term *actor* to refer to individual nodes. Both *P* and *O* are allowed to express their attitudes towards object *X*, furthermore they can express their sentiment towards each other. All relationships under discussion are assumed to be symmetrical. Scenarios (a) through (d) depict balanced situations, where either both actors like each other and agree in their assessment of the object *X* (scenarios (a) and (b)) or both actors disagree in their assessment of the object *X*, but this difference in opinions can be explained by their mutual dislike (scenarios (c) and (d)). Now compare these to scenarios (e) through (h) in the Fig. 3. Scenarios (e) and (f) represent the situation where the actors agree in their sentiment towards the object *X* despite having negative feelings about each other. Even more awkward situation is depicted in scenarios (g) and (h), where actors *P* and *O* apparently like each other but cannot reach a consensus about the attitudes towards object *X*. Such disagreements, as shown by sociological research, can quickly undermine the positive relationship between the actors.

Structural Balance

Most of the analysis of signed graphs depends on the notion of loops. In particular, one is often interested in the sign of particular loops in the signed graph. We will use the term *loop* to describe any closed walk in the graph in which all nodes (except the first and the last) are distinct. The *sign of a loop* is the product of all edges contained in the loop. Since only negative edges change the sign of a loop, and two negative edges cancel each out, an even number of negative edges on the loop will produce a positive loop, and an odd number of negative edges will produce a negative loop. The idea of a signed loop can be further extended to semicycles. A semicycle is a closed sequence of nodes in which every pair of consecutive nodes forming a semicycle is adjacent (in other words, a semicycle is a cycle in which arcs can point in any direction). The *sign of a semicycle* is defined also as a product of signs of arcs.

A signed loop is *stable* if it contains an even number of negative edges. A graph is *stable*, or, in other words, is said to show *structural balance*, if all loops in the graph are stable. Harary (Harary 1953) presents an important finding pertaining to signed graphs:

Harary's Theorem. *A balanced graph can be divided into connected groups of nodes such that all connections between members of the same*



Signed Graphs, Fig. 4 Clustering of a balanced graph

group are positive and all connections between members of different groups are negative.

According to Harary, each group can contain an arbitrary number of nodes and there can be many groups of nodes. A graph is *clusterable* if its nodes can be divided into separate groups such that all positive relationships are happening only within the group and all relationships between groups are only negative. Harary's theorem states that all balanced graphs are clusterable. The opposite does not work, not all clusterable graphs are balanced. The term *structural balance* is used by sociologists and psychologists to refer to groups that are coherent and lack inner tensions between members.

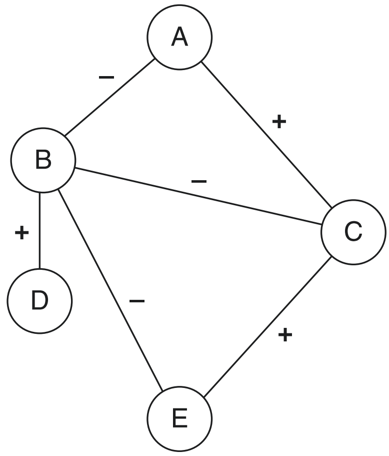
Since structural balance is the key concept in many applications of the signed graph theory, we will provide a simple constructive proof of the Harary's theorem. For the sake of simplicity we will consider a graph with a single connected component.

Proof. We select a random node in the graph and we color this node with white color. Then, we iterate over all remaining uncolored nodes in the graph and we color them according to two simple rules:

If a node n is connected by a positive edge to an already colored node m, node n receives the same color as m.

If a node n is connected by a negative edge to an already colored node m, node n receives the opposite color to m.

If, at any moment, we arrive at a node n that has already been colored, but according to the above rules it should be colored with an opposite color (i.e., a conflict arises), then the entire graph is not balanced. The reasoning behind this simple procedure is the following. While iterating over the nodes in the graph, if we stumble upon a node n that has already been colored, this means that there must be an alternative path leading to the node n from the starting point. According to the Harary's theorem, for the graph to be balanced, each loop in the graph has to have an even number of negative edges. Let us examine in detail the situation in which a conflict in coloring arises (see Fig. 4). There are only two such situations: (a) either we want to assign the node n the same color as the node m to which n is connected (i.e., the edge between m and n is positive) but n is already colored with an opposite color, or (b) we want to assign the node n the opposite color as the node m to which n is connected (i.e., the edge between m and n is negative) but n is already colored with the same color as m. In the first situation, the fact that n is colored differently from m means that there is a loop between m and n with an odd number of negative edges (because the color changes between m and n an odd number of times) and thus m and n should be placed in opposite groups. However, the existence of a direct positive edge between m and n contradicts this; thus, the graph is not balanced. A similar reasoning applies to the second situation. If n has the same color as m, then there is a loop in the network between m and n with an even number of negative edges (this is why color alternates an even number of times on the loop). But m and n cannot be placed in the same group because of



Signed Graphs, Fig. 5 Checking for balance in the graph

the direct negative edge between them. Again, the contradiction proves that the graph cannot be balanced. The generalization of this proof to the graph consisting of several connected components is trivial since it requires simply to apply the above procedure to all components sequentially.

An interesting question arises of how to check efficiently if a given graph is balanced. Since a single loop with a negative sign makes the entire graph unbalanced, one needs to consider loops of length $l = 2,..., n - 1$ sequentially looking for a loop with a negative sign. In order to find the sign of a loop of a given length l, it is sufficient to check the main diagonal of the graph's sociomatrix raised to the power of l. If M^l is the sociomatrix of the signed graph $G_{\pm}$, then the main diagonal of M represents all loops of length l starting and ending at a given node. Consider the simple signed graph $G_{\pm}$ depicted in Fig. 5.

The sociomatrix M for the graph $G_{\pm}$ is:

M

	A	B	C	D	E
A	0	-1	1	0	0
B	-1	0	1	1	-1
C	1	1	0	0	1
D	0	1	0	0	0
E	0	-1	1	0	0

Below we show the consecutive powers of the sociomatrix M. Please observe the values on the main diagonal of the matrix as they contain the signs of all loops of the length $l = 2, 3, 4$ starting and ending in respective nodes.

M²

	A	B	C	D	E
A	2	-1	1	-1	2
B	-1	4	-2	0	-1
C	1	-2	3	-1	1
D	-1	0	-1	1	-1
E	2	-1	1	-1	2

M³

	A	B	C	D	E
A	2	-6	5	-1	2
B	-6	4	-6	4	-6
C	5	-6	4	-2	5
D	-1	4	-2	0	-1
E	2	-6	5	-1	2

M⁴

	A	B	C	D	E
A	11	-10	10	-6	11
B	-10	22	-16	4	-10
C	10	-16	16	-6	10
D	-6	4	-6	4	-6
E	11	-10	10	-6	11

As we can see, the main diagonal of each power of the sociomatrix M is nonnegative, thus we may conclude that the graph $G_{\pm}$ is balanced. The method presented here applies only to undirected signed graphs. The extension of the method to signed digraphs is not trivial. The interested reader will find the detailed description of the method in (Harary 1956, pp. 352–355).

Balance Index

One may ask if it is possible to somehow quantify the amount of imbalance in the graph. In other words, one may wonder how many changes would have to occur in order to make the graph balanced. Several indexes have been proposed that aim at measuring the degree of imbalance. The term *cycle index for balance* has been used to collectively refer to such indexes. The general idea is to find the number of cycles in the graph that have a negative sign (i.e., the number of cycles that violate the balance condition) and to compare this number to the total number of cycles present in the graph. The simplest index (Cartwright and Gleason 1966) divides the number of positive cycles by the total number of cycles in the graph. More elaborate indexing schemes propose to weight each cycle (positive or negative) by the length of the cycle (Harary 1959; Norman and Roberts 1972). Several other measures of structural balance are discussed in (Taylor 1970).

Frustration

Signed graphs can be extended by adding positive and negative values to nodes (in addition to edges). Let each node n have a state $S(n)$, where the state of a node can be either positive (+1) or negative (−1). An edge e is *satisfied* if and only if:

- Edge e is positive and both its endpoints are in the same state
- Edge e is negative and both its endpoints are in opposite states

If an edge is not satisfied, it is called *frustrated*. The minimum number of frustrated edges in any state of the graph is called the *line index for balance* (Harary 1960) or the *frustration index*. Unfortunately, finding the frustration index is computationally hard. Computing the frustration index of the signed graph can be reduced to the maximum cut problem of the graph, which is known to be NP-hard.

Key Applications

The structural balance theory and signed graphs were initially invented to solve the subgrouping problem in social psychology, but they have found applications in other areas as well. Signed graphs are adequate to model opposite relations between objects. The first feasible entities are obviously humans with our complex psychological, social, and anthropological relations. Social sciences are the major application area for signed graphs. They are used to describe dynamics of human sentiments, with the special emphasis on social networks (Tang et al. 2015; Wu et al. 2016; Yang et al. 2012; Yu et al. 2016). In addition to simply predicting and explaining friendship and animosity changes in groups of people (Anchuri and Magdon-Ismail 2012; Antal et al. 2006; Yang et al. 2007a), signed graphs proved to be useful in anthropology and politics (Leskovec et al. 2010a; Tang et al. 2016), structural balance was used in analysis of enmity in tribal wars, political conflicts (Hage 1979), Wikipedia edit wars (Maniu et al. 2011), and international relations (Harary 1961; Moore 1979).

The utility of signed graphs is not limited to social sciences, it is possible to model natural phenomena as signed graphs too. Most notable adaptations can be found in ecology (Ilany et al. 2013), chemistry (Trinajstic 1983), and physics. Frustration index presented in section "Frustration" is used in physics for modeling the ferromagnetism in statistical mechanics using Ising models. Each vertex, representing a molecule, can have a spin-up (+1) or a spin-down (−1) orientation. For each state of molecule spins, there is a particular number of frustrated edges. The total energy of the entire system is proportional to the number of frustrated edges; thus, the preferred state of the lowest energy is the state with the lowest frustration index.

Future Directions

Current development of web-based social networks revealed new possibilities for signed graphs in social network analysis. Recent research in this area includes modeling trust and distrust (Guha et al. 2004), finding friends and foes (Brzozowski et al. 2008), community structure mining (Yang et al. 2007b), or link prediction (Kunegis et al. 2009; Leskovec et al. 2010a) in large signed networks. Signed graphs are being used in the area of sentiment analysis and opinion mining (Proskurnikov et al. 2016; Shi et al. 2016), and the availability of huge corpora of text data containing expressions of opinions provide a fertile research ground. We expect to see more data-driven modeling and analysis of social dynamics based on the data harvested from the social Web. Signed graphs can also be used to improve the quality of network data visualization (Kunegis et al. 2010).

Recommended Reading

Signed graphs are covered thoroughly in the literature, both from the theoretic and application angles. A good starting point is a general book on graph theory, such as excellent text by Harary (1969) or Bondy and Murty (2002). An approach focusing more on the social aspects of networks is presented in seminal textbooks by Wasserman and Faust (1994) and Newman (2010). For a recent survey on various approaches to studying signed networks, the reader should consult (Doreian 2017).

A very detailed summary of social balance research covering over 200 different papers is presented by Taylor (1970). Readers interested in a more anthropological approach to the study of social structure and balance should consult (Hage 1979; Hage and Harary 1983). Our discussion on balance in social structures can be further extended to the notion of clusterability. Concepts of clusterability, ranked clusterability, and transitive tournaments are discussed at length by Holland and Leinhardt (1971). If the reader desires to investigate mathematical properties of signed graphs, she is advised to follow the works of Zaslavsky (1981, 1982).

Cross-References

- ▶ Social Networks and Politics
- ▶ Structural Holes

References

Anchuri P, Magdon-Ismail M (2012) Communities and balance in signed networks: a spectral approach. In: Advances in social networks analysis and mining (ASONAM), 2012 IEEE/ACM international conference on IEEE, pp 235–242

Antal T, Krapivsky P, Redner S (2006) Social balance on networks: the dynamics of friendship and enmity. Physica D 224(1–2):130–136. Dynamics on Complex Networks and Applications

Bondy AJ, Murty USR (2002) Graph theory with applications. Wiley, New York

Brzozowski MJ, Hogg T, Szabo G (2008) Friends and foes: ideological social networking. In: Proceedings of the twenty-sixth annual SIGCHI conference on human factors in computing systems, CHI ‘08, ACM, New York, pp 817–820

Cartwright D, Gleason T (1966) The number of paths and cycles in a digraph. Psychometrika 31(2): 179–199

Cartwright D, Harary F (1956) Structural balance: a generalization of Heider’s theory. Psychol Rev 63(5):277–293

Davis JA (1967) Clustering and structural balance in graphs. Hum Relat 20(2):181–187

Doreian P (2017) Reflections on studying signed networks. J Interdiscip Methodol Issues Sci 2:3.1–3.16

Guha R, Kumar R, Raghavan P, Tomkins A (2004) Propagation of trust and distrust. In: Proceedings of the 13th international conference on World Wide Web, WWW’04, ACM, New York, pp 403–412

Hage P (1979) Graph theory as a structural model in cultural anthropology. Annu Rev Anthropol 8:115–136

Hage P, Harary F (1983) Structural models in anthropology. Cambridge University Press, Cambridge, MA

Harary F (1953) On the notion of balance of a signed graph. Mich Math J 2(2):143–146

Harary F (1956) Structural models: an introduction to the theory of directed graphs. Wiley, New York

Harary F (1959) On the measurement of structural balance. Behav Sci 4(4):316–323

Harary F (1960) A matrix criterion for structural balance. Nav Res Logist Q 7(2):195–199

Harary F (1961) A structural analysis of the situation in the middle east in 1956. J Confl Resolut 5(2):167–178

Harary F (1969) Graph theory. Addison-Wesley, Reading

Heider F (1946) Attitudes and cognitive organization. J Psychol 21(2):107–112

Holland PW, Leinhardt S (1971) Transitivity in structural models of small groups. Small Group Res 2(2):107–124

Ilany A, Barocas A, Koren L, Kam M, Geffen E (2013) Structural balance in the social networks of a wild mammal. Anim Behav 85(6):1397–1405

Kunegis J, Lommatzsch A, Bauckhage C (2009) The slashdot zoo: mining a social network with negative edges. In: Proceedings of the 18th international conference on World Wide Web, WWW’09, ACM, New York, pp 741–750

Kunegis J, Schmidt S, Lommatzsch A, Lerner J, De Luca EW, Albayrak S (2010) Spectral analysis of signed graphs for clustering, prediction and visualization. In: Proceedings of the 2010 SIAM international conference on data mining, SIAM, pp 559–570

Leskovec J, Huttenlocher D, Kleinberg J (2010a) Predicting positive and negative links in online social networks. In: Proceedings of the 19th international conference on World Wide Web, WWW’10, ACM, New York, pp 641–650

Leskovec J, Huttenlocher D, Kleinberg J (2010b) Signed networks in social media. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, pp 1361–1370

Maniu S, Cautis B, Abdessalem T (2011) Building a signed network from interactions in wikipedia. In: Databases and social networks, ACM, pp 19–24

Moore M (1979) Structural balance and international relations. Eur J Soc Psychol 9(3):323–326

Newman M (2010) Networks: an introduction. Oxford University Press, New York

Norman R, Roberts F (1972) A derivation of a measure of relative balance for social structures and a characterization of extensive ratio systems. J Math Psychol 9(1):66–91

Proskurnikov AV, Matveev AS, Cao M (2016) Opinion dynamics in social networks with hostile camps: consensus vs. polarization. IEEE Trans Autom Control 61(6):1524–1536

Sampson S (1968) A novitiate in a period of change: an experimental and case study of relationship. PhD thesis, Cornell University

Shi G, Proutiere A, Johansson M, Baras JS, Johansson KH (2016) The evolution of beliefs over signed social networks. Oper Res 64(3):585–604

Tang J, Chang S, Aggarwal C, Liu H (2015) Negative link prediction in social media. In: Proceedings of the eighth ACM international conference on web search and data mining, ACM, pp 87–96

Tang J, Chang Y, Aggarwal C, Liu H (2016) A survey of signed network mining in social media. ACM Comput Surv (CSUR) 49(3):42

Taylor HF (1970) Balance in small groups. Van Nostrand Reinhold Co, New York

Trinajstic N (1983) Chemical graph theory. CRC Press, Boca Raton

Wasserman S, Faust K (1994) Social network analysis: methods and applications (structural analysis in the social sciences). Structural analysis in the social sciences, 1st edn. Cambridge University Press, Cambridge, MA

Wu Z, Aggarwal CC, Sun J (2016) The troll-trust model for ranking in signed networks. In: Proceedings of the ninth ACM international conference on web search and data mining, ACM, pp 447–456

Yang B, Cheung W, Liu J (2007a) Community mining from signed social networks. IEEE Trans Knowl Data Eng 19(10)

Yang B, Cheung W, Liu J (2007b) Community mining from signed social networks. IEEE Trans Knowl Data Eng 19(10):1333–1348

Yang SH, Smola AJ, Long B, Zha H, Chang Y (2012) Friend or frenemy?: predicting signed ties in social networks. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 555–564

Yu T, Bai L, Guo J, Yang Z (2016) Construct a bipartite signed network in youtube. Big data: concepts, methodologies, tools, and applications: concepts, methodologies, tools, and applications, IGI Global, Hershey, PA, pp 370

Zaslavsky T (1981) Characterizations of signed graphs. J Graph Theory 5(4):401–406

Zaslavsky T (1982) Signed graphs. Discret Appl Math 4(1):47–74

Signed Network

▶ Signed Graphs

Signed Social Networks

▶ Influence Propagation in Social Networks with Positive and Negative Relationships

Similarity Metrics on Social Networks

Cuneyt Gurcan Akcora and Elena Ferrari
DISTA, Università degli Studi dell'Insubria,
Varese, Lombardia, Italy

Synonyms

Categorical data similarity; Network similarity; Profile similarity

Glossary

Homophily	Tendency to create friendships with similar people
Profile data	User-uploaded text-based personal information on social networks
Undirected network	Network where relationships are created by mutual consent of the two involved users

Definition

In the last decade, online social networks have gained millions of users who are daily creating terabytes of personal data. With this amount of data, it quickly becomes impractical to analyze all of the network for solving user-specific problems such as link prediction. At the basis of most of these computations, there is the need of computing similarity between social network users. In this entry, we show how similarity can be computed by using a family of metrics which provide fast, local, and efficient solutions. We classify user-generated social network data into network and profile data and discuss metrics for each type of data. We give particular importance to outline the benefits and shortcomings of the considered metrics and how they are used in current research work.

Introduction

In the literature, the term *similarity* has been used in different meanings (e.g., the short distance between two users) to quantize similarity for different application fields. Some work have attempted to define similarity rigorously (Richter 2007; Ha and Haddawy 2003; Lin 1998). Among these, the four principles by Lin (1998) are widely used to implement similarity metrics on social networks. We will explain these four principles with commonality, differences, maximum similarity, and minimum similarity. In *commonality*, shared commonalities (e.g., race, gender, sex of users) increase the similarity of two users. On the other hand, the more *differences* lead to smaller similarity. Regardless of the number of features, two users are said to have the *maximum similarity* when they are identical in every feature. Similarly, regardless of the number of features, two users are the least similar when they are different in every feature. In current studies, the maximum and minimum similarity values are given as 0 and 1, respectively.

A more rigorous set of properties for similarity metrics can be adopted from distance metrics by considering $similarity = 1 - distance$. For any given distance metric, these properties are (1) symmetry, (2) identity, (3) non-negativity, and (4) triangle equality. In the identity property, $distance(a, b) = 0$, when $a = b$. On social networks this property can have different explanations; on the graph structure $distance(a, b) = 0$ is assumed to be true when the two nodes have the same set of friends, whereas if profile information are considered, two users must have the same values for every profile item. In setting a lower bound for distance, the non-negativity property defines $distance(x, y) \geq 0$ for any user pair. The symmetry condition assures that $distance(a, b) = distance(b, a)$, while in the triangle equality $distance(a, c) \leq distance(a, b) + distance(b, c)$.

The absence of some of these properties can be used to classify different formulas. For example, *quasi-metrics* do not provide the symmetry property, whereas *semi-metrics* do not have the triangle property. Traditionally, the symmetry property is not applicable in directed networks, because directions of edges can lead to different similarity values for a pair of users. In this case, two different values are computed; $sim(a, b)$ denotes the similarity value according to a user a, and $sim(b, a)$ denotes a potentially different value from the perspective of user b.

Similarly, the triangle equality is difficult to achieve for similarity of user profiles because profiles can consist of more than one dimension (i.e., profile item). As a result, although the triangle property holds for one dimension, similarities for three user profiles with multiple dimensions might not adhere to the triangle property. Ideally, any formula that does not carry all these four properties should be called a measure, but researchers still prefer to use the word metric interchangeably with the term measure.

Key Points

Similarity metrics that have been proved efficient and practical on social networks are those that exploit a locality principle in similarity computations. The basic idea underlying such metrics is that, given two social network users, their similarity is computed by observing only a subset of vertices (e.g., friends of the two users) in the social network. This approach restricts required information about the social network to a minimum and reduces the required time of calculations. Even though global measures (e.g., the shortest path between users or the community membership of users) can be used in the same context, they are more costly in time and computational power because they require too much information about the social network. For example, shortest path calculation might require observing friendship links of many users. Moreover, even though the costs can be undertaken, researchers and companies cannot have access to the whole social network data because of privacy issues. Only owners of social networking services can have the complete data that is required to compute global similarity measures. Therefore, local similarity metrics, which are the focus of this entry, provide a simple alternative in the face of these costly issues.

Historical Background

In the 1970s, early attempts at defining similarity metrics involved finding similarities among text-based documents (McGill 1979) that were modeled as a collection of words. Similarity metrics were used to discover relations between documents or rank the documents according to their similarity to a given query. These efforts have resulted in several well-known metrics, such as the Cosine and Jaccard similarities. With the advent of the Internet, researchers have applied document-based similarity metrics to user-generated web items, such as friendships and status posts, to discover relationships between web users. Specifically, similarity research on user-generated data has focused on predicting links (relationships) among users and mining past user behavior to predict future actions.

In the link prediction problem (Liben-Nowell and Kleinberg 2007), similarity of social network users has been exploited to predict new friendships. In this context, high similarity between two social network users is assumed to increase the probability of them creating a new friendship (Spertus et al. 2005). With generalization, this idea has been explored in the homophily theory which states that people tend to be friends with other people who are similar to them along personal attributes, such as gender, race, and religion (McPherson et al. 2001a).

In addition to user characteristics, actions of a user are exploited to predict actions of similar users. This idea has been studied in recommender systems to observe existing item ratings (for movies, books, songs, etc.) of users and predict ratings of unseen items (Melville and Sindhwani 2010). Similar users are assumed to give similar ratings to similar items. From this assumption, similarity of ratings is predicted by finding either similar users (i.e., user based) or similar items (i.e., item based).

Methods

On social networks, user-generated data are classified into two types: profile data refers to user entered textual information, such as hometown, whereas network data refers to created relationships, such as friendships with other users on the social network. Depending on the type of user data, similarity metrics differ in how they model a social network user. By taking this into account, we will first explain how similarity metrics work on network data and then continue to explain metrics for profile data.

Network Similarity

In network similarity metrics, existing user relationships are exploited to find similarity. For example, a big number of shared friends between two users can be assumed to imply their high similarity. Network data that represent relationships can be modeled as a graph $\mathcal{G} = (V, E)$, where each user a in the network is considered a vertex $\upsilon_a \in V$ and a relationship between users a and b is an edge $e_{ab} \in E$ on the graph $\mathcal{G}$. If relationships are established by mutual consent of two users, an edge between them is said to be undirected (the edge has no start and end points), and it can be called a *friendship*. Friendship relations on social networks, such as Facebook, Orkut, and LinkedIn, are modeled with undirected graphs. On an undirected graph, first-level neighbors of a vertex υ_a are a set of vertices $\Gamma(\upsilon_a)$ who share an edge with υ_a (e.g., friends of a). Similarly, second-level neighbors $\Gamma(\Gamma(\upsilon_a))$ (i.e., friends of friends of a) share an edge with first-level neighbors of a. If relationships can be created without mutual consent, an edge is said to be directed; it

Similarity Metrics on Social Networks, Table 1 Network similarity metrics

Measure	Formula	Description
Overlap	$\vert\Gamma(\upsilon_a) \cap \Gamma(\upsilon_b)\vert$	The number of common neighbors
Preferential attachment	$\vert\Gamma(\upsilon_a)\vert \times \vert\Gamma(\upsilon_b)\vert$	Multiplied neighbor counts of both users
Jaccard	$(\vert\Gamma(\upsilon_a) \cap \Gamma(\upsilon_b)\vert)/(\vert\Gamma(\upsilon_a) \cup \Gamma(\upsilon_b)\vert)$	The percentage of shared neighbors over all neighbors
Cosine	$(\vert\Gamma(\upsilon_a) \cap \Gamma(\upsilon_b)\vert)/\left(\sqrt{(\vert\Gamma(\upsilon_a)\vert.\vert\Gamma(\upsilon_b)\vert)}\right)$	The number of common neighbors normalized by multiplied neighbor counts
Adamic and Adar	$\sum_{\upsilon_c \in \{\Gamma(\upsilon_a) \cap \Gamma(\upsilon_b)\}} \frac{1}{Log(\vert\Gamma(\upsilon_c)\vert)}$	Common neighbors who have very few neighbors are given more importance
Point-wise mutual information	$P(\Gamma(\upsilon_a), \Gamma(\upsilon_b)) \times log\left(\frac{P(\Gamma(\upsilon_a), \Gamma(\upsilon_b))}{P(\Gamma(\upsilon_a)).P(\Gamma(\upsilon_b))}\right)$	How much the probability of having the current set of common neighbors differs from the case where neighbors would be added by users on the graph randomly
Katz's measure	$\sum_{p=1}^{+\infty} \beta^p \times NumofPath(\upsilon_a, \upsilon_b, p)$	Similarity is implied by the number and length of paths that connect two users on the graph. Each path $0 < p < +\infty$, whereas $0 < \beta < 1$

starts from the user who initiated the relationship (i.e., source vertex) and ends at the user with whom the relationship was established (i.e., target vertex). For example, popular social networks Twitter and Google+ are directed social networks; when user *a* starts following user *b*, an edge is created, starting from vertex υ_a and ending at vertex υ_b. On directed graphs, the *friendship* term from undirected graphs is replaced with *in-neighbors* and *out-neighbors*. In-neighbors $\Gamma^{-}(\upsilon_a)$ and out-neighbors $\Gamma^{+}(\upsilon_a)$ of a vertex υ_a are defined as target and source vertices of all edges that have vertex υ_a as their source and target vertex, respectively.

Although the similarity metrics that we will discuss are designed to work with the *neighborhood* notion of undirected graphs, they can be applied to directed networks by small modifications. For example, direction of edges can be removed to make the graph undirected. As a popular directed social network, in (Goel et al. 2013) Twitter research has revealed their approach to user similarity, where they use follow lists. Another approach is to consider only one type of edges (out-neighbors or in-neighbors) as neighbors while using the metrics. As these modifications can be used to define neighbors of a user on directed graphs, we will give metric definitions in terms of user neighbors. Assume that a similarity function *sim*(υ_a, υ_b) computes the similarity of users υ_a and υ_b by considering their neighbors $\Gamma(\upsilon_a)$ and $\Gamma(\upsilon_b)$, respectively. We will denote one of their common neighbors with υ_c, i.e., $\upsilon_v \in (\Gamma(\upsilon_a) \cap \Gamma(\upsilon_b))$. High similarity between two users who do not have a relationship will be assumed to increase the probability of them creating a relationship edge. With these definitions, metric formulas that we will explain are given in Table 1. Next we will discuss these network similarity metrics in more details.

Overlap: The overlap measure counts the number of common friends of υ_a and υ_b to compute similarity.

Preferential Attachment: For relationship creation on social networks, the preferential attachment metric reflects the "rich gets richer" notion from sociology. The metric assumes that highly connected vertices (i.e., users who have many neighbors) are more likely to create relationships with each other.

Jaccard ($\mathcal{L}1$ norm): The Jaccard metric counts the number of common neighbors as in the overlap metric, but it normalizes this value by using the total number of neighbors of users υ_a and υ_b.

Cosine ($\mathcal{L}2$ norm): Cosine similarity was originally devised to find the similarity of two documents by computing the cosine of the angle between their feature vectors. When the angle between the two vectors is 0, they are considered identical, and the cosine of the angle equals the maximum similarity value 1.

Adamic and Adar: Like the overlap metric, Adamic and Adar considers common neighbors, but each neighbor's impact on the similarity value depends on the number of its neighbors (Adamic and Adar 2003). If a common neighbor has few neighbors, its impact on the similarity is assumed to be higher. For example, if users υ_a and υ_b are the only neighbors of υ_c , 1/ log (2) is added to the overall similarity. The similarity is computed by summing values from all common neighbors.

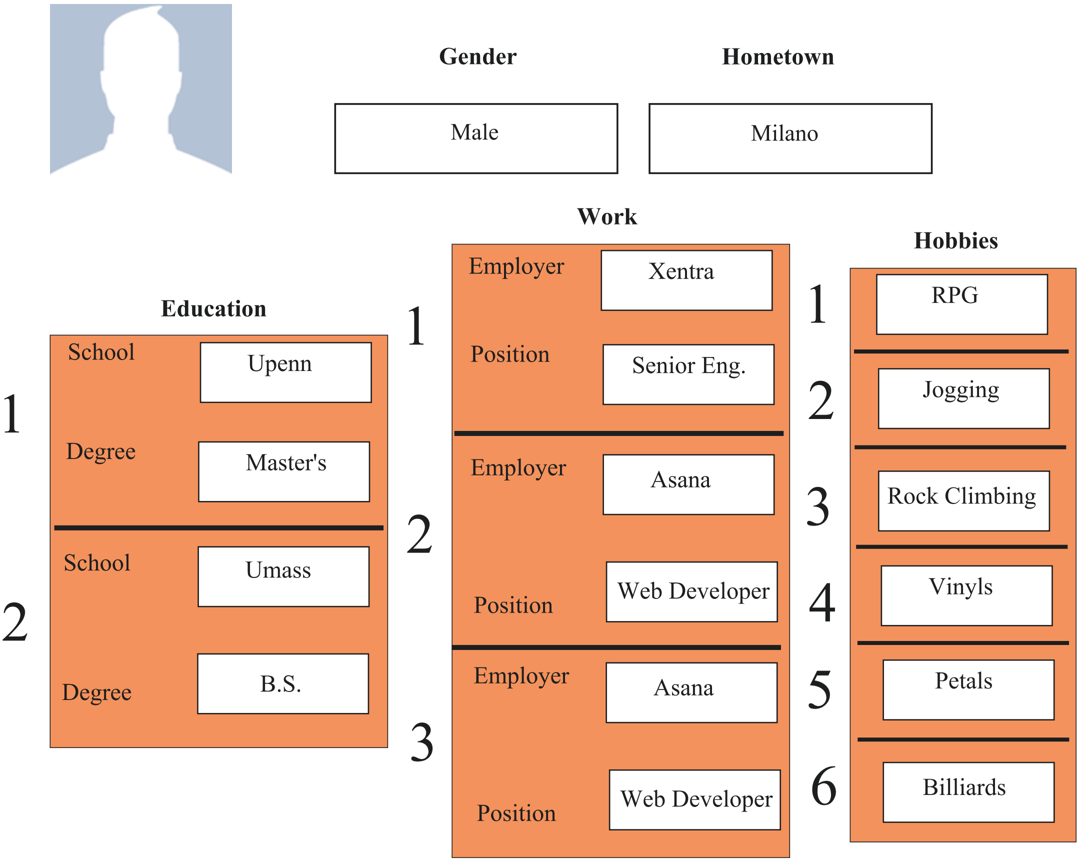
Point-Wise Mutual Information (Positive Correlations): Point-wise mutual information (Bouma 2009) is computed in probabilistic terms where joint probability distribution function $P(\Gamma(\upsilon_a), \Gamma(\upsilon_b))$ computes the probability of a graph vertex $\upsilon_x \in V|x \neq a \neq b$ sharing edges with both υ_a and υ_b, whereas marginal probability distribution functions $P(\Gamma(\upsilon_a))$ and $P(\Gamma(\upsilon_b))$ are probabilities of a graph vertex sharing an edge with υ_a and υ_b, respectively. If edges are assumed to represent friendships, $P(\Gamma(\upsilon_a), \Gamma(\upsilon_b))$ is equal to $\frac{\#\,mutual\ friends}{\#users\ in\ the\ social\ network}$. Similarly, $P(\Gamma(\upsilon_a))$ is equal to $\frac{\#\,friends\ of\ \upsilon_a}{\#users\ in\ the\ social\ network}$. In other words, point-wise mutual information shows whether two users share mutual friends due to randomness. Note that due to computing probability with the total number of users in the social network, point-wise mutual information produces very low average similarity values, because *#users in the social network* can be in millions.

Katz's Measure: Katz's measure (1953) was designed in the 1950s to find the status of a vertex on a graph. The vertex which had the biggest number of shortest paths to the other vertices was considered a central vertex with high status. To compute $sim(\upsilon_a, \upsilon_b)$, Katz's measure finds the number of paths that connect υ_a and υ_b for path length p, $1 < p < +\infty$. The number of paths (i.e., the value of $Num\ Of\ Path(v_a, v_b, p)$) is dampened by a β^p value, where $0 < \beta < 1$. In practice, the β value is chosen as small as 0.005 (Liben-Nowell and Kleinberg 2007). Although p values can be increased to cover a big portion of the graph, usually $p = 2$ or $p = 3$ values are chosen to find similarity, because computations for $p > 3$ contribute very less to the overall value. Note that although the Katz's measure can be used as a global measure with big p values, small p values (e.g., 2 or 3) make it a practical measure for fast, local similarity computations.

In research work, performance of similarity metrics has been compared by making predictions based on similarity values and validating the results (Liben-Nowell and Kleinberg 2007; Spertus et al. 2005). Typically, metrics are used to predict top k relationships (e.g., k most probable future friendships that will be created between users) on graphs at a time t_1, and these predictions are validated at a time $t_2 > t_1$. The performance of a metric can be computed by counting the number of correct predictions.

In Liben-Nowell and Kleinberg (2007), Adamic and Adar has been shown to perform better than preferential attachment, Jaccard, overlap, and Katz's measure, on a scientific coauthorship network. On another social network, Orkut.com, Spertus et al. (2005) have found that cosine similarity performed better than Jaccard and point-wise mutual information metrics.

Despite these comparisons, it is important to understand that each metric has its weaknesses in different application fields. Preferential attachment is widely used in social networks to predict friendships, but unlike Jaccard or cosine similarity its computed value does not reside within [0, 1]. In fact, its max value is only bounded by the total number of users in a social network, because there are no theoretical limits to prevent a user from having every other social network user as a neighbor. However, some social networks may choose to limit the number of neighbors; for example, an undirected network, Facebook, allows up to 5000 neighbors, whereas a directed network, Twitter, does not have such a limit. Because of this, preferential attachment cannot be used to quantify how much *percent* two users are similar. When the graph has many vertices (i.e., the probability of an edge between two users are very small), computed value of point-wise mutual information can be very small, and it cannot be used to define how much *percent* two users are similar. Point-wise mutual information and preferential attachment can be best used in ranking a set of users according to their similarity to a specific user. There are some limitations in using Katz's measure too. In popular social networks, discovering



Similarity Metrics on Social Networks, Fig. 1 An exemplary profile for a social network user. The profile consists of single-valued gender and hometown items, as well as multiple-valued education and work items

edge counts for paths of length 2 or more can be restricted because social networking services do not allow access to social network data. Furthermore, Katz's measure can be costly to compute when the social network is large. Considering these limitations, research work (Jin et al. 2011) has mostly used Jaccard, cosine, or Adamic and Adar in their user similarity computations, because these measures are fast and easier to interpret.

Profile Similarity

Along with network data, profile data constitute the second type of user-generated data on social networks. We will call similarity metrics which work with profile data as profile similarity metrics. On social networks, we will consider profile information as a set of unique items (e.g., hometown, location, education of a user), which can have one or multiple subfields for each value. Figure 1 shows an example of user profile with two education values, where an education value can have more than one subfield (i.e., school and degree). The figure also shows some items which can have only one value, such as gender:male.

After modeling profile data with a set of items, similarity between two users is computed by first finding the similarity of individual item values on the two profiles. For example, similarity of two users according to the gender item compares gender values on the two profiles. If the considered item has many values, or each value has multiple subfields (e.g., the education item on Fig. 1), an aggregation function is required to find item similarity. An overall profile similarity value is determined by aggregating (e.g., by weight averaging) all item similarity values. However in practice, most of the research work on profile similarity consider a user profile to be a set of unstructured keywords. For example, in (Bhattacharyya et al. 2010), Facebook user profiles only consist of user values from the "hobbies" item. Afterward, profile similarity of two users is found by computing item similarity of hobbies on two profiles.

In a more detailed study, similarity of items which have multiple values or subfields has been weight averaged to model the importance of similarity for some items or subfields (Akcora et al. 2011). For example, when user profiles consist of hometown and hobbies fields, hometown similarity can be weighted with a bigger coefficient to show that hometown similarity is more important than hobbies similarity.

When similarity computations are reduced to find item similarities, the type of item values determines the way similarity is computed. Although some items have numerical values (e.g., age, zip code), most of the item values on social networks are text-based categorical data which cannot be ordered on an axis to find similarity/distance of two data points. Using simple approaches, such as string matching, is not efficient because the text represents an identity (e.g., hometown: Barcelona), where partial (n-gram) similarities (e.g., bARcelona:pARis) are trivial. For a detailed study of text similarity, see Gomaa and Fahmy (2013).

Two main approaches are used to find similarity of item values: ontology based (Mika 2005) and social graph based (Akcora et al. 2011). In ontology-based approaches, a graph of entities is created to define their relationships or distance (Cristani and Cuel 2005). For example, considering hometown similarity of three social network users with values Barcelona, Madrid, and New York, an ontology can classify Barcelona and Madrid as Spanish cities, whereas New York is classified as an American city, and compute a higher similarity for users from Barcelona and Madrid. The main disadvantage of this approach is that it requires a reliable ontology which can be difficult to create. Furthermore, as social networks are dynamic, new item values are added in time, and the ontology must be updated frequently.

The social graph-based approach assumes that neighbors (e.g., friends/coauthors) of a user υ_a are similar to υ_a along profile attributes. For example, if hometown of υ_a is Barcelona, we can expect many of its neighbors to be from Barcelona and other Spanish cities. When hometown values of neighbors are observed, another city (e.g., Madrid) can be found similar to Barcelona without explicitly creating an ontology. With this intuition, a user υ_b is said to be similar to υ_a if its hometown is similar to the hometown values of υ_a's neighbors. Furthermore, even when a user has a blank profile, using its neighbors in such a way allows one to compute its similarity with other social network users. The social graph approach has also been found effective for network similarity metrics (Cukierski et al. 2011). The disadvantage of this approach is that neighbors are assumed to be similar to users. This assumption is more applicable in undirected social networks where mutual consent is required to create a relationship edge. However, if the network is undirected, neighbors can have very different characteristics from users, and performance of social graph-based approaches can deteriorate. Another approach is based on using an external data source, such as WordNet to compute similarity even if two profiles do not share any common terms (Spear et al. 2009).

As ontology-based approaches have been extensively studied in semantic web communities (Cristani and Cuel 2005), in the rest of this section we will detail social graph-based approaches. To this end, we will discuss relevant categorical data similarity measures (Boriah et al. 2008) which can be used with the assumptions of social graph-based approaches. Before doing that, we need to introduce some notations and definitions.

Assume that for an item i, we are given a pair of item values i_a and $i_{b.}$ from profiles of υ_a and υ_b, respectively.

From the set of υ_a's neighbors, we create a collection of values, $values(i) = \{\forall i_c |\ \upsilon_c \in \Gamma(\upsilon_a)\}$. We will define three functions over $values(i)$. Function $distinct(i)$ finds the number of distinct values in $values(i)$, and $sup(i_x)$ finds the count of value i_x in $values(i)$, whereas $freq(i_x) = \text{sup}\ (i_x)/|\Gamma(\upsilon_a)|$. With these functions, we will explain the categorical

Similarity Metrics on Social Networks, Table 2 Item similarity measures with their formulas. The two items are identical when $i_a = i_b$. Each measure uses a different formula for identical and nonidentical value pairs

Measure	Formula
Eskin	$1, \text{if } i_a = i_b$ $\frac{distinct(i)^2}{\left(2 + distinct(i)^2\right)}, \text{if } i_a \neq i_b$
Occurrence frequency	$1, \text{if } i_a = i_b$ $\frac{1}{1 + \left(log\left(\frac{\lvert\Gamma(\upsilon_a)\rvert}{sup(i_a)}\right) \times log\left(\frac{\lvert\Gamma(\upsilon_a)\rvert}{sup(i_b)}\right)\right)}, \text{if } i_a \neq i_b$
Lin	$\frac{2 \times log(freq(i_a))}{log(freq(i_a)) + log(freq(i_b))}, \text{if } i_a = i_b$ $\frac{2 \times log(freq(i_a) + freq(i_b))}{log(freq(i_a)) + log(freq(i_b))}, \text{if } i_a \neq i_b$

similarity functions reported in Table 2. Variations of Lin and Eskin measures are excluded for brevity.

Eskin: Eskin's measure (Boriah et al. 2008) assigns 1 in identical cases ($i_a = i_b$), and it penalizes users when their values do not match while there are very few distinct values in $values(i)$.

For example, it punishes users more for mismatches in the gender item (two values with male and female) than it does in the hometown item because hometown can have many more values.

Occurrence Frequency: The occurrence frequency measure assigns 1 to identical value pairs, and it favors mismatches with highly frequent values. If $values(i)$ has two distinct values i_x and i_y, with $i_x = i_a$ and $i_y = i_b$, $sim(i_a, i_b)$ reaches its maximum value.

Lin: Unlike others, Lin's measure (Lin 1998) does not assign 1 to two identical item values. Instead, it assigns high similarity when the two values are highly frequent in $values(i)$. In other words, if item value of υ_b is very frequent among friends of υ_a, $\upsilon_b$is considered very similar to υ_a. For mismatches (i.e., $i_a \neq i_b$), the measure gives less importance to infrequent values.

A comparative evaluation of these measures has been carried out in (Akcora et al. 2011), where the occurrence frequency has been found superior in performance. Overall, choosing social graph-based approaches over ontology-based approaches improves profile similarity results because social graph-based approaches can use profiles of neighbors to infer blank profile items of a user. By doing so, scarcity of data on user profiles can be compensated, and similarity can be computed for more social network users. This gives an edge to social graph-based approaches, because analysis of real-life social networks indicates that a big portion of user profiles (up to 60% for a popular network, Facebook.com (Akcora et al. 2011)) are indeed missing.

Key Applications

A comparison between metrics for profile and network data can be done according to two dimensions: interpretation and availability. In interpreting results, profile data is richer than the network data because it covers more relations between users, and similarity values can be interpreted in terms of items. For example, high similarity between two users can be pointed to their common hometown, education, or religion values. On the other hand, network similarity offers only graph edges as its data, and network similarity results can only be interpreted in terms of being connected on a graph. For example, in Jaccard similarity, two users' similarity can be due to many shared friends, but the metric cannot explain why they share these friends at the first place. In such a case, profile similarity could point out that common friends are due to the shared hometown values.

In availability, network similarity is easier to use because network edges are structured and easier to discover. In comparison, profile data is more scarce and polluted; users might not enter any profile data, or any data they enter might be unstructured. Profile data is also more difficult to find in research data sets because of privacy issues.

Future Directions

In recent years complete academic works have been published on user similarity (Rawashdeh 2015; Han 2015). Usage of similarity metrics has attracted more attention in problems such as link prediction and clone profile detection on social networks. As data grew, research works dealt with scalable similarity computing (Zadeh and Goel 2013). Some research work have also used similarity metrics as an auxiliary method in link prediction where users have not yet generated any actions on social networks (i.e., in the face of cold start) (Leroy et al. 2010) or to predict the risk of interacting with a user in terms of disclosure of personal information (Akcora et al. 2012). A recent direction is to analyze whether users have the same interest (Han et al. 2015) and assign high similarity to those that follow similar accounts.

Despite these works, similarity metrics are still used as black box models without considering their descriptive powers. If this aspect is dully considered, similarity metrics can be used to explain why users are similar and what types of users interact with each other.

Cross-References

- ► Centrality Measures
- ► Components of the Network Around an Actor
- ► Link Prediction

References

Adamic L, Adar E (2003) Friends and neighbors on the web. Soc Networks 25(3):211–230

Akcora C, Carminati B, Ferrari E (2011). Network and profile based measures for user similarities on social networks. In: 2011 I.E. international conference on Information reuse and integration (IRI). IEEE, pp 292–298

Akcora C, Carminati B, Ferrari E (2012) Privacy in social networks: how risky is your social graph? In: 2012 I.E. 28th international conference on Data engineering (ICDE). IEEE, pp 9–19

Bhattacharyya P, Garg A, Wu S (2010) Analysis of user keyword similarity in online social networks. Soc Netw Anal Min 1:1–16

Boriah S, Chandola V, Kumar V (2008) Similarity measures for categorical data: a comparative evaluation. SIAM 30(2):243–254

Bouma G (2009) Normalized (pointwise) mutual information in collocation extraction. In: Proceedings of GSCL conference, pp 31–40

Cristani M, Cuel R (2005) A survey on ontology creation methodologies. Int J Semant Web Inform Syst (IJSWIS) 1(2):49–69

Cukierski W, Hamner B, Yang B (2011) Graph-based features for supervised link prediction. In: The 2011 international joint conference on Neural networks (IJCNN). IEEE, pp 1237–1244

Goel A, Sharma A, Wang D, Yin Z (2013) Discovering similar users on twitter. In: 11th workshop on mining and learning with graphs, Chicago, USA

Gomaa WH, Fahmy AA (2013) A survey of text similarity approaches. Int J Comput Appl 68(13):13–18

Ha V, Haddawy P (2003) Similarity of personal preferences: theoretical foundations and empirical analysis. Artif Intell 146(2):149–173

Han X (2015) Mining user similarity in online social networks: analysis, modeling and applications. PhD dissertation, Evry, Institut national des télécommunications

Han X, Wang L, Crespi N, Park S, Cuevas Á (2015) Alike people, alike interests? inferring interest similarity in online social networks. Decis Support Syst 69:92–106

Jin L, Takabi H, Joshi J (2011) Towards active detection of identity clone attacks on online social networks. In: Proceedings of the first ACM conference on Data and application security and privacy. ACM, pp 27–38

Katz L (1953) A new status index derived from sociometric analysis. Psychometrika 18(1):39–43

Leroy V, Cambazoglu B, Bonchi F (2010) Cold start link prediction. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 393–402

Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. J Am Soc Inf Sci Technol 58(7):1019–1031

Lin D (1998) An information-theoretic definition of similarity. In: Proceedings of the 15th international conference on Machine Learning, vol. 1. San Francisco, pp 296–304

McGill M (1979) An evaluation of factors affecting document ranking by information retrieval systems. Washington: Institute of Education Sciences

McPherson M, Smith-Lovin L, Cook J (2001a) Birds of a feather: homophily in social networks. Annu Rev Sociol 27:415–444

Melville P, Sindhwani V (2010) Recommender systems. Encycl Mach Learn 1:829–837

Mika P (2005) Ontologies are us: a unified model of social networks and semantics. In: International semantic web conference, Springer, pp 522–536

Rawashdeh AY (2015) Semantic similarity of node profiles in social networks. PhD dissertation, University of Cincinnati

Richter M (2007) Foundations of similarity and utility. In: The 20th international FLAIRS conference, Key West, Florida

Spear M, Lu X, Matloff NS, Wu SF (2009) Inter-profile similarity (IPS): a method for semantic analysis of online social networks. In: International conference on complex sciences. Springer, pp 320–333

Spertus E, Sahami M, Buyukkokten O (2005) Evaluating similarity measures: a large-scale study in the orkut social network. In: Proceedings of the 11th SIGKDD. ACM, pp 678–684

Zadeh RB, Goel A (2013) Dimension independent similarity computation. J Mach Learn Res 14(1):1605–1626

Recommended Reading

Anderson A, Huttenlocher D, Kleinberg J, Leskovec J (2012) Effects of user similarity in social media. In: Proceedings of the fifth ACM international conference on Web search and data mining. ACM, pp 703–712

De Meo P, Ferrara E, Fiumara G (2011) Finding similar users in Facebook. Soc Netw Community Behav Model: Qual Quant Meas 4:126

McPherson M, Smith-Lovin L, Cook J (2001b) Birds of a feather: homophily in social networks. Annu Rev Sociol 27:415444

Simple Event

▶ Theory of Probability: Basics and Fundamentals

Simple Walk

▶ Semirings and Matrix Analysis of Networks

Simulated Datasets

Fahimah Al-Awadhi
Department of Statistics and Operations
Research, Kuwait University, Kuwait City,
Kuwait

Synonyms

Gibbs sampler; Markov chain Monte carlo algorithms; Metropolis Hastings; Monte Carlo methods; Statistical simulation

Glossary

GS	Gibbs Sampler
IID	Independent Identically Distributed
MC	Monte Carlo
MCM	Monte Carlo Methods
MCMC	Markov Chain Monte Carlo
MCS	Monte Carlo Simulation
MHA	Metropolis-Hastings Algorithm

Introduction

Simulation is the imitation of the operation of a real-world process or system over time. Simulation has appeared at the very early stages of the development of statistics as a field. Francis Galton invented mechanical devices in 1873 to compute estimators and distributions by means of simulation. His well-known quincunx (Stigler 1986) is a derivation of the Central Limit Theorem for Bernoulli experiments. The randomized experiments of Ronald Fisher (1935) and the bootstrap revolution started by Brad Efron (Efron and Tibshirani 1993) are intrinsically connected with calculator and computer simulations, respectively.

Simulations were used to test a previously understood deterministic problem that has no random variables and no degree of randomness. Statistical sampling was used to estimate uncertainties in the simulations. Monte Carlo simulation (MCS) inverts this approach, solving deterministic problems using a probabilistic analog. An early variant of the Monte Carlo Methods (MCM) can be seen in Buffon's needle experiment, in which π can be estimated by dropping needles on a floor made of parallel strips of wood.

In the 1930s, Enrico Fermi first experimented the MC methods while studying neutron diffusion. In the early 1940s, it was applied in research into nuclear fission. The scientists working on the Manhattan project, making the atomic bomb, had intractably difficult equations to solve in order to calculate the probability with which a neutron from one fissioning uranium atom would cause another to fission. The equations were complicated because they had to mirror the complicated geometry of the actual atomic bomb. The answer had to be right because, if the first test failed, it would be months before there was enough uranium for another attempt. Despite having most of the necessary data, researchers were unable to solve the uncertainty problems using conventional, deterministic methods.

Stanislaw Ulam suggested MCM for evaluating complicated mathematical integrals that arise in the theory of nuclear chain reactions. Von Neumann carried this suggestion to the more systematic development of MC. with the primitive facilities available at the time, Ulam and von Neumann did carry out numerical computations that led to a satisfactory design.

In the 1950s MCM were used at Los Alamos for early work relating to the development of the hydrogen bomb and became popularized in the fields of physics, physical chemistry, and operations research. The RAND Corporation and the US Air Force were two of the major organizations responsible for funding and disseminating information on MCM during this time, and it began to find large applications in many different fields.

Widespread Applications of Simulation

MCM are especially useful for simulating phenomena with significant uncertainty in inputs and systems with a large number of coupled degrees of freedom. Areas of application include:

- Statistics: MCM are generally used for comparing competing statistics for small samples under realistic data conditions and are also used to provide implementations in various fields such as image analysis, signal processing, point processes, econometrics, and surveys.
- Mathematics: To evaluate multidimensional definite integrals with complicated boundary conditions, it is an alternative for the deterministic numerical integration algorithms.
- Physical sciences: MCM are used in computational physics, physical chemistry, quantum systems, and related applied fields. MC molecular modeling is an alternative to computational molecular dynamics.
- Astrophysics: MCM are used in the ensemble models that form the basis of modern weather forecasting. They are also used to model both the evolution of galaxies and the transmission of microwave radiation through a rough planetary surface.
- Engineering: MCM are used for sensitivity analysis and quantitative probabilistic analysis in process design. For example, MCM are applied to analyze correlated and uncorrelated variations in analog and digital integrated circuits in microelectronics engineering.
- Geostatistics: MCM underpin the design of mineral processing flow sheets and contribute to quantitative risk analysis.
- Computational biology: MCM are used in Bayesian inference in phylogeny.
- Finance: To calculate the value of companies and to evaluate investments in projects at a business unit or corporate level, they are also used to calculate the risk and to evaluate financial derivatives and to model project schedules.

Simulation Techniques

Let X = (X_1,. . . , X_n) be IID random variable defined on a suitable sample space $\mathbb{E}$, and assume that each X_i has a known density function π X_i (x_i) defined on $\mathbb{E}$. In many problems, X is high dimensional and evaluation of a function g(X) using π x is a challenging problem. The function under interest will be given by

$$E_{\pi_X}(g(x)) = \int_{x \in \mathbb{E}} g(x)\pi_x dx.$$

Analytical calculation of the above integral is not possible because of the complexity of the distribution function π x $_i$. Simulation inference can be used instead. Suppose, for example, that we have a way to obtain independent samples $x^{(1)}$, $x^{(2)}$. . . from π x, we could then approximate the expectation of g(X,) by the empirical estimate

$$\overline{g}_n = \frac{1}{n}\sum_{i=1}^{n} g\left(x^{(i)}\right).$$

By the strong law of large numbers as n → ∞;

$$\overline{g}_n \overset{a.s.}{\rightarrow} E_{\pi_X}(g).$$

So, if we were able to find an explicit, easily codable function h(U$_1$, U$_2$, . . .) of uniform. (0; 1) variates with values in E and probability distribution the same as π X, then we would evaluate the desired integral as

$$\frac{\sum_{j-1}^{n} g\left(h\left(U_{j1}, U_{j2}\right)\right)}{n},$$

where U $_{jk}$ is a double indexed array of IID uniform variables. It is allowed for h to depend on unboundedly many U variables, as long as the number of such variables required is a random variable with finite expectation. For example, in a one-dimensional integral, we would use random numbers to select points { x $_i$, i = 1,. . . n} in the interval a ≤ x ≤ b and then use the approximation

$$\int_a^b \pi(x)dx \approx \frac{(b-a)\sum_{i=1}^{n} \pi(x_i)}{n}.$$

MCM all follow a similar pattern:

1. Define some domain of inputs $\mathbb{E}$. This just means we have some set of variables in the model and we want to know the range of the values they can take on.
2. Generate inputs randomly, governed by some probability distribution π.
3. Perform some computation on these inputs.
4. Repeat 2 and 3 over and over a very large number of times.
5. Aggregate the results from the previous step into some final computation.

The result is an approximation to some true but unknown quantity.

Examples

Here are some simple examples on MCS.

Example 1 A simple MCS to approximate the value of π could involve randomly selecting points (x $_i$, y $_i$), i = 1, • • •, n in the unit square and determining the ratio $\rho = \frac{m}{n}$, where m is number of points that satisfy $x_i^2 + y_i^2 \leq 1$. Consider a circle inscribed in a unit square. Given that the circle and the square have a ratio of areas that is $\frac{\pi}{4}$, the value of π can be approximated using an MCM as follows:

1. Draw a square, then inscribe a circle within it.
2. Uniformly scatter some objects of uniform size, over the square.
3. Count the number of objects inside the circle and the total number of objects.
4. The ratio of the two counts is an estimate of the ratio of the two areas, which is $\frac{\pi}{4}$. Multiply the result by 4 to estimate π.

In a typical simulation of n = 1,000 sample size there were 787 points satisfying. Using this data, we obtain

$$\rho = \frac{787}{1,000} = 0.787 \text{ and}$$
$$\pi \approx \rho \times 4 = 0.0787^* \, 4 = 3.418.$$

If points are purposefully dropped into only the center of the circle, they are not uniformly distributed, so our approximation is poor. The approximation is generally poor if only a few points are randomly dropped into the whole square. The approximation improves as more points are dropped.

Example 2 Suppose we want to find out the probability that, out of a group of 50 people, 2 people or more people share birthdays. The probability of having ateast two people in the group having the same birthday is equal to $1 - \frac{365!}{365^{50}(365-50)!} = 0.970$, where ! is the factorial operator. Using the MC approach:

1. Pick 50 random numbers in the range [1;365]. Each number represents 1 day of the year.
2. Check to see if any of the 50 are equal.
3. Go back to step 1 and repeat 1,000 times.
4. Report the fraction of trials that have matching birthdays.

Using 1,000 iterations, the probability is approximately 0:864. Obviously, the more times we repeat the experiment, the more precise our result would be. Better than repeating the experiment a 1,000 times, we can easily use a computer to simulate the experiment 10,000 times (or more). Using 100,000 iterations, the simulated result was 0.969.

Example 3 Consider calculating the probability of a particular sum of the throw of two dice. There are 36 combinations of dice rolls. We can manually compute the probability of a particular outcome. For example, there are six different ways that the dice could sum to seven. Hence, the probability of rolling 7 is equal to 6 divided by 36 = 0:167. Using MC approximation:

1. Throw the two dice and record the sum of the output.
2. Go back to step 1 and repeat 10,000 times.
3. Report the fraction of trials for each of the 11 different sum.

If the dice totaled 1,813 times out of 10,000 rolls, we would conclude that the probability of rolling 7 is approximately 0:1813.

The accuracy of an MCS is a function of the number of realizations. That is, the confidence bounds on the results can be readily computed based on the number of realizations. Every time a Monte Carlo simulation is made using the same sample size, it will come up with a slightly different value. The values converge very slowly of the order $O\left(n^{-\frac{1}{2}}\right)$. This property is a consequence of the law of large numbers.

Markov Chain Monte Carlo Methods

Typically, the distribution π X is too complex for direct simulations. Thereupon, the indirect approach of MCMC must be applied. This approach will simulate correlated samples $\{x(_i)\}$ from π X. As the iterations depart more from independence, the number of iterations required for a given degree of accuracy increases. Other algorithms for constructing such transition kernel have been proposed such as importance sampling which involves sampling the points randomly, but more frequently where the integrand is large. One can approximate the integral by an integral of a similar function or use adaptive routines such as stratified sampling, or adaptive umbrella sampling, or the quasi-MCM which uses low-discrepancy sequences.

MCMC methods are widely advocated in a variety of situations where the complexity of the distribution of interest is an issue. In these situations usually the direct sampling from such complicated models is not possible. The key idea of the MCMC methods is to generate an iterative sequence of samples in such a way that it converges in distribution to the model of interest. To implement this strategy, many attempts were made to define algorithms for constructing chains with specified equilibrium distributions. The most common, well-known algorithms for constructing chains with specified equilibrium distributions were defined by Metropolis et al. (1953), Hastings (1970). A wide range of discussion papers on MCMC theory and application can be referred to, for example, Smith and Roberts (1993), Gilks et al. (1996), Robert and Casella (2004), Suess and Trumbo (2010). In this section, we shall briefly discuss in an appropriate framework the theory of the MCMC technique.

To sample from a specified distribution π on $\mathbb{E}$, we construct an MC transition kernel $\mathbb{P}$ (x; A). Let $X_1; X_2; \ldots, X_n$ be random variables. We say that X satisfies a Markov condition if

$$P\left(X_{n+1} = x \mid X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\right) = P\left(X_{n+1} = x \mid X_n = x_n\right).$$

The transition kernel $\mathbb{P}$ is a map, $\mathbb{P}: \mathbb{E} \times \mathbb{E} \to [0;1]$, that implies the target distribution π is a stationary distribution of the chain. The distribution of $X^{(t+1)}$ given $X^{(t)}$ satisfies

$$P\left(X^{(t+1)} \in A \mid X^{(0)} = x^{(0)}, \ldots, X^{(t-1)} = x^{(t-1)}, X^{(t)} = x\right) = \mathbb{P}(x, A).$$

We say that π is the invariant measure (hence equilibrium) of the MC if it satisfies the general balance equation

$\int_{x\in\mathbb{E}}\pi(dx)\mathbb{P}(x,A)=\pi(A),$
for all measurable sets $A\subset\mathbb{E}$.

General balance, π ℙ = π, is also referred to as the global balance.

The conditional distribution of X $^{(t)}$ given X $^{(0)}$ = x $^{(0)}$ is

$$P\left(X^{(t)}\in A|X^{(0)}=x^{(0)}\right)=\mathbb{P}^{t}\left(x^{(0)},A\right),$$

where $\mathbb{P}^t$ denotes the kernel ℙ after iterating it t times. ℙ should π irreducible, aperiodic, positive recurrent and π ℙ = π (Nummelin 1984). A chain is π irreducible if starting at any initial state x ∈ 𝔼, then for all measurable sets A ⊂ 𝔼 with π(A) > 0 there exists t > 0 such that $P(X^{(t)}\in A|\,x^{0}=x)$ > 0. The chain is aperiodic if the chain does not oscillate between different sets of spaces in a regular periodic movement. The term positive recurrent is defined as follows: let τ $_A$ be the first return time to state A ⊂ 𝔼 where π(A) > 0, then we say the π irreducible chain X $^{(t)}$ is recurrent if P(τ $_A$ < ∞) = 1 and is positive recurrent if $E(\tau_A)<\infty$. If the chain is π – irreducible, aperiodic, positive recurrent and if the initial value of X$^{(0)}$ is sampled from π, then all subsequent iterations using MCMC will also be distributed according to π. For example, drawing a number from {1,2,3} with replacement where X $_t$ is last number seen at time t is an MC, but if we draw a number without replacement, then it is not MC. If the initial value of X(^{0}) is sampled from π, then all subsequent iterations using MCMC will also be distributed according to π.

Such methods include the MHA, GS, and the Wang and Landau algorithm. We recall now the most commonly used.

The Metropolis-Hastings Algorithm

This algorithm was first proposed by Metropolis et al. (1953) and extended by Hastings (1970). The algorithm is designed to give samples from a distribution it. It defines a proposal kernel q(x;) to produce a potential new state $x'\in\mathbb{E}$. The proposed candidate x′ is accepted with probability α where

$$\alpha(x,x')=\min\left\{1,\frac{\pi(x')q(x',x)}{\pi(x)q(x,x')}\right\}.$$

If we are currently at time t and x′ is accepted, then X(^{t})= x′ otherwise the chain does not move, i.e., X(^{t}) = x(t− 1).

Formally, the target distribution π is defined with respect to a σ – finite measure. The proposal density q could be defined with respect to a different σ – finite measure from that for π. The transition kernel ℙ(x, x′) using the MHA can be written as

$$q(x,x')\alpha(x,x')\quad\text{if}\ \ x'\neq x.$$

The choice of the distribution q(.,.) is arbitrary provided that q(x; x′) > 0 if and only if q(x′, x) > 0. It is convenient to choose a q that is simple and fast to sample from and for which it is easy to evaluate the acceptance probability. However, the relation between q and′(·) will affect the rate of convergence.

The Gibbs Sampler

The GS was given its name by Geman and Geman (1984) who used it for analyzing Gibbs distributions on a lattice. The algorithm constructs the transition kernel ℙ using the full conditional densities of each component X $_i$, i = 1,. . . n, given the values of the other components X_ $_i$ = { X $_j$; j ≠ i, j = 1,. . . n}. We denote this density by π X $_i$ X|_ $_i$ (x $_i$ | x_ i). Suppose we are at time t and want to update the chain, then (as it is with the MHA) we use either a random sampler or a systematic scan sampler. At each iteration, the random sampler picks a random component say, X $_i$, i ∈ {1. . .n } to update, then the conditional density for $X_i^{(t)}$ becomes $\pi_{X_i|X_{-i}}\left(x_i|\,X_{-i}=x_{-i}^{(t-1)}\right)$. In the systematic scan, we update all the components in turn during one iteration using the marginal conditional densities of the components. In progressing from X $^{(t-1)}$ to X $^{(t)}$, the value of X $_i$ is obtained by sampling from

$$\pi_{X_i|X_{-i}}\left(x_i|\,x_1^{(t)},\ldots,x_{i-1}^{(t)},x_{i+1}^{(t-1)},\ldots,x_n^{(t-1)}\right).$$

Hence, to update X we make random draw from these full conditional densities for each of its components. The iteration is completed when all the components are updated. Hence, the transition probability from x $^{(t-1)}$ to x $^{(t)}$ is given by

$$\mathbb{P}\left(x^{(t-1)},x^{(t)}\right) = \prod_{l=1}^{n}\pi_{X_l|X_{-l}}\left(x_l^{(t)}|x_1^{(t)},x_2^{(t)},\ldots,x_{l-1}^{(t)},x_{l+1}^{(t-1)},\ldots,x_n^{(t-1)}\right).$$

The GS can be regarded as a special case of MHA in which the acceptance rate α is one, meaning that the candidate x $'$ is always accepted.

The Knapsack Example

Given a set of items, each with a weight and a value, determine the number of each item to include in a collection so that the total weight is less than or equal to a given limit and the total value is as large as possible. It derives its name from the problem faced by someone who is constrained by a fixed-size knapsack and must fill it with the most valuable items. To find the most valuable subset of n items that will fit into the knapsack given their weight w $_i$ and value υ $_i$, and subject to knapsack weight limit b.

$z = (z_1, \cdots, z_n) \in \{0, 1\}^n$, z_i means whether we take item i feasible solutions $\mathbb{E} = \left\{z \in \{0,1\}^n; \sum_i w_i z_i \leq b\right\}$. We want to maximize $\sum_i v_i z_i$ subject to z ∈ $\mathbb{E}$:

1. Let the current state be X $_t$ = (z $_1$, . . . z $_n$); we choose j ∈ {1,. . ., n} uniformly at random.
2. Flip z j so Y = (z $_1$, ·· 1 — z $_j$, ··, z $_n$ /.
3. If Y is feasible; that is, the acceptance probability α is high, then set X$_{t+1}$ = Y, else X$_{t+1}$ = X$_t$.

Given a state space $\mathbb{E}$ and a target distribution $\pi = C_b^{-1}\exp\left(b\sum_i v_i z_i\right)$, where C $_b$ is constant. We apply Metropolis algorithm and choose Y ∈ $\mathbb{E}$ randomly using the proposal distribution Q = P [Y = j| X $_t$ = i] = qij. If Y is feasible, it will be accepted with acceptance probability $\alpha = \min\{1, \exp(b\sum_i v_i(y_i - z_i))\}$.

Notice again that this process is an MC because the state we visit next depends only on the state we are currently at and no other state. The n objects z are candidates for inclusion into our random sample. But we must select the members of this set according to some probability Q.

Conclusion

MCS is a very useful mathematical technique for analyzing uncertain scenarios and providing probabilistic analysis of different situations. The basic principle for applying MC analysis is simple and easy to grasp. Various softwares have accelerated the adoption of MCS in different domains including mathematics, engineering, and finance. Various options are available to use MCS in computers. One can use any high-level programming language like C, C++, and Java. R and WinBUGS are free statistical softwares that implement MCMC methods.

Cross-References

- ▶ Gibbs Sampling
- ▶ Simulated Datasets
- ▶ Theory of Probability: Basics and Fundamentals
- ▶ Theory of Statistics: Basics and Fundamentals

References

Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman and Hall, New York

Fisher RA (1935) The design of experiments. Oliver & Boyd, Edinburgh

Geman S, Geman D (1984) Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 6(6):721–741

Gilks W, Richardson S, Speigelhalter D (1996) Markov chain Monte Carlo in practice. Chapman and Hall, London

Hastings W (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equations of state calculations by fast computing machines. J Chem Phys 21:1087–1092

Nummelin E (1984) General irreducible Markov chains and non-negative operators. Cambridge University Press, Cambridge

Robert CP, Casella G (2004) Monte Carlo statistical methods, 2nd edn. Springer, New York. ISBN: 0-387-21239-6

Smith AFM, Roberts GO (1993) Bayesian computation via the gibbs sampler and related MCMC methods. J R Stat Soc B 55:3–23

Stigler SM (1986) The history of statistics: the measurement of uncertainty before 1900. The Belknap Press of Harvard University Press, Cambridge

Suess EA, Trumbo BE (2010) Introduction to probability simulation and gibbs sampling with R. Springer, New York. ISBN: 038740273X

Simulation

▶ Markov Chain Monte Carlo Model

Simulation Modeling

▶ Modeling of Business Processes and Crisis Management

Simulations

Bruce E. Trumbo and Eric A. Suess
Department of Statistics and Biostatistics,
California State University, East Bay, Hayward,
CA, USA

Synonyms

In probability modeling: Monte Carlo procedures, Random sampling with pseudo-random numbers. *In statistical inference*: Bootstraps, (simulated) permutation tests, Markov chain Monte Carlo (MCMC)

Glossary

Beta distributions	The general density function is $f(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$, for $0 < x < 1$ (0 otherwise). We denote a beta distribution by BETA(α, β). Shape parameters: $\alpha > 0$, $\beta > 0$. Note: $\Gamma(1/2) = \sqrt{\pi}$ and, for positive integer k, $\Gamma(k) = (k - 1)!$
Bootstraping	A method of statistical inference based on extensive simulation, often used to make confidence intervals. Takes a large number of *resamples* from the original data (or from a distribution suggested by them) to assess variability. Introduced by Efron (1979) and discussed in Efron and Tibshirani (1998)
Exponential distributions	The general density function is $f(x \mid \lambda) = \lambda e^{-\lambda x}$, for $x > 0$ (0 otherwise). We denote an exponential distribution by EXP (λ). Rate parameter: $\lambda > 0$; the mean and standard deviation are $1/\lambda$
Inverse CDF method	A method of simulating a random variable X from pseudorandom numbers. If X has the cumulative distribution function (CDF) $F_X(x)$, then simulate $X = F_X^{-1}(U)$, where $U \sim$ UNIF (0,1). Also called the quantile method; *quantile* is alternate terminology for the inverse F_X^{-1} of the CDF
Linear congruential generator	A recursive algorithm to generate pseudorandom numbers using modular integer arithmetic
Laplace distributions	The general density function is $f(x \mid \eta, \beta) = \frac{1}{2\beta} \exp\left(-\frac{\mid x - \eta \mid}{\beta}\right)$, for all real x. We denote a Laplace distribution by LAPLACE(η, β). Parameters:

location η (median and mean) and scale $\beta > 0$ (variance $2\beta^2$). Alternative terminology is *double exponential distribution*

Normal distributions — The general density function is $f(x|\,\mu,\ \sigma) = \dfrac{1}{\sigma\sqrt{2\pi}}\exp\left[-\dfrac{1}{2}\left(\dfrac{x-\mu}{\sigma}\right)^2\right]$, for all real x. We denote a normal distribution by NORM $(\mu,\ \sigma)$. Parameters: mean μ and standard deviation $\sigma > 0$. A standard normal distribution has $\mu = 0$ and $\sigma = 1$; its density function is denoted as $\varphi(\cdot)$ and its CDF as $\Phi(\cdot)$

Permutation test — A test in which the null distribution of the test statistic is determined by randomly reassigning outcomes to treatment labels. Also, *exact test*. When the null distribution is difficult to obtain by combinatorial methods, it may be simulated from many random permutations (Eudey et al. 2010)

Pseudorandom numbers (PRNs) — A sequence of numbers generated by a computer algorithm. In practice, PRNs are usually assumed to be random observations distributed UNIF(0,1)

Uniform distributions — The general density function is $f(x|\,\theta_1, \theta_2) = \frac{1}{\theta_2-\theta_1}$, for $\theta_1 < x < \theta_2$ (0 otherwise). We denote a uniform distribution by UNIF(θ_1, θ_2). Mean $\frac{\theta_1+\theta_2}{2}$; variance $\frac{(\theta_2-\theta_1)^2}{12}$

Definition

Simulation is the emulation of a probability model using random numbers accessed with a computer from a high-entropy pool or using PRNs generated by a computer algorithm. In typical applications, a probability model is simulated because important attributes of it are unknown and not convenient (perhaps not possible) to derive by analytic mathematical methods.

Introduction

In practice, almost all simulations require a source of random or pseudorandom numbers (PRNs), taken to be distributed as UNIF(0, 1). Early sections provide some historical background on PRNs currently in general use, information on simulating distributions that are not uniform, and using PRNs in Monte Carlo integration. Later sections illustrate applications of simulation to probability modeling and statistical inference.

Simulations have been used for practical applications throughout the physical sciences for many years. Currently, they are used increasingly in the social sciences, especially in the analysis of large datasets. In applied probability modeling, simulations can test the extent to which simplifying assumptions degrade the accuracy of results and to obtain results that may be difficult to derive analytically. (See Examples 13, 14, 16 and 17.)

In statistical inference, approximate sampling distributions can be checked by simulations to judge the validity of p-values and power computations. (See Examples 19 and 22.) Moreover, some modern methods of statistical analysis depend fundamentally on simulations. (See Examples 23 and 24, and the article on Gibbs Sampling elsewhere in this Encyclopedia.)

Key Points

Vetting PRNs. In most applications, PRNs used for probability simulation are sequences of numbers assumed to be indistinguishable from independent UNIF(0, 1) observations. To verify this assumption, PRNs generated by a particular method are subjected to batteries of benchmark models, known to be difficult to simulate accurately. One collection of such models, called *Diehard*, is due to Marsaglia (1995).

Nonuniform Models. Methods have been developed to use random samples from UNIF (0, 1) to obtain random samples from other distribution families (normal, exponential, Poisson, and so on). They are included in many commercial and open source statistical packages. Although no package contains procedures for all distributions, general methods can be used to simulate distributions not included. We illustrate a few of them in section "Historical Background".

Limitations. A digital computer does not deal with all real numbers, but with a very large and carefully-chosen finite subset of the rational numbers. All simulations based on PRNs begin with rational numbers from a hyperdimensional unit cube, with transformation to the larger set of rational numbers as required. Thus there are some probability models that cannot be satisfactorily simulated on digital computers.

Law of Large Numbers (LLN). The mathematical justification of many simulations relies on the law of large numbers.

Example 1: Repeated rolls of a fair die. A random sequence of equally likely integers 1 through 6 emulates repeated rolls X of a fair die. By the LLN, in a sufficiently long sequence of such integers, the proportion of 2‘s converges to $P(X = 2) = 1/6$ and the running averages $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converge to $E(X) = 3.5$. Estimates improve as n increases.

In R, **`d=sample (1:6,n,repl=T)`** simulates n rolls of a fair die. In one run with $n = 1000$, the additional code **`mean (d==2)`** returned 0.184 ± 0.023 and **`mean(d)`** returned 3.513 ± 0.106. By contrast, results from a run with $n = 1\,000\,000$ were much closer to the theoretical values: 0.166985 ± 0.00073 and 3.499451 ± 0.00335, respectively. With current computers, it is often feasible to use a million iterations in a simulation. ◇

Role of the central limit theorem (CLT). In Example 1, the logical n-vector **`d==2`** has TRUEs and FALSEs, as elements, and its **`mean`** is interpreted as its proportion of TRUEs. With $n = 1000$, the CLT approximates the 95% margins of simulation error for $P(X = 2)$ as

$$P(X = 2) \text{ as } \sqrt{\frac{1}{6}\left(\frac{5}{6}\right)/n} = 0.023,$$

and for E(X) the margin error is $1.96\sqrt{\frac{105}{36n}} = 0.106$ because V(X) = 105/36.

Historical Background

Manipulating Physical Objects. The earliest way to generate "random" numbers for simulation was the straightforward use of physical devices: tossing coins, rolling dice, shuffling decks of numbered cards, and so on. For example, repeated tossing of an ordinary coin may simulate a sequence of independent outcomes, each equally likely to be 0 or 1. However, such physical manipulations must be carefully done in order to get the intended random results (Bayer and Diaconis 1992; Diaconis et al. 2007).

Drawing from High-Entropy Pools. Random numbers have also been generated from *high entropy pools*, sources of "noise" that can be digitized. These have included static on an unused radio frequency, noise in an electronic circuit, and times of human interactions with a keyboard.

Circuit noise was used to generate a million random digits, published in book form by the Rand Corporation (1955) and made available on computer punch cards. These numbers pass various tests for randomness and uniform distribution across the integers 0 through 9. They were suitable for randomizing subjects to experimental groups and for some early simulations, but many modern probability simulations use random numbers at a rate that requires a faster method of generation.

Computer Algorithms. Credit for being the first to use arithmetic algorithms on a computer to generate pseudorandom numbers (PRNs) is often accorded to Stanislaw Ulam and John von Neumann, working together on the mechanics of nuclear fission in the late 1940s. Computers had just then become fast enough that it seemed feasible to use computer algorithms for simulation. Along with others, they planned one of the first computer simulations on probability models of neutron diffusion, as a substitute for doing very difficult mathematical derivations. Some of von Neumann's early attempts to get useful random numbers by arithmetical algorithms were unsuccessful, and at a conference in 1951 he joked that anyone trying this was "living in a state of sin" (von Neumann 1951; Dyson 2012).

However, a lot of "sinning" has gone on since 1951, and it seems that von Neumann was too quick to deprecate the use of algorithms for simulation. Certainly, it is not possible to generate truly random numbers with a computer algorithm. But by working collaboratively through the years, statisticians, mathematicians, and computer scientists have learned that – by being clever and careful – it is possible to generate useful PRNs with algorithms.

Linear Congruential Generators. One method of generating PRNs by algorithm is a LCG, which uses modular integer arithmetic. Let

$$r_{i+1} = ar_i + b \pmod{d},$$

for $i = 1$, 2 , 3, where r_i , a , b , d are integers, a , d , $r_1 > 0$; $b \geq 0$. (That is, r_{i+1} is the remainder when $ar_i + b$ is divided by d.) Then $r_1 < d$ is called the seed and d the modulus of the generator. The period $p \leq d$ of a generator is the number of distinct values it can produce. One important requirement is that a useful generator must have a very large period. Often d is chosen to be a very large prime number. PRNs in the interval (0, 1) are defined by a relationship such as $u_i = (r_i + 1/2)/d$.

Example 2: A trivial linear congruential generator. For a one-time shuffling of a deck of cards, consider a LCG with $a = 20$, $b = 0$, $d = 53$, and $s = 21$, which turns out to have period $d = 52$, scrambling the numbers from 1 through 52. With these constants, the following R program generates a sequence of 60 digits r_i, showing repetition after 52 distinct numbers are produced.

```
r = numeric(60); r[1] = s
for (i in 1:59) {
  r[i+1] = (a*r[i] + b) %% d }
r
```

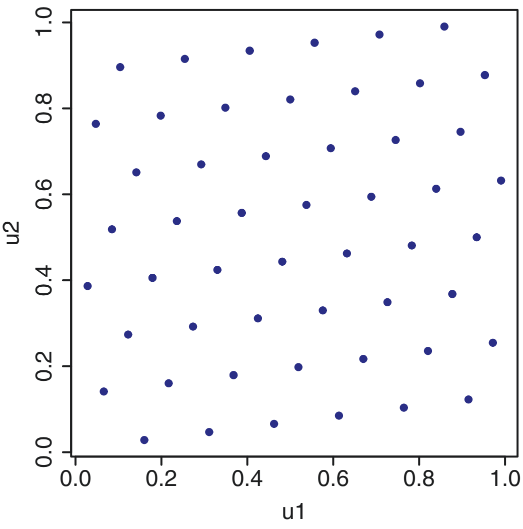
With 13 numbers in each of the first four rows, results are:

```
21 49 26 43 12 28 30 17 22 16  2 40  5
47 39 38 18 42 45 52 33 24  3  7 34 44
32  4 27 10 41 25 23 36 31 37 51 13 48
6  14 15 35 11  8  1 20 29 50 46 19  9
21 49 26 43 12 28 30 17
```

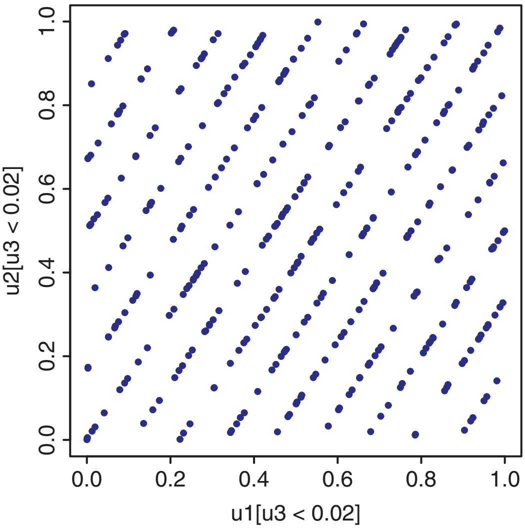
The first iteration finds $20(21) + 0 = 420$; then $420 \pmod{53} = 49$. By using a different seed we could start the sequence with another number between 1 and 52, but the order after that would be the same.

When the vector $\mathbf{u_1} = (u_1, \ldots, u_{59})$ is plotted against $\mathbf{u_2} = (u_2, \ldots, u_{60})$, the ideal result would be dots filling the unit square at random. But LCGs always produce grid patterns. So the goal is for the period to be long enough and the grid to be fine enough that the grid structure is not noticeable. Figure 1 shows the unsatisfactory 2-D plot of this generator. ◇

Example 3: RANDU. The generator RANDU, released in the 1950s for use on IBM mainframe computers, is essentially a LCG with $a = 65539$, $b = 0$, and $d = 2^{31}$. A 2-D plot of



Simulations, Fig. 1 A 2-D plot of PRNs from the trivial congruential generator of Example 2. The grid is about as fine as possible with a period of only $d = 5$



Simulations, Fig. 2 A thin veneer from the front face of a 3-D plot of 20 000 PRNs from the RANDU generator of Example 3. Triplets lie on a few widely separated parallel planes, as shown by this plot of about 400 of them from near the front face

several thousand of its PRNs looks random. However, after several years of widespread use, it was discovered that some simulations based on RANDU gave wrong answers.

A major difficulty is that its 3-D plot has all of its points on about 15 widely separated parallel planes. To illustrate this in two dimensions, Figure 2 shows a thin veneer from the front face of a 3-D plot. One can see where the parallel planes intersect this face of the cube. One says that its PRNs are *equidistributed* in two dimensions, but not in three (Lewis 1986). ◇

Several congruential generators have survived many benchmark tests and are widely used (Gentle 1998).

Mersenne Twister. Matsumoto and Nishimura (1998) introduced the Mersenne twister generator, which has a period of $2^{19937} - 1$ (a Mersenne prime number). Results are equidistributed in 623 dimensions regardless of the starting seed. This generator, based on machine-language recursions, is not a congruential generator. As of 2017, the Mersenne twister is the default or recommended generator in R, SAS, SPSS, and other statistical software. Our simulations use the implementation in R for Windows.

Example 4: Setting seeds for reproducibility. Each time an R session is initiated, a new seed is created for the Mersenne twister. This is done in an essentially unpredictable way by using trailing digits of the current time according to the system clock. The statement **`runif (10)`** simulates ten independent observations from UNIF(0,1). To illustrate the variability of random results, we made two runs with **`mean(runif(10))`**, which gave means 0.5335309 and 0.4577399, respectively; about 95% of such results should be within 0.5 ± 0.18.

However, if we *choose* a seed before a simulation, then we always get the same results. This can be useful in debugging programs or for sharing analyses that involve simulation. For example, if we use **`set.seed(1234)`**, immediately followed by **`mean(runif(10))`**, the result will always be 0.4892264. ◇

Simulating Nonuniform Distributions

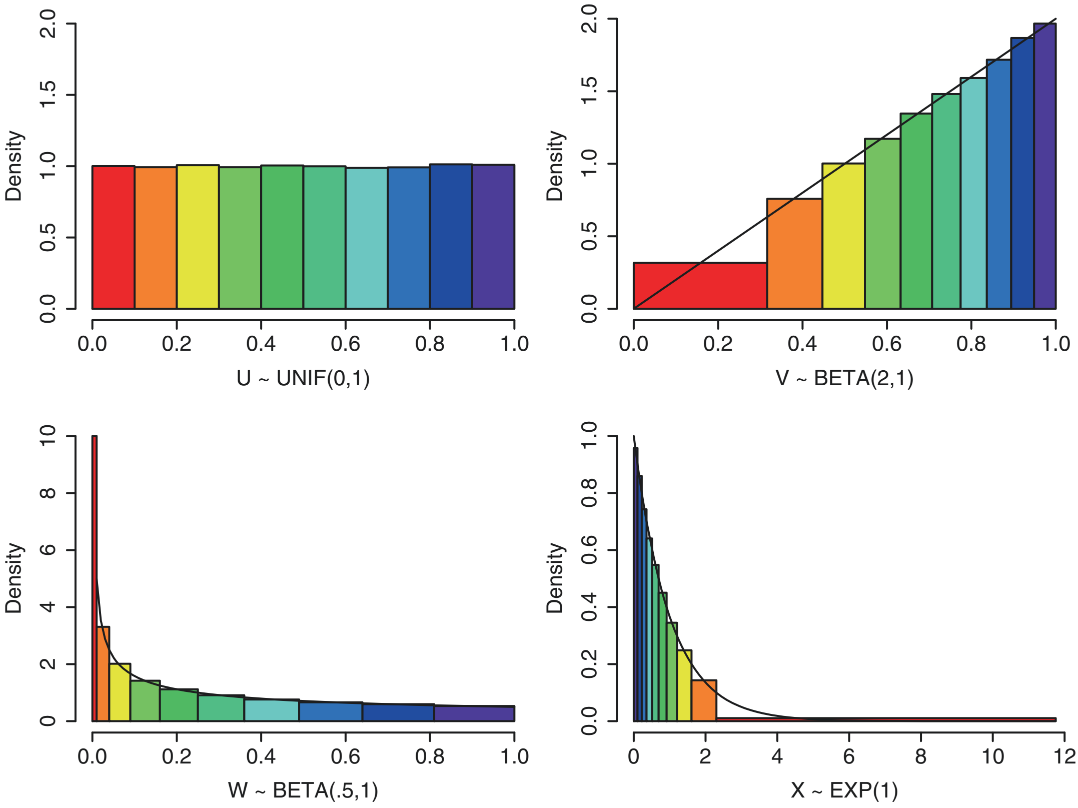
Many methods of generating random variables from various distributions have been discussed in books (e.g., Gentle 1998) and journal articles. We give a few examples of commonly used ones.

Inverse CDF Method

Let the random variable X have CDF $F_X(x) = P(X \leq x)$, for real x, with inverse $F_X^{-1}(u)$, where $0 \leq u \leq 1$, that can be written in a useful form. With $U \sim$ UNIF(0, 1), the random variable $F_X^{-1}(U)$ has the same distribution as X. Thus we can use a random sample U_i, for $i = 1, \ldots, n$, to simulate a random sample of n observations having the distribution of X.

Example 5: Simulating two beta distributions. First, let $V \sim$ BETA(2, 1) with CDF $F_V(v) = v^2$, for $0 < v < 1$. Then $F_V^{-1}(U) = \sqrt{U}$, where $U \sim$ UNIF(0, 1), has the distribution BETA(2, 1). The upper-right panel of Figure 3 shows a histogram of 100 000 realizations of V simulated in this way.

For verification, the R statement **`v=sqrt (unif(10^6))`** simulates a million observations from BETA(2, 1). Then **`mean(v)`** finds their sample mean, returning 0.66644 ±



Simulations, Fig. 3 *Upper left*: Histogram of 100 000 PRNs used for three quantile-method transformations in Examples 5 and 6. *Upper right*: Histogram of $V \sim$ BETA (2, 1). *Lower panels* (different scales) illustrate transformations to obtain $W \sim$ BETA(.5, 1) and $X \sim$ EXP (I). Each bar in each of the four plots represents about 10 000 observations; like colors show images of the same PRNs

0.0005. Similarly, **sd(v)** returns their standard deviation 0.23568. These values are in good agreement with results from calculus for the corresponding population moments: $E(V) = 2/3$ and $SD(V) = \sqrt{1/18} = 0.2357$.

Second, following the same pattern, realizations from the distribution BETA(.5, 1) can be simulated using $W = U^2$, as illustrated in the lower left panel of Figure 3. ◇

Different methods of simulation are used for some members of the beta family. Most statistical computer packages have procedures for simulating distributions from frequently used families. For example, in R the statement **rbeta (n, shape1, shape2)** simulates a random sample of size n from a beta distribution with the designated shape parameters.

Example 6: Simulating EXP(1). In order to simulate $X \sim$ EXP(1), we invert the CDF $F_X(x) = 1 - e^{-x}$, for $x > 0$, to obtain

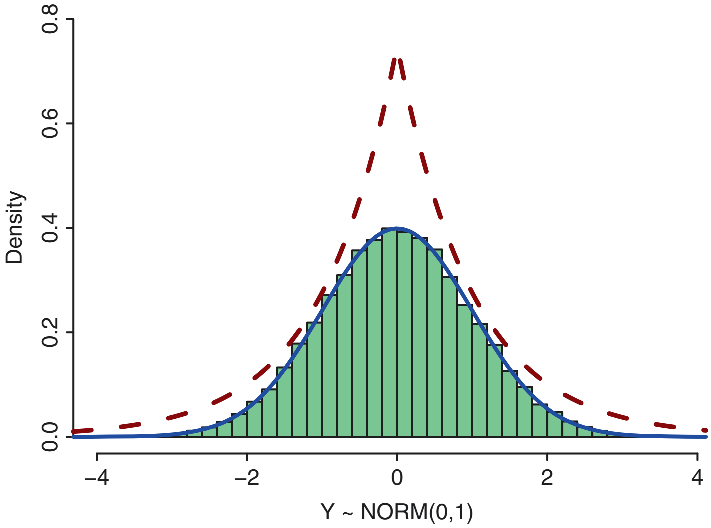
$$X = -\log(1 - U') \sim \text{EXP}(1)$$

where $U' \sim$ UNIF(0,1). However, also $1 - U' = U \sim$ UNIF(0, 1), so we simplify simulation of X as $-\log(U) \sim$ EXP(1). Thus small values of U produce large values of X, which accounts for the reversal of the colors in the lower-right panel of Figure 3. ◇

Acceptance-Rejection Methods

A simple acceptance-rejection method generates independent values of Y with density function $f_Y(t)$, for t in the support of Y. Suppose that $f_X(t)$ is the density function of a random variable X with

Simulations, Fig. 4 Using the acceptance-rejection method of Example 7, 40 052 standard normal observations, shown in the histogram, were accepted from among 60 000 random observations distributed LAPLACE(0, 1)



the same support as Y. To use this method, we must know how to obtain random observations of X and there must be a constant $K > 0$, such that $f_Y(t) \leq Kf_X(t)$, for all t in the support of X and Y. We call $Kf_X(t)$ a majorizing function of $f_Y(t)$. Then $p(t) = f_Y(t)/Kf_X(t)$, with $0 \leq p(t) \leq 1$, is the acceptance probability at t. Corresponding to the value X generated from f_X, we accept $Y = X$ as a generated value of Y with probability $p(X)$; otherwise, we reject (do not use) that value of X.

Example 7: Simulating normal via Laplace. If we have $Q_1,\ Q_2 \sim$ EXP(1), independently, then $X = Q_1 - Q_2 \sim$ LAPLACE(0, 1), with $f_X(t) = 0.5e^{-|t|}$. It is easy to see that $f_Y(t) = \varphi(t) \leq 1.5\ f_X(t)$, where $K = 1.5$ and φ is the standard normal density function. Then the following program, with $m = 60\ 000$ iterations, simulates a sample of about $m/K = 60\ 000/1.5 = 40\ 000$ observations from NORM(0, 1).

```
set.seed(1234); m = 60000
Q1 = rexp(m); Q2 = rexp(m)
x = Q1 - Q2 # Laplace
maj = 1.5*.5*exp(-abs(x))
p.acc = dnorm(x)/maj
acc = p.acc > runif(m)
y = x[acc==1]; mean(acc)
```

Figure 4 shows a histogram of the accepted Y values, with good fit to the standard normal density function. The majorizing function is also shown as a dashed curve. The inflation constant $K = 1.5$ could have been a little smaller, giving a little larger acceptance rate. ◇

The method illustrated in Example 7 can be used quite generally and in any number of dimensions. But sometimes, especially in higher dimensions, it can be difficult to find an efficient majorizing function.

In that case, the Metropolis-Hastings algorithm can be used. Points $\{X_i\}$ are programmed as a random walk through the support of the target random variable Y. At step i, a candidate value X_i is proposed as the next value Y_i. Roughly speaking, according to a specific criterion, results from values X_i that lie in relatively high-density regions of Y tend to be accepted. If a candidate X_i lies in a relatively low-density region, it is more likely rejected, so $Y_i = Y_{i-1}$ and $X_i = X_{i-1}$ (Metropolis et al. 1953; Hastings 1970; Chib and Greenberg 1994).

Special Methods for Normal Distributions

We discuss three methods to simulate random samples from the standard normal distribution NORM(0, 1).

Example 8: Using the Central Limit Theorem. For appropriate U_i and sufficiently large n, the sum $X = \sum_{i=1}^{n} U_i$ is approximately normally distributed. If $U_i \sim^{iid}$ UNIF(0, 1), convergence is quite rapid, so $n = 12$ gives values of X that are difficult to distinguish from NORM(6, 1). Thus the result of the R code **`sum(runif(12))-6`** can be taken as an observation from NORM(0, 1).

This method was widely used on early computers, for which addition and subtraction were by far the fastest operations. A disadvantage is that such Xs are not *exactly* normal; one fault is that generated Xs are always in $(-6, 6.)$ ◇

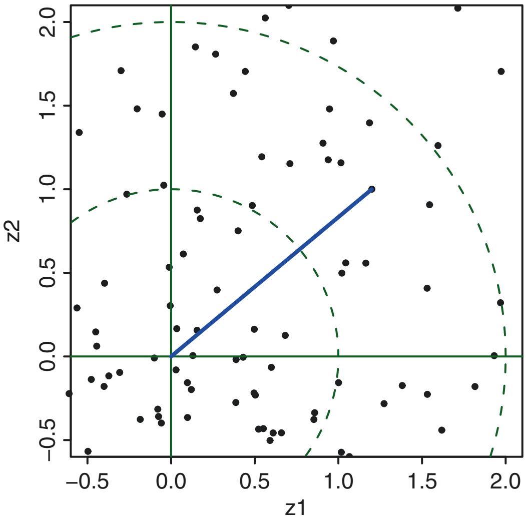
Example 9: Box-Muller transformation. Choose a random point $(Z_1, \ Z_2)$ in the plane with independent $Z_1, \ Z_2 \sim$ NORM(0, 1) and denote its squared distance from the origin (0, 0) as $D^2 = Z_1^2 + Z_2^2$. Then $D^2 \sim$ CHISQ(2) $\equiv$ EXP(1/2), which can be simulated as $-2\ \log(U)$, where $U \sim$ UNIF (0, 1).

Independently, let Θ be the angle from the positive z_1-axis counterclockwise around to the line from (0, 0) to $(Z_1, \ Z_2)$. Then $\Theta \sim$ UNIF $(0, 2\pi)$, so that both D and Θ can be simulated from independent $U_1, \ U_2 \sim$ UNIF(0, 1). See Figure 5.

Finally, converting from polar coordinates $(D, \ \Theta)$ to rectangular coordinates $(Z_1, \ Z_2)$, we have the Box-Muller transformation for generating standard normal observations (Box and Muller 1958).

$$Z_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2),$$
$$Z_2 = \sqrt{-2\log(U_1)}\sin(2\pi U_2).$$

The transcendental functions in the Box-Muller transformation run acceptably fast on modern computers, and several slightly different versions of the method are widely used. Theoretically speaking, results are exactly normal, but because of the limitations of floating point arithmetic, they are truncated to a finite interval, often about $(-7, 7)$. Roughly speaking, the smallest values of U_1 produce the largest absolute values of Z_1 and Z_2, and there is a smallest value of U_1 that avoids underflow. ◇



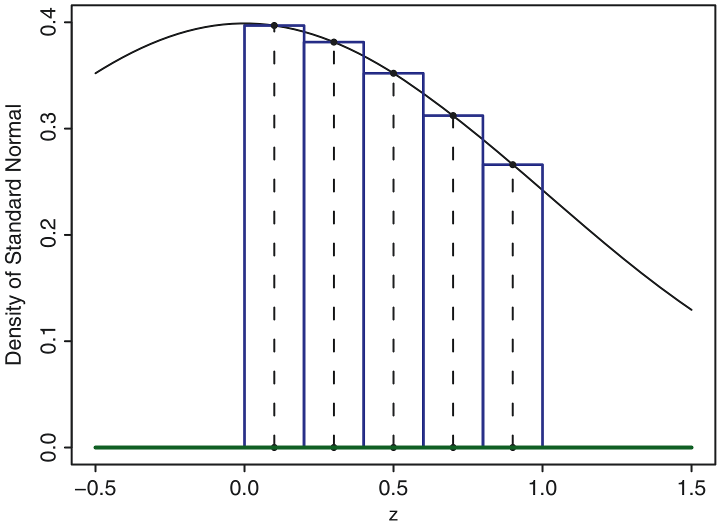
Simulations, Fig. 5 Hits on a target with standard bivariate normal displacements from the origin illustrate the Box-Muller transformation of Example 9. The point at about $Z_1 = 1.2$ and $Z_2 = 1.0$ is at distance $D = 1.56$ from the origin and at an angle $\Theta = 39.8$ degrees from the positive z_1-axis

Example 10: Rational approximation of CDF. While the CDF Φ of the standard normal distribution cannot be written in closed form, it is possible to find piecewise rational functions that approximate Φ and Φ^{-1} well. Wichura (1988) introduced a rational approximation of Φ^{-1} that is about as accurate as double-precision arithmetic can represent. It is used in R for **`qnorm`** and is the default method of generating samples from normal populations. Both **`set.seed(1215);rnorm(1)`** and **`set.seed(1215);qnorm(runif(1))`** return 0.3129521. ◇

Monte Carlo Integration

Simulations can approximate integrals used in a variety of probability models. We show two methods of approximating the integral $J = \int_0^1 \varphi(x)\,dx$, where $\varphi(x) = (2\pi)^{-.5} e^{-.5x^2}$ The CDF $\Phi(x)$ is called **`pnorm (x)`** in R. The value $J = 0.3413$ can be found from printed tables, and the code **`diff(pnorm(c(0,1)))`** returns 0.3413447.

Simulations, Fig. 6 Riemann approximation as in Example 11. The heights of the five rectangles are $\varphi(g_i)$ with $g = (0.1, 0.3, 0.5, 0.7, 0.9)$, so bases are of width $w = 0.2$. The total area of the rectangles is 0.3417, which closely approximates $J = \int_0^1 \varphi(z)dz = 0.3413$. In Example 12, the g_i are randomly chosen



Example 11: Riemann approximation. Historically, normal tables were made by summing areas of rectangles (or trapezoids) that approximate appropriate areas beneath $\varphi(x)$, as suggested by the definition of the Riemann integral. Increasing the number of rectangles provides greater accuracy for this strictly *deterministic* procedure. Figure 6 shows how five rectangles can be used to get $J \approx 0.3417$. The brief R program below, with $a = 0$, $b = 1$ and $m = 10\ 000$ rectangles of width $1/m$, gives an improved result.

```
w = (b-a)/m
g = seq(a+w/2, b-w/2, len=m)
h = dnorm(g); sum(w*h)
```

Grid points g_i are located at the centers of the bases of the m rectangles, and their respective heights are $\varphi(g_i)$. The result is $J = 0.3413447$. ◇

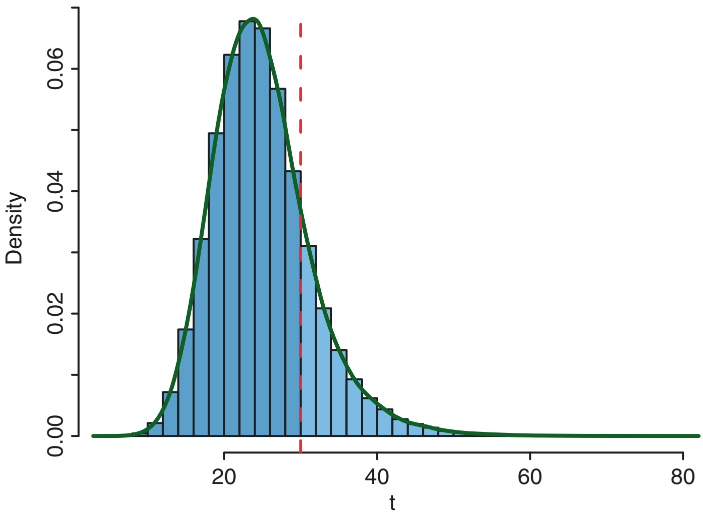
Example 12: Basic Monte Carlo integration. Remarkably, one can successfully approximate J by a *random* procedure, structured similarly to the program in Example 11. Instead of using a deterministic grid of equally spaced values g_i, we use random values from the distribution UNIF(a, b). The R code **runif(m,a,b)** produces m such values. So the essential program change is that the second line of the program above is replaced by **g=runif(m, a, b)**, and **w** is interpreted as the average distance between g_i.

Due to the randomness of the g_i, the results differ from one run to the next, and we use $m = 100\ 000$ iterations for improved accuracy. Results from two successive runs are $J \approx 0.3414103$ and 0.3413173. The margin of simulation error is 0.0003. ◇

The method of Example 12 can also be used to find expectations. For example, let $W \sim \text{BETA}(2, 1)$, with density function $f_W(w) = 2w$, for $0 \leq w \leq 1$. Then $\text{E}(W) = \int_0^1 wf_W(w)\,dw = 2/3$ could be approximated by basic Monte Carlo integration, and similarly for integrals not so easily evaluated by calculus.

Example 13: Monte Carlo in higher dimensions. In one dimension, basic Monte Carlo integration is seldom more accurate than deterministic Riemann integration with the same number m of iterations. But when the domain of integration has three or more dimensions, Monte Carlo integration is often better, and the method is widely used (Liu 2004).

Roughly, the advantage of the Monte Carlo method in higher dimensions is explained as follows. In Riemann approximation over an interval, the only "ragged edge" of the area being approximated is where the rectangles meet the density



Simulations, Fig. 7 The simulated distribution of the total time T needed for a task is found by the Monte Carlo sampling method of Example 14. The region corresponding to $P(T < 30) \approx 0.811$ is emphasized. The *curve* is a density estimator (smoothed histogram) of the simulated distribution

curve. In multidimensional Riemann approximation, ragged boundaries can proliferate to spoil accuracy. By contrast, points scattered at random often fit more usefully into a multidimensional region of integration. Here are two examples for which exact answers are readily available.

(a) Use 100 000 uniformly distributed points to integrate the trivariate uncorrelated standard normal density over the unit cube with vertices (0, 0, 0) and (1, 1, 1). The result of one run was 0.03979981, close to the exact value $J^3 = 0.03977220$.

(b) To integrate this trivariate density over the first octant of the unit ball, we generated points uniformly at random in the unit cube as above and ignored points not precisely within the ball. One run with only about 32 000 accepted points gave 0.02486239, compared with the exact value 0.02484351 from **pchisq(1,3)/8** . ◇

Often a probability model is summarized by a random variable, say T, with an unknown density function, so basic Monte Carlo integration is not useful. However, if we know how to simulate constituent parts of the model, we can use a sampling method to investigate properties of the model.

Example 14: Monte Carlo sampling. Suppose a computer procedure has three independent random components that can be modeled as: $U \sim \text{UNIF}(0, 10)$, $V \sim \text{EXP}(1/5)$, and $W \sim \text{NORM}(15, 3)$ in msec. The total time for the procedure is $T = U + V + W$ and we want to know $P(T < 30)$.

It would be tedious to find the density function of T, but we know how to sample from each of the three constituent distributions involved, so we can simulate 100 000 realizations of T as follows.

```
set.seed(1789); m = 10^5
u = runif(m, 0, 10); v = rexp(m, 1/5)
w = rnorm(m, 15, 3); t = u + v + w
mean(t < 30); mean(t); sd(t)
```

From the last line of code we obtain $P(T < 30) = 0.81116 \pm 0.0025$. As a check that the programming is correct and that the PRNs used in **rexp** and **rnorm** are reliably generated, it is easy to see that $E(T) = 5 + 5 + 15 = 25$ exactly, while the simulated value is 24.997 ± 0.042. Figure 7 shows the simulated distribution of T . ◇

Simulation of Probability Models

In practice, several conditions should be met for the successful simulation of the distribution of a random variable X.

One must be able to:

- Simulate single instances that give rise to X.
- Find an automated way to determine the value of X in each such instance.
- Write a program to generate many realizations of X thus approximating its distribution.

Example 15: Item matching. Consider $n = 12$ account statements ("bills") and 12 envelopes, a proper one for each bill. A clerk inserts the bills into envelopes at random. We use simulation to approximate the distribution of the number X of bills that happen to be put into proper envelopes.

This matching problem is discussed in introductory probability texts (e.g., Blitzstein and Hwang 2015; Feller 1957; Ross 1997). Results known from probability theory include: $\mathrm{E}(X) = \mathrm{SD}(X) = 1$ and $P(X = 0) = \sum_{i=0}^{n} \frac{(-1)^i}{i!}$, for $n \geq 2$. For $n = 12$, we have $P(X = 0) = 0.367879$. Trivially, $P(X = 11) = 0$ and $P(X = 12) = (12!)^{-1}$. So we have some exact results for comparison with simulated ones.

For simulation, the **sample** function in R uses pseudorandom numbers to select from a given population (first argument) a sample of a given size (second argument). The default sampling process is without replacement. If the sample size equals the population size, then the result is a random permutation of the population elements.

We use **1:12** for the vector of integers ordered from 1 through 12 along with sample to obtain one scrambling **p** of bills into envelopes. Then **sum (1:12==p)** counts the number X of bills in proper envelopes. When sum is applied to a logical vector, the result is the number of TRUEs. In the instance shown, the three bills numbered 6, 8, and 10 happened to be properly placed.

```
p = sample(1:12, 12); p
## 4 11 7 12 3 6 5 8 2 10 1 9
1:12 == p # logical vector
#[1] FALSE FALSE FALSE FALSE FALSE TRUE
#[7] FALSE TRUE FALSE TRUE FALSE FALSE
x = sum(1:12 == p); x
## 3
```

By looping through 1 00 000 such random permutations and recording the value of X for each, we get a very good approximation to the distribution of X.

```
m = 10^5; x = numeric(m); n = 12
for (i in 1:m) {
p = sample(1:n, n)
x[i] = sum(1:n == p) }
table(x)/m
```

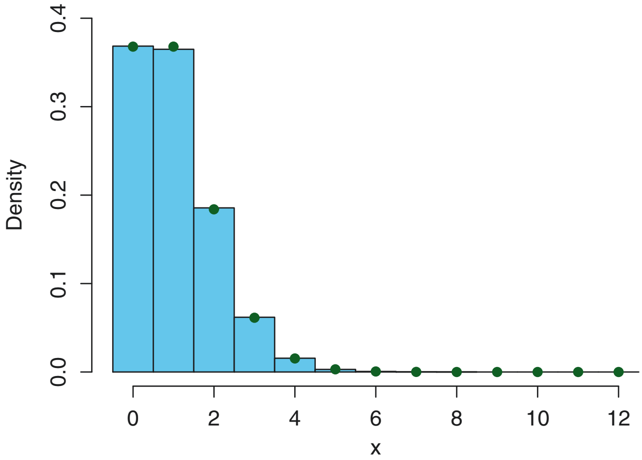
The last line prints probabilities in the distribution of X, accurate to about two places. Specifically, with **set.seed(1234)**, $P(X = 0) \approx 0.36847$. Also, R statements **mean(x)** and **sd(X)** find the sample mean $\overline{X} = 1.00239$ and $S = 1.003759$, which are good approximations of $\mathrm{E}(X) = 1$ and $\mathrm{SD}\ (X) = 1$, respectively. See Figure 8.

Individual probabilities in the distribution of X are accurate to about two places, but the *relative* error may be large for very small probabilities. Although $P(X = 12) = 1/12! \approx 2.1 \times 10^{-9}$, our simulation run happened to give no instances where more than eight bills went into proper envelopes. ◇

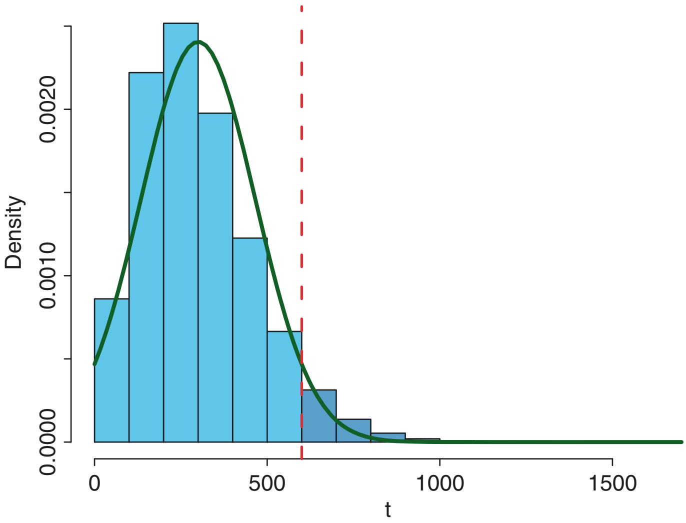
Example 16: Random sum of random variables. Suppose experience has shown the time required to complete a particular kind of search is $X \sim \mathrm{EXP}\ (\beta = 1/50)$, an exponential distribution with rate β and with mean $\mathrm{E}(X) = \mu_X = 1/\beta = 50$ units of time and standard deviation $\mathrm{SD}\ (X) = \sigma_X = 1/\beta = 50$.

Also suppose a compound search task is estimated to require $N + 1$ such searches, where $N \sim \mathrm{POIS}\ (\lambda = 5)$, a Poisson random variable with $\mathrm{E}(N) = \lambda = 5$. Then the total length of time for the compound search is $T = \sum_{i=1}^{N+1} X_i$, where all of the X_i and N are independent. Standard probability formulas give $\mu_T = (\lambda + 1)\mu = 300$, and $\sigma_T^2 = \mathrm{E}(N+1)\sigma_X^2 + \mathrm{V}(N+1)\mu_X^2$, so $\sigma_T = \mu\sqrt{2\lambda + 1} = 165.8312$.

Simulations, Fig. 8 Histogram of the simulated distribution in Example 15 of the number X of bills that happen to go into their proper envelopes. As shown by dots atop bars, X is approximately POIS($\lambda = 1$)



Simulations, Fig. 9 As in Example 16, the simulated distribution of the time T to complete the compound search task. From a million iterations, $P(T > 600) \approx 0.0535 \pm 0.0004$. The approximation NORM(300, 165.64) underestimates right tail probabilities (*solid density curve*)



We want to know the probability that such tasks typically take more than 600 time units to complete. The key statements within the loop of a simulation program are **`n = rpois(1,5)`** and **`t[i] = sum(rexp(n+1, 1/50))`**. The simulation estimates $P(T > 600) \approx 0.0535 \pm 0.0004$. Other simulated results closely match theoretical values: $\mu_T \approx 299.68$ and $\sigma_T \approx 165.64$. See Figure 9. ◇

Example 17: Shuffling cards for randomness. In recreational card games, people often shuffle a deck of cards only two or three times between hands. An analysis based on simulated shuffling shows that this is not sufficient to put the cards into random order (Bayer and Diaconis 1992).

Numbering the cards in a standard deck from 1 through 52, we can use the R statement **`d=sample(1:52, 52)`** to make a random permutation. Here is an example of such a randomized deck.

```
18  6 34 42 52 20 29 41 25 11 14 50 40
37 43 26 19  2 30 16 21 51 31 27 28 17
 3 38  9 39 49  4 10 24 15 13 23 45 48
 7  8 44 22 33  1 36 32 47  5 12 35 46
```

This works well for computer games and simulations, but in a real card game randomizing a deck according to computer output might be awkward.

For a manual shuffle, the dealer cuts the deck somewhere near the middle into two heaps and "riffles" them together into the new order. Starting with cards in order from 1 through 52, we simulate a manual shuffle as follows: (a) Cut the deck into two roughly equal heaps: the left heap with the first N cards, where $N \sim$ BINOM(50, 0.5) + 1, and the rest of the cards in the right heap. Usually, there will be between 20 and 32 cards in the left heap. (b) To simulate the riffle, we randomly choose N out of 52 positions for the cards in the left heap (in order) and put the cards from the right heap into the empty positions. One simulated shuffle might give the result shown below.

```
26 27 28 29  1  2 30 31  3  4 32  5 33
 6 34  7 35  8 36  9 37 10 38 11 39 12
13 14 40 15 41 16 17 42 43 18 44 19 20
45 46 47 48 49 21 50 51 22 23 24 25 52
```

By looking at rising sequences we can see that the deck is not in random order. We say that there are two rising sequences in this arrangement. The first rising sequence from the left heap (emphasized by underlining) starts with cards 1 and 2 and ends with 25. "Jumping back" to the beginning of the shuffled deck, we see that the second rising sequence begins with 26 and ends with 52.

There will always be two rising sequences after a single shuffle (cut and riffle). There will usually be four rising sequences after two shuffles, but perhaps fewer, if two short rising sequences happen to merge into one longer one. By contrast, the random sequence from the **`sample`** function has 27 rising sequences. This is a typical result for a properly randomized deck. (The highest possible number of rising sequences is 52, for a deck in reverse order: (52, 51, . . ., 3, 2, 1).)

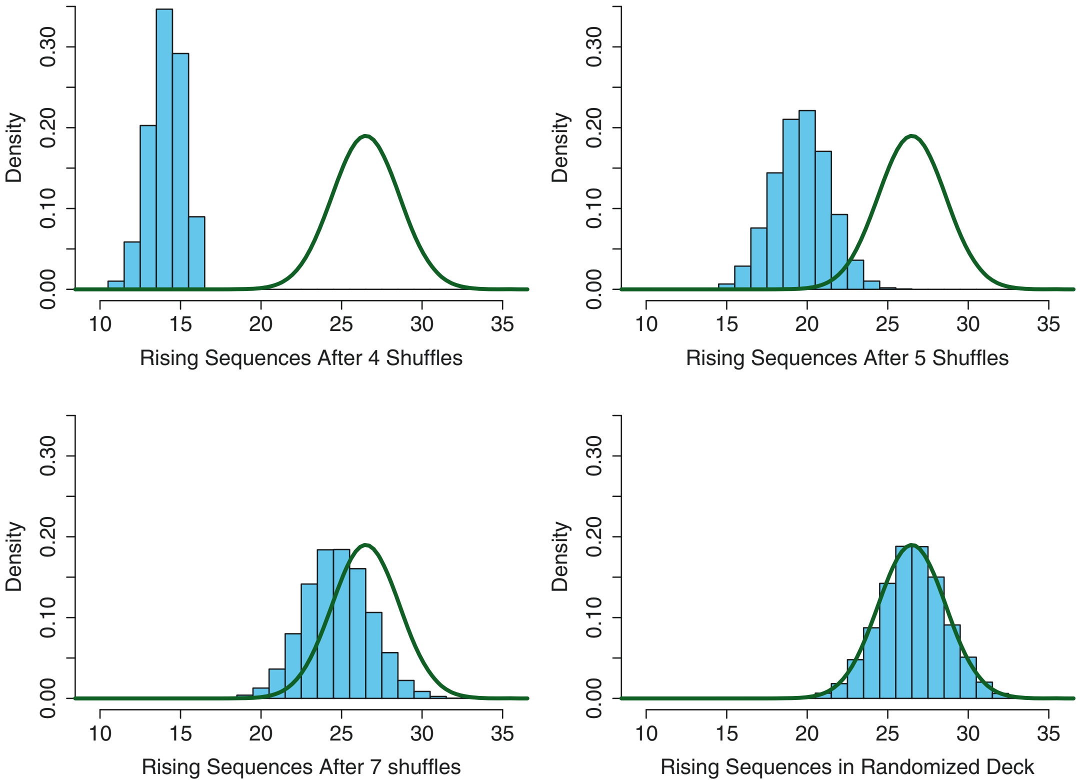
Each time we must "jump back" to get the next number in sequence, that is the beginning of a new rising sequence. Using **`d`** for the earlier randomized deck, we counted its rising sequences **`x`** with statements **`jb=diff(match(1:52,d))<0`** and **`x=sum(jb)+1`**.

By simulating 10 000 random decks, we find that the average number of rising sequences is 26.5 and about 95% of random decks have between 22 and 31, as illustrated in the lower-right panel of Figure 10. In a card game, the purpose of shuffling cards is to keep players from having an unfair advantage by exploiting information stored in the deck from previous play. Figure 10 compares the distributions of rising sequences after four, five, and seven riffle shuffles with the distribution for truly random decks. Shuffling seven times is not quite enough to put it a deck random order. However, as shown in the figure, it seems good enough to thwart players seeking an advantage by memorizing card order from previous play. (See Grolemund and Wickham (2014) for related simulations.) ◇

Example 18: Fairness of a die. We wonder whether a die is fair and roll it $n = 600$ times, observing faces 1 through 6 with frequencies $\mathbf{X} = (113, 92, 91, 94, 96, 108)$, respectively. Under the null hypothesis H_0 that each face has probability $\frac{1}{6}$, the six expected frequencies are $E_i = n\frac{1}{6} = 100$. The usual goodness-of-fit (GOF) statistic is

$$Q = \sum_{i=1}^{6} \frac{(X_i - E_i)^2}{E_i} = 4.3,$$

for which *small* values indicate relatively good fit. Under the null hypothesis, we have Q approximately distributed as CHISQ(5), the chi-squared distribution with degrees of freedom



Simulations, Fig. 10 Histograms of distributions of rising sequences in a deck of cards after 4, 5, and 7 shuffles. In each panel, the curve shows the distribution for truly randomized decks. Six or 7 shuffles should be enough to degrade useful information about the order of the cards from previous play. See Example 17

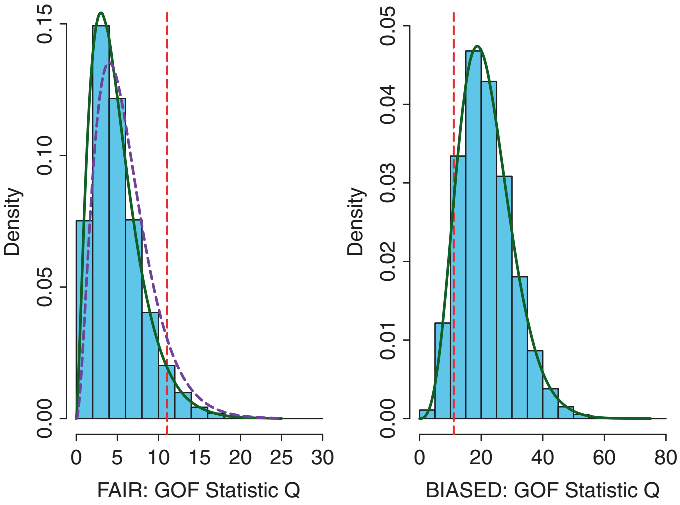
$\nu = 6 - 1 = 5$. The critical value for a test at the 5% level of significance is $q^* = 11.0705$; that is, $P\{Q \geq 11.0705 | H_0\} = 0.05$. Because $Q < q^*$, data from our 600 rolls seem consistent with a fair die.

The continuous distribution CHISQ(5) must be an approximation of the null distribution of Q because Q inherits discreteness from integers X_i. One may wonder how good the approximation is, and whether it is correct to use $\nu = 5$ instead of $\nu = 6$.

We simulate the exact distribution of Q by using the random function **`sample`** to "perform" 100 000 experiments, each with 600 rolls of a fair die, and computing Q for each experiment. The proportion of these simulated values of $Q > q_U = 11.07$ is 0.04991 ± 0.0014. The left panel of Figure 11 shows a histogram of the simulated distribution of Q along with the density curve of CHISQ (5) (which fits) and of CHISQ (6) (which does not).

In this example, there are six categories and 600 rolls of the die, so the expected counts are $E_i = 100$. If some expected counts are less than 5, the GOF statistic Q may not be satisfactorily approximated by a chi-squared distribution. Then we should use simulation to approximate the exact distribution of Q.

Beware of fits "too good to be true." If we are vetting a PRN generator or the programming of the sample function, the chi-squared test becomes two-sided. We need to consider whether the



Simulations, Fig. 11 At *left*: The simulated null distribution of the GOF statistic Q in Example 18. The histogram closely fits the CHISQ(5) density (*solid curve*), but not the CHISQ(6) density (*broken curve*). At *right*: The simulated alternative distribution of Q in Example 19 is closely matched by the density function of CHISQ (5, 16.67); the power of the test against the bias specified in the example is about 90%. Both panels: The vertical line (at 11.07) indicates the critical value of a test at the 5% level

results have a truly random structure. In the same way that we might be suspicious of fudged data if someone reports frequencies (100, 100, 101, 99, 99, 101) for the numbers seen on a die after 600 rolls of a die, we should also be suspicious if we see a high proportion of iterations with $Q < q_L = 1.1455$ because the true probability of that is less than 5%.

Example 19: Power of a GOF test. If H_0 is not rejected in Example 18, it is reasonable to ask whether the GOF test is powerful enough to detect serious bias. Specifically, imagine a die weighted so that faces 1, 2, and 3 each appear with probability 5/36 and the other three faces with probability 7/36. Can the GOF test detect this level of bias? To repeat the previous simulation, but with this biased die, we use the following R code for each of the 100 000 600-roll experiments.

```
bias = c(5, 5, 5, 7, 7, 7)/36
sample(1:6, 600, rep=T, prob=bias)
```

In a run of this modified program, the proportion of simulated values of $Q \geq 11.0705$ is 0.9039 ± 0.0018. Thus the power of the GOF test under an alternative with this amount of bias is 90.4%, and we have a reasonably good chance to detect our weighted die is unfair.

The histogram in the right panel of Figure 11 shows the simulated distribution of Q for such a biased die. For this elementary model, it is natural to assume that the alternative distribution is CHISQ($v = 5$, $\lambda = 16.67$), the *noncentral* chi-squared distribution with degrees of freedom $v = 5$ and noncentrality parameter $\lambda = 16.6667$, which is computed as

$$\lambda = n \sum_{j=1}^{6} \left(p_{0j} - p_{bj}\right)^2 / p_{0j},$$

where $p_{0j} \equiv \frac{1}{6}$ and p_{bj} are the fair and biased probabilities, respectively. In R, **`1-pchisq(11.0705,5,16.6667)`** returns 0.9038948, matching our simulated result.

In more complex GOF tests, it may not be so easy to find the distribution of Q under an alternative hypothesis, but simulation is often straightforward. Because there is no way to show that the die is "absolutely fair," it may be useful to know that data from $n = 600$ rolls would often be enough to establish whether its bias is as bad as indicated by the vector **bias**. ◇

In applications of categorical data analysis, one often wants to know that the null hypothesis is reasonably close to the truth, and simply "failing to reject" the null hypothesis may not be persuasive. Thus power computations for various alternatives may be of particular interest.

Simulation in Statistical Inference

There are various ways in which simulation is used in inference. General information from simulations can guide statistical practice by revealing properties of methods that are not obvious from the underlying theory. Some inferential methods (for example, permutation, GOF, and sequential tests) use simulation to get approximate results in particular applications where it is not feasible to obtain exact theoretical results. Moreover, bootstrapping is fundamentally a method of the computer age, requiring simulation for each application.

Pooled and Welch Two-Sample *t* Tests

Suppose two samples $\{X_i\}$ and $\{Y_i\}$, of sizes n and m, are taken independently from populations distributed $\mathrm{NORM}(\mu_1,\ \sigma_1)$ and $\mathrm{NORM}(\mu_2,\ \sigma_2)$, respectively. Two t tests of $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$ are in common use.

For the pooled two-sample t test one uses the test statistic $T = \left(\overline{X} - \overline{Y}\right)/\mathrm{SE}_p$, where $\mathrm{SE}_p = S_p\sqrt{\frac{1}{n} + \frac{1}{m}}$; assumes that $\sigma_1 = \sigma_2 = \sigma$; and estimates the common variance σ^2 by S_p^2, a weighted average of the sample variances. Then under H_0, the statistic $T \sim \mathrm{T}(\nu)$, Student's t distribution with degrees of freedom $\nu = n + m - 2$. Also under H_0, the p-value of the test is a random variable distributed as $\mathrm{UNIF}(0,\ 1)$.

Example 20: Pooled test, variances equal. Let $\mu_1 = \mu_2 = 100,\quad \sigma_1 = \sigma_2 = 15\ ,\quad n = 10$, and $m = 40$. Figure 12a shows the histogram of simulated p-values for 100 000 pooled tests on samples as specified. If one rejects H_0 for a p-value below 0.05, then the dark blue bar at the left represents the 5141 tests (about 5% of them) for which H_0 was falsely rejected. ◇

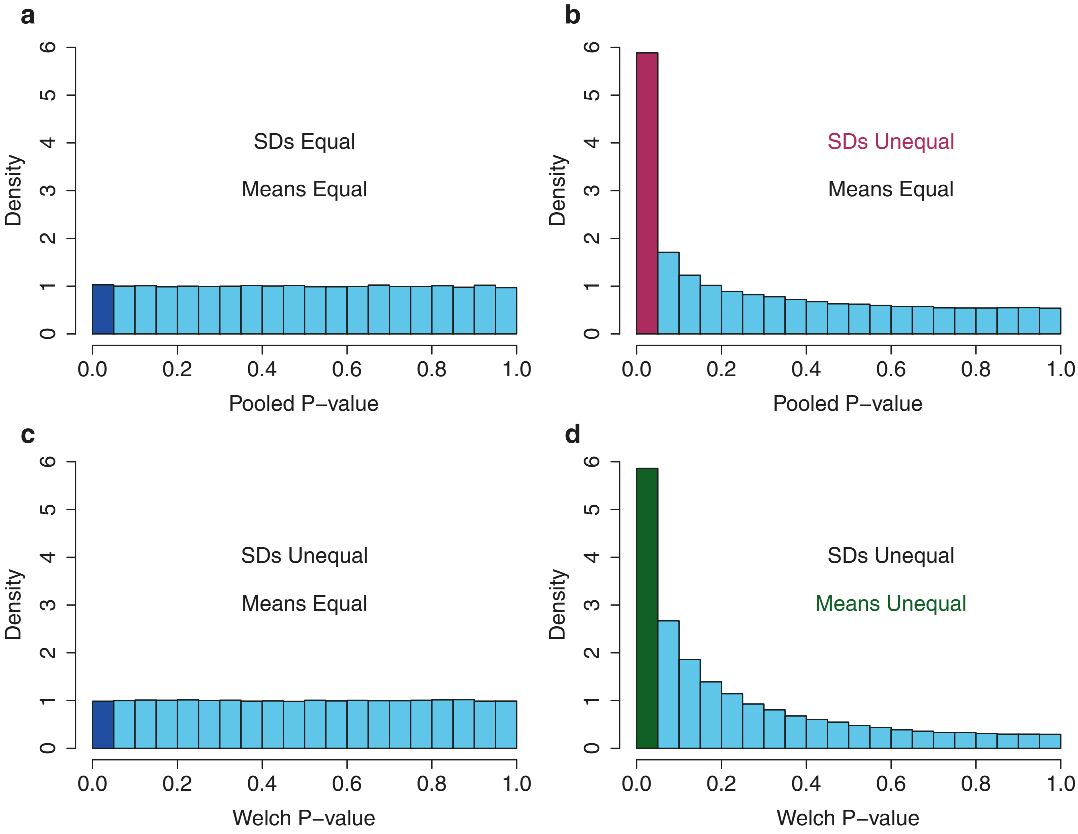
By contrast, if $\sigma_1 \neq \sigma_2$, then T does not have the distribution $\mathrm{T}(\nu)$ and the test may not perform properly. Results can be especially misleading when $n \neq m$ and the smaller sample is from the population with the larger standard deviation.

Example 21: Pooled test, variances unequal. $\mu_1 = \mu_2 = 100,\quad \sigma_1 = 20,\quad \sigma_2 = 5,\quad n = 10$, and $m = 40$. Repeating the simulation of Example 20, but with the unequal standard deviations shown, we find that 29 425 (about 30%) of the tests are incorrectly rejected H_0. In practice, this might cause investigators too often to make false claims of significant differences. See Figure 12b. Ott and Longnecker (2016) § 6.2. ◇

In the Welch two-sample t test one does not assume that $\sigma_1 = \sigma_2$ and uses the test statistic $T' = \left(\overline{X} - \overline{Y}\right)/\mathrm{SE}'$, where $\mathrm{SE}' = \sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}$. Then under H_0 the statistic T' is approximately distributed as $\mathrm{T}(\nu')$, where $\min(n - 1,\ \ m - 1) \leq \nu' \leq n + m - 2$. The precise value of ν' depends on the sample sizes and sample variances Welch (1947).

The theoretical rationale for the Welch test is sound, but it does not reveal details on how the test performs in practice. Numerous simulation studies have investigated these details.

Example 22: Welch test. $\mu_1 = \mu_2 = 100$, $\sigma_1 = 20,\quad \sigma_2 = 5,\quad n = 10,\quad m = 40$. A simulation of 100 000 Welch p-values, with the same parameters as in Example 21, gave the results in Figure 12c. There were 4931 (about 5%) incorrect rejections. So taking the Welch statistic T' to have ν' degrees of freedom gives appropriate p-values. (In most iterations, ν' was near its minimum $n - 1 = 9$.)



Simulations, Fig. 12 p-values for Pooled and Welch 2-sample t tests of Examples 20, 21, and 22. In each panel, the bar for the probability of rejecting at the 0.05 level is shown in a distinct color. In (**b**), H_0 is true, and the area of the tall bar shows an excessive number of rejected tests due to violation of the assumption $\sigma_1 = \sigma_2$. In (**d**), H_0 is false, and the area of the tall bar represents the power of the Welch test

If we repeat the simulation for the alternative $H_a : \mu_1 = 100\ ,\ \ \mu_2 = 110$, the proportion of rejections is appropriately higher, as shown in Figure 12d. ◇

Example 22 shows only two of many simulations that have verified the validity of the Welch test – for H_0 true and not, and across a wide range of sample sizes and ratios σ_1/σ_2. Taken in aggregate, we interpret these simulation studies as evidence that one should always use the Welch test unless there is good prior evidence that $\sigma_1 \approx \sigma_2$ (and even then, use the Welch test if sample sizes are very different). The Welch test is the default in the R procedure `t.test` for two samples (Ruxton 2006).

Permutation Tests

In a permutation test, one can use suitable random permutations of the data to find the distribution of a test statistic under the null hypothesis. As a very simple example, consider a two-sample design in which four independent observations are taken under each of two treatments A and B, and the null hypothesis is that there is no difference between the population distributions of A and B. If all four observations in group A exceed all of the four observations in group B [that is, min(A) > max(B)], then the p-value of a permutation test against the two-sided alternative is $2/\binom{8}{4} = 1/35 \approx 0.029$ (Fisher 1935; Basu 1980).

Example 23: Explosion of Challenger. The US space shuttle Challenger exploded in January 1986. It had been launched from Cape Canaveral, Florida, at the unusually low temperature of 29°F. Concerns before, and especially after, the tragedy centered on the possibility that O-rings, used to seal joints between sections of the rockets against fuel leakage, performed poorly at low temperatures.

Routine scrutiny of spent rockets from the 24 prior launches had revealed some O-ring “incidents” (indications of partial O-ring failure). One version of these data shows incidents per launch categorized by temperature: High (above 65°F) and Low (below 65°F). The question is whether these data show a significantly higher risk of O-ring problems at low temperature (Feynman 1988; Ramsey and Shafer 2002).

```
Low: 1, 1, 1, 3
High: 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
      0, 0, 0, 0, 0, 0, 0, 1, 1, 2
```

A Problematic Dataset. A one-sided Welch t test has p-value 0.038, but this result may be unreliable because the data are so far from normal. A nonparametric Mann-Whitney-Wilcoxon rank sum test is not applicable because there are so many ties. Fortunately, a permutation test is applicable.

Under the null hypothesis that incidents are equally likely in both temperature ranges, it should make no difference if incident counts for individual launches are permuted between the two groups. Suppose we use $D = \overline{X}_{\text{Low}} - \overline{X}_{\text{High}}$ as the measure of difference. Its observed value is $D_{obs} = 1.3$. If we had 0, 1, 2, and 3 in the Low group, that would also give $D = 1.3$. The only rearrangement giving a greater difference than observed would arise from putting 1, 1, 2, and 3 into the Low group to get $D = 1.45$.

There are $\binom{24}{4} = 10626$ ways to do a 4:20 partition of the observations. An elementary combinatorial argument shows that among them there are $10 + 10 + 85 = 105$ ways to get a value of $D \geq D_{obs}$. So the exact p-value of a permutation test is $105/10\,626 = 0.0099 < 0.05$, indicating a significant difference in the rate of O-ring failures due to temperature. If these data had been at hand and correctly analyzed before the Challenger was launched, one can speculate in hindsight that the launch would have been delayed to wait for warmer weather. ◇

Simulated Permutation Test

Especially when sample sizes are large and there are many ties, it can be tedious or difficult to find the exact permutation distribution of a test statistic D under H_0 (or even just the tail probabilities used to decide whether to reject). Nevertheless, we can simulate the permutation distribution, by looping through many random permutations and computing D for each one. This can provide a good approximation of the null distribution of the permutation test and hence of its p-value.

We use an R program, in which the vector **All** contains all 24 observations and the observed difference $D_{\text{obs}} = \overline{X}_{\text{Low}} - \overline{X}_{\text{High}} = 1.3$ is denoted **d.obs = 1:3**. We loop through 10 000 iterations. The key statements, in step **i** of the loop, make a random permutation **p** and compute the corresponding value **d.p[i]** of D_i, as shown below.

```
a = 1:4; b = 5:24
p = sample(All, 24)
d.p[i] = mean(p[a])-mean(p[b])
```

At the end of one such run, the statement **mean (d.p>=d.obs)** gave the approximate p-value 0.0105. This is the proportion of the simulated permutations for which $D_i \geq D_{obs} = 1.3$. It has a ± 0.002 margin of simulation error and very closely matches the exact p-value 0.0099 from combinatorics.

When the population distributions are in doubt, a traditional approach is to use a rank-based nonparametric test. Then the null and alternative hypotheses are often stated in terms of population medians instead of means. Moreover, information is lost when data are reduced to ranks, and ties may cause complications.

Simulated permutation tests are often an attractive alternative. For tests involving location, the metric for the hypotheses can be the mean, trimmed mean, or median, and the original data are used. There may be slight errors in the p-values because of the simulation, but they can be made negligible by doing enough iterations. In practice, especially when sample sizes are small and the metric is median-based, one should always look at the number of unique values of the metric (for example, **`d.p`** in the program just above) to make sure that appropriate p-values are feasible. Eudey et al. (2010) illustrates a wide variety of permutation tests.

Bootstrap Confidence Intervals

Bootstrapping uses simulation, often sampling with replacement from data or from a population distribution with parameters estimated from the data, in order to assess the variability of an estimate. The goal may be to calculate confidence intervals (CIs) or to test hypotheses. This sampling is part of the data analysis and does not create new information about the population; it is called *resampling*. Based on real data, we now illustrate parametric and nonparametric bootstrap CIs (Efron 1979; Efron and Tibshirani 1998).

Example 24: Differences between flood levels. Bain and Engelhardt (1973) quote $n = 33$ years of flood data. Values are differences at flood stage between water levels at two monitoring stations on a river in Wisconsin. Figure 13 shows their histogram along with best-fitting normal and Laplace density functions. Neither seems an ideal fit. With so little data, it is often better to rely on past experience or expert opinion, rather than goodness-of-fit tests, to decide what assumptions about the population distribution may lead to a useful analysis.

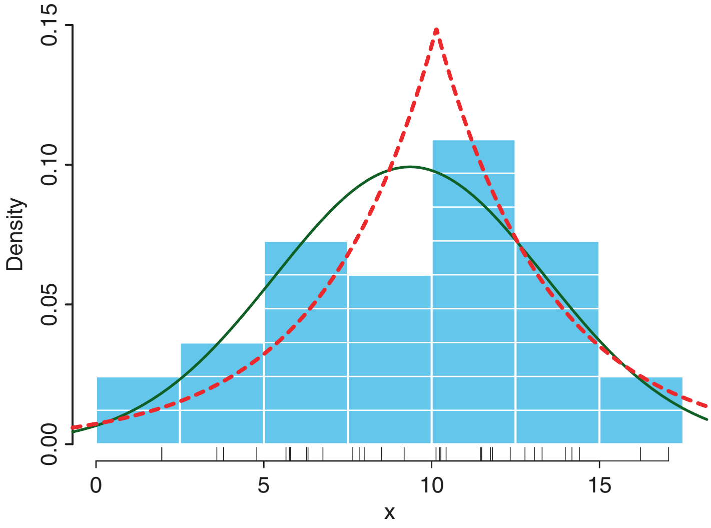
For normal or nearly normal data, it is reasonable to use the pivotal quantity $T = \frac{\overline{X}-\mu}{s/\sqrt{n}}$, where $\overline{X}$ and S are the sample mean and standard deviation, respectively.

Under the assumption of normality, this statistic T has Student's t distribution with $\nu = n - 1 = 32$ degrees of freedom, and the resulting 95% CI for the population mean μ is (7.93, 10.78). ◇

Parametric Bootstrap Confidence Intervals for Laplace Parameters

Bain and Engelhardt (1973) modeled the flood data according to a Laplace distribution. The maximum likelihood estimate (MLE) of η is the sample

Simulations, Fig. 13 Histogram of the flood data in Example 24. Density functions of NORM(9.35, 4.02), solid curve, and LAPLACE (10.14, 3.36), dashed curve, are shown for reference



median $\widehat{\eta} = 10.14$, and the MLE of β is the mean of the absolute deviations from $\widehat{\eta}$,

$$\widehat{\beta} = \frac{1}{n}\sum_{i-1}^{n} \mid X_i - \widehat{\eta} \mid = 3.361.$$

We use the pivotal quantity $D = (\widehat{\eta} - \eta)/\widehat{\beta}$ to obtain a 95% CI for η. The distribution of D does not depend on η or β (but does depend on the sample size n). If we knew the shape of this distribution, we could find values L_D and U_D with

$$\begin{aligned} 0.95 &= P\{L_D \le D \le U_D\} \\ &= P\left\{\widehat{\eta} - U_D\widehat{\beta} \le \eta \le \widehat{\eta} - L_D\widehat{\beta}\right\} \end{aligned}$$

and $\left(\widehat{\eta} - U_D\widehat{\beta}, \widehat{\eta} - L_D\widehat{\beta}\right)$ as a 95% CI for η. Lacking specific information on the distribution of D, we use a parametric bootstrap procedure to find approximate values L_D^* and L_U^* of L_D and L_U, respectively.

Entering the so-called *bootstrap world*, we temporarily treat $\widehat{\eta}$ and $\widehat{\beta}$ as if they were the true values of η and β, respectively. Then we take a large number B of resamples of size $n = 33$ from LAPLACE$\left(\widehat{\eta},\widehat{\beta}\right)$, with parameters *estimated* from the data. For each resample, we compute estimates of η^*, β^*, and then $D^* = (\eta^* - \widehat{\eta})/\beta^*$ (where $*$'s indicate quantities based on resampling). Then we take D_L^* and D_U^* to be quantiles 0.025 and 0.975, respectively, of the B bootstrapped values D^*.

Returning to the *real world*, we view $\widehat{\eta}$ and $\widehat{\beta}$ once again in their roles as the estimates of η and β, respectively, from the original data. The 95% bootstrap CI $\left(\widehat{\eta} - U_D^*\widehat{\beta}, \widehat{\eta} - L_D^*\widehat{\beta}\right)$ is obtained from the program below. The resulting bootstrap CI (8.75, 11.53) from one run agrees reasonably well with approximations in Bain and Engelhardt (1973) and in Kapperman (1975).

```
set.seed(4321); B = 10^5; n=33
eta.hat = 10.14; beta.hat = 3.361
 Z = rnorm(B*n); Q = rexp(B*n)
 x.bt = eta.hat+beta.hat*Z*sqrt(2*Q)
 DTA = matrix(x.bt, nrow=B, ncol=n)
 eta.bt = apply(DTA, 1, median)
 beta.bt = rowMeans(abs(DTA-eta.bt))
 D.bt = (eta.bt - eta.hat)/beta.bt
UL.D.bt = quantile(D.bt,c(.975,.025))
eta.hat - UL.D.bt*beta.hat
```

The additional R code below approximates a 95% CI for β based on the pivotal quantity $V = \widehat{\beta}/\beta$. Here

$$\begin{aligned} 0.95 &= P\left\{L_V \le \widehat{\beta}/\beta \le U_V\right\} \\ &= P\left\{\widehat{\beta}/U_V \le \beta \le \widehat{\beta}/L_V\right\} \end{aligned}$$

and $\left(\widehat{\beta}/U_V^*, \widehat{\beta}/L_V^*\right)$ is a 95% bootstrap CI for β.

```
V.bt = beta.bt/beta.hat
UL.V.bt = quantile(V.bt, c(.975, .025))
beta.hat/UL.V.bt
```

From one run, a 95% bootstrap CI for the scale parameter β is approximated as (2.50, 5.00), which also agrees closely with results in the two papers referenced above.

Nonparametric Bootstrap Confidence Interval

We assume that the observations are randomly sampled from the population of interest and that the population mean μ exists. But we do not make further assumptions about the population distribution (such as normal or Laplace). The nonparametric bootstrap is a computationally intensive nonparametric procedure – fundamentally different from traditional ones based on ranks.

We begin by finding the sample mean $\overline{X}$ and use $W = \overline{X} - \mu$ as a measure of the variability of $\overline{X}$ about μ, in a sample of size $n = 33$. If we knew the distribution of W, we could find values W_L and W_U with

$$\begin{aligned} 0.95 &= P\{W_L \le \overline{X} - \mu \le W_U\} \\ &= P\{\overline{X} - W_U \le \mu \le \overline{X} - W_L\}, \end{aligned}$$

to get $\left(\overline{X} - W_U, \overline{X} - W_L\right)$ as a 95% CI for μ.

However, we do not know the distribution of W, so we enter the bootstrap world to seek a useful approximation of that distribution, in which we use $\overline{X} = \mu^*$ to represent the unknown μ. By contrast with the parametric bootstrap, we draw a large number B of resamples of size $n = 33$ *with replacement* from among the $n = 33$ values *originally observed*. We compute the mean $\overline{X}^*$ of each of the B resamples and find B values $W^* = \overline{X}^* - \mu^*$ Then we denote quantiles 0.025 and 0.975 of these B bootstrapped values as W_L^* and W_U^*, respectively.

Finally, returning to the real world, we take $\left(\overline{X} - W_U^*, \overline{X} - W_L^*\right)$ as an approximate 95% bootstrap CI for μ. In this CI, $\overline{X}$ returns to its original role as the observed mean of the original sample.

With $n = 33$ and $\mathbf{x}$ as the vector of observed data values graphed in Figure 13, the program for the nonparametric bootstrap CI for η is shown below.

```
mean.obs = mean(x); B = 10^5
x.bt = sample(x, B*n, rep=T)
DTA = matrix(x.bt, nrow=B)
avg.bt = rowMeans(DTA)
w.bt = (avg.bt - mean.obs)
UL = c(.975, .025) # for 95% CI
UL.w = quantile(w.bt, UL)
mean.obs - UL.w
```

One run gives the nonparametric 95% bootstrap CI (8.01, 10.70) Additional runs give substantially similar results.

Assumptions affect results: This nonparmetric 95% bootstrap CI (8.01, 10.70) for μ is a little different from the 95% CI (7.93, 10.78) from the t distribution and the parametric 95% bootstrap CI (8.75, 11.53) for η. The assumption that the data are normal or Laplace provides additional information.

We emphasize that the resampling used in bootstrapping (parametric or nonparametric) is part of the data analysis; it produces no new information about the population. Moreover, the fundamental assumption of bootstrapping is that the empirical cumulative distribution function of the sample provides useful information about the cumulative distribution function of the population. This may not be true if the sample size is very small or if the population distribution has very heavy tails (Lunneborg 1999).

Future Directions

1. Quantum computers may soon be generally available, possibly presenting new opportunities for generation and use of (pseudo)random numbers for simulation (Yang 2012).
2. Setting a known seed for a PRN generator can allow exact replication of an individual simulation (See Example 4.) As inferential procedures rely increasingly on simulation methods, publication of seeds used in an analysis might allow others to replicate results and conclusions. Certainly, seeds can be used in this way if the same software package is used on the same platform. It seems worthwhile to investigate to what extent seeds can serve this purpose across operating systems, versions of software, and perhaps even software packages.
3. It is desirable to make information in some government and Internet databases available for public use. Often summary information is provided as medians instead of means, because it is more difficult to deduce private or confidential information of individual participants from medians. Sometimes the databases themselves are made public after jittering individual records with random noise. If the noise is LAPLACE(0, β) with suitable β, it may be possible to preserve anonymity and yet leave medians unchanged. Further investigation is needed (Kinney et al. 2011).
4. In large datasets from the Internet, the independence of observations may be in doubt to an extent that is not typical in small designed experiments. Traditional methods have relied on runs and autocorrelations to measure dependence. But to assess independence in a deck of cards, Bayer and Diaconis (1992) successfully turned to rising

sequences (used by magicians in card tricks for over a century). See Example 17. It seems worthwhile to investigate this and other nontraditional measures of dependence in big datsets.

5. In Internet gambling and actuarial settings, modern analytic techniques are used to monitor the randomness and frequency of extreme events. Rare events are difficult to simulate with small relative error. (See Example 15.) By comparing results from such real-life data with simulations of the theoretical models intended to describe them, it may be possible to improve the theoretical models or the methods of simulating them.

Cross-References

▶ Gibbs sampling
▶ Simulated datasets

Acknowledgments Several of the examples and figures here are adapted from material in the first three chapters of Suess and Trumbo (2010) and in Trumbo (2006). Our perspectives on simulation have been influenced by Braun and Murdoch (2007), Blitzstein and Hwang (2015), Grolemund and Wickham (2014), Gentle (1998), and Venables and Ripley (2002). We thank colleagues, current and former students, three anonymous referees, and our associate editor for useful suggestions.

The simulations in this article used R statistical software (open-source software available without cost from www.r-project.org for use on Windows, Macintosh, or UNIX operating systems). We hope our descriptions and examples of code are sufficiently clear that readers could repeat our simulations using other software. Python open-source software may be the most convenient alternative. Commercial software such as Mathematica, SAS, and Excel can also do simulations, but they may lack analogues of the specialized functions for probability and statistics we have used.

References

Bain LJ, Engelhardt M (1973) Interval estimation for the two-parameter double exponential distribution. Technometrics 15:875–887

Basu D (1980) Randomization analysis of experimental data: the fisher randomization test. J Am Stat Assoc 75(371):575–582. JSTOR 2287648

Bayer D, Diaconis P (1992) Trailing the dovetail shuffle to its lair. Ann Appl Probab 2(2):294–313

Blitzstein JK, Hwang J (2015) Introduction to probability. CRC Press/Chapman Hall, New York

Box GEP, Muller ME (1958) A note on the generation of random normal deviates. Ann Math Stat 29(2): 610–611

Braun WJ, Murdoch DJ (2007) A first course in statistical programming with R. Cambridge University Press, Cambridge

Chib S, Greenberg E (1994) Understanding the Metropolis-Hastings algorithm. Am Stat 49:327–335

Diaconis P, Holmes S, Montgomery R (2007) Dynamical bias in the coin toss. SIAM Rev 49(2):211–235. http://epubs.siam.org/doi/abs/10.1137/S0036144504446436?journalCode=siread

Dyson G (2012) Turing's cathedral: the origins of the digital universe. Vintage Books, New York

Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7(1):126

Efron B, Tibshirani RJ (1998) An introduction to the bootstrap. Chapman & Hall/CRC, Boca Raton

Eudey TL, Kerr JD, Trumbo BE (2010) Using R to simulate permutation distributions for some elementary experimental designs. J Stat Educ 18(1). http://www.amstat.org/publications/jse/v18n1/eudey.pdf

Feller W (1957) Introcution to probability theory and its applications, 2nd edn. Wiley, New York

Feynman RP (1988) What do you care what other people think? W W Norton, New York

Fisher RA (1935) The design of experiments. Oliver & Boyd, Edinburgh

Gentle J (1998) Random number generation and Monte Carlo methods. Springer, Berlin/Heidelberg/New York

Grolemund G, Wickham H (2014) Hands-on programming with R. O'Reilly Media, Sebastepol

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57(1):97–109. JSTOR 2334940

Kapperman RF (1975) Conditional confidence intervals for double exponential distribution parameters. Technometrics 17:233–235

Kinney SK, Reiter JP, Reznek AP, Miranda J, Harmin R, Abowd JM (2013) Toward unrestricted public use of business microdata: The Longitudinal Business Database. Int Stat Rev 79(3):362–384

Lewis PAW (1986) Graphical analysis of some pseud0-random number generators. Technical report NP-55586025, US Naval Postgraduate School, Monterrey

Liu JS (2004) Monte Carlo strategies in scientific computing. Springer, Berlin/Heidelberg/New York

Lunneborg CE (1999) Data analysis by resampling: concepts and applications. Duxbury, Pacific Grove

Marsaglia G (1995) The Marsaglie random number CDROM, including the Diehard battery of tests of randomness. Department of Statistics, Florida State U. http://stat.fsu.edu/pub/diehard/

Matsumoto M, Nishimura T (1998) Mersenne twister: a 623-dimensionally equidistributed uniform pseudorandom number generator. J ACM Trans Model Comput Simul (TOMACS) 8(1):3–30
Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equations of state calculations by fast computing machines. J Chem Phys 21(6):1087–1092
von Neumann J (1951) Various techniques used in connection with random digits. Natl Bur Stan Appl Math Ser 12:3638. https://dornsifecms.usc.edu/assets/sites/520/docs/VonNeumann-ams12p36-38.pdf
Ott RL, Longnecker M (2016) Introdution to statistical methods and data analysis, 7th edn. Cengage, Boston
RAND Corporation (1955) A million random digits with 100,000 normal deviates. Free Press, Glencoe
Ransey FL, Schafer DW (2002) The statistical sleuth: a course in methods of data analysis, 2nd edn. Duxbury, Pacific Grove
Ross SM (1997) Introduction to probability models, 6th edn. Academic Press, San Diego
Ruxton G (2006) The unequal variance t-test is an underused alternative to Student's t-test and the MannWhitney U test. Behav Ecol 17:688–690
Suess EA, Trumbo BE (2010) Introduction to probability simulation and Gibbs sampling with R. Springer, Berlin/Heidelberg/New York
Trumbo BE (2006, 2006) Random numbers from nonrandom arithmetic, STATS. J Am Stat Assoc (46):23–27. Alexandria
Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, Berlin/Heidelberg/New York
Welch BL (1947) The generalization of "Student's" problem when several different population variances are involved. Biometrika 34(12):28–35
Wichura MS (1988) Algorithm AS 241: the percentage points of the normal distribution. J R Stat Soc Ser C 57(3):477–484
Yang Y (2012) Quantum computing and quantum information. Stat Sci 27(3):373–394

Singular Value Decomposition

▶ Principal Component Analysis
▶ Semi-Discrete Decomposition

Singular Value Decomposition = Principal Component Analysis

▶ Eigenvalues: Singular Value Decomposition

SIoT

▶ Social Internet of Mobile Things and Decision Support Tools

Situational Web Applications

▶ Web Mash-Ups

Small Communities in Social Networks

Isaac Jones
CIDSE, Arizona State University, Tempe, AZ, USA

Synonyms

Local community detection; Tiny communities

Glossary

Community	A group of nodes in a network that exhibit a shared characteristic or abnormally dense interconnection
Resolution Limit	A mathematical constraint that prevents evaluation techniques from considering communities under a certain size, usually because the combination of two communities is scored higher than the two apart
Small Community	A group of nodes in a network that fits the definition of a community, but has low membership (e.g., 10 nodes in a million-node network)

Definition

Small Community Detection is the process of finding small communities in networked data that are often overlooked or aggregated by traditional community detection methods.

Social Networks are platforms, or simply data sets, that consist of connections between individuals.

Introduction

Modern online social networks are vast and complex with billions of users and even more connections between users. When traditional community detection is performed on these networks, often communities that are small in size are neglected. Paradoxically, multiple studies of community sizes in real networks have shown that these small communities, in fact, vastly outnumber the large communities. Consider, for example, a high school graduating class. Even among small classes, the students of that class fracture into many small groups whose members are more closely connected than with the rest of their class. There are strong parallels between this high school class example and many other real-world social networks, like workplaces, residential neighborhoods, and even academic collaborations. By finding these small groups in addition to the larger groups, techniques that rely on connections between users, like product recommendation or collaborative filtering, can be improved by taking these small, dense communities into account.

At face value, the transformation of the community detection problem by simply adding constraints on the community sizes does not seem like a big step. However, if we consider some of the most popular approaches to community detection, this does significantly affect the problem. Many approaches, most notably modularity optimization, are structurally biased against the detection of small communities (Barthélemy 2007). In modularity, this is referred to as the *resolution limit*. Even the map equation has a resolution limit, though the research on this limit is still ongoing. In fact, any global objective function for measuring community module quality must have either a resolution limit or a resolution parameter (Barthélemy 2007). Even techniques like spectral clustering that do not use a global objective function cost more and more computational power as the communities get smaller.

Key Points

Small community detection has its roots in standard community detection with an added complicating criterion that detected community sizes to be small. This is a complicating criterion because methods for detecting and evaluating community assignments are structurally biased against these small communities. The categories of approaches to small community detection parallels those used for standard communities, but modifications are necessary to avoid resolution limits. These modifications make the algorithms unique and interesting, and their results allow researchers in many fields to better utilize the results.

Historical Background

Detecting communities in social networks has a rich history as a problem, dating back to the early days of social network analysis though Newman's work in the area. As the topic matured, one strategy for detecting communities became dominant, hierarchical clustering with the modularity metric used to determine the best network partition. Further research into modularity and community detection in general later showed, however, that the metric has a *resolution limit*, a size threshold below which the metric prefers to group two communities together. Complicating this issue is the fact that this resolution limit is based on the size of the entire network, not the size of the communities involved. Thereafter, researchers in community detection expanded the approaches used in community detection to either avoid or mitigate the resolution limit. Since the resolution limit problem demonstrated that small communities have been understudied, many of these approaches have explicitly sought out small communities.

Small Communities in Social Networks

Problem Definition

In order to rigorously discuss small communities, we must first define the Small Community Detection problem. As implied previously, small community detection is a variant of the standard community problem. In community detection, we say that a *community assignment* is an assignment of nodes to communities such that each community consists of one or more nodes and each node is assigned a community. These community assignments, called C, inform the creation of links between nodes, which we represent as an adjacency matrix, A. Thus, in traditional community detection, the task is to find a set of community assignments, C, that maximally explain the adjacency matrix. In this formulation, an individual community, C_i, has no limitations on its size. Communities themselves are only loosely defined as a group of nodes whose internal connections outweigh their external connections. In small community detection, we keep this definition, but add an additional constraint; that the size of each community, $|C_i|$, be minimized.

Approaches

The structural hurdles of resolution limits have not stopped researchers from developing techniques that can detect small communities, however. Generally, these techniques can be broken into a few categories. In this section, we will explore these categories individually and then next section will provide specific examples of techniques from each category.

Community Expansion

Community Expansion, also called Seed Set Expansion, is a detection technique that relies on expanding small precursor communities into larger, fully realized communities. Often, these techniques start with a set of seed nodes or a set of extremely small communities (only 3–4 nodes) and then use some heuristic to incrementally add new nodes to the set until no more nodes satisfy the heuristic or some other terminating condition is reached. Many methods of this type have controllable heuristics, so the final size of the communities detected can be controlled by carefully adjusting the parameters of the heuristic.

Other techniques of this type use a labelling system, assigning each node in the network a label, again based on a controllable heuristic, and then using those labels to build communities. While these methods may seem very similar, upon further inspection we see that the two step process allows the latter category of techniques to use a larger number and type of heuristics for node labelling. Conversely, the former category is frequently more scalable and can often operate when there is incomplete network information.

Spectral Clustering

Harkening back to one of the original forms of community detection, the Shi-Malik algorithm, spectral clustering approaches perform community detection by projecting the adjacency matrix into an alternative space. Frequently, but not exclusively, this is the eigenvector space. Classically, spectral clustering algorithms have been used to divide the network into two portions. Unfortunately, it divides the network into exactly two partitions, so does not scale particularly well to very large networks with a large number of communities, especially when those communities are small.

By redefining the problem of spectral clustering to local spectral clustering, these techniques can be applied to small subgraphs of the network in order to isolate one community at a time. By performing spectral clustering on a selected subgraph, one group of nodes is considered the community, where the remaining nodes are considered the noncommunity nodes. In this way, a broad, overly general algorithm can section a large graph into subgraphs and then a local spectral clustering method can specifically tease out the communities from those subgraphs. This has the added advantage of reducing the time complexity of spectral clustering, which is very poor on large network but manageable, especially through parallelization, on subgraphs.

Network Reweighting

The final strategy for small community detection we will discuss directly combats the resolution limit problem referenced earlier. In the base formulation of modularity, and many other global optimization methods, the network is assumed to be unweighted but the techniques can easily handle weighted networks. Thus, a number of reweighting techniques have been applied to attempt to circumvent resolution limits by weighting edges between nodes based on their likelihood of connecting members of the same community.

These techniques use a wide variety of measures to inform their reweighting schemes. Like community detection techniques, these can range in complexity and be global or local in nature. A local measure might be as simple as average node degree between the two edge endpoints or as complex as simulated random walk occurrence frequency. Global techniques for edge reweighting also range in complexity.

Techniques

In this section, we will cover the basics of a few techniques from each category of approaches discussed above. Under Community Expansion, we will discuss the SCAN algorithm (Xu 2007), which uses the two-pass labelling/clustering system to assign labels to nodes and then create a network partition. Under Spectral Clustering, we will discuss the LOSP++ algorithm (He 2016), which uses a breadth-first algorithm to collect a community and its neighborhood and then a spectral transformation to chip away the extra nodes. Lastly, Network Reweighting will be represented by Infomap (Rosvall 2008), which uses random-walk reweighting to influence the optimization process for the Map Equation.

The SCAN Algorithm

The SCAN algorithm, as an exemplar of seed set expansion methods techniques in general, follows a two-step process for finding communities and that process can be configured to find small communities. The two steps are, roughly, seed set identification and community expansion. Some techniques omit the first step and focus on expansion, but SCAN does not. In SCAN, identifying a seed set from which to expand is by analyzing the neighborhoods of two nodes at the endpoints of a particular edge. This neighborhood similarity is used, along with the similarity of other edges, to determine if a particular node is a “core” node, indicating its neighborhood is highly self-similar and thus densely connected. Using these densely connected neighborhoods as a seed set is a good starting point for building communities of interest. In SCAN, the neighborhood similarity threshold can be controlled in order to ensure that, in our case, the resultant cores are very dense and thus also small. In addition to the similarity threshold, SCAN also has a density threshold that ensures core nodes have a minimum number of sufficiently similar neighbors, which helps to control final community sizes.

The next step of the algorithm, expansion, occurs after the seed set has been identified and handles to assignment of noncode nodes to the most appropriate community. Here, SCAN’s algorithm is particularly appropriate for detecting small communities, as it allows noncore nodes to be classified as “hub” or “outlier” nodes if they connect to many or no communities. This removes the need to assign every node to a community, which allows small communities to remain small and connected only to those other nodes which are appropriate. Other techniques use other methods to isolate these outliers and hub nodes, though discussion of these techniques is best left to their individual authors.

As an example of seed set expansion techniques, SCAN is notable for its relatively simple algorithm that provides strong results and is adaptable to many scenarios through manipulation of its two thresholds. In addition, since SCAN does not use a global measure of partition goodness, it is not susceptible to the resolution limits that plague techniques that rely on these measures.

Spectral Clustering

In general, spectral techniques can be identified by their use of alternate space projections to simplify the computation of community assignments. One of the first examples of such a technique, the Shi-Malik algorithm, uses an eigenvector projection to partition a network into two partitions. The technique we will discuss here, LOSP++, builds upon the ideas of the Shi-Malik algorithm while ensuring that the technique can discover small communities and remains scalable to very large network sizes. LOSP++ is also notable since it is a seed set technique. The algorithm relies on having a reasonable seed set provided in order to it to find an optimal community assignment for each node. Thus, LOSP++ could be considered a spectral seed set expansion technique, though that does not capture the subtleties accurately.

The LOSP++ algorithm has, generally speaking, two steps that it executes for each seed in the seed set, eventually finding one community for each seed set. The first phase, the expansion phase, attempts to collects a reasonable estimate of the seed's community using a high-recall approach that also captures many extraneous nodes. This approach is based on a breath-first search of the network with a limited maximum path length from the seed node. In addition, the local search only admits new nodes to the local subset if the nodes have a high proportion of inward-facing edges. This helps to filter out the hyperpopular nodes that characterize social networks and prevent the local subset from exploding in size, even under a constrained path length. By setting the maximum distance low and maintaining a high inward-facing ratio, the LOSP++ can be controlled to restrict the candidate network to a small network and thus a small final community.

The second phase of LOSP++ is the spectral phase, which reduces the local subnetwork found in phase one to its most essential components using Krylov subspace projection. This transforms the community detection problem into one of finding a cut that retains as dense and valuable as possible a community around the seed node. Since the subnetwork is much smaller that the base network, the spectral projection computation can be performed much more quickly. As an example of spectral methods, LOSP++'s approach is relatively light on spectral computation, but it does effectively blend the two methods to use the strengths of each while minimizing the disadvantages.

The Infomap Algorithm

Lastly, the Infomap algorithm takes a global approach to community detection through the map equation and information theory. Though the map equation is a global measure of partition goodness, Infomap is more resistant to resolution limit concerns due to its two-level approach. Broadly, this two-level approach uses information theory to capture the flow of information around the network and create small clusters and then uses the map equation to optimize the modules. By simulating random walks on the network, the Infomap algorithm can determine the best clustering of node sets. Since small communities are, by their nature, very dense, there is a natural tendency to for random walks to stay within small communities. After creating a large number of clusters, the second level of the map represents the connections between clusters. Dividing the map this way allows the clusters in the first level of the map to contain a large amount of information. Infomap works well on large networks with small communities precisely because of this two-level structure. Unlike modularity, which considers the entire network simultaneously, the map equation considers the relationships between communities separately from the composition of communities, which allows the technique to capture small communities without agglomerating them together (though this is not always the case).

As an example of global clustering techniques, Infomap shows the potential for these techniques to work around the resolution limits inherent in global optimization. Even modularity has a number of proposed techniques for avoiding its resolution limit, and we encourage interested readers to consider these techniques if they are appropriate for the problem at hand.

Key Applications

Community detection, and more specifically small communities, has a number of applications that allow researchers in the field to apply their techniques to other problems. For example, finding small groups of tightly connected, highly similar people has value in the fields of product recommendation where members of a community are likely to have similar tastes and interests. Similarly, this can be applied to friend recommendation, where it is likely that similar people who already have friends in common are likely to know or like one another. In addition, small communities demonstrate this property to an enhanced degree, since the members of the community are more tightly connected and similar that the members of large communities. Therefore, knowing memberships of communities of various sizes not only allows for membership to influence other problems, but allows that influence to have degrees of impact corresponding to the size of the community.

Network structures have historically been used to represent much more than just social networks, however. These structures can be used to represent many data sets, from biology to marketing and small community detection has application for all of them. In biology, protein interactions can be represented as a network and small community detection can be used to isolate important sections of the networks for detailed study. In marketing, product purchases can be represented as a network through frequent itemset interaction or through features like "customers who viewed this item also bought" on Amazon. Any problem that has a network representation of the data and finds value in the discovery of small, densely connected subnetworks can benefit from the application of small community detection techniques.

Future Directions

Future directions for the field of small community detection have a wide variety of possibilities in both the theoretical and application space. Theoretically, researchers aim to improve detection results, finding more communities of high quality in data sets without ground truth and more closely matching ground truth in those with such assignments. In addition, the resolution limit problem continues to receive attention. If this is solved, the door opens to apply many more techniques from general detection to the problem of detecting small communities. The application of small communities also continues to be explored, as researchers find more and more problems with meaningful network representations. Each new domain provides unique challenges, as a network representation in necessary, but the resultant communities must also be interpretable and actionable, otherwise the effort is wasted.

Cross-References

- ▶ Community Detection and Analysis on Attributed Social Networks
- ▶ Community Detection and Recommender Systems
- ▶ Community Detection in Social Network: An Experience with Directed Graphs
- ▶ Community Detection: Current and Future Research Trends
- ▶ Community Evolution

▶ Community Structure Characterization
▶ Competition Within and Between Communities in Social Networks
▶ Connecting Communities
▶ Data Mining Techniques for Social Networks Analysis
▶ Detecting and Identifying Communities in Dynamic and Complex Networks: Definition and Survey
▶ Dynamic Community Detection
▶ Evolving Social Graph Clustering
▶ Extracting and Inferring Communities via Link Analysis
▶ Game-Theoretic Framework for Community Detection
▶ Link Dynamics and Community Formation in Social Networks
▶ Models for Community Dynamics

References

Barthélemy M (2007) Resolution limit in community detection. Proc Natl Acad Sci 104:36–41

He K (2016) Local spectral diffusion for robust community detection. Twelfth workshop on mining and learning with graphs

Rosvall M (2008) Map of random walks on complex networks reveal community structure. Proc Natl Acad Sci 105:1118–1123

Xu X (2007) SCAN: a structural clustering algorithm for networks. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 824–833

Small-World Networks

▶ Networks in Geography

Smart Social Objects

▶ Social Internet of Mobile Things and Decision Support Tools

Smartphone Proximity Networks

▶ Assessing Individual and Group Behavior from Mobility Data: Technological Advances and Emerging Applications

SMS

▶ Microtext Processing

SNA

▶ NodeXL: Simple Network Analysis for Social Media
▶ R Packages for Social Network Analysis

SNA Software

▶ Tools for Networks

Snap.py

▶ Python Packages for Networks

SNS

▶ Social Networking on the World Wide Web

SNSs

▶ Privacy and Disclosure in a Social Networking Community

Social Analysis

► Social Web Search

Social Anthropology

► Social Network Analysis in an Age of Digital Information

Social Behaviors

► Behavior Analysis in Social Networks

Social Bookmarking or Tagging

Johann Stan[1,2] and Pierre Maret[3]
[1]Univ. Lyon, UJM-Saint-Etienne, CNRS, Laboratoire Hubert Curien UMR 5516, Saint Etienne, France
[2]ISCOD – Institut Henri Fayol, Ecole Nationale Supérieure des Mines, Université de Lyon, UJM-Saint-Étienne, CNRS, Saint-Étienne, France
[3]UMR 5516 Laboratoire Hubert Curien, Université de Lyon, CNRS, Saint- Étienne, France

Synonyms

Bookmarks; Social Tags

Glossary

Annotation	See Tag
Folksonomy	Whole set of tags that constitutes an unstructured collaborative knowledge classification scheme in a social tagging system
Resource	In the context of this work, a multimedia content (e.g., text file, photos, videos, web page) available on the Internet. A resource is generally identified by a Unique Resource Identifier (URI) which enables its access using the REST protocol
Social bookmarking system	Web-based systems allowing users to describe resources with tags
Social bookmark	Tag in the form of a link to a resource (e.g., web page) that is intentionally stored, and possibly shared, by an identified individual on a social bookmarking system, in which individuals can attach tags
Tag	A descriptive keyword entered by a human individual with the objective to describe a resource (e.g., a photo, a web page). It is also called an annotation or user-generated content

Definition

Social bookmarking systems (SBS) are web-based systems allowing users to describe resources with annotations, also called tags. The fundamental unit of information in a social bookmarking system consists of three elements in a triplet, represented as user, resource, and tag (Cattuto et al. 2006). This triplet is also called a tag application (an instance of a user applying a tag to a resource; this is also referred to as a tag post) (Sen et al. 2006). The combination of elements in a tag application is unique. For example, if a user (also known as tagger) tags a paper twice with the same tag, it would only count as one tag application. Resources can mean different things for different social bookmarking systems. In the case of del.icio.us, a resource is a website, and in the case of CiteULike, it is an academic paper.

Introduction

Social bookmarking (or tagging) is a means for describing resources. The social bookmarks are tags attached to a resource, with the main objective to describe said resource. They describe the context or the meaning of such artifacts. Social bookmarks can be of multiple forms, depending on the semantic structure they rely on.

The manipulation of web resources involves tasks such as description, retrieval, reuse, presentation, and search. All these tasks need a layer of prior knowledge, which is represented by the social bookmarks, which can be composed of different types of annotations.

Key Points

Tag (or annotations) may be either structured, semi-structured, or unstructured. Tags tend to be short. Hashtags are one-word annotations. Tagging became popular on social networks.

Historical Background

The emergence of the so-called Web 2.0 (from 2004) gave rise to user-generated contents (UGCs) and therefore to web repositories of UGCs. However, Wikipedia (2001) was already launched in 2001, and it was one of the first public crowd-sourced websites. This free encyclopedia has been allowing anyone to edit the content of any article. Whereas this openness has implied many disputes on pages related to controversial subjects (e.g., facts about presidential candidates just before election, about historical events, companies, etc.), it has grown to become a major and useful reference, covering many languages. This encyclopedia has been translated to a semantic database called DBpedia (Bizer et al. 2009) since 2007, enabling its user-generated content to be machine readable, so that computer programs (and mash-ups) can leverage knowledge facts by formulating precise queries.

Even though Wikipedia has been opened to any voluntary contributions, contributors are still few, compared to the number of readers. Participating in social bookmarking sites, like Delicious (2003), has become more popular, as the contribution process was quicker, simpler, and more personal. After creating an (free) account on the site, users can immediately bookmark web pages that they want to keep, because they enjoy them, they want to be able to easily find them later, and they (often) want to share them with other people. In order to make bookmarked web pages more easy to find later, users are invited to annotate them with "tags," unconstrained words (in any language, without even spell-checking) that subjectively reflect the apparent nature, function, category, and context of those web pages (Golder and Huberman 2006a). Web pages bookmarked (and tagged) by several people are thus described by a "tag cloud," a displayed set of tags. The size of a tag depends on the number of people who used this tag to describe this page.

As any URL-located resource can be bookmarked on social bookmarking sites, these descriptions can apply on various types of entities represented by those resources. For example, tags given to a page that presents a car are most probably associated to the car, than to the page/site itself. Now that web pages exist for almost anything on earth (e.g., people, objects, places, events, etc.), social bookmarking is a promising paradigm for gathering crowd-sourced descriptions and classifications of virtual and real entities.

More specific repositories also exist to represent and describe real-word entities and discover their involvement with people's activities. Concerning music, MetaBrainz Foundation (2001) can identify the name and interpreter of a song from a sampled audio (e.g., recorded with a microphone), and tags given by people to songs and artists are gathered on websites like Last.fm (2002), which also maintains a history of the last songs that users listened to. Image-sharing websites like Flickr (2004) can be considered as social bookmarking applied to photographs, as it is possible to tag one's own and other people's photographs, including the time and geographical location where the picture was taken.

Additionally, real-world places are described, reviewed by people, and geographically located on various websites (and their mobile applications) such as Yelp and Qype (2004) and (2006).

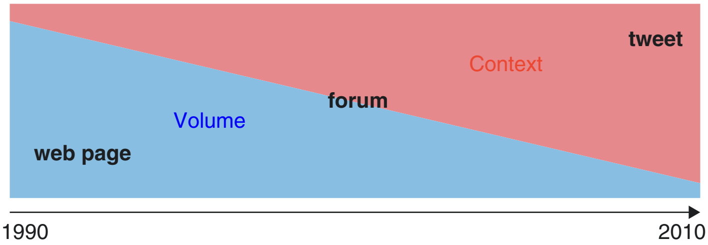
Rattenbury et al. (2007) have proven that names of places and events can also emerge by analyzing the frequency and temporal distribution of tags associated to geolocated pictures. Most websites cited above expose public feeds that one can subscribe for being aware of last updates and/or APIs that allow computer programs to query information, given specific criteria (e.g., information about a place and a topic, at a given time range).

Thousands of other APIs are referenced on sites like ProgrammableWeb. Also note that tags are not directly available on all the websites cited above, but keywords can be identified from the user-generated content they feature. It is also possible that pages from those sites are tagged on Delicious.

Types of Annotations

Annotations may be either structured, semi-structured, or unstructured:

1. *Structured annotations*. In this case, the terms employed in the annotation are regulated by a common domain vocabulary that must be used by the members of the system. These types of annotations are currently not used in the majority of social platforms because a domain vocabulary containing the necessary terms for the annotations is needed. Although such an approach has many advantages (e.g., absence of synonyms, absence of differences in pronunciation), this is not the natural way to describe resources in Web 2.0 platforms, as the domain is not well defined, and, therefore, it is very difficult to build such vocabularies and to establish a consensus for each term used. At the same time, the use of semantic annotations would be cumbersome for people, as it is time-consuming and requires additional cognitive effort to select concepts from existing domain ontologies. In addition, semantic annotations work well in systems where the domain is well defined (e.g., a system for sharing knowledge about human genes (Yeh et al. 2003)), but, in social bookmarking systems, this is not the case, as the shared content is generally very heterogeneous, as people can discuss without limits (i.e., covers multiple domains with no regularities and relations).
2. *Semi-structured annotations*. In contrast, semi-structured annotations, such as social tags, are widely used or photo tagging and bookmarking (e.g., the annotation of a web page). These annotations are generally freely selected keywords without a vocabulary in the background. However, we consider them to be semi-structured, as they represent an intermediate approach between semantic annotations (i.e., annotations that are based on concepts from domain ontologies) and free-text annotations. Besides, such collections of tags converge to a structured data organization, called a folksonomy (Gruber 2007). This consists of a set of users, a set of free-form keywords (called tags), a set of resources, and connections between them. As folksonomies are large-scale bodies of light-weight annotations provided by humans, they are becoming more and more interesting for research communities which focus on extracting machine-processable semantic structures from them. These underlying data clouds of collaborative tagging systems enable Internet users to annotate or search for resources using custom labels instead of being restricted by predefined navigational or conceptual hierarchies (e.g., ontologies).
3. *Unstructured annotations*. Finally, a more recent form of annotations is represented by free-text annotations, also called social awareness streams, composed of status updates or microposts (Naaman et al. 2010a). This can be found in the majority of social networks and microblogging systems and primarily consists of free texts in the form of short messages describing a resource, a finding, an impression, a feeling, a recent activity, a mood, or a future plan. The limitations of this practice from the viewpoint of information retrieval and knowledge management are similar to that of social



Social Bookmarking or Tagging, Fig. 1 Evolution of content production on the Internet: from structured documents to microposts

tagging, as users have complete freedom in the formulation of these messages. It is important to mention that in social awareness streams, the produced content often contains the described resource itself, in the form of an integrated hyperlink. A common practice is either to express an opinion about the resource (e.g., web page) or to provide its short summary for the community. Since Internet took over Usenet as the main computer-based means of communication, it has gone through several stages: read-only web, with large pieces of information close to a magazine article size; read-write web, or Web 2.0, with forums mimicking Usenet, exchanging pieces of information up to half a page in size; blogging, close to the web page model but with a shift in authorship toward the general public; and microblogging, based on very short messages (e.g., 140 characters on Twitter). This shift from large, authoritative information to very short and amateur information is contemporary with the mobility evolution, with the more user-friendly web-enabled devices (e.g., the iPhone) emphasizing a particular factor, the context in which information is written. This has blurred the distinction between information and messaging, as all information on Twitter is in fact a message to followers, and all messages may be shared, thus creating information. Events and documentation on the contrary are becoming more distinct: in the traditional newspaper information model, documentation is delivered with events in a single article; in the Twitter-driven model, events are tweeted, and the user is meant to seek information in more reliable and static sources, such as Wikipedia. An example of such a shift is the growing use of Twitter in the scientific community, contrasting strongly with the process of peer-reviewed publication (Fig. 1).

The Decreasing Size of Annotations

An interesting issue about free-form posts in social platforms is their short size, which emerged as a simple, convenient way to communicate about activities or share findings. The size limitation of such posts, defined by the majority of such platforms, is mostly due to the fact that users can in this way follow hundreds of friends in real time, without an important time investment. Also, this light-weight form of communication enables users to easily broadcast opinions, activities, and status (Java et al. 2007; Naaman et al. 2010b).

Common practices emerged to reduce the length of messages and to help users to rapidly identify the messages relevant for them. Thus, hashtags are used to identify posts relevant to a specific event or a specific topic. Also, common ways are used to synthesize information: include the source web page or reduce the amount of stop words in order to gain place for the informative terms (keywords and named entities). These practices largely depend also on the targeted user community, which can vary from a small family to the world at large.

The same applies to the composition of such posts, where common practices emerged as new means to better identify posts relevant to a specific event, also called “hashtags,” or common ways to synthesize an information, such as including the source web page or reducing the amount of stop words in order to gain place for the informative terms (keywords and named entities). These practices largely depend also on the targeted user community, which can vary from a small family to the world at large. Microposts are often called "social signals" (Mendes et al. 2010) and users of such systems "social sensors" (Sakaki et al. 2010), as they can be useful to detect important events in a given location, such as an earthquake.

Knowledge Management in Social Bookmarking Systems

A category of annotations in social platforms is semi-structured, also called social tags. Social bookmarking systems have become extremely popular in recent years. Their underlying data structures, known as folksonomies (Mathes 2004), consist of user-tag-resource triples.

Folksonomies contain peoples' structural knowledge about documents. A person's structural knowledge has been defined as the knowledge of how concepts in a domain are interrelated from the individual's point of view. According to Mathes (2004), an important aspect of a folksonomy is that it is comprised of terms in a flat namespace: that is, there is no hierarchy and no directly specified parent-child or sibling relationships between these terms. There are, however, automatically generated "related" tags, which cluster tags based on common URLs. This is unlike formal taxonomies and classification schemes where there are multiple kinds of explicit relationships between terms. These relationships include functions like broader, narrower, as well as related terms. These folksonomies are simply the set of terms that a group of users tagged content with; they are not a predetermined set of classification terms or labels.

Folksonomies claim to have many advantages over controlled vocabularies or formal taxonomies. Tagging has lower costs because there is no complicated, hierarchically organized vocabulary to learn and adapt to its own one. Users simply create and apply tags. According to Wu et al., "Folksonomies are inherently open-ended and therefore respond quickly to changes and innovations in the way users categorize content" (2006). Collaborative tagging is regarded as democratic metadata generation where metadata is generated by both the creators and consumers of the content. Folksonomies can be divided into broad folksonomies, which allow different users to assign the same tag to the same resource, and narrow folksonomies, in which the same tag can be assigned to a resource only once.

The question of why folksonomies are successful has been the subject of several studies in the literature. An important argument for this is the fact that the feedback loop is tight (Mathes 2004), i.e., once the user assigns a tag to an item, the cluster of items with identical or similar tags can be immediately retrieved. This can help the user decide whether to keep the tag or change it to a similar or different one. The scope of such a cluster can be expanded by showing all items from all users in the system which are tagged with the same tag. By viewing the result set, the user can decide how to better adapt the tag to the group norm or to have better visibility in the community for the tagged resource. The issue of how to influence the group norm was also studied by Udell. This tight feedback loop leads to a form of asymmetrical communication between users through metadata. The users of a system are negotiating the meaning of the terms in the folksonomy, whether purposefully or not, through their individual choices of tags to describe documents for themselves.

A folksonomy eases collaboration. Groups of users do not have to agree on a hierarchy of tags or detailed taxonomy; they only need to agree, in a general sense, on the "meaning" of a tag enough to label similar material with terms for there to be cooperation and shared value. Although this may require a change in vocabulary for some users, it is never forced, and as Udell discussed, the tight feedback loop provides incentives for this cooperation.

The main problems of social tagging systems include ambiguity, lack of synonyms, and discrepancies in granularity (Golder and Huberman 2006b). An ambiguous word, e.g., apple, may refer to the fruit or the computer company, and this in practice can make the user retrieve undesired results for a given query. Synonyms like lorry and truck, or the lack of consistency among users in choosing tags for similar resources, e.g., *nyc* and *new york city*, makes it impossible for the user to retrieve all the desired resources unless he/she knows all the possible variants of the tags that may have been used. Different levels of granularity in the tags may also be a problem: documents tagged “java” may be too specific for some users, but documents tagged “programming” may be too general for others.

Several attempts have been made to uncover the structure of this kind of data organization. Basic formal models of folksonomies include that of Mika (2007) and Hotho et al. (2006). Mika proposes a model based on *tripartite hypergraphs* while Hotho et al. on *triadic context* (term used in formal concept analysis). We present in the following the formal model of Mika, one of the most cited models in the literature for the representation of these structures.

As said before, a folksonomy is an association of users, annotations, and resources. The corresponding three disjoint sets of vertices are considered by Mika in the formal model: the set of actors (users) A, the set of concepts (tags) C, and the set of resources O (e.g., photos, videos, or web resources, like bookmarks, websites, etc.). Since in a social tagging system, users tag objects with concepts, ternary relations are created between the user, the concept, and the object.

This resulting tripartite hypergraph can be transformed into several bipartite graphs, each having a very specific meaning, like AC – the graph that associates actors and concepts – CO, the graph that associates concepts and objects, and AO, the graph that associates actors and resources.

Abel (2008) investigates the benefits of additional semantics in folksonomy systems. Additional context can be provided to the tagging activity with an extension of the tripartite model, i.e., an association of the user, the tag, and the tagged resource that describes more precisely the particular tagging activity. For example, time stamp helps to categorize tags in a temporal manner; the mood the user had when tagging the resource would allow to qualify opinions expressed in a tag. Other information, like background knowledge about the user, would allow to have information about the reliability of the tagger. The GroupMe! folksonomy system is proposed, which is a new kind of resource-sharing system for multimedia web resources. A first extension of previous models is the introduction of the term group, which is a finite set of related resources. The folksonomy model in GroupMe! can be thus formalized in the following manner (we note with F the folksonomy model): $F = (U, T, IR, G, Y)$, where U, T, R, and G are finite sets that contain instances of users, tags, resources, and groups. $IR = R \cup G$ is the union set of resources and the set of groups.

Wu et al. (2006) identify the key challenges in collaborative tagging systems. The three identified challenges are the following: (i) the identification of communities, i.e., groups of users with similar interests, (ii) preventing information overload by filtering out high-quality documents and users (e.g., experts in a domain), and (iii) how to create scalable, navigable structures from folksonomies. Folksonomies are criticized to have flaws that formal classification systems are designed to eliminate, including polysemy, words having multiple related meanings, and synonymy, multiple words having the same or similar meanings.

Information Retrieval from Folksonomies: Social Information Retrieval

In the previous section, we have seen the general definition and structure of folksonomies, the data organization in social tagging systems. In this section, we go further and review existing techniques of information retrieval in folksonomies.

The biggest challenge in folksonomies is information retrieval, i.e., the question of how to efficiently rank items (e.g., tags, resources, users) for a given user query. In traditional Internet applications, the search and navigation process serves two vital functions: retrieval and discovery. Retrieval incorporates the notion of navigating to a particular resource or a resource containing particular content. Discovery incorporates the notion of finding resources or content interesting but therefore unknown to the user. The success of collaborative tagging is due in part to its ability to facilitate both these functions within a single user-centric environment. Reclaiming previously annotated resources is both simple and intuitive, as most collaborative tagging applications often present the user's tag in the interface. Selecting a tag displays all resources annotated by the user with that tag. Users searching for particular resources they have yet to annotate may select a relevant tag and browse resources annotated by other users. However, the discovery process can be much more complex. A user may browse the folksonomy, navigating through tags, resources, or even other users. Furthermore, the user may select one of the results of a query (i.e., tag, resource, or user) as the next query itself. This ability to navigate through the folksonomy is one reason for the popularity of collaborative tagging.

In order to provide efficient retrieval mechanisms, a formal model of folksonomies is required. There are several models in the literature, e.g., that of Mika (2007) and Hotho et al. (2006). Mika proposes a model based on *tripartite hypergraphs*, while Hotho et al. on *triadic context* (term used in formal concept analysis).

Hotho et al. adapt the well-known PageRank algorithm in order to apply it on folksonomies, called *FolkRank*. The impossibility of applying *PageRank* has its origins in the fact that a folksonomy is different from the web graph (undirected triadic hyperedges instead of directed binary edges). By modifying the weights for a given tag, FolkRank can compute a ranked list of relevant tags.

The original formulation of PageRank (Brin and Page 1998) reflects the idea that a page is important if there are many pages linking to it and if those pages are important themselves (recursive aspect of importance). The distribution of weights can thus be described as the fixed point of a weight-passing scheme on the web graph. This idea was extended in a similar fashion to bipartite subgraphs of the web in HITS (Kleinberg 1999) and to n-ary directed graphs (Xi et al. 2004). The same underlying principle is employed for the ranking scheme in folksonomies. The basic notion is that a resource which is tagged with important tags by important users becomes important itself. The same holds, symmetrically, for tags and users. Such a ranking schema can help the emergence of a common vocabulary in collaborative tagging systems, by recommending to the user tags that have a bigger visibility in the community and that is also semantically close to the user-defined tag.

Abel et al. (2008) perform an in-depth analysis of ranking algorithms specially designed for folksonomies: FolkRank, SocialSimRank (Bao et al. 2007), and SocialPageRank and adapts them to the GroupMe! social bookmarking system, where an additional dimension is added to folksonomies, i.e., groups of resources.

Gemmell et al. (2008) propose a method to personalize a user's experience within a folksonomy using unsupervised clustering of social tags as intermediaries between a query and a set of items. Terms in the query are weighted based upon their affinities to particular clusters to help disambiguate queries.

Bao et al. (2007) propose different algorithms, such as SocialSimRank and SocialPageRank to optimize web search using social annotations. The underlying hypothesis of the proposed algorithms are the following: (i) social annotations about web pages are good summarizations of the given web page and can be used for efficient computation of similarity between a search query and a web page and (ii) the amount of annotations assigned to a web page is a good indication of its popularity.

Vocabulary Construction and Emergence of Semantics

In this section, we present different approaches for extracting and constructing a hierarchical structure of tags in collaborative tagging systems. Several papers proposed different approaches to construct conceptual hierarchies from tags collected from social websites. Mika (2007) uses a graph-based approach to construct a network of related tags, projected from either a user-tag or object-tag association graphs. Although there is no evaluation of the induced broader/narrower relations, the work provides a good suggestion to infer them by using betweenness centrality and set theory. Other works apply clustering techniques to keywords expressed in tags and use their co-occurrence statistics to produce conceptual hierarchies (Brooks and Montanez 2006; Zhou et al. 2007).

Brooks and Montanez (2006) argue that hierarchical structures which seem to match that created by humans can in fact be inferred from existing tags and articles in collaborative tagging systems. This may imply that folksonomies and traditional structured representations are not so opposed after all; rather, tags are a first step in helping an author or reader to annotate her information. Automated techniques can then be applied to better categorize specific articles and relate them more effectively to other articles. The method used is agglomerative clustering and consists of the following steps: the comparison of each tag cluster to every other tag cluster, using the pairwise cosine similarity metric. Each article in cluster one is compared to each article in cluster two, and the average of all measurements is computed. The two closest similarity clusters from the list of tag clusters are removed and replaced with a new abstract tag cluster, which contains all of the articles in each original cluster. This cluster is annotated with an abstract tag, which is the conjunction of the tags for each cluster.

This procedure is followed until there is a single global cluster that contains all of the articles. By recording the order in which clusters are grouped into progressively more abstract clusters, a tree that shows the similarity of tags can be constructed. Plangprasopchok and Lerman (2009) proposes a different approach for constructing folksonomies from user-specified relations on Flickr by statistically aggregating tags from different collections. This approach uses the shallow hierarchies created through the collection-set relations on Flickr. Authors argue that partial hierarchies are a good source of information for generating folksonomies and propose a simple statistical approach to resolve hierarchical relation conflicts in the aggregation process.

Another approach for the extraction of hierarchical semantics from social annotations is proposed by Zhou et al. (2007). A probabilistic unsupervised method is proposed, called deterministic annealing. This method performs a top-down approach on the flat tag space, beginning with the root node containing all annotations and splitting it to obtain clusters with narrower semantics.

Cattuto et al. (2008) perform an analysis on a large-scale snapshot of the popular social bookmarking system Delicious. To provide a semantic grounding of the folksonomy-based measures, tags of Delicious are mapped to synsets of WordNet (Markines et al. 2009) and use the semantic relations of WordNet to infer corresponding semantic relations in the folksonomy. In WordNet, the similarity is measured by using both the taxonomic path length and a similarity measure by Jiang and Conrath (1997) that has been validated through user studies and applications (Budanitsky and Hirst 2006). The use of taxonomic path lengths, in particular, allows to inspect the edge composition of paths leading from one tag to the corresponding related tags, and such a characterization proves to be especially insightful. Co-occurrence is a measure that extracts from the folksonomy a graph for tags, where edges are weighted with the number of times they co-occur (i.e., tags on the same resource).

The results can be taken as indicators that the choice of an appropriate relatedness measure is able to yield valuable input for learning semantic term relationships from folksonomies, i.e., (i) synonym discovery, (ii) concept hierarchy extraction, and (iii) the discovery of multi-word lexemes. The cosine similarity is clearly the measure to choose when one would like to discover synonyms. Cosine similarity delivers not only spelling variants but also terms that belong to the same WordNet synset. Both FolkRank and co-occurrence relatedness yields more general tags. This could be a proof that these measures provide valuable input for algorithms to extract taxonomic relationships between tags.

Key Applications

The main application domain is the area of social networking systems. However, this technique related to bookmarking has been already used in information retrieval and classification systems.

More generally, it is used in enterprise knowledge management.

Future Directions

The social bookmarking activity became very popular on the web from the second part of the 2000 decade with the rise of Web 2.0 and social networks. On the user side, the free-text bookmaking strategy may not change in the next years, and freedom of users will even increase. The main novelties will certainly arise from the back office processing of the annotations. One can expect higher precision and an increasing dynamicity in the management of vocabularies, special thanks to large-scale tag collection process, higher processing power, learning algorithms, and better consideration of the semantic.

Cross-References

- ▶ Folksonomies
- ▶ Tag Clouds
- ▶ Sentiment Quantification of User-Generated Content
- ▶ Semantic Social Networks Analysis

References

Abel F (2008) The benefit of additional semantics in folksonomy systems. In: PIKM '08: Proceedings of the 2nd PhD workshop on information and knowledge management. ACM, New York, pp 49–56. https://doi.org/10.1145/1458550.1458560

Abel F, Henze N, Krause D (2008) Ranking in folksonomy systems: can context help? In: CIKM '08: Proceeding of the 17th ACM conference on information and knowledge management. ACM, New York, pp 1429–1430

Bao S, Xue G, Wu X, Yu Y, Fei B, Su Z (2007) Optimizing web search using social annotations. In: WWW '07: Proceedings of the 16th international conference on world wide web. ACM, New York, pp 501–510

Bizer C, Lehmann J, Kobilarov G, Auer S, Becker C, Cyganiak R, Hellmann S (2009) Dbpedia – a crystallization point for the web of data. J Web Sem 7(3):154–165

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30 (1-7):107–117. https://doi.org/10.1016/S0169-7552(98)00110-X

Brooks CH, Montanez N (2006) Improved annotation of the blogosphere via autotagging and hierarchical clustering. In: Proceedings of the 15th international conference on world wide web (WWW 2006). Edinburgh, Scotland, pp 625–632

Budanitsky A, Hirst G (2006) Evaluating wordnet-based measures of lexical semantic relatedness. Comput Linguist 32(1):13–47. https://doi.org/10.1162/coli.2006.32.1.13

Cattuto C, Loreto V, Pietronero L (2006) Semiotic dynamics in online social communities. The European Physical Journal 46:33–37

Cattuto C, Benz D, Hotho A, Stumme G (2008) Semantic grounding of tag relatedness in social bookmarking systems. In: ISWC '08: Proceedings of the 7th international conference on the semantic web. Springer-Verlag, Berlin/Heidelberg, pp 615–631. https://doi.org/10.1007/978-3-540-88564-1_39

Delicious social bookmarking system (2003) http://delicious.com

Flickr – image and video hosting website (2004) http://flickr.com

Gemmell J, Shepitsen A, Mobasher M, Burke R (2008) Personalization in folksonomies based on tag clustering. In: Proceedings of the 6th workshop on intelligent techniques for web personalization and recommender systems

Golder SA, Huberman BA (2006a) Usage patterns of collaborative tagging systems. J Inf Sci 32(2):198–208. https://doi.org/10.1177/0165551506062337

Gruber T (2007) Ontology of folksonomy: a mash-up of apples and oranges. Int J Semant Web Inf Syst 3(2):1–11

Harvesting social knowledge from folksonomies (2006). Novel systems and models

Hotho A, Jschke R, Schmitz C, Stumme G (2006) Information retrieval in folksonomies: Search and ranking. In: Proceedings of the 3rd European semantic web conference, LNCS, vol 4011. Springer, Budva/Montenegro, pp 411–426

Java A, Song X, Finin T, Tseng B (2007) Why we Twitter: understanding microblogging usage and communities. In: Proceedings of the joint 9th WEBKDD and 1st SNA-KDD workshop 2007

Jiang J, Conrath D (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the international. conference on research in computational linguistics, pp 19–33

Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632. https://doi.org/10.1145/324133.324140

Last.fm – music website (2002) http://last.fm

Markines B, Cattuto C, Menczer F, Benz D, Hotho A, Stumme G (2009) Evaluating similarity measures for emergent semantics of social tagging. In: WWW '09: Proceedings of the 18th international conference on world wide web. ACM, New York, pp 641–650. https://doi.org/10.1145/1526709.1526796

Mathes A (2004) Folksonomies – cooperative classification and communication through shared metadata. Computer Mediated Communication – LIS590CMC. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html. Accessed 25 Feb 2013

Mendes PN, Passant A, Kapanipathi P, Sheth AP (2010) Linked open social signals. In: WI-IAT '10: Proceedings of the 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology – volume 01. IEEE Computer Society, Washington, DC, pp 224–231. https://doi.org/10.1109/WI-IAT.2010.314

MetaBrainz Foundation (2001) Musicbrainz – the open music encyclopedia. http://musicbrainz.org

Mika P (2007) Ontologies are us: a unified model of social networks and semantics. Web Semant 5(1):5–15. https://doi.org/10.1016/j.websem.2006.11.002

Naaman M, Boase J, Lai CH (2010a) Is it really about me?: message content in social awareness streams. In: CSCW '10: Proceedings of the 2010 ACM conference on computer supported cooperative work. ACM, New York, pp 189–192. https://doi.org/10.1145/1718918.1718953

Naaman M, Boase J, Lai CH (2010b) Is it really about me?: message content in social awareness streams. In: CSCW '10: Proceedings of the 2010 ACM conference on computer supported cooperative work. ACM, New York, pp 189–192. https://doi.org/10.1145/1718918.1718953

Plangprasopchok A, Lerman K (2009) Constructing folksonomies from user-specified relations on Flickr. In: Quemada J, Len G, Maarek YS, Nejdl W (eds) WWW. ACM, New York, pp 781–790

Qype – local directory service with social networking and user reviews (2006) http://qype.com

Rattenbury T, Good N, Naaman M (2007) Towards automatic extraction of event and place semantics from Flickr tags. In: SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, pp 103–110. https://doi.org/10.1145/1277741.1277762

Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: WWW '10: Proceedings of the 19th international conference on world wide web. ACM, New York, pp 851–860. https://doi.org/10.1145/1772690.1772777

Sen S, Lam SK, Rashid AM, Cosley D, Frankowski D, Osterhouse J, Harper FM, Riedl J (2006) Tagging, communities, vocabulary, evolution. In: CSCW '06: Proceedings of the 2006 20th anniversary conference on computer supported cooperative work. ACM, New York, pp 181–190. https://doi.org/10.1145/1180875.1180904

Wikipedia knowledge base (2001) http://wikipedia.org

Wu H, Zubair M, Maly K (2006) Harvesting social knowledge from folksonomies. In: Proceedings of the seventeenth conference on hypertext and hypermedia. ACM Press, New York, pp 111–114

Xi W, Zhang B, Chen Z, Lu Y, Yan S, Ma WY, Fox EA (2004) Link fusion: a unified link analysis framework for multi-type interrelated data objects. In: WWW '04: Proceedings of the 13th international conference on World Wide Web. ACM Press, New York, pp 319–327

Yeh I, Karp PD, Noy NF, Altman RB (2003) Knowledge acquisition, consistency checking and concurrency control for gene ontology (go). Bioinformatics 19(2):241–248

Yelp – local directory service with social networking and user reviews (2004) http://www.yelp.com

Zhou M, Bao S, Wu X, Yu Y (2007) An unsupervised model for exploring hierarchical semantics from social annotations. In: Aberer K, Choi KS, Noy NF, Allemang D, Lee KI, Nixon LJB, Golbeck J, Mika P, Maynard D, Mizoguchi R, Schreiber G, Cudr-Mauroux P (eds) ISWC/ASWC, Springer, Lecture notes in computer science, vol 4825, pp 680–693

Social Capital

Roger Leenders
Department of Organization Studies, Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, The Netherlands

Synonyms

Capital; Goodwill; Human capital; Social networks; Trust

Glossary

Human Capital	One's stock of competencies, knowledge, skill, education
Political Economics	The study of the relationship between politics and economics in society
Social Capital	Productive resources residing in and resulting from social networks
Social Liability	Obstructive resources residing in and resulting from social networks

Definition

By analogy with notions of physical capital and human capital, "social capital" refers to the features of social organization that facilitate coordination and cooperation for mutual or individual benefit. According to sociologist James Coleman (1990), “Like other forms of capital, social capital is productive, making possible the achievement of certain ends that would not be attainable in its absence. . . . In a farming community . . . where one farmer got his hay baled by another and where farm tools are extensively borrowed and lent, the social capital allows each farmer to get his work done with less physical capital in the form of tools and equipment.”

Social capital has been referred to as “the glue that holds society together” and is centrally concerned with the value and implications of relationships as a resource for social action. It is often considered to be the contextual complement to human capital. Social capital theory contends that returns to intelligence, education, and seniority depend considerably on a person’s location in the social structure of a market or hierarchy. While human capital refers to individual ability, social capital refers to opportunity (Burt 1997).

Introduction

Over recent years, the concept of social capital has become one of the most popular exports from sociological theory into everyday language (Portes 1998). As a term, “social capital” has become one of sociology’s trendiest terms, both in academic literature and popular publications. There seems to be a contagious quality to the concept’s predominant focus on positive aspects of human interrelationships. In addition, the concept is attractive to many by providing a broad framework that focuses on nonmonetary capital as a source of influence and prosperity. Below, the concept’s roots and historical development will be discussed and challenges and future directions are laid out.

Key Points

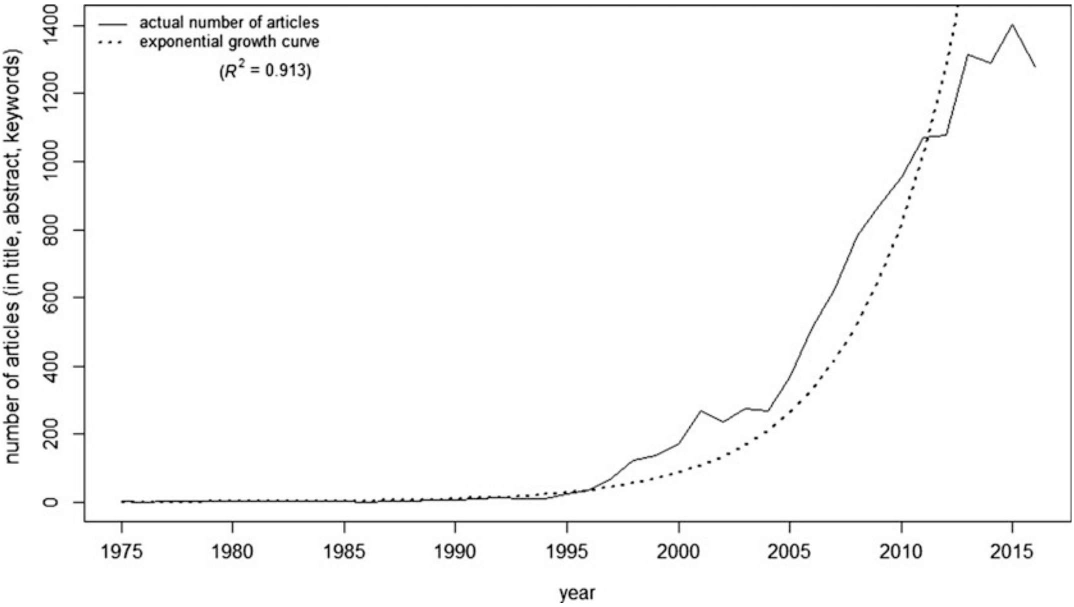
The concept of “social capital” has a history of over a century and has been adopted by a great variety of scientific fields. Many social science researchers and policy makers may have embraced the term because it provides a hardnosed economic feel while restating the importance of the “social” (Halpern 2005). At the same time, it lacks a generically accepted definition and agreement on its measurement.

One of the key merits of social capital is that it shifts the focus of analysis from the behavior of individual agents to the pattern of relations between agents. Closely linked to this is that the social capital concept links micro-, meso-, and macrolevels of analysis (Coleman 1990; Schuller et al. 2001), which makes social capital inherently a multilevel concept. Regardless of the scientific tradition the user of the concept comes from, the key idea behind social capital is that the (dynamic) pattern of interaction among actors (which can be individuals, groups, organizations, institutions, countries, et cetera) yields some (typically profitable) outcome, through the connectedness of these actors. This makes the collection of network ties that actors maintain into a valuable resource in itself, hence the term “social capital.”

Historical Background

Although the active use of the concept dates back to the nineteenth century, the social capital concept only became popular in the 1980s and this popularity accelerated in a major way in the 1990s. Figure 1 shows the number of academic articles abstracted in Scopus (1975–2016) with “social capital” in the title, abstract, and keywords. I have added an exponential growth curve, which clearly fits very well; this highlights that explosive growth of the concept (even when taking into account that the number of journals abstracted in Scopus has also increased over the years). Until around 1980, there were only very few papers that featured social capital. In the late 1990s, the World Bank started a Social Capital Initiative, a program with the aim of defining, monitoring, and measuring social capital (Grootaert and Van Bastelaer 2001).

Apart from economics and sociology, which are the concept’s original habitats, “social capital” has been adopted by a very wide range of



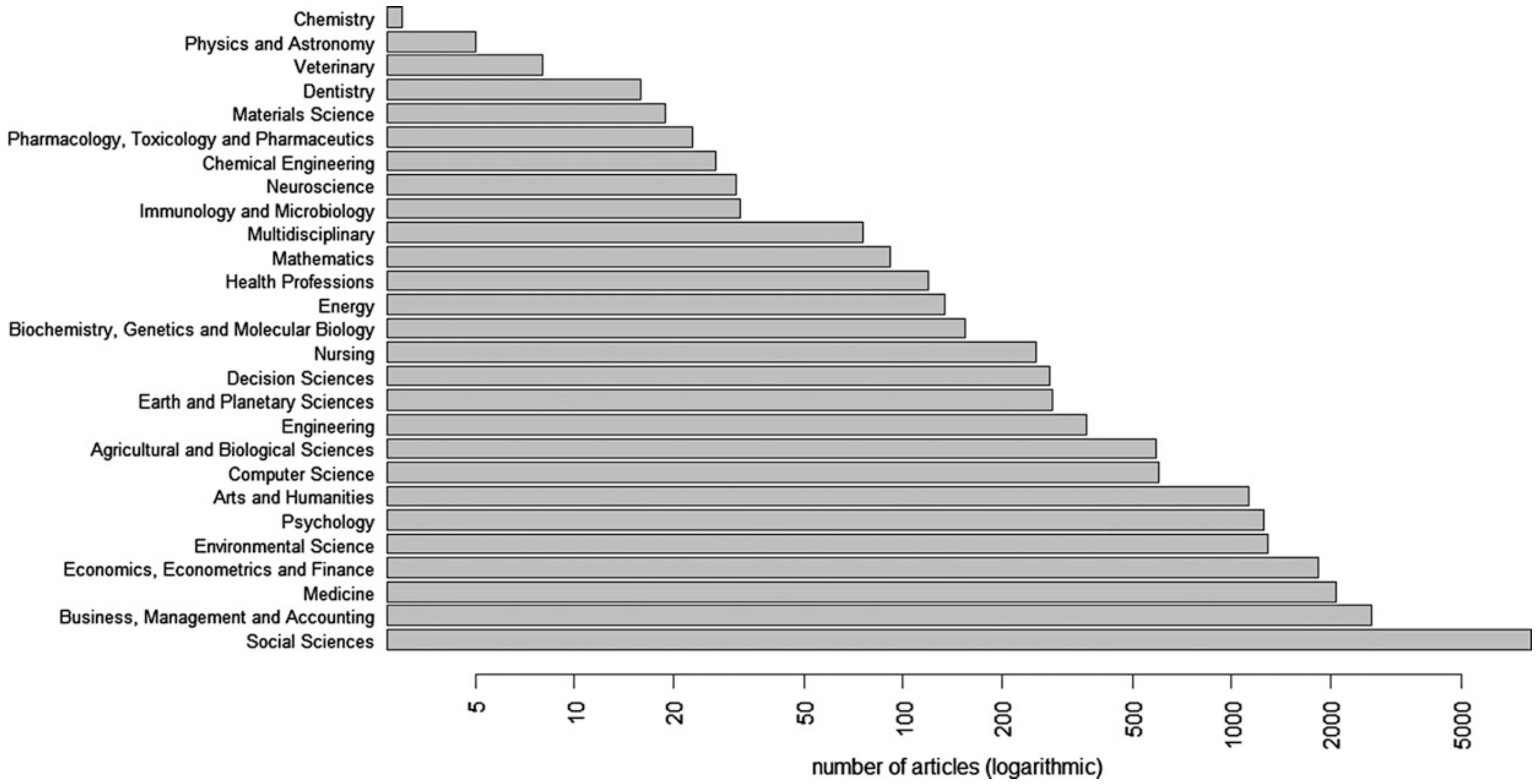
Social Capital, Fig. 1 The social capital concept in the academic literature (1975–2016)

disciplines. Figure 2 gives an overview of the fields (as categorized by Scopus) that published papers with "social capital" in their abstract. The figure shows that an astounding variety of academic disciplines publish papers that employ the concept, including the disciplines "biochemistry," "earth and planetary science," "agricultural science," "computer science," and "medicine" – not exactly social sciences. The lion's share of these papers considers social capital in terms of the ability of actors to secure benefits by virtue of membership in social networks or other social structures. Kwon and Adler (2014) conclude that the social capital concept has become mainstream, now that its basic thesis – that social ties can be efficacious in providing information, influence, support, solidarity – is no longer in dispute in any of these fields. They suggest that the specific topics that interest social capital researchers today becomes increasingly discipline specific, causing social capital citations to have arrived at an inflection point. Indeed, this is what Fig. 1 suggests. As Kwon and Adler (2014) contend, this is not a signal of decreasing interest in social capital, but a signal of scientific success and maturity. Clearly, the concept has caught on.

The Evolution of the Concept

There seem to be two largely separate histories of the "social capital" concept. The first starts in the late nineteenth century and runs until the beginning of the twentieth. The second starts around 1980 and is ongoing. Although today's use of the social capital concept differs from the way it was originally developed over a century ago, it is instructive to consider at least a little bit of the concept's original heritage.

Political economists were the first to use "social capital" in their writings. Alfred Marshall used the term in 1890 (Marshall 1890). Before that, so had John Bates Clark (1885), Henry Sidgwick (1883), and Karl Marx (1867). In their work, they opposed what they regarded as the unsocial point of view of classical political economy. In contrast to the "individual" point of view of capitalists, social capital was "capital from the social point of view." Social capital was seen as an aggregate of tools, inventions, improvements in land, roads, bridges, the organization of the State, and the skill and ability of humans. Also, immaterial elements were added to the concept, such as "goodwill" (Farr 2004, p. 22). As I will show later



Social Capital, Fig. 2 The scientific fields that study social capital in this entry, goodwill is still at the core of contemporary views of social capital.

The way in which the political economists of the nineteenth century thought about social capital paints a lively picture of corporations, trade unions, friendly societies, brotherhoods, guilds, communes, and cooperatives of endless variation. Through their joint ties, these cooperatives attempted to increase wages, share wealth, and render mutual aid (Farr 2004). Perhaps surprisingly, this picture is in many ways quite close to the social networks approach to social capital that has become the concept's dominant focus in contemporary research.

Farr (2004, p. 25) eloquently describes how the political economists' approach to social capital relates to its contemporary treatment:

> The political economists of the nineteenth century. . . . took capital from the social point of view. Today's social capitalists, apparently, take "the social" from capital's point of view. The one reflected an age coming to terms with capital, the other an age coming to capital for its terms. Then, "social capital" expressed an explicit antithesis to an unsocial perspective upon capital, now, an implicit antithesis to a noncapitalist perspective on society. "Social capital" was once a category of political economy in a period of its transformation, now one of economized politics, expressing the general dominance of economic modes of analysis in society and social science. But, in the long view, these perspectives may not be logical antinomies so much as two sides of the same coin. Both, surely, sought or seek to comprehend the social relations constitutive of modern capitalist societies, and to position capital as their governing asset. And both, significantly, did so in the very terminology of "social capital."

The contemporary stream of social capital studies flows primarily from the works of Pierre Bourdieu, James Coleman, and Robert Putnam. In *The Forms of Capital*, Bourdieu (1986) distinguishes between three forms of capital: economic capital, cultural capital, and social capital. To Bourdieu, social capital is made up of social obligations and connections. It is "the aggregate of the actual or potential resources linked to possession of a durable network of more or less institutionalized relationships of mutual acquaintance and recognition – or in other words, to membership in a group." In Bourdieu's view, social capital can be broken into two elements: the size of one's connections and the volume of capital (economic, cultural, or symbolic) in these connections' possession. To Bourdieu, social capital refers to "mutual acquaintance and recognition" (Bourdieu and Wacquant 1992). Bourdieu's social capital cannot be reduced to economic or cultural capital, nor is it independent of them: rather, it acts as a multiplier for economic and cultural capital.

For Coleman, social capital consists of "a variety of different entities with two elements in common: they all consist of some aspect of social structure, and they facilitate certain actions of actors – whether personal or corporate actors – within the structure" (Coleman 1990). One of Coleman's well-known examples is that of the Jewish diamond traders of New York. These merchants were able to have their diamonds appraised through their local networks without the need to resort to costly legal contracts (to safeguard against being cheated), because of the strength of the ties between their community members and the ready threat of exclusion if trust was violated. As a result, their social networks enabled the traders to increase their economic advantage (Coleman 1988). Coleman's approach derives from his interest in drawing together the insights from two disciplines: economics and sociology. Where in Bourdieu's work social capital serves to multiply economic and cultural capital, in Coleman's work an important function of social capital is in the multiplication of human capital. For example, Coleman extensively suggested that social capital has a profoundly beneficial effect on the acquisition of educational credentials (Schuller et al. 2001).

The fundamental difference between the Bourdieu and Coleman definitions lies in how and why the social processes develop. To Bourdieu, social processes are constrained by underlying economic organization; in his view, the potential of profit is the very reason for the solidarity that makes group existence possible. In fact, Bourdieu argues that these processes may become habitualized and then become reinforced by "habitus." To Coleman, on the other hand, they are created by the free will of individuals. In Coleman's approach, social capital is created by rational, purposeful individuals. As they attempt to maximize their individual opportunities, individuals freely choose to build networks to further their self-interest. Coleman views social capital as a form of contract: individuals must have trust that others will reciprocate their actions and will feel some social obligation to do so.

The disparity in the definition of social capital between Coleman and Bourdieu has consequences in the way social capital needs to be measured. An analysis based on Bourdieu's definition would need to include an understanding of the material conditions driving the formation of social processes. A Coleman-esque analysis needs only to consider motivation at the individual (or aggregated individual) level.

One of Coleman's chief contributions to the social capital literature may be in his relatively straightforward approach to the concept, which attracted widespread attention among social researchers. Bourdieu's work became popular only after it had been translated from French to English. Coleman's work has probably shaped the contemporary debate more than that of any other social capital author. Since it has been so prominent, it has also been widely criticized. Important criticism comes from Portes (1998) who charged Coleman with using a "rather vague definition" that "opened the way for re-labelling a number of different and even contradictory processes as social capital" (Portes 1998, p. 5). In particular, Portes argued for the need to draw a clear line between membership of social structures on the one hand and the resources acquired through such membership on the other.

Despite the reservations that have been voiced regarding his work, Coleman's contributions have been both influential and significant (to both practice and theory). Although now overshadowed by Robert Putnam in the wider public policy debate, Coleman has arguably had much greater influence over scholarship in the debate so far (Schuller et al. 2001, p. 8) and justifiably so.

Probably the currently most well-known author on social capital is Robert Putnam, who has appeared in televised talk shows, was invited to a Camp David meeting hosted by Bill Clinton, and was even featured in People magazine. The social capital concept is now used in a great many fields and Robert Putnam undoubtedly is the author whose work is cited across a wider range than any other (Schuller et al. 2001). Much of his work's popularity is due to the use of the compelling metaphor "Bowling Alone," characterizing the transformation of American social and political life during the postwar era (Putnam 1995, 2000). His work deployed the example of bowling as an activity which used to be highly associational, not only a source of recreational pleasure but also of social interaction. The metaphor refers t Putnam's observation that people are increasingly bowling outside of organized groups (I will come back to this below).

In this work, Putnam argued that a decline in civic culture was occurring in the USA since the 1960s, an idea that resonated with many of his readers. Controlling for political ideology, tax revenues, and several other conditions, Putnam concluded that the best predictor of governmental performance was a strong local tradition of civic engagement, measured by social capital variables such as "membership in voluntary organizations" and "voter participation in elections." Putnam was certainly not the first to call attention to the disintegration of American civic culture, but his work is clearly distinct from earlier authors through its specific focus on the eroding of social capital. According to Putnam, social capital "refers to the collective value of all 'social networks' and the inclinations that arise from these networks to do things for each other." In other words, it refers to features of social organization, such as trust, norms, and reciprocity, that can improve the efficiency and effectiveness of society by facilitating participants to act together to pursue shared objectives. Like Coleman, Putnam's definition strongly relies on networks and social linkages, but Putnam aggregates the social capital of individuals to a "collective social capital" of a population, state, or community. Putnam's main argument is that social capital is a key component to building and maintaining democracy; he notices its decline by, among other things, lower levels of trust in government and lower levels of civic participation. He makes the claim that television ("the only leisure activity where doing more of it is associated with lower social capital") and urban sprawl ("every 10 min of commuting reduces all forms of social capital by 10%") have had a significant role in making the USA far less "connected." In his analysis, Putnam focuses on the creation of civic norms, which lead to socioeconomic order; this is basically the reverse of Bourdieu's description of the relationship.

Notwithstanding the "celebrity status" his work gained him, Putnam's work has been extensively criticized. Putnam's arguments have been criticized as being circular and tautological: "social capital is simultaneously a cause and an outcome. It leads to positive outcomes, such as economic development and less crime, and its existence is inferred from the same outcomes" (Portes 1998). Other criticism (e.g., McLean et al. 2002) relates to his lack of sound empirical measures (of both social capital and many of the other variables in his empirical applications), inconsistent and incomplete derivation of his causal statements, the presence of implicit ideological underpinnings, and historical inaccuracy. In a 1998 special issue of the *American Behavioral Scientist*, several of his key empirical results were reexamined, many of which found no or only limited support.

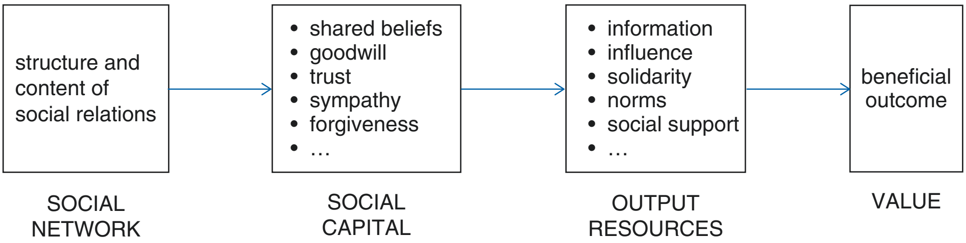
Perhaps the strongest points of criticism has been raised by Boggs (2001), who argues that "the author's iconic status does not prevent his book from being so conceptually flawed and historically misleading that it would seem to require yet another large tome just to give adequate space to the needed systemic critique." Boggs makes compelling arguments that social capital is not on the decline at all, for example, because "Putnam fails to consider the spread of newer, in many ways more interesting, civic phenomena over exactly that same time span – not only social movements but thousands of selfhelp and new-age groups, religious movements, and community organizations (resource centers, clinics, bookstores, periodicals, public interest groups, tenants associations, and so forth) often spawned by the larger movements" (Boggs 2001, p. 286). Boggs argues that social capital in fact has resurfaced, but in new (often apolitical) forms. He even takes issue with Putnam's Bowling Alone metaphor, which Putnam derived from the observation that participation in bowling leagues had declined by about 40%, a sign to Putnam that social networks were eroding. However, bowling activity might as well have been rechanneled to even more socially interactive sports like golf or soccer (Lemann 1996). In addition, Boggs argues that people simply switched from bowling in leagues to bowling in more informal groups of friends and relatives: it would be quite rare for people to actually bowl alone.

Basically, every aspect of Putnam's work has been criticized, sometimes dogmatically so and overly harsh. Indeed, Putnam's measures and concept of social capital are relatively weak and lack definition and consistency. Putnam's work popularized "social capital" quickly and across a great many disciplines. It was inevitable that many authors would blindly copy Putnam's less-than-fully-sound approach, yielding a surge of social capital-based papers that lack rigor and consistency. However, it is hardly fair to blame Putnam for this. Rather, he deserves some praise for bringing an academic concept to the political agenda and the general public in an easily understandable way. It is now up to the (scientific) community to develop sound definitions, measures, and causal models that bring the concept further.

A Model of Social Capital

The explosion of social capital research is having a predictable consequence: the term is proliferating meanings. It has been applied in so many different contexts that it has lost any distinct meaning. The concept has therefore been characterized as a "wonderfully elastic term" and a "notion that means many things to many people" that has taken on "a circus tent quality" (Adler and Kwon 2002, p. 18). As a consequence, social capital may be at a risk of ending up as a metaphor only.

The commonalities of most definitions of social capital are that they focus on social relations that have productive benefits. In the remainder of this entry, I will adopt a definition and a conceptual broad model of social capital that includes



Social Capital, Fig. 3 A conceptual model of social capital

both social capital sources and outcomes, multiple levels of aggregation, allows room for multidimensionality and multidisciplinarity, and can be extended to incorporate "time" as a variable in the social capital process. Of course, as with any definition, its leniency and agility is in how it is interpreted and applied, so I can make no claim that this definition fits with every research project in any discipline. In fact, it is unlikely that a definition of social capital can be made that fits that bill and would still be useful. My definition is:

> Social capital refers to the social resources that accrue to an actor (or a set of actors). Its source lies in the structure and content of the actor's social relations. These social resources facilitate the attainment of goals of the actor (or set of actors).

This definition is largely based on the definitions provided by Gabbay and Leenders (1999) and Adler and Kwon (2002). The term "social resources" refers to goodwill, norms, sympathy, trust, forgiveness, and shared beliefs that make alters willing or more likely to share resources that facilitate the attainment of some goal. As Adler and Kwon (2002) put it: "if goodwill is the *substance* of social capital, its *effects* flow from the information, influence, and solidarity such goodwill makes available." In the current definition, the information and influence (or other resources made available through the goodwill within an actor's social ties) need to support the attainment of some goal for it to constitute to social capital. The reason for this is that one can easily argue that available information that one may not need or is unable to understand cannot fruitfully be considered capital. Social capital requires the use or mobilization of the actor's social resources in *purposive actions* (Lin 2001).

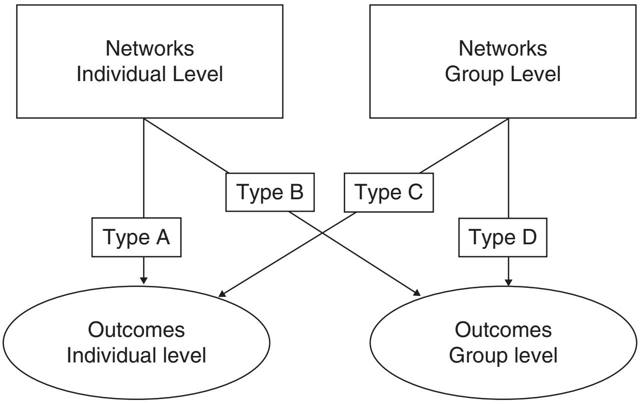
This view on social capital is depicted in Fig. 3. While it is consistent with a large proportion of the social literature, authors vary in their implementation of it. For example, the model categorizes "norms" as flowing from the goodwill that is inherent in the social network. However, there are also authors who view norms as inherent in social networks themselves – the presence of certain norms in a network can even be a reason for an actor to join the network – and these norms then make people likely to share knowledge with their alters; this would put norms in the social capital box. For this entry, it is not necessary to have a full-blown categorization that fits with the entire literature (which would be impossible, anyway) or with which all researchers would agree (which is equally outside the realm of possibility). What is important is that it fits with *most* of the literature and addresses the basic process that underlies social capital production and effects.

A Multilevel Concept

Much debate in the social capital literature deals with whether social capital resides at the individual level of aggregation or whether it is a group-level phenomenon. James Coleman focused mainly, though not exclusively, on the individual: a person's set of social ties provides that person with benefits. Robert Putnam, on the other hand, mainly focuses on the distribution of social ties within societies and studies how this high-level social structure produces outcomes at the level of the community.

The level-of-analysis discussion relates to two separate questions: what is the level of analysis at

Social Capital, Fig. 4 Two-level model of social capital



which the social ties reside and is their outcome accrued to the group or to the individual? A third, related, question addresses whether the social capital process is driven by *individual* goal seeking behavior or whether it is driven by *communities/ collectives* that have preferred outcomes they seek to fulfill.

When studying the social capital literature, one can only conclude that questions regarding the level at which the relevant social ties or their outcomes reside are unnecessarily restrictive and ignorant of empirical reality. Just as an individual can mobilize his personal contacts' resources for purposive action, so can a formal organization activate various resource networks to achieve its goals (Knoke 1999). There is ample evidence that individuals benefit from their own individual-level social networks as well as from the ties maintained by collectives they are part of. Similarly, collectives such as organizational groups draw the fruits from both their own connectivity with other collectives and the ties maintained by (some of) their individual members. In fact, fine-grained analyses indicate that more levels than just two are often relevant. Figure 4 shows the multilevel nature of social capital, with only two levels for simplicity. It is based on Gabbay and Leenders (2001), who classify the multilevel character of the literature into four main categories.

Type A refers to the lion's share of social capital research performed in organizational settings. Social structure and outcomes are both considered at the level of the individual. A typical example is the analysis by Ron Burt (1992), who shows that managers with disconnected networks achieve faster promotions to managerial positions. Other typical examples include studies on how people mobilize their array of direct and indirect relationships to accomplish personal goals such as finding jobs and achieving upward mobility (Granovetter 1973). Although he doesn't use the term social capital explicitly, Granovetter's (1973) argument is entirely about social capital; the mechanism Granovetter discusses is as follows. The friendship ties that people maintain provide them with alters who have the goodwill (and good will) to provide them with valuable information (for example, about possible interesting job opportunities). Friendship ties tend to vary in strength; consider the individual John who has drinks with some of his friends every night of the week, whereas John interacts with other friends only once or twice a year. The former set of friends (i.e., those who have strong ties with John) is likely to have the many of the same friends as John, whereas the latter set of friends (who are connected to John through weak ties) will likely socialize with many others than John does. The larger the set of weak friendship ties an individual has, the more varied the information will be that will reach this individual. Granovetter's work showed that individuals with many weak ties are therefore more likely to find a suitable job or be upwardly mobile. Thus, the *social network* of individuals provides them with the *social capital* that makes *output resources* (in this case, information) available that bestows them with the *value* of increased opportunities on the job market.

Social capital research of Type B refers to the benefits a collective (e.g., a company) draws from networks of individuals. For example, trustworthy relationships between employees of a firm and the employees of a bank may make it easier for that firm to secure a loan from that bank. Law firms, accounting firms, and consulting agencies considerably draw upon the networks senior consultants have with their clients to bring business to the firm. In firms, successful innovation often requires the firm to bring information about the market into the firm as well as new technology and other resources (such as financial resources). In many firms, much of this is achieved through ties that individual employees maintain with actors outside the firm; they then distribute these resources to the places in the firm that might need them. The effectiveness of this process for the firm highly relies on the number and quality of the ties that these employees maintain and the goodwill and knowledgeability of their network partners.

Type C research refers to situations where networks of corporations or other groups confer advantages to individuals. Examples include joint research and development projects between two firms that create new job opportunities for the individuals working in these positions or that produce the knowledge necessary to do one's job better and become eligible for a bonus or promotion. The networks of consulting firms can assist (junior) consultants in bringing in new projects, and the ties maintained by an academic department can be of great use to a junior academic in need of specialized expertise or research funding.

Finally, in Type D, organizations draw advantages from their own interorganizational networks. Joint venture relationships or joint marketing efforts, allowing for economies of scale or increased expertise, are examples of this type. Through interfirm relations, firms can gain timely and affordable access to new technology. For example, high prestige semiconductor firms tend to establish license alliances in which they gain the rights to produce and sell the proprietary technologies of competing organizations. It is because of their ability to certify the initiatives of other organizations (startups, in particular) that high prestige firms will gain access to the endeavors of others. The correspondence between prestige and access implies that prestigious firms enjoy a powerful positional advantage.

Whether social capital is seen from the group level or the individual level, Lin (2001) contends that all scholars remain committed to the view that, at the heart of things, it is the interacting members who make the maintenance and reproduction of social capital possible.

The Dark Side

Even though the predominant sentiment is that social networks are beneficial to individuals and groups, there is an increasing realization that there are profound negative sides to them as well. This is often referred to as "negative social capital," "the dark side of social capital," or, more in keeping with the "capital" part of the concept's name, "social liability" (Gabbay and Leenders 2001; Leenders and Gabbay 1999).

An example is violent or criminal gang activity that is encouraged through the strengthening of intragroup relations: this brings social liability to society. In a related fashion, the benefits that individuals can receive by virtue of membership in ethnic or religious communities can be experienced by others as exclusion from the same social and economic benefits. Tight coethnic bonds allow the restriction of the best jobs to members of the in-group, thus requiring the intervention of impersonal public agencies to break up the holds of these bonds and open up opportunities for others (Portes 2014).

Alternatively, membership in certain groups may require individuals to submit to group norms and obligations that reduce individual autonomy. An example of this is ISIS/Daesh, which creates highly embedded ties for anyone joining it, making it extremely difficult for anyone who wants to opt out and leave the organization.

Social liability shows why it pays off to explicitly relate social capital to goals or other outcomes: the same mechanism can provide outcomes that are productive for one goal but harmful for the achievement of another. For example, dense ties in a network of an R&D team provide the team members with quick access to knowledge, assisting the team in being efficient (social capital). However, research also shows that this comes at the expense of reduced levels of team creative performance: strong solidarity with in-group members may overembed actors, which reduces the flow of new ideas into the group, resulting in parochialism and inertia (Adler and Kwon 2002; Gargiulo and Benassi 1999).

Similarly, social structures can be beneficial to the fulfillment of a particular outcome at one point in time but become a liability later. An example of Type A research, Gargiulo and Benassi (1999) showed that relational structures that were helpful to managers in the past, later increased the number of coordination failures for which they were responsible. The network had become a constraint, impeding their performance. In his study on network marketing, Gabbay (1997) found that, for some entrepreneurs, strong ties combined with structural holes were beneficial at the initial stages of their business but were harmful for future expansion.

Grapevines – informal, person-to-person communications network of employees which are not officially sanctioned by the organization – are sources of rumors and gossip that spread quickly throughout an organization (Ellwardt et al. 2012). Management decisions may travel through grapevines days ahead of their official announcement. Because they feel threatened by it, managers often try to suppress the grapevine but find themselves confronted by a nearly impossible exercise. Grapevines and gossip networks, examples of individual-level social structure, can have detrimental effects on organization-level well-being and productivity (Type B).

Another source of potential organizational social liability is related to the resilience of personal networks. Managers in charge of (re)designing business processes often experience difficulties in breaking through the power structures that exist among the firm's employees. As a result, many attempts to redesign organizational processes fail or can only be implemented after long and painful struggles between higher management and employees (Type C).

An interesting case of social liability refers to the situation in which network ties that each provide social capital, together lead to negative outcomes. As an example, consider the situation in which multiple teams are jointly engaged in a new product development project, each focusing on a part of the overall design. A team whose task is affected by the task of another team will ask the relevant information from the second team (Gokpinar et al. 2010). These informal exchanges aggregate into a network of information flows between teams, which largely matches the network of task interdependencies between those teams. Considering that teams are typically dependent on the information of a larger set of teams, it is likely that two interdependent teams (say, teams A and B) have an overlap in the set of teams (say, C1, C2, C3, et cetera) that they also share information with. From a social capital point of view, most analysts would expect that the presence of these common third parties increase the knowledge-sharing norms and frequency in the network. However, Gargiulio and Sosa (2016) argue, thought-provokingly, that such common third parties may in fact induce interdependent teams to neglect exchanging information on their technical interdependencies. Further, they show that efforts to prevent such information-sharing neglect can itself result in coordination disruptions that are also undesirable, negatively affecting the performance or durability of the affected components and subsystems (Sosa et al. 2015).

At the fully organizational level (Type D in Fig. 4), long-standing relationships with customers may stifle the firm by monopolizing a disproportionate share of its resources, inhibiting the firm from forming relationships with alternative customers. Similarly, dense long-lived ties with other firms often effectively create blinders, reducing the firm's ability to see new (technological) developments that occur outside the firm's constrained field of vision.

Is It Capital?

By now, it will be clear to the reader why "social" is part of the term "social capital." A discussion that addresses the history and roots of the social capital concept, starting from its use in nineteenth-century political economics, also has to spend at least a few lines on the question of whether social capital is "capital," a question that increasingly appears in academic social capital articles. There appear to be two valid answers to this question, the second perhaps being the most to the point (except, perhaps, to economists).

The first answer addresses the nature of capital itself. Social capital does exhibit a number of characteristics that distinguish it from other forms of capital. Unlike physical capital, social capital can accumulate as a result of its use. Moreover, unlike financial capital, social capital erodes when it is not used. On the other hand, similar to other forms of capital, social capital is not costless to produce, requiring an investment that can be significant (Adler and Kwon 2002; Knoke 1999). The trusting relationships among the members of a sports club or professional organization can require years of meeting and interacting to develop (Grootaert and Van Bastelaer 2001). In addition, like all other forms of capital, social capital is a long-lived asset into which other resources can be invested, with the expectation of future benefits. It is also both appropriable and convertible and can be a substitute for or complement other resources. Based on these arguments, Adler and Kwon (2002) conclude that social capital "falls squarely within the broad and heterogeneous family of resources commonly called capital."

The second potential answer to the question is: "who cares?" The key attribute of capital is that it is an accumulated stock from which a stream of benefits flows. The view that social capital is an asset – that is, that it represents genuine capital – means that it is more than just a set of social organizations or social values. On the output side, it shows how things are getting done in society. On the input side, it shows that it requires a genuine investment to make society prosper. This is important, both from a conceptual and societal point of view. Aside from the intellectual joy the "is it capital" debate can undoubtedly provide to academists at cocktail parties, the social capital literature is probably best served by spending our efforts on developing better ways to measuring social capital and on improving the empirical and analytical rigor in social capital papers than by a debate about the semantic accuracy of the concept's name.

Key Applications

It will be clear from the discussion above that the social capital concept has pervaded scientific thinking across a great many and a great variety of disciplines. Likewise, the practical application of the concept is wide-ranging as well. Below, I will briefly outline four social capital applications that are among the most contemporary: mental and physical health, societal coherence, terrorism, and online networks.

Mental and Physical Health

Much research has been performed on the question how and to what extent social networks affect people's mental and physical health. The concept of social capital has become prominent in health promotion and health research (Nyqvist et al. 2013). Social networks have been found to be an important health resource (Kawachi et al. 2008; Nummela et al. 2009), in terms of both and individual's physical well-being and his mental state. Helliwel and Putnam (2004) refer to social capital as one of the "most robust correlates of subjective well-being. . .more important than economic differences when explaining life satisfaction differences."

A particular aspect of mental health, there is increasing research connecting social capital to happiness (e.g., Leung et al. 2011; Rodríguez-Pose and von Berlepsch 2014). For example, Helliwel et al. (2014) studied how the social capital affects the ability of people to deal with crises in ways that maintain and even improve experienced well-being.

The association between social capital and health is especially relevant because of the increase of the relative number of older people in society. Being socially integrated in society is an important health source, but older people are at greater risk of losing their partner and friends than younger people. This means that their social capital is more fragile: their volume of network ties decreases as network partners pass away and network partners also become less physically and mentally able to provide help, advice, and support. Therefore, social capital research increasingly focuses on ways in which people can remain connected to society (including both older adults and younger, able, others) as they age.

Societal Coherence

Traditionally, social capital thinkers – whether we consider the political economists of the late nineteenth century or contemporary popular social capital thinkers like Robert Putnam – have been concerned with the way social interaction acts as a glue to society. The idea is that the more people interact, and the more everybody is connected to everybody (within, say, six or seven steps), the more coherent, cohesive, and sustainable a society is.

More recently, researchers and policy makers have begun to look at the negative consequences of this social glue. The main observation is that social ties are not equally divided among all members of society; rather, the distribution tends to be extremely skewed. Regardless of whether one considers friendship networks on Facebook, offline friendships, corporate board memberships, or most other types of network ties, the number of ties maintained by most people is fairly close to the average, whereas a small group of people have an excessively large number of ties. This creates a ruling class and invokes inequality in society. The problem with this is that network inequality can be cumulative: high-status people benefit more from their networks than low-status people; this increases the status of high status members even further, at the expense of the (relative) status of the low-status individuals. In brief, the rich get richer, while the poor get poorer/remain poor. This effect is known as the "Matthew effect" (Merton 1968), after the "parable of the talents" in the biblical Gospel of Matthew. Updated analyses of how skewed the power distribution in the USA is found at http://whorulesamerica.net.

Other high-impact consequences of the glue that social capital brings to society is in its bringing and holding together of social movements (Edwards 2013). This can increase stability in society but has even more power to overturn current stability to establish new norms and policy; for example, in way the Arab Spring was supported by social capital in the Arab countries (e.g., Achilov 2013).

Terrorism

Closely related to the way in which social capital can spur social movements is the role of social capital in supporting terrorist movements. Daesh/ISIS is known for the way in which it uses its social capital to draft new recruits into the organization and the way in which social capital is then used to encapsulate these recruits inside the organization, so they become (further) brainwashed. In addition, the tight-knit internal networks make it hard for anyone to leave the organization, as every move one makes can and is monitored, with short paths to commanding members of the organization. Although little has been disclosed about it, it appears that governmental intelligence and law enforcement agencies are actively collecting massive amounts of data to uncover (and find vulnerabilities in) terrorist networks (e.g., Perliger and Pedazur 2011; Wu et al. 2014; Raymond 2015).

Online Networks

A fairly new application of the social capital concept relates to the changing nature of social relations in modern life. The works of Bourdieu, Coleman, and much of Putnam's work addressed social relations in a "bricks-and-mortar" world, in which social relations were largely created and maintained in a face-to-face manner. Especially over the last decade, social relations increasingly reside in cyberspace as well, and our social environment is transformed into a "clicks-and-mortar" world. Increasingly, social ties are built or maintained on Facebook, LinkedIn, Twitter, and other electronic platforms that are now frequently referred to as "social networks." One can easily see that the claim that social capital is declining can be refuted if one goes beyond the traditional interpersonal offline networks and includes network ties that live in cyberspace. However, can cyber ties be seen as equal to physical ties? In a recent meta-analysis, Liu et al. (2016) found that the involvement of people in social network sites does not straightforwardly yield social capital, the way one would expect time investment in offline interaction does.

We may have to reconsider findings from earlier research. For example, weak ties may no longer provide such a strong informational advantage when most job openings can easily be found by a single click of the mouse. It is conceivable that investment needed to crea online social capital is lower than the investment needed to build offline social capital; at the same time, the social capital (or social liability) that one draws from online ties may also differ from those drawn from offline relationships. An example of this type of question is the research by Colombo et al. (2015), who find that the online social relationships that entrepreneurs build by investing in projects hosted in the Kickstarter online community are fundamental to attracting backers and raising capital in one's own, later, online crowdfunding campaigns.

At any rate, social capital researchers cannot deny the increasing and pervading importance of cyberrelations if they are to study social capital in today's society and need to rethink their causal models and social capital measures.

Future Directions: Challenges

Although it was first used in the nineteenth century, social capital is still relatively immature as a concept, especially in its contemporary use. Its rapid proliferation has allowed a diversity of approach, definition, measurement, and causal logic (Schuller et al. 2001). Social capital is used in an extraordinarily wide range of disciplines. One consequence of this is that it is still largely unclear how social capital should be measured. Where such a diversity of definition exists, it is inevitable that an equivalent heterogeneity of measures is used (Schuller et al. 2001). A main challenge for the concept is for its users to develop useful and analytically sound and stable measures of social capital (and of the other parts of the social capital model). It is unlikely that any time soon a measure of social capital will (or can) be developed that is acceptable or useful to the wide range of contemporary social capital analysts, but one would hope that at least the empirical and statistical rigor of social capital research would be improved in the near future.

A second challenge relates to the multilevel aspect of social capital (see Fig. 4). The huge majority of studies stay within a single level, probably because it provides theoretical simplicity and restricts empirical complexity. However, many of the social capital outcomes that organizations enjoy, are in fact due to the networking done by its employees (Brand et al. 2017). A rich understanding of the way in which social networks and social capital are associated requires researchers to make the multilevel nature of the concept core to their theory and methods.

Third, as our world become increasingly connected (through online interaction, migration, holiday traffic, et cetera), it becomes important to further consider the cultural aspects of social networks and social capital. A given network structure might yield social capital in one culture, but not in another. More importantly, what happens when individuals or organizations from different cultures engage in mutual networking? The interaction could yield social capital to one party, but social liability to the other, purely based on each of their cultural heritages, norms, and embeddness. This type of research is already done, but is not well-developed yet.

Finally, an important challenge for the social capital literature is how to deal with temporal issues. Social networks are dynamic, those residing in cyberspace perhaps even more so. With social relations being dynamic, it is inevitable that social capital and its outcomes will experience dynamics as well. In general, there is a dearth of time-based theories in the social sciences (Leenders et al. 2016). Statistical models for network dynamics are now publicly available. However, appropriate theories of network dynamics are still lacking. For the rest of the social capital framework of Fig. 3, both theory and statistical models are missing almost entirely.

Conclusion

Social capital research addresses issues that are important to everyone, everyday. It addresses questions related to interpersonal trust (Bakker et al. 2006), quality of relationships in different contexts, and about equality and inequality in society (Kwon and Adler 2014). Even in academic fields like sociology or economics, there are only few topics that so consistently address issues that are of direct importance to every human being.

A successful future for the social capital literature requires an interdisciplinary approach that bridges some of the current different disciplinary perspectives. Political scientists, sociologists, and anthropologists tend to approach the concept of social capital through analysis of norms, networks, and organizations. Economists, on the other hand, tend to approach the concept through the analysis of contracts and institutions and their impacts on the incentives for rational actors to engage in investments and transactions. Each of these views has merits and the overarching challenge is to take advantage of the complementarities of the different approaches (Grootaert and Van Bastelaer 2001, p. 8). In this manner, we can turn the current proliferation of approaches, often seen as a weakness of the concept and threat to its viability, into a strength, providing the social capital literature with a bright and productive future.

Cross-References

- ▶ Community Evolution
- ▶ Counterterrorism, Social Network Analysis in
- ▶ Human Behavior and Social Networks
- ▶ Intraorganizational networks
- ▶ Link Dynamics and Community Formation in Social Networks
- ▶ Online Communities
- ▶ Personal Networks: Ties, Space, and the Internet
- ▶ Structural Holes
- ▶ Trust in Social Networks

References

Achilov D (2013) Social capital, Islam, and the Arab spring in the middle east. J Civ Soc 9(3):268–286

Adler PS, Kwon S-W (2002) Social capital: prospects for a new concept. Acad Manag Rev 27:17–40

Brand MJ, Croonen EPM, Leenders RTAJ (2017, forthcoming) Entrepreneurial networking: a blessing or a curse? Differential effects for low, medium and high performing franchisees. Small Business Economics

Bakker M, Leenders RTAJ, Gabbay SM, Kratzer J, van Engelen J (2006) Is trust really social capital? Knowledge sharing in product development projects. Learn Organ 13(6):594–605

Boggs C (2001) Social capital and political fantasy: Robert Putnam's "bowling alone". Theory Soc 30:281–297

Bourdieu P (1986) The forms of capital. In: Richardson JG (ed) Handbook of theory and research for the sociology of education. Greenwood Press, New York, pp 241–258

Bourdieu P, Wacquant LJD (1992) An invitation to reflexive sociology. University of Chicago Press, Chicago

Burt RS (1992) Structural holes: the social structure of competition. Harvard University Press, Cambridge

Burt RS (1997) The contingent value of social capital. Adm Sci Q 42:339–365

Clark JB (1885) The philosophy of wealth. Ginn, Boston

Coleman JS (1988) Social capital in the creation of human capital. Am J Sociol 94:95–120

Coleman JS (1990) Foundations of social theory. Belknap Press of Harvard University Press, Cambridge

Colombo MG, Franzoni C, Rossi-Lamastra C (2015) Internal social capital and the attraction of early contributions in crowdfunding. Entrep Theory Pract 39(1):75–100

Edwards B (2013) Social capital and social movements. The Wiley-Blackwell encyclopedia of social and political movements. Wiley-Blackwell, New York

Ellwardt L, Wittek R, Wielers R (2012) Talking about the boss. Group Org Manag 37(4):521–549

Farr J (2004) Social capital. Pol Theory 32:6–33

Gabbay SM (1997) Social capital in the creation of financial capital: the case of network marketing. Stipes Publishing, Champaign

Gabbay SM, Leenders RTAJ (1999) CSC: the structure of advantage and disadvantage. In: Leenders RTAJ, Gabbay SM (eds) Corporate social capital and liability. Wolters-Kluwer Academic Publishers, New York, pp 1–14

Gabbay SM, Leenders RTAJ (2001) Social capital of organizations: from social structure to the management of corporate social capital. In: Gabbay SM, Leenders RTAJ (eds) Social capital of organizations, research in the sociology of organizations. JAI Press, New York, pp 1–20

Gargiulo M, Benassi M (1999) The dark side of social capital. In: Leenders RTAJ, Gabbay SM (eds) Corporate social capital and liability. Wolters-Kluwer Academic Publishers, Norwell, pp 298–322
Gargiulo M, Sosa ME (2016) Common third parties and coordination disruptions in new product development organizations. J Prod Innov Manag 33(2):132–140
Gokpinar B, Hopp WJ, Iravani SMR (2010) The impact of misalignment of organizational structure and product architecture on quality in complex product development. Manag Sci 56(3):468–484
Granovetter MS (1973) The strength of weak ties. Am J Sociol 78:1360–1380
Grootaert C, Van Bastelaer T (2001) Understanding and measuring social capital: a synthesis of findings and recommendations from the social capital initiative. In: World bank social capital initiative working paper 24. http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTSOCIALDEVELOPMENT/EXTSOCIAL-CAPITAL/0,contentMDK:20194767~menuPK:418848~pagePK:148956~piPK:216618~theSitePK:401015,00.html
Halpern D (2005) Social capital. Polity Press, Cambridge
Helliwell JF, Putnam RD (2004) The social context of well-being. Philos Trans-R Soc London Series B Biol Sci 359(1449):1435–1446
Helliwell JF, Huang H, Wang S (2014) Social capital and well-being in times of crisis. J Happiness Stud 15(1):145–162
Kawachi I, Subramanian SV, Kim D (2008) Social capital and health. Springer, New York
Knoke D (1999) Organizational networks and corporate social capital. In: Leenders RTAJ, Gabbay SM (eds) Corporate social capital and liability. Wolters-Kluwer Academic, Norwell, pp 17–42
Kwon S-W, Adler PS (2014) Social capital: maturation of a field of research. Acad Manag Rev 39(4):412–422
Leenders RTAJ, Gabbay SM (1999) Corporate social capital and liability. Kluwer Academic, Boston
Leenders RTAJ, Contractor NS, DeChurch LA (2016) Once upon a time: understanding team processes as relational event networks. Organ Psychol Rev 6(1):92–115
Lemann N (1996) Kicking in groups. Atl Mon (10727825) 277:22–26
Leung A, Kier C, Fung T, Fung L, Sproule R (2011) Searching for happiness: the importance of social capital. J Happiness Stud 12(3):443–462
Lin N (2001) Social capital: a theory of social structure and action. Cambridge University Press, Cambridge
Liu D, Ainsworth SE, Baumeister RF (2016) A meta-analysis of social networking online and social capital. Rev Gen Psychol 20(4):369–391
Marshall A (1890) Principles of economics. Macmillan, London
Marx K (1867) Das Kapital: Kritik der Politischen Ökonomie. Marx-Engels Werke, Dietz Verlag, Berlin
McLean SL, Schultz DA, Steger M (2002) Social capital: critical perspectives on community and “bowling alone”. New York University Press, New York
Merton RK (1968) Social theory and social structure. Free Press/Simon and Schuster, New York
Nummela O, Sulander T, Karisto A, Uutela A (2009) Self-rated health and social capital among aging people across the urban–rural dimension. Int J Behav Med 16(2):189–194
Nyqvist F, Forsman AK, Giuntoli G, Cattan M (2013) Social capital as a resource for mental well-being in older people: A systematic review. Aging & Mental Health 17(4):394–410
Perliger A, Pedahzur A (2011) Social network analysis in the study of terrorism and political violence. Polit Sci Polit 44(1):45–50
Portes A (1998) Social capital: its origins and applications in modern sociology. Annu Rev Sociol 24:1–24
Portes A (2014) Downsides of social capital. Proc Natl Acad Sci 111(52):18407–18408
Putnam RD (1995) Bowling alone: America's declining social capital. J Democr 6:65
Putnam RD (2000) Bowling alone: the collapse and revival of American community. Simon & Schuster, New York
Raymond DJ (2015) Combating Daesh: a socially unconventional strategy. Naval Postgraduate School, Monterey
Rodríguez-Pose A, von Berlepsch V (2014) Social capital and individual happiness in Europe. J Happiness Stud 15(2):357–386
Schuller T, Baron S, Field J (2001) Social capital: a review and critique. In: Baron S, Field J, Schuller T (eds) Social capital: critical perspectives. Oxford University Press, Oxford, pp 1–38
Sidgwick H (1883) The principles of political economy. Macmillan, London
Sosa ME, Gargiulo M, Rowles C (2015) Can informal communication networks disrupt coordination in new product development projects? Organ Sci 26(4):1059–1078
Wu E, Carleton R, Davies G (2014) Discovering bin-Laden’s replacement in al-Qaeda, using social network analysis: a methodological investigation. Perspect Terrorism 8(1):57–73

Social Capital in Business

► Entrepreneurial Networks

Social Capital of Public Managers

► Managerial Networking

Social Classification

► Folksonomies

Social Collaborative Media

► Social Collaborative Media in Software Development

Social Collaborative Media in Software Development

Didi Surian[1,2] and David Lo[3]
[1]School of Information Technologies, University of Sydney, Sydney, NSW, Australia
[2]Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, NSW, Australia
[3]School of Information Systems, Singapore Management University, Singapore, Singapore

Synonyms

Collaborative project; Project management media; Social collaborative media; Software development

Glossary

Collaborative media
The forms of media which enable people (or participants) to collaborate and exchange information. Examples of collaborative media discussed in this entry are application software and online tools/media

Software projects
Projects in which software products are planned, developed, and monitored. Software projects have defined goals and methodology to follow in order to achieve the goals

Software development process
Process which is related to the development steps in software projects including designing, planning, management, and monitoring in realizing the project goals

Definition

In this entry, we discuss various collaborative media which are commonly used among software developers. We start by discussing common *communication channels* developers used. These communication channels are discussed in two groups: public and enterprise-wide media. We then elaborate *project management media* in coordinating and managing project activities. Finally, we discuss a number of *online knowledge resources*, i.e., collaborative/individual knowledge resources and social networks.

Introduction

In the past decade, a wide variety of collaborative media have been utilized by software developers in their projects. Regardless of the size of projects, typically multiple software developers participate in a software project. Interactions among software developers, which include communication and project coordination, play an important role in determining project outcome. These interactions are often supported by multiple collaborative media.

Communication, project management, and knowledge management are three main things that play important roles in determining project success (Kraut and Streeter 1995; Surian et al. 2010, 2013; Frese and Sauter 2014). Understanding different collaborative media that can effectively facilitate communication, project management, and knowledge management is an important first step before starting any software project. We believe that in order to facilitate effective interactions to achieve project goals, software developers should choose the correct collaborative media that suit the need of team members and the project itself.

Key Points

Collaboration is a key factor in software development. Collaborations could be manifested in various ways such as asking others for help, sharing ideas, exchanging information (Deal 2009), etc. Interactions among software developers often become a cornerstone in a software project as more and more software projects are developed in a collaborative fashion involving more than one developer. Moreover, many software projects today are developed by developers spread around the globe working on distributed teams. Even in the extreme case where a project only involves one developer, he/she may need to communicate with others to get feedback and help. This phenomenon creates an environment where information exchange process becomes crucial to the success of projects.

Collaborative media constitute the development process, especially in the planning, management, and monitoring steps where *active involvement* of the participants/collaborators is essential and being part of the process itself. Likewise, it is very nature that any methodology in software projects may still be redefined and refactored along the development process. As the participants become the focus of process, therefore, communication and information flow among the participants have to be carefully treated.

Historical Background

Software development activities need both individual commitments and socio-technical interactions (Kilamo et al. 2015). Therefore, a collaboration among software developers in a software project could be interpreted as a collective effort of a group of people who work together aiming for a common goal (i.e., project success) and sharing a common passion to improve the quality of their work by regular interactions. This definition puts collaborations among software developers to share common characteristics as collaborations in other communities of practice (Wenger and Snyder 2000), where:

1. Each member shares a common domain of interest with other members and shows a commitment to help one another.
2. Each member interacts, communicates, and builds relationships with the other members.
3. Each member engages with the other members and is involved in various problem-solving activities.

Numerous tools are available today to support the collaboration activities in a software development process (Singer and Schneider 2012; Storey et al. 2010). These tools come with various features to help developer to interact more effectively. The incorporation of social dimension to the software development process creates what is referred to as a “participatory culture” or “collaborative software development” (Whitehead et al. 2010) and which is also one important aspect in popular software development methodologies such as *Agile* (Larman 2004) and *DevOps* (Erich et al. 2014).

Communication Media

Public Media

In the early years of the twenty-first century, interactions among software developers mostly use basic voice-based and text-based communication technologies, e.g., telephone and SMS (Thurlow et al. 2004). While telephone and SMS are still among the most widely used communication tools, in recent years Internet-based communication tools such as instant messaging, video chat, e-mail, etc. gain much traction and are also widely used. Instant messaging services are prevalent nowadays as they provide a convenient and practical way for one-to-one interaction and group chat. Some of them provide multiple ways for their users to access the services (e.g., via web browsers and apps) and support multiple communication mode (e.g., text and voice calls). Some examples of these tools, which are publicly available, are Google Hangouts, Facebook messenger, Yahoo! messenger, Skype, eBuddy, ICQ, Tencent QQ (or known as QQ), Gadu-Gadu, Meebo, WhatsApp, Paltalk, Xfire, Viber, WeChat, LINE, KakaoTalk, Kik messenger, Tango, Nimbuzz, Hike messenger, MessageMe, Digsby, Adium, Pidgin, etc.

While instant messaging services are mainly used as communication tools, several of them have been equipped with file-sharing and document-editing capability, which allow developers to better collaborate with one another. Several examples of file-sharing/file-hosting services provided by big companies are Google Drive, Dropbox, Microsoft OneDrive, etc.

Enterprise-Wide Media

The use of social collaborative media to foster communication among employees has been implemented in many enterprises (DiMicco et al. 2008). This practice is generally known as "enterprise social networking" or "enterprise 2.0" (Makkonen and Virtanen 2015). While communication among employees through this media may lead to informal nonwork conversations, it has been shown that informal nonwork-related conversations (or commonly termed as *e-cheap talk*) help to promote cooperation and trust in software development teams (Wang and Redmiles 2015). The following list presents examples of social collaborative media developed by several companies:

1. IBM: Notes and Domino (e-mail client with social collaboration and business applications), Connections (business social network platform), Connections Cloud S1 (e-mail, instant messaging, online document editing, web conferencing, file-sharing and social business services), Sametime (instant messaging with integrated voice, data, and video), Forms (data collection)
2. Box Inc.: Box (file sharing)
3. Citrix: Podio (online collaboration), OpenVoice (audio conference), GoToMeeting (online meetings)
4. Atlassian: Confluence (document collaboration), HipChat (team chat, video, file sharing)
5. Microsoft: SharePoint and Yammer (company's private social network)
6. Jive: Jive-n (social collaboration), Jive Daily (company's social community), Jive Chime (company's instant messaging), Jive Circle (company's expertise finder)
7. Salesforce: Chatter (company's social network with file-sharing and expertise finder capabilities)
8. SAP: Jam Collaboration (cloud-based social collaboration)

Project Management Media

Software development process is a series of actions and steps including planning, organizing, staffing, directing, and controlling a software project to realize a working software system (Thayer and Yourdon 2000). This process includes a requirement to document, track, and record all project activities including project design and information changes. Face-to-face meeting and note keeping (by meeting notes, e-mails, etc.) are mostly used in facilitating project development. These two methods may be effective for small-scale software projects with small number of participants/collaborators. However, many software projects nowadays involve a large number of developers and stakeholders and have fast-changing needs. Furthermore, many software projects today are developed by software developers from geographically diverse locations.

In the past decades, more and more tools have been built to allow developers to better coordinate and manage projects. These tools incorporate various features to maximize the productivity and efficiency of developers in a software development process. One example tool that is commonly used in developing software is an integrated development environment (IDE). Some IDEs are platform independent and highly support collaborative software development. The social aspect of IDE has been shown to support effective coordination in a software team (Dourish and Bellotti 1992; Treude and Storey 2009).

In recent years, many web-based source code management platforms arise and are publicly available. Several of these platforms include SourceForge.net (http://sourceforge.net/), GitHub, Bitbucket, and many more. SourceForge.net was introduced as a platform which hosts many source code repositories with various features such as integrated issue tracking, discussion forum, and many more. Another web-based source code management system is GitHub. GitHub is based on Git system which enables distributed revision control to support parallel workflows and version tracking. Several previous studies have studied the collaboration among GitHub users (Dabbish et al. 2012; Marlow et al. 2013). Similar to GitHub, Bitbucket is also a web-based source code management platform that supports both Mercurial and Git. GNU Savannah is a collaborative software development management system which also supports Git, Subversion, and many other distributed revision control systems. Launchpad, Tigris.org (http://tigris.org/), and BerliOS (Berlin Open Source) are some other examples of software management platforms for open-source software. Some platforms are equipped with additional features such as crowdfunding (e.g., Bountysource) and with Agile project management feature (e.g., JavaForge). RubyForge is another platform dedicated to the Ruby programming language.

Online Knowledge Resources

With the emergence and growth of the Internet, many online knowledge resources are publicly available today. Some of these knowledge resources are individually managed, while others are community-driven services. In this section we discuss several online knowledge resources, i.e., (i) collaborative/individual knowledge resources and (ii) social networks.

Collaborative/Individual Knowledge Resources

A collaborative/individual knowledge resource is a website that allows a user or a group of users to add, delete, or modify a piece of content. An example of this resource is a wiki. A wiki is built using a wiki software/engine such as MediaWiki. Generally, each programming language may have its own wiki as a knowledge base for its users, such as wiki.python.org (https://wiki.python.org/) which provides knowledge resources for Python users. Some online sites such as code.wikia.com (http://code.wikia.com/) provide knowledge resources for several programming languages such as Java, Python, C++, Sawfish, etc.

Another knowledge resource media are news websites such as Slashdot, reddit, and Hacker News. Several studies have analyzed these news websites as a collaborative media (Lampe and Resnick 2004; Kunegis et al. 2009; Gilbert 2013). Slashdot was developed initially as a news website, but often discussions on Slashdot are valuable knowledge resources for software developers. Another example of a community-driven news website is reddit, which also allows discussions/comments by its users. Hacker News is a similar site focusing on computer science and entrepreneurship.

A blog (or weblog) is often designed by a single individual who would like to post something on the Internet. There are also multiauthor blogs where a group of people collaboratively write, manage, and edit the contents of a blog. Blogs are commonly used in software developer community as a documentation media and a discussion forum (Park and Maurer 2009; Parnin et al. 2013; Blood 2004; Cayzer 2004). Blogs are one of the important media which support software development as many developers often share their personal experiences such as API usages, bugs, solutions, and tutorials on their blogs to the benefit of many others.

A question-and-answer (Q&A) website is another form of knowledge resource which effectively supports knowledge transfer among people (crowdsourced knowledge) in a question/answer format. Like a discussion forum, people post questions on a Q&A website, and other users would answer the questions. Several Q&A websites which are frequently used as a knowledge sharing medium by software developers are Stack Overflow, Google Groups, Quora, Yahoo! Answers, WikiAnswers, etc. A recent study by Barua et al. shows that the discussion on a Q&A website varies, and discussions on some topics may lead to other discussions in some other topics (Barua et al. 2014). Wang et al. investigated interactions among software developers on Stack Overflow

(Wang et al. 2013). A study by Vasilescu et al. shows that the question-and-answer activity rates of developers correlate with their code changing activities in GitHub (Vasilescu et al. 2013).

Social Networks

Online social networks have also been used as collaborative media for software developers. The heterogeneous social structure of social networks contributes and plays an important role in determining how far and fast information spreads (Romero et al. 2013). Software developers have been using online social networks and creating virtual communities among them. One example of social networks is a microblogging service such as Twitter. Twitter has been used in software development process such as for sharing idea, discussing software/project issues, promoting blogs or other knowledge resources, etc. Several studies have investigated Twitter as one of the important media supporting software development (Bougie et al. 2011; Tian et al. 2012; Prasetyo et al. 2012; Wang et al. 2014). While some software developers are not hesitant to discuss their projects on public social networks, others are concerned on potentially disclosing confidential information on such networks; thus a number of social networks have been built for internal use in a company such as Confluence, Yammer, Jive, tibbr, etc.

Key Applications

Various types of social collaborative media have been used in practice to support the software development process. From the use of more common communication technologies such as telephone and SMS (Thurlow et al. 2004) to more specific tools such as collaborative-supported IDE, if it is chosen carefully by the development team, social collaborative media have been shown to be very useful to support the development process (Treude and Storey 2010). Our discussion in this entry only focuses on software development process; however, the application of social collaborative media could also be applied to other projects that involve collaborations among people.

Future Directions

As more and more collaboration media are available to support software developers in their software projects, there are several promising future research directions. Some collaboration media feature a ranking of their users (gamification). It is interesting to investigate the role and impact of high-status developers (developers who gain high rank based on their contributions), especially to project success/failure. Another interesting study would be an analysis on collaboration patterns on public collaboration media and internal-use (enterprise-wide) collaboration media. Yet another interesting direction is to develop automated systems to promote further collaborations on such media, e.g., by making recommendations or by sending suitable messages tailored to one's needs.

Cross-References

- ▶ Collaboration Patterns in Software Developer Network
- ▶ Social Interaction Analysis for Team Collaboration

References

Barua A, Thomas SW, Hassan AE (2014) What are developers talking about? An analysis of topics and trends in stack overflow. J Emp Soft Eng 19(3):619–654

Blood R (2004) How blogging software reshapes the online community. Commun ACM 47(12):53–55

Bougie G, Starke J, Storey MA, German DM (2011) Towards understanding Twitter use in software engineering: preliminary findings, ongoing challenges and future questions. In: Proceedings of the 2nd International Workshop on Web 2.0 for Software Engineering (Web2SE), pp 31–36

Cayzer S (2004) Semantic blogging and decentralized knowledge management. Commun ACM 47(12): 47–52

Dabbish L, Stuart C, Tsay J, Herbsleb J (2012) Social coding in GitHub: transparency and collaboration in an open software repository. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW), pp 1277–1286

Deal A (2009) A teaching with technology white paper: collaboration tools. Available online https://www.cmu.edu/teaching/technology/whitepapers/CollaborationTools_Jan09.pdf

DiMicco J, Millen DR, Geyer W, Dugan C, Brownholtz B, Muller M (2008) Motivations for social networking at work. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW), pp 711–720

Dourish P, Bellotti V (1992) Awareness and coordination in shared workspaces. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW), pp 107–114

Erich F, Amrit C, Daneva M (2014) A mapping study on cooperation between information system development and operations. PROFES 8892:277–280

Frese R, Sauter V (2014) Improving your odds for software project success. IEEE Eng Manag Rev 42(4):125–131

Gilbert E (2013) Widespread underprovision on Reddit. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW), pp 803–808

Kilamo T, Leppänen M, Mikkonen T (2015) The social developer: now, then, and tomorrow. In: Proceedings of the 7th International Workshop on Social Software Engineering (SSE), pp 41–48

Kraut RE, Streeter LA (1995) Coordination in software development. Commun ACM 38(3):69–81

Kunegis J, Lommatzsch A, Bauckhage C (2009) The Slashdot zoo: mining a social network with negative edges. In: Proceedings of the 18th International Conference on World Wide Web (WWW), pp 741–750

Lampe C, Resnick P (2004) Slash(dot) and burn: distributed moderation in a large online conversation space. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp 543–550

Larman C (2004) Agile and iterative development: a manager's guide. Addison-Wesley, Boston

Makkonen H, Virtanen K (2015) Social capital approach on Enterprise 2.0: a multiple case study. Technol Analy Strat Manag 27(10):1212–1225

Marlow J, Dabbish L, Herbsleb J (2013) Impression formation in online peer production: activity traces and personal profiles in GitHub. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW), pp 117–128

Park S, Maurer F (2009) The role of blogging in generating a software product vision. In: International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE), pp 74–77

Parnin C, Treude C, Storey MA (2013) Blogging developer knowledge: motivations, challenges, and future directions. In: Proceedings of the 21st International Conference on Program Comprehension (ICPC), pp 211–214

Prasetyo PK, Lo D, Achananuparp P, Tian Y, Lim EP (2012) Automatic classification of software related microblogs. In: Proceedings of the 28th IEEE International Conference on Software Maintenance (ICSM), pp 596–599

Romero DM, Tan C, Ugander J (2013) On the interplay between social and topical structure. In: the 7th international conference on weblogs and social media (ICWSM)

Singer L, Schneider K (2012) Influencing the adoption of software engineering methods using social software. In: International Conference on Software Engineering (ICSE), pp 1325–1328

Storey MA, Treude C, van Deursen A, Cheng LT (2010) The impact of social media on software engineering practices and tools. In: Proceedings of the FSE/SDP workshop on Future of software engineering research (FoSER), pp 359–364

Surian D, Lo D, Lim EP (2010) Mining collaboration patterns from a large developer network. In: Proceedings of the 17th Working Conference on Reverse Engineering (WCRE), pp 269–273

Surian D, Tian Y, Lo D, Cheng H, Lim EP (2013) Predicting project outcome leveraging socio-technical network patterns. In: Proceedings of the 17th European Conference on Software Maintenance and Re-engineering (CSMR), pp 47–56

Thayer RH, Yourdon E (2000) Software engineering project management, 2nd edn. Wiley-IEEE Computer Society Press, Los Alamitos

Thurlow C, Lengel L, Tomic A (2004) Computer mediated communication. Sage, London

Tian Y, Achananuparp P, Lubis IN, Lo D, Lim EP (2012) What does software engineering community microblog about? In: Proceedings of the 9th IEEE Working Conference on Mining Software Repositories (MSR), pp 247–250

Treude C, Storey MA (2009) How tagging helps bridge the gap between social and technical aspects in software development. In: Proceedings of the 32nd International Conference on Software Engineering (ICSE), pp 12–22

Treude C, Storey MA (2010) Awareness 2.0: staying aware of projects, developers and tasks using dashboards and feeds. In: Proceedings of the 32nd International Conference on Software Engineering (ICSE), pp 365–374

Vasilescu B, Filkov V, Serebrenik A (2013) StackOverflow and GitHub: associations between software development and crowdsourced knowledge. In: International Conference on Social Computing (SocialCom), pp 188–195

Wang S, Lo D, Jiang L (2013) An Empirical Study on Developer Interactions in StackOverflow. In: Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC), pp 1019–1024

Wang X, Kuzmickaja I, Stol KJ, Abrahamsson P, Fitzgerald B (2014) Microblogging in open source software development: the case of Drupal and twitter. IEEE Softw 31(4):72–80

Wang Y, Redmiles D (2015) Cheap talk, cooperation, and trust in global software engineering: an evolutionary game theory model with empirical support. Empir Soft Eng 21:2233–2267

Wenger EC, Snyder WM (2000) Communities of practice: the organizational frontier. Harv Bus Rev 78(1):139–146

Whitehead J, Mistrík I, Grundy J, van der Hoek A (2010) Collaborative software engineering: concepts and techniques. Springer, Berlin/Heidelberg

Social Communication Network: Case Study

Bin Liang[1], Bo Xu[1], Deqing Yang[2], Qi Liu[1] and Yanghua Xiao[1]
[1]School of Computer Science, Fudan University, Shanghai, China
[2]School of Data Science, Fudan University, Shanghai, China

Synonyms

Call network; Communication network; Interaction network; Mobile network; Social interaction

Glossary

BC	Betweenness centrality
CC	Closeness centrality
DC	Degree centrality
IM	Instant messaging
SNA	Social network analysis
SP	Shortest path

Definition

Social network is formally defined as a set of social actors that are connected by one or more types of relations (Wasserman and Faust 1994). Social actors can be individuals, groups, organizations, and even any units that can be connected to other units such as web pages, blogs, emails, instant messages, families, journal articles, neighborhoods, classes, sectors within organizations, positions, or nations (Furht 2010).

Social communication network is one of the most important social networks. In a social communication network, social actors are mostly persons, and the relationship between them is established for the purpose of communication. In a social communication network, social actors use communication tools such as mobile devices (mobile phone or other smart devices such as iPad), instant message software (WhatsApp, WeChat, Snapchat, etc.), email, and so on to communicate with each other. Social communication networks can be classified into different categories in terms of communication tool and network infrastructure. Typically, those running on telecom network with mobile phones as clients include mobile call network and short message network. Those running on Internet with PCs or smart devices as clients include instant messaging network, email network, and online social network (Twitter, Facebook, Instagram, etc.).

Introduction

Social network analysis (SNA) is one of significant steps towards understanding the behavior of actors in the networks. The first step of SNA is characterizing the structural properties of the networks. In general, different structural properties imply different principles of users' behaviors. Understanding user behavior is critical for the success of applications built upon these networks. Social communication networks underlie our daily lives. All of us are living in social communication networks. Thus, our communication behavior pattern is certainly embedded in these social communication networks. Hence, SNA on social communication networks is of special importance for user behavior understanding. After understanding the properties of these networks, the next key step is leveraging these properties for a successful application.

The purpose of this article is twofold. First, we showcase the common structural properties of social communication networks. Second, we showcase the applications on these networks.

Key Points

The structural properties of social communication networks in general can be explored from the following aspects: (1) social ties, (2) node strengths, (3) shortest paths and diameter, (4) centrality, and (5) assortativity. We show that social communication networks exhibit similar properties to a general social network but with some exceptions. For example, the degree of a typical social network follows power-law distribution. While in social communication network, the degree distribution follows Double Pareto-Log-normal (DPLN) distribution.

We have witnessed many successful applications on social communication networks including (1) economic development evaluation, (2) spammer detection, (3) email worm defense, and (4) social friend recommendation. The diversity of social ties in social communication networks is positively correlated to economic development (Eagle et al. 2010) which allows us to evaluate the regional economic development by the social tie diversity of inhabitants in the region. In general, social communication networks reflect people's interaction in social lives, but they may also include some spammers who generate garbage information. An email worm is a malicious Internet-borne program that uses your email account and address book to copy itself and spread to other account holders. By identifying the structure of the network, the state-of-the-art systems can successfully resist against the email worms. As justified through extensive empirical studies, the features of network structure and user behavior are generally more significant than content-based features to find or rank authorities/influencers in online social communication networks. This task is not only the key issue of media spread, but also useful to another important application in social networks, i.e., friend recommendation.

Historical Background

Social communication network analysis has been studied for a long time. It dates back to the experiment that was made by social psychologist Stanley Milgram in 1967. He selected two target people and found some volunteers to let them send letters from one target person to another by using their own social relationships. Some letters were successfully delivered from one target person to another target person. He found that the average distance of the success delivery is 6, implying that any two people are linked to each other on average via a chain with "six-degrees-of-separation." However, the experiment data are very small; the number of successful experiments is only 300. Hence, the reliability of the experimental result is an issue, which can be solved by statistical analysis on large-scale social communication network available nowadays.

Structural Measures on Social Communication Networks

In this part, we will review some important aspects to characterize social communication networks, including tie strength, node strength, shortest paths, centrality, and assortativity.

Tie Strength

Edges in a social communication network represent the social ties between two social actors. Typical social ties in social communication networks include sending messages, calling, and sending email.

The strength or weight of a tie between person *i* and person *j*, denoted by *tie(i, j)*, can be quantified by the aggregate time that *i* and *j* spent on the communication with each other or by the total number of communication times between them. These weights are denoted by w_{ij}^D (total duration of communication) and w_{ij}^N (total number of communication times), respectively.

Node Strengths

Based on tie strengths, node strengths can be defined as $s_i^N = \sum_{j \in N(v_i)} w_{ij}^N$ or $s_i^D = \sum_{j \in N(v_i)} w_{ij}^D$, where *N*(*vi*) is the neighbors of *i*. s_i^N represents the aggregate number of communication times. s_i^D represents the aggregate communication duration.

Shortest Paths

The shortest path between two nodes is one of simple paths with minimal length between them. The *diameter* of a network is the longest shortest path length over all node pairs. In general, it is hard to calculate the exact diameter on a large network due to its quadratic computational complexity. Diameter can be approximated with affordable cost (Magnien et al. 2009). It was found that the average shortest path on mobile social network of a city in China is 5.75 (Dong et al. 2009), which confirms the "six-degrees-of-separation" theory.

Centrality

Centrality measures the importance of users in social communication networks. There are several centrality measures, such as degree centrality, betweenness centrality, closeness centrality, and PageRank.

Degree Centrality

Degree of a node is the number of its connections. In social communication networks, it represents the number of contacts the user has. Hence, degree is a natural choice to measure the importance or the activity of the user. The nodes with higher degree are more important.

Betweenness Centrality

The betweenness of a vertex i is defined as the fraction of shortest paths that pass through i. More specifically, it is defined as

$$b_i = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(i)$ is the number of those paths that pass through i.

Closeness Centrality

The closeness of a node is the inverse of the average distance in the network from the node to all other nodes. Closeness reveals how long it takes for information to spread from one individual to others in the network. High-scoring node tends to have shorter shortest paths to other nodes in the network.

PageRank

PageRank (Brin and Page 1998) is also a useful centrality measure. It was first used by Google Search to rank websites in their search engine results. A high-scoring node tends to be linked from other important nodes or it is highly linked. The equation is as follows:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

where $p_1, p_2, \ldots, p_N$ are the nodes, $M(p_i)$ is the set of nodes that link to p_i, $L(p_j)$ is the number of outbound links on node p_j, and N is the total number of nodes.

Assortativity

A network is assortatively mixing if the nodes in the network that have many connections tend to be connected to other nodes with many connections (Newman 2002). That is, people with many friends are connected to others who also have many friends. This gives rise to degree-degree correlations in the network, implying that the degrees of two adjacent nodes are not independent (Onnela et al. 2007). The average nearest neighbors degree of a node vi is $k_{nn,i} = \frac{1}{k_i} \sum_{j \in N(v_i)} k_j$, where kj is the degree of vj. By averaging this over all nodes in the network of a given degree k, one can calculate the average degree of the nearest neighbors for degree k, denoted by $\langle k_{nn}|k \rangle$ (Pastor-Satorras et al. 2001). The network is assortatively mixing if $\langle k_{nn}|k \rangle$ increases with k and disassortatively mixing if it decreases as a function of k. On edge-weighted networks, weighted average nearest neighbor degrees are also used to characterize strength-strength correlations. There are two typical weighted versions: $k_{nn,i}^{N} = \frac{1}{s_i^N} \sum_{j \in N(v_i)} w_{ij}^{N} k_j$ and $k_{nn,i}^{D} = \frac{1}{s_i^N} \sum_{j \in N(v_i)} w_{ij}^{D} k_j$. It was found that in a typical social communication network, the degrees of two adjacent nodes are strongly correlated, while the strengths of two adjacent nodes in most cases are not (Onnela et al. 2007).

Mobile Call Network

Mobile phones are widely used in our daily lives. According to the statistics of the International Telecommunication Union, at the end of 2014, there are more than seven billion mobile subscriptions. Ericsson also forecasted that mobile subscription will reach 9.3 billion in 2018. In a mobile call network, each vertex is a mobile phone user, and each edge between two users means that they have at least one mobile call. There are several interesting structural properties and application in mobile call network.

Distribution

Power-law distribution is frequently observed in the real world. For example, a common property of many large real networks is their power-law degree distribution. This feature was found to be a consequence of two generic mechanisms: (1) the network grows continuously by the addition of new vertices and (2) new vertices attach preferentially to well-connected vertices. Several works (Saramaki and pekkaOnnela 2007; Nanavati et al. 2006) studied typical mobile call networks and found that their distributions with respect to degree and many other measures also follow the power-law distribution. However, the study on a larger mobile call network which consists of a million users and a hundred million calls shows that most distributions of this network significantly deviate from power-law and lognormal distribution but fit better to a less-known distribution: Double Pareto-Lognormal (DPLN) distribution (Seshadri et al. 2008). The distributions following DPLN include the number of phone calls per customer, the total talk time per customer, and the distinct number of calling partners per customer. Their study further reveals that the DPLN distributions can be consistently observed for networks in different snapshots.

Social Tie Diversity

Social networks form the backbone of social and economic life. Theoretical work suggests that the structure of social relations between individuals may affect personal life or economic development. For example, it was found that weak acquaintance relationships rather than close friendships are more helpful to find a job (Granovetter 1973, 1983). This is well known as weak tie theory. Eagle et al. (2010) found that the economical development is positively correlated to the diversity of social ties in a mobile call network. They use the following steps to study this relationship between network structure and economic development.

Step 1. Mobile network construction. They collected the national mobile call logs on August in 2005 in the UK. The data contain more than 90% of all mobile phones, which cover more than 99% of the populations and business landlines in the country. The constructed network consists of 65 million nodes and 368 million edges.

Step 2. Measuring diversity of social ties. They use Shannon entropy (Shannon 2001) to quantify diversity. They propose two diversity metrics: social diversity and spatial diversity. Social diversity of person *i* is defined as

$$D_{\text{social}}(i) = \frac{-\sum_{j=1}^{k} p_{ij}\log\left(p_{ij}\right)}{\log(k)}$$

where *k* is the number of *i* 's contacts and *pij* is the proportion of ’s total call volume that involves *j*. Spatial diversity of person *i* can be similarly defined as

$$D_{\text{spatial}}(i) = \frac{-\sum_{a=1}^{A} p_{ia}\log(p_{ia})}{\log(k)}$$

where *A* is the total number of telephone exchange areas and *pia* is the proportion of time *i* spends communicating with *a*-th of exchange area.

Step 3. Analysis. In this step, they compare the social tie diversity to economic development measured by IMD (Index of Multiple Deprivation) of UK in 2004 UK. IMD is a composite measure of relative prosperity of 32,482 communities encompassing the entire country, based on income, employment, education, health, crime, housing, and the environmental quality of each region. They found that the ranks of both social and spatial network diversity scores are positively correlated to IMD rank. For example, in Stoke-on-Trent, one of the least prosperous regions in the UK has one of the lowest diversity scores in the country.

Short Message Network

Short messages are sent from one mobile phone to another. This inherently is a network with users as vertices and edges as message-sending relationships. Short message has been one of the fastest-growing telecom value-added services worldwide. Due to its own characteristics, there are some special applications on it. In this section, we will showcase an SMS antispam system on this network.

As we know, short message service has greatly changed our lives. On some occasions, we prefer to short message rather than phone call to communicate with others. However, an accompanying problem is that message spam has also grown fast. Unsolicited and unwanted commercial advertisements may be sent as messages to mobile phone users. In some cases, fraud messages and rumor messages may be sent over the network.

Many solutions have been proposed to overcome this problem. Wang et al. (2010) uses spammers' behavior features and temporal features to detect spammers.

To distinguish legitimate users from spammers, they summarize many behavior patterns of spammers and normal users. In general, spammers tend to send a large number of messages to legitimate users. The legitimate users in general will not reply to an unknown phone number. Legitimate user's messaging targets are probably their friends, while spammer's messaging targets are mostly strangers. Furthermore, in a given period, a spammer usually sends only one message to one recipient. These social features can be quantified by out degree, mean weight on out edges, variance of weight on out edges, one weight ratio, reply ratio, partner ratio, and edge ratio (Wang et al. 2010).

Spammers can be divided into fraudulent senders and unauthorized advertisement agencies. They used temporal patterns to distinguish fraudulent senders from advertisement agencies. Fraudulent senders always submit a large number of messages in a short time period. Unauthorized advertisement agencies submit messages at a low frequency, and legitimate users submit messages at a medium frequency.

All these behavior features can be regarded as noncontent features which were also proved effective to detect spammers in SMS by other scholars. For example, Xu et al. (2017)) examined the effectiveness of various content-less features that range from network and to time-oriented categories by the classification models of SVM and KNN. Zhang et al. (2017) compared the performance of many machine learning algorithms on spammer detection by integrating multiple user behavior attributes. They found that random forest is a good choice to balance the tradeoff of the precision rate and the recall rate, and in an acceptable time.

Instant Messaging Network

Instant messaging programs, such as Microsoft MSN, ICQ, Yahoo Messenger, Tencent QQ, Skype, are very widely used in personal and business communications. A recent report (Leskovec and Horvitz 2008) estimated that approximately 12 billion instant messages are sent each day. These instant messaging tools imply instant messaging networks, where each vertex is a user, and each edge represents the contact relationship between users. Unlike other social communication networks, people tend to use informal language, loose grammar, abbreviations, and minimal punctuation in instant messages.

As a typical instant messaging network, MSN network was investigated in Leskovec and Horvitz (2008). They use anonymized data capturing a month of high-level communication activities in MSN system. They have found the following interesting facts.

First, they found that birds of the same feather flock together. People with similar properties tend to communicate with each other. For example, people with similar ages, the same languages, and geographically close locations tend to communicate with each other more frequently and longer. One of exceptions is gender. People tend to converse more frequently and with longer durations with those with opposite gender.

Second, they found that the instant messaging network is well connected and well clustered with 99.9% of the users belonging to the largest connected component, and the average clustering coefficient is 0.137. The average shortest path length among Messenger users is 6.6, which is half a link more than “6-degrees-of-separation.”

Third, they found that instant messaging network is very robust against intentional attack. They used different attack measures, such as average number of sent messages per user’s conversation, average duration of user’s conversation and so on, and simulate the intentional attach on the network.

Email Network

Email is a highly effective communication tool. It is inexpensive and only requires Internet connection. Hence, email network is one of most important social communication networks. However, email network is prone to some security issues.

“Email worms” (Zou et al. 2004) are one of the major Internet security threats for our society. There are many different types of worms (Weaver et al. 2003). One typical email worm works as follows: When an email user clicks a worm program in the attachments of a worm email, the worm program will find all the email address stored on this computer and sends the copies of itself to other users. Email worms spread on the email network, which is one of great security challenges to manage email networks.

Newman et al. (2002) found that there is little that computer system administrators can do to control the spread of a virus in the world at large through the study on the email network reconstructed from emails in a university. There are two main methods to defend against the “email worms”: random vaccination and targeted vaccination. According to Newman’s results, random vaccination has little effect on virus spread, while targeted vaccination seems pretty good. The effectiveness of vaccination strategy obviously depends on the network structure. Zou et al. (2004) investigated the influence of three topologies: power law, small world, and random graph. They found that on power-law topology, email worms spread more quickly, and targeted vaccination is more effective.

Microblogging Network

In recent years, microblogging systems have become the most popular communication tool in online social network, such as Twitter and Weibo. With the aid of microblogging systems, the web users not only release all kinds of posts ranging from public news to personal activities, but also share their opinions and spread information. There are many hot research spots in microblogging social networks, including how to find authority/influencers, spammer detection, information diffusion.

Weng et al. (2010) revealed that the presence of “reciprocity” can be explained by phenomenon of homophily. Based on their findings, they proposed an extension of PageRank algorithm to measure the influence of Twitter users, namely *TwitterRank*, which measures user influence by taking both the topical similarity between users and the link structure into account.

Hu et al. (2013) investigated how to collectively use network and content information to perform effective Twitter spammer detection and furthermore, proposed an optimization formulation that models the social network and content information in a unified framework. They also found that leveraging knowledge from various media, e.g., emails, SMS, and the web, can detect the spammers in microblogging more precisely through the empirical studies on Twitter datasets (Hu et al. 2014).

Through developing an efficient parameter fitting technique and applying the model to the emergence of URL mentions in Twitter, Myers et al. (2012) discovered that the information tends to “jump” across the network, which can only be explained as an effect of an unobservable external influence on the network. Specifically, only about 71% of the information volume in Twitter can be attributed to network diffusion, and the remaining 29% is due to external events and factors outside the network.

Key Applications

In general, analysis on social communication networks allows us to understand users' communication behavior. Specifically, these networks are helpful in the following applications. First, they can be used for the evaluation of regional economical development. The positive correlation between economical development and diversity of social ties in mobile call networks can be used for this application. Second, they can be used for spammer detection. Spammers have different features in various networks including short message network, email communication network, microblogging network, and so on, which allows us to detect spammers. Third, they can be used for friend recommendation in all online social platform on which users with similar properties tend to communicate with each other. Fourth, they can be used for resisting email worm attacks. Finally, as the focus of media spread, discovering influencers in social communication networks is also sensitive to both the network structure and user behavior features, of which the solution can be further used to accomplish social friend recommendation.

Future Directions

These social communication networks allow us to understand human behavior better. However, previous research on social communication network can be extended in many directions. First, social communication networks are inherently evolving. Investigation on the evolution pattern is more important in many real applications. Second, social communication networks contain abundant heterogeneous information. For example, users in instant messaging networks and microblogging networks have much profile information. How to employ the heterogeneous information for the analysis of these networks is one of promising directions?

Cross-References

▶ Mobile Communication Networks

References

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. In: International conference on world wide web, vol 30. Elsevier Science Publishers, Amsterdam, pp 107–117

Dong Z, Song G, Xie K, Wang J (2009) An experimental study of large-scale mobile social network. In: Proceedings of the 18th international conference on World wide web. ACM, pp 1175–1176

Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. Science 328(5981):1029–1031. https://doi.org/10.1126/science.1186605

Furht B (2010) Handbook of social network technologies and applications. Springer, New York

Granovetter M (1973) The strength of weak ties. Am J Sociol 78(6):1360–1380

Granovetter M (1983) The strength of weak ties: a network theory revisited. In: Sociological theory. Wiley, New Jersey, pp 201–233

Hu X, Tang J, Zhang Y, Liu H (2013) Social spammer detection in microblogging. Int J Confer Artif Intell 433–435:2633–2639

Hu X, Tang J, Liu H (2014) Leveraging knowledge across media for spammer detection in microblogging. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM, pp 547–556

Leskovec J, Horvitz E (2008) Planetary-scale views on a large instant-messaging network. In: Proceedings of the 17th international conference on World Wide Web, WWW '08, Beijing, China. ACM, New York, pp 915–924

Magnien C, Latapy M, Habib M (2009) Fast computation of empirically tight bounds for the diameter of massive graphs. J Exp Algorithm 13:10:1.10–110:1.9

Myers SA, Zhu C, Leskovec J (2012) Information diffusion and external influence in networks. In: ACM SIGKDD international conference on knowledge discovery and data mining pp 33–41

Nanavati AA, Gurumurthy S, Das G, Chakraborty D, Dasgupta K, Mukherjea S, Joshi A (2006) On the structural properties of massive telecom call graphs: findings and implications. In: Proceedings of the 15th ACM international conference on Information and knowledge management, CIKM '06. ACM, New York, pp 435–444

Newman MEJ (2002) Assortative mixing in networks. Phys Rev Lett 89(20):208701

Newman M, Forrest S, Balthrop J (2002) Email networks and the spread of computer viruses. Phys Rev E 66(3):035101

Onnela JP, Saramaki J, Hyvonen J, Szab G, de Menezes MA, Kaski K, Barabsi AL, Kertsz J (2007) Analysis of a large-scale weighted network of one-to-one human communication. New J Phys 9(6):179

Pastor-Satorras R, Vazquez A, Vespignani A (2001) Dynamical and correlation properties of the internet. Phys Rev Lett 87(25):258701

Saramaki J, pekkaOnnela J (2007) Structure and tie strengths in mobile communication networks. Proc Natl Acad Sci 104(18):7332–7336

Seshadri M, Machiraju S, Sridharan A, Bolot J, Faloutsos C, Leskove J (2008) Mobile call graphs: beyond power-law and lognormal distributions. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, Las Vegas, Nevada, USA, KDD '08. ACM, New York, pp 596–604

Shannon C (2001) A mathematical theory of communication. ACM SIGMOBILE Mobile Comput Commun Rev 5(1):3–55

Wang C, Zhang Y, Chen X, Liu Z, Shi L, Chen G, Qiu F, Ying C, Lu W (2010) A behavior-based smsantispam system. IBM J Res Dev 54(6):3

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Structural analysis in the social sciences. Cambridge University Press, Cambridge

Weaver N, Paxson V, Staniford S, Cunningham R (2003) A taxonomy of computer worms. In: Proceedings of the 2003 ACM workshop on rapid malcode. Washington, DC, USA, ACM, pp 11–18

Weng J, Lim EP, Jiang J, He Q (2010) TwitterRank: finding topic-sensitive influential twitterers. In: ACM international conference on web search and data mining, pp 261–270

Xu Q, Xiang EW, Yang Q, Du J, Zhong J (2012) Sms spam detection using non-content features. IEEE Intell Syst 27(6):44–51

Zhang B, Zhao G, Feng Y, Zhang X, Jiang W, Dai J et al. (2017). Behavior analysis based SMS spammer detection in mobile communication networks. In: IEEE international conference on data science in cyberspace, pp 538–543

Zou C, Towsley D, Gong W (2004) Email worm modeling and defense. In: Proceedings, 13th international conference on computer communications and networks, ICCCN 2004. Chicago, Illinois, IEEE, pp 409–414

Social Communities Competition

▶ Competition Within and Between Communities in Social Networks

Social Computing

▶ Collective Intelligence for Crowdsourcing and Community Q&A

▶ Social Interaction Analysis for Team Collaboration

▶ Social Provenance

▶ Web Communities Versus Physical Communities

Social Content Search

▶ Social Web Search

Social Data Analysis

▶ Temporal Analysis on Static and Dynamic Social Networks Topologies

Social Engineering

▶ Reconnaissance and Social Engineering Risks as Effects of Social Networking

Social Exchange

▶ Exchange Networks

Social Factor

▶ User Behavior in Online Social Networks: Influencing Factors

Social Graph Dataset

▶ Social Network Datasets

Social Grids

▶ Social Network Analysis and Organizational Multimodal Representation

Social Group Evolution

► Community Evolution

Social Groups in Crowd

Jarosław Wąs[1] and Krzysztof Kułakowski[2]
[1]Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering, Department of Applied Computer Science, AGH University of Science and Technology, Kraków, Poland
[2]Faculty of Physics and Applied Computer Science, Department of Applied Informatics and Computational Physics, AGH University of Science and Technology, Kraków, Poland

Synonyms

Family groups in crowd; Meso-scale structures in crowd

Glossary

Crowd	A temporary gathering of persons
Dyad	A group consisting of two persons
Small group	A group enough for all members to interact simultaneously. It is possible for all members to communicate or be acquainted with each other
Triad	A group consisting of three persons

Definition

According to different authors, a social group is a set of people with a common fate, with a direct interaction between them, with a social relationship between them, or who consider themselves as members of the same social category.

A crowd is a large group of people, gathered at one time and place, connected by a common aim.

A social group in crowd is defined as two or more human beings, who are allocated in the crowd and who are connected by and within social relationships.

Most frequently crowd consists of a set of social groups like couples, groups of friends or families, etc.

Introduction

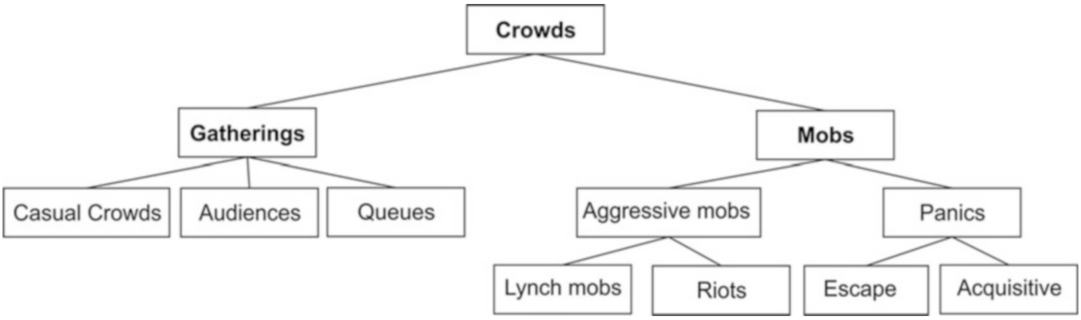
The occurrence of social groups in human crowds is a very common phenomenon. It is estimated that, depending on the situation, about 50–75% of people walk in groups and hold together in a crowd (Aveni 1977; Moussaïd et al. 2010).

A group in crowd is interpreted as two or more persons who are connected by interpersonal relationships. We can distinguish several methods of analyzing crowd dynamics and crowd behavior: from the macroscopic level when the crowd is treated as a whole, through the microscopic level when we consider the behavior and dynamics of individuals, and finally the analysis of the behaviors of particular groups of people in the crowd – the meso-level.

It seems that the meso-level analysis of crowd is crucial in terms of crowd behavior – seen as a whole (Moussaïd et al. 2010).

Crowd Classification

Social groups are part of vast majority of crowds. What is a crowd? Forsyth (2005) defines crowd as “a temporary gathering of individuals, who share a common focus on interest.” The occurrence and character of these groups depend on the type of crowd. According to Forsyth (2005), one can distinguish two different types of crowd: *gatherings* and *mobs*. Both of these types of crowds are different from the perspective of a situational context: gatherings mean more ordered aggregation of persons like audiences, queues, or street crowds, while a mob is described as an acting, disordered crowd, often aggressive in character. In



Social Groups in Crowd, Fig. 1 Crowd classification according to Forsyth (2005)

some cases, in a social group or crowd, the deindividuation phenomenon may occur, as described by Zimbardo (1969). In this situation, one can observe loss of self-awareness, reduced responsibility, loss of self-regulation, and emotional and impulsive behavior of individuals (Fig. 1).

Modeling and Simulation of Crowd Dynamics

Models of crowd dynamics and crowd behavior are used for simulations of evacuation, simulations of mass events, design of pedestrian traffic in public utility facilities, and, finally, in the entertainment industry (in the creation of movies, games, and special effects).

One can distinguish two main kinds of crowd dynamics models: macroscopic, where pedestrians are considered as fluid particles in hydrodynamics equations (Henderson 1974) and microscopic approach when pedestrians are considered as individuals or groups (Köster et al. 2011). Actually, most of the crowd dynamics models are based on the microscopic approach, as it entails the mapping of behavior of particular individuals or groups.

The most common method of microscopic modeling of crowd dynamics is social force model (Helbing and Molnár 1995). In this model, time and space are continuous. The model is based on differential equations equivalent to the Newton's second law of dynamics. There, each pedestrian i with mass of m_i moves according to the following equation:

$$m_i \frac{dv_i}{dt} = F_a$$

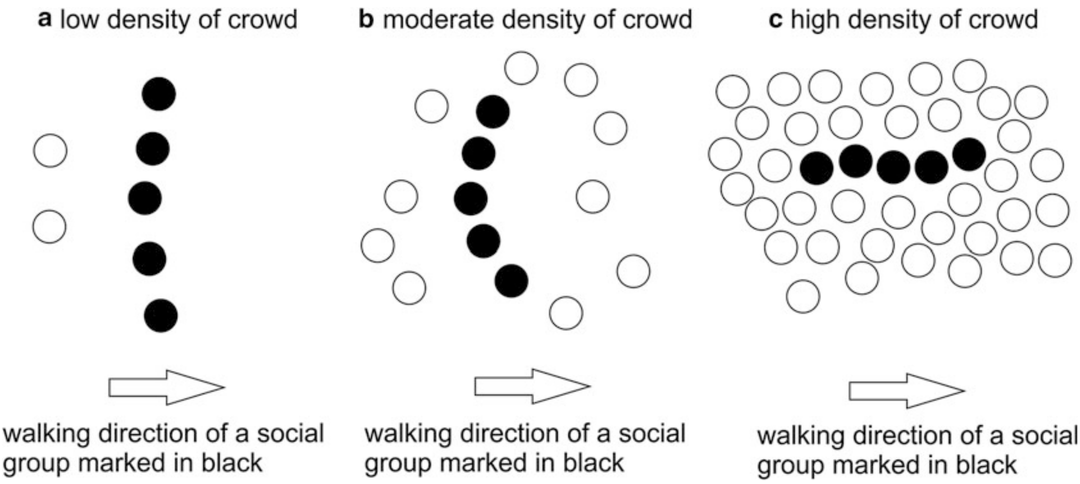
where
v_i – actual velocity of pedestrian i
m_i – mass of pedestrian i
F_a – the vector of forces, which takes into account personal desire force and interaction forces

The method has a lot of variants and extensions. The most important fact is that in using some variants of the method, we can take into account group attraction forces, and we can map group dynamics and group behavior (Moussaïd et al. 2010).

Another popular method of crowd dynamics modeling is cellular automaton (CA). It is a rule-based dynamical model, where time and space are discrete. Majority of implementations of CA models are interpreted as multi-agent systems. Pedestrians represented as autonomous agents are allocated in a lattice. Agents move on the lattice according to a transition rule f that modifies the configuration C_t of the agents allocated in the lattice (their environment) at certain interval Δt:

$$C_{t+1} = f(C_t)$$

The transition function f can be implemented using floor field (FF), which is defined on a supplementary lattices. Floor field is a set of rules assigned to the lattice, which take into account


a low density of crowd
b moderate density of crowd
c high density of crowd
walking direction of a social group marked in black
walking direction of a social group marked in black
walking direction of a social group marked in black

Social Groups in Crowd, Fig. 2 Exemplary patterns of walking small, social groups for low, moderate, and high density of flow. Members of walking social group are marked in black

different parameters, which determinate a type of the floor field: distance from a pedestrian to an aim (static floor field), following predecessors of a pedestrian (dynamic floor field) (Burstedde et al. 2001) or anticipating potential collisions (anticipation floor field) (Suma et al. 2012).

In modeling and simulation of crowd, a very important issue is the validation of flow, both qualitative and quantitative. A crucial stage of quantitative validation is the fundamental diagram, i.e., the relation between flow and density or velocity and density. The simulated scenario should be compared to experimental results (Zhang et al. 2011).

Most of the models based on cellular automata assume that the crowd is made up of individuals (Burstedde et al. 2001). It was demonstrated recently that it is possible to define rules of behavior for groups as well (Köster et al. 2011; Bandini et al. 2011).

Social Group Behavior

In a crowd we can often observe a situation, when a group of people intentionally walk together (for instance, family members, couples, or friends). It is in opposition to a different situation, when several proximate pedestrians fortuitously walk close to each other (Moussaïd et al. 2010). How to recognize these cases? Recognizing groups in crowd is not a trivial task from pattern recognition point of view. One can point out a correlation clustering procedure in pedestrians' trajectories as a possible methodology (Solera et al. 2015)

Social groups in the crowd behave differently depending on crowd density, size, and purposes of the particular group or type of crowd.

At low densities of crowd, group members usually tend to walk side by side creating a line perpendicular to the walking direction (Fig. 2a). When the density increases, the linear walking formation is bent forward, turning it into a V-like pattern (Fig. 2b). These spatial patterns can be well described by a model based on *social communication* among group members (Moussaïd et al. 2010). In very high densities, V-like patterns are transformed into a lane aimed toward the direction of motion (Fig. 2c).

The speed of the group is related to the density of the crowd and the size of the group. Speed of movement in high densities can be estimated based on the fundamental diagram (the relation between density and flow) (Seyfried et al. 2005; Chattaraj et al. 2009). It should be stressed that in smaller densities, there is a rule that the larger the group, the smaller the velocity of motion, and this relationship between size of the group and its velocity is linear (Klüpfel 2007; Moussaïd et al. 2010).

If we consider crowd as a social network with individuals represented as nodes of a graph (Fig. 2), then the social groups are interpreted as subgraphs, namely, network motifs (Milo et al. 2002; Juszczyszyn et al. 2009). Network motifs in this case may include two nodes (dyad) and three nodes (triads), up to a dozen nodes. One can notice complex patterns in crowd, among groups and individuals according to Hall's theory of proxemics (Hall 1966). Such patterns can be observed during different situations (Liu et al. 2016).

Sociological Aspects

Sociological aspect-specific properties and identity of a group as opposed to an individual have always been a matter of interest. *Senatores boni viri, senatus autem bestia* (senators are good men, but the Senate is an evil beast) was a common opinion in ancient Rome; this sentence indicates a contemporary knowledge belief held at the time that the Roman senate was different (worse) than a set of its members. The ability to control crowds is a necessary skill for politicians, artists, military, and religious leaders. Besides their personal charisma, their methods include rhetorics and dedicated institutions, like army, church, and theater. In more modern times, this role is played also by media. In the recent 10 years, interpersonal communication became possible by means of Internet and smartphones. The latter technology allows members of a large crowd interact in real time; this kind of communication transforms a crowd to an autonomous system. Its power has been demonstrated during a series of events known as the Arab Spring.

The idea of collective thinking has been of interest for centuries. The very concept of democracy relies on the belief that decisions taken by many people can be ahead of those by a single ruler. It makes sense to ask what is the quality of decisions made by a crowd? According to a traditional notion derived from Gustave Le Bon, crowd is irrational. Being part of the crowd, an individual gains feeling of power and loses responsibility. On the contrary, individualistic theories deny the existence of anything like collective thinking. Often, a famous statement by Floyd H. Allport is quoted: “The individual in the crowd behaves just as he would behave alone only more so.” Today, both these positions are rejected as unsupported by experience.

If crowd is different from a sum of individuals, interaction between them is an issue of primary importance. During this interaction, some individuals appear to be more influential than others. Once such a leader is able to create an impression of unanimity within some group; the group accepts his leadership. Simultaneously, the group itself is established, with identity defined by the content of accepted messages. It is clear that this acceptance depends, among others, on relations between personal features of individuals and leaders. During the process of group formation, the group identity continues to evolve. According to Jonathan H. Turner, the direction of this evolution is such as to minimize intragroup differences with respect to intergroup differences. Also, group decisions are more extreme and more risky than initial attitudes of group members; the effect is known as group polarization.

In a longer time scale, the very existence of groups has far-reaching consequences. However, immediate group formation in crowd can also be observed in situations of emergency, when communication is limited to a given area, as within the hearing range. According to the emergent norm theory by Ralph Turner and Lewis Killian, group members perceive their group as unanimous. As they follow the group action, the illusion of unanimity can become a self-fulfilling prophecy. If this is the case, we expect homogeneity of group behavior. On the contrary, differences between groups tend to grow and can lead to an intergroup hostility, even if the group formation is a purely random process.

Social Groups and Evacuation

Individual persons, as well as social groups, are the constitutive units of emergency evacuations.

When considered from the perspective of an evacuation study, a social group is characterized by several types of characteristics (Santos and Aguirre 2005):

- Operational context of a group (characteristic of their environment),
- Individual characteristics of members (age, sex, physical fitness, health and competences, etc.)
- Density understood as a function of physical space occupied by the group and the size of the group
- Relationships among the members including leadership, communication channels and cohesiveness, etc.

When a social group is faced with an emergency that makes it necessary to evacuate, the key parameter is the decision-making (Aveni 1977). One of the most important determinants of evacuation timing is the size of a group. The larger the group, the more difficult is to take the decision to begin evacuation as a response to the emergency. It should be stressed that "in the large group there will be more variation and differences of opinion and relevant experiences about what to do that must be reconciled before the emergent norm is created" (Santos and Aguirre 2005). Response time for an emergency is significant component of evacuation time, and it is determined by occurrence and characteristics of social groups. Recent studies on different sizes and different types of groups in crowd during evacuation confirmed that "evacuation efficiency decreases with the increase of the group size" (You et al. 2016; Gorrini et al. 2016).

Social groups also strongly influence the evacuation effectiveness during movement phase of evacuation, because members of the groups often create blocks (cluster patterns). In practice, single individuals (not members of the group) who want to overtake the group have great difficulties to do this, especially in constrained spaces like narrow corridors, stairways, etc. because they are exposed to "the set of norms and new statuses guiding the behavior of these collectivities which they cannot evade"(Santos and Aguirre 2005).

In extreme cases (caused by real or imagined reasons), evacuation situation may cause sudden, overpowering terror called panic, when an individual or the whole group is affected at once. In this case, relationships within the specified group have a large impact on behavior (Templeton et al. 2015). During panic, it is often possible to observe antisocial behaviors, but strong relationship within a group often leads to altruism and strong cooperation in the group according to the sentence "families survive together or die together" cited by (Köster et al. 2011).

Group Structure of Crowd

Usually a mere observation of a crowd does not allow inferring about existence and content of groups there, and dedicated tools are necessary. More than often, these tools are borrowed from statistical mechanics. In particular, the concept of modularity was proposed by Mark Newman. This quantity allows to evaluate if a given group structure is statistically meaningful.

Suppose we have a weighted network. Its nodes are pedestrians, and the links describe the similarities between the nodes. In particular, we can measure the trajectories of pedestrians; the value *w(l, j)* assigned to the link between nodes *l, j* is an absolute value of the Pearson's correlation coefficient between pedestrians *l* and *j*. The correlation can be calculated for positions or velocities or both. Suppose that we have a proposition of the network structure. This means that all nodes are divided into groups. The modularity *Q* is defined (Blondel et al. 2008) as

$$Q = \frac{1}{m} \sum_{lj} \left\{ w(l,j) - \frac{k(l)k(j)}{m} \right\} \delta(l,j)$$

where $k(j) = \sum_{l} w(l,j)$, $m = \sum_{lj} w(l,j)$, and $\delta(l,j)$ is equal to one if nodes l, j belong to the same group according to the proposed division; otherwise it is zero.

The challenge is to find the proposed division which gives the maximal value of Q. For large networks this task is NP-complete, then it cannot be treated with exhaustive methods. Instead, approximate algorithms have been proposed. One of them – the so-called agglomerative method – is to connect two nodes which give the largest Q; subsequent nodes are added according to the same rule. Starting from N separated nodes, we end up with a single connected cluster. Somewhere at this path, Q has a maximum; this is the approximated partition. This, however, does not prove that it is statistically meaningful. If the maximal value of Q is at least 0.3, our confidence increases.

We recommend Fortunato (2010) for a review.

Key Applications

Real-time monitoring of large gatherings supported by a software able to identify collective motion and interpersonal correlations should be helpful for predictions and prevention of stampede disasters, like the one in Duisburg, Germany, in 2010.

Future Directions

An interdisciplinary research conducted by sociologists, psychologists, physicists, computer scientists, and fire and transportation engineers can advance our understanding of mutual influence of majority and minority in crowd. In particular, the social mechanisms which rule this influence are not known yet (Brown 2000). Analysis of data on crowd dynamics collected during large gatherings is an example of a research strategy which can build bridges between social theory, field experiments, and applications.

Cross-References

- ▶ Collective Intelligence: Overview
- ▶ Human Behavior and Social Networks
- ▶ Motif Analysis
- ▶ Simulations
- ▶ Spatiotemporal Proximity and Social Distance

Acknowledgments The authors acknowledge financial support of FP7 project SOCIONICAL, No 231288.

References

Aveni AF (1977) The not-so-lonely crowd: friendship groups in collective behavior. Sociometry 40(1):96–99

Bandini S, Rubagotti F, Vizzari G, Shimura K (2011) A cellular automata based model for pedestrian and group dynamics: motivations and first experiments. Parallel Computing Technologies, pp 125–139

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech: Theory Exp 2008(10):P10008

Brown R (2000) Group processes: dynamics within and between groups. Blackwell, Malden

Burstedde C, Klauck K, Schadschneider A, Zittartz J (2001) Simulation of pedestrian dynamics using a two-dimensional cellular automaton. Physica A 295 (3-4):507–525

Chattaraj U, Seyfried A, Chakroborty P (2009) Comparison of pedestrian fundamental diagram across cultures. Adv Complex Syst 12(3):393–405

Forsyth DR (2005) Group dynamics. Wadsworth, Belmont

Fortunato S (2010) Community detection in graphs. Phys Rep 486(3–5):75–174, 486(4–5):75–174

Hall ET (1966) The hidden dimension. Doubleday, Garden City

Helbing D, Molnár P (1995) Social force model of pedestrian dynamics. Phys Rev E 51(5):4282–4286

Henderson LF (1974) On the fluid mechanics of human crowd motion. Transp Res 8(6):509–515

Gorrini A, Vizzari G, Bandini S (2016) Age and group-driven pedestrian behaviour: from observations to simulations. Collect Dyn 1–16, July 2016: https://doi.org/10.17815/CD.2016.3

Juszczyszyn K, Musial K, Kazienko P, Gabrys B (2009) Temporal changes in local topology of an email-based social network. Commun Inf 28(6):763–779

Klüpfel H (2007) The simulation of crowd dynamics at very large events. Calibration, empirical data, and validation. In: Pedestrian and evacuation dynamics. Springer, Berlin, pp 285–296.

Köster G, Seitz M, Treml F, Hartmann D, Klein W (2011) On modelling the influence of group formations in a crowd. Contemp Soc Sci: J Acad Soc Sci 6(3):397–414

You L, Hu J, Gu M, Fan W, Zhang H (2016) The simulation and analysis of small group effect in crowd evacuation. Phys Lett A 380(41):3340–3348, ISSN 0375-9601, https://doi.org/10.1016/j.physleta.2016.08.012

Liu X, Song W, Fu L, Lv W, Fang Z (2016) Typical features of pedestrian spatial distribution in the inflow process. Phys Lett A 380(17):1526–1534, ISSN 0375-9601, https://doi.org/10.1016/j.physleta.2016.02.028

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298(5594):824–827

Moussaïd M, Perozo N, Garnier S, Helbing D, Theraulaz G (2010) The walking behaviour of pedestrian social groups and its impact on crowd dynamics. PLoS One 5(4): e10047. https://doi.org/10.1371/journal.pone.0010047

Santos G, Aguirre BE (2005) A critical review of emergency evacuation simulation models. Workshop on building occupant movement during fire emergencies, pp 27–52

Seyfried A, Steffen B, Klingsch W, Boltes M (2005) The fundamental diagram of pedestrian movement revisited. J Stat Mech: Theory Exp 2005(10)

Solera F, Calderara S, Cucchiara R (2015) Socially constrained structural learning for groups detection in crowd. IEEE Trans Pattern Anal Mach Intell 38(5):995–1008

Suma Y, Yanagisawa D, Nishinari K (2012) Anticipation effect in pedestrian dynamics: modeling and experiments. Physica A 391(1–2):248–263

Templeton A, Drury J, Philippides A (2015) From mindless masses to small groups: Conceptualising collective behaviour in crowd modelling. Rev Gen Psychol 19(3):215–229. https://doi.org/10.1037/gpr0000032

Zimbardo PG (1969) The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. Nebraska symposium on motivation, pp 237–307

Zhang J, Klingsch W, Schadschneider A, Seyfried A (2011) Transitions in pedestrian fundamental diagrams of straight corridors and T-junctions. J Stat Mech: Theory Exp 6:4

Recommended Reading

Helbing D, Johansson A (2010) Pedestrian, crowd and evacuation dynamics. Encycl Complex Syst Sci 16:6476–6495

Social History of Computing and Online Social Communities

Joseph M. Kizza and Li Yang
Department of Computer Science and Engineering, The University of Tennessee-Chattanooga, Chattanooga, TN, USA

Synonyms

Crime in online communities; Ethics; Privacy; Social media; Social networks

Glossary

OSNs	Online social networks are social networks with underlining electronic communication infrastructure links enabling the connection of the interdependencies between the network nodes
mOSNs	Mobile OSNs are newer OSNs that can be accessed via mobile devices and can deal with the new mobile context
IMN	Instant messaging network supports real-time communication between two or more individuals
SNS	Social networking services

Definition

It is almost unimaginable that a modern person can live a meaningful life today without a mobile device as a conduit to an online social mesh of friends. These online social "gatherings" have slowly replaced the traditional face-to-face social gatherings that make us humans. While these online ecosystems are now packed with all sorts of interesting items that keep members coming back and new ones enrolling, the basic element of "presence" which transforms into "telepresence" in the virtual gatherings of any social gathering remains the same. The history of this amazing transformation of social gatherings mimics the history of social computing, the focus of this entry. The development of the different media of social gatherings and communication is linked with computer technological development. In fact the types of these social media developed in tandem with the development in computing technology. The history of social computing cannot be discussed comprehensively without talking about these online media. And these online social media cannot be justifiably discussed without investigating individual rights and how these media affect participants' individual attributes. Therefore, ethical, privacy, and security issues in these ecosystems are all involved in protecting personal privacy. On the central point of ethical implications of life in the social network, unlike in the traditional network, governance is not centralized, but community based with equally shared authority and responsibility by all users. But the mechanisms are not yet defined, and where they are being defined, it is still too early to say whether they are effective. The complexity, unpredictability, and lack of central authority are further enhanced by a virtual personality, anonymity, and multiple personality. These three characteristics are at the core of the social and ethical problems in online social networks in particular and cyberspace in general; the larger and more numerous these communities become, the more urgent the ethical concerns become.

Introduction

Social networks are at the core of social computing! In this discussion, therefore, the history of social computing is going to be discussed through the prism of social networks and their evolution into online social ecosystems, as we have them today. So a social network is a theoretical network where each node is an individual, a group, or an organization that independently generates, captures, and disseminates information and also serves as a relay for other members of the network. This means that individual nodes must collaborate to propagate the information in the network. The links between nodes represent relationships and social interactions between individuals, groups, organizations, or even entire society.

The concept of social networking is not new. Sociologists and psychologists have been dealing with and analyzing social networks for generations. In fact social networks have been in existence since the beginning of human. Prehistoric man formed social networks for different reasons including security, access to food, and the social well-being.

As Joseph Kizza (2013) observes, social networks begin with an individual reaching out to another individual or group for a social relationship of sorts, and it snowballs into a mesh of social relationships connecting many individuals and/or groups. In general, social networks come in all sizes and are self-organizing, complex, and agile depending on the nature of relationships in its links. As they grow in size, social networks tend to acquire specific elements and traits that make them different. These traits become more apparent as the network size increases. The type of social interactions, beliefs, and other traits usually limit the size of the social network. It is important to note that as a social network grows big, it tends to lose the nuances of a local system; hence, if certain qualities of the network properties are needed, it is better to keep the size under control. There are different social networks, but our discourse and focus will be on online social networks.

Online Social Networks (OSNs)

As computing technology developed, many types of social networks started evolving into online social networks. **Online social networks** (OSNs) are social networks but differ from all other social networks in that they depend on a mesh of underlining electronic communication infrastructure links enabling the connection of the interdependencies between the network nodes. The discussion in this entry will focus on these OSNs. In particular, we will focus on two types of online social networks (Kizza 2013):

The traditional OSNs such as Facebook and Myspace. Many of these can be accessed via mobile devices without the capability of dealing with mobile content.

The mobile OSNs (mOSNs) which are newer OSNs that can be accessed via mobile devices and can deal with the new mobile context.

The interdependency between nodes in the OSNs supports social network services among people as nodes. These interdependencies as relations among people participating in the network services define the type of OSNs.

Types of Online Social Networks

The growth of the OSNs over the years since the beginning of digital communication saw them evolving through several types. Let us look at the most popular types using a historical chronology (Kizza 2013):

Chat network. The chat network was born out of the digital chatting anchored on a *chat room*. The chat room was and still is a virtual room online where people "gather" just to chat. Most chat rooms have open access policies meaning that anyone interested in chatting or just reading others' chats may enter the chat room. People can "enter" and "exit" any time during the chats. At any one time, several threads of the public chats may be going on. Each individual in the chat room is given a small window on his or her communication device to enter a few lines of chat contributing to one or more of the discussion threads. This communication occurs in real time, and whatever one submits to the chat room can be seen by anyone in the chat room. Chat rooms also have a feature where a participating individual can invite another individual currently in the public chat room into a private chat room where the two can continue with limited "privacy." To be a member of the chat room, you must create a user name and members of the chat room will know you by that. Frequent chatters will normally become acquaintances based on user names. Some chat room software allows users to create and upload their profiles so that users can know you more via your profile.

Although chat rooms by their own nature are public and free for all, some are monitored for specific compliance based usually on attributes like topics under discussion.

With the coming of more graphical-based online services, the use of chat room is becoming less popular especially to youth.

Blog network. Another online social network is the bloggers network. "Blogs" are nothing more than people's online journals. Avid bloggers keep diaries of daily activities. These diaries sometimes are specific on one thread of interest to the blogger or a series of random logs of events during a specific activity. Some blogs are comment on specific topics. Some bloggers have a devoted following depending on the issues.

Instant messaging network (IMN), The IMN supports real-time communication between two or more individuals. Like chat rooms, each participant in the IMN must have a user name. To IM an individual, one must know that individual's user name or screen name. The initiator of the IM is provided with a small window to type the message, and the recipient is also provided with a similar window to reply to the message. The transcript of the interchange is kept scrolling up both users' screens. Unlike the chat room however, these exchanges of short messages are private. Like in chat networks, some IMN allows users to keep profiles of themselves.

Online social networks (OSNs). These are a combination of all the network types we have discussed above and other highly advanced online features with advanced graphics. There are several of these social networks including Facebook, Twitter, Myspace, Friendster, YouTube, Flickr, and LinkedIn. Since these networks grew out of those we have seen before, many of the features of these networks are those we have discussed in the above networks. For example, users in these networks can create profiles that include their graphics and other enclosures and upload them to their network accounts. They must have a user name or screen name. Also communication, if desired, can occur in real time as if one is using chat or IM capabilities. In additional to real time, these networks also give the user the delayed and archiving features so that the users can store and search for information. Because of these additional archival and search capabilities, network administrators have fought with the issues of privacy and security of users as we will see later in this entry. As a way to keep users' data safe, profiles can be set to a private setting, thus limiting access to private information by authorized users.

Online Social Networking Services

An online social networking service is an online service accessible via any internet-enabled device with the goal of facilitating computer-mediated interaction among people who share interests, activities, backgrounds, or real-life connections. Social networking services (SNS) offer users functionalities for identity management (i.e., the representation of the owner, e.g., in form of a profile) and enable furthermore to keep in touch with other users (and thus the administration of own contacts) (Koch et al. 2007).

Most online social network services consist of:

User profile management: People construct user profile in social networks for a particular group of audience or a particular task. The profile is used and managed as a social identity that they used to present to each other and analyze each other.

Social or business links of interests: Users of social networks can search experts or peers based on different criteria such as interest, company, or name. They can also proactively receive recommendations for contacts of interests from social networks.

Context awareness: This helps to identify common backgrounds of users in social networks. For example, users could have common contacts, common interests, the same university, or the same company. Context awareness helps to build trust among users, which are essential for a successfully collaboration (Kramer 1999).

Contact management: This combines all functionalities that manage and maintain users' personal network. Examples include tagging people and access restrictions to profile in social networks.

Network awareness: This includes any change or update of users in one's personal network. This includes awareness of indirect communication, News Feeds, and user notification.

Exchange: This enables information sharing directly (e.g., messages) or indirectly (e.g., photos or messages via bulletin boards). Examples of exchange in social networks include messages and photo albums.

Currently, the most popular online social network services fall in categories that range from friends based, music and movie, religion, business, and many other interests. In each of these categories, let us give a sample of the current services:

General and friend-based social networks – Facebook

Myspace
Hi4

Movie and music social networks

Last.fm
Flixster
iLike

Mobile social networks

Dodgeball
Loopt
Mozes

Hobby and special interest social networks

ActionProfiles
FanIQ

Business social networks

LinkedIn
XING
Konnects

Reading and books social networks

Goodreads
Shelfari
LibraryThing

The Growth of Online Social Networks

OSNs have blossomed as the Internet exploded. The history and the growth of OSNs have mirrored and kept in tandem with the growth of the Internet. At the infant age of the Internet, computer-mediated communication services like Usenet, ARPANET, LISTSERV, and bulletin board services (BBS) helped to start the growth of the current OSNs as we know them today. Let us now see how these contributed to the growth of OSNs.

BITNET was an early world leader in network communications for the research and education communities and helped lay the groundwork for the subsequent introduction of the Internet, especially outside the US (Fox 2000). Both BITNET and Usenet, which were invented around the same time in 1981 by Ira Fuchs and Greydon Freeman at the City University of New York (CUNY), were both "store-and-forward" networks. BITNET was originally named for the phrase "Because It's There Net," later updated to "Because It's Time Net" (Fox 2000). It was originally based on IBM's VNET e-mail system on the IBM virtual machine (VM) mainframe operating system. But it was later emulated on other popular operating systems like DEC, VMS, and Unix. What made BITNET so popular was its support of a variety of mailing lists supported by the LISTSERV software (ICANN 2005).

BITNET was updated in 1987 to BITNET II to provide a higher bandwidth network similar to the NSFNET. However, by 1996, it was clear that the Internet was providing a range of communication capabilities that fulfilled BITNET's roles, so CREN ended their support and the network slowly faded away (ICANN 2005).

Bulletin board system (**BBS**). A bulletin board system (BBS) is software running on a computer allowing users on computer terminals far away to login and access the system services like uploading and downloading files and reading news and contribution of other members through e-mails or public bulletin boards. In "Electronic Bulletin Boards, A Case Study: The Columbia University Center for Computing Activities," Janet F. Asteroff (Evolving the High Performance Computing and Communications Initiative to Support the Nation's Information Infrastructure—Executive Summary 2013) reports that the components of computer conferencing that include private conferencing facilities, electronic mail, and electronic bulletin boards started earlier than the electronic bulletin board (BBS). Asteroff writes that the concept of an electronic bulletin board began from 1976 through ARPANET at schools such as the University of California at Berkeley, Carnegie Mellon, and Stanford University. These electronic bulletin boards were first used in the same manner as physical bulletin boards, i.e., help wanted, items for sale, public announcements, and more. But electronic bulletin boards soon became, because of the ability of the computer to store and disseminate information to many people in text form, a forum for user to debate on many subjects. In its early years, BBS connections were via telephone lines and modems. The cost of using them was high; hence, they tended to be local. As the earlier form of the World Wide Web, BBS use receded as the World Wide Web grows.

LISTSERV. It started in 1986 as automatic mailing list server software which broadcasts e-mails directed to it to all on the list. The first LISTSERV was conceived of by Ira Fuchs from *BITNET* and Dan Oberst from EDUCOM (later EDUCAUSE) and implemented by Ricky Hernandez also of EDUCOM, in order to support research mailing lists on the *BITNET* academic research network (Kizza 1999).

By the year 2000, LISTSERV ran on computers around the world managing more than 50,000 lists, with more than 30 million subscribers, delivering more than 20 million messages a day over the Internet (Kizza 1999).

Other online services. As time went on and technology improved, other online services come along to supplement and always improve on the services of whatever was in use. Most of the new services were commercially driven. Most of them were moving toward and are currently on the web. These services including news, shopping, travel reservations, and others were the beginning of the web-based services we are enjoying today. Since they were commercially driven, they were mostly offered by ISPs such as AOL, Netscape, Microsoft, and the like. As the Internet grew, millions of people flocked onto it, and the web and services started moving away from ISP to fully fledged online social network companies like Facebook, Flicker, Napster, LinkedIn, Twitter, and others.

Gaining Knowledge from Social Networks

When more and more people are making their opinions available in social networks, it is possible to find out about the opinions and experiences of those in the vast pool of people that are neither our personal acquaintances nor well-known professional critics. Figuring out "What other people think" has always been an important piece of information for most of us during the decision-making process. Organizations are attempting to extract insights from opinions of their consumers for revenue increase and competitiveness improvement. The Twitter as an example of a social network consists of 40 million Twitter users, including billions of tweets, more than 1 billion relationships between users, and millions of posts, hashtags, URLs, and emoticons. Through analyzing and exploiting the Twitter data, it is possible to formulate and answer a variety of interesting problems/questions, such as the trending topics, brands, and pop culture, to assess the sentiment or popularity around any area of interest, followers count, tweet counts by catalog, and more. For instance, the problems or questions related to Twitter may be “What’s the twitter traffic distribution by hours, days, weeks, months, and years?”, “Sort all the URLs twitted in descend order”, “What background color Twitter users like most?”, “Who is the person who twitted the most in the three year period?”, “Who is the Twitter user who has the most followers by month and year?”, “Which geographic location has the most Twitter users?”, and so forth.

Ethical and Privacy Issues in Online Social Networks

Privacy is a human value consisting of a set of rights including solitude, the right to be alone without disturbances; anonymity, the right to have no public personal identity; intimacy, the right not to be monitored; and reserve, the right to control one’s personal information, including the dissemination methods of that information. As humans, we assign a lot of value to these four rights. In fact, these rights are part of our moral and ethical systems. With the advent of the Internet, privacy has gained even more value as information has gained value. The value of privacy comes from its guardianship of the individual’s personal identity and autonomy.

Autonomy is important because humans need to feel that they are in control of their destiny. The less personal information people have about an individual, the more autonomous that individual can be, especially in decision-making. However, other people will challenge one’s autonomy depending on the quantity, quality, and value of information they have about that individual. People usually tend to establish relationships and associations with individuals and groups that will respect their personal autonomy, especially in decision-making.

As information becomes more imperative and precious, it becomes more important for individuals to guard their personal identity. Personal identity is a valuable source of information. Unfortunately, with rapid advances in technology, especially computer and telecommunication technologies, it has become increasingly difficult to protect personal identity.

Privacy Issues in OSNs

Privacy can be violated, anywhere including online social network communities, through intrusion, misuse of information, interception of information, and information matching (Web Surpasses One Billion Documents 2000). In online communities, intrusion, as an invasion of privacy, is a wrongful entry, a seizing, or acquiring of information or data belonging to other members of the online social network community. Misuse of information is all too easy. While online, we inevitably give off our information to whoever asks for it in order to get services. There is nothing wrong with collecting personal information when it is authorized and is going to be used for a legitimate reason. Routinely information collected from online community members, however, is not always used as intended. It is quite often used for unauthorized purposes, hence an invasion of privacy. As commercial activities increase online, there is likely to be stiff competition for personal information collected online for commercial purposes. Companies offering services on the Internet may seek new customers by either legally buying customer information or illegally obtaining it through eavesdropping, intrusion, and surveillance. To counter this, companies running these online communities must find ways to enhance the security of personal data online.

As the number and membership in online social networks skyrocketed, the issues of privacy and security of users while online and the security of users’ data while off-line have taken center stage. The problems of online social networking have been exhibited by the already high and still growing numbers especially of young people who pay little to no attention to privacy issues for themselves or others. Every passing day, there is news about and growing concerns over breaches in privacy caused by social networking services. Many users are now worried that their personal data is being misused by the online service providers.

As the growth in online social networks continues unabated, the coming in the mix of the smart mobile devices is making the already existing problems more complex. These new devices are increasing the number of accesses to OSNs and increasing the complexity of the privacy issues, including (Wresch 1996):

The presence of a user. Unlike in the most traditional OSNs where users were not automatically made aware of the presence of their friends, most mobile OSNs (mOSNs) now allow users to indicate their presence via a "check-in" mechanism, where a user establishes their location at a particular time. According to Wresch (1996), the indication of presence allows their friends to expect quick response, and this may lead to meeting new people who are members of the same mOSN. Although the feature of automatic locate by oneself is becoming popular, it allows leakage of personal private information along two tracks: the personal information that may be sent and the destination to which it could be sent.

Location-based tracking system (LTS) that are part of our mobile devices. This is a feature that is widespread in the mobile environment. However, users may not be aware that their location can be made known to friends, and friends of friends who are currently online on this mOSN, their friends in other mOSNs, and others who may lead to leakage of personal information to third parties.

Interaction potential between mOSNs and traditional OSNs. According to Wresch (1996), such connections are useful to users who, while interacting with a mOSN, can expect some of their actions to show up on traditional OSNs and be visible to their friends there. However, a lot of their personal information can leak to unintended users of both the traditional OSNs and the mOSNs.

In addition to almost free access to a turn of personal data on OSNs, there is also a growing threat to personal data ownership. For example, who owns the data that was altered or removed by the user which may in fact be retained and/or passed to third parties? Fortunately users are beginning to fight for their privacy to prevent their personal details from being circulated further than they intended it to be. For example, Facebook's 2006 News Feed and Mini Feed features are designed to change what Founder and CEO Mark Zuckerberg called Facebook's old "encyclopedic interface," where pages mostly just list off information about people, to the current stream of fresh news and attention content about not only the user but also the user's friends and their activities (Walsh 2013). The first, News Feed, brought to the user's home page all new activities on all friends and associate links including new photos posted by friends, relationship status changes, people joining groups, and many others, thus enabling the user to get an abundance of information from every friend's site every day. Although these features adhered to Facebook's privacy settings, meaning that only people a user allowed to view the data were able to see it, it still generated a firestone from users across the world. Over 700,000 users signed an online petition demanding the company to discontinue the feature, stating that this compromised their privacy (Walsh 2013). Much of The criticism of The News Feed was that it gave out too much individual information.

Since online social networks are bringing people together with no physical presence to engage in all human acts that traditionally have taken place in a physical environment. As these cyber communities are brought and bound together by a sense of belonging, worthiness, and the feeling that they are valued by members of the network, they create a mental family based on trust, the kind of trust you would find in a loving family. However, because these networks are borderless, international in nature, they are forming not along well-known and traditional identifiers such as nationalities, beliefs, authority, and the like but by common purpose and need with no legal jurisdiction and no central power to enforce community standards and norms.

Strengthening Privacy in OSNs

As more and more people join OSNs and now the rapidly growing mOSNs, there is a growing need for more protection to users. Chew et al. suggest the following steps needed to be taken (Chew et al. 2013):

Both OSN and mOSN applications should be explicit about which user activities automatically generate events for their activity streams.

Users should have control over which events make it into their activity streams and be able to remove events from the streams after they have been added by an application.

Users should know who the audience of their activity streams is and should also have control over selecting the audience of their activity streams.

Both OSN and mOSN application should create activity stream events which are in sync with user expectation.

Other suggestions that may help in this effort are:

Use secure passwords.

User awareness of the privacy policies and terms of use for their OSNs and mOSNs.

Both OSNs and mOSNs providers should devise policies and enforce existing laws to allow some privacy protection for users while on their networks.

Ethical Issues in Online Social Communities

Online social communities including online social network are far from the traditional physical social communities with an epicenter of authority with every member paying allegiance to the center with a shared sense of responsibility. This type of community governance with no central command, but an equally shared authority and responsibility, is new, and a mechanism needs to be in place and must be followed to safeguard every member of the community. But these mechanisms are not yet defined, and where they are being defined, it is still too early to say whether they are effective. The complexity, unpredictability, and lack of central authority are further enhanced by Kizza (2013):

Virtual personality: You know their names, their likes, and dislikes. You know them so well that you can even bet on what they are thinking, yet you do not know them at all. You cannot meet them and recognize them in a crowd.

Anonymity: You work with them almost every day. They are even your friends; you are on a first-name basis, yet you will never know them. They will forever remain anonymous to you and you to them.

Multiple personality: You think you know them, but you do not because they are capable of changing and mutating into other personalities. They can change into as many personalities as there are issues being discussed. You will never know which personality you are going to deal with next.

These three characteristics are at the core of the social and ethical problems in online social networks in particular and cyberspace in general; the larger and more numerous these communities become, the more urgent the ethical concerns become. With all these happening in online social networks, the crucial utilitarian question to ask is what is best way and how we can balance the potential harms and benefits that can befall members of these online social networks and how if possible to balance these possibilities. Of late, the news media has been awash with many of these online ills and abuses, and the list is growing including potential for misuse, cyberbullying, cyberstalking, and cyber harassment, risk for child safety, psychological effects of online social networking, and free speech.

Security and Crimes in Online Social Communities

Online crimes, in tandem with the growth of computing and telecommunication technologies, are one of the fastest-growing types of crimes, and they pose the greatest danger to online communities, e-commerce, and the general public in general. An *online crime* is a crime like any other crime, except that in this case, the illegal act must involve either an Internet-enabled electronic device or computing system either as an object of a crime, an instrument used to commit a crime, or a repository of evidence related to a crime. Also online crimes are acts of unauthorized intervention into the working of the telecommunication networks and/or the sanctioning of authorized access to the resources of the computing elements in a network that lead to a threat to the system's infrastructure or cause a significant property loss. The International Convention of Cyber Crimes and the European Convention on Cyber Crimes both list the following crimes as online crime (Kizza 2005):

Unlawful access to information
Illegal interception of information
Unlawful use of telecommunication equipment
Forgery with use of computer measures
Intrusions of the public switched and packet network
Network integrity violations
Privacy violations
Industrial espionage
Pirated computer software
Fraud using a computing system
Internet/e-mail abuse
Using computers or computer technology to commit murder, terrorism, pornography, and hacking

As we discussed before, the online contents are accessible from different locations without noticeable delay. Because of the decentralized architecture of the Internet, personal publication through the web becomes more feasible and affordable, while still maintaining a high exposure to the target audience. At the same time, the lack of regulations makes the online social community a pretty free realm where the geographical border dims in the online communities. Information can be spread anonymously with little interference from governments via the online community. Costs of the community are relatively low compared with other media. Various communities benefit from the online features of the community. We will analyze a dark web as a case study here to illustrate how terrorist/extremist organizations and their sympathizers exchange ideology, spread propaganda, recruit members, and plan attacks. The terrorists, extremists, and their sympathizers can benefit from web techniques and online communities. They exchange ideology, spread propaganda, recruit members, and even plan attacks through the online community. Especially, because of the ubiquity of the online community, the previously isolated terrorists/extremist cells are able to collaborate more efficient than any time before and to form a more compact community virtually. Dark webs contain rich information about the dark groups, such as ideologies, recent topics, and news.

Several research works have been conducted to analyze web of terrorist cells or criminal activities. M. Sparrow (1991) applied social network analysis to criminal activities and observed three problems associated with criminal network analysis. They are incompleteness of analyzing data as a result of missing nodes and links that the investigators will not uncover, fuzzy boundaries resulting from the difficulty in deciding who to include and who not to include, and the dynamic property of analyzed networks. V. E. Krebs (2001) uses public information reported in major newspapers such as the *New York Times* and the *Wall Street Journal* to map networks of terrorist cells. Their research unrevealed a picture of a covert network after the tragic events of September 11, 2001. P. Klerks (2001) describes the development of criminal network analysis. The approaches start from manual analysis. An analyst constructs an association matrix by identifying criminal associations from raw data. Then a graphic-based approach is proposed to automatically generate graphical representation of criminal networks. Recently social network analysis has been used to provide more advanced analytical functionality to assist crime investigation. J. Xu and H. Chen (2005) and Koch et al. (2007) use data mining techniques to reveal various structures and interactions within a network.

Discovering topics from dark websites helps in developing effective combating strategies against terrorism or extremists. The latent or topics are buried in large-scale web pages and hosted by dark websites. This work employs information retrieval (IR) techniques to discover hidden topics in a known dark web, such that the discovered latent topics can provide insights into social communities.

Modeling text corpora extracted from websites help find short description of a topic such that essential statistical relationships are preserved from for the basic tasks such as classification, summarization, and similarity judgment (Sparrow 1991). In the field of information retrieval, a basic vocabulary of words or terms is chosen, and each document in the corpus is reduced to a vector of real numbers, each entry representing ratios of word counts. In the popular *tf-idf* scheme (Krebs 2001), term frequency (tf) count is compared to an inverse document frequency (idf) count, which measures the number of occurrences of a word in the entire dataset. The tf-idf scheme generates a term-by-document matrix X whose columns contain the tf-idf values for each of the documents in the corpus. Latent semantic indexing (LSI) (Deerwester et al. 1990) is proposed to further reduce description length and reveal more inter- or intra-document statistical structure. LSI uses a singular value decomposition of the X matrix to identify a linear subspace in the space of tf-idf features that capture most of the variance in the collection.

Hofmann (1999) presented probabilistic LSI (pLSI) model to model each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of topics. Thus, each word is generated from a single topic, and different words in a document may be generated from different topics. While Hofmann's word is a useful step toward probabilistic modeling of text, it provides no probabilistic model at the level of documents. Yang et al. (2009) discovered latent topics from the dark web by latent Dirichlet allocation (LDA) (Blei and Jordan 2003) which improves upon pLSI by placing a Dirichlet prior on topic distribution to reduce overfitting and bias the topic weights from each document toward skewed distributions with few dominant topics.

Conclusion

The growth of online social communities, emanating from the old social gatherings of days before computing, has given us all a bonanza to and means to access information in amazing ways. Online communities have created opportunities for us unprecedented in the history of human where one individual can reach millions of others anywhere on the globe in seconds. The history and development of computing has made all this possible. However, with the easiness and abundance of resources at our disposal availed to us by online communities, there has also been evils that have been enabled by these large ecosystems. To be able to safeguard personal privacy, security, and dignity, we must pay special attention and develop protocols and best practices that must make everyone in these communities safely enjoy the experiences presented in these ecosystems. The battle is not yet worn, and the way forward is not clear yet just because the next move in new technologies is not predictable.

Cross-References

- ▶ Consequences of Publishing Real Personal Information in Online Social Networks
- ▶ Ethical Issues Surrounding Data Collection in Online Social Networks
- ▶ Online Social Network Privacy Management
- ▶ Privacy Preservation and Location-based Online Social Networking
- ▶ Topology of Online Social Networks
- ▶ User Behavior in Online Social Networks: Influencing Factors

References

Blei N, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022 MATH

Bylaws for Internet Corporation for Assigned Names and Numbers (ICANN) (2005). http://www.icann.org/general/bylaws.htm. Retrieved Apr 2013

Chew M, Balfanz D, Laurie B (2013) (Under)mining privacy in social networks, Google Inc. http://w2spconf.com/2008/papers/s3p2.pdf. Retrieved Apr 2013

Deerwester S, Dumais S, Landauer T, Furnas G, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):39–407

Evolving the High Performance Computing and Communications Initiative to Support the Nation's Information Infrastructure—Executive Summary (2013). http://www.nap.edu/openbook.php?record_id=4948. Retrieved Apr 2013

Fox R (2000) News track: age and sex. Commun ACM 43(9):9

Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the twenty-second annual international SIGIR conference, Berkeley

Kizza JM (1999) Ethical and social issues in the information age. Springer, New York

Kizza JM (2005) Computer network security. Springer, New York MATH

Kizza JM (2013) Ethical and social issues in the information age, 5th edn. Springer, London

Klerks P (2001) The network paradigm applied to criminal organizations: theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands. Connections 24(3): 53–65

Koch M, Richter A, Schlosser A (2007) Services and applications for IT-supported social networking in companies. Wirtschaftsinformatik 6(49): 448–455

Kramer RM (1999) Trust and distrust in organizations: emerging perspectives, enduring questions. Ann Rev Psychol 50:569–598

Krebs VE (2001) Mapping networks of terrorist of cells. Connections 24:43–52

Sparrow MK (1991) The application of network analysis to criminal intelligence — an assessment of the prospects. Soc Netw 13:251–274

Walsh M (2013) Court backs student on Facebook page criticizing teacher, NewsWeek. http://blogs.edweek.org/edweek/school_law/2010/02/court_backs,student_on_faceboo.html. Retrieved Apr 2013

Web Surpasses One Billion Documents: Inktomi and NEC Research Institute Complete First Web Study (2000) Inktomi News & Events, Jan 2000

Wresch W (1996) Disconnected: haves and have-nots in the information age, "Information Age Haves and Have-Nots." Rutgers University Press. ISBN-10:0813523702

Xu J, Chen H (2005) Analyzing and visualization. Commun ACM 48:100–107

Yang L, Liu F, Kizza JM, Ege RK (2009) Discovering latent topics from dark websites. In: IEEE symposium on computational intelligence in cyber security, IEEE Xplore, Nashville

Social Indexing

▶ Folksonomies

Social Influence

▶ Web Communities Versus Physical Communities

Social Influence Analysis

Tiziana Guzzo, Fernando Ferri and Patrizia Grifoni
Institute of Research on Population and Social Policies - IRPPS, National Research Council -CNR, Rome, Italy

Synonyms

Social networks members; Social networks users

Glossary

Confounding variables Unknown variables exist (e.g., common location, gender, school, and several other external factors), which may cause friends to behave similarly with one another

Correlation factor Correlation between variables is a measure of how well the variables are related. The most common measure of correlation in statistics is the Pearson correlation

Edge-reversal test Reserves the direction of all edges. Social influence spreads in the direction specified by the edges of the graph, and hence reversing the edges should intuitively change the estimate of the correlation

Homophily	A user in the social network tends to be similar to his/her connected neighbors
Induction	An action of a user is triggered by an action of another user
Selection	People tend to create relationships with other people who are already similar to them
Shuffle test	Shuffles the activation time of users. It is based on the idea that influence does not play a role, and then the timing of activation should be independent of the timing of activation of others
The influence maximization problem	Aims to identify an initial set of users in a social network that could maximize the spread of influence such that other users will adopt the new product in the shortest time

Definition

Social influence refers to change of a person's behavior after an interaction with other people, organizations, and in general society. It consists of the process by which the individual opinions can be changed by the influence of other individual(s) (Friedkin 1998). It is characterized by three main features:

- Conformity, that occurs when an individual expresses a particular opinion in order to meet the expectations of a given other, though he/she does not necessarily hold that belief that the opinion is appropriate
- Power, that is the ability to force someone to behave in a particular way by controlling his/her outcomes
- Authority, that is the power that is believed to be legitimated by those who are subjected to it

Webster's dictionary defines influence as "the power or capacity of a person or things in causing an effect in indirect or intangible ways." It could be defined as the combination of all things that may change or have some effects on a person's behavior, thoughts, actions, or feelings. It can be represented by peer pressure, persuasion, marketing, sales, and conformity.

This phenomenon in social networks refers to the behavioral change of individuals affected by others in a network. Social influence analysis in online social networks, studies people's influence by analyzing the social interactions between its members.

Introduction

Three broad categories of social influence were identified by Kelman (1958): (i) compliance, when people appear to agree with others while keeping their dissenting opinions private; (ii) identification, when people are influenced by someone who is liked and respected, such as a famous celebrity; and (iii) internalization, when people accept a belief or behavior and agree both publicly and privately.

The social environment and personal interactions have powerful effects on human behavior that in fact is always influenced by each other.

In literature three types of reference group influences are identified: informational influence, utilitarian influence, and value-expressive influence (Park and Lessig 1977; Bearden and Etzel 1982):

- The informational influence acts when individual would like to improve its knowledge and have best and useful information in order to optimize its choices (Kelman 1961).
- The utilitarian influence is based on the compliance process and acts when individual would like to satisfy a group's expectation in order to achieve a favorable reaction from it (Kelman 1961).
- The value-expressive influence is based on the identification process and acts when individual would like to be similar to the group in order to belong to it (Kelman 1961).

The exponential growth of online social networks such as Facebook, Twitter, MySpace, Flickr, and Pinterest, Instagram is playing an important role in shaping the users' behavior on the Web. Fowler and Christakis (2008) introduced the theory of three degrees of influence to explain the great influence that social networks have on people' behavior. According to them, people have an influence on friends which in their turn influence their friends, meaning that actions can influence people they have never met. They claim that "everything we do or say tends to ripple through our network, having an impact on our friends (one degree), our friends' friends (two degrees), and even our friends' friends' friends (three degrees). Our influence gradually dissipates and ceases to have a noticeable effect on people beyond the social frontier that lies at three degrees of separation."

The probability to be influenced by an influencer depends on four factors:

- Relevance (the right information): the user's information needs have to coincide with the influencer's expertise.
- Timing (the right time): information has to be delivered when the user needed it.
- Alignment (the right place): few channel of overlap between the user and the influencer there must be.
- Confidence (the right person): users have to trust the influencer with respect to his/her information needs.

Historical Background

Social relationships are key components of human life, and they have been historically connected to time and space limitations; these restrictions have been partially removed with the Internet diffusion. In particular, the emergence of social networks has created a new social dimension where individuals can increase their social awareness interacting with old and new friends; share information about data, products, and services; and be more informed about different aspects of everyday lives anywhere and anytime. The interest in social network studies has been growing massively in recent years. Psychologists, anthropologists, sociologists, economists, and statisticians have given important contributions, making it actually an interdisciplinary research area. In the last years several methods to collect and visualize network data have been developed in order to analyze relationships between people, groups, and organizations.

In a social network, members (nodes associated with others nodes) are influenced by others for various reasons. Social influence is a directional effect from node A to node B. Some nodes can have intrinsically higher influence than others due to network structure. Social Network Analysis is the study of social relations among a set of actors (nodes). The nodes in the network are the people and groups, while the links show relationships or flows between the nodes. The analysis allows "to determine if a Virtual Social Network is tightly bounded diversified or constricted, to find its density and clustering, and to study how the behaviour of network members is affected by their positions and connections" (Scott 2000). The importance of a node in the network is measured by its centrality. The three most important individual centrality measures are (http://www.orgnet.com/sna.html):

- The degree centrality refers to the number of direct connections a node has.
- The betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes.
- The closeness centrality that is the inverse of farness, which in turn is the sum of distances to all other nodes.

A node with high centrality is usually more highly influential than other nodes. According Katona et al. (2011), demographics and user's position can predict their influential power on their neighbors. Social Network Analysis analyzes which members are individuals or peripheral in a network; it identifies bonding and bridging and who has influence in the network. Many mathematical techniques are available to measure networks (Wasserman and Faust 1994). Hoppe and Reinelt (2010) demonstrate how to use these metrics to understand and evaluate specific leadership networks.

Farrow and Yuan (2011) explored the strength of network ties to show how Facebook influences the attitudes of the alumni to volunteer for and make charitable gifts to their alma mater fortifying consistency between attitude and behavior.

Social influence analysis aims at qualitatively and quantitatively measuring the influence of one person on others. There are different methods and algorithms for measuring social influence, and they will be analyzed in the following sections.

Qualitative Measures for Analyzing Social Influence

According to Anagnostopoulos et al. (2008), influence of a person on another can act for three reasons: (i) induction, (ii) homophily, and (iii) confounding variables (factors). They applied statistical analysis on data from a large social system in order to identify social influence as a source of correlation between the actions of individuals with social ties. They proposed two tests, the Shuffle test and the Edge-Reversal test, to identify induction as cause of social correlation when the time series of user actions is available. This approach is based on the assumption that timing of actions should matter if induction is a likely cause of correlation.

Goyal et al. (2010) followed a similar approach, proposing to establish relationship between users by scanning log of user action. According to them the influence probability between two users is determined by common actions and time issues. The approach based on homophily was followed by Crandall et al. (2008) that used cosine similarity to compute the similarity between two people. They proposed a probabilistic model which samples activities of people based on their history and those of their neighbors and a background distribution. This concept was stressed also by Matsuo and Yamamoto (2009). They studied user's behavior on an e-commerce site and found that users generally trust other users who have similar behavior with them. Other studies analyzed the correlation between social similarity and influence. Singla and Richardson (2008) studied the probability of relationships between two users by measuring their similarity. According to them users with common features (age, gender, zip code, word, and queries issued) chat more likely to each other; then influence probabilities could be estimated by user's similarity.

These studies used different approaches to analyze influence probabilities, but they did not address the issue of identifying influential users of the network. This issue will be analyzed in the next section introducing studies that used quantitative measures.

Quantitative Measures for Analyzing Social Influence

The problem to quantify the strength of social influences and differentiate social influences from different angles (topics) was addressed by Tang (2009). They studied the topic-based social influence analysis on large networks. The goal was to simultaneously analyze nodes' topic distributions (or user interests), similarity between nodes (users), and network structure. They proposed a Topical Factor Graph (TFG) model to incorporate all information into a unified probabilistic model and present Topical Affinity Propagation (TAP) for model learning.

Most studies about social influence analysis considered positive interactions (agreement, trust) between individuals; Li et al. (2011) also considered negative relationships (distrust, disagreement) between individuals and conformity of people (the inclination of a person to be influenced). They proposed an algorithm called CASINO (Conformity-Aware Social INfluence cOmputation) which quantifies the influence and conformity of each individual in a network by utilizing the positive or negative relationships between individuals. This algorithm consists of three phases. In the first phase, a set of topic-based subgraphs that represent the social interactions associated with a specific topic are extracted from a social network. In the second phase, the edges (relationships) between individuals are labeled with positive or negative signs. Finally, in the third phase, the influence and conformity indices of each individual in each signed topic-based subgraph are computed.

The problem of dynamic social influence analysis was addressed by Wang et al. (2011). They proposed a pairwise factor graph (PFG) model to quantify the influence between two users in a large social network. Different types of factor functions capture information such as users' attribute information, social similarities/weights, and network structures, which form the basic components of the factor graph model. An algorithm was designed to learn the model and make inference to obtain all the marginal probabilities. They further proposed a dynamic factor graph (DFG) model to incorporate the time information.

Domingos and Richardson (2001) first studied the problem of which individuals is necessary to target to have a large cascade of further adoptions. The problem was considered in a probabilistic model of interaction; heuristics were given for choosing customers with a large overall effect on the network. Kempe et al. (2003) faced the same issue of choosing influential sets of individuals by formulating it as a discrete optimization problem and proposing an approximation algorithm that was applicable to general cases. This optimization problem has a complexity NP and "the greedy algorithm can guarantee the influence spread within (1-1/e) of the optimal influence spread."

Kimura and Saito (2006) propose shortest-path-based influence cascade models and provide efficient algorithms to compute influence spread. However, these algorithms are not scalable for large graphs; to solve the problem Chen et al. (2009, 2010) designed a new heuristic algorithm. This algorithm, scalable to millions of nodes and edges, allows controlling the balance between the running time and the influence spread of the algorithm. With respect to the work of Kimura that used simple shortest paths on the graph, which are not related to propagation probabilities, Chen used maximum influence paths and local structures such as arborescences.

Key Applications

In social networks very important is the effect of "word of mouth," since idea, opinions, and recommendations propagate very quickly and with an exponential grow. This concept is very frequently applied in different fields like marketing, recommendations, healthcare, and politics.

Many companies have recently started to capture data on the social interaction between consumers in social networks, with the objective of understanding and leveraging how this interaction can generate social influence. Consumers can really modify their opinions about products and/or services according to the social influence process; this process also impacts on knowledge diffusion about products and services. Social network emerges as one of the most authoritative and influential sources of knowledge about products and services related to the area of interest of a community. They have the aptitude to generate knowledge sharing among consumers and facilitate the collaboration and exchange of ideas among consumers. In this context, viral marketing involves customers in commercial strategies for recommending commercial products to their friends through the customer social networks. According to De Bruyn and Gary (2004), viral marketing is a "consumer-to-consumer (or peer-to-peer) communication, as opposed to company-to-consumer communications, to disseminate information about a product or a service, hence leading to its rapid and cost-effective market adoption." In this context the problem of the influence maximization that aims to identify individuals to target to have a large cascade of further adoptions assumes a great relevance. Several studies introduced in the previous section (such as Domingos and Richardson 2001; Kempe et al. 2003; Kimura and Saito 2006; Chen et al. 2009, 2010) addressed this issue.

The emergence of e-commerce has led to the development of recommender system, a personalized information filtering technology used to identify a set of items that will be of interest to a certain user. Mao et al. (2012) explored social influence for item recommendation. Previous approaches mostly incorporated social friendship into recommender systems by heuristics. They captured quantitatively social influence and proposed a probabilistic generative model, called social influenced selection (SIS), extracting social influence and preferences through statistical inference.

Moreover, they developed a new parameter learning algorithm based on expectation maximization (EM) to face the problem of multiple layers of hidden factors in SIS.

Social networks are rapidly transforming also the healthcare field. People are more and more connected to the Web in order to search, share, and exchange information and find support from other people. According to a 2008 survey carried out by Icrossing, the Internet has been the most used source to find information about health and wellness in the previous 12 months. Patients, thanks to the Web, can share the same illness with people all over the world and feel themselves less alone. In addition according to Edelman's study (2008), people have more trust in a person with which they can identify themselves than business, government, and media subjects. Which is the impact of these activities on heath conditions? A study of Christakis and Fowler found that health status can be influenced by the health status of the neighbors. How do you manage the inaccurate information disseminating on health social networks? Some studies were carried out to identify influential users in order to optimize the spread of health information (Krulwich and Burkey 1995; Zhang et al. 2007).

Another emerging key application of social influence on social networks is the political field.

A strength of the first Obama election campaign was his strategic use of social media. Analysts are now studying the impact of tools such as Facebook, Twitter, and YouTube had on election results. A recent Pew Research study (Rainie and Smith 2012) analyzed politics on social networks and found that users after discussing a political issue or reading posts about it on these sites change their points of view and political involvement. Bond et al. (2012) hypothesized that voting behavior is significantly influenced by messages on Facebook. They found that political self-expression, information seeking, and real-world voting behavior of millions of users were directly influenced by messages. This had an indirect effect through social contagion also in the users' friends and friends of friends. Close friends had four times more influence than the message itself. Furthermore, they stated that "online mobilization works because it primarily spreads through strong-tie networks that probably exist offline but have an online representation."

Fowler (2005) based on observational data found that behavior of each act of voting spreads through the network generating on average an additional three votes.

Future Directions

Social influence analysis studies are in its beginnings, and so in the future more methods and techniques will be developed. A challenge for future works will be to develop efficient, effective, and quantifiable methods for analyzing the persuasion and influence phenomenon within social networks. Until now, studies have mainly focused on conceptual models and small-scale simulations. In the future as online social networks enable for the first time to measure social influence over a large population, they should include more large-scale data mining algorithms to analyze social network data. It will allow having more realistic results for large-scale applications in different fields and in different social and informational settings.

Cross-References

- ▶ Centrality Measures
- ▶ Friends Recommendations in Dynamic Social Networks
- ▶ Mobile Communication Networks
- ▶ Modeling Social Preferences Based on Social Interactions
- ▶ Origins of Social Network Analysis
- ▶ Political Networks
- ▶ Questionnaires for Measuring Social Network Contacts
- ▶ Recommender Systems: Models and Techniques
- ▶ Recommender Systems Using Social Network Analysis: Challenges and Future Trends
- ▶ Social Recommendation in Dynamic Networks
- ▶ Social Recommender System
- ▶ Trust in Social Networks
- ▶ User Behavior in Online Social Networks: Influencing Factors

References

Anagnostopoulos A, Kumar R, Mahdian M (2008) Influence and correlation in social networks. In: KDD'08, Las Vegas, pp 7–15

Bearden WO, Etzel MJ (1982) Reference group influence on product and brand purchase decisions. J Consum Res 9:183–194

Bond RM, Fariss CJ, Jones JJ, Kramer ADI, Marlow C, Settle JE, Fowler JH (2012) A 61-million-person experiment in social influence and political mobilization. Nature 489:295–298

Chen W, Wang Y, Yang S, (2009) Efficient influence maximization in social networks. In: KDD'09, Paris

Chen W, Wang C, Wang Y (2010) Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: KDD'10, Washington, DC

Crandall D, Cosley D, Huttenlocher D, Kleinberg J, Suri S (2008) Feedback effects between similarity and social influence in online communities. In: KDD'08, Las Vegas, pp 160–168

De Bruyn A, Gary LL (2004) A Multi-Stage model of word of mouth through electronic referrals. eBusiness Research Center Working Paper

Domingos P, Richardson M (2001) Mining the network value of customers. KDD'01, San Francisco, pp 57–66

Edelman (2008) Edelman trust barometer. http://www.edelman.com/trust/2008. Accessed 18 Sept 2012

Farrow H, Yuan YC (2011) Building stronger ties with alumni through Facebook to increase volunteerism and charitable giving. J Comput Mediat Commun 16(3):445–464

Fowler JH (2005) Trunout in a small world. In: Zuckerman AS (ed) The Social logic of politics: personal networks as contexts for political behavior. Temple University Press, Philadelphia, pp 269–287

Fowler JH, Christakis NA (2008) The dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham heart study. Br Med J 337:a2338

Friedkin N (1998) A structural theory of social influence. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511527524

Goyal A, Bonchi F, Lakshmanan LVS (2010) Learning influence probabilities in social networks. In: Proceedings of the proceedings of the third ACM international conference on Web search and data mining, New York. ACM, New York, pp 241–250

Hoppe B, Reinelt C (2010) Leadersh Q 21:600–619

Icrossing (2008) iCrossing's how America searches: health and wellness. Retrieved 10 May 2009

Katona Z, Zubcsek PP, Sarvary M (2011) Network effects and personal influences: the diffusion of an online social network. J Mark Res 48(3):425–443

Kelman H (1958) Compliance, identification, and internalization: three processes of attitude change. J Confl Resolut 2(1):51–60

Kelman HC (1961) Processes of opinion change. Public Opin Q 25:57–78

Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence through a social network. In: KDD'03. Washington, DC, pp 137–146

Kimura M, Saito K (2006) Tractable models for information diffusion in social networks. In: Proceedings of the 10th European conference on principles and practice of knowledge discovery in databases, Berlin, pp 259–271

Krulwich B, Burkey C (1995) Contact finder: extracting indications of expertise and answering questions with referrals. In: Symposium on intelligent knowledge navigation and retrieval, Cambridge, pp 85–91

Li H, Sourav SB, Sun A (2011) CASINO: towards conformity-aware social influence analysis in online social networks. In: Proceedings of the 20th ACM conference on information and knowledge management, CIKM 2011, Glasgow, October 24–28

Mao Y, Xingjie L, Wang-Chien Lee (2012) Exploring social influence for recommendation – a generative model approach. In: Proceeding SIGIR'12 proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval, Portland, pp 671–680

Matsuo Y, Yamamoto H (2009) Community gravity: measuring bidirectional effects by trust and rating on online social networks. In: Proceedings of the 18th international conference on world wide web. Madrid. ACM, New York, pp 751–760

Park W, Lessig VP (1977) Students and housewives: differences in susceptibility to reference group influence. J Consum Res 4:102–110

Rainie L, Smith A (2012) Politics on social networking sites. Pew Research Center's internet & American Life Project. http://pewinternet.org/~/media/Files/Reports/2012/PIP_PoliticalLifeonSocialNetworkingSites.pdf. Accessed 4 Sept 2012

Scott J (2000) Social network analysis: a handbook. Sage, London

Singla P, Richardson M (2008) Yes, there is a correlation: from social networks to personal behavior on the web. In: WWW'08, Beijing, pp 655–664

Tang J (2009) Social influence analysis in large-scale networks. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. Paris. ACM, New York, pp 807–816

Wang C, Tang J, Sun J, Han J (2011) Dynamic social influence analysis through time-dependent factor graphs. In: International conference on advances in social networks analysis and mining Kaohsiung, Taiwan, 25–27 July 2011, IEEE Computer Society, Los Alamitos

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge

Zhang J, Ackerman M, Adamic L (2007) Expertise networks in online communities: structure and algorithms. In: Proceedings of the 16th international conference on World Wide Web, Banff. ACM, New York, pp 221–230

Social Influence Propagation

▶ Influence Maximization Model

Social Influence Propagation Model

▶ Mathematical Model for Propagation of Influence in a Social Network

Social Interaction

▶ Incentives in Collaborative Applications
▶ Networks at Harvard University Sociology
▶ Origins of Social Network Analysis
▶ Privacy and Disclosure in a Social Networking Community
▶ Privacy Issues for SNS and Mobile SNS
▶ Social Communication Network: Case Study

Social Interaction Analysis for Team Collaboration

Ognjen Scekic, Mirela Riveni, Hong-Linh Truong and Schahram Dustdar
Distributed Systems Group, Institute for Information Systems, TU Wien, Vienna, Austria

Synonyms

Collaboration analysis; Collaboration metrics; Collaboration platforms; Collective adaptive systems; Crowdsourcing; Human-based services (HBS); Interaction patterns; Process mining; Rewarding; Social computing; Social trust; Socio-technical systems; Task assignment; Team collaboration; Team formation

Glossary

Actor	Entity (human or computer) possessing a capability to act intelligently and process specific assignments (activities/ tasks)
Atomic task	Task that can be handled by an individual actor
CAS	Collective adaptive system
Collaboration system (platform)	Information system supporting execution of collaborative processes
Collaborative process (collaboration)	Joint effort of a (limited) number of actors with the goal of performing a task. A collaborative process has a limited duration and requires coordination among actors (due to task dependencies)
Composite task	Task that must be handled by multiple actors due to size or complexity. A composite task can be broken down into atomic tasks
CSCW	Computer-supported cooperative work
HBS	Human-based service
Metric	Precisely defined, context-specific measure of some properties
QoD	Quality of data
QoS	Quality of service
SOA	Service-oriented architecture
SOC	Service-oriented computing
Task assignment	The art to divide a (composite) task into (sub)tasks and assign them to appropriate actors
Task	Piece of work to be solved, typically complex enough to require knowledge or processing power of a large number of individual actors
Team formation	Process consisting of identifying appropriate actors for performing all atomic tasks and establishing of internal coordination and functioning rules in the team

Team	Set of actors taking part in a collaborative process. Team lifetime is considered equal to the lifetime of the collaborative process
WS	Web service

Definition

In this article, we look into different types of collaboration systems. We describe team structures and discuss different forms of collaborations they support. In particular, we focus on interaction processes that are supported by the system and discuss different metrics used to describe and analyze such systems. We discuss three main team collaboration types: static, ad hoc, and open collaboration. We then focus on interaction analysis and discuss appropriate interaction metrics.

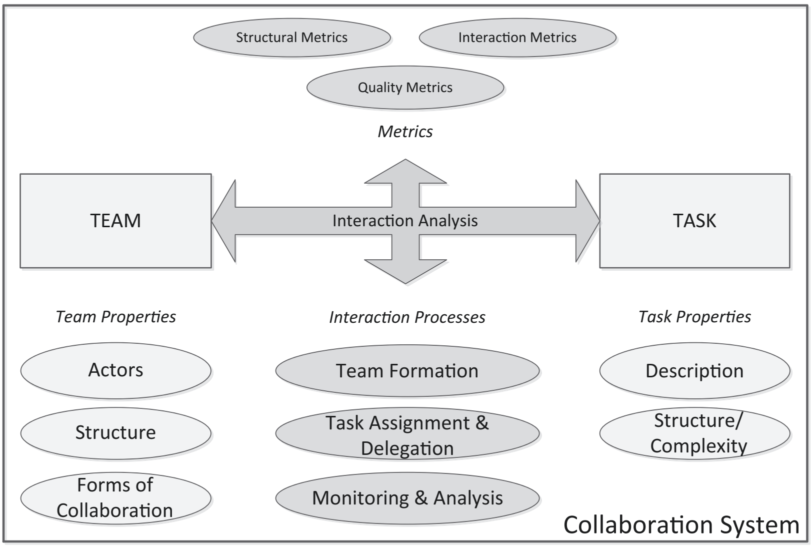
Introduction

With the advent of Web 2.0 and social networks, millions of users around the world were given the opportunity to collaborate, share ideas, and coordinate their efforts easier than ever before. These developments lead to an increased interest to exploit these opportunities, both in the research community and in the industry. Such collaborative efforts are supported by different types of *collaboration systems,* providing automated or semiautomated actor management (e.g., modeling, reputation, and rewarding), task management (e.g., modeling, creation, division, scheduling, aggregation, and monitoring), and process execution environment (including actor communication and coordination).

Collaboration systems enable different collaboration types. Depending on the type of system and type of problem to be solved, different team structures are possible. The team structure guides the interactions and collaboration among team members and consequently plays an important role in a team's performance. Figure 1 depicts the fundamental elements of a collaboration system that we discuss in this article.

Key Points

- *Human computation systems* – Systems in which human actors perform assigned tasks in a precisely defined sequence (e.g., by following an algorithm). The execution is explicitly



Social Interaction Analysis for Team Collaboration, Fig. 1 Elements of a collaboration system controlled and coordinated by the system and expected to yield precise results (Law 2011).

- *Workflow management systems* – Systems that allow modeling of tasks and their execution scenarios. Notable representatives of such systems are the various business process management (BPM) systems. Although tasks can be performed by human actors, the traditional understanding of the notion of a workflow system does not include an integrated management of human-performed tasks.
- *Mixed systems* – Systems where both human and computer actors process the tasks. Humans are deeply integrated into the system, making both types of actors first-class citizens of it. The decision on who processes a particular task can be made by the system. While computer-performed tasks are accurate, employing humans for certain tasks requires dealing with uncertainties both in terms of human behavior and the quality of results.
- *Crowdsourcing systems* – Systems in which the task is offered, rather than assigned explicitly, to an unknown and usually large group of people who can freely accept and perform the tasks.

Historical Background

The idea of combining research on how humans work, communicate, and cooperate and the research on how computer systems can efficiently support such collaborations led to the creation of an interdisciplinary research area known as *computer-supported cooperative work* (CSCW) in the 1980s (Grudin 1994). Initially, the research was focused on small-scale collaborations, e.g., within companies or interest groups. With the wide adoption of Internet technologies, service-oriented architectures (SOA), mobile and cloud computing, and especially social networks, nowadays it is possible to carry out large-scale collaborations, possibly involving thousands of collaborators across boundaries of multiple organizations and countries. In the absence of an agreed naming taxonomy, a number of nuanced terms are commonly used in literature to generally denote systems and platforms for managing collaborations: *social computing systems*, *socio-technical systems*, *hybrid collective adaptive systems (hCAS)*, and *cyber-human systems.* Some examples of today's well-established types of such collaboration systems are given as follows.

Team Collaboration Analysis

Team Properties

We consider three important team properties: (a) *actors* making up the team, with their different skills, qualities, and personalities; (b) *structure,* representing the relationships, interactions, scale, and elasticity of the actor population; and (c) different *forms of collaboration* among the actors.

Actors and Team Structure

Actor teams are usually modeled as undirected or directed (multi)graphs with nodes representing people or teams of people and edges representing relationships between them (Newman 2010). Often, the edge is associated with a set of properties quantifying the interaction between the two nodes it connects and annotated with a context, representing the type of the relationship (e.g., social, professional collaboration, trust). Therefore, a team network can be modeled as a graph consisting of nodes representing actors, sets of skills forming their profiles, and edges representing relationships and associated contexts of relationships (Caverlee et al. 2008).

Depending on the lifetime and the scale of the observed actor set, we can consider different structural properties with varying relevance. Small-scale actor teams are usually assembled relying on the compatibility of the actors and their relationships, allowing the delegation of task coordination to the actors themselves. Typical examples are expert teams where actors have previous mutual collaboration experience, of friend cliques, where the common social fabric is a promise of a successful collaboration. The lifetime of such teams is short, and the structure is mostly static and determined at assembly time. Large-scale collaborations exhibit properties of populations – an extended lifetime and complex structure changing dynamically throughout execution time, thus requiring efficient automated mechanisms for coordination. *Scale* and *turnover* (Lee and Paine 2015) are metrics used to describe the magnitude and volatility (or stability) of such collaborations. For open collaborations, the turnover represents the rate of actors joining/leaving the collaboration. For managed collaborations, the scale and turnover are important inputs for determining the next *elastic* actions to take (Riveni et al. 2014).

Forms of Computer-Supported Team Collaboration

Static Collaboration Static collaboration is characterized by well-defined, long-lasting/repetitive processes (tasks), executed by human actors with specific assigned roles. Such kind of collaborations is usually found in companies that encode and execute their daily business use cases as business processes by using workflow technologies. This collaboration type makes no use of the underlying social networks connecting the actors to alter or enhance the collaboration in any way. As such, this approach works well only in cases where the predictability of the process execution is high and where no adaptability is required.

Ad Hoc Collaboration Unlike static collaboration, the ad hoc collaboration is suitable when performing highly dynamic tasks that change in time or complex tasks that occur only once and are not repeated. In this type of collaboration, tasks are initially defined, but the actors performing them are provisioned only at runtime. Ad hoc collaborations often cross organizational boundaries and are distributed in nature – in terms of software services used, in terms of actors executing the tasks, as well as in terms of control. Actor provisioning can be fully automated or partially performed by the actors themselves, often relying on social and other underlying networks connecting the actors.

Open Collaboration In open collaborations, a task can be actively shaped by the actors. The actors (often belonging to a professional community or an interest-based community) contribute freely to the task resolution during runtime. A task is not strictly assigned to a particular actor, but instead it is editable by (m)any community members upon their wish. In this case, the coordination between the actors can affect the quality of the task (Kittur and Kraut 2008). Data quality is controlled by the system itself and/or by a designated entity, often relying on the feedback information from data users. Open collaboration is particularly suitable for longer running, best-effort tasks, with no strict quality and time constraints, but requiring distributed know-how.

Open-source development, Wikipedia, and community-based Q&A websites are among the best examples of open collaboration. Examples of open collaboration enabling technologies and platforms include cloud services (e.g., Amazon EC2), sharing and collaboration platforms (e.g., Dropbox, Google Docs, Mendeley, and some more secure and private ones such as ownCloud and ownDrive), and open-source repositories (e.g., GitHub, GitLab, SourceForge).

Most existing systems employ a combination of different collaboration types, attempting to reduce the respective limitations of individual types and offer more versatile collaborations. CrowdLang platform (Minder and Bernstein 2012) combines static and ad hoc collaboration by integrating human-provided services into workflow systems. The workflow is not fully static but can be designed as needed by (re)combining a number of generic (simple) collaborative patterns (e.g., iterative, contest, collection, divide and conquer). The CrowdComputer platform (Tranquillini et al. 2015) combines static and open collaboration. The tasks are executed following a workflow, but the tasks are split into atomic tasks offered to individual workers through different “tactics” (e.g., marketplace, auction, mailing list). The SmartSociety platform (Scekic et al. 2015) supports combining all three collaboration types, allowing the actors to actively participate in determining and executing the team and the workflow (collaboration plan).

Task Properties

Task Description

Considering the general nature of the tasks that can be handled by a team composed of human actors, describing tasks precisely and unambiguously is extremely challenging. The difficulty lies in expressing the information that needs to be interpreted by each actor in the same way. At the same time, the effort required to properly interpret a task's objectives must be considerably smaller compared to the effort required to perform the task itself. In practice, the tasks can be described, *informally* and *formally*:

- Informally describing tasks means expressing the required outcomes in natural language, accompanied with simple examples. This approach is usually taken by today's crowdsourcing platforms that handle simple tasks. Also, informal description may be preferred in cases where tasks require aesthetic judgment or when the required outcome of the task is too vague to be expressed more precisely (e.g., on websites running creativity contests).
- Formally describing tasks means employing a specific notation that precisely defines how the task should be processed and what should the outcome be. Formal task description is usually used in specific environments, most notably in business process modeling (BPM). Initial versions of the most prominent business control-flow languages, such as BPEL, did not support specification and invocation of human interactions. An extension to BPEL, known as BPEL4People, was proposed in 2005 to allow modeling of human interactions within business processes by introducing the concept of *people activities*. A people activity can be described according to the WS-HumanTask specification. In this way, humans can be internally represented as Web services and integrated into the system.

Task Structure and Complexity

Task structure directly influences the team structure. Different task structures and complexities demand specific types of collaboration in terms of communication form, coordination protocols, adaptation schemes, and outcome type. Subtask interdependencies are one of the fundamental factors determining the task structure and task complexity. One basic task structure categorization is that into *parallel* and *sequential* tasks. Parallel tasks contain subtasks that can be executed independently in parallel, while a sequential task is composed of subtasks whose execution must follow a strict order. A subtype of sequential tasks is *iterative* tasks, where the output of one actor is given as input to another actor for subsequent task execution. An experiment and analysis of parallel and iterative approaches in open systems can be found in Little et al. (2010).

Apart from subtask interdependencies, other nonstructural factors can influence a task's complexity, such as the following: (a) number of atomic tasks, (b) growth (Dustdar and Bhattacharya 2011) (the number of atomic tasks can grow in runtime, necessitating team-size adaptability), and (c) task cardinality (tasks can be designed to be executed by one or many actors in one-to-one, many-to-one, many-to-many, and few-to-one fashion). See Quinn and Bederson (2011) for details and examples.

Interaction Processes

Team Formation

The problem of team formation consists of selecting suitable actors to perform a given task (out of a larger group of available actors) and organizing them in a collaborative structure. The first problem with identifying "suitable" actors is that suitability is highly context dependent and difficult to define precisely. Furthermore, suitability can have many different aspects. For example, the minimal suitability requirement for an actor is to possess the skills to perform the task. But, at the same time, for a successful teamwork, factors like trust, motivation, experience, and personal relations with other team members can be equally important.

Initially, the research focused on locating individual best-matching actors for a required set of skills and other individual properties. However, a group of top individuals does not guarantee the quality of their collaboration. Subsequent research efforts began taking into account the underlying social relations among the actors (e.g., friendships, managerial relations, previous business interactions, interests, connectedness, and social trust). Finally, recent systems aim to include human actors as first-class citizens allowing them to actively influence at runtime the team formation process (Rovatsos et al. 2015).

After selecting suitable actors, the following step in ensuring a successful collaboration is setting up a collaborative organization and environment. Although collaboration patterns in a team often resemble those in the underlying social networks, other factors like coordination cost, user preferences, and context are also important.

Whichever the properties considered, they are always measurable and quantifiable, meaning that the problem of team formation can be ultimately expressed as an optimization problem where we want to optimize certain performance aspects of the team as a whole (speed, quality, cost, and response time) while respecting the fixed constraints.

In general, team formation can be:

- *Self-organizing* – The actors themselves lead the team formation in a collective-intelligence fashion and set up the collaboration environment. The system assists the process (e.g., by enforcing negotiation rules, counting the votes) but does not make decisions on actor participation.
- *Centralized* – Team formation and setting up of collaborative environment is managed by the system, including the decision on participating actors.

Wikipedia and open-source community are striking examples of how self-organizing teams can perform well. The assumption is that the actors taking part in collaboration will perform best if they are given the possibility to modify and adapt the collaborative environment. This includes also the initial team formation. For example, in Gaston and DesJardins (2005) the authors investigate a system that enables actors to locally modify their collaborative environment according to their social network preferences (i.e., to *rewire* the local network topology) with the goal of achieving globally noticeable, collective performance improvement.

The most problematic aspect of self-organizing teams is the discrepancy between local and global effects. Although we rely on the collective intelligence of the actors, in practice, actors may not know how or when to modify the local network to achieve global improvements, since their actions are based upon their partial views only.

Centralized team formation is entirely handled by the system. Internally, the system can employ an algorithm or human actors to assemble the team:

- *Human-managed team formation* relies on human actors offering their referrals and recommendations via Web services, thus leveraging crowdsourcing techniques to identify the best candidates from their social networks. An example of such a system is PeopleCloud (Lopez et al. 2010).
- *Algorithmic team formation* relies on an algorithm to select actors and assemble the team. A lot of research efforts have been directed in this sense, producing a number of different algorithms. In Schall and Dustdar (2010) and Schall et al. (2012), the authors modify the well-known page ranking algorithms PageRank and HITS to identify the best team members, based on their previous interactions. In Lappas et al. (2009) and Anagnostopoulos et al. (2012), the goal is to minimize the total coordination cost of the newly established team, while in Dorn and Dustdar (2010), the optimal team is chosen as a trade-off between skill coverage and actor connectivity. Anagnostopoulos in Anagnostopoulos et al. (2010) presents an algorithm for forming minimum size teams with minimum workload that satisfies required team skills. Kargar et al. present team formation algorithms with communication and personnel cost minimization in An et al. (2013). Sagar in Sagar (2012) has presented a formal model for task assignment and finding team collaborators in social

networks by using shortest distance in the network. In Caverlee et al. (2010), the social trust between the team members is regarded as the most important factor in forming efficient collaborations.

Task Assignment and Delegation

Routing and Delegations Task delegation mechanisms are being explored as forms of coordination and load balancing in human computation. The concept of *social routing* is introduced in Dustdar and Gaedke (2011) as a form of delegation of tasks by task owners to actors from their social, professional, and other context-based community networks or the crowd. The so-called social router can be a software service that actually does the task forwarding across different types of networks depending on the requirement of the actor wishing to delegate the task.

Historical data on delegations (e.g., the executed/delegated tasks ratio, frequently used delegates) can serve as a good indicator of actor's role, performance, and social relationships. For example, a high number of task delegations testify a coordinating/managing role. However, the same information, if interpreted properly, is a potential indicator of actor's laziness. Moreover, it was shown (Sun et al. 2014) that actors favor the familiarity with the delegates over their assessed suitability for the given task. Delegation data can be used as metrics in actor selection and team formation algorithms. Moreover, delegation measures can be used in trust inference mechanisms. For example, Riveni et al. (2015) present a trust model where the rate of successfully executed delegated tasks is included in the definition of an actor's reliability metric, which in turn is included in the actors trust score. On the other hand, if the receivers of delegated tasks are considered trustworthy, new trust-based links will be created between the delegator and the delegates (Skopik et al. 2010).

Delegation Patterns in Business Process Activities The four main delegation patterns, detailed in Kloppmann et al. (2005), are:

- *Nomination* pattern allows predefined actor (s) to decide to whom to assign a task.
- *Escalation* pattern allows transfer of responsibility for task execution to other human actors when the originally assigned actor cannot meet task's time constraints. When an escalation is triggered, actors designated as escalation recipients are notified and allowed to decide how to proceed with the task execution, possibly reassigning it to another actor. When dealing with uncertainties related to human processing, this is an inevitable mechanism that must be supported.
- *Chained execution* pattern forces the actors to perform a specific sequence of actions, where the concrete actions may be determined only in runtime. The actors can be assigned a new action only after completing the previously assigned one. This pattern allows implementing a "bag of tasks" kind of parallelism, with each actor repeatedly removing tasks from the bag.
- *Four eyes principle* pattern allows two actors to take a public or a private decision on the same issue independently *(separation of duties*).

Algorithmic Task Life Cycle Management In cases when subtasks are clearly delimited and subtask dependencies are static and do not change in time, parallelizing a task execution is fairly easy. Some application domains, such as crowdsourcing systems, are characterized by exactly such properties.

This has led researchers to dedicate a lot of effort to automate task life cycle management transparently for the programmer, by developing a number of programming language extensions/libraries that work on top of existing commercial crowd-sourcing systems, such as Amazon Mechanical Turk. The extensions are typically able to automatically split a task; to assign/offer the subtasks to the actors in the crowd respecting the dependency, cost, and time constraints; and to merge the processed subtasks into the final resulting task. Additionally, automated quality control processes may be also offered. Most commonly, these are based on peer reviews or on a combination of redundant processing and majority rule. For example, an image that needs to be tagged may be submitted to multiple actors, but the aggregated result will contain only tags suggested by multiple independent actors. The data quality requirements can have a direct influence on task assignment, as they may introduce assignments not explicitly required by the user but performed transparently by the system. In fact, the main purpose of algorithmic handling of task assignment is exactly to move the burden of task life cycle management from the user to the system.

Collaboration systems can manage task assignments automatically throughout the entire execution time, repeating them when needed. For example, Little et al. (2009) show a system offering the possibility of iterative task execution, by reassigning previously processed tasks a number of times in order to improve the final quality of work by incrementally building upon previous work. In Marcus et al. (2011), a system can autonomously decide when to assign pleasing tasks to specific actors in order to motivate/reward them.

Another major advantage of algorithmic task assignment is the cost optimization. For large-scale collaborations, the system is able to assign the tasks in such a way to reduce the coordination costs better than human managers could do. For example, the task can be assigned to actors possessing similar professional skills and backgrounds, or the system can adjust task prices and time allotments based on the feedback obtained from monitoring data (Barowy and Berger 2012).

Collaboration Monitoring and Analysis

Monitoring and analyzing collaborative processes is necessary to gather important metrics about the performance of teams and actors and the quality of processed tasks. Such metrics are then used to detect bottlenecks, improve performance, and decide on appropriate compensation of the actors. As these metrics play a fundamental role in determining overall collaboration efficiency and costs, every collaboration system must support some kind of monitoring and analysis functionalities.

Monitoring can be performed during the runtime of a collaborative process *(active monitoring*) or it can be performed post-runtime, e.g., by *log mining*. Log mining is usually considered a part of more complex analysis processes, known as *workflow/process mining* (van der Aalst 2011; Zhang and Serban 2007).

Active monitoring is suitable for detecting anomalies that require quick responsive actions and team adaptations. An example of a system capable of monitoring and analyzing SOA-based collaborative processes can be found in Truong and Dustdar (2009).

Log mining, on the other hand, is used to gather less obvious information about the internal functioning of the team, since it considers the backlog of all recorded actions performed during previous collaborations. This allows discovery and prediction of critical execution paths, expected workload distribution, actor performance, and identification of previously unknown collaborative social networks, e.g., the network of most trusted colleagues or the groups of workers that together collaborate most efficiently as a team.

Metrics lie at the very heart of any monitoring and analysis process. In the following section, we present an overview of the metrics used by different collaboration systems.

Collaboration Metrics and Patterns

Metrics used in collaboration systems can be divided into three major categories:

- *Structural metrics* – defining the mathematical properties of the social/collaborative network connecting the actors.
- *Interaction metrics* – defining various properties of individual actors or actor groups, emerging as the result of past interactions.
- *Quality metrics* – defining quality criteria for actor performance and for task outcome data.

Structural Metrics and Network Patterns

Structural metrics and network patterns are based on mathematical properties of the social graph connecting the actors in a collaboration network (team). They provide useful insights into the functioning and self-organization of actors in a team. Structural metrics are well researched. In the following, we briefly mention some of main structural metrics:

- *Centrality measures* – include various metrics that identify the importance of an actor within a network in different contexts of importance. Some of the *most important centrality metrics are* degree centrality, closeness centrality, betweenness centrality, *and* eigenvector.
- *Structural groups* – They refer to various group patterns that can be identified within networks, such as *core* (denoting a subset of actors within a network where each actor is connected to at least *k* other actors within the same subset), *k-component* (denoting a subset of actors in which each two actors are connected by at least *k* independent paths), and *clique* (denoting a subset of actors all directly connected to each other).
- *Transitivity and reciprocity* – Transitivity reflects the "friend-of-a-friend" concept, i.e., if an actor *a* is connected by an edge to another actor b, and *b* is connected to c, then *a* is also connected to c. Reciprocity, on the other hand, denotes the probability that actor *b* points to actor *a* if actor *a* points to *b*.
- *Similarity* – It is defined by *structural equivalence* and *regular equivalence* metrics. See Newman (2010).

Details about all these and other metrics, as well as about ranking algorithms, can be found in Newman (2010).

Interaction Metrics

Interaction metrics can be defined at two levels: *individual level* (targeting individual actors) and *group level* (targeting multiple actors or the entire team). Individual interaction metrics describe a property of an individual actor that is shaped by the interaction in which the actor has participated. Group interaction metrics describe properties of particular interactions between actors, possibly including the collaboration as a whole.

Certainly, the most important actor-level metrics are *skill coverage* and *trust.* Skill coverage represents a degree to which an actor or a team possesses necessary skills to perform a task. This metric is important because it describes how much a team's set of skills deviates from the optimal one for a given task. The problem of matching skills is equivalent to the problem of functional matching in Web service compositions.

Trust, as a computational concept, was formalized in Marsh (1994) and since then it has been seen as a metric of great importance for selection of appropriate actors during the team formation phase. Trust is defined as an indicator of an actor's expectation about another actor's future behavior based on knowledge from previous interactions, and which inherently involves a degree of uncertainty about this behavior and its outcomes. Trust is highly context dependent, and one actor may have information about several scope-specific trust values for another actor. A scope can be the membership in a professional network, social network, or a collaboration team. Inferring trust is important in several cases:

- For actor discovery and team formation algorithms, when determining actor suitability for specific tasks
- For team optimization, adaptation, and risk management purposes
- For delegation mechanisms, e.g., when selecting a collaborator that may be a part of the extended team structure for the purpose of load balancing in cases of unexpected load

We can distinguish three types of trust based on the types of actors and interactions that are taken into account for its inference:

- *Local trust* or *direct trust* (sometimes also called *private reputation*) – firsthand trust, inferred from the outcome of an actor's previous interactions with the trustee
- *Recommendations* – secondhand trust inferred from the outcome of past interactions between a well-trusted entity and the trustee
- *Global trust* or *reputation* – aggregated community trust, inferred from outcomes of past interactions between third-party actors and the trustor (Skopik et al. 2009)

Other actor-level metrics include *task familiarity* and *team familiarity* (Espinosa et al. 2007). These are especially important for open collaboration where the system cannot assign a task to appropriate and trusted actors. If some of the actors within an open collaboration are already familiar with other actors, the coordination will be positively affected.

Team familiarity is important in large teams where effective team coordination is more difficult. Team familiarity is a function of multiple other metrics such as quality of prior interactions with a coworker, prior belonging to the same team, and prior experience with the same team structure and organization. Hence, this measure is closely related to *trust*.

Homophily is another metric closely related to team familiarity, where an actor chooses a collaborator based on profile similarity. Fazel-Zarandi et al. (2011) conclude that homophily (e.g., gender, same interests, tenure status) is considered more important in choosing research collaborators than, for example, their qualification level.

Task familiarity is best explained with an example of open source software development team. The bigger the number of interdependent modules, the more complex is the task. This increases the amount of information to be processed by human actors; thus, it is important that actors have a reasonable amount of task familiarity. Details of a model for performance analysis of teams based on task familiarity and team familiarity can be found in Espinosa et al. (2007).

Group-level metrics describe performance properties of a collaboration. One of the fundamental metrics describing collaborations is the team size. The bigger the number of collaborating actors, the more communication and coordination among them is needed. For example, in Kittur and Kraut (2008), the authors use Wikipedia to analyze how the number of editors and the coordination methods affect the article quality in terms of accuracy, completeness, and clarity.

A metric indicating *interaction intensity* between an actor and other important actors is measured in specific interaction contexts. It is used in the aforementioned *DSARank* ranking algorithm (Schall and Dustdar 2010).

The relevance of the connections to important actors is the most important factor in determining the reputation of an actor. The reliability of the feedback information in reputation systems depends on the reputation of actors providing the feedback. Reputation information is valuable when an actor lacks information based on direct experiences with another actor. However, when this information is available and appropriate, the private or direct trust weights more than trust values based on reputation data. In this case the weight of data from direct interactions should be determined by calculating the minimum number of direct/local trust or rating values that should be maintained by an actor for the actor providing the service/executing a task (Noorian et al. 2012).

Collaboration cost is an important metric because of its direct business influence. This metric takes into account not only the price of task processing paid to the actors, but rather the total costs, including the communication and coordination costs. It is used as the basis for the cost optimization algorithms, as shown in the following systems – Quirk (Marcus et al. 2011) and AUTOMAN (Barowy and Berger 2012).

Automatically discovering *collaboration patterns* naturally occurring among actors opens up a possibility to identify particularly (un)successful collaboration groups or execution sequences. This information can in turn be used to optimize collaborative process. Identifying collaboration patters is one of the central topics of process mining.

Quality Metrics

Quality of Data (QoD) Metrics As collaboration systems deal with various human-performed tasks, and the data quality primarily depends on the type of tasks, trying to develop a general set of quality metrics makes little sense. For example, metrics listed in Table 1, such as data completeness, freshness, and accuracy, are well-known metrics, but their definition is highly dependent on the goal of their use. Instead, different metrics are developed for particular application domains (e.g., Hu et al. (2007) and de La Robertie et al. (2015) present models to assess article quality in Wikipedia). However, it is exactly the fact

Social Interaction Analysis for Team Collaboration, Table 1 Overview of metrics and patterns used in collaboration systems

Structural metrics		Centrality measures (Degree, closeness, betweenness, eigenvector, etc.)
		Structural groups (Cores, components, cliques)
		Transitivity, reciprocity
		Similarity, equivalence
Interaction metrics	Actor level	Trust, reputation functional/skill coverage
		Task familiarity, team familiarity
	Group level	Structural groups
		Team size
		Link quality, interaction intensity
		Collaboration patterns (Delegations, escalations, redundant processing, iterative processing, etc.)
Quality metrics	Quality of data (QoD)	Uncertainty, completeness, accuracy, freshness, relevancy etc.
	Performance	Availability, response time, success rate, etc.
	Rewarding and incentives	Effort, productivity, quality of work

that humans participate in the collaborative processes that introduces a concept common to all the application areas – that of *uncertainty* or *inaccuracy* (Parameswaran and Polyzotis 2011). The main sources of uncertainty are caused by the dynamic and unexpected behavior of humans: humans make mistakes, are subjective, and exhibit malicious/dysfunctional behavior. Thus, approaches for dealing with uncertainty should be included in supporting systems.

Different research communities deal with uncertainty differently. However, all approaches rely on some probability metrics that quantify our belief that a single task is performed correctly. In principle, all approaches can be divided into two categories:

- *Optimistic approaches* – Processed tasks are returned along with a confidence (accuracy) estimate. The data user accepts the results but must be aware that a certain percentage of the results will be wrong.
- *Pessimistic approaches* – The system applies various mechanisms for error detection and correction and usually resubmits the task to multiple actors until the merged result satisfies the required quality threshold.

Actor performance quality metrics are similar to the "traditional" Web service metrics, like average execution time, number of invocations, and availability. On the group and collaboration level, these metrics measure and predict the existence of various invocation patterns, i.e., the probabilities that certain services will be called in a particular order with respect to other services. A detailed discussion on interaction metrics can be found in Truong and Dustdar (2009).

Incentives and rewarding are important and effective mechanisms for indirectly influencing quality and motivation of human actors in collaborations. The principal metrics in use in today's computer-supported collaboration systems are:

- *Effort* – It measures an actor's determination to perform a task. The main purpose of this metric is to provide a way to compare the performance of both experienced and inexperienced actors. For example, an inexperienced actor may put in a lot of his time and resources only to perform a task worse or slower than an experienced actor. However, for the purpose of incentivizing, a higher effort level should be compensated with a higher reward, because it will ultimately lead to better experienced actors.
- *Productivity* – It expresses the number of units processed in a time period. This metric is suitable for piecework and easily quantifiable tasks (e.g., bug reporting, image tagging, text translation). Kasunic (2008) has defined effort and productivity metrics for software projects.
- *Quality of work* – This metrics expresses the quality of the working process of an actor. It should not be confused with the quality of data (QoD) of processed tasks. This metric is used to assess actors when the task's QoD cannot be

easily determined or when it cannot say much about the actor. For example, actors that help other actors, waste less resources, provide creative ideas, or take responsibility should be also rewarded. In such cases, the subjective opinions of other relevant actors (i.e., *peers*) can be used to quantify these elusive actor qualities.

In order to acquire the rewarding metrics, collaborative systems use different *evaluation methods*, relying both on human and machine actors:

- Individual evaluation methods
  - *Quantitative methods* – They represent a quantitative measurement of an individual actor's contribution as measured by the system itself. Such metrics can represent the number of processed tasks, average speed, responsiveness, acceptance rate, etc. These methods are considered fair and cheap to implement, but unfortunately they are applicable only in cases where actors work on easily quantifiable tasks.
  - *Subjective methods* – In cases where the quality of work is a property understandable to humans only, a quantitatively expressed subjective assessment by a human actor replaces a quantitative metric measured by the system itself. This is the case with artistic, design, and engineering tasks. The advantages are the simplicity and cost, but a serious drawback is the inevitable lack of objectivity.
- Group evaluation methods
  - *Peer evaluation methods* – They are used to express an aggregated opinion of an interest group. The members of evaluation group usually express their votes by scoring tasks or actors on a fixed scale or by investing amounts of virtual credits expressing their confidence (placing bets). The quality and effectiveness of these methods are influenced by the size of the composition of the evaluation group.
  - *Indirect evaluation methods* – In certain situations, human actors can be evaluated by comparing the status of the artifacts they previously produced with the status of the artifacts produced by other members of the same community. The artifacts can be Web pages, projects, articles, photos, and programming code. These comparisons are usually performed with the help of sophisticated algorithms. An example is the Google's PageRank algorithm, impact factor for scientific publications, or Klout's algorithm for measuring social network influence. Advantages and disadvantages of these methods are dependent on the properties of the algorithm.

Key Applications

Taken individually, many of the described techniques and algorithms have found practical used in today's systems: Structural and interaction metrics are used in social network platforms and recommender systems for determining trusted groups, which in turn are used for improving the quality of recommendations. The same techniques are at the same time a useful tool for sociological, ethnological, medical, and forensic research. The described evaluation methods are used in software development industry and crowdsourcing platforms, where a nonautomated evaluation would be impractical due to the scale of the collaborative effort. The task assignment and delegation patterns are used in workflow management systems, which are standard tools for automating complex or critical business processes in medium and large companies. Many of the team formation algorithms are adapted from the algorithms originally used in service-oriented architectures for service composition.

The wish of the authors of this article was to draw attention to the prospective benefits that these techniques can bring when used together in the context of emerging collaborative systems. These systems (e.g., Scekic et al. 2015; Minder and Bernstein 2012; Tranquillini et al. 2015) need to manage the entire cycle of human participation and will thus require in future Kittur et al. (2013) application of many of the presented methods and metrics and will hopefully drive the development of novel ones.

Future Directions

Although a considerable amount of work is done in the area of interaction analysis in social networks, there is much less work conducted on team-based metrics and analysis. Many open questions still remain to be tackled. Some of them are (i) understanding the interdependencies between metrics for better analysis of different collaboration systems, testing, and evaluating these team-based metrics and (ii) utilizing these metrics in the most appropriate way for task adaptation. Another future research direction in team collaboration in mixed systems is to develop metrics that can be used to compare human and software-based actors.

Cross-References

- ▶ Behavior Modeling in Social Networks
- ▶ Collaboration Patterns in Software Developer Network
- ▶ Collective Intelligence: Overview
- ▶ Community Detection and Recommender Systems
- ▶ Community Detection in Social Network: An Experience with Directed Graphs
- ▶ Community Detection: Current and Future Research Trends
- ▶ Extracting and Inferring Communities via Link Analysis
- ▶ Incentives in Collaborative Applications
- ▶ Link Dynamics and Community Formation in Social Networks
- ▶ Modeling Social Behavior
- ▶ Modeling Social Preferences Based on Social Interactions
- ▶ Network Structure Analysis
- ▶ Recommender Systems based on Social Networks
- ▶ Reputation Systems
- ▶ Rewarding
- ▶ Social Groups in Crowd
- ▶ Social Influence Analysis
- ▶ Web Service Composition
- ▶ Web Service Infrastructure Patterns

References

An A, Kargar M, ZiHayat M (2013) Finding affordable and collaborative teams from a network of experts. In: Proceedings of the 13th SIAM international conference on data mining, Austin, 2–4 May 2013, pp 587–595. https://doi.org/10.1137/1.9781611972832.65

Anagnostopoulos A, Becchetti L, Castillo C, Gionis A, Leonardi S (2010) Power in unity: forming teams in large-scale community systems. In: Huang, Jimmy & Koudas, Nick & Jones, Gareth J. F. & Wu, Xindong & Collins-Thompson, Kevyn and An, Aijun. CIKM, pp 599–608

Anagnostopoulos A, Becchetti L, Castillo C, Gionis A, Leonardi S (2012) Online team formation in social networks. In: Proceedings of the 21st international conference on World Wide Web-WWW '12. ACM Press, New York, p 839. http://dl.acm.org/citation.cfm?id=2187836.2187950

Barowy D, Berger E (2012) AUTOMAN: A platform for integrating human-based and digital computation. http://www.cs.umass.edu/~emery/pubs/AutoMan-UMass-CS-TR2011-44.pdf

Caverlee J, Liu L, Webb S (2008) Socialtrust: tamper-resilient trust establishment in online communities. In: Proceedings of the 8th ACMIEEECS joint conference on digital libraries. ACM, pp 104–113. http://portal.acm.org/citation.cfm?id=1378889.1378908

Caverlee J, Cheng Z, Eoff B, Hsu CF, Kamath K, Kashoob S, Kelley J, Khabiri E, Lee K (2010) SocialTrust++: building community-based trust in social information systems. In: The 6th international conference on collaborative computing: networking, applications and worksharing, CollaborateCom 2010, Chicago, 9–12 Oct 2010, pp 1–7. https://doi.org/10.4108/icst.collaboratecom.2010.40

de La Robertie B, Pitarch Y, Teste O (2015) Measuring article quality in wikipedia using the collaboration network. In: Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015, ASONAM 15. ACM, New York, pp 464–471. https://doi.org/10.1145/2808797.2808895

Dorn C, Dustdar S (2010) Composing near-optimal expert teams: a trade-off between skills and connectivity. In: On the move to meaningful Internet systems: OTM 2010, pp 472–489. http://www.springerlink.com/index/H434570G12787H57.pdf

Dustdar S, Bhattacharya K (2011) The social compute unit. IEEE Internet Comput 15(3):64–69. https://doi.org/10.1109/MIC.2011.68. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5755601

Dustdar S, Gaedke M (2011) The social routing principle. IEEE Internet Comput. Vol. 15, No. 4. pp. 80–83, https://doi.org/10.1109/mic.2011.97

Espinosa JA, Slaughter SA, Kraut RE, Herbsleb JD (2007) Familiarity, complexity, and team performance in geographically distributed software development. Organ Sci 18(4):613–630. https://doi.org/10.1287/orsc.1070.0297

Fazel-Zarandi M, Devlin HJ, Huang Y, Contractor N (2011) Expert recommendation based on social drivers, social network analysis, and semantic data representation. In: Proceedings of the 2nd international workshop on information heterogeneity and fusion in recommender systems, HetRec 11. ACM, New York, pp 41–48. https://doi.org/10.1145/2039320.2039326

Gaston ME, DesJardins M (2005) Agent-organized networks for dynamic team formation. In: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems – AAMAS '05, p 230. https://doi.org/10.1145/1082473.1082508. http://portal.acm.org/citation.cfm?doid=1082473.1082508

Grudin J (1994) Computer-supported cooperative work: history and focus. Computer 27(5):19–26. https://doi.org/10.1109/2.291294. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=291294

Hu M, Lim EP, Sun A, Lauw HW, Vuong BQ (2007) Measuring article quality in wikipedia: models and evaluation. In: Proceedings of the sixteenth ACM conference on conference on information and knowledge management, CIKM 07. ACM, New York, pp 243–252. https://doi.org/10.1145/1321440.1321476

Kasunic M (2008) A data specification for software project performance measures: results of a collaboration on performance measurement. Technical report, Carnegie Mellon University, Software Engineering Institute. http://books.google.at/books?id=Un3NSgAACAAJ

Kittur A, Kraut RE (2008) Harnessing the wisdom of crowds in wikipedia. In: Proceedings of the ACM 2008 conference on computer supported cooperative work – CSCW '08. ACM Press, New York, San Diego, pp 37–46. http://portal.acm.org/citation.cfm?doid=1460563.1460572

Kittur A, Nickerson JV, Bernstein M, Gerber E, Shaw A, Zimmerman J, Lease M, Horton J (2013) The future of crowd work. In: Proceedings of the 2013 conference on computer supported cooperative work, CSCW '13. ACM, pp 1301–1318. https://doi.org/10.1145/2441776.2441923

Kloppmann M, Koenig D, Leymann F, Pfau G, Rickayzen A, Schmidt P, Trickovic I (2005) WS-BPEL extension for people (July):1–18. http://public.dhe.ibm.com/software/dw/specs/ws-bpel4people/BPEL4People_white_paper.pdf

Lappas T, Liu K, Terzi E (2009) Finding a team of experts in social networks. In: Proceedings of the 15th ACM international conference on knowledge discovery and data mining, vol 7120(4). p 467. http://portal.acm.org/citation.cfm?doid=1557019.1557074

Law E (2011) Defining (human) computation. In: Proceedings of the Workshop on Crowdsourcing and Human Computation, in conjunction with CHI

Lee CP, Paine D (2015) From the matrix to a model of coordinated action (MoCA): a conceptual framework of and for CSCW. In: Proceedings of the 18th ACM conference on computer supported cooperative work &#38; social computing, CSCW 15. ACM, New York, pp 179–194. https://doi.org/10.1145/2675133.2675161

Little G, Chilton LB, Miller R, Goldman M (2009) TurKit: tools for iterative tasks on mechanical turk. IEEE. https://doi.org/10.1109/VLHCC.2009.5295247. http://ieeexplore.ieee.org/ lpdocs/epic03/wrapper.htm?arnumber=5295247

Little G, Chilton LB, Goldman M, Miller R (2010) Exploring iterative and parallel human computation processes. In: Proceedings of the 28th of the international conference extended abstracts on human factors in computing systems – CHI EA '10, p 4309. https://doi.org/10.1145/1753846.1754145. http://portal.acm.org/citation.cfm?doid=1753846.1754145

Lopez M, Vukovic M, Laredo J (2010) PeopleCloud service for enterprise crowdsourcing. In: 2010 I. E. International conference on services computing, pp 538–545. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5557275

Marcus A, Wu E, Karger DR, Madden S, Miller RC (2011) Platform considerations in human computation. In: Workshop on crowdsourcing and human computation

Marsh SP (1994) Formalizing trust as a computational concept. University of Stirling, Stirling

Minder P, Bernstein A (2012) Crowdlang: a programming language for the systematic exploration of human computation systems. In: Aberer K, Flache A, Jager W, Liu L, Tang J, Guret C (eds) Social informatics, LNCS, vol 7710. Springer, Berlin/Heidelberg, pp 124–137

Newman MEJ (2010) Networks: an introduction. Oxford University Press, New York, NY, USA

Noorian Z, Fleming M, Marsh S (2012) Preference-oriented QoS-based service discovery with dynamic trust and reputation management. In: Proceedings of the 27th annual ACM symposium on applied computing – SAC '12, p 2014. https://doi.org/10.1145/2245276.2232111. http://dl.acm.org/citation.cfm?doid=2245276.2232111

Parameswaran A, Polyzotis N (2011) Answering queries using humans, algorithms and databases, Tech. rep. Stanford University. http://ilpubs.stanford.edu:8090/986/

Quinn AJ, Bederson BB (2011) Human computation: a survey and taxonomy of a growing field. In: Proceedings of the 2011 annual conference on human factors in computing systems – CHI '11. ACM Press, New York, p 1403. http://dl.acm.org/citation.cfm? id=1978942.1979148

Riveni M, Truong HL, Dustdar S (2014) On the elasticity of social compute units. In: Jarke M, Mylopoulos J, Quix C, Rolland C, Manolopoulos Y, Mouratidis H, Horkoff J (eds) Advanced information systems engineering. Lecture notes in computer science, vol 8484.

Springer International Publishing, Cham, Switzerland, pp 364–378 https://doi.org/10.1007/978-3-319-07881-6_25

Riveni M, Truong HL, Dustdar S (2015) Trust-aware elastic social compute units. In: Trust- com/BigDataSE/ISPA, IEEE, vol 1, pp 135–142

Rovatsos M, Diochnos D, Craciun M (2015) Advances in social computing and multiagent systems. In: 6th International workshop on collaborative agents research and development, CARE 2015 and Second international workshop on multiagent foundations of social computing, MFSC 2015, Istanbul, Turkey, May 4, 2015, Revised Selected Papers. Springer International Publishing, Cham, chap Agent Protocols for Social Computation, pp 94–111. https://doi.org/10.1007/978–3–319-24804-2_7

Sagar AB (2012) Modeling collaborative task execution in social networks. In: Potdar V, Mukhopadhyay D (eds) CUBE. ACM, pp 664–669. http://dblp.uni-trier.de/db/conf/cube/ cube2012.html#Sagar12

Scekic O, Schiavinotto T, Diochnos DI, Rovatsos M, Truong HL, Carreras I, Dustdar S (2015) Programming model elements for hybrid collaborative adaptive systems. In: 2015 I.E. Conference on Collaboration and Internet Computing (CIC), pp 278–287. https://doi.org/10.1109/CIC.2015.17

Schall D, Dustdar S (2010) Dynamic context-sensitive PageRank for expertise mining. In: Proceedings of the second international conference on social informatics, Springer, Berlin/Heidelberg, pp 160–175. http://dl.acm.org/citation.cfm?id=1929326.1929338

Schall D, Skopik F, Dustdar S (2012) Expert discovery and interactions in mixed service-oriented systems. IEEE Trans Serv Comput 5(2):233–245. https://doi.org/10.1109/TSC.2011.2. http://doi.ieeecomputersociety.org/10.1109/TSC.2011.2; http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5710867

Skopik F, Schall D, Dustdar S (2009) The cycle of trust in mixed service-oriented systems. In: 35th Euromicro conference on software engineering and advanced applications, pp 72–79. https://doi.org/10.1109/SEAA.2009.20. http://ieeexplore.ieee.org/lpdocs/epic03/ wrapper.htm?arnumber=5349860

Skopik F, Schall D, Dustdar S (2010) Modeling and mining of dynamic trust in complex service-oriented systems. Inf Syst 35(7):735–757. https://doi.org/10.1016/j.is.2010.03.001. http://linkinghub.elsevier.com/retrieve/pii/S0306437910000153

Sun H, Srivatsa M, Tan S, Li Y, Kaplan LM, Tao S, Yan X (2014) Analyzing expert behaviors in collaborative networks. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, KDD 14, ACM, New York, pp 1486–1495. https://doi.org/10.1145/2623330.2623722

Tranquillini S, Daniel F, Kucherbaev P, Casati F (2015) Modeling, enacting, and integrating custom crowdsourcing processes. ACM Trans Web 9(2):1–43

Truong HL, Dustdar S (2009) Online interaction analysis framework for ad-hoc collaborative processes in SOA-based environments. Framework 260–277. https://doi.org/10.1007/978-3-642-00899-3_15. http://www.springerlink.com/index/u1075446t8qr727q.pdf

van der Aalst WMP (2011) Process mining. Springer, Berlin/Heidelberg. https://doi.org/10.1007/978–3–642-19345-3. http://www.mendeley.com/research/no-title-avail/; http://www.sciencedirect.com/science/article/pii/ S0166361503001945; http://www.springerlink.com/index/10.1007/978-3-642-19345-3

Zhang P, Serban N (2007) Discovery, visualization and performance analysis of enterprise workflow. Comput Stat Data Anal 51(5):2670–2687. https://doi.org/10.1016/j.csda.2006.01.008 http://linkinghub.elsevier.com/retrieve/pii/S0167947306000132

Social Internet of Mobile Things and Decision Support Tools

Monica Wachowicz[1], Sangwhan Cha[1] and Chiara Renso[2]

[1]University of New Brunswick, People in Motion Lab, Fredericton, Canada

[2]ISTI Institute of National Research Council, Pisa, Italy

Synonyms

SIoT; Smart Social Objects; Social network of IoT devices; Social Virtual Objects

Glossary

Decision support system According to Geoffrion's definition, a DSS has six characteristics (Geoffrion 1983): (1) Is designed to solve ill- or semi-structured problems, i.e., where objectives cannot be fully or precisely defined; (2) Has an interface that is both powerful and easy to use; (3) Enables the user to combine models and data in a flexible manner;
(4) Helps the user explore the solution space (the options

	available to them) by using the models in the system to generate a series of feasible alternatives; (5) Supports a variety of decision-making styles, and easily adapted to provide new capabilities as the needs of the user evolve; (6) Allows an interactive and recursive process in which decision-making proceeds by multiple passes, perhaps involving different routes, rather than a single linear path
Internet of Things (IoT)	Smart-connected things with the ability (1) to be identifiable (anything identifies itself), (2) to communicate (anything communicates), and (3) to interact (anything interacts) without human intervention
Social Internet of things (SIoT)	An IoT where things are capable of establishing social relationships with other things, without human intervention
Social Internet of Mobile Things (SIoMT)	Is a specific architecture for the Internet of things which includes functionalities required to integrate mobile things into a social network
Mobility data	Is any type of large volume data sets – structured and unstructured data – containing the information about the positions of a moving IoT device. It is usually represented as trajectories
Social data	Unstructured data created and shared by individuals through social networks such as Facebook, LinkedIn, Twitter, and Foursquare. It is usually represented as graphs
Data streaming	Is the transfer of data at a steady high-speed rate for making sure that enough data is being continuously received without any noticeable time lag. It is the real-time human perception of the data
Streaming analytics	Data mining techniques that allow us to analyze streaming data from the Internet of things (IoT) in real time
Cloud computing	A kind of network-based computing to facilitate the sharing of on-demand computing resources so that users are able to manage these resources without knowing the complexity and details of the underlying infrastructure. The cloud computing includes Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS)
Hadoop	An open-source MapReduce framework for distributed systems. The Hadoop system can be scaled up to thousands of computers corresponding to needs of data storage and processing power with fault tolerance

Definition

The Internet of Things has been largely recognized as a global network interconnecting humans with RFIDs, sensors, actuators, smartphones, computers, buildings, home/work appliances, cars, and any other device with the goal of unlocking a new combination of applications and services (Biswas 2014; Welbourne et al. 2009). From a data management perspective, IoT consists of devices acting as providers and consumers of data needed for supporting an application or service. From a networking point of view, IoT is a system architecture that supports point-to-point communications, preferably in real time. Due to the large device heterogeneity, data streams from IoT systems need to be analyzed in real time for enabling users to combine models and data in a flexible manner in decision support systems (Zheng et al. 2010). The main goal is to generate actionable intelligence to accelerate decision-making.

The social Internet of things originates from the idea of a social network of “friendly” things, where a level of trustworthiness exists for leveraging the type of interaction among IoT devices that are friends or not. Atzori et al. (2012) propose five elementary relational models based on Fiske’s theory (Fiske 1992) for modeling how IoT devices can interact without human intervention. They can be described as one of the following:

- Parental object relationship: homogeneous IoT devices belonging to one person.
- Ownership object relationship: heterogeneous IoT devices that collaborate because they belong to the same person.
- Co-location object relationship: homogeneous and heterogeneous IoT devices from different people located at the same place but unlikely to communicate and interact with each other.
- Co-work object relationship: IoT devices collaborate to provide a common application to a group of people.
- Social object relationship: collaboration is established when IoT devices belonging to different people come into contact, sporadically or continuously.

Mobile IoT devices add more complexity to these models because the relationships will vary through time because of their locations. In other words, the above relationships between IoT devices will be dependent on their proximity. The research challenge is how to manage these relationships because they will support a variety of decision-making styles and need to be adapted to the mobility patterns which are exhibited by these devices.

Introduction

The growth of Internet of Things brings the promise of a wide range of new decision support systems due to the expected 57 billion smart-connected devices by 2025, from which 80% are expected to be mobile things such as autonomous cars, wearables, and drones (Gartner and Huang 2016). The integration of social networks with the Internet of Mobile Things is revolutionizing the traditional field of decision-making analysis, not only to scale up to the expected massive real-time data volumes but also to address complex questions related to streaming analytics for understanding change, trends, duration, and evolution.

The social Internet of Things is not a new concept. In fact, the idea of friendship between devices was first introduced by Holmquist et al. (2001), when the design of Smart-Its-enabled devices (i.e., wireless sensors) was proposed to integrate sensing, processing, and communication to facilitate awareness of any device’s proximity, independent of the computing architecture. Friends could take two devices they would like to connect and move them together, by holding them in their hands, briefly waving or shaking them.

In the social Internet of Mobile Things, “things” connected to the Internet are clearly distinguished from the “things” participating in social networks through their connection to the Internet and their mobility on demand. By generating massive data sets that are continuously streaming as they occur in real time, it will be possible to explore the evolution dynamics of the urban social sphere; to predict the spreading of sentiments, opinions, and diseases; and thus to understand in real time the evolving borders of the community structure of a city.

For example, behavior recognition in smart homes often employs graphical models like hidden Markov chains. By combining them with contextual information about location and time, the performance of these models can be boosted (e.g., see Chua et al. 2009). Such cross-fertilizations are clearly identifiable in the previous work in cognitive vision (Dubba et al. 2010), where the demonstrated interactions and integrations of techniques from machine learning, inductive logic programming, and location intelligence may serve as a blueprint for the construction of hybrid intelligent decision support systems in the near future. The research challenge is twofold: (1) how the social relationships should be modeled for mobile IoT devices without human intervention and (2) what computing architecture is needed to integrate sensing, processing, analytics, and communication.

Key Points

The social Internet of Mobile Things plays an important role in improving our understanding of social process over space and time, such as influence, trust, and information spreading. Decision-making needs to take into consideration different sorts of relationships between mobile IoT devices, thereby extracting valid, novel, and useful patterns in networks ranging from transportation networks and World Wide Web to social networks (Xu et al. 2016). This area is rapidly evolving to provide examples of new techniques and applications, leading to future research directions.

Developing the next generation of decision support systems based on the social Internet of mobile things relies on harvesting relevant socio-mobility data from different sources, processing these data accordingly to user requirements, running the algorithms as the data arrives, exploring and visualizing the information, and, finally, storing the data streams into an appropriate database. Previous research has been focused on batch processing using parallel platforms (Riedy and Bader 2013). This is mainly because supporting real-time streaming data in decision support systems is still in its infancy. One of the major challenges of developing real-time streaming processing comes from developing a relevance index to retrieve stream data that is actually needed for the decision-making analysis.

Another major challenge of developing the next generation of decision support systems comes from designing a cloud-based architecture in conjunction with an analytical workflow. In terms of architecture, relying on a single server or single processing computing unit to analyze stream data in real time is the most inefficient way in terms of processing speed, scalability, reliability, and stability (Hua et al. 2015; Song and Kim 2013). The alternative is cloud computing, which is based on a collection of independent distributed computers that appears to the users of the system as a single coherent system (Lee 2015). Very few attempts are found in the literature in implementing a cloud-based architecture for supporting decision support systems.

Finally, another challenge is to develop an analytical workflow that provides a set of analytical steps that are designed to perform tasks such as data ingestion, data processing/data exploration, data visualization, and data storing in a cloud-based system. Barker and Hemert (2007) argue that workflows should be reused, refined over time, and shared with other scientists in the field. They must be fully reproducible, indicating where the data originated, how it was altered, and which components and what parameter settings were used. This will allow other scientists to reconduct the analysis in different decision support systems.

Historical Background

The high rates of incoming stream data requires new techniques for efficient data ingestion. Magdy et al. (2014, 2015) propose an aggregation technique to update a set of counters for each incoming microblog keyword over a hopping window of the last T time units. It can digest arrival rates up to 64,000 microblogs/sec. For data processing, nearest-neighbor queries can be performed using ZIP distribution properties to minimize the number of visited microblogs/keywords and, as a result, proving an average query latency of 4 msec.

Thuraisingham (2014) proposes a scalable feature selection and extraction solution that leverages a cloud computing architecture. The preliminary results indicate that depending on the availability of cluster nodes, the running time for feature extraction and selection can be reduced by a factor of m, where m is the number of nodes in the cloud cluster. Based on this work, Abrol et al. (2015) have adapted the same architecture for a variety of applications, including security, marketing, law enforcement, healthcare, emergency response, and finance.

Streaming data analytics in real time has been made possible with Apache Storm and Apache Spark recently. In Solaimani et al. (2014), the Apache Spark has been chosen for their system architecture to analyze stream data in real time because of the high processing speed obtained from detecting anomaly for multisource. Their framework monitors VMware performance stream data (e.g., CPU load, memory usage, etc.) continuously. It collects these data simultaneously from all the VMwares connected to the network. The empirical results show that Spark processes a tuple much quicker than Storm on average.

Since Apache Spark performs data-parallel computations while Apache Storm performs task-parallel computations (Ericsson Research 2015), Apache Storm is more suitable for our proposed streaming data analytics workflow. However, more research is needed on how to build a system architecture for streaming data analytics based on the use of Apache Storm versus Apache Spark.

Streaming analytics for the social Internet of Mobile Things is challenging because it requires efficient querying and retrieval of data produced rapidly (Yang et al. 2012; Raina et al. 2014). There are important issues to be considered in order to build an analytical workflow, to design a system architecture for a variety of applications of decision support systems, as well as to deploy the distributed computing resources to handle real-time updates. In order to exploit the power of large-scale supercomputing without the need to invest into expensive operational infrastructure costs for building the efficient decision support systems, Osman et al. (2013) propose cloud and distributed computing should be always considered as the best options.

In Veen et al. (2015), which is the closest approach to our research in terms of system architecture, distributed real-time sentiment analysis for social media streams is proposed. They focus on distributed data mining system with data stream processing using distributed learning algorithms on top of Apache Storm. Since they deal with all data coming from data sources for streaming data analytics on the limited memory size, it is not applicable to our system architecture. Our approach here is that streaming data is filtered out by a relevance index before data processing. However, there is a possibility to use their approach in our future research if data filtered by the relevance index becomes too large to be processed on the proposed cloud architecture.

SIoMT Platform

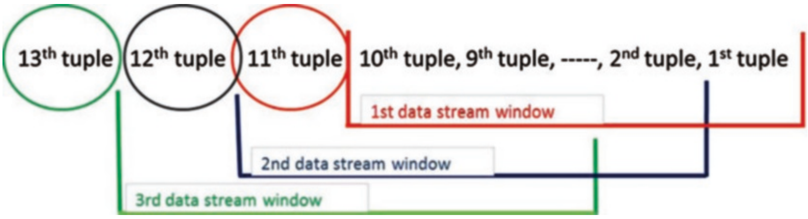
We describe a possible architecture for the SIoMT, which includes the functionality required to integrate mobile things into a social network.

Relevance Index

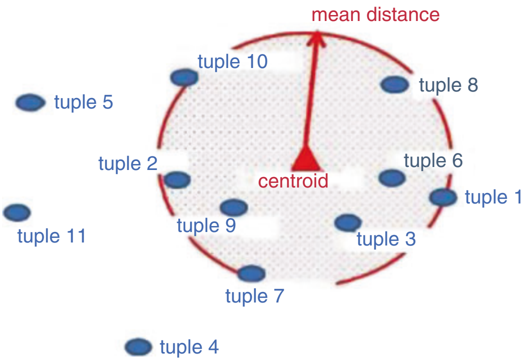
The continuous output of a mobile IoT device can be represented as a sequence of unbounded tuples of the form (a_1,x_1, y_1, t_1); (a_2,x_2,y_2, t_2); (a_3, x_3,y_3, t_3) . . . .(a_n, x_n,y_n, t_n) where an attribute, image, or a text *a* is generated continuously in time *t*, sometimes having geographical coordinates *x* and *y*.

The relevance index is computed to determine whether the streaming tuples are relevant to a specific relationship (i.e., parental object; ownership object, co-location; co-work object and social object relationships). This is achieved by creating data stream windows containing only the most recent N tuples. In other words, the stream data is treated as a possibly infinite sequence of tuples that captures many of the characteristics of such massive data sets, as distinct from standard data sets that can be loaded into a traditional RDBMS and be analyzed offline.

The intuition is that the spatial proximity of the tuples can be used to compute a relevance index, which is a numeric measure to indicate how much relevant the arriving tuples from the mobile IoT devices are. The index is continuously computed using a sliding window every time a new tuple arrives as shown in Fig. 1. It consists of calculating the centroid which is a simple heuristics equivalent to the mean center of a tuple point distribution. Geometrically, it is the average X and Y coordinates for a spatial point distribution as given by the equation below (see Eq. 1).



Social Internet of Mobile Things and Decision Support Tools, Fig. 1 Sliding data stream windows



Social Internet of Mobile Things and Decision Support Tools, Fig. 2 The mean distance computation computed for the first data stream window

$$\overline{X}_{Coord} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \overline{Y}_{Coord} = \frac{\sum_{i=1}^{n} Y_i}{n} \tag{1}$$

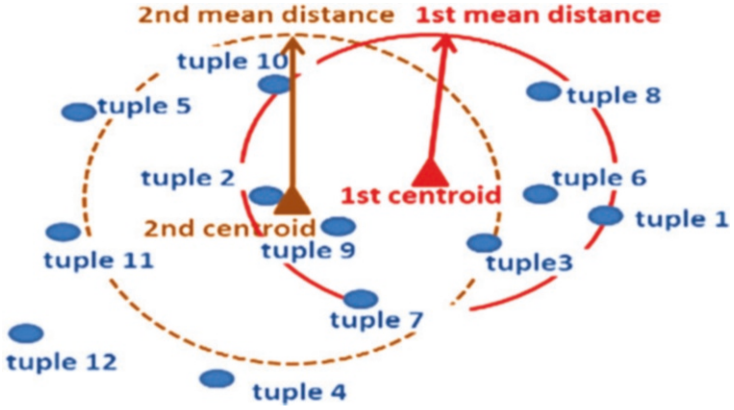
Once the location of a centroid is known, the mean distance is calculated as the average of the greatest and least distances between the centroid and a tuple point belonging to the same data stream window (Fig. 2).

If the distance between the centroid and a tuple point is smaller than the mean distance, the tuple is considered relevant. In other words, the data being generated from mobile IoT devices within such a circle will be used for the decision-making analysis. In the example illustrated in Fig. 3, it means that the tuples 1, 2, 3, 6, 7, 8, 9, and 10 are relevant. On the other hand, if the distance between the centroid and a tuple point is greater than the mean distance, the tuple is considered not relevant for the decision-making analysis. In our example, it means that the tuples 4, 5, and 11 are discarded. This process is repeated every time a new stream data window is computed with the arrival of a new tuple. Figure 3 illustrates this process when a new tuple 12 arrives in the cloud, and as a result, a second centroid is calculated as well as a second mean distance, and the tuples 4, 5, 6, 8, 11, and 12 are discarded.

The computation of the relevance index is shown by the pseudo-code algorithm below.

Algorithm: Relevance Index

```
Input : $ST_i$ /* $i^{th}$ Streaming Data from a mobile IoT device */
Output : status /* fully-hydrated objects */
begin
Load $ST_i$
Let listLat = null;
Let listLong = null;
Let geoLocation = status.getGeoLocation ();
If ( geoLocation != null)
latitude = geoLocation.getLatitude();
longitude = getLocation.getLongitude ();
if (listLat.size () <10)
listLat.add(latitude);
listLong.add(longitude);
else
if (getRadius () >= getDistanceIoTdevices (latitude, longitude)
this._outputCollector.emit(new Values (satus));
listLat.remove(0);
```



Social Internet of Mobile Things and Decision Support Tools, Fig. 3 The mean distance computation computed for the second data stream window

```
listLong.remove(0);
getLatMean();
getLongMean();
getDistanceTwoGeoCoordinates
(latitude1, longitude1, latitude2,
longitude2)
end
```

The preliminary results are promising on paving the way to support relevance indexes for decision support systems in the future. Future research will be toward developing real-time topic derivation or topic modeling to improve the relevance index in terms of semantics. A semantic relevance index could be used to classify the stream tuples according to different relationships (i.e., parental object; ownership object, co-location; co-work object and social object relationships). Toward this end, there is a need to develop appropriate dictionaries that can support different relationships that can occur at a particular place.

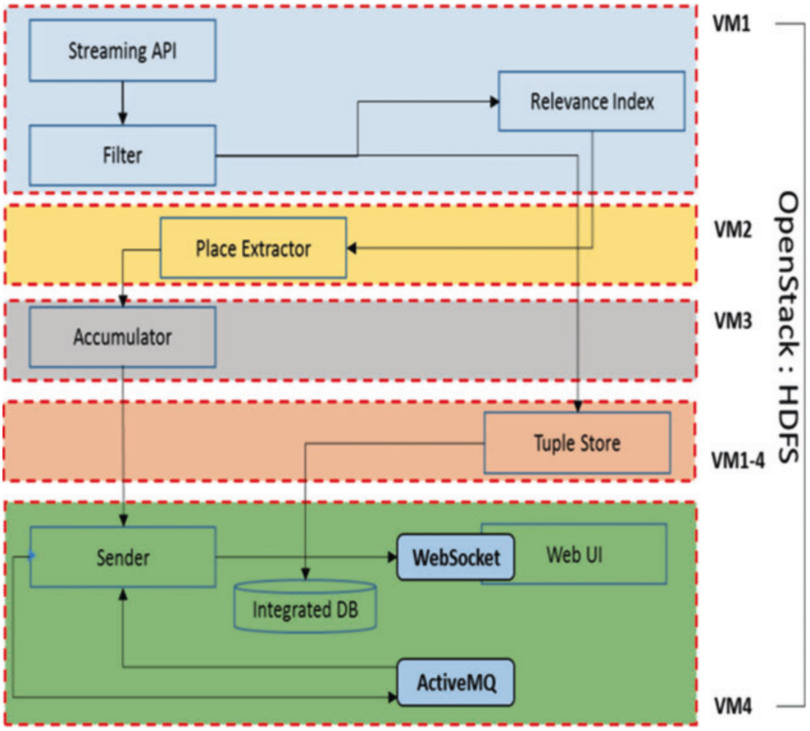
IoT Cloud Architecture

The proposed analytics workflow was first introduced in Cha and Wachowicz (2015). Our cloud infrastructure is provided by Compute Canada that is a partnership with regional organizations ACENET, Calcul Québec, Compute Ontario, and WestGrid to lead the acceleration of research innovation by deploying state-of-the-art advanced research computing (ARC) systems, storage, and software solutions in Canada (Compute Canada 2015). Each virtual machine consists of 8 VCPUs, 450 GB total disk, and 32 MB ram. The main advantage to use the Compute Canada's cloud infrastructure is that the system can be scaled up and down as needed. We have implemented four Storm clusters installed in four virtual machines (VMs) in an OpenStack cloud based on a Linux environment. OpenStack is an open-source cloud computing platform for private and public clouds, focusing on infrastructure as a service (IaaS) (Babu 2015).

Although there are several stream processing frameworks available, Apache Storm was selected because it is open source, has an active community, and is efficient for streaming data processing (Veen et al. 2015). Storm distinguishes between two kinds of nodes in a cluster. They are the master node and the supervisor node. The master node runs the so-called Nimbus daemon that is responsible for assigning tasks to the worker nodes, monitoring the clusters for failures, and distributing the computing code of an algorithm around the VM.

The worker nodes run an instance of the so-called Supervisor daemon. This daemon listens for work assigned (by Nimbus) to the node it runs on and starts/stops the worker processes as necessary. Each worker process executes a subset of a topology. In addition to its own components, Storm relies on a ZooKeeper cluster (consisting of one or more ZooKeeper servers) to perform the coordination between Nimbus and the Supervisors (Noll 2013).

In addition to its own components, Storm relies on a ZooKeeper cluster (consisting of one or more ZooKeeper servers) to perform the coordination



Social Internet of Mobile Things and Decision Support Tools, Fig. 4 Overview of our OpenStack cloud architecture

between Nimbus and the Supervisors (Noll 2013). Zookeeper is a maintaining service for a Hadoop cluster, which provides an open-source distributed configuration service, a synchronization service, and a naming registry.

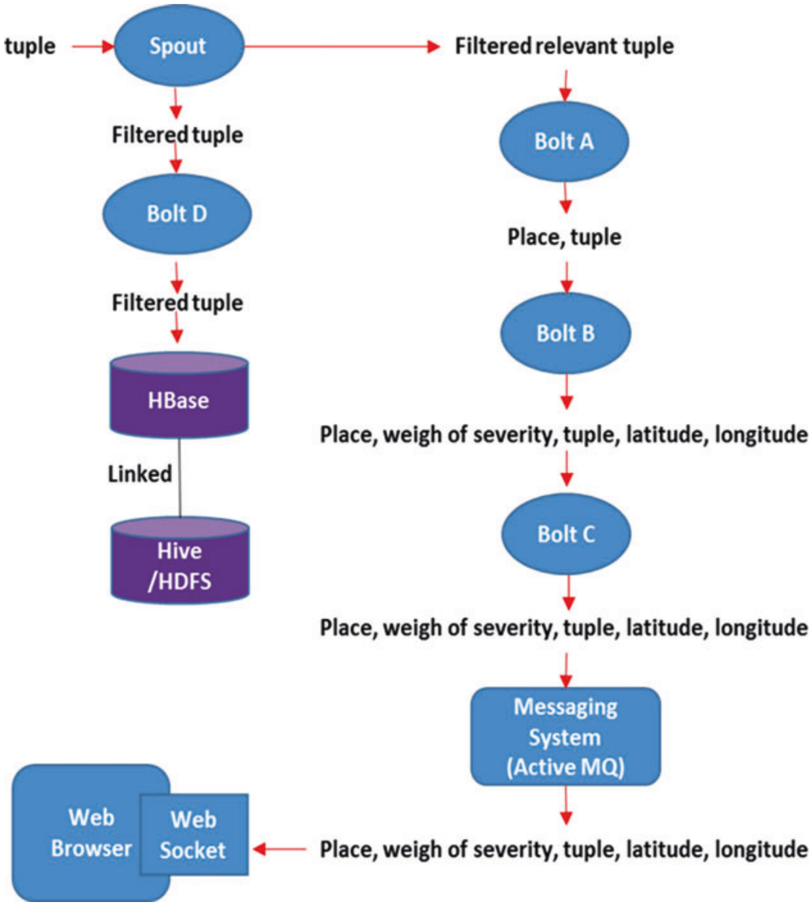
The Storm topology consists of two components: spout and bolt. The spout of Storm is used for pulling out large data streams from the messaging system queue or streaming API and sending them to a bolt of Storm which consumes any number of input streams, does some processing, and possibly emits new streams to other bolts. In our cloud architecture, each spout and bolt of Apache Storm is assigned to one VM by a Nimbus node. Therefore, a Storm cluster consists of a nimbus node, three supervisor nodes, and three Zookeeper servers. The failure of a single Zookeeper server could cause the shutdown of the whole Storm cluster.

As illustrated in Fig. 4, VM1 is used as a spout node, which handles the data ingestion and data processing tasks. The data ingestion is implemented using a streaming API which allows high-throughput near-real-time access to public and protected user-generated contents such as Twitter, Instagram, and Flickr (Fig. 5).

The selected relevant tuples of the relevance index of the spout are sent to the place extractor of the bolt A in VM2, which processes the relevant tuples for reverse dynamic geocoding using a Bing Maps API. The aim is to associate coordinates on the map with an address or a place name in real time.

The geocoded tuples of the bolt A are sent to the accumulator of the bolt B in VM3, which performs parsing to analyze the string of characters in order to associate groups of characters with the syntactic units of the underlying grammar. The final result consists of a place information, the relevance index, the content of tuple, and the geographical coordination. They are sent to the sender of a bolt C in VM4.

The sender sends a tuple as message through ActiveMQ and WebSocket to web browsers for


tuple
Spout
Filtered relevant tuple
Filtered tuple
Bolt A
Bolt D
Place, tuple
Filtered tuple
Bolt B
HBase
Place, weigh of severity, tuple, latitude, longitude
Linked
Hive /HDFS
Bolt C
Place, weigh of severity, tuple, latitude, longitude
Messaging System (Active MQ)
Web Browser
Web Socket
Place, weigh of severity, tuple, latitude, longitude

Social Internet of Mobile Things and Decision Support Tools, Fig. 5 Data flow among components

data visualization. In order to store entire tuples into HBase/HDFS, tuples coming from the filter in VM1 are emitted to the tuple store of the bolt D in VM1–VM4.

In order to make our topology run on four VMs, the bolt D needs to have four tasks and executors. For further analytics in batch processing, the HBase table has to be linked with a Hive Store.

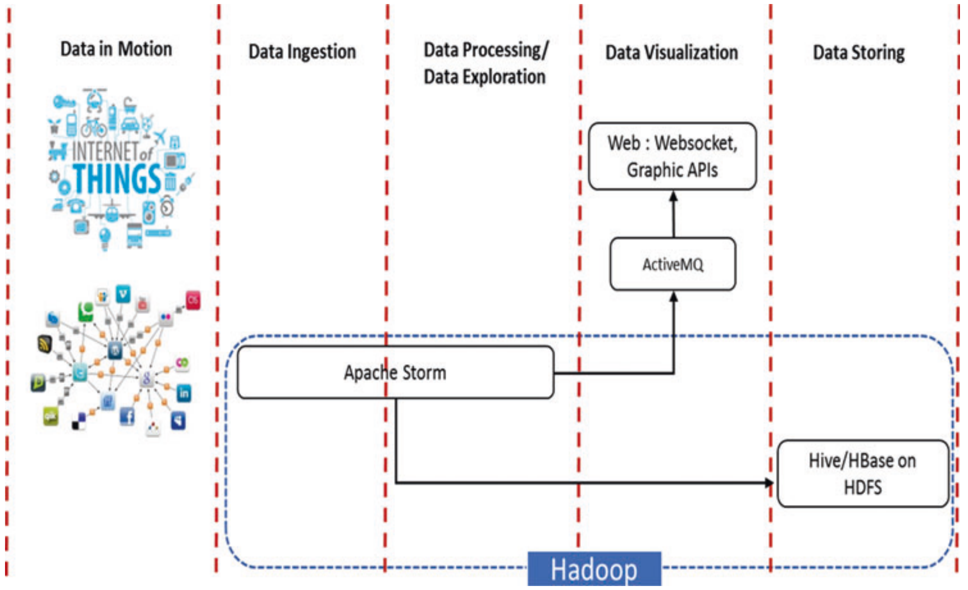
We used the “HBaseStorageHandler” to register the HBase table, which is “tuple_events,” with a Hive metastore. For the schema mapping, the SerDe property mapping, which performs one-to-one mapping between HBase and Hive table, was used. For example, the “id” column in the Hive table “tuple_events” is mapped to “i” column in the “events” column family of the HBase table “tuple_events”; Fig. 6 illustrates entire data flow among components.

The example of linking the HBase table with the Hive Store is illustrated using the following query statement:

CREATE EXTERNAL TABLE tuple_events(key string, id string, user_tags string, download_url string, page_url string, longitude string, latitude string, date_taken string, valdate_uploaded string) STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'

WITH SERDEPROPERTIES ("hbase.columns.mapping" = ":key,events:i, events:u,events:d,events:p,events:lo,events:la, events:dt,events:du") TBLPROPERTIES ("hbase. table.name" = "tuple_events");

We used the “HBaseStorageHandler” to register the HBase table, which is “tuple_events,” with a Hive metastore. For the schema mapping, the SerDe property mapping, which performs one-to-



Social Internet of Mobile Things and Decision Support Tools, Fig. 6 Overview of the streaming analytical workflow

one mapping between HBase and Hive table, was used. For example, the "id" column in the Hive table "tuple_events" is mapped to "i" column in the "events" column family of the HBase table "tuple_events"; Fig. 5 illustrates entire data flow among components.

Streaming Analytics

Tsai et al. (2014) point out that "it is much easier to create data than to analyze data." This is particularly applicable to SIoMT because finding the mobility patterns from relevant IoT data allows us to take the actionable decisions based on new insights. The analytics workflow described here was first introduced in Cha and Wachowicz (2015). In this workflow, real-time streaming analytics consists of five tasks: data ingestion, data processing, data exploration, data visualization, and data storing in a cloud-based architecture based on the Hadoop ecosystem.

Figure 6 provides an overview of how the tasks and the Hadoop-related projects are implemented to support them.

The data processing task makes direct access to stream data in raw formats from different mobile IoT devices. This task can also transform well-structured data as well as unstructured data into human readable structured data using algorithms, which can perform aggregation, simplification, conversion, and sorting.

The Apache Storm is selected to support both data ingestion and processing tasks. There could be multiple steps for complex stream transformations, which require multiple bolts. The results from the processing tasks are kept in memory and updated in real time.

The data exploration task is focused on applying data mining algorithms for discovering mobility patterns. However, most of the algorithms available today cannot be applied directly to process the large amount of data generated from SIoMT, and they therefore have to be redesigned to be properly applied to SIoMT.

Clustering is one of the most well-known data mining algorithms, and it follows the idea of grouping together similar objects such that the intra-cluster similarity is maximized, while it minimized the inter-clustering similarity. An example of the use of clustering for IoT is to group users based on their behavior to provide better tailored services. Approaches based on distributed clustering are typically used in wireless sensor networks

(WSNs). In this context, data mining algorithms suffer from some dilemmas, especially energy conservation to maximize the lifetime sensors. Many new clustering algorithms (e.g., LEACH; see Tsai et al. 2014) have been developed for the WSNs that are probably the most common devices to be found on the SIoMT.

Dynamic clustering algorithms take into consideration the characteristics of the SIoMT by exchanging information between nodes (Kardeby 2011): each node can select an interested sensor node based on the current needs of the application dynamically. Since things may be expected to make decisions by themselves, the design of clustering algorithm for the SIoMT needs to take into account the use of active network to create collaborative, multi-hop, and dynamic interaction or to exchange information between objects (Tsai et al. 2014). It is therefore important for clustering algorithms to forward information to a network and other distant objects.

Interestingly, other research work focuses on the behavior of a social network. Here things are the smartphones, personal digital assistants, or other personal devices. The clustering algorithm can be therefore used to find out the social relationships between the users and mobile nodes. The clustering results can be used to predict the trace of mobile nodes and to provide the awareness information to the users or to find out latent communities existing in the SIoMT which will be useful for connecting things to things, people to things, or people to people (Tsai et al. 2014).

We are currently exploring spectral clustering algorithms for discovering mobility patterns (Tran & Wachowicz 2017). Normally they consist of three main tasks which are (1) constructing the similarity matrix and the graph Laplacian for a given set of data tuples having a local symmetric and nonnegative weight associated to each edge that connects a node, (2) computing the whole set of eigenvectors of for the graph Laplacian using power methods, and (3) clustering the graph nodes according to applying the k-means clustering technique to the first k eigenvectors of the graph usually applying a standard clustering algorithm such as k-means.

Spectral clustering algorithms typically start from local information encoded in a weighted graph on the data points and cluster according to the global eigenvectors of the corresponding (normalized) similarity matrix (Nadler and Galun 2006). Computing eigenvalues is important to explain data tuples because the second largest eigenvalue of a graph can provide us with information about expansion and randomness properties of mobility patterns, and the spectrum of the similarity matrix provides a useful invariant.

With the advent of massive data sets, most of the recent research proposed in spectral clustering has been focused on reducing the run time complexities of these tasks. Instead of computing the whole set of eigenvectors of the random walk graph Laplacian, the authors Lin and Cohen (2010) and Thang et al. (2013) propose the use of the power method to compute the largest "pseudo-eigenvector" of a random walk similarity matrix by using power method and applying k-means clustering method to this largest "pseudo-eigenvector" in order to uncover the patterns. Random walks have been used to reach unknown data points of different data sets by generating random clusters (Lovász 1993). This makes this approach a candidate to clustering SIoMT data. However, each pseudo-eigenvector is a linear combination of all original eigenvectors of the random walk similarity matrix, including the information not only from the "useful" eigenvectors but also from the "noise" eigenvectors. In Ye et al. (2016), the independent component analysis (ICA) is proposed to tackle this problem, and the kurtosis-based selection strategy is used to increase cluster partitioning. This proposed solution is promising, but even simple computations require prolonged runtimes.

Key Applications

SIoMT has been recognized as a revolution in ICT during the past few years, and various applications such as smart homes, environmental monitoring, healthcare, and smart cities have emerged in the market (Yang et al. 2014; Li et al.

2011). This requires SIoMT platforms to be responsive in nature and anticipate users' needs according to different contexts they are in by means of intelligent components, devices, and applications.

Analyzing SIoMT data can be rephrased in *how*, *when*, *where*, and *why* things move. This knowledge gives to a decision-maker – who could be a traffic manager or an urban planner – a better understanding on how to increase the efficiency of energy systems and the delivery of services ranging from utilities to retailing in cities and to improve communications and transportation.

The variety of relationships expected to occur between mobile IoT devices are fueling exponential growth in data volumes. This exponential growing of SIoMT data is comparable to a similar growing rate of unstructured data generated by social networks such as Facebook, Google+, Twitter, and Foursquare. This big data revolution will continue over the next decade and beyond with opportunities that will include:

– *Smart governance*: enhancement of citizen participation in the decision-making process, creation of new public and social services, support for transparent governance, and advance political strategies and perspectives.
– *Transport and ICT*: improve local accessibility and develop sustainable, innovative, and safe transport systems.
– *Smart cities*: IoT is becoming more and more popular in the smart cities context, where several kinds of sensors are collecting huge amount of heterogeneous data about the living cities. A proper decision system in this context is essential to take advantage of these data and help local administrators to take the most appropriate decisions to improve the citizen's quality of life.
– *Natural resources*: reduce pollution, support environmental protection, and achieve a sustainable resource management.
– *Quality of life*: create new cultural, housing, and education facilities and improve health conditions and individual safety.

Future Directions

There have been few SIoMT platforms developed for supporting decision support systems. They are usually designed bearing in mind common goals for IoT deployment such as streaming analytics, quality controlling, and maintenance. Therefore, more research is needed to design SIoMT platforms that can be easily chosen according to the type of sensor data and analytics required to support different decision-making analysis.

Cross-References

▶ Cloud Computing
▶ Clustering Algorithms

Acknowledgments This research has been funded by the NSERC/Cisco Industrial Research Chair in real-time mobility analytics and partially supported by H2020 project BASMATI (GA 723131).

References

Abrol S, Rajasekar G, Khan L (2015) Real-time stream data analytics for multi-purpose social media applications. Inf Reuse Integ (IRI):25–30

Atzori L, Iera A, Morabito G, Nitti M (2012) The social internet of things (siot)–when social networks meet the internet of things: Concept, architecture and network characterization. Comput Netw 56(16):3594–3608

Babu *P. apache* licenced open source. Available via https://github.com/P7h/StormTweetsSentimentD3Viz. Cited 18 Aug 2015

Barker A, Hemert JV (2007) Scientific workflow: a survey and research directions. In: Parallel processing and applied mathematics. Springer, Heidelberg, pp 746–753

Biswas AR (2014) IoT and cloud convergence: Opportunities and challenges. In: 2014 I.E. World Forum on Internet of Things (WF-IoT). IEEE, Piscataway, pp 375–376

Cha S, Wachowicz M (2015) Towards real-time streaming analytics based on cloud computing. Inter J Big Data (IJBD) 2:28–40

Chua SL, Marsland S, Guesgen HW (2009) Behaviour recognition from sensory streams in smart environments. In: Australasian conference on artificial intelligence. Springer, New York/London 5866, p 666–675

Compute Canada. Available via https://www.computecanada.ca/. Cited 20 Aug 2015

Dubba KSR, Cohn AG, Hogg DC (2010) Event model learning from complex videos using ILP. In: Proceedings ECAI. IOS Press, 215,p 93–98

Ericsson Research blog. Available via http://www.ericsson.com/research-blog/data-knowledge/apache-storm-vs-spark-streaming/. Cited 8 Sept 2015

Fiske AP (1992) The four elementary forms of sociality: framework for a unified theory of social relations. Psychol Rev 99:689–723

Gartner G, Huang H (2016) Progress in Location-Based Services 2016. Springer Heidelberg

Geoffrion AM (1983) Can OR/MS evolve fast enough? "Source for six essential characteristics of DSS". Interfaces 13:10

Holmquist LE, Mattern F, Schiele B, Alahuhta P, Beigl M, Gellersen W (2001) Smart-its friends: a technique for users to easily establish connections between smart artefacts. In: International conference on ubiquitous computing. Springer, Berlin/Heidelberg

Hua Y, He W, Liu X, Feng D (2015) SmartEye: Realreal-time and efficient cloud image sharing for disaster environments. In: The proceedings of IEEE conference on computer communications. IEEE, Piscataway, pp 1616–1624

Kardeby V (2011) Automatic sensor clustering: connectivity for the internet of things. Licentiate thesis, Mid Sweden University, Department of Information Technology and Media, Sundsvall

Lee I. Software system. Available via http://www.cis.upenn.edu/~lee/07cis505/Lec/lec-ch1-DistSys-v4.pdf. Cited 4 Sept 2015

Li X, Lu R, Liang X, Shen X, Chen J, Lin X (2011) Smart community: an internet of things applications. IEEE Commun Mag 49(11):68–75

Lin F, Cohen W (2010) Power iteration clustering. In: Proceedings of the 27th international conference on machine learning (ICML-10). p 655–662

Lovász L. (1993) Random walks on graphs. Combinatorics, Paul erdos is eighty 2: 1–46

Magdy A, Alsrabi L, Harthi SA, Muslesh M, Ghanem TM, Ghani S, Mokbel MF (2015) Demonstration of Taghreed: a system for querying, analyzing, and visualizing geotagged microblogs. Data Eng (ICDE) 1:1416–1419

Magdy A, Aly A, Mokbel M, Elnikety S (2014) Mars: realreal-time spatio-temporal queries on microblogs. Data Engineering (ICDE). 1238–1241

Nadler B, Galun M (2006) Fundamental limitations of spectral clustering. In: Advances in neural information processing systems. 1017–1024

Noll M (2013) Running Hadoop on Ubuntu Linux (Multi-Node Cluster). Retrieved from http://cs.smith.edu/dftwiki/images/MichaelNollHadoopTutorial2.pdf on February 2017

Osman A. EI-Refaey M, Elnaggar A (2013) Towards real-time analytics in the cloud. In: The proceedings of the 9th IEEE World congress on services. p. 428–435

Raina I, Gujar S, Shah P, Desai A, Bodkhe B (2014) Twitter sentiment analysis using apache storm. In: the international journal of recent technology and engineering (IJRTE) 3(5): 23–26

Riedy J, Bader D (2013) Multithreaded community monitoring for massive streaming graph data. In: 2013 I.E. 27th international parallel and distributed processing symposium workshops & PhD forum. 1646–1655

Solaimani M, Iftekhar M, Khan L, Thuraisingham B (2014) Spark-based anomaly detection over multi-source VMware performance data in real-time. In: IEEE symposium on computational intelligence in cyber security (CICS) 1–8

Song M, Kim MC (2013) RT2M: Real-time Twitter trend mining system. In: the proceedings of international conference on Social Intelligence and Technology 64–71

Tsai CW, Lai CF, Chiang MC, Yang LT (2014) Data mining for internet of things: a survey. IEEE Commun Surv Tutorials 15(1):77. First Quarter

Thang ND et al (2013) Deflation-based power iteration clustering. Appl Intell 39(2):367–385

Thuraisingham B (2014) Secure sensor semantic web and information fusion. Texas University, Dallas Richardson

Tran L, Wachowicz M (2017) Spectral Clustering for discovering location patterns of mobile IoT devices. Transactions in GIS, submitted

Veen JS, Waaij B, Lazovik E, Wijbrandi W, Meijer RJ (2015) Dynamically scaling apache storm for the analysis of streaming data. In: the proceedings of 1st international conference on big data computing service and applications 154–161

Welbourne E, Cole G, Gould K, Rector K, Raymer S, Balazinska M, Borriello G (2009) Building the internet of things using RFID: The RFID ecosystem experience. In: the journal of IEEE internet 13(3)

Xu K, Qu Y, Yang K (2016) A tutorial on the internet of things: from a heterogeneous network integration perspective. IEEE Network, 30(2):102–8

Yang G, Xie L, Mantysalo M, Zhou X, Pang Z, Da Xu L, Kao-Walter S, Chen Q, Zheng L (2014) A health –IoT platform based on the integration of intelligent packaging, unobtrusive Bio-sensor, and intelligent medicine box. IEEE Trans Ind Inf 10(4):2180–2191

Yang X, Ghoting A, Ruan Y, Parthasarathy S (2012) A framework for summarizing and analyzing twitter feeds. In: the proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining 370–378

Ye W, et al. (2016) FUSE: full spectral clustering. In: 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)

Zheng VW, Zheng Y, Xie X, Yang Q (2010) Collaborative location and active recommendations with GPS history data. In: Proceedings of the 19th international conference on World Wide Web 1029–1038

Recommended Reading

Granville V (2014) Developing analytical talent: becoming a data scientist. Wiley Press, Hoboken

Li S, Da Xu L, Zhao S (2015) The internet of things: a survey. Inf Syst Front 17(2):243–259

Vermesan O, Friess P, Guillemin P, Gusmeroli S, Sundmaeker H, Bassi A, Jubert IS, Mazura M, Harrison M, Eisenhauer M, Doody P. Internet of things strategic research roadmap. Vermesan O, Friess P, Guillemin P, Gusmeroli S, Sundmaeker H, Bassi A, et al. (2011) Internet of things: global technological and societal trends 1: 9–52

Memon N, Xu JJ, Hicks DL, Chen H (2010) Data mining for social network datadata. Annals of information systems, vol 12. Springer, New York/London, pp 47–74

Yu S, Lin X, Misic J, Shen X (2015) Networking for big data. CRC Press, Boca Raton

Social Knowledge Network

▶ Automatic Document Topic Identification Using Social Knowledge Network

Social Location

▶ Role Discovery

Social Media

▶ Mapping Online Social Media Networks
▶ Mining Trends in the Blogosphere
▶ NodeXL: Simple Network Analysis for Social Media
▶ Privacy Issues for SNS and Mobile SNS
▶ Reconnaissance and Social Engineering Risks as Effects of Social Networking
▶ Social History of Computing and Online Social Communities
▶ Social Media and Social Networking in Political Campaigns/Movements
▶ Social Media, Definition, and History
▶ Social Media Policy in the Workplace: User Awareness
▶ Social Provenance
▶ Topology of Online Social Networks
▶ Web Communities Versus Physical Communities

Social Media Analysis

▶ Analysis and Mining of Tags, (Micro)Blogs, and Virtual Communities
▶ Collective Intelligence: Overview
▶ Sentiment Analysis in Social Media
▶ Sentiment Analysis of Reviews
▶ Twitris: A System for Collective Social Intelligence

Social Media Analysis and Mining

▶ Social–Spatiotemporal Analysis of Topical and Polarized Communities in Online Social Networks

Social Media Analysis for Monitoring Political Sentiment

Cristina Bosco and Viviana Patti
Computer Science Department, University of Turin, Turin, Italy

Synonyms

Elections; Microblogging data; Opinion mining; Political debates; Politics; Sentiment analysis; Stance detection; Tracking political sentiment

Glossary

NLP	The acronym for natural language processing is used to indicate a field of artificial intelligence, also known as computational linguistics, concerned with the simulation of linguistic competence by computers that involve the interaction between computers and human natural

languages. It addressed several challenges, like natural language understanding and natural language generation, and tasks, like the identification of named entities within a text, the analysis of morpho-syntactic structure of sentences, or the analysis of opinions and sentiment (sentiment analysis) expressed in a message

Sentiment analysis
Sentiment analysis, which is also known as opinion mining, indicates the use of NLP and text analysis techniques for detecting subjective information in user-generated contents. These techniques have been widely applied on texts from microblogs, that is, reviews of products (e.g., mobile phones, shoes, cameras) and services (e.g., hotels, restaurants, spas), and social media. According to the data where sentiment analysis is applied the goal also varies, ranging from marketing to customer service, to prediction, to political position of a population, to polarity detection

Social media
Social media are computer-mediated technologies for viewing, creating, and sharing information among individual users and communities. Even if the term indicates different services, it is mainly used for the microblogs like Twitter and Facebook where user-generated contents, such as text posts or comments, digital photos or videos, as well as data generated through all online interactions, are posted and shared among users. Social media use web-based and mobile technologies on smartphones, tablets, and computers to create highly interactive platforms through which individuals, communities, and organizations can share, cocreate, discuss, and modify user-generated content or premade content posted online. They introduce substantial and pervasive changes to communication between businesses, organizations, communities, and individuals

Stance detection
Stance detection consists in automatically, by sentiment analysis techniques, determining from text whether the author is in favor of the given target, against the given target, or whether neither inference is likely. While humans can deduce from the text the speaker’s position toward the target, to successfully detect stance, automatic systems often have to detect and analyze other information that may not be present in the text but available as metadata. For example, considering that stance detection has been mainly applied in political texts from social media, only having knowledge about the author of a message we can interpret it in the right way as against or in favor of the addressed target of the post

Definition

The activity of monitoring political sentiment and opinions has been addressed by applying, to texts and evidences from different sources and media, techniques developed in the context of research areas ranging from sociology to linguistics. The recent trends in monitoring political sentiment consists in considering texts from social media, mainly microblogging data, and users’ digital traces, and in exploiting computational techniques for extracting information about the political landscape in the offline world. In particular, techniques such as sentiment analysis and opinion mining, developed within the context of computational linguistics for extracting several kinds of information about humans’ behavior, are here specifically declined for monitoring political contents and are gradually considered as especially useful for this purpose, with possible different focuses: detecting stance of users, detecting the polarity of messages expressing opinions about candidates in political elections, or detecting communities around polarized opinions in political debates.

Introduction

In the last decade, the birth and development of social media have changed our communication strategies and behavior, having an especially strong impact on politics. Social media are indeed low-cost platforms for connecting people which can be exploited also for promoting engagement among users in the political base of politicians or linking candidates with voters. Therefore, several works devoted to monitoring political sentiment are currently based on collections of digital trace data, i.e., documents that record the activities of users in digital services, the most of which are in textual form usually associated with metadata about author, time, or location. These works show that texts from these platforms, often called "user-generated contents," can be usefully exploited for detecting and monitoring opinions and sentiments about politics.

Nevertheless, the richness and complexity of communication about politics and the dynamics featuring the diffusion of these kinds of contents make crucially challenging the development of novel forms of computational analysis which can sustain the traditional ones in order to improve the collection of knowledge about political behavior of population and related sentiment. The computational techniques applied for monitoring political sentiment are forms of natural language processing, usually indicated as *sentiment analysis* (SA) or with the synonym expression *opinion mining*, where are usually clustered statistical and linguistic methods.

The result that the scholar wants to obtain studying social media determines the task to be performed on data, which can vary from identifying the user influence on a debate or discussion held in social media to the detection of the disposition of a user toward the given party or position in a debate, which is reported in literature as *stance detection*. This latter is a particular declination of the general task of sentiment polarity detection within politics, and it is similar to the aspect-based analysis of sentiments which is commonly applied in consumer reviews in order to detect a particular feature of a review's subject, like cost, battery, or durability of a mobile phone.

As usually when a linguistic content is addressed, it should be also taken into account the variety of languages which have been addressed until now, since the communication of political sentiment is also influenced by the language exploited by users.

Regardless of the motivation for the exploitation of social media in communicating political contents, several studies show that users' behavior is strongly determined by their sociodemographic features, like age, geolocation, or party affiliation. For instance, an inclination to adopt social media communication can be more often detected in younger (politicians, parties' supporters, and public users) and in those with urban constituencies than in older and those with rural constituencies (Jungherr 2016). No clear evidence emerges with regard to the impact of genre.

According to these findings, it is well shared among scholars the opinion that social media users expressing opinions about politics are not representative of the entire population, while they represent a small politically interested and partisan group of it. This is confirmed also by the well-attested fact that only a minority of users posts the most of political messages during elections and debates, while most post only a few.

Several different perspectives can be assumed for describing how social media analysis is applied for monitoring political sentiment. We adopted three main directions: the source from where analyzed textual data are extracted (e.g., Twitter and Facebook) also filtered per author producing the analyzed posts (politicians, social media users, and so on), the specific task or forms of less or more automated analysis to be performed on data (e.g., polarity detection, stance detection, community detection), and the application field. For what concerns the application field, most of the efforts have been addressed by scholars in two main directions, that is, the analysis and prediction of electoral results and the analysis and detection of communication dynamics of political debates, as we will discuss in section "Key Applications."

Key Points

Some main aspect must be taken into account in order to understand what monitoring political sentiment currently means applying more or less computational approaches.

First, each form of analysis is applied on a previously collected dataset extracted from the huge amount of digital traces available on the Internet and other related information. This means that scholars exploit computational techniques for streaming and then filtering (mainly textual) data associated with a variety of metadata that link them with author time and other features that make them unique.

Second, a variety of analyses are applied to the collected dataset in order to classify them in different ways and according to different perspectives, e.g., all the data produced in a specific day or by a particular user, how many posted during an event or related to a topic or time or location, and how many related to a political debate or an electoral campaign.

Finally, different tasks can be performed on data, e.g., polarity detection, stance detection, and community detection, in order to extract a particular kind of knowledge we are interested in. A task usually relies on different forms of information which can be derived from the analysis of texts and of metadata composed and observed together. For instance, having knowledge both about the author and on the text of a message, we can detect its meaning or extract the sentiment that it expresses.

Historical Background

In the last decade, the birth and development of social media have changed our communication strategies and behavior. An especially strong impact social media are bringing about in politics, as observed, e.g., in Lassen et al. (2010), by promoting the diffusion of ideas and cues, but also determining a stronger polarization of views and creating networks among politics and public users that share similar positions and beliefs (Hong and Kim 2016) ("echo chamber" effect), and having not only positive effects on online and offline forms of political participation (Theocharis and Lowe 2016).

As a side effect, the ever-growing number of messages posted on social media platforms has progressively negated the possibility of a single person reading them to gain an overall perspective on the discussed political topics. This motivated the increasing need for more or less automated forms of analysis to be applied on the contents that users generate in social media. They allow us to understand the role played in politics by target entities, to determine how public sentiment is shaped identifying more impacting tweets and authors, and to identify streaming toward opinions and political positions or the main cues involved in a debate.

More traditional analyses based on linguistic features have been usually applied to other forms of political communication; see, e.g., Conoscenti (2011), for a study about the language exploited by Barack Obama, while computational techniques can be exploited for monitoring political sentiment also in texts not generated within social media networks; see, e.g., Balahur et al. (2009), where opinion classification is applied to a corpus of texts on a political topic extracted from an American Congressional debate.

Social Media Analysis for Monitoring Political Sentiment

Datasets and Annotated Corpora

Sources and Authors of Social Media Messages

Data about politics can be extracted in principle from several different sources, according to the topic to be addressed, and also when the focus is mainly on social media, the influence of events happening in the other "traditional" media (like TV and newspapers) must be taken into account. Nevertheless, two main motivations make Twitter the main source for the extraction of data about politics: the widespread use of the microblog in political campaigns and debates by politicians, parties, and public users and the comprehensive access that it allows to the data, offering room for the development of methods of research with digital trace data in general.

A portion of studies presenting analysis of social media is also devoted to monitoring how different groups of users express their political sentiments, limiting the range of the observed data to those generated by specific users. For instance, in Lassen et al. (2010), the authors analyze the exploitation from 2006 of social media by the members of Congress for addressing electoral goals. The high costs of television advertising and the lack of time to be spent in forms of public talks often lead members to use them only during campaign seasons. Communication by social media can be instead used all around the year regardless of place and time at no cost. Moreover, Twitter can reach a meaningful portion of a member's district, not only the declared followers, which can play the role of opinion leaders by conveying what they read on Twitter to a broader audience.

Using Hashtags for Data Collection

For drawing attention to topics that vary from politicians to slogans to be promoted during election campaigns (Bermingham and Smeaton 2011; Mohammad et al. 2015; Sang and Bos 2012), several new hashtags are daily created by users. Hashtags are single words or expressions (with words not separated by spaces) preceded by the symbol "#" well known in Twitter, which allow users to create communities of people interested in the same topic, by making it easier for them to find and share information related to it (Cunha et al. 2011).

The wide diffusion of hashtags in Twitter can vary according to the topic addressed in posts but also to the language and users' community. As shown in the study presented by Gotti et al. (2014) on a parallel English and French corpus from Twitter, on average more than 50% of tweets contain at least one hashtag in both languages, while no less than 8.1% of English tokens and 6.6% of French tokens used in posts are hashtags.

Not all the hashtags generated are destined to be exploited and shared by large people groups. It is indeed known in literature (Gotti et al. 2014) that, regardless of language, hashtags exhibit a Zipfian distribution, i.e., numerous hashtags are used very rarely, and a few are used extensively across tweets. This kind of distribution is confirmed in datasets like the ones studied by Bosco et al. (2016b) and Stranisci et al. (2016) which focus on political debates. In these works the hashtags *#mariagepourtous* and *#labuonascuola* were selected, since they were the more frequently exploited, respectively, by French Twitter users, for making widely known information and opinions about the bill on the homosexual wedding, and by Italian Twitter users, for extending the debate on the reform of Italian education and school sector.

In general, when a user exploits an existing hashtag, he/she wants to be recognized as belonging to the group using it, to be accepted within the dialogical and social context growing around the topic (Chiusaroli 2012), but not necessarily in order to assume the same opinion about the content of the hashtag. For instance, *#mariagepourtous* has been used by people expressing both positive and negative opinions about homosexual wedding in France.

Data selection based on hashtags can be usefully exploited in building corpora for monitoring political sentiments. By selecting hashtags as main filtering criterion, it is possible to easily collect several arguments and different opinions expressed by the persons interested in political debates, circumscribing the collection to posts which are really related to them and, therefore, of interest. The specific sociopolitical topics linked to the hashtags we selected, together with their large diffusion, make them especially adequate for the study of the dynamics of communication in social media and for the comparison with what happens in other media about the topics represented in Twitter by using the hashtag.

Annotated Corpora

Building collections of linguistic data, known as *corpora*, is a widespread practice in computational linguistics, which allows the detection and analysis of various kinds of information. Several corpora have been developed which include data from politics often associated with forms of annotation devoted to make explicit knowledge relevant for monitoring political sentiment too, as shown in the rest of this section.

In Mohammad et al. (2015), a corpus for analyzing electoral tweets is developed by experimenting the usefulness of annotating different layers of fine-grained information such as the sentiment, the emotions, the purpose or intent behind the tweet, and the style of the tweet. They collected English tweets labeled with a set of hashtags pertaining to the 2012 US presidential elections, like, e.g., *#election2012*, *#election*, *#campaign2012*, and *#president2012*. The corpus has been labeled with a multilayer annotation scheme concerning different aspects: sentiment (positive or negative), emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust), purpose (to point out a mistake, to support, to ridicule, etc.), and style (simple statement, sarcasm, hyperbole, understatement). The tweets were annotated by relying on crowdsourcing platforms.

In the context of political debates on reforms, recently a set of sentiment corpora have been developed taking a cross language perspective (Bosco et al. 2016a, b; Stranisci et al. 2016). The set includes Italian, French, and Spanish/Catalan tweets which focus on controversial political reforms highly debated in Italy, France, and Spain, respectively. A homogeneous methodology of analysis and annotation for sentiment analysis and opinion mining has been applied, which collocates the recent trends toward computational semantics-oriented frameworks. It is oriented toward a holistic comprehension of messages and considers all parts of them as intimately interconnected and explicable only by reference to the whole, by involving a global notion of communication, which includes, e.g., context, themes, and dialogical dynamics in order to detect the affective content even when not directly expressed by words, e.g., when the user exploits figurative language. In particular, a multilayered annotation scheme has been proposed for marking the presence of subjectivity, polarity, and irony in tweets. The approach has been applied on texts from different sociopolitical debates, namely, the debate on the homosexual wedding in France, the reform of the education sector in Italy, and the one concerning the independency of Catalonia in Spain, testing the relative independence of the approach from topic and language and preparing the ground for future crosslinguistic comparisons. In all debates analyzed, two different sides can be detected, i.e., pro and con *MariagePourTous* and *LaBuonaScuola*. Moreover, since in this context it is interesting to encode a more fine-grained knowledge, which is related to more specific targets within the debates showing the relationship linking these targets and the opinions and stance about them, an annotation for such aspects has been proposed. Notice that, as a difference with respect to the works reported about sociopolitical debates, or debates' sites, where posts are linked to one another by either rebuttal or agreement within a tree, here the main data source is Twitter, where posts may be only linked by metadata, like author or time. The Italian corpus (Stranisci et al. 2016) extends the available Italian Twitter data in the domain of political communication. Indeed it has been created with an annotation scheme compatible with that one proposed for the sentiment polarity classification (SENTIPOLC) shared task (Basile et al. 2014) for sentiment analysis in Italian (http://di.unito.it/sentipolc14), held during the last edition of the evaluation campaign for Italian NLP tools and resources (Attardi et al. 2015).

Analysis and Tasks

Social media provide a way for expressing opinions about different aspects of the political debate. From this kind of user-generated content, it is possible to discover relevant information under several perspectives.

Sentiment Analysis in the Political Domain

One of the most interesting research areas concerns to investigate how people expose their feelings, evaluations, attitudes, and emotions, and these kinds of aspects are the subject of interest of sentiment analysis. Generally speaking, sentiment analysis can be defined as the task of identifying the subjectivity and the polarity (positive vs. negative semantic orientation) of a text, by exploiting natural language processing and text analysis. In particular, the main tasks associated with sentiment analysis range from a general polarity detection task to more fine-grained challenges, as recently the interest on studying finer-grained and different facets of sentiment in texts is growing, leading to the development of new tasks such as *aspect-based sentiment analysis* (Pontiki et al. 2015) and *stance detection* (Mohammad et al. 2016b), which has been proposed also as shared tasks in the context of international evaluation campaigns such as SemEval (http://en.wikipedia.org/wiki/SemEval).

Most of the works carried on so far in this area focus their analysis on English datasets and rely on the use of lexical and affective resources which are widely available for English (Nissim and Patti 2017). Other languages, such as, for instance, French and Italian, can be still considered among under-resourced languages with respect to the availability of affective lexicons and resources for sentiment analysis. Nevertheless, as the interest in sentiment analysis is increasingly growing, in the last few years many efforts have been devoted to the development of new annotated data to be exploited in this area; see, e.g., Fraisse and Paroubek (2014a, b) for French, while for Italian, let us mention, among the existing resources, the Senti-TUT corpus (Bosco et al. 2013), which has been exploited together with the TWITA corpus (Basile and Nissim 2013) in the sentiment polarity classification shared task (Attardi et al. 2015; Basile et al. 2014). SENTIPOLC's dataset includes tweets with politics as topic.

Detecting Stance in Political Debates

Online debates are a large source of informal and opinion-sharing dialogue on current sociopolitical issues, and several works rely on finer-grained sentiment analysis techniques to analyze politics. Among these works some are dedicated in particular to the classification of users' stance, i.e., the detection of positions pro or con a particular target entity that users assume within debates, applied to data from microblogging platforms such as Twitter, or from websites ranging from forums to other dedicated platforms like convinceme.net. In these studies, the interest is usually focused on dual-sided debates (Somasundaran and Wiebe 2010; Sridhar et al. 2014), where two possible polarizing sides can be taken by participants, and on expressed forms of subjectivity that can be the signal of stance, like, e.g., arguing, since supporting their side, people not only express their sentiment but also argue about what is true. On this line, also the social news website Reddit (http://www.reddit.com) has been recently taken as an object of study (Wallace et al. 2015), since it comprises many, often polarized, user communities – called subreddits – centered around specific topics of interest, which constitute a natural source of data for the analysis of debates on controversial issues.

Stance detection, formalized as the task of identifying the speaker's opinion toward a particular target, has recently attracted the attention of researchers interested in sentiment analysis in Twitter, with a natural focus on the political domain (Mohammad et al. 2016a, b, c). The idea is that, when the real interest is on monitoring sentiment in a specific political debate, stance detection does not only provide useful information for improving the performance of a sentiment analysis system but can also help to better understand the way in which people communicate ideas in order to highlight their point of view toward a particular target entity. Moreover, being able to detect stance in user-generated content can provide useful insights to discover novel information about social network structures. Even if detecting stance in social media could become a helpful tool for various sectors of society, such as journalism, companies, and government, it is evident that it has politics as an especially good application domain. In fact, focusing on stance is particularly interesting when the target entity is a controversial issue, e.g., political reforms (Bosco et al. 2016b; Stranisci et al. 2016), or a polarizing person, e.g., candidates in political elections, and we observe the interaction between polarized communities.

The SemEval-2016 Task 6: Detecting Stance in Tweets (http://alt.qcri.org/semeval2016/task6/) was the first shared task on detecting stance from tweets, explicitly or implicitly expressed. In Mohammad et al. (2016b), the task is described as follows: *Given a tweet text and a target entity (person, organization, movement, policy,* etc.*), automatic natural language systems must determine whether the tweeter is in favor of the target, against the given target, or whether neither inference is likely*. Stance detection is of course related to sentiment analysis, but there are significant differences. In a classical sentiment analysis task, systems have to determine if a piece of text is positive, negative, or neutral. Instead, in stance detection, systems have to determine the favorability *toward a given target entity of interest*, where the target may not be *explicitly* mentioned in the text. Let us introduce an example extracted from the training set of SemEval-2016 Task 6: *Support #independent #BernieSanders because he's not a liar. #POTUS #libcrib #democrats #tlot #republicans #WakeUpAmerica #SemST.* The target of interest here is "Hillary Clinton." The tweeter expresses a positive opinion toward an adversary of the target. Consequently we can infer that the tweeter expresses a negative stance toward the target, i.e., he/she is likely unfavorable toward Hillary Clinton. As can be noticed, this tweet does not contain any explicit clue to find the target. Organizers of SemEval-2016 Task 6 annotated near to 5,000 English tweets for stance toward six commonly known targets in the United States: "atheism," "climate change is a real concern," "feminism movement," "Hillary Clinton," "legalization of abortion," and "Donald Trump." Furthermore, they annotated the dataset also for sentiment, in order to explore the relation between sentiment and stance (Mohammad et al. 2016c). An interactive visualization of the dataset has been also provided (http://www.saifmohammad.com/WebPages/StanceDataset.htm), a useful tool to explore the stance-target combinations present in the annotated dataset and the relations between stance and sentiment. Notice that two of the targets considered in order to evaluate stance detection systems in the shared task were Hillary Clinton and Donald Trump, i.e., the candidates who won the Party Presidential Primaries for the Democratic and Republican parties, respectively, in 2016. Studying these targets is an attracting topic of research due to the impact of the use of social media during the political campaign for the 2016 US presidential elections. Both targets have been the focus also of different research, for instance, in Schumacher and Eskenazi (2016), the authors exploited another data source, studying their speeches during the 2016 political campaign.

Analysis of Contextual Features

Other kinds of analysis which can be interesting in the social media context concern the study of nonlinguistic features of the observed data, in order to detect the pragmatic nature of communicative behavior of users in exploiting subjective and evaluative language for expressing opinions about a political target. The *users' behavior* can be investigated both in the perspective of the single user and in that of the community, by observing how the opinion of users can be relevant within the debate and, therefore, identifying among them the most influential (Lai et al. 2015).

Also data features that are strongly related to the passage of time can shed light on the context of the political debate and can be precious elements for a deeper analysis of the political sentiment. For instance, relevant information can be extracted by observing the frequency of some hashtags, which strongly characterize a political debate, in defined time slots, and the relationships between this frequency and the events happened during the same time slots and spread by other media, like TV, newspapers, and parliamentary debates. When Twitter is the data source, this is allowed by the association with the tweets of metadata showing time and user. As already mentioned hashtags are widely used in order to circumscribe a collection to posts which are relevant for a political debate of interest. Hashtags provide also a way to label and monitor emerging trends, and by assuming this perspective is interesting to observe the *life* of the hashtags: how they propagate among Twitter users, the time when they are newly proposed, and then when they are negotiated and accepted (if they are) by the community and finally not more used. Let us notice that some hashtag is also used without the context of Twitter and assumes the status of a sort of *formula* to be used in language of media also other than social (Krieg-Planque 2009). This phenomenon has been observed in the study described by Bosco et al. (2016b) and Lai et al. (2015) with respect to the hashtag *#mariagepourtous*, which has been used also in newspapers: in the corpus NEWS-MariagePourTous, the linguistic expression "mariage pour tous" appeared in newspaper texts from June 2011, but it is only from March 2012 that also the hashtag can be found in media other than Twitter, showing that *#mariagepourtous* became a linguistic device used by a community that spans beyond Twitter.

Key Applications

Several works in sentiment analysis focused in the last few years on political domain as debated in various ways within the variety of social media.

The analysis of political debates in social media has been often related to election campaigns and became quite popular (Bermingham and Smeaton 2011; Mohammad et al. 2015; Sang and Bos 2012; Tumasjan et al. 2010), since there is a lot of interest in developing tools to automatically gauge the political sentiment in order to predict the election outcome. Also political reforms are often widely debated in social media, and some works focused also on aspects concerning the political polarization (Conover et al. 2011), which in some contexts can also be accompanied by an alarming amount on hate speeches. Others addressed the issues related to the arguments accompanying the political messages, which can be a precious element to be used as actionable knowledge by policymakers.

Tracking Political Elections

Most of the works involving social media data and elections have focused on tasks such as election forecasting. The analyses of political elections reported in literature refer to voting procedures held in a variety of countries in the last few years; see, e.g., Bermingham and Smeaton (2011), Boutet et al. (2012), Chung and Mustafaraj (2011), Gayo-Avello (2012), and Tumasjan et al. (2010). Twitter, in particular, behaves like a pervasive tool in election campaigns, since all the entities involved in politics and related communication, i.e., candidates, parties, journalists, and an increasing portion of the public, are using this microblog to comment on, interact around, and research public reactions to politics. For this motivation Twitter is currently the main source exploited in studies about elections, whose results are surveyed by Jungherr (2016). This study classifies the monitoring of social media based on electoral campaigns in three main investigated areas: the use of Twitter by parties and candidates, the use of Twitter by politically vocal publics, and the use of Twitter during and in reaction to mediated events.

As far as the use of Twitter by candidates and parties is concerned, a higher inclination to adopt social media communication has been detected in politicians of minorities and in opposition with respect to members of governing parties, since they probably feel they have few other ways to communicate their party message to the voting public (Lassen et al. 2010). Nevertheless, the effects of the use of Twitter by politicians or single users are not at present time well understood, also considering that motivations for that use vary from personal communication to promoting party's activities to proposing candidate's website to discuss political arguments.

Considering that several studies confirm that users' behavior is strongly determined by socio-demographic features, like in particular age and location (Jungherr 2016), it is widely shared among scholars the opinion that social media users represent a small younger politically interested and partisan group of population.

By contrast with candidates and parties that sustain their positions with posts in microblogs, public users exploit social media for posting communications in reaction to political events and for debating them often by contrasting opinions expressed by candidates and parties. Nevertheless, users mainly interact with other users of the same political conviction, retweeting their messages and clustering around their same preferred hashtags.

Although research increasingly shows Twitter metrics not to be representative for public opinion at large, for political elites Twitter seems to become an informal barometer of public opinion. Nevertheless, it should be also observed that some scholar criticizes the soundness and reliability of results proposed in research on monitoring political sentiment, in particular in predicting electoral results. The predictive power of Twitter regarding elections has been indeed exaggerated, and the related research problems still lie ahead (Gayo-Avello 2012). These studies are valuable as post hoc analysis, rather than as predictions, and also because there is not a shared standard metric for counting votes and other relevant features, the sentiment analysis techniques are often applied according to a black box strategy, and the demographic features of the involved population are neglected. The development and application of more sound approaches for monitoring political sentiment in electoral contexts are at issue, e.g., in works where beyond sentiment polarity (positive or negative) are taken into account more subtly expressed emotions (e.g., joy, sadness, anger) and also the purpose behind posts (e.g., to support, to ridicule) or the style of tweets (e.g., simple statement, sarcastic, ironic) is detected (Mohammad et al. 2015).

Monitoring Debates on Political Reforms

Political debates, like those developed by public users and politicians around topics impacting on different social aspects of everyday life, are increasingly developed also in social media. Among the recent cases, let us cite, for instance, the debate held in Italy for the reform of education area (Stranisci et al. 2016) or that in France about homosexual wedding (Bosco et al. 2016b), which have been better described in section "Annotated Corpora."

Nevertheless, while Twitter allows a detection of user opinion and sentiment that can be usefully exploited for determining the position of a single user, it makes harder following a debate where a sort of dialogue is developed among different users and the meaning of posts can be only understood within an adequate context of discourse. In practice, the dialogical nature of debates makes harder monitoring political sentiment in these cases using social media where a notion of a discourse thread is missing and can be only reconstructed a posteriori. For this motivation several studies about political debates are based on texts from different sources (Balahur et al. 2009) or the analysis of social media posts is supported by data obtained from these latter.

Political Polarization Analysis

Another interesting application area concerns the possibility to analyze the influence social media have in shaping the networked public sphere and in facilitating the communication between communities with different political orientations. On this line some works focused on aspects concerning the political polarization in Twitter (Conover et al. 2011; Skilters et al. 2011).

Future Directions

Future interesting directions concern the possibility to go toward a deeper analysis of political sentiment, by exploiting fine-grained knowledge which refers not only to the sentiment polarity expressed in the texts and to the target of the sentiment/opinion but also to the aspects of the target that are discussed and to the presence of figurative devices such as irony and sarcasm, which can be possibly used to express negative sentiment in an implicit way. Overall, such fine-grained knowledge will enable a deeper understanding of the political issue of interest, but also, hopefully, the improvement of the performance of systems trained on corpora annotated with such new knowledge for what concerns the detection of political sentiment. Finally, political debates can also be accompanied by an alarming amount on hate speeches. Therefore, also analyzing these aspects will become important in the future in order to manage to understand and contrast this phenomenon.

Aspect-Based Sentiment Analysis in Political Domain

A further interesting matter of future work could be to explore also the stance w.r.t. different aspects of a political target entity. This means to perform a sort of aspect-based sentiment analysis in a political domain, e.g., a tweeter can be in favor of Hillary for aspects related to "health" but not for other aspects. On the same line, another finer-grained feature that could be interesting to explore as future line of research is related to arguments exploited by users in order to support their positions about a controversial issue in the context of a political debate. On this line, it could be interesting to explore the possibility to develop multilayer linguistic resources, where a further layer of annotation related to arguments is added to the sentiment and topic layers, in order to explore possible fruitful relationships among sentiment- and argument-related information.

Dealing with Figurative Language

Since humor, irony, and sarcasm are very prominent in tweets in particular posted during mediated events, the analysis of social media for reliably monitoring political sentiment should be able to deal with this kind of linguistic figurative devices (Bosco et al. 2013). As observed by Ranade et al. (2013), online debates differ from public debates because participants assert their opinion toward either side something ignoring discourse coherence and generally using strong degree of sentiment words including insulting or sarcastic remarks for greater emphasis of their point. The frequent exploitation of figurative language devices in social media and website like Reddit (Wallace et al. 2015), especially in the political domain, is described in several papers, among which are Bosco et al. (2013), Davidov et al. (2011), Maynard and Greenwood (2014), and Reyes et al. (2013), and has been addressed in the SemEval 15–11 shared task on *Sentiment Analysis of Figurative Language in Twitter* (Ghosh et al. 2015). Dealing properly with the presence of ironic devices is indeed crucial when the goal is to analyze the polarity of the opinions on a topic of interest, and in political domains, irony is very often used in conjunction with a seemingly positive statement, to reflect a negative one, due to a phenomenon known in literature as *polarity reversal* (Bosco et al. 2013*)*. Therefore, the issue has to be tackled in order to correctly label the polarity of an opinion, going beyond the literal meaning of the posts.

On this line, it could be also interesting to investigate how to fruitfully combine information about stance and information about the presence of figurative devices in tweets, such as irony and sarcasm, since the use of such devices is very frequent in political debates.

Monitoring Hate Speech Online

The transformative potential of social media systems brings together many challenges. One of the biggest challenges is to maintain a balance between freedom of expression and the defense of human dignity, as such new possibility of expressions opens the way to discourses that are harmful to certain groups of people. This challenge manifests itself at different levels, and online hate speech has been rapidly recognized as one of the serious problem in this context by citizens and politicians in many countries (Silva et al. 2016).

Cross-References

- ▶ Actionable Information in Social Networks, Diffusion of
- ▶ Behavior Analysis in Social Networks
- ▶ Election Forecasting: Scientific Approaches
- ▶ Human Behavior and Social Networks
- ▶ Legislative Prediction with Political and Social Network Analysis
- ▶ Microblog Sentiment Analysis
- ▶ Mining Social Media Data
- ▶ Political Networks
- ▶ Sentiment Analysis in Social Media
- ▶ Sentiment Analysis of Microblogging Data
- ▶ Social Media and Social Networking in Political Campaigns/Movements
- ▶ Social Networks and Politics

References

Attardi G, Basile V, Bosco C, Caselli T, Dell'Orletta F, Montemagni S, Patti V, Simi M, Sprugnoli R (2015) State of the art language technologies for Italian: the EVALITA 2014 perspective. J Intell Artif 9(1):43–61

Balahur A, Kozareva Z, Montoyo A (2009) Determining the polarity and source of opinions expressed in political debates. In: Gelbukh A (ed) Computational linguistics and intelligent text processing. Springer, Berlin, pp 468–480

Basile V, Nissim M (2013) Sentiment analysis on Italian tweets. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2013), Association for Computational Linguistics, Atlanta, Georgia, pp 100–107

Basile V, Bolioli A, Nissim M, Patti V, Rosso P (2014) Overview of the Evalita 2014 SENTIment POLarity classification task. In: Proceedings of EVALITA 2014. Pisa University Press, Pisa, pp 50–57

Bermingham A, Smeaton AF (2011) On using Twitter to monitor political sentiment and predict election results. In: Proceedings of the workshop on sentiment analysis where AI meets Psychology (SAAIP), IJCNLP 2011. Chiang Mai, pp 2–10

Bosco C, Patti V, Bolioli A (2013) Developing corpora for sentiment analysis: the case of irony and Senti–TUT. IEEE Intelligent Systems 28(2):55–63

Bosco C, Lai M, Patti V, Rangel Pardo FM, Rosso P (2016a) Tweeting in the debate about Catalan elections. In: Proceedings of the international workshop on Emotion and Sentiment Analysis at LREC2016, European Language Resources Association, Portoroz, Slovenia, pp 67–70

Bosco C, Lai M, Patti V, Virone D (2016b) Tweeting and being ironic in the debate about a political reform: the French annotated corpus twitter-mariagepourtous. In: Proceedings of the tenth international conference on Language Resources and Evaluation (LREC 2016). ELRA, Portoroz, pp 1619–1626

Boutet A, Kim H, Yoneki E (2012) What's in your tweets? I know who you supported in the UK 2010 general election. In: Proceedings of the international AAAI conference on web and social media, Association for the Advancement of Artificial Intelligence, Dublin, Ireland, pp 411–414

Chiusaroli F (2012) Scritture brevi oggi. tra convenzione e sistema. In: Chiusaroli F, Zanzotto FM (eds) Scritture brevi di oggi. Università Orientale di Napoli, pp 4–44

Chung J, Mustafaraj E (2011) Can collective sentiment expressed on Twitter predict political elections? In: Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, San Francisco, California, pp 1770–1771

Conoscenti M (2011) The reframer: an analysis of Barack Obama's political discourse (2004–2010). Bulzoni, Roma

Conover MD, Ratkiewicz J, Francisco M, Goncalves B, Flammini A, Menczer F (2011) Political polarization on Twitter. In: Proceedings of the fifth international AAAI conference on weblogs and social media, Association for the Advancement of Artificial Intelligence, Barcelona, Spain, pp 89–96

Cunha E, Magno G, Comarela G, Almeida V, Goncalves MA, Benevenuto F (2011) Analyzing the dynamic evolution of hashtags on Twitter: a language-based approach. In: Proceedings of the workshop on language in social media (LSM 2011). Association for Computational Linguistics, Portland, pp 58–65

Davidov D, Tsur O, Rappoport A (2011) Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In: Proceedings of the CONLL'11. Portland, pp 107–116

Fraisse A, Paroubek P (2014a) Toward a unifying model for opinion, sentiment and emotion information extraction. In: Proceedings of the ninth international conference on language resources and evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, pp 3881–3886

Fraisse A, Paroubek P (2014b) Twitter as a comparable corpus to build multilingual affective lexicons. In: Proceedings of the LREC'14 workshop on building and using comparable corpora. European Language Resources Association (ELRA), Reykjavik, pp 17–21

Gayo-Avello D (2012) I wanted to predict elections with Twitter and all i got was this lousy paper. CoRR abs/1204.6441. http://arxiv.org/abs/1204.6441

Ghosh A, Li G, Veale T, Rosso P, Shutova E, Reyes A, Barnden J (2015) Semeval-2015 task 11: sentiment analysis of figurative language in Twitter. In: Proceedings of the international workshop on semantic evaluation (SemEval-2015), co-located with NAACL and *SEM

Gotti F, Langlais P, Farzindar A (2014) Hashtag occurrences, layout and translation: a corpus driven analysis of tweets published by the Canadian government. In: Proceedings of ninth international conference on language resources and evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, pp 2254–2261

Hong S, Kim SH (2016) Political polarization in Twitter: implications for the use of social media in digital governments. Government Information Quarterly, 33(4):777–782

Jungherr A (2016) Twitter use in election campaigns: a systematic literature review. Journal of Information technology and politics 13(1):72–91

Krieg-Planque A (2009) La notion de "formule en analyse" du discours. Cadre théorique et méthodologique. Presses universitaires de Franche-Comté, collection Annales littéraires

Lai M, Bosco C, Patti V, Virone D (2015) Debate on political reforms in Twitter: a hashtag-driven analysis of political polarization. In: IEEE Data Science and Advanced Analytics (DSAA 2015), Institute of Electrical and Electronic Engineers, Paris, France, pp 1–9

Lassen DS, Brown AR, Riding S (2010) Twitter: the electoral connection? Social Science Computer Review 29(4):419–436

Maynard D, Greenwood M (2014) Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In: Proceedings of the ninth international conference on Language Resources and Evaluation (LREC'14). ELRA, Reykjavik, pp 4238–4243

Mohammad SM, Zhu X, Kiritchenko S, Martin J (2015) Sentiment, emotion, purpose, and style in electoral tweets. Information Processing Management 51(4):480–499

Mohammad SM, Kiritchenko S, Sobhani P, Zhu X, Cherry C (2016a) A dataset for detecting stance in tweets. In: Proceedings of the tenth international conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association, Portoroz, Slovenia, pp 3945–3952

Mohammad SM, Kiritchenko S, Sobhani P, Zhu X, Cherry C (2016b) Semeval-2016 task 6: detecting stance in tweets. In: Proceedings of the international workshop on semantic evaluation, SemEval'16. Association for Computational Linguistics, San Diego, California, pp 31–41

Mohammad SM, Sobhani P, Kiritchenko S (2016c) Stance and sentiment in tweets. CoRR abs/1605.01655. http://arxiv.org/abs/1605.01655

Nissim M, Patti V (2017) Semantic aspects in sentiment analysis, Chapter 3. In: Fersini E, Liu B, Messina E, Pozzi F (eds) Sentiment analysis in social networks. Elsevier, Cambridge MA, pp 31–48

Pontiki M, Galanis D, Papageorgiou H, Manandhar SI (2015) Semeval-2015 task 12: aspect based sentiment analysis. In: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015). Association for Computational Linguistics, Denver, Colorado, pp 486–495. http://www.aclweb.org/anthology/S15-2082

Ranade S, Sangal R, Mamidi R (2013) Stance classification in online debates by recognizing users' intentions. In: Proceedings of the SIGDIAL 2013 conference. Association for Computational Linguistics, Metz, pp 61–69

Reyes A, Rosso P, Veale T (2013) A multidimensional approach for detecting irony in Twitter. Language Resources Evaluation 47(1):239–268

Sang ETK, Bos J (2012) Predicting the 2011 Dutch senate election results with Twitter. In: Proceedings of the workshop on semantic analysis in social media. Association for Computational Linguistics, Stroudsburg, pp 53–60

Schumacher E, Eskenazi M (2016) A readability analysis of campaign speeches from the 2016 US presidential campaign. CoRR abs/1603.05739. http://arxiv.org/abs/1603.05739

Silva L, Mondal M, Correa D, Benevenuto F, Weber I (2016) Analyzing the targets of hate in online social media. In: Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016), Association for Advancements on Artificial Intelligence, Cologne, Germany, pp 687–690

Skilters J, Kreile M, Bojars U, Brikse I, Pencis J, Uzule L (2011) The pragmatics of political messages in Twitter communication. In: Garcia-Castro R, Fensel D, Antoniou G (eds) ESWC workshops. Lecture notes in computer science, vol 7117. Springer, Berlin, Germany, pp 100–111

Somasundaran S, Wiebe J (2010) Recognizing stances in ideological on-line debates. In: Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text. Association for Computational Linguistics, Los Angeles, pp 116–124

Sridhar D, Getoor L, Walker M (2014) Collective stance classification of posts in online debate forums. In: Proceedings of the ACL joint workshop on social dynamics and personal attributes in social media. Association for Computational Linguistics, Baltimore, pp 109–117

Stranisci M, Bosco C, Hernandes Farìas DI, Patti V (2016) Annotating sentiment and irony in the online Italian political debate on #labuonascuola. In: Proceedings of the tenth international conference on language resources and evaluation (LREC 2016). European Language Resources Association, Portoroz, Slovenia, pp 2892–2899

Theocharis Y, Lowe W (2016) Does Facebook increase political participation? Evidence from a field experiment. Information, Communication & Society 19 (10):1465–1486

Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with Twitter: what 140 characters reveal about political sentiment. In: Proceedings of the fourth international AAAI conference on weblogs and social media. Association for Advancement on Artificial Intelligence, Washington D.C., pp 178–185

Wallace BC, Choe DK, Charniak E (2015) Sparse, contextually informed models for irony detection: exploiting user communities, entities and sentiment. In: Proceedings of the 53rd annual meeting of the ACL and the 7th international joint conference on natural language processing of the Asian Federation of Natural Language Processing, Association for Computational Linguistics, Beijing, China, pp 1035–1044

Social Media and Social Networking in Political Campaigns/Movements

James Ponder and Paul Haridakis
School of Communication Studies, Kent State University, Kent, OH, USA

Synonyms

Facebook; Internet; MySpace; Online; Social media; Social network; Social networking; Twitter; YouTube

Glossary

Social media	Web-based platforms that allow users to (1) create and maintain a public, private, or semiprivate profile within a given domain; (2) acquire, share, and post content available only to themselves or with other users; and (3) manage a list of connections with other users
Social networking sites	Web-based platforms that allow users to (1) create and maintain a public, private, or semiprivate profile within a given domain; (2) manage a list of connections with other users; and (3) acquire, share, and post information with other users
Political campaigns	An organized effort to influence the decisions and feelings of a given constituency with the ultimate aim of influencing policy or policy making for a government
Political movements	A group of people that work together to influence policy or public sentiment about a given issue/topic or series of issues/topics with the ultimate goal of facilitating legal and policy changes

Definition

Social networking sites and social media are now important parts of political campaigns/movements that are used to organize, orient, and engage volunteers, supporters, and voters.

Introduction

Social networking sites and social media have become standard tools used by political campaigns/movements to organize, orient, and engage volunteers, supporters, and voters. In their relatively brief history in the political arena, they have made a significant and lasting impact on how campaigns and movements function. In this chapter, we examine how political campaigns and movements have used these channels, discuss research regarding the effects of using social networking sites and social media, and discuss how these technologies have changed how political campaigns and movements operate in the twenty-first century.

Key Points

Social networking sites and social media are common tools used in politics. In regard to political information, these tools have been shown to facilitate social interaction, online and offline political activity, and increase awareness of certain issues. However, researchers must still seek out how these new channels influence knowledge and voting decisions.

Historical Background

Social networking has been a mainstay of academic investigation for over 70 years. However, the advent of SNS and social media has changed the way people develop and maintain their social networks. Starting in 2004, with Howard Dean, political campaigns began to leverage these tools for connection to engage constituencies and supporters to influence elections, party platforms, and even the make-up of governing bodies. In the United States, Barack Obama leveraged these tools to raise over $750 billion in funds and organize over 8 million grassroots campaign volunteers in each of his election bids. On a global scale, the Arab Spring was largely successful due to SNS' ability to organize geographically dispersed populations quickly.

More recently, governments have sought to restrain social networking sites and social media by developing laws that limit what types of activity can occur on these sites, while simultaneously seeking to use these sites to identify potential threats. It is apparent that social media and SNS are moving into the spotlight for activists, researchers, and governments.

S

Social Media and Social Networking in Political Campaigns/Movements

Mainstream mass media channels are primary sources of political information. This information typically moved from political elites such as the news media and political parties/candidates to the general population. However, in the twenty-first century, social networking sites (SNS) have altered this dynamic, providing new outlets that allow audience members to spread and share information with professionals and with each other.

People now use their social networking tools to access information, connect with likeminded others, engage in debate (civil and uncivil), watch and upload political videos, push political agendas, engage in social activism, and respond in real time to political situations. Candidates also use social media to reach voters directly, provide raw footage for use by politically interested groups, attack opponents, garner support from their supporters, raise funds, and even name running mates. SNS also provide people with additional avenues for exchanging information. People with access to SNS are not limited to more one-way directional mass communication for political information. They are not limited to face-to-face discussion with those with whom they have strong ties and weak ties. They can blog, tweet, text message, tag videos to share with others, and/or become Facebook friends with similar others.

People have also used SNS to institute political changes in and around the globe, serving as a primary means to connect people and coordinate their efforts in the uprisings and revolutions in the Middle East, Russia, Northern Africa, and South America. With such a vast array of applications and outcomes in the political arena, SNS have become a central component of the political process, serving as venues for candidates to connect with their constituents as well as places for citizens to engage in political activities. In this chapter, we discuss the role of SNS in politics. While the use of social media is a global phenomenon, in this entry, we use examples largely from the United States, where major social networks such as Facebook and Twitter first emerged. The focus will be on three applications of SNS: the use in campaigns, the use by and effects on voters during political campaigns, and the use by citizens to advance social and political change.

Social Networks and Political Campaigns

While the mainstream media have always been primary sources of political information, we have long known that when political information is shared by people within their social networks, it is often more influential than when it comes directly from the media (e.g., Lazarsfeld et al. 1944). For example, during the 1940 US Presidential campaign, Lazarsfeld et al. found that interpersonal discussion with people in their social networks helped voters make up their mind about the candidate for whom to vote. Those interpersonal sources who were particularly influential were referred to as opinion leaders. These opinion leaders were people who tended to obtain information directly from the media and then share it with others in their interpersonal networks (referred to as followers).

Over the years, researchers came to recognize that the flow of communication within political and other social systems was more complex than a simple “two-step” flow. Later, investigations expanded on the complexity of the flow and diffusion of news and political information and the role of the media and interpersonal communication in the diffusion of information and ideas. We have come to understand the media are particularly effective at getting information to people, but real attitudinal and behavioral changes occur when people share it through interpersonal and group networks. The growth of SNS in the twenty-first century have provided new channels for that interpersonal and social connection, thereby functioning not only as a viable venue for attitude and behavior influence, but also as a place where people can access information easily.

Other investigations considering the relationship between membership in social networks and political involvement include arguments that the membership stimulates a collective interest in politics (Schlozman et al. 1995), makes people available to elites for mobilization (Leighley 1996), and helps people learn skills that make participation easier (Schlozman et al. 1995). More recently, scholars have found that social networks are a rapid way to disseminate innovative information and values in a society (Gibson 2001). At the same time, not all social networks are the same. Some social networks are more homogenous than others. In homogenous networks, where people tend to share similar political attitudes and views, political discussion may occur more often. Discussion among group members can help people form and/or reinforce political views and positions and affect political participation (for review of some research, see Cho 2005; Mutz 2002). It is important to consider differences in the types of political discussion that occurs among members of homogeneous and heterogeneous networks, because people do discuss politics with members of political in-groups and with members of political out-groups (Ponder and Haridakis 2015).

Thus, prior research suggests that social interaction can expose people to a different set of politically-relevant information and stimuli than they possess individually which can be used to develop or reinforce pre-existing political attitude. Since individual understanding, information, resources, and ability are inherently limited, this means that social interaction provides people with other opportunities to accumulate resources, such as information, that lower the barriers to political participation. Consequently, participation in social networks supplement (rather than supplant) the person's resources and abilities that make participation likely (McClurg 2003). McClurg (2003) found that social interaction has a twofold influence on likelihood to participate in politics (i.e., vote). First, he found that social interaction in these networks exerts a positive and statistically precise effect on participation, but only when it is politically-relevant. Second, this effect exists even after controlling for membership in organized groups, which indicated that formal and informal social interactions have theoretically distinct effects on involvement. While we do have a rather thorough understanding of nonmediated social networks, we have yet to fully understand how online social networks influence the voting public. So how does the advent of the Internet and the different structures inherent in it influence these social networks?

Online Social Networks

The rise of the Internet and its constitutive parts allowed people to maintain and expand their social ties with others. For example, as early as the 1980s, before the Internet was widely diffused, interactive communities such as political bulletin boards emerged on the Internet. Some, like the California city of Santa Monica's PEN project, were designed for interactive exchange among citizens and between citizens and city officials about issues including political matters.

Later iterations of SNS include a multitude of different platforms such as Facebook, LinkedIn, Twitter, video sharing sites such as YouTube, WhatsApp, Periscope, Telegram, Diaspora, and Snapchat. By the 2012 US Presidential campaigns, all mainstream candidates were using a variety of these platforms. In fact, the ultimate nominees of their respective parties for President (Hillary Clinton and Donald Trump) used Twitter to announce their choices of their Vice-Presidential running mates.

In addition to political candidates, a variety of SNS permit potential voters to select which specific site best fits their own individual wants. For instance, users can form various groups that support particular candidates or issues, seek out political information, engage in online discussions with others about issues or candidates, blog about political issues, share videos (both permanent and nonpermanent videos), and livestream events.

One major debate between scholars has been whether people use online communities, such as those fostered by SNS, to expose themselves to new ideas, evolve their attitudes and beliefs, or reinforce existing attitudes and beliefs (for a more in-depth discussion, see boyd 2008). The answers to such questions probably are largely the result of how individual participants use SNS. SNS can be used by small groups of homogeneous people to inform or coordinate actions. They also can be used for mass communication to reach large groups of people. Regardless of how they are used, SNS provide users the ability to generate content, share content, and serve as a portal to a variety of print and video sources.

Finally, scholars have questioned whether use of SNS and other forms of social media actually net actionable political behavior, such as mobilizing the public, or whether they merely serve to pacify citizens with largely ineffectual forms of political participation (e.g., commenting on message boards that no one reads, sending e-mails that are easily deleted). However, recent research has brought about surprising results that help identify the changing role of online activity.

Social Networking Sites and Political Activity

In the 10+ years that SNS have been used for political purposes, scholars have examined the relationship between online behavior and offline behavior. Researchers have shown that people who use SNS for information about the candidates or to discuss politics are more likely to engage in civic and political activities online such as signing petitions (Abroms and LeFebvre 2009) and donating money (Baumgartner and Morris 2010; Valenzuela et al. 2009). There has also been strong historical evidence that suggests online activities can lead to offline activities such as volunteering for a candidate (Abroms and LeFebvre 2009), talking with others about candidate preferences, and voting (Valenzuela et al. 2009). However, scholars were largely unsure how widespread these behaviors were as most results were largely a result of statistically significant, but small differences between groups. In an effort to identify a causal link between SNS use and political activity, Turcotte et al. (2015) found that people read news stories recommended by their political opinion leaders via social media, return to that news outlet for information at a later time, and trust that outlet for political information.

Several recent studies (e.g., Boulianne 2015; Vaccari et al. 2015) have also lent support for the argument that social media and SNS actually serve to encourage participation in political activities and campaigns. In a meta-analysis, Boulianne (2015) found that some of the differences in existing literature on social media's impact on political participation can be attributed to sample size (smaller samples are less likely than large samples to produce a significant result), sampling methodology (snowball sampling is more likely than random sampling to produce insignificant results), and sampling frames (panel designs are more likely than cross-sectional designs to produce significant results). Overall, the metadata suggested social media use is positively associated with participation in civic and political affairs both online and offline.

While the Internet and SNS can enhance participation, the requisite amount of political participation necessary for a healthy democracy or for political stability is debatable. For example, some have argued that a highly active population acting through targeted social networks can be disruptive to established order and can advance multiple and sometimes conflicting political agendas (Jowett and O'Donnell 2012). Internet and SNS use can allow for greater fragmentation. There also remains a digital divide. Some have argued that the Internet and SNS can widen the gap not just between those who have access or not, but also between those who are more politically active and those who are less active, because the former are the ones most likely to take advantage of these technologies. However, it is important to note that while SNS may have this effect, TV remains, at least in the US, the most used medium for political information (Smith 2011). Thus, the potential of SNS for robust public participation and dialog still has not supplanted the more traditional media structure, though it has complemented it. Recently, Bimber et al. (2014) proposed an interesting twist on the role of political participation and digital media, suggesting that political interest serves as an important mediator between social media use and political participation such as donating toward campaigns, working for a party during a campaign, and discussing politics. This particular position provides opportunities for future researchers to parse out the complex relationship between social media use and ultimately political participation.

Finally, SNS and social media have been tied to increased levels of online political participation. Often called "clicktivism" or "hactivism," this new form of political participation generally refers to online behaviors of a political nature, including, but not limited to, signing online petitions; hacking, or breaking into a computer system, for a politically or socially motivated purpose; spamming politicians; organizing protests; Google bombing; online parody and satire; and even circumventing news blackouts facilitated by governmental overreach of news organizations. However, results on whether these online forms of political activity result in offline political activities/effects is mixed, depending on the social context and nature of the activism.

SNS Use and Attitudinal Effects

Scholars also have examined the relationship between SNS use and its effects on a person's attitudes toward politics. For instance, Vitak et al. (2011) found a positive relationship between a person's levels of political interest and engaging in political activities on Facebook.

Valenzuela et al. (2009) found that SNS use was positively related to higher levels of social trust. Further, Hanson et al. (2011) found a negative relationship between a person's level of political cynicism and SNS use. They suggested that these sites offered their users the ability to interact with others thereby potentially reducing people's levels of cynicism. Further, they proposed that SNSs offered facilitated interaction with like-minded others that could increase understanding of politics and political issues while at the same time increasing feeling of political self-efficacy. However, without further investigation, it is impossible to know if this is a sustainable relationship, an artifact of a new technology, or a reaction to the messages used by the candidates of this particular election.

Once again, these links suggest a relationship between SNS use and attitudinal outcomes. Due to the relatively recent rise of SNS in political campaigns, the causal linkage between SNS use and its effects on attitudes has not yet been examined sufficiently.

Social Networking Sites and Political and Social Change

The ubiquitous nature of the Internet and the growth of satellite communications and mobile technologies such as cell phones have increased connection among citizens on a domestic and global level. The Internet's role in politics has garnered a great deal of inquiry from scholars, the public, social movements, and popular press. Its popularity, when coupled with increasing globalization, has begun yielding interesting results in politics.

More recently, its long-lauded potential for broadening participatory democracy has started to be used by social movements and political change groups. Further, social protest movements and political parties have leveraged SNS' ability to aggregate people (via membership) and messages (via hashtags) to facilitate dialogues about issues affecting people and raise awareness of their cause. For instance, the Black Lives Matter movement has leveraged the ubiquitous nature of cell phones in the United States to raise awareness about police brutality and racial injustice by encouraging citizens to record and/or share stories of injustice, institutionalized racism, and brutality via social media, and SNS to fuel a twenty-first century civil rights movement (Day 2015). In particular, Black Lives Matter uses hashtags, shared posts of SNS, Tumblr posts and videos, and Vines, Instagram, and Periscope feeds to share images about events that were largely ignored by mainstream media. Further, Black Lives Matter leaders like Dee Ray McKesson arm themselves with these stories, images, and videos whenever they are invited to speak to any media outlet/news organization. Finally, Black Lives Matter uses SNS and social media's membership lists and message aggregational features to organize offline protests and marches (Cobb 2016; Stephen 2015).

This use has caused certain governments to take action to curb citizen access and use of the Internet to maintain the status quo. Numerous countries have shut down access to all or parts of the Internet to quell rebellions/unrest (e.g., Myanmar in 2007, China in 2009, Iran in 2009). Google also acquiesced to China's demands to censor content in order to operate there. Other countries such as Russia, Belarus, and the United States have enacted laws that force Internet service providers to allow government officials to monitor

SNS and other Internet use to prosecute citizens for their online activities (Solon 2015).

In recent years, this potential for political action and participation via social media and mobile technologies has been very visible in their use for communication during uprisings and protests around the globe. Facebook, Twitter, and text messaging via mobile phones were used to coordinate political protests and/or uprisings in Ukraine in 2004, and Moldova in 2009. Additionally, mobile devices and social media were used during the "Arab Spring" revolutions that spread across North Africa and the Middle East in 2011 to not only coordinate efforts by protestors, but to also send images to foreign news organizations to bring awareness to political oppression in these countries. While a number of factors led to the downfall of several authoritarian regimes, social media such as Facebook, YouTube, and Twitter provided avenues spread political ideas and foster and advance social and political movements and change. Therefore, while it is hard to argue that revolutions during Arab spring were social media-generated, when much of the planning and orchestration occurred offline, social media did play a role in the political change.

More recently, other political revolutions have leveraged social media to facilitate social and political change in the Middle East. The Islamic State (also called ISIS, ISIL, or Daesh), has used SNS, social media, and phone apps to recruit volunteers, radicalize them, disseminate propaganda, spread news, intimidate populations, and even plan attacks. Indeed, Winter and Bach-Lombardo (2016) discussed how ISIS's social media strategies have shaped the definition of what it means to be a millennial Muslim. Zehra (2015) explained that ISIS uses darknet communities, peer-to-peer networks, and potentially even video game communities to spread information to sympathizers and members without leaving an online trail. Indeed, ISIS's complex social media strategy has somewhere between 46,000 and 70,000 active Twitter accounts, in addition to Tumblr pages, Facebook pages, online magazines, Skype accounts, Weibo accounts, and LinkedIn accounts to name a few. ISIS also employs a Twitter application that enable ISIS turn volunteers' social media accounts into puppet Twitter accounts that tweet out messages supportive of the cause. This creates the perception that ISIS's message is more widespread than it is. Finally, ISIS routinely moves from one SNS and/or social media platform to another whenever their accounts are suspended (which, by some accounts can happen in as quickly as a few days).

In addition to those who challenge mainstream parties or those currently in power, SNS are also used by mainstream parties and campaigns to advance their agendas. At times, political candidates use SNS to go around mainstream media channels to get information from and communicate directly with constituents, supporters, or potential supporters (Jowett and O'Donnell 2012). While traditional Internet sites provide voters with an avenue for getting political information on candidates at their discretion, SNS such as Twitter and Facebook give candidates an opportunity to connect with voters in a way much different than just setting up their own websites to be found when voters want it. SNS permit candidates to become part of voters' social networks and communicate their messages directly to them, making campaigns potentially much more interactive (Gilmore 2011). Gilmore (2011) suggested that such media may help disadvantaged groups and their candidates who are less well established compete with those who are more entrenched and have significantly more resources for campaigning.

Like candidates, citizens and interest groups are using social media in progressive movements such as the environmental movement, United for Change and Occupy Wall Street. In the US, the tea party movement and tea party leaders have used SNS to encourage others to join the movement as well as pressure politicians on certain issues such as government spending and taxation (Barrow 2012). These movements use SNS to disseminate ideas among their members, coordinate movements, and connect with mainstream media outlets. However, when broad-sweeping change does occur, it is usually due to older methods such as marches, pressure from constituents, mainstream media coverage, and rallies. Previously those pushing for societal change had to rely on print media such as flyers, manifestos, and pamphlets. SNS have provided new vibrant forms of mass, group, and interpersonal communication to spread ideas more quickly and efficiently. Clearly, SNS do provide new forms of connection. However, because there is little empirical research on the uses and effects of SNS for political communication specifically, we do not really understand all of their positive and negative effects at this time.

Conclusion

Newer social media, like media before them, are tools for accessing and sharing information. They have only been around for a little more than a decade, but, they are now part of the media landscape. How they fit within that landscape, generally, and their role in politics, specifically is still being explored. They are used in social and political movements, and by candidates, interest groups, campaigns, and advertisers to reach potential voters, sympathizers, and volunteers. They have been used by voters to seek out political information, political entertainment, and to help them make up their minds, and engage in their own political activities. Needless to say, more research is needed on individual differences and desires of users and the social and political contexts in which SNS use occurs and the mechanism by which SNS leads to attitude and behavioral changes. Until more is known about such use and effects, it is difficult to hypothesize or generalize about the effects of SNS use and/or the functions of different platforms for accessing SNS sites.

Understandably, there is still little empirical research demonstrating their effects on voter turnout, candidates selected, and impact on voters. That which does exist has tended to focus on campaigns in more developed nations where social media are widely used and diffused. Many of these nations also have somewhat stable democratic governments. While there have been some initial forays into the sociopolitical factors that influence SNS use and effects, these include the extent of a population's access to SNS, the power structure of the different countries in which the SNS are used, the extent of government control exerted, the homogeneity of the populace using social media, the type of political activity, and a host of other socio-political factors, much more needs to be conducted before any clear understanding arises.

Key Applications

Social media and social networking in political campaigns and movements are standard tools used by political campaigns/movements to organize, orient, and engage volunteers. They have been used, in many cases, to circumvent existing channels for political information – whether the goal is to inform, persuade, or energize, these tools allow citizens the opportunity engage with politics in a relatively safe environment. Further, political campaigns have begun to use these sites to invigorate disenchanted populations and bring them into the political process. Finally, they have been used to raise awareness on issues that traditional mainstream media do not cover.

Future Directions

It has been argued that real activism requires strong ties among people within a social system/network. Groups who use social media for real activism need large numbers of people beyond their more intimate social networks to muster the political will and numbers to effectuate change. Gladwell (2010) argued that SNS really build weak not strong ties and may not foster long-term relationships necessary to govern once change has occurred. However, with so many candidates, political parties, and political and social movements using SNS to further their cause, it appears that, at the very least, SNS do provide opportunities to allow others a seat at the governance table. How these technologies will be used to govern has yet to be determined.

SNS and other Internet sites also can be used by group to hide their true identities making it hard for those who access them to make informed decisions about the credibility of sources. This may make SNS potentially strong tools for spreading misinformation and propaganda. Therefore, future scholars should not only examine the effects of SNS use but also the accuracy and truthfulness of the information presented to SNS users.

SNS use may assist people in engaging in more political activities (e.g., voting, protesting), and becoming more polarized. However, as of yet, it is impossible to determine if SNS are responsible for increased political activity and polarization, or if people who are already politically active and polarized use these sites to reinforce such behavior and attitudes.

Future research also needs to explore the extent of fragmentation in SNS. One of the key aspects of SNS is that they can connect people together in groups. However, in the process of connecting with others in a group, people also may disconnect from those not associated with the group. Therefore, future scholars should examine how this connection and disconnection influences the public. Specifically, does this connection lead to negative view of those not associated with the group? Can it lead to greater in-group bias and out-group derogation?

Perhaps one of the most negative effects of SNS use is selective exposure. Specifically, when a person strongly identifies with a particular political party or movement, and he or she acquires most of his or her information through trusted (and more likely like-minded) political channels, does that have an impact of his or her own perceptions of reality? In particular, is he or she more likely to believe messages from politically similar people or organizations (in-group members) and disbelieve messages from politically dissimilar people or organizations (out-group members)?

Additionally, many of those who use SNS in political campaigns may be more likely to be politically active. However, what is unknown is if this activity is inherently good for democracy. Scholars should examine SNS use and determine if these sites are good for democracy and lead to productive change, if the use of these sites may foster partisanship and polarization and other unproductive changes, and/or if there is a mixture of both productive and unproductive changes as a result of using SNS.

Cross-References

- ▶ Competition Within and Between Communities in Social Networks
- ▶ Connecting Communities
- ▶ Counterterrorism, Social Network Analysis In
- ▶ Crime Prevention, Dataveillance, and the Regulation of Information Communication Technologies
- ▶ Cross-Platform Social Network Analysis
- ▶ E-Government
- ▶ Misinformation in Social Networks: Analyzing Twitter During Crisis Events
- ▶ Political Networks
- ▶ Social Media Analysis for Monitoring Political Sentiment

References

Abroms LC, Lefebvre RC (2009) Obama's wired campaign: lessons for public health communication. J Health Commun 14:415–423

boyd d (2008) Can social network sites enable political action? Int J Media Cult Politics 4:241–244. https://doi.org/10.1386/macp.4.2.241_3

Barrow B (2012, August 13) Tea party evolves, achieves state policy victories. Retrieved from: http://news.yahoo.com/tea-party-evolves-achieves-state-134945533.html

Baumgartner JC, Morris JS (2010) MyFaceTube politics: social networking web sites and political engagement of young adults. Soc Sci Comput Rev 28:24–44. https://doi.org/10.1177/0894439309334325

Bimber B, Cantijoch Cunill M, Copeland L, Gibson R (2014) Time digital media and political participation: the moderating role of political interest. Soc Sci Comput Rev 33:21–42

Boulianne S (2015) Social media use and participation: a meta-analysis of current research. Inf Commun Soc 18:524–538. https://doi.org/10.1080/1369118X.2015.1008542

Cho J (2005) Media, intepersonal discussion and electoral choice. Commun Res 32:295–322. https://doi.org/10.1177/0093650205275382

Cobb J (2016, March 14) The Matter of Black Lives. A new kind of movement found its moment. What will its future be? *The New Yorker*. Retrieved from:

http://www.newyorker.com/magazine/2016/03/14/where-is-black-lives-matter-headed

Day E (2015, July 19) #BlackLivesMatter: the birth of a new civil rights movement. Retrieved from: https://www.theguardian.com/world/2015/jul/19/blacklivesmatter-birth-civil-rights-movement

Gibson JL (2001) Social networks, civil society, and the prospects for consolidating Russia's democratic transition. Am J Polit Sci 45:51–68

Gilmore J (2011) Ditching the pack: digital media in the 2010 Brazilian congressional campaigns. New Media Soc 14:617–633. https://doi.org/10.1177/1461444811422429

Gladwell M (2010, October 4) Small Change: Why the revolution will not be tweeted. *The New Yorker*. Retrieved from: http://www.newyorker.com/reporting/2010/10/04/101004fa_fact_gladwell

Hanson G, Haridakis P, Wagstaff A, Sharma R, Ponder J (2011) The 2008 Presidential campaign: political cynicism in the age of Facebook, MySpace and YouTube. Mass Commun Soc 13:584–507. https://doi.org/10.1080/15205436.2010.513470

Jowett G, O'Donnell V (2012) Propaganda and persuasion, 5th edn. Sage, Thousand Oaks

Lazarsfeld PF, Berelson B, Gaudet H (1944) The people's choice. Columbia University Press, New York

Leighley J (1996) Group membership and the mobilization of political participation. J Polit 58:447–463

McClurg SD (2003) Social networks and political participation: the role of social interaction in explaining political participation. Polit Res Q 56:449–464. https://doi.org/10.1177/106591290305600407

Mutz DC (2002) The consequences of cross-cutting networks for political participation. Am J Polit Sci 46:838–855

Ponder JD, Haridakis PM (2015) Selectively social politics: the differing roles of media use on political discussion. Mass Commun Soc 18:281–302. https://doi.org/10.1080/15205436.2014.940977

Schlozman KL, Verba S, Brady H (1995) Participation's not a paradox: the view from American activists. Br J Polit Sci 25:1–36

Smith A (2011, March 17) The internet and political news sources. Retrieved from Pew Internet & American Life Project website: http://pewinternet.org/Reports/2011/The-Internet-and-Campaign-2010/Section-2.aspx

Solon O (2015, August 31) Russia's fist just clenched around the Internet a little tighter. Retrieved from: http://www.bloomberg.com/news/articles/2015-08-31/russia-internet-law-tests-facebook-google-and-other-foreign-firms

Stephen B (2015, November) Get up, stand up. Wired. Retrieved August 4, 2016 from: http://www.wired.com/2015/10/how-black-lives-matter-uses-social-media-to-fight-the-power/

Turcotte J, York C, Irving J, Scholl RM, Pingree RJ (2015) News recommendations from social media opinion leaders: effects on media trust and information seeking. J Comput-Mediat Commun 20:520–535. https://doi.org/10.1111/jcc4.12127

Vaccari C, Valeriani A, Barbera P, Bonneau R, Jost JT, Nagler J, Tucker JA (2015) Political expression and action on social media: exploring the relationship between lower and higher-threshold political activities among Twitter users in Italy. J Comput-Mediat Commun 20:221–239

Valenzuela S, Park N, Kee KF (2009) Is there social capital in a social network site?: Facebook use and college students' life satisfaction, trust, and participation. J Comput-Mediat Commun 14:875–901. https://doi.org/10.1111/j.1083-6101.2009.01474.x

Vitak J, Zube P, Smock A, Carr C, Ellison N, Lampe C (2011) It's complicated: Facebook users' political participation in the 2008 Election. Cyberpsychol 14:107–114. https://doi.org/10.1089/cyber.2009.0226

Winter C, Bach-Lombardo J (2016, February 13) Why ISIS propaganda works: and why stopping it requires that governments get out of the way. *The Atlantic*. Retrieved from: http://www.theatlantic.com/international/archive/2016/02/isis-propaganda-war/462702/

Zehra R (2015) ISIS and its use of technology in modern day terrorism. Retrieved August 4, 2016 from: http://www.igyaan.in/121469/isis-technology-terrorism/

Recommended Reading

Al-Momani M (2011) The Arab "Youth Quake": implications on democratization and stability. Middle East Law Govern 3:159–170. https://doi.org/10.1163/187633711X591521

boyd d m, Ellison NB (2007) Social network sites: definition, history, and scholarship. J Comput-Mediat Commun 13(1):210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

Bradley T (2008, December 5) Final Fundraising Figure: Obama's $750M. Retrieved from: http://abcnews.go.com/Politics/Vote2008/story?id=6397572&page=1&singlePage=true

Brown J, Broderick AJ, Lee N (2007) Word of mouth communication within online communities: conceptualizing the online social network. J Interact Mark 21(3):2–20. https://doi.org/10.1002/dir.20082

Hindman M (2005) The real lessons of Howard Dean: reflections of the first digital campaign. Perspectives 3:121–128

Kavanaugh AL, Zin TT, Rosson MB, Carroll JM, Schmitz J, Kim BJ (2007) Local groups online: political learning and participation. Comput Supported Coop Work 16:375–395. https://doi.org/10.1007/s10606-006-9029-9

Smith SD (2008, November 9) How many volunteers did Obama have?. Message posted to https://my.barackobama.com/page/community/post/trishaifw/gGxZYv/commentary

Sweetser KD, Kaid LL (2008) Stealth soapboxes: political information efficacy, cynicism and uses of celebrity weblogs among readers. New Media Soc 10:67–91. https://doi.org/10.1177/1461444807085322

Vargas JA (2008, November 20). Obama raised half a billion online. *The Washington Post*. Retrieved from: http://voices.washingtonpost.com/44/2008/11/obama-raised-half-a-billion-on.html

Social Media Mining

► Network Data Collected via Web

Social Media Policy in the Workplace: User Awareness

Mohd Heikal Husin[1] and Jo Hanisch[2]
[1]Service Computing, School of Computer Sciences, Universiti Sains Malaysia, Gelugor, Pulau Pinang, Malaysia
[2]Strategic Information Management Group, School of Computer and Information Science, University of South Australia, Adelaide, SA, Australia

Synonyms

Impacts of policy; Legal issues; Risks; Social media; Training

Glossary

Awareness	Knowledge or perception of a situation or fact
Bureaucratic	Having the characteristics of a bureaucracy or a bureaucrat
CRM	Customer relationship management
Governance	The act of governing and relates to decisions that define expectations or verify performance
Longitudinal research	A research study that involves repeated observations/interviews over a period of time
Policy	There are two main types of policies – public policies and private policies. In this paper, the research focuses on the private policies or organizational policies which are limited on available resources as well as legal coercion

Definition

Organizational policy and social media are two of the most highly discussed topics within organizations today, especially within governments. Policy is typically described as a principle or process to guide decisions in order to achieve rational outcomes or to address evident problems (von Solms and von Solms 2004). The difference between a policy and a procedure is that a policy will contain the "what" and "why," while a procedure contains the "what," "how," "where," and "when" (Colebatch 2006, pp. 313, 317). Policies are generally adopted by a board or senior management body within an organization (Wergin 1976). They guide senior management in making both subjective-based decisions (on the relative merits of factors which are difficult to objectively test, such as work-life balance) and objective-based decisions (operational in nature and easier to objectively test, such as security policy) (Wergin 1976).

Social media are web-based applications that have emerged from outside the organization. They provide an interactive and open approach to collaboration and communication mainly through the use of Web 2.0. Mergel and Schweik (2012) highlight that Web 2.0 derives its power from users for all activities, and this indirectly differentiates Web 2.0 from other standard technologies implemented within organizations, such as CRM or any other management systems. This poses several problems, including the following: (1) Controlling the level of openness that is needed within government organizations. This is due to the lack of an effective regulatory framework or policy within organizations associated with social media. (2) The existence of information leakage as employees might accidentally share confidential information about their work through social media. This information could either damage or threaten the reputation of the department. (3) Lengthy bureaucratic approval processes, especially when it comes to allowing employees to either access new information or when providing information to other users, such as citizens and nongovernmental organizations. While this process provides a security barrier for the organization, it also deters the interaction among employees who are interested in the new information or providing information and feedback to other users. Placing too many restrictions contradicts the rationale for using social media.

There have been some developments in lowering risks of employing social media tools within organizations. Husin and Hanisch (2011a) propose a policy development framework, highlighting the important components within an effective social media policy. Osimo (2008) presents lessons learned, such as enabling authentication policies and partnering with certain Web 2.0 applications instead of centrally implementing all applications, and Tapscott et al. (2007, p. 18) developed steps to manage change for new governance designs that lead to innovative and agile processes within governments through social media.

While the available research provides good arguments in terms of the importance of planning implementation for social media as well as the development of an effective social media policy, the relationship between users' awareness of a social media policy and the adoption rates of social media remains largely unexplored. So the research question that has been developed for this paper is: "How does the awareness of a social media policy influence the use of social media among users in an organization?"

This entry considers why user adoption is important and why a policy is essential for organizations to maintain control over a new technology. The overview of steps that were undertaken for the research is provided, leading to the results of the longitudinal research. The conclusion includes a summary of the research outcomes and the relevance of the research to an organization which intends to develop an effective policy.

Introduction

The successes of a technology implementation within organization are dependent on the users. Rogers (2003, pp. 171, 177) highlights that users could either adopt or reject a technology, and these decisions are based on either a need or an awareness of a technology. Adoption rates are expected to be lower within a working environment where key decisions are made through different parties within a department (Bajwa et al. 2005; Onyechi and Abeysinghe 2009). The complexity of the decision is related to different processes and policies depending upon the organization (Shumarova and Swatman 2008).

There are also a number of reasons why adoption among users could be affected such as low trust levels (Johnston 2007), corporate culture, and the requirement of more training for new technology which is something that users would mainly try to avoid (Husin and Swatman 2010). So in order to limit the barriers to adoption, an effective policy is crucial.

When an organization implements a new technology, the need for an effective organizational policy is essential in order to provide a sense of security for the organization (Husin and Hanisch 2011a). But more often than not, such policies tend to be looked over by users (Althaus et al. 2008). This may be attributed to the generally extensive comprehensiveness of policies (especially in the public sector) and the associated perception of lack of relevance for the user (Althaus et al. 2008). So an appropriate policy development process is essential in ensuring that users have an understanding of the organization's rules and in ensuring that the policy contains important components (Hrdinova et al. 2010; Husin and Hanisch 2011b; Woodford 2005). This also allows the policy to be developed more effectively, while maintaining relevance to the users from the perspective of their work.

For an authoritative-based organization such as in public sector, policies are usually viewed in three ways (Althaus et al. 2008, p. 6):

1. As an authoritative choice. Clearly viewed as the method for government to exercise their power and guarantee results through a series of hierarchical decisions.
2. As a hypothesis. All policies go through an iterative process or "error making" which enhances and changes the policy to be more effective.
3. As an objective of government action. Policies act as a guide for a department to achieve the intended results (Moule and Giavara 1995).

So, it is natural for employees within the public sector to view the authoritative choice as essential while using a social media tool, but this should not be the case as mentioned by Husin and Hanisch (2011a) and Hrdinova et al. (2010) due to the flexibility and openness that the tool promotes for an organization.

This research considers the social media policy from an organizational perspective (refer to Glossary) and aims to identify the influences that a user's awareness of the policy has on their social media usage.

Historical Background

Due to the nature of ethical requirements, the government agency that participated in the research remains anonymous. For the purpose of this paper, it will be referred to as Agency A. Agency A was in the early stages of implementing a standardized social media platform which would be accessible to all their employees for their daily activities. The researcher conducted semiformal interviews as well as quick questionnaires during the platform training session for employees.

As the research was a longitudinal approach over the period of 2010 and 2011, the interview sessions were conducted with ten users from within Agency A with two interviewees continually participating due to their role with the implementation process for the social media platform. The interview included questions concerning their opinions about social media, examples of usage of social media, and their awareness of the social media policy. It should be noted that Agency A consists of a number of internal departments with many of the interviewees spread across different locations.

During the initial implementation process, Agency A conducted training for their employees, involving an hour and a half of "hands-on" time with the social media platform. The participation from the employees was encouraging enough for the training sessions to be held every month since October 2010. The quick questionnaires were circulated at the end of five randomly selected training sessions which brought the total of 81 respondents. The questionnaires asked participants about their level of social media usage within their own departments, tools which they deem useful for the respondents, and their expectations from using social media in their work.

Key Research Result

The results are based on the analysis conducted through the interviews as well as the quick questionnaires. The levels of awareness for the existence of a social media policy among the respondents from the training sessions are high as evident from Table 1.

Table 1 shows the overall level of awareness among the questionnaire respondents at 39.51%, with female respondents at 33.33% and male respondents at 6.17%. A main reason for the stark contrast of gender number is due to a high number of female employees (62% out of the total employees) within Agency A. Nevertheless, the results still indicate that there is awareness for the social media policy among employees along with

Social Media Policy in the Workplace: User Awareness, Table 1 Level of awareness for social media policy

		Awareness of social media policy			Total	Percentage
		N/A	No	Yes		
Gender	F	22	10	27	59	72.84
	M	4	13	5	22	27.16
Total		26	23	32	81	100

Social Media Policy in the Workplace: User Awareness, Table 2 Position versus social media usefulness

		Awareness of social media policy			Total	Total percentage
		N/A	No	Yes		
Position of user	Employee	12	18	19	49	60.49
	Middle M	7	14	8	19	23.45
	N/A	4	1	0	5	6.18
	Other	0	0	2	2	2.47
	Senior M	3	0	3	6	7.41
Total		26	23	32	81	100

a number of employees unaware of the policy (28.39%). The results from the respondents about their positions and whether they found the social media platform useful in their work are shown in Table 2.

The majority of the respondents were on the employee (60.49%), the middle management (23.45%), and senior management (7.41%) level, while the remainder was reluctant to disclose their position. This shows that the interests in using the social media platform in daily work activities are still evident (39.51%) even with the existence of the social media policy. This is a good sign as it shows that higher management supports the use of the social media platform. As some of the feedback from the interview sessions states:

> More higher management should use the platform so it gives a sense to employees, that yes, the platform is an official tool for them to use. (Interviewees 6 and 7)

As Agency A consisted of different departments, the level of social media usage is quite varied. The research found that most of the departments that are using social media are aware of the social media policy. Table 3 shows the different levels of usage and awareness.

The department which was "going ahead" with utilizing social media in their daily activities had a higher awareness of the policy (13.58%) which is followed by the department which is "trying out" the tools (12.35%). But coincidentally, the latter department also had the highest level of unawareness for the social media policy (8.64%). The interesting result was that departments who were "fully using" had the lowest number of social media policy awareness (1.23%) compared to the other levels of usage. In the interview sessions, majority of the participants were aware of the social media policy and have either read or had a quick review of the policy. More than 70% of the participants were still using the social media platform frequently without any indifference to the policy. The participants were using the platform to communicate ideas, comment on nonwork-related information, and even share common interests with their colleagues.

An example of the social media usage was by Interviewee 6 where an employee was looking for

Social Media Policy in the Workplace: User Awareness, Table 3 Level of usage within departments versus awareness

		Awareness of social media policy				
		N/A	No	Yes	Total	Total percentage
Level of social media usage	Fully using	8	2	1	11	13.58
	Going ahead	3	3	11	17	20.99
	I am not sure	4	6	7	17	20.99
	N/A	1	3	0	4	4.94
	Planning	1	1	3	6	7.4
	Trying out	9	7	10	26	32.1
Total		26	23	32	81	100

an available meeting room through the social media platform. As Agency A is spread across different locations with different meeting room sizes; traditionally, the employee would need to either email or contact via telephone the appropriate parties to find an available meeting room. But instead, the employee accessed the social media platform and sent a mass broadcast for assistance via the available micro-blogging tool. Within 30 min, the employee had a reply from another employee located in a different location who had booked a meeting room for the initial employee.

Even with the clear benefits of the social media platform in Agency A, there are a few employees who are using the platform mainly because the tool's usage is mandatory by their department. Even though the employees were quite happy with using the social media platform for any work-related activity, there is not much interest among the employees in any social activity that comes with the social media platform within working hours. From an analysis of the interviews, some participants recalled that they were paid to work and not to socialize during working hours (Interviewees 9 and 10).

> It doesn't matter to me if someone wants to share their interest but I don't see the point of doing so in working hours. (Interviewee 9)

The interesting point is that the participants who mentioned the quotes above are highly interested in the social media policy available in Agency A. In a way, they view the policy as a useful guide for how they should interact on the social media platform.

Key Applications

From the results shown in the previous section, there are different types of users on the platform within the same agency. This has led to some interesting results that not only indicates the various awareness levels of policy but highlights that the usage level within the different departments of Agency A may also influence the users. It is essential for any agency especially Agency A to implement certain approaches that instills a positive trend of utilizing the platform and an acceptable awareness of the existing social media policy in the agency. Some of the approaches may include providing different level of training to the users as well as highlighting the usefulness of the platform among the senior management.

Future Directions

The results show that balance is needed to cater for different users of the social media platform. On one side, there is the socially based user (highly interactive and willing to share information), while the other is the work restrictive-based user (critical only on work-related issues, with no interest in the social side). Both user categories have awareness of the policy but vary in their usage of social media.

The results indicate that influencing factors on the uptake of social media include the level of training, the ability to use social media for work-related activities, and the level of use by senior management, as an example and reassurance to all employees.

Awareness of policies appears varied across the departments in Agency A which is predicted to occur in a large organization. But the variedness of awareness, especially in the departments which were designated as "fully using," was quite surprising as it was expected that awareness would generally be high. Hence, the departments which have made the decision to "go ahead" with social media have employees with higher awareness of the policy than those departments which are "fully using" social media. This is where the effectiveness of the organizational social media policy is needed as well as the dissemination of the policy and its repercussions in practice. As Bridgman and Davis (2003) suggest, there needs to be a bridge between technical expertise and policy domain. Organizational policy, which focuses on social media, needs to be developed with due diligence as employees are dependent on the policy to guide them.

Cross-References

► Legal Implications of Social Networks
► Social Media

Acknowledgments The authors would like to thank Agency A for providing the research with the opportunity to interview and conduct short questionnaires with their employees.

References

Althaus C, Bridgman P, Davis G (2008) The Australian policy handbook, 4th edn. Allen & Unwin, New South Wales

Bajwa DS, Lewis LF, Pervan G, Lai VS (2005) The adoption and use of collaboration information technologies: international comparisons. J Inf Technol 20(5):130–140

Bridgman P, Davis G (2003) What use is a policy cycle? Plenty, if the aim is clear. Aust J Publ Adm 62(3):98–102

Colebatch HK (2006) What work makes policy? Policy Sci 39(4):309–321

Hrdinova J, Helbig N, Peters CS (2010) Designing social media policy for government: eight essential elements. Center for Technology in Government. University of Albany, Albany

Husin MH, Hanisch J (2011a) Social media and organisation policy (SOMEOP): finding the perfect balance. In: The 19th European conference on information systems – ICT and sustainable service development, Helsinki, 9–11 June 2011

Husin MH, Hanisch J (2011b) Utilising the social media and organisation policy (SOMEOP) framework: an example of organisational policy development within a public sector entity. In: The 19th European conference on information systems – ICT and sustainable service development, Helsinki, 9–11 June 2011

Husin MH, Swatman PMC (2010) Removing the barriers to Enterprise 2.0. In: 2010 I.E. international symposium on technology and society, UoW, Wollongong, pp 275–283

Johnston K (2007) Collaborative filtering and e-business: is Enterprise 2.0 one step forward and two steps back? Electron J Knowl Manag 5(4):411–410

Mergel I, Schweik CM (2012) The paradox of the interactive web in the U.S. public sector. Public service, governance and web 2.0 technologies: future trends in social media. IGI Global, Hershey

Moule B, Giavara L (1995) Policies, procedures and standards: an approach to implementation. Inf Manag Comput Secur 3(3):7–16

Onyechi GC, Abeysinghe G (2009) Adoption of web based collaboration tools in the enterprise: challenges and opportunities. In: 2009 international conference on the current trends in information technology (CTIT), Dubai, pp 1–6, 15–16 Dec 2009

Osimo D (2008) Web 2.0 in Government: why and how? Institute for Prospective Technological Studies (IPTS), JRC, European Commission, EUR, vol 23358. Seville, p 57

Rogers E (2003) Diffusion of innovation, Paperback edn. Free Press, a division of Simon and Schuster, New York

Shumarova E, Swatman PA (2008) Informal ecollabora-tion channels: shedding light on 'shadow CIT';. In: 21st bled eConference: overcoming boundaries through multi-channel interaction, Bled, pp 371–394, 15–18 June 2008

Tapscott D, Williams AD, Herman D (2007) Government 2.0: transforming government and governance for the twenty-first century. New Paradigm White Paper

von Solms R, von Solms B (2004) From policies to culture. Comput Secur 23(4):275–279

Wergin JF (1976) The evaluation of organizational policy making: a political model. Rev Educ Res 46(1):75–115

Woodford MD (2005) Central bank communication and policy effectiveness. SSRN eLibrary, Wyoming

Social Media, Definition, and History

Andreas M. Kaplan
Department of Marketing, ESCP Europe Business School, Paris, France

Synonyms

Blog; Content communities; Facebook; Foursquare; Microblog; Second life; Social media; Twitter; User-generated content; Web 2.0; Wikipedia; Word-of-mouth; YouTube

Glossary

Ambient awareness	Awareness created through regular and constant reception and/or exchange of information fragments through social media
MMORPG	Massively multiplayer online role-playing game
Mobile social media	Group of mobile marketing applications that allow the creation and exchange of user-generated content
UCG	User-generated content

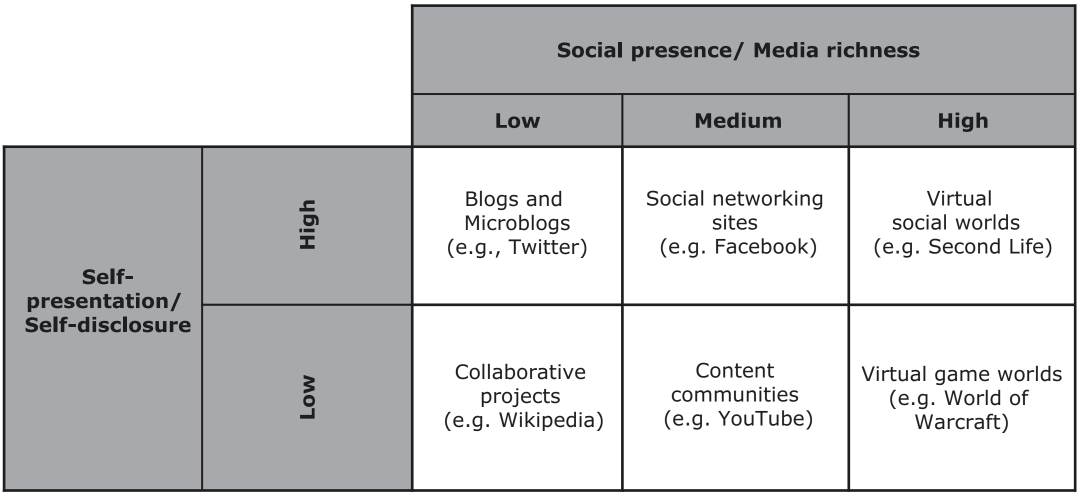
Definition

Social media are defined as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content" (Kaplan and Haenlein 2010, p. 61). Indeed, many have mastered the social media landscape successfully, showing the potential of these applications to yield impressive results. In politics, e.g., social media communications were a key element in Barack Obama's presidential campaign, which led to his first election in 2008. Many states and public administrations make use of Facebook, Twitter, and similar platforms, including the European Union, which aims to create a feeling of European identity among its citizens through social media (Kaplan 2014). Within the entertainment industry, stars such as Britney Spears have built their communication strategies completely around social media (Kaplan and Haenlein 2012b). Even the higher education sector might be close to disruption due to the arrival of digital elements (Kaplan and Haenlein 2016; Pucciarelli and Kaplan 2016), such as massive open online courses (MOOCs) and small private online courses (SPOCs). Social media applications (cf. Fig. 1) – including collaborative projects, microblogs/blogs, content communities, social networking sites, and virtual worlds – have become part of the standard communication repertoire for most corporations and organizations.

Many would probably identify the advent of Facebook, Twitter (Kaplan and Haenlein 2011), and YouTube as the beginning of social media. But, contrary to this belief, the creation and exchange of user-generated content existed long before. The aim of this short essay is to provide a brief sketch of the key developments in social media history, its roots, and its future evolutions.

First Era, 1980s: Arrival of Social Media

The arrival of social media applications coincides with the Internet's first use by private individuals. In fact, a big part of the Internet started as nothing more than so-called newsgroups where individuals could view, discuss, and post bulletin board-like messages to numerous categories. Often these newsgroups were focused on technical issues but they also covered cultural topics such as science fiction or similar. Usenet, established in 1980 by Tom Truscott and Jim Ellis from Duke University, was the most popular discussion system at that time and can be seen as the direct forerunner of the category "Internet forum" which is similar to collaborative projects. These bulletin board systems quickly developed into real discussion groups by allowing individuals to create and exchange user-generated content with each other. Also the first virtual game worlds came up during this era of social media: in 1980, Multi-User Dungeon, the



		Social presence/ Media richness		
		Low	Medium	High
Self-presentation/ Self-disclosure	High	Blogs and Microblogs (e.g., Twitter)	Social networking sites (e.g. Facebook)	Virtual social worlds (e.g. Second Life)
	Low	Collaborative projects (e.g. Wikipedia)	Content communities (e.g. YouTube)	Virtual game worlds (e.g. World of Warcraft)

Social Media, Definition, and History, Fig. 1 Classification of social media (for more details, see Kaplan and Haenlein (2010), p. 62)

first so-called massively multiplayer online role-playing game (MMORPG) and precursor of virtual game worlds (Kaplan and Haenlein 2009) such as the World of Warcraft, was introduced by Roy Trubshaw and Richard Bartle from Essex University.

Second Era, 1990s: Fading of Social Media

During the second era of social media, user-generated content heavily lost in importance due to the fact that more and more companies started to make use of the Internet for their purposes. With industry giants such as Amazon or eBay arriving in 1995 and conquering the web with their corporate websites, the social media applications from the first era seemingly faded away. Despite the fact that social media went by unnoticed by the general public, more and more people started to have their own blogs during the second era and used them to publicly account of their personal lives. While the term "weblog" was introduced by Jorn Barger not before the end of 1997, blogs existed already in the beginning of the 1990s. Its short form "blog," by the way, was coined by Peter Merholz, who jokingly broke the word weblog into the phrase we blog on his own blog in 1999.

Third Era, 2000s: Rising of Social Media

With the dot-com bubble bursting in 2001, social media came back into the game and started to recapture the virtual sphere. Wikipedia (Kaplan and Haenlein 2014) started on January 15, 2001, with the simple sentence: "Hello world. Humor me. Go there and add a little article. It will take all of five or 10 minutes." On February 4, 2004, Marc Zuckerberg launched Facebook, originally located at thefacebook.com, changing it to the current web address not before 2005. Founded on February 14, 2005, YouTube's first video entitled "Me at the zoo" showed cofounder Jawed Karim at the San Diego Zoo and was uploaded on April 23 of the same year. And Twitter, launched on July 15, 2006, started out with its first tweet 4 months earlier on March 21 sent by cofounder Jack Dorsey typing "Just setting up my Twtr." All of these four social media applications lived an enormous success story and today belong to the top ten websites worldwide.

Fourth Era, 2010s: Mobilizing of Social Media

The fourth era of social media is characterized by the arrival of so-called mobile social media

(Kaplan 2012) such as Foursquare, i.e., social media accessed via a mobile device. These new mobile forms have turned computer-based social media, despite their young age, already into traditional social media. Geolocalization and increased time sensitivity are two of the features offered by mobile devices. Both provide mobile social media applications with increased opportunities compared to computer-based ones. For example, with mobile social media, one is aware not only of one's friends' plans but also of their current location and might just go and see them. Ambient awareness, defined as "awareness created through regular and constant reception, and/or exchange of information fragments through social media" (Kaplan 2012, p. 132), is an equally important concept within the area of mobile social media. Since this era just started, it is difficult to say more about its potential evolution for the moment. However, some futuristic, but not impossible, scenarios already arise on the horizon, e.g., facial recognition could make it feasible to take somebody's picture with a cell phone and compare it to social networking sites. A match could give the name and other details about this individual.

This brief sketch of the key developments in social media history showed that these applications started earlier than one would have thought, i.e., in the 1980s. Applications such as Facebook or YouTube can actually be seen as the Internet going "back to the roots" when the power was with the individual users instead of with big companies (Kaplan and Haenlein 2012a). Social media retransformed the Internet to what it was initially intended for – a platform to create and exchange user-generated content.

Future Directions

A potential topic of the future might involve the so-called blockchain technology. What started as a mathematical idea to generate electronic cash without the need for formal institutions (like central banks) and which resulted in the creation of Bitcoin (Nakamoto 2008) can today easily be extended to other forms of information that need to be stored in a secure manner. Such information includes land ownership certificates (especially in countries where no central repository of such information exists), but also possession of precious metals or academic diplomas. The specific way in which blockchains are maintained and modified ensures that information stored in them is essentially free of being manipulated in a fraudulent way. In combination with (mobile) social media, blockchains will allow for new inventions in the digital sphere which we probably cannot even imagine yet.

Cross-References

▶ Facebook's Challenge to the Collection Limitation Principle
▶ Flickr and Twitter Data Analysis
▶ Gaming and Virtual Worlds
▶ Location-Based Social Networks
▶ Virtual Goods in Social Media
▶ Wikipedia Collaborative Networks

References

Kaplan AM (2012) If you love something, let it go mobile: mobile marketing and mobile social media 4×4. Bus Horiz 55(2):129–139

Kaplan AM (2014) European management and European business schools: insights from the history of business schools. Eur Manag J 32(4):529–534

Kaplan AM, Haenlein M (2009) The fairyland of second life: about virtual social worlds and how to use them. Bus Horiz 52(6):563–572

Kaplan AM, Haenlein M (2010) Users of the world, unite! The challenges and opportunities of social media. Bus Horiz 53(1):59–68

Kaplan AM, Haenlein M (2011) The early bird catches the news: nine things you should know about micro-blogging. Bus Horiz 54(2):105–113

Kaplan AM, Haenlein M (2012a) Social media: back to the roots and back to the future. J Syst Inf Technol 14 (2):101–104

Kaplan AM, Haenlein M (2012b) The Britney Spears universe: social media and viral marketing at its best. Bus Horiz 55(1):27–31

Kaplan AM, Haenlein M (2014) Collaborative projects (social media application): about wikipedia, the free encyclopedia. Bus Horiz 57(5):617–626

Kaplan AM, Haenlein M (2016) Higher education and the digital revolution: about MOOCs, SPOCs, social media and the Cookie Monster. Bus Horiz 59 (4):441–450

Nakamoto S (2008) Bitcoin: a peer-to-peer electronic cash system. https://bitcoin.org/bitcoin.pdf

Pucciarelli F, Kaplan AM (2016) Competition and strategy in higher education: managing complexity and uncertainty. Bus Horiz 59(3):11–320

Social Navigation

▶ Social Web Search

Social Network

▶ Human Behavior and Social Networks
▶ NodeXL: Simple Network Analysis for Social Media
▶ Social Media and Social Networking in Political Campaigns/Movements
▶ Social Phishing
▶ Social Provenance
▶ User Behavior in Online Social Networks: Influencing Factors

Social Network Actors

▶ Network Actors Within Entrepreneurial Networks: The Current State of Research

Social Network Analysis

▶ Guess
▶ NetMiner
▶ Pajek and PajekXXL
▶ Social Recommendation in Dynamic Networks
▶ Topology of Online Social Networks
▶ UCINET

Social Network Analysis and Organizational Multimodal Representation

Magdalena Bielenia-Grajewska
Intercultural Communication and Neurolinguistics Laboratory, Department of Translation Studies, Faculty of Languages, Institute of English, University of Gdansk, Gdansk, Poland

Synonyms

Company multimodal identity; Corporate linguistic image; Corporate linguistic personae; Lattices and ties; Social grids

Glossary

Organizational multimodal representation	A type of organizational image created and maintained by using linguistic and nonlinguistic tools
Social network analysis (SNA)	An approach used to study the characteristic features of social networks

Definition

The concept of *organizational multimodal representation* has been created by the author to denote all linguistic and nonlinguistic elements that shape the way organizations view themselves and are perceived by the broadly understood stakeholders. In a previous contribution devoted to the topic of identity (Bielenia-Grajewska 2014), the author focused on the linguistic dimension of organizational representation. Taking into account the growing importance of different senses in creating and sustaining organizational representation, the verbal side has been enriched with other modes of encoding and decoding identity.

In the literature, different terms are used to describe how companies are perceived by workers and stakeholders. The term *company identity* is applied to denote the meanings covered by corporate identity and organizational identity (Bielenia-Grajewska 2014). Taking the target audience into consideration, corporate identity represents how stakeholders look at the company, whereas organizational identity is related to employees' views on their work (Cornelissen 2008). Analyzing the process of identity creation, corporate identity is related to the role of managers in effective communication, whereas organizational identity is observed during informal interactions (Rughase 2006).

The proposed term *company linguistic identity* aims to encompass different tools used by various organizational stakeholders in diversified settings. As discussed by Bielenia-Grajewska (2014), company linguistic identity can be observed by applying social network analysis. This perspective highlights that the linguistic representation is formed by social networks. It also allows one to study the way that social networks determine and sustain linguistic tools in organizations.

The proposed new concept of *organizational multimodal representation* goes beyond the linguistic dimension of legal entities, drawing one's attention to other senses and tools used in creating and enriching the identities of organizations. Social network analysis can be briefly defined as an approach used to study the characteristic features of social networks. Notions such as nodes, ties, links, and edges are studied to determine the types and characteristics of social networks. Yang et al. (2017) stressed that network data encompass at least two datasets: a regular dataset (nodelist) where nodes are the elements of analysis and a dataset that defines the relationship between units (often called an adjacency matrix or edgelist, depending on the characteristics and the presentence of rows, columns, and ties).

Introduction

Social network analysis (SNA) may be applied in different ways. Crossley et al. (2015) stressed that there are three approaches to SNA: ego-net analysis, whole network analysis, and two-mode analysis. An ego-net is the network created around a given social network. This network encompasses the other actors to which an ego has relations and ties. In organizational settings, this may involve information sharing and economic exchange. Whole network analysis is used to study large communities or populations, whereas two-mode analysis focuses on the networks existing between two types of nodes.

SNA can be used to study organizational multimodal representation in different ways. In a previous contribution on social networks and identities, Bielenia-Grajewska (2014) stressed that SNA investigates process-related notions, such as communication, innovation, and knowledge transfer. This processual approach facilitates the study of identity that is not fixed. Another important determinant is the technology that influences social networks (Travica 1999). Technology is crucial for networks from both internal and external perspectives. A direct understanding is connected to the technological parts within human bodies (e.g., artificial limbs). General and metaphorical perspectives are connected to the way technology determines different networks inside a human being, related to biology, emotion, feelings, and social interactions (Michael and Michael 2008). As Bielenia-Grajewska (2014) discussed, technology is also important in shaping linguistic networks, offering tools and channels to communicate with those who are physically far away from interlocutors.

Key Points

Social network analysis can be used to observe organizational multimodal representation at both individual and social levels. SNA can be used to study the way that social networks create and sustain multisensual organizational communication. In that way, it is possible to observe how a given way of communication (e.g., the linguistic one) is encoded and decoded by social networks. This occurs similarly in the analysis of nonverbal communication by observing how gestures are used in communities. Apart from this centric focus, SNA offers a holistic approach to a studied phenomenon, showing how multimodal representation results in a third form with distinctive features that do not result directly from the substituting elements.

In addition, SNA analysis can be conducted at both individual and social levels. As Bielenia-Grajewska (2010, 2014) highlighted, the personal dimension can be studied by examining ethnic, national, and professional identities and their influence on the linguistic performance of an individual. The organizational level can be observed by studying interactions at micro (worker), meso (company), and macro (country/continent) linguistic levels. These domains can be investigated from an individual perspective, by observing the networks that create them and paying attention to network dynamics.

Historical Background

Research devoted to social networks can be traced back to the nineteenth century and the papers of Auguste Comte, Ferdinand Tönnies, Norbert Elias, Emile Durkheim, and Gustave Le Bon. George Simmel studied how networks shape the lives of human beings. Jacob Moreno and Helen Jennings worked on a technique called *sociometry*, which provided a visual representation of individuals and their contacts (Yang et al. 2017).

From the 1930s to 1970s, scientists used metaphors such as “fabric” or “web” to denote the interweaving and interlocking characteristics of social relationships. In the 1970s, researchers started to use technical terms and specialist applications that led to the creation of SNA concepts (Scott 2017). In the 1990s, many publications, tools, and professional associations related to networks appeared in the market (Gamper and Reschke 2010). At the same time, an interest in discourse, identity, and SNA could be observed (Mische 2011). As Bielenia-Grajewska (2014) stressed, the popularity of discursive approaches in organizational studies is linked to postmodernist perspectives, focusing on the role of changeability and fluidity in organizational interactions.

Proposed Solution and Methodology

Organizational representation can be studied through the prism of SNA by focusing on aspects such as the properties of ties and contact characteristics (Ferligoj and Hlebec 1999). As Bielenia-Grajewska (2014) highlighted, the different properties of linguistic interactions can be studied in organizational settings. For example, the length of communicative acts and their frequencies can be observed to determine how these notions affect organizational linguistic performance. Organizational relationships and their influence on organizational personae can be investigated by focusing on strong and weak ties and their roles in communication and knowledge flows. Another notion that can be studied is the border in organizational settings. As Fombrun (1982) discussed, professions or positions on an organizational ladder can serve as organizational frontiers. Another option is to focus on the individual feelings of workers and how they perceive themselves within organizations.

Different methods can be used to explore the notion of multimodal organizational representation. Bielenia-Grajewska (2014) stressed that corporate linguistic identity can be examined by taking into account individual and societal factors. The personal dimension encompasses aspects such as one’s opinion on organizational culture, factors connected with the work environment, and benefits linked with using a given language. The engagement of workers in linguistic networks is connected with the benefits one may have in a given network. The social dimension encapsulates such ideas as corporate language policy, corporate communication, and corporate hierarchy, which determine network formation and network selection.

Moreover, company linguistic identity may be observed through the prism of the dual relationship between workers and organizations. First, an organization with its own lingo shapes the linguistic performance of its workers. In each community, including professional ones, words and phrases (characteristic of individual performance and organizational culture) may be acquired by new employees. Their function is mainly to make communication more efficient and strengthen organizational bonds. Secondly, workers and the ways they speak shape the company linguistic identity because they add new expressions and phrases to the corporate linguistic repertoire. This can be supplemented with perspectives such as micro (individual), meso (company), and macro (environment) linguistic network levels, studied as independent systems as well as the sets of interactions among them (Bielenia-Grajewska 2010, 2013a). Another way is to focus on the role of linguistic tools in shaping representation.

Bielenia-Grajewska (2015b) discussed company linguistic identity through the prism of metaphors. Their role is studied by applying the 3 Ps model of company linguistic identity and its metaphorical dimension. This model facilitates the understanding of the role of symbolic language in organizational facets such as personnel, products, and purchasers. For example, in the case of organizational multimodal representation, the proposed discussion can be enriched with the study of pictorial metaphors in organizational settings. This perspective encompasses the study of the symbolic nature of pictures and drawings present in organizational settings. In addition, researchers may investigate the role of scents and odors in the way organizations are viewed by stakeholders. As is commonly known, some companies use different types of aromas associated with festivities to stimulate purchasing behaviors. Some companies opt to use the same scent in every shop of their chain to evoke the same connotations among customers, no matter where the shop is situated.

Another option is to apply the perspective of homogeneous and heterogeneous networks. Homogeneous linguistic networks are formed among users sharing the same languages (national or professional). These networks lack a linguistic barrier because communicators share the same linguistic background. Heterogeneous networks, on the other hand, include people with a variety of linguistic backgrounds (Bielenia-Grajewska 2012). The linguistic side of organizational representation can be enriched with studies on other dimensions of organizations. It should be stated that communication is not restricted to its linguistic interactions, but also includes the nonverbal side. As Bielenia-Grajewska (2013b) discussed, office artifacts such as professional clothing, business cards, and corporate architecture are crucial in organizational representation.

Interest in multimodality can be traced to different contributions representing different domains. The multimodal perspective reflects the approach advocated by social semiotics, for example, with different types of semiotic resources analyzed in research. With a focus on the concept of organizational representation, multimodal identity has been studied in the literature from different perspectives. Anzellotti and Caramazza (2017) described how visual and auditory information can be used to recognize one's identity. They focused on different brain parts and their roles in observing and understanding visual or auditory stimuli as well as multimodal messages. Zamparini and Lurati (2016) stressed the role of multimodal metaphors in shaping company identity and discussed how distinctive organizational identities in the winery sector were formed by verbal and visual elements. Staying within the same sector of the industry, Foroni et al. (2017) investigated whether individuals can detect different terroirs in wine by researching wine olfactory discrimination ability.

Multimodality as a concept can be denoted by different terminology. Remley (2017) drew attention to the differences in terms depending on the discipline. For example, neurobiologists use the term *multisensory integration.* Bielenia-Grajewska (2017) also highlighted that approaches such as synesthetic prisms or hybrid perspectives offer complex discussion on different types of organizational representation. In addition, the multidimensional character of communication is visible in such concepts as heteroglossia and synechism. The former, introduced by Bachtin, gave rise to multiple terms that also explain identity, such as the *heteroglossic linguistic identity of companies* (Bielenia-Grajewska 2013c). The heteroglossic linguistic identity of companies is represented in, among others, the multilingualism of employees, language policies of companies, and specialized jargon.

Organizational multimodal representation can be analyzed in different ways by taking SNA into account. One way is to apply methodologies and tools used in organizational discourse analysis (Fairhurst and Uhl-Bien 2012) or critical discourse analysis. Moreover, Soleymani et al. (2017) discussed the application of sentiment analysis in the study of opinions on products, brands, and organizations. The growing role of social media in the life of individuals and organizations has led to the popularity of multimodal sentiment analysis, focusing on written, facial, and vocal elements that can be analyzed. They also stressed that the advancements in different fields of technology and novel ways of communication result in new methodological methods. For example, sentiment analysis can be used to investigate human–human and human–ECA interactions by studying avatars and virtual agents as well as different types of interactions. Jewitt (2016) focused on the concept of multimodal ensembles, represented as different modes engaged in a communicative event, with all modes contributing to the complete meaning of a message. Thus, the approach can focus on different senses at the same time, engaging different stimuli and reactions simultaneously.

Key Applications

The social network method can be used to study corporate relations from different perspectives. One can examine the diversified networks underlying corporate communication and how these networks shape organizational representation. SNA can be used to investigate how different senses influence communication in organizations. As Bielenia-Grajewska (2017) discussed, SNA stresses the role of internal and external entities in creating and sustaining social networks, focusing on people or organizations and the relationships between them. For informal communication, SNA may be applied to observe the circulation of gossip or rumors, as well as the role of nodes and ties in exchanging organizational data in an informal way. Thus, SNA may help one to understand the main determinants of the position and role of informal interactions in a given organization. The next factor for the popularity of SNA in studies on organizational communication is the growing role of technology and the formation of organizational virtual communities.

SNA also can be used along with different network methods, such as the actor-network theory (ANT). This theory focuses on the role of human and nonhuman entities in interactions, highlighting that not only human beings shape the way organizations perform. For example, Bielenia-Grajewska (2015a) emphasized the role of ANT in studying informal communication by examining human and nonhuman entities in organizational interactions. Nonhuman elements determine the creation and development of social bonds. For example, the type of furniture as well as desk and chair arrangement may stimulate or hinder informal interactions. The way an office is planned and organized into smaller spaces can also provide information on organizational hierarchy and corporate borders. Technology, which is visible mainly as office computers and telephones, also creates modern interactions among the broadly understood stakeholders and is responsible for the speed of information sharing (in both positive and negative ways) among workers and customers.

Future Directions

In the future, the network approach is expected to be even more visible in studies on the linguistic performance of modern organizations because of the growing interconnectedness of companies and their dependence on other organizations such as cooperatives, suppliers, customers, and competitors. Moreover, rapid developments in the sphere of neuroscience should lead to the application of neuroscientific tools in different areas, giving rise to new domains such as neuromanagement, international neurobusiness, international neurostrategy, neuroentrepreneurship, neuroethics, and neuromarketing (Bielenia-Grajewska 2013d). Thus, future research may focus more on the interrelationship between multimodality and neuroscience. Recent studies on multimodal communication have focused on the role of neuroscience in examining persuasive messages (Remley 2017). Thus, the co-existence of different domains and technology is expected to increase in future research on multimodal representation.

Cross-References

- ▶ Collection and Analysis of Relational Data in Organizational and Market Settings
- ▶ Entrepreneurial Networks
- ▶ Interorganizational Networks
- ▶ Learning Networks
- ▶ Managerial Networking
- ▶ Online Identities and Social Networks in Organizational Settings
- ▶ Social Network Analysis in Organizational Structures Evaluation

References

Anzellotti S, Caramazza A (2017) Multimodal representations of person identity individuated with fMRI. Cortex 89:85–97

Bielenia-Grajewska M (2010) The linguistic dimension of expatriatism-hybrid environment, hybrid linguistic identity. Eur J Cross Cult Compet Manage 1(2/3): 212–231

Bielenia-Grajewska M (2012) Linguistic aspects of informal learning in corporate online social networks. In: Dennen VP, Myers JB (eds) Virtual professional development. IGI Publishing, Hershey, pp 93–112

Bielenia-Grajewska M (2013a) Corporate linguistic rights through the prism of company linguistic identity capital. In: Akrivopoulou C, Garipidis N (eds) Human rights and risks in the digital era: globalization and the effects of information technologies. IGI Publishing, Hershey, pp 275–290

Bielenia-Grajewska M (2013b) Office artifacts. In: Smith V (ed) Sociology of work: an encyclopedia. SAGE, Thousand Oaks, pp 645–646

Bielenia-Grajewska M (2013c) The heteroglossic linguistic identity of modern companies. Manage Bus Adm Cen Eur 21(4(123)):120–131

Bielenia-Grajewska M (2013d) International neuromanagement: deconstructing international management education with neuroscience. In: Tsang D, Kazeroony HH, Ellis G (eds) The Routledge companion to international management education. Routledge, Abingdon, pp 358–373

Bielenia-Grajewska M (2014) Social network analysis and company linguistic identity. In: Alhajj R, Rokne J (eds) Privacy and security issues in social network. Encyclopedia of social network analysis and mining. Springer, New York, pp 1827–1830

Bielenia-Grajewska M (2015a) The role of gossip in the creation and miscreation of company identity from the perspective of actor-network-theory. Int J Actor Netw Theory 7(4):41–57

Bielenia-Grajewska M (2015b) Company linguistic identity and its metaphorical dimensions. Purchasers, personnel and products through the perspective of metaphors. In: Holden N, Michailova S, Tietze S (eds) Routledge companion to cross-cultural management. Routledge, Abingdon, pp 170–179

Bielenia-Grajewska M (2017) Informal communication. In: Scott CR, Lewis L (eds) The international encyclopedia of organizational communication. Wiley, Chichester, pp 1223–1239

Cornelissen JP (2008) Metaphor. In: Thorpe R, Holt R (eds) *The SAGE Dictionary of Qualitative Management Research*. SAGE Publications Ltd, London, pp 128–129

Crossley N, Bellotti E, Edwards G, Everett MG, Koskinen J, Tranmer M (2015) Social network analysis for ego-nets: social network analysis for actor-centred networks. SAGE, Thousand Oaks

Fairhurst GT, Uhl-Bien M (2012) Organizational discourse analysis (ODA): examining leadership as a relational process. Leadersh Q 23(6):1043–1062

Ferligoj A, Hlebec V (1999) Evaluation of social network measurement. Soc Networks 21(2):111–130

Fombrun CJ (1982) Strategies for network research in organizations. Acad Manag Rev 7(2):280–291

Foroni F, Vignando M, Aiello M, Parma V, Paoletti MG, Squartini A, Rumiati RI (2017) The smell of terroir! Olfactory discrimination between wines of different grape variety and different terroir. Food Qual Prefer 58:18–23

Gamper M, Reschke L (2010) Soziale Netzwerkanalyse. Eine Interdisziplinäre Erfolgsgeschichte. In: Gamper M, Reschke L (eds) Knoten und Kanten: Soziale Netzwerkanalyse in Wirtschafts- und Migrationsforschung. Transcript Verlag, Bielefeld, pp 13–54

Jewitt C (2016) Multimodal Analysis. In: Georgakopoulou A, Spilioti T (eds) The Routledge handbook of language and digital communication. Routledge, Abingdon, pp 69–84

Michael K, Michael MG (2008) Homo Electricus and the continued speciation of humans. In: Quigley M (ed) Encyclopedia of information ethics and security. IGI Global, Hershey, pp 312–318

Mische A (2011) Relational sociology, culture and agency. In: Scott J, Carrington PJ (eds) The Sage handbook of social network analysis. SAGE, Thousand Oaks, pp 80–98

Remley D (2017) The neuroscience of multimodal persuasive messages: persuading the brain. Routledge, New York

Rughase O (2006) Identity and strategy: how individual visions enable the design of a market strategy that works. Edward Edgar Publishing, Cheltenham

Scott J (2017) Social network analysis. SAGE, London

Soleymani M, Garcia D, Jou B, Schuller B, Chang SF, Pantic M (2017) A survey of multimodal sentiment analysis. Image Vis Comput 65:3–14

Travica B (1999) New organizational designs: information aspects. Ablex Publishing Company, Stamford

Yang S, Keller FB, Zheng L (2017) Social network analysis: methods and examples. SAGE, Thousand Oaks

Zamparini A, Lurati F (2016) Being different and being the same: multimodal image projection strategies for a legitimate distinctive identity. Strateg Organ 15 (1):6–39

Social Network Analysis and Recommendation

► Location-Based Social Network Analysis

Social Network Analysis in an Age of Digital Information

Mariann Hardey
Institute for Advanced Research in Computing (iARC), University of Durham, Durham, UK

Synonyms

Analysis; Consumer; Digital; Linkages; Networks; Relationships; Social anthropology

Glossary

Consumer	An individual or organization that uses a commodity or service
Digital data	Digitally sourced and/or published statistics or items of information
Research	Systematic inquiry or investigation
Social network	Facilitates communication, provided by a network of related linkages

Definition

The term "social network analysis" (SNA) provides an increasingly overarching research context for scholars, as well as policy makers, industry, commercial organizations, and the public sector. It refers not only to an integrated set of theoretical concepts and analytic methods but intends to explain a whole set of relationships and their variations as these are defined by "a specific set of linkages" and their "additional property" (Mitchell 1969, p. 2).

Introduction

This chapter reports on the description and development of SNA in an age of digital information. The intention is to succinctly capture for this ESNAM, second edition, the most recent formulation in business and media theory.

Key Points

SNA is used to extract network features and concerns three main areas: first, the relationships between individuals and groups, second to examine how those relationships arise and are sustained, and third to understand the consequences of the relationships. In doing so, SNA uses these factors to predict, for example, economic success, political instability, access to new opportunities, levels of depression, spread of disease, group dynamics, and concentration of power, along with health and life longevity.

Historical Background

The concept of the structural relationship and measurement of "social network analysis" have its roots in the now classic community studies, class relationships, and social anthropology (Banton 1966; Fortes 1949, 1953; Bell and Newby 1971) and predate the digital age. Among the first to discern mathematical patterns was the work of Bavelas (1948) who helped to develop a mathematical model of group structure of human relationships. The coming of the digital information age provides a powerful new model for social structures and potential for change in the study and analysis of individual and community structures and the interlocking relationships. Most recently, the design of social network visualization has helped to create exciting representations of data to capture social phenomenon. To date, one of the best examples is a 3-D *visualization* of carbon gas in the atmosphere using data from NASA's Orbiting Carbon Observatory-2 satellite (Fleur, reporting for the New York Times, 2016). Research on whether search or social media can predict an outcome has motivated researchers to seek to "improve the value of data" (Lazer et al. 2014, p. 1203). In addition, the value of big data is often held up as an exemplary use of social network analysis offering novel statistical approaches to concept-level interactive influence effects. Social scientists have traditionally been interested in not only the analysis of social data but critical of its collection. Indeed, social network analysis has undergone a profound change at levels of epistemological enquiry, ethical processes, the constitution of knowledge, and research engagement, along with the categorization and nature of understanding the digital age. Just as Beer and Burrows note, such perspective stakes out new ways of knowing, and as researchers we might (cautiously) celebrate a new social life of data.

SNA and Ubiquitous Connectivity

One outcome of the burgeoning territory of digital and continuous development of ubiquitous connectivity has meant that new informational infrastructures and data production provide a rich and diverse series of network contexts that have significant implications for scholars and digital researchers. Recognized as one of the earliest social network researchers, Barnes's work asks a pertinent question that remains as relevant today: when collecting data, on social relations that may not hold any obvious limits, where does the researcher set their boundaries? (1979, p. 414). The study of mediated social networks and increasingly digital spaces has a well-established concentration as a field of study in sociology (cf. Wellman 1983; Rainie et al. 2012). However, the capture of such data deserves further scrutiny that calls into question the acquisition and observation of such linkages that do not remain fixed into place and are these equally accessible to all observers. The issue of technology, and especially digital and social media platforms, brings social network analysis into contact with debates about the nature of its technology, heritage, culture, and processes. These are debates that are particularly challenging in terms of the observations that we may conduct in society and analysis that any researcher would seek to adopt. Indeed the persuasive actor-network theory (ANT) provides an additional explanation of networks that are beyond the agency of the social actors alone (see Callon 1987; Latour 1996), where networks themselves become additional vehicles for social analysis and evaluation. A key question to raise here is to ask; what new possibilities (and limitations) come from this knowledge about the social domain? Social network analysis should enable researchers to quantify and analyze many more social spaces. However, as boyd and others caution, such data are still subjective, and claims to objectivity should be viewed with suspicion: "[. . .] there is a tendency to claim [work] as the business of facts and not interpretation (boyd and Crawford 2012, p. 10). The method of extraction, control, storage and analysis are clearly impactful on the yield and validity of social network analysis. For example, the information harvested from individuals from social media that is fed into commercial operations and systems of analytic sorting and predictive data may be framed as "[. . .] raw streams of observation to be gathered and then processes and systematized" (Cohen 2015, p. 16). The raw data has a high marketing value in terms of how commercial companies buy and sell on consumer data. For the researcher, this requires that they pay attention to the specificities of technologies and how they are varied and composed of new networked spaces. The objectives of SNA include attention on the consequences of digital data and devices and the effect on the knowledge practices, methods, and analytical procedures for academic researchers. For example, data extracted from a social network site (SNS) requires exploration of the quality of data, which are likely to share qualities with those of longer-standing methods such as survey data, and also differences due to the materialities, capacities, and mediating productivity of the technologies. When we refer to SNA, this means the mapping of specific networks that connect social and other relations along with the methods used to assemble the data and to analyze this knowledge.

SNA has undergone two major conceptualizations. The first is established by the social anthropologist Radcliffe-Brown and the procedures he has for describing the structures of society (cf. Oliver 1958). This is essentially an operational derived system since it is couched in terms of the practice and measurement of the interlocking relations through which social connections are organized. The conceptual strength of Radcliffe-Brown's perspective is that this developed into new ideas about the significance of informal relations and interpersonal connections that laid the groundwork for community studies. Although there may be variations in the perceptions of connectedness, which confound measuring the impact of social networks, this concentration on the significance of relations correlates to a vast (and impactful) amount of empirical research that continues today. The second major conceptualization of SNA has been sparked, in part, by the emphasis on the significance of digital networks in everyday life and visibility of "networking" in organizational, business, and management terms. Many have found SNA useful in their observation of the correlation of the principle measurement of network dynamics and structure with key concepts in assessing social structures, for example, the analysis of community, centrality, cliques, density of ties, expanding niche communities, and privacy and networked publics, along with the abstraction of data from different social contexts. One of the greatest challenges facing researchers today is the ever-growing mountain of real-time data.

Key Applications

There are some important points for any researcher to discern and place into context for social network analysis. First, it is crucial to establish, at least at an individual level, the role that technologies have in the formal production of social practices and relationships. This is to scrutinize how, and in what way, technologies may surround research development and be deeply bound to political and economic products, as well as cultural artifacts (see Fine and Kleinman 1983). Second, the researcher must think about the social interpretation of network construction. This means the contextualizing of everyday and seemingly mundane social interactions within networks and groups of networks. Alongside these social relationships, consumption and consumerism take on significance. We live, as Mats Alvesson notes, in a society where trendy jargon, media appeal, and "looking good" define the successes of individuals, groups, and organizations. It is natural to extend this as an escalation of expectations that are part of the "gilt edge of life" (Alvesson 2013, p. 188), into networks that provide a growing focus for individuals to pursue high-status employment and to seek out socially prestigious others. Third, in understanding social networks and stripping away the elements of analysis, it is necessary to reflect on Barnes's question and to put in place some context for boundaries, whether these are a temporal positioning or shaded by cultural, social, political, or commercial significance.

The key applications of SNA are concurrent with the acquisition of more sophisticated technologies to collect, cleanse, and evaluate network data. This reflects emergent changes in the characteristics of social networks that consist of dimensions relating to correlative patterns of network attributes. It stands to reason that the more diversified the network experiences, the more attributes and different number of dimensions the researcher has available for adopting SNA concepts. For example, the schema representing a cluster of temporally related updates in a social app such as Snapchat could be interpreted as a social network describing a typical sequence of communication and set of relations between young users. In essence, this episodic schema may represent a repertoire of network situations that the technology has stored and where SNA may be applied to map out various configurations of relations into new analytical configurations. The methodological procedure and opportunities from SNA, like concepts, are directly related to the environmental and social situation – often dynamic and experiential. In other words, SNA allows researchers to deal with a complex set of time-space-sequenced actions and the combination of social relations. The variance in the use and adaptation of SNA is due in part to researchers' perceptions of the relative efficacy of network-related and social-related variables having significance for their investigation. A basic motivating factor underlying SNA use is that the social relations reveal a paradigm or context that has not been previously encountered. Here, we can propose five key attributes found to have general relevance to researchers seeking to adopt SNA:

1. *Relative categorization*: the degree to which researchers may be interested in the strength and/or weakness of social relations (positively related to adaptability within networks)
2. *Compatibility*: the extent to which social relations are consistent with past values and existing experience of connectedness
3. *Complexity*: the degree to which social relations and their relative value are difficult to understand or to use
4. *Observationally*: the extent to which social relations are visible to others
5. *Differentiation*: the degree to which social relations may be revealing of difference and the separation of ties

These propositions suggest that the contemporary researcher has to be more in tune than ever before with the factors believed to underlie SNA in that she must be attentive to the source of data and contextual sorting. At least three types of sorting may present challenges. First, to perform her role, the researcher conducts their own "sort" of the data, and it is likely she will undertake a variety of data-cleansing actions. Second, digital social relations in particular present their own version of sorting based on the number of categories the individual is performing. Third, the analytical extraction of the data is designed to enhance the perceived social relations and way in which these may be evaluated. To implement SNA, sets of operational measures are necessary along with the perceived intercorrelation for possible empirical measures and respective attribute sets. One approach would consist of the analysis of a group of individuals in a given social network domain to generate characteristics of relations and situations.

Within the informational economy of the digital age, contemporary practices of social network processing constitute a new type of research and direction for SNA. Often it is up to the researcher to find ways to control a repository of raw materials that are undergoing constant fragmentation and productive activity. The raw materials typical today are in flux, consisting of information identifying or relating to networks and the public domain of individual and personal data. As a methodological construct, SNA shapes practices of appropriation and use of social network information in two complementary ways. First, SNA constitutes points of connection and network information as publically available and potentially valuable: as a product of the information age and pool of digital materials that may be freely appropriated as inputs to SNA. This framing supports the reorganization of sociotechnical networks in ways directed toward data extraction and reappropriation. Second, SNA constitutes personal information harvested across social networks as raw. This context creates an important backdrop for knowledge production and for the analytical logic frame that designates this technique as an appropriate tool applied to the abundance of networks and extraction of data. The processes of resource extraction, the activities of identifying data sources, of collecting and processing network information require an enabling set of research constructs. In continuing to explore the physical and conceptual entailments of SNA is to appreciate emerging patterns of relations based on prediction and targeted extraction that profoundly influence our understanding of social networks. For the early adopters, the act of *doing SNA* has its roots in an era of global exploration, imperialism, privatization, and conquest. More recently, the context of an informational and digitally connected set of relations have shaped research and depend centrally on the idea of the public domain of data that acts as a repository of raw materials upon which future investigators and authors can build.

Future Directions

Every researcher needs to justify such efforts as part of a theoretical and methodological account and to make explicit their instrument handling and data processes.

There are, no doubt, additional vulnerabilities and opportunities that are geared to the stimulation of social network analysis. The points made here offer an overall condition that is specific to digital scientific interest that has been often overcelebrated without critical treatment and that may lead us to new interpretations and encounters with new forms of data. Moreover, as the technology becomes cheaper and more sophisticated, individuals, companies, organizations, and campaign activities increasingly benefit from this kind of statistical monitoring. We live in an information-rich data age from which SNA affords powerful and expressive representations of real-world interactions and offers different kinds of network insights. When we consider the future directions for SNA, the research process is strongly linked to hitherto impactful networks and the contents available for harvesting, cleansing, and capture. Contemporary descriptions about the relative value of network data are intimately tied to the commercial future of personal data processing and ownership. Advertised in business and marketing brochures, SNA has become an essential product to offer as part of client services. For example, several agencies are using SNA to identify key nodes or "influencers" to affect brand perception. Therefore, SNA also represents a new data industry opening up along with the many uses of new techniques to discover and track network data through the use of so-called spyware and dataveillance (for a comprehensive review of the enforcement actions regarding digital privacy legislation, see Cohen (2015)). Researchers using SNA should be aware of the absence of regulatory framework to cover the complexity of network tracking. In the USA, the Federal Trade Commission had advised to constrain the use of automated tagging and tracking protocols and to regulate unfair and deceptive practices in commerce (Solove and Hartzog 2014, p. 583). The activity of communication providers has also come under scrutiny with consumers often surprised to learn that companies like Apple, Google, Microsoft, and others are tracking digital network activities by means of embedded, often invisible, software and smart apps. The incidence of privacy regulation and scrutiny will continue to be impactful for SNA. In the past, network data and analysis required formal notification from the individual in order to install and configure data sharing. Today, it appears that the formal construct enabling ongoing data tracking is downplayed, and information businesses have powerful incentives to configure networked data in ways that make enrolment near automatic and straightforward. Thus, in the contemporary marketplace of network data, such things as disclosure, data dashboards, and user privacy are being sublimated and, in some extreme cases, used to justify network data as free of appropriation and official scrutiny. Any researcher who is interested in SNA should be attentive to the Global Network Initiative that states, "privacy is a human right" and (therefore) "[. . .] important to maintaining personal security, protecting identity and promoting freedom of expression in the digital age" (Global Network Initiative, "Principles," https://www.globalnetworkinitiative.org/principles/index.php#19.). The point here is that individuals must appreciate the opportunities for data collection alongside the privacy implications of networked information harvesting and analysis.

The drive to explore and understand social networks has produced a pattern that I identify as 'SNA-dual direction: first, initial extensions of network tracking, discovery, and mapping via a two-pronged strategy of development and customization and followed, second, by the magnification of networks as data are harvested, consolidated, analyzed, and stored. In other research contexts, the initiatives for social network data processing are framed as big data projects aimed at improving our understanding of, for example, living standards, the prospects of the world's healthcare resources, or mapping crime data. Among scholars, researchers, politicians, and activists, complex debate continues to unfold about whether SNA initiatives, and particularly the inclusion of big data, provide robust outcomes to understand impactful social networks. To put this point another way, the danger is to treat SNA as monolithic, and the commercial intensity to harvest and profit from the public domain of social network data is being intensively played out. This fairly mundane observation has significant implications for understanding the future of SNA, especially the emergence of the networked self and the emerging commercial companies that make a profit out of network resources. The popularity of large companies such as Apple, Google and Facebook underscore the significance of network-structured relationships and the way in which personal data is driving profitability in the networked economy. In summary, the networks of connection that characterize SNA and the emerging digital data industry represent new strategies through which resources extracted from social networks become both research and marketable assets and as sources of competing advantage.

Conclusion: The Power of SNA

The context of SNA reflects a pervasive interest in the spread of network patterning into areas such as digital social interaction, digital political advertising, news and current events coverage, search, business, and marketing. The shift to algorithmic and software-sorted modes of network information productions has deepened preexisting distributional analysis of network data, and researchers along with policy makers are beginning to give more attention to the profound impact of network data. The popularity and interest in SNA reflects a network of crowds and crowded data economy through which social capital is often viewed as a hierarchal conception that sits fundamentally in tension with the freedom and openness of personal data. For individuals, the change from private and closed information to open networks and resources is a primary state. As we become used to the automated extraction and tracking of our own network data, we are also in danger of falling victim to the Dark Side of Social Networking. Thus framed, the challenge is not simply that SNA facilitates the commodification and commercial production of network data, or that it enables hidden data tracking, but that there are varying degrees to which researchers must act to protect themselves and to put in place a firm foundation in the extraction and safeguarding of social network data long term.

Cross-References

- ▶ Arts and Humanities, Complex Network Analysis of
- ▶ Community Evolution
- ▶ Dark Side of Online Social Networks: Technical, Managerial, and Behavioral Perspectives
- ▶ E-Commerce and Internet Business
- ▶ e-Government
- ▶ Online Privacy Paradox and Social Networks

- ▶ Quality of Social Network Data
- ▶ Social Capital
- ▶ Social Network Datasets
- ▶ Spatial Networks
- ▶ Visualization of Large Networks

References

Alvesson M (2013) The triumph of emptiness: consumption, higher education, and work organization. Oxford University Press, Oxford

Banton M (ed) (1966) The social anthropology of complex societies. Tavistock, London

Barnes JA (1979) Network analysis: orientating notion, rigorous technique or substantive field of study. In: Laumann EO, Marsden PV, Prensky D (1989) (eds) The boundary specification problem in network analysis. Research methods in social network analysis, vol 61. p 87

Bavelas A (1948) A mathematical model of group structure. Hum Organ 7:16–30

Bell C, Newby H (eds) (1971) Community studies. George Allen and Unwin, London

boyd D, Crawford K (2012) Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inform Commun Soc 15 (5):662–679

Callon M (1987) Society in the making: the study of technology as a tool for sociological analysis. Soc Constr Technol Syst 550:83–103

Cohen JE (2015) The networked self in the modulated society. In: crossroads in new media, identity and law, pp 67–79. Palgrave Macmillan, UK

Fine GA, Kleinman S (1983) Network and meaning: an interactionist approach to structure. Symb Interact 6 (1):97

Fortes M (1949) The web of kinship amoung the Tallensi. Oxford University Press, Oxford

Fortes M (1953) Analysis and description in social anthropology. Adv Sci 38:190–201

Latour B (1996) On actor-network theory: a few clarifications. Soz Welt 47:369–381

Lazer D, Kennedy R, King G, Vespignani A (2014) The parable of Google flu: traps in big data analysis. Science 343(6176):1203–1205

Mitchell JC (1969) The concept and use of social networks. Bobbs-Merrill

Oliver D (1958) An ethnographer's method for formulating descriptions of "Social Structure". Am Anthropol 60(5):801–826

Rainie H, Rainie L, Wellman B (2012) Networked: the new social operating system. MIT, Cambridge

Solove DJ, Hartzog W (2014) The FTC and the new common law of privacy. Columbia Law Rev 114:583

St. Fleur, N. (2016). Visualising the indivisible drivers of climate change. The New York Times. 16 Dec. [online] https://www.nytimes.com/2016/12/16/science/carbon-dioxide-satellite.html?_r=0

Wellman B (1983) Network analysis: some basic principles. Social Theory 1(1):155–200

Recommended Reading

Berkowitz SD (1982) An introduction to structural analysis. Butterworth, Toronto

Scott J (2012) Social network analysis, 3rd edn. Sage, London

Wu X, Zhu X, Wu GQ, Ding W (2014) Data mining with big data. IEEE Trans Knowl Data Eng 26(1):97–107

Social Network Analysis in Organizational Structures Evaluation

Radosław Michalski[1] and Przemysław Kazienko[2]
[1]Department of Computational Intelligence, ENGINE – The European Centre for Data Science, Wroclaw University of Science and Technology, Wrocław, Poland
[2]Department of Computational Intelligence, ENGINE – The European Centre for Data Science, Faculty of Computer Science and Management, Wroclaw University of Science and Technology, Wrocław, Poland

Synonyms

Corporate hierarchy; Enterprise management; Organizational design; Organizational network analysis; Social network analysis in organizations

Glossary

HR	Human resources
Organizational Chart	A diagram representing the formal structure in the organization
ONA – Organizational	The analysis of the organization which focuses on the relationship between the

Network Analysis	informal social network and the formal structures in the organization: organizational charts, process definitions, and others
OSN – Organizational Social Network	An informal social network, which was created from the data collected within the organization like e-mail logs, phone call records, surveys, and others
MSN – Multilayered Social Network	The social network which consists of multiple layers; each of them represents different type of information used as a source for creating the network layer
SNA	Social network analysis

Definition

Although typical social network analysis (SNA) may bring interesting results while being applied in an organizational environment, it is a very promising to have these results compared with the organization itself. This allows gaining additional knowledge about the whole organizational environment. The reason for that is caused by the fact that each organization at the moment of performing social network analysis possesses a more or less structured hierarchy which regulates the information and workflow in the formal way. Simultaneously, members of the organization maintain the informal social network by contacting and collaborating with each other. It means that the comparison of formal and informal structures may enhance the knowledge about the employees by means of their role in both networks. In other words, one may say that the goal of this comparison is to check the discrepancy between the visible (defined, official) and the invisible (unofficial social network) structures in the organization.

Results of such analysis may bring the answer on some of the following questions: Are the organization members well placed in the organizational chart? Is the organization maximizing its information and decision flow efficiency by using the recent organizational design? and Are there any substantial discrepancies in the employees' formal role in the organization and their placement in the informal social network? Obtaining answers on those questions may lead to gaining the competitive edge.

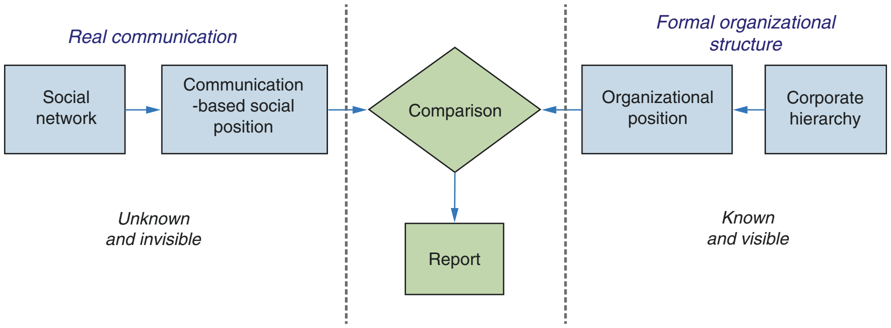
This kind of analysis is a key part of ONA – organizational network analysis, which may be described as a framework for understanding formal organizations (Knoke 2001).

Introduction

The possibility to collect the formal structure of the organization and its communication logs or collaboration traces has enabled the researchers the possibility to create a new field of social network analysis. By performing the comparison of both social networks – the formal and informal ones – some findings from such analysis may be useful for organization managers and the human resources departments.

While finding a chance to gain a competitive advantage, organizations are searching for the solutions that would enable them to beat their market opponents. It may be crucial to discover, among various ways of increasing company effectiveness, the company's own potential, hidden in the social network of the organization. The knowledge derived from this area, if properly extracted and interpreted, may lead to various positive effects in organization management (Palus et al. 2010; Song and van der Aalst 2008). The general idea of SNA application in organizational structure evaluation is illustrated in Fig. 1.

Managers may often ask the question about the proper alignment of their employees in the organization structure. The problem may be particularly important in fast-growing organizations, where medium-level management team may be chosen without prior adequate preparation and without the use of proper HR tools, e.g., start-ups. Companies, in which some of the employees



Social Network Analysis in Organizational Structures Evaluation, Fig. 1 The idea of comparing formal and visible organizational structure with informal hidden social network based on real communication, based on Michalski et al. (2011)

are awaiting retirement, are another example where the knowledge about real worker position may be crucial. If a company decides to search internally for the successor or replacement of anybody, social network analysis may become helpful for such a task. It may also be helpful in extracting some prospective problems in the company, like managers avoiding communication with other employees within their units.

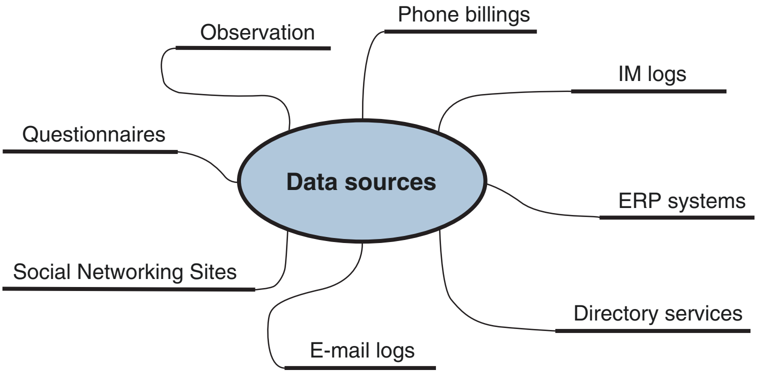
The problem of a proper organizational design is considered as a crucial one in corporate management (Daft 2009), because it strongly influences the information and decision flow. In that case, managers should permanently observe the internal structure of the company in terms of bottlenecks, overhead, or other possible problems. However, the task of performing such analyses manually becomes virtually impossible, due to the fact that the amount of information exchanged in the typical organization is too big even to be just observed. On the other hand, this process may be automated and more or less quantified by application of the social network analysis and using already existing information, such as organization charts and communication logs. Still, the final results should be treated individually, because some of the information may not be included in the analysis.

Key Points

To benefit from SNA applied to the organization, a number of steps have to be carried out. There is a need to obtain both formal and visible and informal and invisible structures of the organization, see Fig. 1.

Overall, in case of organizational structure evaluation, the most important information and the starting point is the organization chart. In the simplest scenario, the organization may use the functional design where every department is responsible for different tasks, each organization member belongs to only one department, and all the departments form a hierarchy, which starts from the management board (Daft 2009). However, nowadays, some more complicated scenarios also apply like process-oriented or matrix structures. These require some more effort while being analyzed, and they introduce additional limitations as well, which will be discussed later. The organization chart itself may also be represented by a graph that eases the further comparison of both structures. The informal part of the organization is often a multilayered social network, which was built by using variety of data sources available in the company. Some of these data sources are presented in Fig. 2.

The process of evaluation of the organizational structure actually begins after performing some preliminary analysis understood as a feasibility study and conducting matching of entities of both networks, described later on. The overall result of the analysis is the report presenting the structural difference between formal and informal networks in the organization.



Social Network Analysis in Organizational Structures Evaluation, Fig. 2 A choice of possible data sources for social network extraction in the organization (Michalski et al. 2011)

Summarizing, the whole analytical process consists of the following steps:

Source data preprocessing
Social network building
Network measure calculation
Social network and corporate hierarchy comparison

A complete process and framework that enables evaluating the organizational structure in the company by using SNA will be further depicted by using the example of the Enron company case.

Historical Background

The idea of performing the SNA in organizations is not completely new (Tichy et al. 1979); however, since the introduction of this idea, the general SNA measures and metrics were developed intensively (Wasserman and Faust 1994). At the beginning, organizational network analysis (ONA) was more oriented to discovering key players in the organization without matching organizational social networks (OSN) to the organizational charts and finding the structural holes. Later on, some experiments related to uncovering the informal structure in the organizational social network were performed (Borgatti and Molina 2003), but the real enabler and accelerator in the research was the Enron case. It became especially famous worldwide in 2001 due to financial manipulation scandal. The Enron e-mail dataset was made public by the Federal Energy Regulatory Commission during its investigation (Klimt and Yang 2004). The Enron official hierarchy structure still remains publicly unavailable. However, there are some sources which can provide information concerning plenty of job positions of selected employees together with their department or division (Rowe et al. 2007). This led to a number of analyses which focused on the relationship between formal and informal structures (Rowe et al. 2007; Diesner et al. 2005; Hossain 2009; Borgatti and Molina 2003). Nevertheless, any new research is limited by the data availability – to fully evaluate the organization by means of the social network, as it was previously described, not only the social network data but also organizational details are needed. As a result, not all the organizations are happy to disclose this information limiting further research opportunity. On the other hand, there is also the research conducted in which the formal structure is not needed (Fire and Puzis 2016), yet it also provides limited results compared to the full comparison.

Analysis of Organizational Structures

A social network, which is built on the basis of employees' communication logs, may be found useful in the evaluation of formal organizational structures existing in the company. The communication-based social network can provide information about social network leaders, communication gaps, and anomalies. However, the problem is what factors in the social network analysis results should be considered as important ones and useful in further company management decisions. Another problem is how to perform such analysis in order to ensure its acceptable meaningfulness and representativeness.

Although this sort of analyses is mostly performed for business companies, other types of organizations may benefit by performing such a study as well. Moreover, if a company has introduced some organizational changes in order to reach some goals that should also result in communication changes, these kinds of comparisons become essentially useful for validation of such changes.

The variety of communication forms, which are already used in organizations, such as e-mails, instant messaging systems, ERP systems, and landline and mobile phones, allow to build the social network as a multilayered one; each layer corresponds to another communication channel. All these data sources facilitate gathering more information about the whole communication in the organization and include more organization members. Additionally, by applying separate weights to these layers, some of them may be interpreted as more important or more *social* ones.

Although the formal structure influences all the communication between members of the organization, yet there still exists the space for the informal communication. However, there is no easy way to distinguish both types of communication without complex and resource-consuming analysis. But still, as it will be shown, such a distinction is not necessarily needed to perform meaningful and useful analyses.

In this entry, the typical social network analysis in organizational structure evaluation is presented. Especially, the role the preliminary part of the whole process – the feasibility study – is underlined and ideas regarding the analysis of the dynamics of the organizational network are introduced. To complete the study, selected limitations of the proposed approach are presented as well. The case study on the Enron company depicts usability of the proposed method.

Feasibility Study

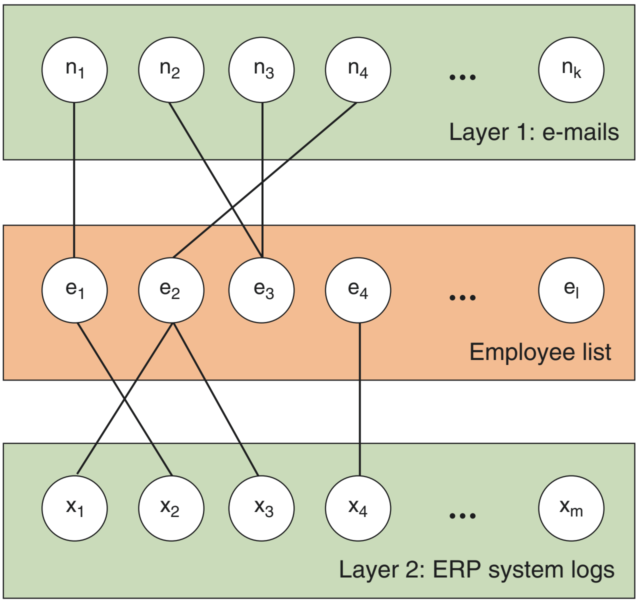
As it was mentioned in the "Introduction" section, the first step of the process is the source data preprocessing which may be also treated as a separate feasibility study, because this part may even disqualify all source data or part of the company from further analysis.

There are at least five important factors to be taken under consideration while obtaining the data and performing data preprocessing:

The choice of data sources
Adjustment of the period of the analysis
Extraction of the official organizational structure
Matching informal social network entities to employees
Employees using external communication channels (or simply not registered by the organization)

To obtain the comprehensive communication social network, it is worth to consider multiple data sources: phone calls, e-mails, instant messaging systems, or even ERP and workflow management systems. These will be used to build a multilayered social network (MSN), a directed and weighted graph with weights representing the importance and intensity of relations between users. The chance to gather more valuable results grows up by using more than one data source, but not without consequences. First of all, if there are multiple data sources, the researcher should try to obtain the data of the same or at least similar time frame. It is caused by the fact that the social network evolves and the relations (in that case weights of the graph edges) change over time. It is nearly impossible to create a representative multilayered social network by using the data from different periods, so the nonoverlapping periods shall be cut off or only the common part of the datasets should be utilized.

When designing the study, a researcher faces the problem of choosing the period of the analysis. The shorter the period, the chance for matching most of the nodes in the graph decreases. However, for the longer periods (years), it is necessary to tackle with the problem of the probably smaller importance of old communication compared to the new one. Typically, it is solved by assignment of higher weights



Social Network Analysis in Organizational Structures Evaluation, Fig. 3 Mapping social network actors to employee list (Kazienko et al. 2011)

to newer communication (Kazienko et al. 2009). On the other hand, some other difficulties are partially overcome for the longer periods: holidays and longer illness absences. In general, the period of half a year should be considered as representative enough. However, it is also expected that the formal structure of the company would not change strongly over time of analysis. Otherwise, it would be hard to compare both the social network and the formal structure. Concluding, the above limitations clearly show that the selection of the most suitable time frame for the analysis may be challenging.

Yet one more problem is related to the organizational hierarchy. Paradoxically, such kind of relatively well-defined corporate relations, i.e., the organizational structure, can be hardly extracted automatically, because, depending on the size and profile, the company may have no need to maintain the full company structure drill-down from board through departments up to a single employee in their IT systems. That is why it may be necessary to convert organizational structure, taken from official documents, into a graph, where nodes represent employees and edges – employee-supervisor relations.

Another important problem is the need for finding the same entities in the datasets, as presented in Fig. 3. If we consider the employees as entities, there should exist a mapping between any entities in the other data sources to the employee. However, due to the nature of communication systems, such a mapping may not exist at all. To present one of the examples, let us assume that albeit every warehouse worker has got his individual e-mail account, all these workers may use the same one account in the instant messenger (IM) system. In that case, the researcher should decide whether the IM layer will be analyzed deeper, the IM layer will be discarded completely, or only the warehouse will be excluded from the analysis. Alternatively, the whole warehouse department will be treated as a single node, and only its relationships with other departments will be studied. Another source of potential problems is the fact that the company may use multiple aliases for a single employee, for instance, in the e-mail system – in that case they should be merged and mapped to the single entity in the employee list. Of course, it is also possible that an employee would not use some communication systems, which actually is not a problem for the analysis – a node may be isolated in some layers.

There might be some employees using external communication techniques (supervisory board, expats), and some employees may be represented by other ones, as it often happens for top-level managers – they are substituted by their assistants for writing and sending e-mails. This should be respected at the preprocessing stage.

The problems discussed above are to be found in most organizations. However, they may be overcome and there is still the chance to conduct reliable studies. However, some more limitations may be recognized in the organization itself and they were described in the "Limitations" section.

Building the Social Network and the Corporate Hierarchy

The process of building social network consists of choosing the graph type (directed or undirected) and weight calculation method for edges linking nodes. In general, various methods of relationship valuation may be applied including multilayered (multigraph) concepts, in which two nodes are connected by means of multiple edges. However, in the Enron dataset utilized in the use case, only one layer exists, so a single-layered social network has been built. A directed and weighted graph may be created from e-mail logs using the following formula for the weight of an edge between node *i* and *j*:

$$w_{ij} = \frac{\sum e_{ij}}{\sum e_i}, \qquad (1)$$

where $\sum e_{ij}$ is the number of e-mails sent by node *i* to node *j* and $\sum e_i$ is a total number of e-mails sent by *i*. It means that weight w_{ij} focuses on the local neighborhood of an employee rather than on global network characteristic. As it was mentioned earlier, it is also possible to extend the above approach by applying the importance of correspondence in terms of time. Then, each email would not be counted as 1 ($e_{ij} = 1$) but as a fraction of 1 depending on its time stamp-smaller values for older messages: $e_{ij} = 1/\lambda^k$, where λ ∊0.0; 1] is constant, e.g., 0.8, and *k* is the period index (0 – for the newest period, 1 – for the previous one, and so on). Obviously, instead of numbers of e-mails exchanged, some other communication logs may be used like phone records or IM chats.

After building the social network, it is also required to obtain the organizational hierarchy. Paradoxically, sometimes it can be even more difficult than creation of social network. It refers especially informal and vague organizational structures. Especially for larger companies, their hierarchy may be extracted from internal phone books or other catalogues.

Introducing the Network Measures

A variety of measures were computed to reflect different aspects of the importance of the node in the social networks (Wasserman and Faust 1994; Scott 2000). They were combined into a single value – social score, as presented in Rowe et al. (2007) and Palus et al. (2010) by including:

(a) E-mails count – the number of e-mails a user has sent and received.
(b) Average response time – the time elapsed between a user sent an e-mail and later received a response e-mail from that same person. The exchange of this nature is considered a "response" only if a received message succeeds a sent message within three business days.
(c) Response score – a combination of the number of responses and average response time.
(d) Number of cliques – the number of maximal complete subgraphs that the account belongs to.
(e) Raw clique score – a score computed using a size of the given account's clique set. Bigger cliques are worth more than smaller ones; importance increases exponentially with size.
(f) Weighted clique score – a score computed using the importance of the people in each clique, which is computed strictly from the number of e-mails and the average response time.
(g) Centrality degree – count of the number of ties to other actors (nodes) in the network.
(h) Clustering coefficient – likelihood that two associates of a node are also linked with themselves.

(i) Mean of the shortest path length from a specific vertex to all vertices in the graph.
(j) Betweenness centrality – reflects the contribution of a given node in all shortest paths connecting all pairs of nodes, i.e., how important is a node in linking other nodes.
(k) *Hubs-and-authorities* importance – refers to the algorithm proposed in Kleinberg (1999).

Above measures are then weighted and normalized to a [0, 100] scale, as presented in Rowe et al. (2007). Obviously, some other measures can be utilized in a given organization according to the needs and data availability.

Hierarchical Position

It is possible to find people who are higher or lower in the hierarchy for each employee in the corporate hierarchy. The Hierarchical Position (HP) is a measure that denotes the importance of an employee within the company (Kleinberg 1999). For each user *i* in a company C, there is a sum of hierarchical differences *D* between *i* and any other user *j* in the company normalized by the total number of other users.

$$HP(i) = \frac{\sum_{j \in C \wedge j \neq i} D(i,j)}{m-1} \tag{2}$$

The hierarchical difference *D(ij)* is computed as follows:

$$D(i,j) = \begin{cases} 1, & \text{if } i \text{ is higher in the hierarchy than } j \\ 0, & \text{if } i \text{ and } j \text{ are at the same level of the hierarchy} \\ -1, & \text{if } i \text{ is lower in the hierarchy than } j \end{cases} \tag{3}$$

At first, the Kendall's rankings comparison method was used to compare two rankings (Kendall and Gibbons 1990). It compares the nodes in pairs, i.e., the positions of pair nodes within both rankings. If the position of node A is related to the position of node B in both rankings monotonically in the same direction (lower or higher in the both hierarchies), then this pair is well correlated. It is assumed that when the level in hierarchy is the same within the pair, then it does not matter whether they are in different positions in the second ranking. Kendall's *t* rank correlation coefficient is a value from the [−1, 1] range, where 1 means that two rankings are perfectly correlated and −1 means that they are completely different (in the opposite order).

It is impossible to distinguish the importance of departments, e.g., whether the Director of Northwest is higher in the hierarchy than the Director of Fundamental Analysis; see Fig. 4. Thus, analyses in the Enron use case were not performed globally, but locally at department level.

Discovering the Organizational Level of Employees

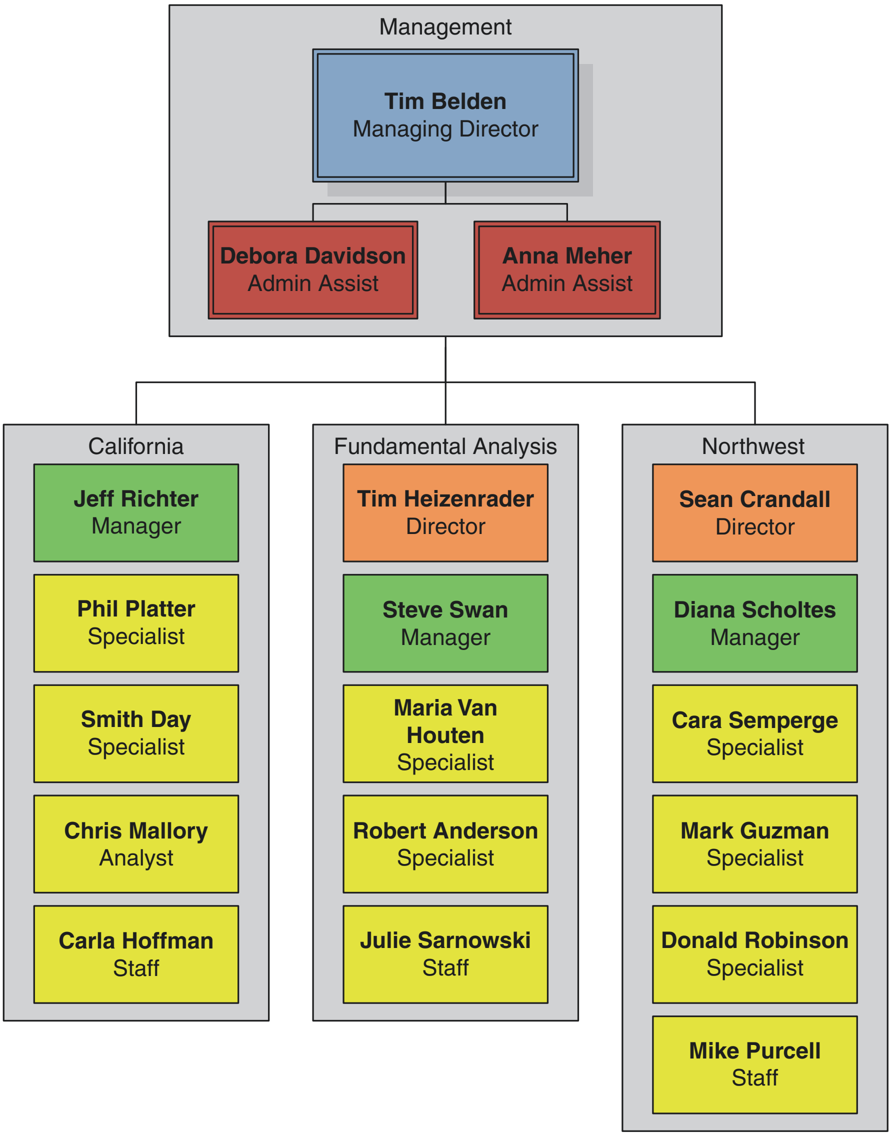
The structural node measures within the social network can be utilized to predict hierarchy level of particular nodes – employees. Some of these structural features may be more while the other less correlated with the organizational level. As a result, the level of a given person in the organizational hierarchy may be discovered based only on this person's centrality measures in the social network.

Some of these typical centrality measures were compared for the Enron employees; see the use case described in the following section. The results are presented in Table 1.

The above analysis clearly shows that some measures like in-degree centrality and centrality eigenvector are able to identify the level of the employee with quite good accuracy and that there exists the general relation between the employees' social network position and the corporate hierarchy placement.

Enron Use Case

The use case for evaluation of the organizational structure will be presented based on the Enron dataset that contains e-mail communication between employees. This e-mail corpus is extracted from mailboxes of 150 Enron employees, mostly senior management. In total, they contain



Social Network Analysis in Organizational Structures Evaluation, Fig. 4 A fragment of the Enron corporate hierarchy (Palus et al. 2010)

517,430 e-mail messages. Because this is the only available communication channel, only one layer in the social network, a single-layered social network, may be built (Klimt and Yang 2004).

Having the social network built, the organizational hierarchy must be identified. There is an Excel file with a list of over 160 employees and their job title available at Shetty and Adibi (2004). Many of them do not exist in the Enron Corpus, though. Using this list, four groups from Enron North American West Power Traders are chosen – it is possible to distinguish levels of hierarchy by matching them with job titles. Since only a part of hierarchy is available, the most complete part of it has been taken for further analysis. The extracted hierarchy is presented in Fig. 4.

The list of Enron employees sorted by their *social score* (see "Introducing the Network Measures" section) is presented in Table 2. The *HP* measure (Eq. 2) and *Position* column indicates official hierarchy structure. It can be seen very clearly that social scores of the management is far higher than the others.

Social Network Analysis in Organizational Structures Evaluation, Table 1 The accuracy of management level matching while using various social network metrics (Michalski et al. 2011)

	Percentage of the management level employees matched	Percentage of regular employees matched
In-degree centrality	67	85
Out-degree centrality	50	77
Centrality betweenness	33	69
Centrality closeness	33	69
Clustering coefficient	17	62
Centrality eigenvector	67	85

The diagram of *Hierarchical Position* should be descending, but there are deep structural holes, as presented in Fig. 5.

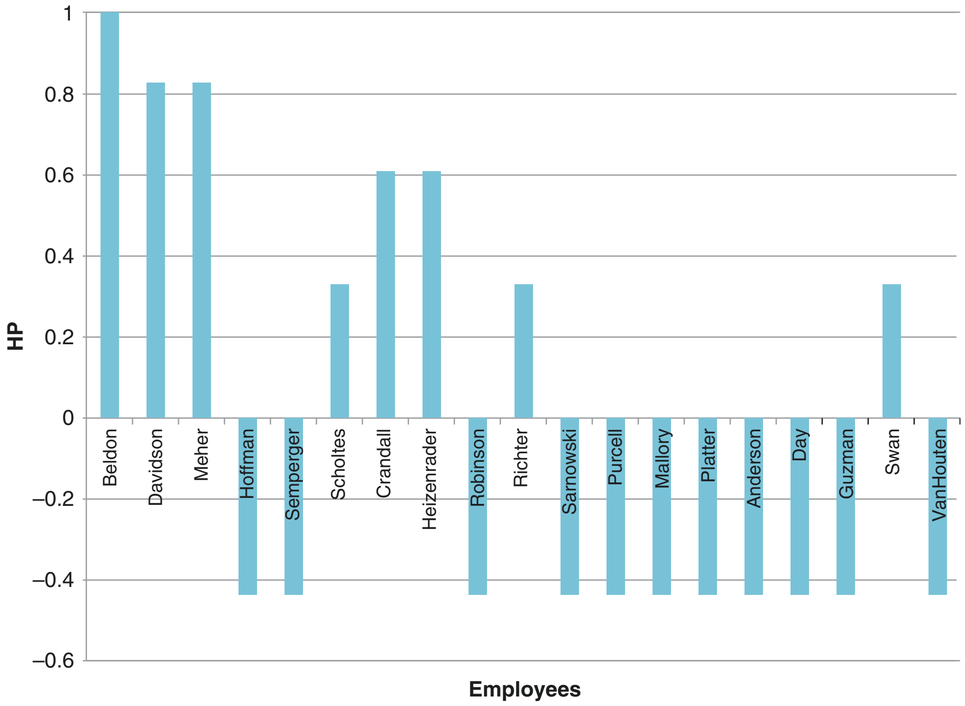
The summary of Kendall's correlation coefficient between the official hierarchy (ordered by *HP*) and the one derived from the social network (ordered by *social score*) for chosen departments is presented in Table 3.

The main problem with the Enron dataset is lack of information about direct hierarchy structure; only partial information was known. However, the analysis shows the rankings are very similar with Kendall's rank over 0.6 with management department perfectly identical (Kendall's rank of 1).

An interesting fact is that all employees who are lower in the hierarchy than who comes from the social network are women. There are 7 women among 19 analyzed employees, and there are 5 female workers in the top 6 of the social ranking, while 4 have been classified as lowest-level employees according to the hierarchy

Social Network Analysis in Organizational Structures Evaluation, Table 2 Social measures for Enron employees sorted by the social score (Palus et al. 2010)

Name	Surname	Position	Lv1	HP	Degree	Betweenness	Hubs	Clustering	Social score
Tim	Beldon	Managing Director	1	1.00	83	370.35	0.04	0.40	75.68
Debora	Davidson	Admin assist	2	0.83	66	278.35	0.04	0.41	63.51
Anna	Meher	Admin assist	2	0.83	62	260.94	0.04	0.42	62.84
Carla	Hoffman	Staff	5	−0.44	55	143.98	0.04	0.49	61.67
Cara	Semperger	Specialist	5	−0.44	63	82.96	0.03	0.52	53.68
Diana	Scholtes	Manager	4	0.33	45	21.44	0.03	0.70	53.31
Sean	Crandall	Director	3	0.61	42	40.04	0.03	0.62	43.64
Tim	Heizenrader	Director	3	0.61	33	19.45	0.02	0.71	35.56
Donald	Robinson	Specialist	5	−0.44	27	6.67	0.02	0.81	33.03
Jeff	Richter	Manager	4	0.33	25	12.80	0.02	0.74	32.53
Julie	Sarnowski	Staff	5	−0.44	28	25.94	0.02	0.63	32.14
Mike	Purcell	Staff	5	−0.44	24	5.02	0.02	0.79	30.36
Chris	Mallory	Analyst	5	−0.44	27	9.92	0.02	0.76	30.19
Phil	Platter	Specialist	5	−0.44	33	34.34	0.02	0.63	27.90
Robert	Anderson	Specialist	5	−0.44	8	0.15	0.01	0.96	20.06
Smith	Day	Specialist	5	−0.44	6	0.00	0.01	1.00	20.00
Mark	Guzman	Specialist	5	−0.44	18	6.84	0.01	0.75	19.97
Steve	Swan	Manager	4	0.33	9	0.20	0.01	0.93	19.55
Maria	VanHouten	Specialist	5	−0.44	7	0.11	0.01	0.95	19.44



Social Network Analysis in Organizational Structures Evaluation, Fig. 5 Hierarchical positions of Enron employees sorted by social score (Palus et al. 2010)

(these are marked green in Table 2). There are three possible reasons of such case. Firstly, a wrong assumption has been made while ranking job titles. Secondly, there can be a simple but important reason that women are underestimated and maybe should occupy higher company positions. The last, but not less probable, is that women may be more likely to gossip than men, and this fact is disrupting the process of proper social network extraction. It is possible that the real reason is the combination of these three.

Social Network Analysis in Organizational Structures Evaluation, Table 3 Kendall's correlation coefficient for each department between official hierarchy and the social network (Palus et al. 2010)

Department	Kendall's correlation coefficient
Management (official vs. SN)	1.0
California (official vs. SN)	0.8
Fundamental Analysis (official vs. SN)	0.6
Northwest (official vs. SN)	0.6

Limitations

It must be clearly stated that the comparative analysis may be applied in the more effective way to companies with the stable (probably functional) organization design (Daft 2009), because other designs, such as matrix or horizontal ones, would not allow to create a hierarchy chart easy comparable with the social network ranks.

However, while performing such an analysis, there is also the need to consider ethical aspects of performing such studies inside the organization (Borgatti and Molina 2003). That is why every result of the evaluation should be individually interpreted and discussed. In particular, the access to organizational communication logs may be treated as violation of privacy protection restrictions. Sometimes, it may be necessary to obtain individual employee permissions to process the data.

It is also related to very important limitation: this kind of analyses require to process data, which may be considered as very sensitive for the organization. Even while applying anonymization procedures on the data, there is a big chance to map entities to real employees, especially while analyzing the organizational hierarchy. In fact, the organization must be trustful and convinced to share this kind of information with researchers.

Key Applications

The idea of matching organizational structure and the social network may be regarded as another possible way to improve overall company management. The idea focuses on the comparison of calculated node position ranks using chosen measures within the organization structure.

The positive results in comparison of formal structure with communication-based social network may mean that the similar level of managers and regular employees was properly assigned to their management levels. However, if the results differ significantly, some more sophisticated analysis might be needed to answer the question why real-life communication and hierarchy do not necessarily cover organization chart. The reasons may be different: (i) not the most important social network source has been analyzed, (ii) due to the profile of the company, communication between company members have nothing to do with their formal position, (ii) the relations change too fast to give stable point of view, or (iv) the company chose inappropriate persons to hold some management positions.

There might be also another usage of proposed concepts. The choice of new leaders in the organization can be supported by the application of the described set of methods, i.e., through recognition as prospective candidate for managers those employees who belong to the higher level of the management team in based on unofficial communication (having compared to the formal organization structure).

There is one more, more controversial, application field of the considered methodology. If someone wants to uncover organizational hierarchy, e.g., for crime groups, or at least wishes to know possible managers of this organization using available communication logs (phone records registered by the telecom company), they may discover organization managers in the easier, faster, and safer (passive) way. This may be used by the police in their investigations.

Despite all the techniques regarding core data analysis that may be very ambitious for SNA experts, the real challenge for companies is to properly interpret and make valuable use of the achieved corporate SNA results.

Future Directions

A very promising field in ONA is related to the dynamics of the organization. At monitoring the social network in organizations and calculating the HP values, one may discover that in some parts of the organization some problems arise even before they will be officially mentioned.

The other usage of the above approach is the ability to observe how fast the just introduced organizational changes are influencing the communication social network. It could be used for validation of motivation programs.

Cross-References

- ▶ Anomaly Detection
- ▶ Anonymization and De-anonymization of Social Network Data
- ▶ Classical Algorithms for Social Network Analysis: Future and Current Trends
- ▶ Collection and Analysis of Relational Data in Organizational and Market Settings
- ▶ Ethics of Social Networks and Mining
- ▶ Intraorganizational Networks
- ▶ Managerial Networking
- ▶ Multilayered Social Network
- ▶ Network Actors Within Entrepreneurial Networks: The Current State of Research
- ▶ Process of Social Network Analysis

- ▶ Ranking Methods for Networks
- ▶ Role Identification of Social Networkers
- ▶ Social Network Analysis and Organizational Multimodal Representation
- ▶ Top Management Team Networks

Acknowledgments This work was partially supported by the National Science Centre, Poland, the decisions no. DEC-2016/21/B/ST6/01463 and 2015/17/D/ST6/04046.

References

Borgatti SP, Molina JL (2003) Ethical and strategic issues in organizational social network analysis. J Appl Behav Sci 39(3):337–349

Daft RL (2009) Organization theory and design, 10th edn. Cengage Learning, Cincinnati

Diesner J, Frantz TL, Carley KM (2005) Communication networks from the Enron email corpus "It's always about the people. Enron is no different". Comput Math Organ Theory 11(3):201–228

Fire M, Puzis R (2016) Organization mining using online social networks. Netw Spatial Econ 16(2):545–578

Hossain L (2009) Effect of organizational position and network centrality on project coordination. Int J Proj Manag 27(7):680–689

Kazienko P, Musial K, Zgrzywa A (2009) Evaluation of node position based on email communication. Control Cybern 38(1):67–86

Kazienko P, Michalski R, Palus S (2011) Social network analysis as a tool for improving enterprise architecture. In: KES-AMSTA 2011, the 5th international KES symposium on agents and multi-agent systems – technologies and applications. Lecture notes in artificial intelligence, LNAI 6682. Springer, Berlin/Heidelberg, pp 651–660

Kendall MG, Gibbons JD (1990) Rank correlation methods, 5th edn. Edward Arnold, A Division of Hodder & Sloughton, London

Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632

Klimt B, Yang Y (2004) Introducing the Enron corpus. In: CEAS 2004, 1st conference on email and anti-spam, Mountain View

Knoke D (2001) Changing organizations: business networks in the new political economy. Westview Press, Boulder

Michalski R, Palus S, Kazienko P (2011) Matching organizational structure and social network extracted from email communication. In: BIS 2011, 14th international conference on business information systems. Lecture notes in business information processing, LNBIP 87. Springer, Berlin/Heidelberg, pp 197–206

Palus S, Bródka P, Kazienko P (2010) How to analyze company using social network? In: WSKS 2010, the 3rd world summit on the knowledge society, Corfu, 22–24 Sept 2010. Communications in computer and information science, vol 111. Springer, Berlin/Heidelberg, pp 159–164

Rowe R, Creamer G, Hershkop S, Stolfo SJ (2007) Automated social hierarchy detection through email network analysis. In: Proceedings of the 9th WebKDD and 1st SNAKDD 2007 workshop on Web mining and social network analysis WebKDD/SNAKDD 2007. ACM, New York, pp 109–117

Scott J (2000) Social network analysis: a handbook, vol 3, no 5. Sage, Thousand Oaks, p 208

Shetty J, Adibi J (2004) Ex employee status report. http://www.isi.edu/~adibi/Enron/Enron_Employee_Status.xls

Song M, van der Aalst WMP (2008) Towards comprehensive support for organizational mining. Decis Support Syst 46(1):300–317

Tichy NM, Tushman ML, Fombrun C (1979) Social network analysis for organizations. Acad Manag Rev 4 (4):507–519

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge

Social Network Analysis in Organizations

▶ Social Network Analysis in Organizational Structures Evaluation

Social Network Analysis Packages

▶ Tools for Networks

Social Network Anonymity

▶ Privacy in Social Networks, Current and Future Research Trends on

Social Network Data

▶ Sociograph Representations, Concepts, Data, and Analysis

Social Network Datasets

Jérôme Kunegis
Institute for Web Science and Technologies,
University of Koblenz-Landau, Koblenz,
Germany

Synonyms

Network dataset; Social graph dataset

Glossary

Bipartite	A network is bipartite when it contains two distinct node types, and all edges connect a node of the first type with a node of the second type
Directed	A network is directed when each edge has an orientation, i.e., each edge explicitly goes *from* one node *to* another node
Timestamps	When a network has timestamps, the creation time of each edge is known
Undirected	A network is undirected when its edges do not have an orientation
Unipartite	A network is unipartite when it contains a single node type
Weighted	A network is weighted if its edges are labeled with edge weights, for instance, rating values

Definition

A social network dataset is a dataset containing the structural information of a social network. In the general case, a social network dataset consists of persons connected by edges. Social network datasets can represent friendship relationships or may be extracted from a social networking Web site (Kunegis 2013). Social network datasets are widely used, not only in the area of social network analysis but also in the areas of data mining, Web science, and network analysis as the basis for various kinds of research.

Introduction

In order to study social networks, social network datasets are necessary. Thus, the availability of social network datasets are of crucial importance in all disciplines covering social networks. Beyond the area of social network analysis, social networks are studied in such diverse fields as data mining, Web science, network science, recommender systems, and many more. In fact, the majority of research being performed in these fields takes the form of the analysis of a social network and its usage as the basis of further analyses. Therefore, an increasing number of social network datasets are used in the literature, of which more and more are openly available.

Since the success of social media platforms such as Facebook and Twitter with the general public, these social networks have been increasingly studied, and accordingly a high number of datasets of these sites are available.

Historical Background

Historically, sociologists and anthropologists have studied social networks either theoretically by constructing corresponding models, or have observed or conducted surveys and then assembled social network datasets by hand. As an arbitrary example, the article *Cultures of the Central Highlands, New Guinea* by Read (1954) lists, in the form of a table, the 55 relationships between 16 tribes of the Central Highlands in New Guinea.

While a network of 16 nodes is perfectly correct in that it faithfully represents that actual relationships between tribes, its 55 edges are too few for performing statistical analyses. For instance, a common network analysis tool is the *degree distribution*, in which the number of nodes having a specific number of neighbors (the degree) are counted, resulting in the observation of *power laws*. These power laws reflect the fact that the number of nodes with *n* neighbors is proportional to $n^{-\gamma}$, for some constant γ. When applied to the network of New Guinean tribes, power laws are not observed. This does not mean however that the tribal network is in any way special. It does just mean that the network is too small to observe power laws. In fact, the larger a network, the easier it is to make statistically significant observations. Thus, small networks assembled by hand are too small for data mining applications. Instead, larger networks must be used.

Large social network datasets have become possible with the World Wide Web. With the availability of online social networking sites, large social network datasets have become available for research and other purposes. Nowadays, a large selection of large network datasets can be used, although many datasets are still proprietary and only available to the research divisions of social networking companies. Additionally, social media is used to collect other types of networks, for instance, rating graphs, consisting of ratings by users of items, or communication networks, consisting of individual messages such as emails sent between users.

Social Network Datasets

A social network datasets is mathematically a graph, with optionally additional structure. In the simplest case, a social network dataset is simply a graph

$$G = (V, E)$$

in which the vertex set V represents the users and the edge set E represents the friendships. In these kinds of datasets, each edge $\{i, j\}$ is undirected as are the friendships on Facebook (http://www.facebook.com) rather than directed as is the *follow* relationship on Twitter (twitter.com). Also, these kinds of network datasets allow only a single edge between two nodes.

In the following subsections, we will first describe basic statistics and analyses that can be computed and performed with social network datasets, and then describe additional forms of structure associated with social network datasets.

Social Network Datasets, Table 1 The possible structural features of social network datasets. Each of these features is described in one subsection

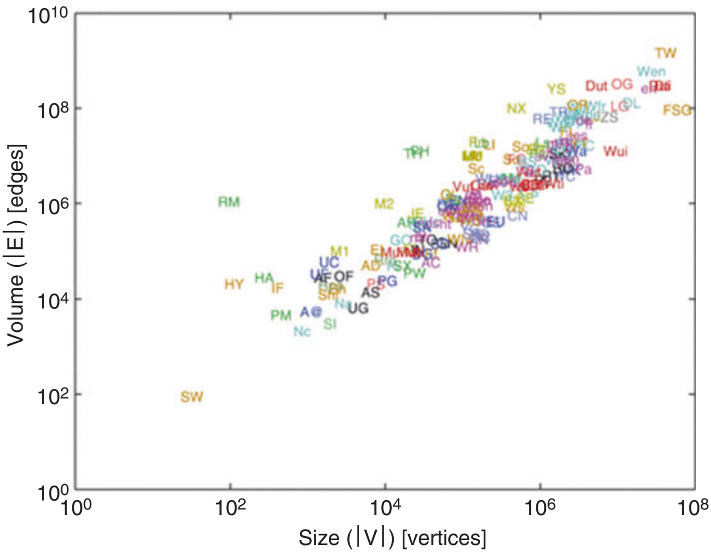
Feature	Description
Directed network datasets	Each edge is directed
Bipartite network dataset	There are two node types; each edge connects two nodes of different type
Network datasets with multiple edges	Multiple edges are permitted between any node pair
Signed network datasets	Edges can be positive or negative
Rating network datasets	Each edge represents a rating and is thus annotated with a rating value
Temporal network datasets	Each edge is annotated with an edge creation time, allowing the evolution of the network to be studied
Multirelational network datasets	Multiple edge types exist
Typed networks	There are multiple node and edge types

The possible structural features of network datasets are summarized in Table 1.

Network Dataset Statistics

Trivial statistics of a social network dataset are the number of nodes $|V|$, which is also called the *size* of a network, and the number of edges $|E|$, also called the *volume* of a network. The size of social network datasets range from a dozen forpre-Internet networks from anthropology and sociology to several hundred thousands for large social networking sites such as Face-book (Backstrom et al. 2012) and Twitter (Kwak et al. 2010). Figure 1 shows an overview of the network datasets from the KONECT project (Kunegis 2013), a collection of network datasets of typical sizes.

Other common statistics are described in the following. We must note that not all notations are established: While graphs almost universally written as $G = (V, E)$, the average degree, for instance, may be denoted by several symbols. The notation we use here represents a reasonable choice in symbols, although it is not universal.



Social Network Datasets, Fig. 1 A typical collection of network datasets from social media, from the KONECT project (Kunegis 2013). Each *letter code* represents one network dataset

The *average degree* is used as a statistic and ranges from 1 to about 100 in the most dense networks datasets. The average degree can be defined as

$$d = 2 \mid E \mid / \mid V \mid .$$

The *fill* is the proportion of edges to the number of total possible edges. The fill can be defined as

$$f = \mid E \mid / \left(\frac{1}{2} \mid V \mid (\mid V \mid - 1) \right).$$

Both the average degree and the fill are sometimes called the *density* in the literature.

The size of the *largest connected component* is sometimes given as a social network statistic, although social network datasets are often connected, making this statistic equal to the network size.

The *clustering coefficient c* equals the probability that two friends of a single persons are themselves friends. The clustering coefficient is thus a number between zero and one. A high clustering coefficient in social networks is used as an indication that a network is a *small-world network* (Watts and Strogatz 1998).

The *algebraic connectivity a* is defined as the second-smallest eigenvalue of the social network's Laplacian matrix **L** (Fiedler 1973). The algebraic connectivity is zero when the network is not connected; otherwise, it is larger than zero. The algebraic connectivity is used to measure the connectivity of a network.

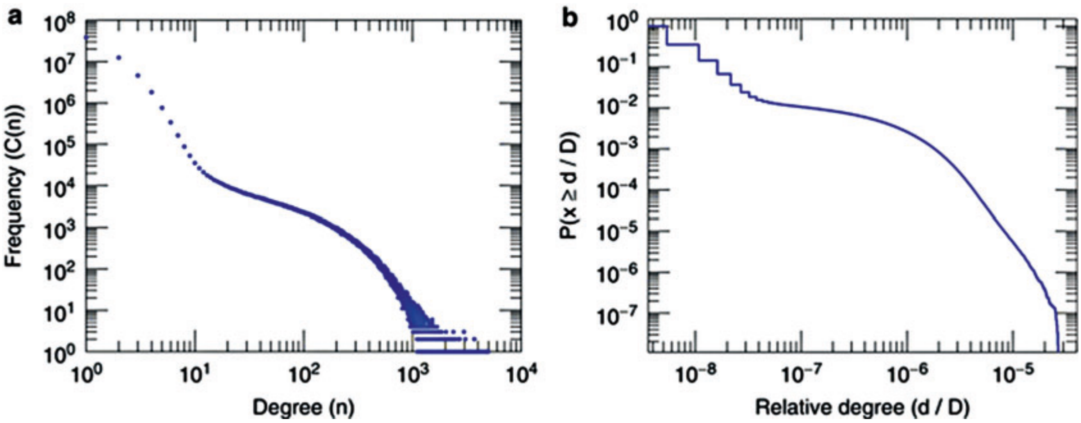
The *spectral norm* $\|\mathbf{A}\|_2$ of a network equals the largest absolute eigenvalue of its adjacency matrix. The spectral norm is used as a measure of the *size* of a network, complementing the volume.

The *diameter* δ of a network equals the length of the longest path in the network. A small value of the diameter is used in conjunction with the clustering coefficient to characterize a social network as a *small-world network* (Watts and Strogatz 1998). As a robust replacement of the diameter, the following measures are often used:

- The 90% effective diameter equals the number of edges one must take to reach 90% of all nodes, on average.
- The mean path length is defined as the average of the distance between all node pairs.

Typical values for the diameter range from 4 to 6.

The list of social network statistics is much longer, and new statistics are constantly introduced in the literature.



Social Network Datasets, Fig. 2 The simple and cuplots are shown on a doubly logarithmic scale. (**a**) Degree mulated degree distributions of a subset of the Facebook distribution. (**b**) Cumulated degree distribution social network dataset from Gjoka et al. (2010). Both

Network Dataset Analyses

Social network datasets are used as the basis for a large number of analysis types. In this section, we review several very common types of analyses.

The *degree distribution* represents the distribution of the degree values, i.e., the number of neighbors in the social graphs, over all nodes in the networks. The degree distribution can be visualized in several ways, of which the most common is by far the simple degree distribution plot, and another is the cumulated degree distribution plot. Figure 2 shows the two types of plots for a subset of the Facebook social network (Gjoka et al. 2010). Both plots use a doubly logarithmic scale. Both degree distribution plots are typically used to point out a *power law*, i.e., the observation than the number of nodes with degree n is proportional to $n^{-\gamma}$ for a constant γ.

The second plot type we show is the hop plot. The hop plot shows, for each possible distance n, the average number of nodes at distance n from any nodes in the network. The hop plot can be used to read out the diameter, mean average path length, and 90% effective diameter of the network. The plot can also be used to measure the median path length in the network, which corresponds to the 50% effective diameter. The hop plot is expensive to compute. Figure 3 shows the hop plot of the online social network of users of an online community of students from the University of California at Irvine (Opsahl and Panzarasa 2009).

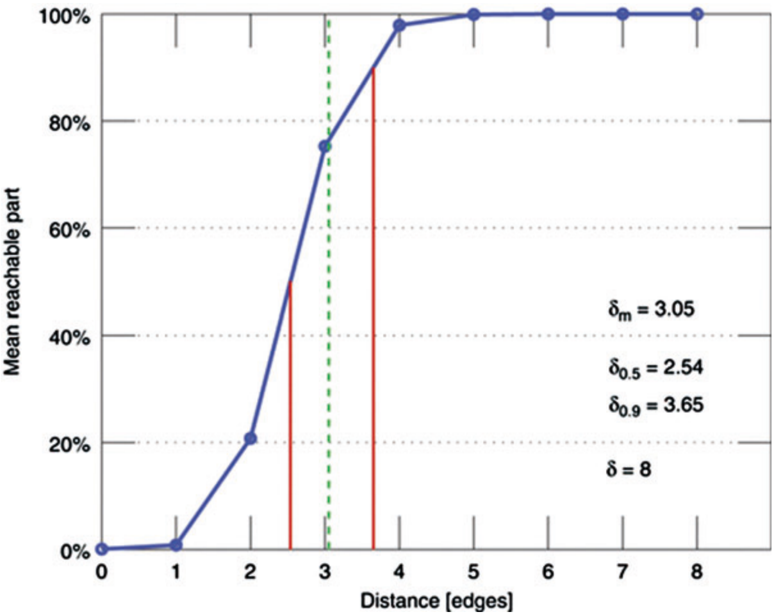
Directed Network Datasets

Some social networks have directed edges. An example are trust networks: The fact that person A trusts person B is independent of the fact the person B trusts person A. Thus, trust networks are directed and have directed edges.

Mathematically, directed networks are written as $D = (V, A)$, in which D stands for *digraph* (an abbreviation of *directed graph*) and A is the set of arcs (or *directed edges*). A directed edge between nodes i and j is usually denoted.(i, j) in contrast to the notation $\{i, j\}$ used for undirected graphs.

As an example for statistics specific to directed networks, the largest connected component can be extended to the largest strongly connected component. The strongly connected component of a directed network is defined as the largest set of nodes in the networks in which every node is reachable from every other node, using only directed paths.

In directed networks, two degrees are defined: the outdegree and the indegree. Thus, in addition to the usual degree distribution, the outdegree



Social Network Datasets, Fig. 3 The hop plot of the online social network of users of an online community of students from the University of California at Irvine (Opsahl and Panzarasa 2009)

distribution and the indegree distribution can be defined. An example of an analysis in which both distribution behave differently are power laws: Indegree distributions follow much more often power laws than outdegree distribution.

Another key feature of directed networks are reflected in algebraic graph theory, i.e., those methods that represent the social network as a matrix. In an undirected social network, the adjacency matrix $\mathbf{A}$ defined as $\mathbf{A}_{ij} = 1$ when $\{i, j\}$ is an edge and $\mathbf{A}_{ij} = 0$ when otherwise is symmetric. In directed networks, the matrix $\mathbf{A}$ is not symmetric. Therefore, methods based on its eigenvalue decomposition must be modified. In an undirected network, the adjacency matrix can be decomposed as $\mathbf{A} = \mathbf{U}\Lambda\, \mathbf{U}^{\mathrm{T}}$, in which Λ contains the real eigenvalues of $\mathbf{A}$. In undirected networks, this is not possible, and it is necessary to use either a non-orthogonal eigenvalue decomposition (leading to complex eigenvalues) or another matrix decomposition altogether.

Bipartite Network Datasets

A bipartite network is a network in which the set of nodes V can be partitioned into two sets V_1 and V_2 such that all edges connect a node in V_1 with a node in V_2. Social networks themselves are rarely bipartite. However, other networks extracted from social media are bipartite, for instance, user-item rating graphs or user-group inclusion graphs.

In bipartite networks, the clustering coefficient c is trivially equal to zero, since a bipartite network contains no triangles. Other network statistics and analyses must be extended to be used. In many cases, a statistic can be computed for the nodes in V_1 and V_2 separately. For instance, the average degrees of nodes in V_1 and V_2 can be defined. As another example, the largest connected component in a bipartite network contains a certain number of node from each of V_1 and V_2.

Network Datasets with Multiple Edges

In some social network datasets, multiple edges are allowed. An example is an email communication network, in which the nodes are the users and each edge represents a sent email. In these types of networks, also noted $G = (V, E)$, E does not represent a set but instead a multiset. Thus, two nodes, i and j, can be connected by multiple edges, for instance, denoting multiple emails that have been sent. Analogously, directed networks with multiple edges can be defined.

Most network statistics can be applied to networks with multiple edges without problem. For instance, the degree is defined as the number of edges adjacent to a vertex, counting multiple edges as such. The resulting degree distributions, as an example, can be tested for power laws.

When representing a social network with multiple edges as an adjacency matrix, the multiplicities are used as entries. In other words, the entry $\mathbf{A},{}_{ij}$ is defined to equal the number of edges between i and j, even when no edges connect the two nodes. The resulting adjacency matrix can be used in nearly all algebraic graph theoretical methods used for simple graphs.

Signed Network Datasets

Some social networks contain both positive and negative edges. An example are social networks with friendship and enmity links, such as the social network from the Slashdot technology news Web site, in which users can mark other users as *friends* and *foes* (Kunegis et al. 2009).

Mathematically, a network with positive and negative edges is modeled as a signed graph $G = (V, E,\sigma)$, in which σ is the sign function, mapping the edges in E to the set $\{-1, +1\}$.

The extension of social network statistics to signed graphs is not trivial. Using the example of the degree, each vertex of a undirected signed graph can be defined to have two degrees: the positive degree $d^+ (i)$ counting the number of positive incident edges and the negative degree $d^-(i)$ counting the number of negative incident edges. Another way of defining degrees consists in subtracting the number of incident negative edges from the number of positive incident edges, giving the signed degree

$$d(i) = d^+(i) - d^-(i).$$

Analogous signed definitions can be given, for instance, for the clustering coefficient (Kunegis et al. 2009).

The adjacency matrix of signed graphs is typically defined as a $-1/0/+1$ matrix $\mathbf{A}$ defined as $\mathbf{A}_{ij} = \sigma(\{i, j\})$ when $\{i, j\}$ is an edge and $\mathbf{A}ij = 0$ otherwise.

Rating Network Datasets

Rating network datasets are networks in which the edges represent ratings. As an example, if users of a dating site can rate other users, the resulting social network is a directed rating network between users. The much more common case however is that of bipartite rating networks, in which users rate items, for instance, movies (GroupLens Research 2006), songs (Yahoo! Labs 2011), or jokes (Goldberg et al. 2001). Ratings are typically numerical and given on a *rating scale*, the most common one ranging from 1 (dislike) to 5 (like).

To extend network statistics to rating networks, the ratings can be used as weights. However, care must be taken. in the example of the 1-to-5 rating scale, since the adjacency matrix is defined to contain the value zero for node pairs that are not connected, this would imply that a dislike of weight one counts as more than no rating of weight zero. Thus, it is typical to subtract the overall mean rating from all rating values and use the resulting numbers as weights in the adjacency matrix. Since the resulting matrix contains positive and negative values, a rating network can always be interpreted as a signed graph.

Temporal Network Datasets

A common type of study in social network analysis consists in observing the evolution of a network. in order to observe the evolution of a network, temporal information must be known. in the simplest case, edge arrival times are known for all edges, allowing one to reconstruct the network at any timepoint. All network statistics mentioned before in this article can be analyzed temporally, by computing them in function of time. The result can give insight into the processes of graph evolution. As an example, several network statistics which capture the notion of *diversity* of a network in different ways have been shown to decrease over time in a majority of social and other networks (Kunegis et al. 2012).

Multirelational Network Datasets

Signed and rating networks can be generalized to multirelational networks. in multirelational networks, any number of edge types are allowed (Greene and Cunningham 2009). For instance, edge types can be *friend, relative*, or *coworker*. Multirelational networks may be alternatively called *heterogeneous networks*.

Since the meaning of the edge types depends on the specific network, no simple generalization of network measures to multirelational networks is possible, beyond ignoring edge types. if the strength of each relationship type can be assessed, these values can be used as weights to compute the degree of nodes or as entries in the adjacency matrix. Another complication with multirelational networks are the structural properties of the various relationship types, which can vary. For instance, one relationship type can be directed, while another one is undirected.

Although multirelational social network datasets are available from various sources, only very few studies consider these types of networks generically. One example is given in Lippert et al. (2008).

Typed Networks

A further extension of multirelational networks are typed networks, in which in addition to multiple edge types, multiple node types are allowed. An example is given when a social network is combined with a user-item rating network. Such a network contains users and items as nodes, and ratings and friendships as edges. Each edge type must thus connect two nodes of a given type. The edge and node types of a typed network can be summarized by an entity-relationship (ER) diagram. A bipartite network, for instance, can be modeled as a typed network in which the entity-relationship diagram consists of two nodes connected by a single edge. Typed networks are often used but seldom modeled as such. Examples of studies using typed networks are those combining social and collaborative recommenders (Adomavicius et al. 2005).

A further generalization of typed networks results in semantic networks, whose only constraints are that it consists of triples, and in which the fundamental difference between nodes and edges is removed at lower level and modeled as part of the network itself.

Practical Considerations

Several practical issue have to be dealt with when using social network datasets. First, a social network dataset that may be incomplete are biased due to the way it was aggregated. Then, legal considerations may be necessary for using datasets. Finally, varying data formats may affect usage.

Bias Due to Data Extraction

An ideal social network dataset is generated directly from the database of a social network company. Such a dataset is complete, and all statistics computed with it reflect that actual social relationships among the users. in practice however, most social network datasets are crawled by scientists from the social networking sites. Thus, they may be incomplete, corrupted, and reflect different parts of a social network at different timepoints. These biases can have a drastic effect in analyses performed on them. For instance, if degree distributions are studied in a social network where users with zero friends are excluded due to the way the data was crawled, the resulting average degree will be wrong. Other statistics however will not be affected, for instance, the diameter of the network.

Typical biases in social network datasets are the exclusion of nodes with small degree, the omission of everything except the largest connected component, and the fact that parts of the network were crawled at different times, resulting in a social network dataset that has never existed in that form at any timepoint.

Legal Considerations

Due to the sensitive nature of social networks, most social networking companies do not publish their datasets. Thus, datasets are usually crawled, putting the publication and usage of these datasets in a legal gray area. As an example of a large dataset of the Twitter social network which included user names was retracted from its Web site from the researcher that was involved, due to complaints from Twitter. Nevertheless, many social network datasets are available online, and many studies are performed on them. Well-known newly created social networks are crawled soon after they gain a sizable market share, as shown by the example of Google Plus (SchiöBerg et al. 2012).

Data Formats

There is no unified data format for the publication of social network datasets. The formats that are used can be classified into those that try to be efficient, those that try to make it easy to combine the datasets with other datasets, and those that make it easy to access the dataset from a large number of programming languages and environments.

An example of an efficient format, both in terms of runtime and memory usage, is the binary format used by Boldi and Vigna (2004). An example of a format that makes it easy to combine a social network with other types of data is given by all social networks published as RDF. An example of social network datasets published in a format that is optimized for easy access from many programming languages is given by the tab separated value format used in KONECT (Kunegis 2013).

Key Applications

Applications of social network datasets are too numerous to cite and cover almost all aspects of data mining, information retrieval, recommender systems, Web science, and increasingly social sciences such as sociology.

Future Directions

New applications of social network datasets are published continuously. New network datasets are also published regularly. A trend in the recent years has been the aggregation of social network datasets into collections, for instance, in the Stanford Network Analysis Project (SNAP) (Leskovec 2010) and in the Koblenz Network Collection (KONECT) (Kunegis 2013). Another trend is the migration toward more interoperable formats, in line with the Link Open Data initiative.

Cross-References

- ► Linked Open Data
- ► Sources of Network Data
- ► Web Archives

Acknowledgments All plots in this article are from the KONECT project (Kunegis 2013). The author of this work has received funding from the European Community's Seventh Frame Programme under grant agreement no. 257859, ROBUST.

References

Adomavicius G, Sankaranarayanan R, Sen S, Tuzhilin A (2005) Incorporating contextual information in recommender systems using a multidimensional approach. ACM Trans Inf Syst 23(1):103–145

Backstrom L, Boldi P, Rosa M, Ugander J, Vigna S (2012) Four degrees of separation. In: Proceedings of the web science conference, Evanston, pp 45–54

Boldi P, Vigna S (2004) The WebGraph framework I: compression techniques. In: Proceedings of the international World Wide Web conference, New York, pp 595–601

Fiedler M (1973) Algebraic connectivity of graphs. Czechoslov Math J 23(98):298–305

Gjoka M, Kurant M, Butts CT, Markopoulou A (2010) Walking in Facebook: a case study of unbiased sampling of OSNs. In: Proceedings of the conference on computer communications, San Diego, pp 2498–2506

Goldberg K, Roeder T, Gupta D, Perkins C (2001) Eigentaste: a constant time collaborative filtering algorithm. Inf Retr 4(2):133–151

Greene D, Cunningham P (2009) Multi-view clustering for mining heterogeneous social network data. Technical report, University College Dublin

GroupLens Research (2006) MovieLens data sets. http://www.grouplens.org/node/73

Kunegis J (2013) KONECT – the Koblenz network collection. konect.uni-koblenz.de

Kunegis J, Lommatzsch A, Bauckhage C (2009) The Slashdot Zoo: mining a social network with negative edges. In: Proceedings of the international World Wide Web conference, Madrid, pp 741–750. http://uni-koblenz.de/~kunegis/paper/kunegis-slashdot-zoo.pdf

Kunegis J, Sizov S, Schwagereit F, Fay D (2012) Diversity dynamics in online networks. In: Proceedings of the conference on hypertext and social media, Milwaukee, pp 255–264. http://userpages.uni-koblenz.de/kunegis/paper/kunegis-diversity-dynamics-in-online-networks.pdf

Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? In: Proceedings of the international world wide web conference, Raleigh, pp 591–600

Leskovec J (2010) Stanford network analysis project. http://snap.stanford.edu/

Lippert C, Weber SH, Huang Y, Tresp V, Schubert M, Kriegel HP (2008) Relation prediction in multi-relational domains using matrix factorization. In: Workshop on structured input structure output, Vancouver

Opsahl T, Panzarasa P (2009) Clustering in weighted networks. Soc Netw 31(2):155–163

Read KE (1954) Cultures of the Central Highlands, New Guinea. Southwest J Anthropol 10(1):1–43

Schiöberg D, Schneider F, Schiöberg H, Schmid S, Uhlig S, Feldmann A (2012) Tracing the birth of an OSN: social graph and profile analysis in Google+. In: Proceedings of the web science conference, Evanston

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442

Yahoo! Labs (2011) KDD Cup from Yahoo! Labs. http://kddcup.yahoo.com/

Social Network History

▶ Networks at Harvard University Sociology

Social Network Mining

▶ Research Designs for Social Network Analysis

Social Network of IoT Devices

▶ Social Internet of Mobile Things and Decision Support Tools

Social Network Privacy

▶ Anonymization and De-anonymization of Social Network Data

▶ Privacy in Social Networks, Current and Future Research Trends on

Social Network Randomization

▶ Privacy in Social Networks, Current and Future Research Trends on

Social Network Representation

▶ Sociograph Representations, Concepts, Data, and Analysis

Social Network Security

▶ Socialbots

Social Network Sites

▶ Privacy and Disclosure in a Social Networking Community

▶ Social Networking on the World Wide Web

Social Network-Based Recommender Systems

▶ Relational Network Classification and Its Applications in Recommender Systems

Social Networking

▶ Identifying Spam in Reviews
▶ Social Media and Social Networking in Political Campaigns/Movements

Social Networking for Open Innovation

Milan Stankovic, Saman Musacchio and Philippe Laublet
Université Paris-Sorbonne, Paris, France

Synonyms

Innovation crowdsourcing platforms; Open innovation social networks; Web problem-solving platforms

Glossary

Open Innovation
A paradigm stating that companies can and should use both external and internal ideas to boost their innovation. This includes both "outside-in" (calling on external knowledge for use internally) and "inside-out" (when unused or under-performing internal knowledge is promoted outside company walls) approaches. The term has been attributed to Henry Chesbrough, professor at the University of California, Berkeley (USA)

Problem Solving on the Web
Using Web user connectivity to collaborate (e.g., the 2009 Polymath Project) or answer open-problem challenges as individuals (e.g., P&G Connect, Innocentive, Hypios)

Crowdsourcing
An approach that involves outsourcing tasks to a distributed group of people, both online or offline. This can involve the mass collaboration of thousands of individuals to accomplish one overall task (e.g., the Galaxy Zoo project, which recruited over 200,000 online volunteers to classify galaxies), or the competition of thousands of individuals in an open call for solutions (e.g., the X-Prize, Innocentive, Hypios)

Social Networking
Engaging in social activities online involving but not solely based on connected platforms like Twitter, Facebook, LinkedIn, and Pinterest

Semantic Web
A Web of interconnected self-describing data structures interpretable by machines. It represents an extension of the Web of connected pages, in which data resources are connected among each other with typed links thus forming a giant typed graph

Serendipity
The discovery of relevant but unintended and unexpected facts, phenomena, and ways of thinking

Definition

Social networking for Open Innovation is the practice of using virtual social networks to identify and engage with participants, often external to a company, as part of a larger process to develop innovative solutions and products or acquire R&D.

The current methodology of online Open Innovation problem-solving platforms involves broadcasting problems to either undefined (e.g., the Web) or very specific communities (e.g., brand consumers, specific solver communities). Though this "push-out" method does produce results, it is found to generate excessive noise and limit the involvement of certain users.

However, coupling data collected on social networks (user profiles, comments) with various articles by the same users (publications, resumes) can allow the "crowdsourcer" – OI platform or company – to create an ad hoc global virtual community to address specific issues. This reduces noise, since only relevant solution providers will be identified, and increases resolution probability, as these individuals have not opted into specific communities and are wide ranging (in both areas of interest and geography).

Furthermore, and contrary to traditional "push-out" methods, these solution providers are personally contacted (via social networks, email, and sometimes phone calls) by a dedicated team to generate interest for the problem that needs to be resolved. Using this "pull-in" method, the overall success rate of problem resolution has increased significantly.

Introduction

It was in 2009 that Fields medalist Tim Gowers decided to use his blog to find a new combinatorial proof to the density version of the Hales-Jewett theorem – in other words, to solve a very complex mathematical problem. More of social experiment on his part, he decided to put the question out in the open and see how long it would take experts, collaborating online, to crack it. The Polymath project1, as it is now called, solved the problem in 37 days, with over 800 contributions from 27 people. It even led to two papers published under the name D.H. J. Polymath. It is a true example of collaborative innovation powered by online social networks.

Collaboration is in our DNA. From the study of animal groups to Georg Simmel's extensive research on social geometry, it is clear that social networks, from family units to tribes, to the nineteenth century's great urban centers, have been critical to all cultural, social, or scientific advancement. These networks are now more powerful than ever through the new tools of interconnectivity offered by the Web of the twenty-first century.

And these tools cater to every aspect of collaboration, from universal user-generated encyclopedias (Wikipedia) to sharing documents both outside (Google Docs) and inside (Share-point) company walls. Harnessing the power of effective online collaboration through blogs, forums, community networks, open problem-solving platforms, or social platforms is still a challenge today, even more so for large organizations, whose internal tools lack the connectivity enjoyed by their employees outside company walls. In fact, the flexibility, speed, and efficiency of temporary online collaborations – even those found in multiplayer games – are forcing companies to question internal processes and adapt from the outside in.

And the crux of the problem is the rapid identification of groups or individuals who are best qualified to solve any given challenge. In this paper we argue that a truly Open Innovation approach of using external social networks as basis for such ad hoc groups is key to successful problem resolution. Yet to create the most efficient groups, existing social networks must be broken down to eliminate both homophily and propinquity. Adequate distance from the subject matter, bridges, and weak ties are highly important in promoting serendipity and guarantee that novel and often surprising solutions are submitted or that transferable processes already applied in different disciplines are quickly located.

We argue that the Semantic Web is best able to help discover the most relevant keywords for identifying such individuals on existing social networks, and from traces (papers, resumes, etc.) found online. The group can then be asked to collaborate or answer an open call for solution. This can be directly handled by the company or outsourced to a dedicated problem-solving platform.

Key Points

Using the Web for Open Innovation:

- Speeds up the process of innovation
- Gives access to external know-how
- Leads to unexpected results and solutions coming from unexpected domains
- Reduces the cost of acquisition of research
- Reduces the need for managing multiple contacts with several academic actors and gives one point of access to research

Historical Background

In the increasingly competitive market that characterizes the world economy today, the need to develop innovations quickly has become a Holy Grail for many companies. The Open Innovation (OI) model emerged as a response to the limitations of traditional innovation models, involving mainly internal research departments siloed in their respective areas of expertise. The traditional model was perceived as unsatisfactory mostly in terms of efficiency and heterogeneity of solutions considered. For Henry Chesbrough, who introduced the term in 2003, "open innovation is the use of purposive inflows and outflows of knowledge to accelerate internal innovation, and expand the markets for external use of innovation, respectively. The open innovation paradigm assumes that firms can and should use external ideas as well as internal ideas, and internal and external paths to market, as they look to advance their technology" (Chesbrough et al. 2006). According to existing literature, three key processes of the Open Innovation can be differentiated (Enkel et al. 2009): "outside-in" (the use of external resources), "inside-out" (the realization of profits from the commercialization of sleeping patents), and the "coupled" (co-creation with partners). This paper focuses on the outside-in processes completed by Open Innovation platforms.

Although these three classifications are perfectly in line with current company models, the idea that Open Innovation in its simplest form did not exist prior to 2003 is a fallacy. Recent literature argues that "open" practices have been applied before in companies in different ways (Trott and Hartmann 2009). Furthermore, open calls for solutions, for the benefit of organizations and governments, have been common throughout history, from the Longitude prize (1714) to the invention of canned foods (Napoleon's Food Preservation Prize 1795) or the development of submarines (The Confederate Prize for Inventions that Sink or Destroy Union Ships 1861).

The successful application of Open Innovation practices has been well documented by companies like IBM, P&G, Intel, Cisco Systems, DuPont, Lucent, or Philips (Sari et al. 2007). What follows are a few examples of how one open innovation practice in particular, crowd-sourcing, is adopted by some of today's leading firms.

1. Branded Platforms with Corporate Needs

   Probably the most famous example of such a platform is Procter & Gamble's (P&G) Connect & Develop site. The company formulates specific needs and posts them on its website in order to invite innovators and researchers to submit potential solutions. One challenger posted on P&G's platform was the need to develop a lipstick that would glow for 4 h, much longer than today's standard lipsticks. While this is not a problem that anyone in cosmetics would find surprising, the company with such a solution would gain an essential edge over its competitors.
2. Corporate Communities for Discussion of Questions and Needs

Another approach is the one taken by Clorox with Clorox Connect: building a platform where innovators and researchers can sign up to discuss issues with employees of the company, in a forum led by a corporate community manager. The downside to this type of platform is that it competes with specialized science wikis, forums, or specialized social networks like Research Gate, Academia, or university intranets.

3. Communities for Customer Co-Creation

   As opposed to (2), this type of platform addresses customers or fans of a company's products. An example of this was Lego Mindstorms, which during its lifetime led to the commercialization of several products. While this kind of initiative tends to lead to highly motivated and (nearly) self-driven communities, it is not applicable to every company.

4. Corporate Idea Boxes

   Shell's Gamechanger exemplifies this type of platform. While it is rather successful, all ideas that have led to products and process improvements have come from inside the company.

5. Platforms That Centralize Open Problems

   As opposed to corporate platforms, websites like Innocentive or Hypios list problems from a number of companies. This attracts individuals who are generally interested in solving problems, wherever they may arise. For a researcher who wants to maximize the chances of finding a problem that he/she can solve, such platforms are highly attractive. These emerging open innovation platforms are trying to leverage Web technology and most notably its social aspects to help innovation occur faster and more efficiently. Those services rely on social networks to diffuse innovation challenges, engage with experts, and boost collaboration.

Web Technologies for Open Innovation

Hypios, a Web-Based Marketplace for Solutions

Hypios, a French solution marketplace launched in 2009, best exemplifies the latest methods available for Web-based innovation. Companies with R&D problems (called seekers) use Hypios to externalize their problems to an ad hoc group of experts (called solvers), who then submit novel and often unexpected solutions. Karim Lakhani, HBS professor and leading academic expert on the subject, calls this method "problem-broadcast." R&D departments usually have expertise in a specific area and approach problems from a certain perspective. Yet it is evident that across the world – thus somewhere on the Web – there are people with different perspectives who can approach the problems differently and suggest truly novel solutions. The goal of a marketplace for solutions is to ensure that R&D problems reach the right people on the Web. One of the initial observations made was that companies constantly reinvented the wheel, simply because they didn't know where to look for existing plans for wheels – or because they were too scared that their competitors could find out that they were working on the wheel. Yet the truth is that most of their competitors are also working on the wheel. In the words of Kevin McFarthing, who implemented Open Innovation at Reckitt Benckiser: "R&D problems that would surprise your competitors are very rare."

The people-centric approach of Hypios makes it possible to identify explicit solutions (e.g., in publications or patents) as well as "incorporated solutions," ones that have not been made public but that can be provided by individuals if you ask them. The ability to find such "sticky" and implicit knowledge is a key advantage of identifying people rather than existing explicit solutions.

Semantic Web Technologies for Open Innovation on the Web

Although there have been studies on online search (Parkes 2007), the work done in relation to the use of social networks and new technologies by innovation intermediaries are far and between. What is of particular interest to our research is that the Social Web and Semantic Web are powerful tools that can be used for building and maintaining relationships of dispersed social communities and thus create and expand networks to produce synergies through combined interactions of users (Breslin et al. 2009). While the Social Web has already been introduced, the Semantic Web, on the other hand, should be clearly defined.

The Semantic Web is an extension of the current Web in which any content is tagged with a more precise meaning to enable machines to process it and thus better answer human queries.

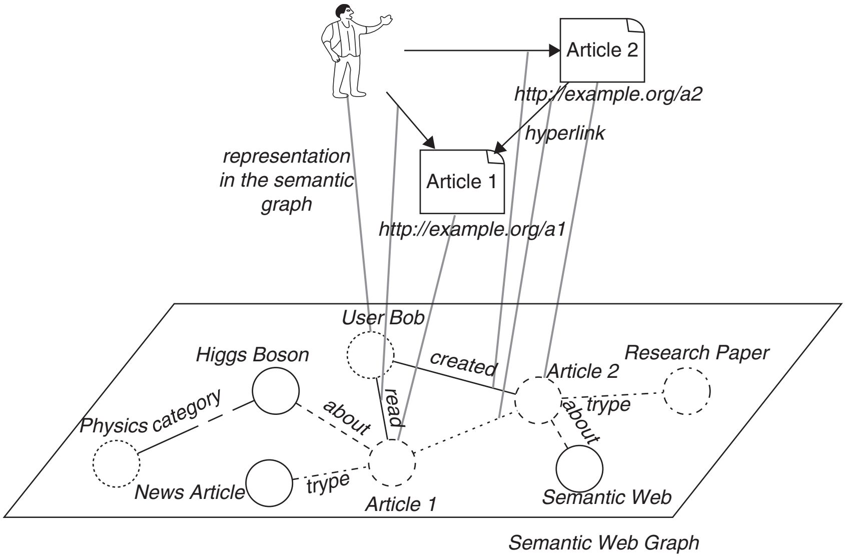
In addition to the plain textual, visual, and additive content that dominated the initial Web, the Semantic Web creates structured, self-describing articles that are published and interconnected on the Web. In this section, we will discuss how Semantic Web technologies, as a complement to the Social Web, can be used to enhance the problem-solving processes of Open Innovation platforms.

Expert Identification

The possibility of using the Web as a source for identifying experts has already been explored in literature. Web resources that users create or interact with have been used to assess expertise for such tasks as human resource management and finding assistance in e-learning scenarios.

Recently, a new trend has emerged in regard to how data is published on the Web: Linked Data (Bizer et al. 2009), which is now seen as an integral part of the Semantic Web. In contrast to representing data in the form of regular Web pages, Linked Data publishes information in a more structured format with a semantic overlay. Linked Data publishing is increasingly widespread, and even large data providers like Facebook are turning to specific forms of semantic data representation standards (http://developers.facebook.com/docs/opengraph/).

Several possibilities of currently available user data in Linked Data form have already been explored to identify experts (Stankovic et al. 2010), and further improvements will be made once data publishers accept richer forms of expressing expertise-related data (Aleman-Meza et al. 2007). The benefits of the Linked Data formats are primarily in the rich structure of typed nodes and links. When user activities, such as interaction with Web content, are represented as Linked Data, the result is a rich structure of traces, which clearly identify the users' interests and knowledge. In the example represented in Fig. 1, we can see a user interacting with two articles on the Web. The user read Article 1 and created Article 2. Modern Social Web applications using Semantic Web standards would store and selectively publish data about those activities in the form of a Semantic Web Graph. In this graph, a node representing a user would be connected to nodes representing Web content, which are further connected to nodes representing topics of interest. The relationship of the user with the content would determine the strength and the nature of his relationship with the topics of the content. Given the diversity of those links, it is possible to weigh the importance of certain topics



Social Networking for Open Innovation, Fig. 1 User activities and their traces in the Semantic Web Graph for a given user differently in different situations. For instance, in our Fig. 1, when searching for user qualifications, we would consider the topic "Semantic Web" more important, because the user created content on this topic, and when searching for his interests, we would likely pick "Higgs Boson," because the user read content on this topic. In our previous research, we have constructed a system Hy.SemEx (Stankovic et al. 2011) that relies on the diversity of link types to deliver a better expert identification engine, adapted to different needs and different situations.

Identifying experts for Open Innovation practices (especially when dealing with the Hypios platform) is different than for simple HR needs or similar queries. For OI, it is essential to find potential problem solvers that are not necessarily the best-ranked experts, with rich expertise in the given problem area (Jeppesen and Lakhani 2009). It is therefore especially important to adapt the way experts are selected in order to create a broad base of individuals with relevant, yet distant areas of expertise from the context of a specific innovation problem.

This is how Semantic Web-based expert identification technology can be used to reach the outside solvers, including the ones on the margin of the area in question, and to encourage the transfer of knowledge between those fields. In doing so, it enables "cross-sectorial problem solving." Likewise, by identifying experts in peripheral fields, this technology helps determine the graph of social behavior between relevant solvers on the Web and thus identifies "weak ties."

Semantic Keyword Discovery

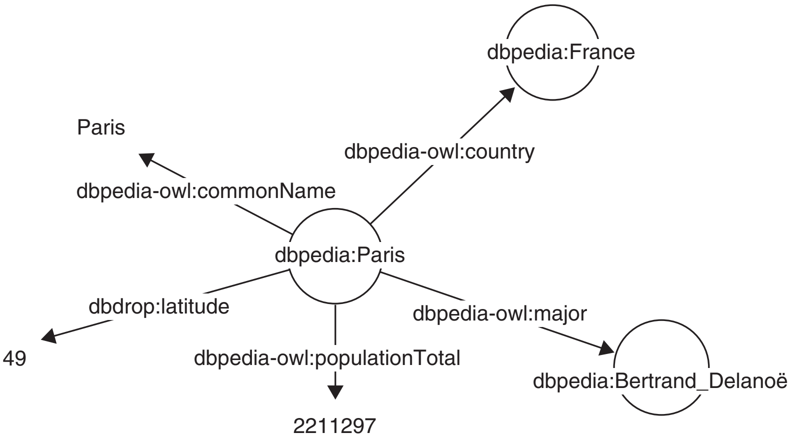
Semantic keyword discovery extends the standard matching of documents by keywords, with a notion of semantic proximity of keywords. By going beyond exact matches, it enlarges the space of possibilities. This particular property of semantic keyword matching fills a real need in Open Innovation models.

Different communities use different words to express the same or similar concepts. Thus coming from one community of practice and using one's own words to express an innovation problem heavily limits its reach in different areas. Present technologies exist to find synonyms and words of similar meaning, based on taxonomies of concepts (Ziegler et al. 2006) and word cooccurrence (Cilibrasi and Vitanyi 2007). Such existing approaches have limitations, however, as they focus on providing relevant suggestions and often neglect the need for serendipity and discovery that are essential to OI scenarios. Novel approaches that use Linked Data sources, such as DBPedia.org, to make meaningful connections between concepts in the area of music (Passant 2010) and enable the discovery of unexpected, but relevant concepts give hope that such sources might also serve to establish a notion of semantic proximity of concepts that would be more open to serendipity.

When using a problem description to identify profiles of potential experts, semantic matching makes it possible to find experts who not only work in the exact discipline of the problem but also in areas that are semantically relevant. One of the primary motivations for using a broader matching approach is to decontextualize the problem from its context and thus from the language prevalent in a particular area of expertise and increase the diversity of submitted solutions promised by a truly OI process.

For example, let us imagine an innovation problem related to detecting cable joints underground. Solving the problem requires expertise that would allow one to construct a device capable of precisely detecting cable joins by scanning the surface of the ground. A standard approach for finding experts capable of solving this problem would be to consider different keywords related to cables, electricity, and metal detection. However, a semantic approach to keyword discovery might provide less expected keywords, for instance, those related to the bone and vascular joints in the body. Those topics might seem unexpected, but in fact experts working on medical scanning equipment might have knowledge that can be transferred to the cable joint problem.

In order to deliver the functionality of semantic keyword discovery for its problem-solving platform, Hypios has developed a unique solution based on Semantic Web data structures, called Hy.Proximity (Stankovic et al. 2011). Hy.Proximity uses DBpedia, a semantic version of Wikipedia,



Social Networking for Open Innovation, Fig. 2 A part of DBpedia graph

to find concepts related to a number of initial concepts of interests. DBpedia is composed of a rich structure of concepts and their typed links. In a small part of the semantic graph featured in Fig. 2, the concept "Paris" is connected to the concept "France" with the link of type "country." When calculating the proximity score of two nodes, Hy.Proximity takes advantage of the different types of links that connect them and gives them different weights. In addition to structural futures of the graph, the semantic nature of links and nodes open numerous possibilities for constructing fine-tuned recommender systems. Hy.Proximity exploits those possibilities to deliver concept recommendations that are both relevant and unexpected for the user, encouraging the discovery of then unknown, but relevant knowledge.

We have compared the performance of our system against state-of-the-art keyword recommendation approaches. While an exhaustive evaluation has already been made (Damljanovic et al. 2012), we present here a couple of results and examples to illustrate the usefulness of the Semantic Web-based approach. For instance, we compared our system, Hy.Proximity, with the AdWords tool (https://adwords.google.com/o/KeywordTool) for keyword suggestion used to help advertisers better design their online promotional campaigns. A main difference of the Semantic Web-based approach used by Hy.Proximity is that AdWords applies a statistical approach, looking for words that co-occur in search queries and Web documents. In our evaluations, the two systems performed similarly on relevance of their proposed keywords, but in terms of unexpectedness of relevant suggestions, Hy.Proximity outperformed AdWords.

To illustrate the usefulness of suggestions that Hy.Proximity provides, what follows is an example of keyword suggestions obtained from our system and from Google AdWords. We have run both systems to obtain keyword suggestions that would help us advertise an innovation problem to an audience of experts, potential problem solvers. The problem in question deals with Kaolin extraction and issues with current mining techniques. The initial keywords we used for suggestions were Kaolinite, Drying, Mining, Separation process, Settling, Filter press, Brazil, Mill (grinding), Tailings, and Redox. The suggestions provided by the two systems are given in Table 1 While the keyword obtained through the treatment of the semantic DBpedia graph (Hy.Proximity) concerns specific topics likely to be used by experts and includes diverse topics (some topics related to mining and some related to similar processes used in other industries, such as filtering using filter paper), the topics provided by Google AdWords represent combinations of terms often used together on the Web. Their utility is more in reminding the user of known notions than in enriching him with unknown concepts. The Semantic Web structures thus play an important role in opening the audience to unexpected disciplines from which knowledge transfer can be expected – a key feature of the Open Innovation approach.

Social Networking for Open Innovation, Table 1 Keyword suggestions obtained from Hy.Proximity and AdWords

Hy.Proximity	AdWords
Drying	Dry eyes
Induced gas flotation	Dry cleaners
Souders-Brown equation	Dry shampoo
API oil-water separator	Chem dry
Dissolved air flotation	Dry tortugas
Froth flotation	Dry scalp
Aqueous two-phase system	Flights to Brazil
Gas separation	Text mining
Adduct purification	Dry-erase board
Liquid-liquid extraction	Mining companies
Acid-base extraction	Filter press
Spinning cone	Dry rot
Vapor-liquid separator	Cheap flights to Brazil
Settling	Brazil holidays
Flotation process	Brazil travel
Sublimation apparatus	Mining jobs Australia
Filter paper	Salvador, Brazil
Azeotrope	Dry ice blasting
Supercritical fluid extraction	Dry eye syndrome
Fluid extract	Dry suit

Key Applications

Using a semantic graph for identifying relevant, unexpected, or distant links can have applications in every industry – from advertising to recommendation algorithms. Yet our interest here is how useful this technology can be for solving complex innovation and R&D problems and thus accelerating research and reducing time to market.

The main problem with existing OI open problem-solving platforms is that the target group of solvers has either too much heterogeneity or too much homophily. In the first case, broadcasting a problem to random solvers will not necessarily ensure the positive and rapid resolution of a problem. Furthermore, no matter the group size, it will be limited to solvers who have opted into one of these platforms. The second case exemplifies what already plagues internal research departments. Broadcasting a problem in airplane aerodynamics to experts in the field will seldom lead to novel, unexpected solutions and discourage any technological cross-pollination.

A semantic approach for identifying experts, coupled with an outreach process, solves both problems at once. By using relevant keywords to identify experts, an ad hoc group of solvers – a network – can be created from profiles anywhere in the world, in real time, and for every specific problem. This group will have enough heterogeneity – relevant heterogeneity – to offer novel solutions, limit noise, and encourage the "systematic serendipity" of ideas.

Future Directions

Social Web technologies for Open Innovation have so far mostly addressed the field of open problem solving. However, the field of Open Innovation is much wider. Different actors in the OI world still remain to be connected and their collaboration facilitated by Social Web technologies. For instance, apart from the need to connect with problem solvers, companies that adopt OI approaches also need to connect to peer companies with whom they could codevelop innovative products. Furthermore, there is also a need for better interactions within their ecosystem of suppliers, consultants, and partners. Social networks can still play a role to encourage a paradigm shift at this level.

Social networks may also prove critical in the design of novel ways of engaging with consumers and guiding the companies' innovation towards impulses coming from social networks. Many brands already maintain an online presence via social networks to promote their products and control their image. A stronger connection between innovation research and consumer content on these social networks could prove useful in the future, as the use of social networks by customers becomes increasingly ubiquitous.

Cross-References

- ▶ Collective Intelligence for Crowdsourcing and Community Q&A
- ▶ Creating a Space for Collective Problem-Solving

- ▶ Linked Open Data
- ▶ Online Communities
- ▶ R&D Networks

Acknowledgments The work of Milan Stankovic has been partially funded by ANRT (French National Agency for Research and Technology) under the grant number CIFRE N 789/2009.

References

Aleman-Meza B, Bojars U, Boley H, Breslin JG, Mochol M, Polleres A, et al (2007) Combining RDF vocabularies for expert finding. Lect Notes Comput Sci, vol 4519. Springer, p 235. Retrieved from http://www.springerlink.com/index/p6u10781711xp102.pdf. Accessed 1 Feb 2013

Bizer C, Heath T, Berners-lee T (2009) Linked data – the story so far. Int J Semant Web Inf Syst 5:1–22. (Special issue on Linked data)

Breslin J, Passant A, Decker S (2009) The social semantic web. Springer, Heidelberg

Chesbrough HW, Vanhaverbeke W, West J (eds) (2006) Open innovation: researching a new paradigm. Oxford University Press, Oxford

Cilibrasi RL, Vitanyi PMB (2007) The Google similarity distance. IEEE Trans Knowl Data Eng 19(3):370–383. https://doi.org/10.1109/TKDE.2007.48

Damljanovic D, Stankovic M, Laublet P (2012) Linked data-based concept recommendation? Comparison of different methods in open innovation scenario. In: Proceedings of extended semantic web conference (ESWC 2012), Heraklion

Enkel E, Gassmann O, Chesbrough H (2009) Open R&D and open innovation: exploring the phenomenon. R&D Manag 39(4):311–316

Jeppesen LB, Lakhani KR (2009) Marginality and problem solving effectiveness in broadcast research. Organ Sci 20. Retrieved from http://dash.harvard.edu/handle/1/3351241. Accessed 1 Feb 2013

Parkes DC (2007) Online mechanisms. In: Nisan N, Toughgarden T, Tardos E, Vijay VV (eds) Algorithmic game theory. Cambridge University Press, Cambridge, pp 411–439

Passant A (2010) dbrec – music recommendations using DBpedia. In: Patel-Schneider PF, Pan Y, Hitzler P, Mika P, Zhang L, Pan JZ, et al (eds) Proceedings of the 9th international semantic web conference – ISWC 2010, Shanghai, vol 1380, Springer, 1–16

Sari V, Pekka S, Marko T (2007) Implementation of open innovation paradigm cases: Cisco systems, DuPont, IBM, Intel, Lucent, P&G, Philips and Sun Microsystems. Research report 978-952-214-478-2, 189, Lappeenranta

Stankovic M, Wagner C, Jovanovic J, Laublet P (2010) Looking for experts? What can linked data do for you? In: Proceedings of linked data on the web 2010, on WWW 2010, Raleigh. Retrieved from http://events.linkeddata.org/ldow2010/papers/ldow2010_paper19.pdf. Accessed 1 Feb 2013

Stankovic M, Breitfuss W, Laublet P (2011a) Discovering relevant topics using DBPedia. In: Proceedings of the web intelligence conference (WI2011), Lyon

Stankovic M, Jovanovic J, Laublet P (2011b) Linked data metrics for flexible expert search on the open web. In: Proceedings of 8th extended semantic web conference (ESWC 2011), Heraklion. Springer

Trott P, Hartmann D (2009) Why 'open innovation' is old wine in new bottles. Int J Innov Manag 3(4):715–736

Ziegler C-N, Simon K, Lausen G (2006) Automatic computation of semantic proximity using taxonomic knowledge categories and subject descriptors. In: Proceedings of the 15th ACM international conference on information and knowledge management, CIKM'06, Arlington. ACM, New York, 465–474. Retrieved from: http://doi.acm.org/10.1145/1183614.1183682

Social Networking in the Telecom Industry

Laurent-Walter Goix[1] and Fabio Luciano Mondin[2]
[1]Telecom Italia S.p.A, Milan, MI, Italy
[2]Telecom Italia S.p.A, Torino, TO, Italy

Synonyms

Federated social web; Rich communication suite; Telco operators

Glossary

ENUM	E.164 NUmber mapping
FSW	Federated social web
GSM	Global system for mobile communications GSMA GSM association
IETF	Internet engineering task force
IM	Instant messaging
IMAP	Internet message access protocol
IMS	IP multimedia subsystem
IP	Internet protocol
MMS	Multimedia messaging service
OMA	Open mobile alliance

S

OTT	Over-the-top
POP	Post office protocol
PSN	Personal social network
RCS	Rich communication suite
SIP	Session initiation protocol
SMS	Short message service
SMTP	Simple mail transfer protocol
SN	Social network
SNEW	Social NEtwork web
VAS	Value added services
W3C	World wide web consortium

Definition

Soon after the rise of the Social Web on the Internet, pervasiveness, in particular mobile access, has been fostering the adaptation – and evolution – of the entire telecom industry. In this essay we illustrate how mobile devices, and consequently telco networks, have been tremendously evolving to support this trend through various approaches from competition to cooperation, backed up by examples of massive societal usage. We also report the continuum flow of technical activities ongoing in that area currently moving to the concept of federation of social networks, or interoperability, through standardization efforts, specification work from the industry and the Web community, and the deployment of early solutions and open-source projects.

Social Networks (SN) have introduced a new paradigm of communication/content exchange between users that have tremendously boosted the telecom industry over the last years through the massive adoption of smartphones and the explosion of broadband and mobile Internet accesses worldwide.

However, although widely used from mobile devices (e.g., 50% of monthly Facebook users), SN services are not using mobile assets efficiently. Together with the many over-the-top mobile applications available nowadays, they are harming mobile network infrastructures due to heavy signalling traffic.

The related ecosystem is growing fast, driven mainly by the Web/enterprise industry, and moving towards standardization and regulatory institutions. Federation of Social Networks is the future of the Social Web that is expected to create brand-new business opportunities based on interoperable communities of all kinds.

This evolution can also be seen as a concrete potential opportunity for operators to leverage (back) their customer base relying on the phone number as a trusted user identity and on its reputation to protect customer privacy and ensure "data portability." At large scale, the success of such a service depends on the federation/peering amongst operators, at least at national level, as for the GSM service.

Introduction

While chat rooms (remember IRC) and instant messaging were the first services that gave birth to users and identities (although usually fake) within the Internet, their real-time constraints have been superseded by the advent of the current popular social networks and their chronological stream of activities. The epoch-making turn was made when Facebook launched the news feed in September 2006 (Marshall 2006) where users could see what their friends were doing at latest....

With a new (tele)communication paradigm that would sweep away real-time communication patterns (although still used to some extent) allowing people to keep in touch with friends anytime anywhere, with no need for contemporaneity, the "always-on" reachability of the wall de facto created a virtual representation of each user.

Since then the "wall" has become for its owner a history of private/public activities giving control of their outreach in an implicit manner: with respect to traditional (telco) systems, target audience is not defined one by one but grouped together in a list or circle, which is resolved by the central (dispatching) entity.

It appears clear that Social Networks have been facilitating many-to-many communications with respect to telco messaging systems (e.g., SMS). Besides, considering that posting a message on a

Social Network has little to no cost one can understand why this mechanism allowed the massive widespread of information at a worldwide level. In some cases a single post can reach millions of users able to interact with each other up to the point of drastically impacting the society, like for earthquake prevention or popular revolutions. Such scenarios are later described in this essay.

Key Points

The knowledge user relationships, together with the real user identity, have become over the years the most valuable artifacts of the Social Web and has led to many battles between telcos and OTTs on how to leverage, or obtain, this information.

Furthermore the operator's assets such as network-based user authentication and location or mobile push mechanisms are key elements for which applications, and more recently device operating systems, have been designing and implementing alternative solutions that in some cases still are suboptimal and harm the operator's network infrastructures and the device battery through heavy signalling, awaiting standards more friendly to the telecom industry.

In parallel the explosion of walled-garden Social Networks has fragmented, and in some cases replicated, users and their relationships based on their interests, unavoidably calling for interoperability (also called "federation") of social networks to avoid isolation (and the failure of closed social network tentatives of telcos). This same popular trend of vertical social networks has created privacy concerns by users in trusting their service providers, which may not have proven tracking records and which has further led to some self-regulation principles by regulation authorities.

Historical Background

It is widely believed that communication can be considered a primary need for human beings, such as eating or sleeping, and that is probably why they keep searching for better ways to (tele) communicate.

In the Beginning Was the SMS. . .

Even if social networking is quite a new concept, telecommunications systems exploiting the same principles have been used for years in many different ways and contexts. The Short Message Service (SMS), which was first used in 1993 (MobilePronto 2010), despite of its simplicity can be considered the first forefather of (mobile) social networks. Somehow SMS clearly showed the need of a direct, short, effective, and asynchronous way to communicate with friends, which can be found in its closest relatives. The SMS usage exploded also as a (near) real-time service and incentivized the specification of the MMS (Multimedia Messaging Service) standard, introduced in 2002 (Mobile Phones Uk 2012) to support multimedia content including photos and animations, which never reached the same success due to early interoperability problems on devices and high costs for users.

In the meantime on the Web side, Internet Relay Chat (IRC), allowing many users to exchange text in real time in a chat room, led to instant messaging, thus making another step towards Social Networks. Instant messaging reduced consistently the number of contacts and the audience per single message, but increased their quality and, most of all, introduced the concept of a "presence" status, a virtual real-time "marker" of our online availability in the instant messenger.

The Blogger's Dream. . .

As the Internet grew (Internet World Stats 2012), IM clients such as ICQ before and MSN and Yahoo! Messenger later went extremely popular, leading to a variety of specifications aimed at standardizing and interconnecting those types of systems.

Meanwhile, the Web community was experimenting different forms of communications. Bulletin boards became the best way of discussing about very specific topics, while weblogs (later called blogs) evolved from their initial idea of "online diary" to something more related to opinion and journalism, turning each blogger into a potential Pulitzer winner. Besides the "illusion" of blogs, some things became very clear: users were becoming content "prosumers" initiating the "Web 2.0" era, but their audience was smaller than the one they imagined, maybe due to the missing link with contacts, messaging, and audience/privacy control.

The Dawn of Social Networking...

Social Networking is the form of communication fitting best this need: a profile, in which users can put their content and show with whom they are in contact.

The way to Facebook, by far the most successful social network as of writing, was started in 1997 by SixDegrees and passed through Friendster, Myspace, and Linkedin. The aces in the hole for Facebook were probably the insertion of the Facebook Wall, which somehow overtook the concept of online "presence" with an always-available "virtual presence" concept, together with the "Facebook Platform," opening the social network to third-party applications.

Diverging Interests

When looking at what is happening between telcos and over-the-top service providers (OTT), one could consider it an "epic battle between Ying and Yang." On one side OTTs are dedicated and structured to offer innovative services but may not have proven track records in managing personal information, while on the other side telco operators control the network and its assets and have a deep knowledge of their real users and their relationships but lack of rapid processes as their strategy is still focused on optimizing their network.

OTT services are usually offered over an untrusted domain by recent lightweight start-ups with very efficient and restricted process: users get to know about the specific service offered by that company, quickly subscribe, and start to use it, but there is no certainty about the real user's identity. This may also explain why Facebook, Google, and other OTT services are trying to get more user data such as the actual full name of their users or their phone number (Smith 2012). Indeed the competition is not really centered around the number of subscribers or active users to a specific service, rather to the quality of these users. Being able to profile users has become the key success factor for service providers who often request additional personal data and permissions to perform social data mining on profile and communication data. In a world in which online services are free of charge, revenues come mainly from advertising and their value increases the more it fits with the profile the advertising target.

Furthermore, OTTs count much on network capabilities that are not under their control. Usually their services benefit much of "always-on" users and have further contributed to accelerate the deployment (and subscription) of broadband and mobile Internet devices and infrastructures over the last decade (OMT 2012), also due to a viral effect amongst users & their friends.

Instead telco operators know their users very well: they get personal data when customers subscribe their contracts and also know the most active contacts of a user through voice or text communications, but usually have regulatory restrictions to leverage this data for any other purpose. Additionally, telcos are traditionally large companies (especially incumbent operators) still getting their revenue mainly from the voice service (Patuano 2012) and used to complex (thus slower) processes to accommodate high availability of their network together with regulatory compliance.

Which Solutions?

It appears clear that OTTs and telcos are nearly pulling in opposite directions. OTTs often perceive operators as "carriers": from their perspective operators should offer high-quality data connectivity to their customer and should not compete with them on services, which is typically what telcos want to avoid as the price of mobile Internet connection is lowering (also due to regulatory agencies) and so are revenues; offering high-quality, affordable services and exploiting user phone number and identity are something mobile operators perceive as an opportunity, a way to escape from the dreaded "bit pipe" fate.

One possible solution could be taking this competition to a higher level. Both OTT and telcos seem to be aware of the importance of user profiling (and of the related privacy issues), and both know very well how a complete user profile should largely exploit the user relationships. This can explain the "raison d'être" of Facebook's social graph and be even more evident looking at some used cases and applications: recently many OTT applications helped by the device evolution are trying to catch information from the user's address book. This is, for example, the case of WhatsApp (2012) in which users are identified by their phone number (requested at sign-up) and buddies are automatically discovered through the address book. The WhatsApp intuition is to use the address book as a social network: one is in contact with many people in many different social networks, but the people one really keeps in contact with is probably a part of her/his address book.

On the other side, telcos, who easily have access to the user's phone number, are trying to exploit it socially although unsuccessfully. Many operators have tried to build their own social networks, e.g., Vodafone360, Orange Pikeo, and Telefonica KeTeKe, but none of these became truly popular, probably due to the absence of integration with other social networks and from the cold start problem of achieving critical mass.

The ace in the hole could be in making social network a commodity and move the competition from the mere "existence" of the social network service to the quality of the provided service.

New Approaches to Social Networking

Over the last years new approaches have emerged from the telecom industry to position themselves with respect to OTT. This is also related to the tremendous evolution of smartphones and data plans that fostered the wide adoption of social networking on mobile. According to Microsoft, in March 2011, 91/of mobile Internet access was to socialize, and over 1/3 of FB's 600 M users used it from mobile (Microsoft 2011). In June 2012, 57/of Facebook's 950 M users were mobile according to Facebook itself.

A first cooperative approach is in recognizing SNs as the owners of the user identity, contacts, and social interactions. This approach can vary from a plain "proxying" (e.g., through aggregation) to contractual partnership (sometimes exclusive). The aggregation (or gateway) functionality is nowadays a popular feature provided by telco operators or embedded within device operating systems themselves that do not have a strong relationship with a specific SN, but rather offer their own customers to connect to their favorite SN.

Social Network Aggregation services are popular entry doors to the social activities of users having multiple accounts over the Internet. They acquire messages, status feeds, content, and friends from various stand-alone SNs and aggregate all information in one point (device and/or server). Some specific – valuable – applications/features can cause users to migrate from isolated Social Networks to an Aggregation Site/Service. For example, some of them also offer cross-posting capabilities to simultaneously update all user accounts. This has become very popular on smartphones as well where the most popular Social Networks are integrated in operating systems such as iOS or Android to offer these features as native capabilities to users. Some tentatives have also emerged to design "social smartphones" that are explicitly focused on SN interactions (Inqmobile 2012). In some cases a business alliance is established with a specific SN to facilitate access from mobile devices (such as KDDI with GREE) in Japan (Fujimura and Yamaguchi 2011), although such a tighten relationship could eventually lead to fragmentation in "isolating" those mobile users from their friends on other operators that partnered to an alternative SN. This can become particularly risky in case of homegrown SN that is more popular within a single country than global players such as Facebook or Twitter.

Yet in other cases telcos have adopted even stronger strategies, by buying an existing SN to internalize know-how. This happened with Spain's Telefonica buying the national Tuenti in

2010 (Butcher 2010) to target "local" mobile SN services for youngsters (where telco can help) and leverage an already-popular and well-established SN (contrary to other tentatives of building one from scratch) to grow further. This challenging approach aims at growing and merging the SN user base with the telco customer base (potentially also abroad) while closing it de facto to "external" users and creating isolation.

Interestingly some telcos have realized the need to evolve these approaches of providing their own SN and open to noncustomers. DoCoMo in Japan has open its community service to other Japanese operators (Akimoto 2011), and more recently Telefonica announced the global availability of Tuenti (Lunden 2012). These strategies are clearly tentatives to keep the community alive and overcome the isolation created by the customer-only approach at the time where mobile users, further assisted by number portability, are attracted by many offers to keep switching operators. While the validity of this approach still remains to be assessed as regards the long term, it clearly calls for interoperability and walled gardens end.

Indeed alternative paths have been studied over the last years for a long-term solution beyond establishing or partnering with walled-garden SNs. Such paths aim at defining standard specifications for the popular service that has now become SNs. Indeed in the meantime, SNs evolved from a niche youngsters playground to a must-have service for the telco industry now nearly to a commodity and a new global societal way of communicating and exchanging content. The telco standard community through the GSMA association has for several years been defining the RCS (Rich Communication Suite) specifications (RCS 2012) based on the SIP-based IMS infrastructure that focuses on real-time communications such as chat, file, & video sharing with other RCS-enabled users. Recently RCS was universally branded "Joyn" (Joynus 2012) and is being deployed and offered commercially by some telcos in Germany and Spain mainly. In this context, collaboration and the simultaneous launch of the interoperable service by multiple operators within the same country is essential for its success, following the lessons learned by the GSM cellular communication standard (in the positive sense) and the failure of stand-alone attempts of SNs.

In parallel, the Web community has also been moving towards standardization: while the telco industry could leverage its standardization and interoperability experience in communication services, the Web industry has been inventing the SN paradigm and improving it through various initiatives. This naturally led some large companies in the field as well as self-initiated initiatives to start building a "Federated Social Web" based on well-known Web technologies where the telecom industry is already active. This approach however is targeting the "wall-based" asynchronous & implicit communication paradigm, which is slightly different – and actually complementary – to the RCS-based communication scenarios. Eventually both these worlds will merge and some telco standardization initiatives already have been working in that sense (e.g., OMA Social Network Web). By participating in the standardization activities, the telco industry, manufacturers & telcos, can also improve the architecture and protocols to be optimized for networks and over-the-air communications by leveraging well-known assets.

The Way to Standardization (and Regulation)

Why?

While the Web is becoming increasingly social, social networking itself is heavily fragmented due to the multitude of disparate services, implementing a "walled-garden" approach as reported above. This limits interaction & sharing between users belonging to different Social Networks (SN).

Furthermore privacy problems arise as global SN providers reside in different countries than their users and, besides legal implications, may not have proven track records in managing personal information. Users are requesting to have more control on sharing their own data or for the "right to be forgotten."

Besides consumers, businesses rely on popular SN (e.g., Facebook, Twitter) to promote themselves through a Social Media strategy, in the form of pages, advertising, and other initiatives. This ensures popularity but provides limited control over the community itself, to customize, manage, or animate it, or to get statistics, besides moving users away from the enterprise's official website.

Alternatively, creating their own user community as a stand-alone website typically results in being isolated from those SNs and remains a niche with little profit expectations.

Such enterprises, but also public administrations, are demanding for self-managed communities that can maximize brand awareness and allow users to join while still be connected to their friends and other SNs.

Federation (or interoperability) is a proven solution for this type of issues and also a natural evolution of popular societal trends set by a few stand-alone competing initiatives that eventually need to collaborate. In the recent Internet history, the email communication system is a track record of such an evolution from proprietary systems (RFC808) to global standards (POP, IMAP, & SMTP to cite a few).

But interoperability is also a native concept within telecom industry (operators are interconnected for telco services, including IP-based, e.g., MMS, with well-defined procedures for global routing of phone numbers including ENUM).

The Evolution of the Social Web Ecosystem

Between 2008 and 2010 many initiatives within the Social Web have been dedicated to aggregation as a way to limit market fragmentation: FriendFeed. Such an approach is now showing strong limitations.

Starting 2010 this community promoted SN interoperability (or federation), similarly to email systems, to overcome silos and provide users back in control of their own identity & personal information. Some commercial platforms (e.g., Ning) allow users to "easily" set up their own SN in a hosted environment.

More recently, various initiatives ranging from stand-alone projects (Diaspora, Vodafone OneSocialWeb) to community based (OStatus, OpenSocial) or even commercial platforms (SocialEngine) now provide solutions to self-create & host one's own SN.

The Benefits of Interoperability

It is reasonably foreseeable that Federated Social Networks are the future of the Social Web. In this context users can communicate with each other across domains through global identifiers (whose syntax is similar to email addresses) without the need for replicating accounts. User data portability becomes easier so that users can choose their favorite social network and migrate. From a systemic perspective, such a distributed approach also provides major scaling & robustness of the overall Social Web avoiding single points of failure. For the telecom industry, such interoperability is also a benefit, besides an opportunity. For telcos it allows to leverage their existing customer base to offer SN communication paradigm, letting their subscribers interact with friends across different SN/operators similarly as with calls/SMS. By being involved in the definition of such specifications, it also allows to leverage mobile assets and ensure network optimization. For example, it can enable users to reuse their phone number as social identity or for authentication, which is seamlessly recognized and asserted by the operator's network. Furthermore by standardizing the core interaction features of the social networking communication paradigm, the migration across platforms provided by vendors should become seamless and further allow telcos to differentiate by providing specific rich features (e.g., games) beyond the "basic" interoperability. On the other hand device manufacturers can provide smartphones that can seamlessly connect to any social networking service irrespective of their provider, thus allowing users to easily switch devices & SNs.

The Current Standardization Landscape

As anticipated above, the Web community has the leading expertise on the SN world driving most of the specification work.

In particular, large enterprise software players such as IBM are leading the OpenSocial specification work and its reference open-source implementation work (Apache Shindig project), mostly targeting enterprise social containers. Similarly, Google (who initiated the OpenSocial work), Facebook, and others are either coauthors or early adopters of some specifications related to social data models or federation protocols.

In parallel, most of the biggest Web players are involved in the related standardization bodies or industry fora such as W3C, IETF, and the OpenSocial Foundation. While the latter has long-term expertise in designing client-server specifications for Social Networking, the IETF is currently focused on refining discovery protocols and social network global identity. Within W3C several Community Groups were activated in 2012 that act as large discussion forums mainly targeting the "federation" aspects, anticipated in 2005 by the Social Web incubator group that started to investigate privacy concerns and a distributed approach (SW 2005).

Regarding federation specifically in 2010, OStatus created a Web-based specification targeted to interconnection of social networks by combining together several other draft specifications related to protocols and data models for exchanging social information. This created a de facto early reference for initial implementations from the open-source community and for the upcoming standards.

In the telecom standardization landscape, "Mobile Social Networking" (SNEW 2011) is viewed as a bridge between the SN Web community and the mobile world. OMA has been recently working at a specification called SNeW (Social Network Web) that targets this bridge with an end-to-end vision from the customer perspective.

Indeed current "mobile" version of SNs suffers from lack of mobile specificities on various aspects: frequent usage of polling instead of push notifications, no reuse of mobile identity/ authentication, poor user experience in case of loss of connectivity or roaming (differed delivery not possible), and no integration with SMS/MMS or other traditional communication mechanisms.

In addition, most of the current open specifications are not addressing an end-to-end approach: OpenSocial or OStatus are in fact focused only on a specific type of interactions (respectively client-server and server-server) with a lack of consideration for interworking of such specifications.

Towards Regulation

As described above, standardization initiatives, and the Web industry, are focusing on solutions (protocols, data models, & architectures) for social network interoperability. In this context, increasing care is given to tackle data privacy issues from a technical perspective, in particular with respect to discovery, sharing, and deletion of users' data.

Over the past years, various legal cases have been targeting SNs on leaks and breaches in managing user's data privacy, typically under the jurisdiction of the SN's home country that may bypass institutions or even violate local laws of their users.

Since 2008 the European Commission has been working with SN providers on a concept of self-regulation to overcome the duration of a European legislation process in that field. The basic idea is for SNs to self-declare their compliance with "safe principles" that target young people protection. Most of the current popular (and mostly non-EU) SNs have provided such a declaration, further explaining how they implemented it (EU-selfreg 2011). Such declarations have been assessed periodically (latest in 2011) by the European Commission through an independent assessment on nine social networking sites (EU-report 2001).

In January 2012, Viviane Reding, Vice-President of the European Commission, EU Justice Commissioner, has further announced her/his commitment to give back users the control over their personal data (EU-dataprotection 2012):

You will have an effective "right to be forgotten" so that you can remove your personal information from any site if you so wish;

Web operators must provide 'privacy by default'. The default settings for all services should be the most privacy-friendly;

You will have the right to know how your personal data will be used and where your consent is required, you must give it explicitly;

You will be able to move your personal data from one service provider to another more easily ("data portability");

Organizations processing your personal data must inform you as soon as possible if your data has been compromised;

Your personal data will enjoy the same level of protection if it is transferred outside the EU as applies within the EU – vital in this age of instant global data flows.

Although not yet effective, this statement is clearly attempting to relaunch the debate in overcoming the current limitations of the privacy laws in place in most countries regarding digital identity & related data privacy.

Key Applications

The Potential of Mobile Social Networking

It is a fact that SNs are more and more influencing our daily life as they are powerful and independent sources of information, so powerful to be used for earthquake prevention, and so independent to help in spreading news and organizing protests during the Arab spring.

Surveys (Huang 2011) showed that nearly 9 in 10 Egyptian and Tunisian used Facebook and Twitter to organize protests and get news. In such a context, rapidly evolving and changing in which the main media were under control by the government, SNs got a key role, due to their speed and independence. It has been shown (Stepanova 2011; Ellis 2011) that the Twitter updates were faster than the media updates (and widespread because of low flat costs for mobile Internet, starting $8 in Egypt).

Twitter's speed is being exploited also by another application, aiming at reducing the number of victims caused by earthquakes. It has been seen (Sakaki et al. 2012) that it is possible to use Twitter to detect target events such as earthquakes by using each Twitter user as a sensor revealing data in real time. Such an earthquake reporting system has been really developed in Japan where the earthquakes are more frequent.

Social Network Analysis of Telecom Data

The idea to consider social network services as a field of convergence for services has been already taken into account by many players. A proposal is to identify social networks over the Telco Networks (Galindo et al. 2008). Each communication media can be the starting point for a network of people, and discovering and exploiting this information can be a valuable opportunity for telco operators.

Social Network discovery can be performed by analyzing user's call and SMS history (Tomar et al. 2010) in order to understand which people in our address book users are more in touch with. The basic idea standing behind this approach is to discover the social graph underneath the network and exploit this to empower the provided services.

Future Directions

Nearly related to the concept of FSW stands the idea of Personal Social Networks (PSN): once a technology is able to offer users interconnected social networks, there is theoretically no constraint on the dimension of the social network.

The idea standing behind PSN is to have a trusted environment for user's data. The user publishes his/her data on the personal social network, and the federation becomes a way to share data with users belonging to different SNs (personal or not). The advantage is that users can publish their data on a system, which is under their direct control and thus are free to turn off at any time, a technology that could comply to the EU Directives about digital oblivion and data portability. Besides the directives about privacy already mentioned earlier in this essay, the European Commission has shown growing interests about this topic which is standing behind projects such as di.me (Di.me consortium 2010), related to personal services, and Societies (ICT-Societies.eu), related to community smartspaces.

In particular, di.me also relates to semantics, a popular research topic beyond social networks, where a precursor can be seen in SMOB (Passant et al. 2008) as early semantic microblogging tool. The basic idea is to have any social information semantically described in a machine understandable language (such as RDF). This gives the possibility to augment content with external content (e.g., provided by Linked Open Data) and thus to provide users the content they are really searching for through enriched semantic queries (Rodriguez et al. 2012).

Cross-References

- ▶ Actionable Information in Social Networks, Diffusion of
- ▶ Assessing Individual and Group Behavior from Mobility Data: Technological Advances and Emerging Applications
- ▶ Community Evolution
- ▶ Connecting Communities
- ▶ Exchange Networks
- ▶ Futures of Social Networks: Where Are Trends Heading?
- ▶ Interorganizational Networks
- ▶ Linked Open Data
- ▶ Location-Based Social Networks
- ▶ Location-Based Online Social Networks: Location-Based Online Social Media, Location-Based Online Social Services
- ▶ Misinformation in Social Networks: Analyzing Twitter During Crisis Events
- ▶ Mobile- and Context-Aware Applications of Social Networks
- ▶ Mobile Communication Networks
- ▶ Multiple Social Networks, Data Models and Measures for
- ▶ New Intermediaries of Personal Information: The FB Ecosystem
- ▶ Online Privacy Paradox and Social Networks
- ▶ Personal Networks: Ties, Space, and the Internet
- ▶ Privacy and Disclosure in a Social Networking Community
- ▶ Privacy in Social Networks, Current and Future Research Trends on
- ▶ Privacy Issues for SNS and Mobile SNS
- ▶ Semantic Social Networks
- ▶ Social Communication Network: Case Study
- ▶ Telecommunications Fraud Detection, Using Social Networks for
- ▶ Trust in Social Networks

References

Akimoto A (2011) Japan no.1 cellphone carrier's official social network now opened to other two. asiajin.com, 26 Apr 2011. http://asiajin.com/blog/2010/04/26/japan-no-1-cellphone-carriers-official-social-network-now-opened-to-other-two/. Last access 20 Dec 2012

Butcher M (2010) Tuenti looks like it will go to Telefonica for $99 million. techcrunch.com, 4 Aug 2010. http://techcrunch.com/2010/08/04/tuenti-looks-like-it-will-go-to-telefonica-for-e75-million/. Last access 20 Dec 2012

Di.me consortium (2010) http://www.dime-project.eu/en/home/dime/project/contenido.aspx. Last access 20 Dec 2012

Ellis W (2011) The role of information and communication technologies in shaping the Arab Spring. POLI-340, Nov 2011. http://www.scribd.com/doc/75905411/The-Role-of-Information-and-Communication-Technologies-in-Shaping-the-Arab-Spring. Last access 20 Dec 2012

EU-dataprotection (2012) Protection of personal data. http://ec.europa.eu/justice/data-protection/index_en.htm. Last access 20 Dec 2012

EU-report (2001) Implementation of the safer social networking principles for the EU, Sept 2001. http://ec.europa.eu/information_society/activities/social_networking/eu_action/implementation_princip_2011/index_en.htm. Last access 20 Dec 2012

EU-selfreg (2011) Safer social networking: the choice of self-regulation. http://ec.europa.eu/information_society/activities/social_networking/eu_action/selfreg/index_en.htm. Last access 20 Dec 2012

Fujimura N, Yamaguchi Y (2011) Gree, KDDI Sue DeNA Amid Japan social-network competition. Bloomberg.com, 21 Nov 2011. http://www.bloomberg.com/news/2011-11-21/gree-kddi-sue-dena-amid-japan-social-network-competition-2-.html. Last access 20 Dec 2012

Galindo LA et al (2008) The social network behind telecom networks. Position paper for W3C workshop on the future of social networking. http://www.w3.org/2008/09/msnws/papers/telefonica-business-operator.pdf. Last access 20 Dec 2012

Huang C (2011) Facebook and Twitter key to Arab Spring uprising: report. The National, June 2011. http://www.thenational.ae/news/uae-news/facebook-and-twitter-key-to-arab-spring-uprisings-report. Last access 20 Dec 2012

ICT-Societies.eu. http://www.ict-societies.eu/. Last access 20 Dec 2012

Inqmobile (2012) http://www.inqmobile.com. Last access 20 Dec 2012

Internet World Stats (2012) Internet growth statistics. http://www.internetworldstats.com/emarketing.htm. Last access 20 Dec 2012

Joynus (2012) http://www.joynus.com/. Last access 20 Dec 2012

Lunden I (2012) Tuenti, Telefonica's answer to Facebook and Twitter, opens up to users worldwide. techcrunch.com, 11 July 2012. http://techcrunch.com/2012/07/11/

tuenti-telefonicas-answer-to-facebook-and-twitter-opens-up-to-users-worldwide/. Last access 20 Dec 2012

Marshall M (2006) Facebook launches “News Feed” and “Mini Feed” – as YouTube invades turf. venturebeat.com, Sept 2006. http://venturebeat.com/2006/09/05/facebook-launches-news-feed-and-mini-feed-as-youtube-invades-turf/. Last access 20 Dec 2012

Microsoft (2011) Mobile Stats 2011. Microsoft Tag, Mar 2011. http://tag.microsoft.com/Libraries/Blog/mobile-marketing-and-advertising-landscape.sflb.ashx. Last access 1 June 2012

Mobile Phones Uk (2012) What is MMS? http://www.mobile-phones-uk.org.uk/mms.htm. Last access 20 Dec 2012

MobilePronto (2010) The history of SMS text messaging. http://www.mobilepronto.org/en-us/the-history-of-sms.html. Last access 20 Dec 2012

OMT (2012) Social media vs smartphone usage in Europe. Online marketing trends, Feb 2012. http://www.onlinemarketing-trends.com/2012/02/social-media-vs-smartphone-usage-in.html. Last access 20 Dec 2012

OStatus (2010) http://ostatus.org/. Last access 20 Dec 2012

Passant et al (2008) Microblogging: a semantic web and distributed approach. In: SFSW

Patuano M (2012) Telecom Italia 1H 2012 results. Telecom Italia Webcasting. http://telecomitalia.web casting.it/1H2012-ondemand/files/SlideMarcoPatuano1H2012.pdf. Last access 20 Dec 2012

RCS (2012) http://www.gsma.com/rcs/. Last access 20 Dec 2012

RFC808 (1982) Postel J, Mar 1982. http://www.rfc-editor.org/rfc/rfc808.txt, Appendix A. Last access 20 Dec 2012

Rodriguez RO et al (2012) LODifying personal content sharing. In: EDBT conference, Berlin

Sakaki T, Okazaki M, Matsuo Y (2012) Tweet analysis for real-time event detection and earthquake reporting system development. IEEE Trans Knowl Data Eng. IEEE Computer Society Digital Library. IEEE Computer Society

Smith G (2012) Now Facebook wants your mobile number: social network to ask 900 million users for phone details to, ‘prevent’ Linkedin-style hack. The Daily Mail. http://www.dailymail.co.uk/sciencetech/article-2159672/Facebook-ask-900-million-users-phone-details-prevent-LinkedIn-style-hack.html. Last access 20 Dec 2012

SNEW (2011) OMA brings interoperability of social networks to the mobile world. http://www.openmobilealliance.org/comms/technical/msn_overview.htm. Last access 20 Dec 2012

Stepanova E (2011) The role of information communication technologies in the “Arab Spring”, May 2011. Ponars Eurasia. http://www.gwu.edu/~ieresgwu/assets/docs/ponars/pepm_159.pdf. Last access 20 Dec 2012

SW (2005) A standards-based, open and privacy-aware social web. http://www.w3.org/2005/Incubator/socialweb/XGR-socialweb-20101206/. Last access 20 Dec 2012

Tomar V et al (2010) Social network analysis of the short message service, TICET. http://www.ee.iitb.ac.in/~karandi/pubs_dir/conferences/vikrant_himanshu_karandikar_vinay_swati_prateek_ncc10.pdf. Last access 20 Dec 2012

WhatsApp (2012) http://www.whatsapp.com. Last access 20 Dec 2012

Social Networking on the World Wide Web

Qingpeng Zhang[1,2], Dominic DiFranzo[3] and James A. Hendler[4]
[1]Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong, SAR, China
[2]Shenzhen Research Institute of City University of Hong Kong, Shenzhen, China
[3]Department of Electronics and Computer Science, University of Southampton, Southampton, UK
[4]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, USA

Synonyms

SNS; Social network sites; Social networking services; Social networking sites

Glossary

Collective intelligence	Shared or group intelligence that emerges from the collaboration, collective efforts, and competition of many individuals and appears in consensus decision making (http://en.wikipedia.org/wiki/Collective_intelligence)
Crowdsourcing	The practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, and especially from an online community, rather than from traditional

employees or suppliers (http://www.merriam-webster.com/dictionary/crowdsourcing)

Human flesh search, HFS Is the phenomenon of distributed researching using Internet media such as blogs, forums, and microblog (Zhang et al. 2012)

Microblog A broadcast medium that exists in the form of blog, with content that has a typically smaller size

Social computing, computational social science Computational facilitation of social studies and human social dynamics, as well as the design and use of information and communication technology technologies that consider social context (Wang et al. 2007)

Social media The forms of Web interactions among people through which Web users create, and share information, personal messages, ideas, etc. in virtual communities and SNS

Social networking site, SNS A platform for Web users to build and maintain social networks and share interests, activities, and/or real-life connections

Web 2.0 The websites that use advanced technology beyond the static pages to enhance the social networking applications

Web science The socio-technical science of understanding the complex, cross-disciplinary dynamics driving development on the Web (http://tw.rpi.edu/web/concept/WebScience)

Definition

Enabled by the Internet, people have been connected on various social network sites. In this entry, we perform a brief review of the social network sites and services worldwide.

Introduction

The rising popularity and use of social computing technologies has not only connected people in new and interesting ways but has also generated vast amounts of data on human crowd behavior. This is allowing researchers to view and study crowds and communities at a scale never before possible. The growth of social networking sites (SNSs) has dramatically changed the way people communicate, collaborate, and maintain their social connections. Social networking on the Web has also enabled the emergence of crowdsourcing and collective intelligence sites, allowing for new economies and workflows to develop. In the past decade, SNSs have played an important role in current movements all around the world. This entry reviews the history of social network sites on the Web and summarizes research on SNSs. This entry also includes a review of Chinese SNSs, which has not been fully taken into consideration previously because of the isolation of SNSs in the Chinese Mainland.

Key Points

This entry reviews the history of social network sites and state-of-the-art research on SNSs. In particular, we review Chinese SNSs, which have not been fully covered by the literature.

Social Networking on the World Wide Web

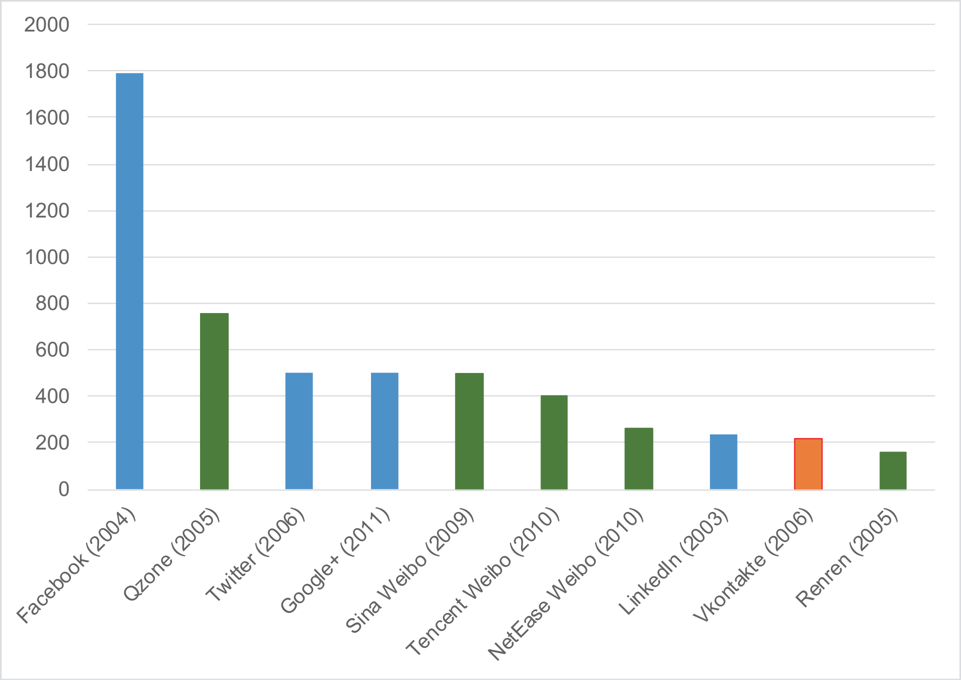
Global SNSs at a Glance

Social networking sites have become predominant in the age of the World Wide Web. The burst of social networking sites (SNSs) has dramatically changed the way people communicate, collaborate, and maintain their social connections. SNSs provide platforms and interfaces that enable people to follow and communicate with their friends, families, and other social connections. The sizes of SNSs have been growing rapidly. Boyd and Ellison reviewed SNSs in 2006 and defined SNS as “web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system” (Boyd and Ellison 2007). We feel this definition is still appropriate to describe the phenomena and covers newer SNSs (especially microblogging sites like Twitter). Figure 1 and Table 1 show the top ten largest SNSs according to the number of registered users (date of count and resources listed in Table 1). In Table 1, we use “SNS” to refer to traditional SNSs as defined in Boyd and Ellison (2007) and use “microblog SNS” to annotate Twitter and other equivalent services. Figure 2 is from the June 2017 edition of the World Map of Social Networks as reported in the Vincos Blog (http://vincos.it/world-map-of-social-networks/). It shows a map of the most popular social networking sites by country, according to Alexa traffic data (http://www.alexa.com/).

As shown in Figs. 1 and 2, and Table 1, Facebook is the largest SNS in the world, with over a billion registered users. Following, Twitter is the third largest SNS and the largest microblog site, with over 500 million registered users. These two US-originated SNSs are dominating the social networking services all over the world, except in a small number of countries, which either have very strong SNSs of their own or have limits to the access of Facebook and Twitter.

Chinese SNSs

In this entry, we use China to refer Chinese Mainland, which does not include Taiwan, and China’s Special Administrative Regions of Hong Kong and Macau. The censorship of Internet in Chinese Mainland is stricter. As of July 2013, the number of Internet users in China is nearly 600 million, which consist over one fifth of the global Internet



Social Networking on the World Wide Web, Fig. 1 Ten largest SNSs (with launched year) according to the number of registered users, as in late 2012 to early 2013. Color represents the origin. *Blue*, USA; *green*, China; *red*, Russia

Social Networking on the World Wide Web, Table 1 Ten largest SNSs

Name	Registered users (in million)	Time counted	Year launched	Original country	Type	Source
Facebook	1,150	Mar 2013	2004	USA	SNS	http://investor.fb.com/releasedetail.cfm?Rel easeID=761090
QZone	610	May 2013	2005	China	SNS	http://www.tencent.com/en-us/content/at2013/attachments/20130515.pdf
Twitter	500	Mar 2013	2006	USA	Microblog SNS	http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html
Google+	500	May 2013	2011	USA	SNS	http://googleblog.blogspot.com/2012/12/google-communities-and-photos.html
Sina Weibo	500	Dec 2012	2009	China	Microblog SNS	http://news.xinhuanet.com/tech/2013-02/21/c_124369171.htm
Tencent Weibo	400	Dec 2012	2010	China	Microblog SNS	http://www.techweb.com.cn/internet/2012-04-24/1183131.shtml
NetEase Weibo	260	Oct 2012	2010	China	Microblog SNS	http://tech.163.com/12/1018/18/8E49Q121000915BF.html
LinkedIn	238	May 2013	2003	USA	SNS	http://press.linkedin.com/about
Vkontakte	220	Aug 2013	2006	Russia	SNS	http://vk.com/catalog.php
Renren	160	Aug 2012	2005	China	SNS	http://life.renren.com/

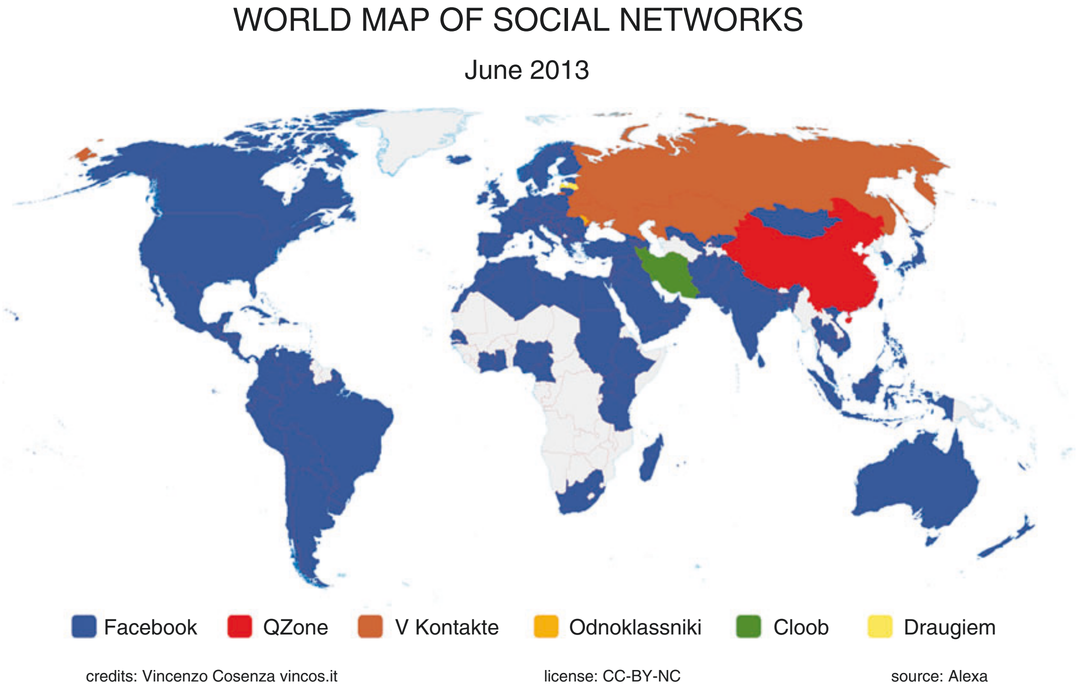
users (CNNIC 2013; ITU 2013). Among them, 78.5% primarily use mobile phone to surf the Internet (CNNIC 2013). The very large number of Chinese Internet users has enabled the birth of many SNS giants in China (five of the top ten largest SNSs in the world). As the access to several SNSs (including Facebook and Twitter) has been limited in China (most require use of a virtual private network to visit) since 2009 (Facebook was also blocked for a small time period several times before 2009), the Chinese SNSs do not directly compete with American SNSs. It is worth noting that the restriction to visit oversea SNSs is not the only reason for the growth of Chinese SNS giants. The biggest SNS in China, QZone, was founded in 2005, just after Facebook was born. In fact, a lot of foreign Internet services (like ICQ, the first Internet-wide instant messaging service in the late 1990s) were defeated by their Chinese equivalents (which were usually improved and optimized for Chinese users) due to a variety of reasons (which will be discussed later in this entry). In addition, Chinese Internet users use online forums extensively. There are also some novel and unique Chinese SNSs which have no equivalents to any foreign services (like douban.com, launched in 2005). The more inherent cultural factors of this phenomenon are yet to be analyzed.

The History of SNSs

As there is a comprehensive review of SNSs prior to 2006 (Boyd and Ellison 2007), we briefly summarize the history of SNSs before 2006 and then concentrate on a more recent activity.

Before 2000 (Early Days)

Launched in 1997, SixDegrees.com is usually cited as the first recognizable SNS. It contains the basic functions of SNS, including listing friends and maintaining personal profiles. There were millions of users in SixDegrees in the late 1990s, before it was closed in 2000. The founder of SixDegrees thought that, at that time, Internet users did not have many friends online and people usually did not want to meet strangers (Boyd and



Social Networking on the World Wide Web, Fig. 2 The most popular SNSs by country, according to Alexa traffic data (From June 2017 edition of the World Map of Social Networks; http://vincos.it/world-map-of-social-networks/; accessed September 25, 2017)

Ellison 2007). The next important SNS is LiveJournal, which was launched in 1999.

Around the same time, SNSs also emerged in Asia, for example, Cyworld in South Korea (launched in 1999, with SNS features added in 2001) and Tencent QQ, which was launched in 1999 as a Chinese equivalent of ICQ, and added SNS-type features known as QZone in 2005. Due to Tencent dominating the instant messaging field in China (about 800 million active accounts), QZone, the SNS service for QQ users, has been one of the biggest SNSs in the world since its birth in 2005 (QZone was once the largest SNS before Facebook bypassed it) (Boyd and Ellison 2007). QZone has integrated a lot of features, first as a blog site, and then music and photo sharing, and even some microblog-similar features before microblogging became popular in China. However, these features did not bring overwhelming success for QZone in either SNS or blogging, despite the very large number of users. Tencent even produced separate SNS (Tencent Pengyou) and microblog (Tencent Weibo) services to compete with other Chinese SNSs. Many of these services failed, but some eventually became popular. For example, Tencent Weibo, a microblog service, is now #6 largest SNS in terms of the number of registered users.

2001–2004 (Burst of SNSs)

Ryze.com was launched in 2001 as a business-driven SNS. Following Ryze, Friendster was launched in 2002 as a social networking complement to Ryze, aiming to help build friend-of-friend connections. Friendster was the top SNS until 2004, when it was overtaken by MySpace. Friendster found a second home in Asia, as a social gaming site, with over 90% traffic coming from Asia. After the success of Friendster, a large number of SNSs were launched. Among them, LinkedIn, an SNS for professional connections and networks, became one of the most successful and largest SNSs. At the same time, media-sharing sites like YouTube and Flickr started to incorporate more SNS-type features. MySpace was launched in 2003 and because of a variety of both technical and societal difficulties Friendster was facing (one of the major difficulty was with the ill-equipped databases), many Friendster users migrated to MySpace and other SNSs (Boyd and Ellison 2007). MySpace employed the (musical) bands-and-fans dynamic to attract both bands and fans to join in and communicate. This strategy fostered marketing in SNS. Several years later, a similar and more well-designed strategy was successfully adopted by Sina to promote their microblog service in China, Sina Weibo (will be discussed later). MySpace also allowed users to add and modify HTML elements into their profiles to generate more personalized MySpace pages. In 2004, there were more and younger Internet users joining MySpace. They joined MySpace mainly because they would like to connect with their favorite bands. The "word-of-mouth" effect quickly spread in the teens' world, and MySpace grew very fast during this period. During this time period, IT giant Microsoft also released their SNS service in 2004, MSN Spaces, which became very popular outside the USA (e.g., China). However, MSN Spaces closed in 2011.

Since 2004 (Facebook and Its Equivalents)

In early 2004, Facebook was launched as an SNS only for Harvard students. It opened to other college students and then high school students and corporate networks in 2005. Eventually, Facebook moved to an open signup (to users older than 13) in 2006. The social ties in Facebook are mutual. Users have their profile pages and news feeds for their home pages to highlight the updates of users' activities. Each user has a "wall," which summarizes updates of his or her friends. Facebook also added the "like" feature so that users can express that they like others' content. Facebook also has messaging functions, and the mobile app of Facebook enables users to communicate with their Facebook friends without having to sit in front of a computer. The major difference between Facebook and MySpace is that Facebook requires users to give their true identity (at least before the open signup). In addition, Facebook allows developers to produce "applications" and "games" for Facebook to allow users to personalize their pages and have more fun with their friends. Facebook also revealed its "Facebook Platform" (with "Graph API" as a core) in May 2007 to allow developers to read and edit the data of Facebook, especially the social graph. Facebook overtook MySpace in April 2008 and eventually became the largest SNS worldwide. It is also the most popular SNS in English-speaking countries. Facebook had 500 million users in July 2010 and quickly doubled it to one billion users in October 2012. Today, Facebook is not only an SNS for many people. It is an integrated social platform for almost everyone in many countries around the world (see Fig. 2).

Because Facebook requires users to use their real names, privacy has been a big concern. In November 2007, Facebook implemented its advertising system, Beacon. It used the data of Facebook users and advertised to friends of users using the history of purchases they made, causing a backlash of criticism. It was shut down in a month. In 2009, Facebook enabled users to choose which parts of their profile can be viewed by everyone, though the name and profile photo are always accessible to public.

Facebook is not only the largest SNS in America but also the largest SNS in Europe. However, Facebook is not dominant in Russian-speaking countries. VKontakte (VK) was launched in Russia in 2006. It was first only for college students and then opened to public. It quickly grew and became the second largest SNS in Europe.

In 2011, Google launched its SNS, Google+, after the failure of Google Buzz. Google+ has been described as a combination of Facebook and Twitter, with an aim to attract users from both sites. Google+ has its unique "circles" for users to organize their friendship information. "Circle" enables Google+ to have "social layers" and enhanced a major property that Facebook and Twitter lacked, making users' updates and messages visible to only a subgroup of their contacts, instead of pushing their information to everyone connected to them (Facebook and Twitter have since added their own variants of this capability). Google+ reached 500 million users in May 2013, making it the fourth largest SNS worldwide. However, there are many reports saying that Google+ is a "ghost town," with a large number of registered users but few activities (Gonzalez et al. 2013). Gonzalez et al. conducted a comprehensive empirical study of Google+, looking at its topological properties and evolution patterns. They found that the stable connectivity features of Google+ network were very similar to Twitter and different from Facebook, indicating that the use of Google+ was more like the messaging propagation in Twitter, rather than pairwise relations in Facebook. They also found that the user is not actively engaged in Google+ network, as compared with Twitter and Facebook (Gonzalez et al. 2013). More research on Google+ has focused on the privacy issue, taking a closer look at its "circle" function.

During the same time period, Renren (formerly known as Xiaonei, literally "on-campus network") was launched in China in 2005. It is widely known as the Chinese equivalent of Facebook. Similar to Facebook and VKontakte, Xiaonei was first only open to college students. In August 2009, Xiaonei was renamed to Renren (literally "everyone's network"), in order to expand its user size. Renren has been competing with Kaixin001 since the latter was launched in 2008. Kaixin001 first aimed at "white collars" (educated people performing professional, managerial, or administrative work in office) and then changed their strategy to compete with Renren for all users. Both SNSs have their own user groups. They have been losing active users since the birth of Sina Weibo, a microblog service. As it now stands, Sina Weibo and other microblog services are losing their active users (CNNIC 2013), with the emergency of WeChat (we will discuss Chinese SNS later in this entry in more detail).

The burst of SNSs has also attracted the attention of researchers. The rich data generated by SNSs provided ideal test-beds for research. In 2007, Wang et al. revisited the term "social computing" and gave a new definition to refer the research on both the design of social software (which was coined as "social computing" in 1994 by Doug Schuler) (Schuler 1994) and the study of social systems using computational science methodologies (Wang et al. 2007). In 2009, another similar term "computational social science" appeared in Science Magazine (Lazer et al. 2009). Computational social science refers to the second part of social computing, and both terms became popular and widely used among researchers studying SNSs.

Besides fostering the birth of social computing and computational social science, the bursts of SNSs also facilitated the growth of several other domains, including young fields like network science (Barabási 2013) and Web science (Shadbolt et al. 2013) and mature fields like data mining (Han et al. 2006), machine learning (Bishop and Nasrabadi 2006), and natural language processing (Manning and Schütze 1999). In particular, the study of various social networks formed by SNSs has been one of the most active research topics following pioneering work defining properties such as scale-free and small world networks (Barabási 2013). This research on SNSs expanded earlier small-scale (tens or hundreds of nodes and edges) survey-based social network analysis to very large scale, usually from thousands to millions of nodes and edges. The nodes in these social networks were typically a unique user ID in SNS, and the edges between nodes represented different types of social connections/interactions, including directed or undirected friendship, message exchange, and comment and reply, which normally indicate the social structure and the information propagation in SNS. During the past decade, researchers have studied almost every popular SNS, including the blogosphere, Facebook, Google+, Renren, various media-sharing SNSs, and Q&A SNSs (please refer to section "Recommended Reading" for typical publications of these SNSs). The social network analysis (SNA) studies revealed many interesting aspects of the social systems and dynamics of SNSs. We summarize a few typical research results. (We summarize the results briefly below. For more details, please refer to section "Recommended Reading" for source publications of the results.)

In most SNSs, people found that a small portion of the users were controlling the communications and information spread in SNSs, and people are easily connected with each other via "traveling" through those key users, known as hubs.

Users were clustered around different topics, and in certain events (like political elections and revolutions), users were polarized into two or more big clusters, with few interactions in between.

Researchers also conducted temporal and spatial analysis on the conversations in SNSs.

There are many successful algorithms being developed to discover the subcommunities in SNSs based on social networks.

In addition, topic models and other probabilistic models have been employed to further explore the implicit subcommunities.

Furthermore, the privacy and trust issues in SNSs have also been studied.

Researchers have conducted empirical studies of the use of SNSs in social movements and performed experiments of using SNS for social mobilization.

For behavioral and social science researchers, various theories in social network can be validated with the "big data." Among them, balance theory ("the enemy of my enemy is my friend") was one of the most intuitive and early studied theories, and it was found to hold in most SNSs.

The strength of weak ties and the relevant structural holes theory have also been validated in SNSs, showing that users in the broker position of social networks formed by SNS have the advantage to be more innovative and productive because they have access to various fresh ideas.

For details and a more comprehensive review of state-of-the-art research on SNSs, please refer to other entries of the "Encyclopedia of Social Network Analysis and Mining".

Since 2006 (Twitter and Its Equivalents)

The birth of Twitter in 2006 changed the cyberspace again. Twitter created a new form of SNS named a microblog, in which users post short messages (up to 140 characters) via the Web, smartphone apps, email, mobile phones, and instant messages. Different from other SNSs, the relationship in Twitter is not reciprocal, meaning that a user can follow other users, and a user can be followed by others without following them. The followers of a user in Twitter can view the messages (named as "Tweets") from the user. The followers can reply or retweet this user's tweets. Twitter users use @ to mention a Twitter user and hashtag # to represent a topic of the tweet.

Since its launch, Twitter quickly became one of the most visited websites as "the SMS of the Internet" and the largest microblog site worldwide (though its Chinese equivalent is close). As compared to traditional blogs and SNSs, microblogging is a faster method to communicate, share quick thoughts, and report news. In addition, the frequency of updating a microblog is usually much higher than traditional blogs and SNSs. These features made Twitter and other microblogging services distinct. In a recent review, Murthy describes Facebook as to "keep ties between users active and vibrant," while Twitter is used to seek the "accumulation of more and more followers who are aware of a user's published content" (Murthy 2013).

The use of Twitter in China is limited. Twitter was not popular in China before being blocked. The first Chinese microblog was Fanfou.com, which was launched in May 2007. The number of Fanfou users was around one million in 2009. Largely because of riots that happened in certain parts of China, Twitter and Facebook were blocked in July 2009 and have been limited in access since then. Fanfou and some other microblogs were also blocked for a while in July 2009. Chinese IT giant Sina.com grasped this opportunity and launched Sina Weibo in August 2009 (1 month after Twitter was blocked). "Weibo" means "microblog" in Chinese and Sina registered weibo.com. Therefore, people usually use Weibo to refer to Sina Weibo. Sina had its unique marketing strategy – Sina invited celebrities to sign up to Sina Weibo and communicate with their fans. This strategy worked very well and Sina Weibo quickly became the largest microblog service in China. Within a year, Twitter's other

Chinese equivalent, Tencent Weibo, NetEase Weibo, and Sohu Weibo, started to grow along with Sina Weibo. Sina Weibo's competitors also tried to pay some celebrities so that these celebrities would only use their service to post microblogs. However, the Sina Weibo community had already grown to a large number of users, who had also connected to their friends and families and constructed their networks and thus did not want to turn to another platform. Some celebrities even flew to Sina Weibo to be more visible. Therefore, although other Chinese microblogs have successfully built their own communities (which are also large scale), they do not really threaten Sina Weibo, which is still dominating the Chinese microblog world.

In the West, Facebook still has been growing since Twitter was born. People are using Facebook and Twitter for different purposes. However, in China, traditional SNSs quickly lost active users, and many of Chinese SNSs became ghost towns after Sina Weibo's launch. There is a sign that it may also happen for Sina Weibo 3 years after its birth. Tencent (the company who produced QQ and QZone) launched WeChat in 2011. WeChat was first a multimedia (text, voice, video) messaging software. However, Tencent soon added its SNS features "Moments" into WeChat. Moments is a user timeline similar to Facebook. WeChat now has over 400 million active users, and many Weibo users moved to WeChat. Although Moments of WeChat is growing very fast, currently most WeChat users are still using it solely for messaging purposes. Therefore, we do not include it in the ranking of SNSs (Fig. 1 and Table 1). According to a report by GlobalWebIndex in January 2013, the number of active Weibo and traditional SNS (like Renren) users decreased significantly in 2012, when Twitter, Facebook, and Google+ were still increasing (http://www.pingwest.com/twitter-the-fastest-growing-social-platform/). This decline is likely attributed to the changing dynamic between WeChat and its competitors.

An early and highly cited empirical study of the topology and intention of Twitter was published in 2007, finding users use Twitter to talk about their daily activities and to seek/share information (Java et al. 2007). Since 2008, due to these unique features, Twitter and other microblog services have quickly become the key social media and SNS for not only in daily conversation and chats but for news reporting (i.e., discussing breaking news, report news, political elections), business (i.e., marketing, advertising), emergent events (i.e., disasters, protests, and terrorist attacks), and social movements (i.e., Occupy Movement, Arab Spring, civil wars) as well. The recent research on SNSs has largely focused on microblogs. Another reason that microblogs are now the key datasets for research is because it is easier to retrieve data as compared to other SNSs like Facebook and Renren. The two biggest microblogs Twitter and Sina Weibo both have open APIs that allow people to retrieve all kinds of data, usually with limits in the volume of data to be retrieved or the number of requests to the server. Kwak et al. analyzed a Twitter network of 41.7 million users, 1.47 billion social relations, and 106 million tweets with 4,262 topics and conducted a series of quantitative analyses on the data to reveal the difference between the Twitter network with other SNSs (Kwak et al. 2010). The research that has been done on traditional SNSs like Facebook and MySpace has been repeated with Twitter data, and more novel research has been conducted to answer many interesting research questions that could not be answered before. People have explored whether the information diffusion seen in Twitter was due to social connections or external resources, the roles of Twitter in information diffusion, the formulation and organization of groups in protest and revolutions, emerging distributed group chats on Twitter, and so forth (please refer to section "Recommended Reading" for corresponding publications). Currently, Twitter is the most frequently used data for researchers in social computing and computational social science, and Sina Weibo is playing the same role in Chinese academia.

Since 2004 (Crowdsourcing and Collective Intelligence)

Collective intelligence is defined as the intelligence emerged from the communication, collaboration, and competition of a group of individuals. The term was first coined by sociologists, who studied the swarm intelligence of insects, birds, mammals, bacteria, etc. (Lévy and Bonomo 1999; Bonabeau 2009). With the advances of SNSs, massive collaboration among a large number of users around the world has become a reality. People can collaborate online to work on the same task and solve problems. For example, Wikipedia is "a collaborative edited, multilingual, free Internet encyclopedia supported by the non-profit Wikimedia Foundation" (http://en.wikipedia.org/wiki/Wikipedia). The 30 million articles in 287 languages of Wikipedia were written by volunteers all over the planet. Anyone has access to edit almost every article of it (Glott et al. 2010). Wikipedia has been one of the top ten most popular websites according to Alexa (http://www.alexa.com/).

Online forums were the first big platform for collective intelligence. The use of online forums for collective intelligence ranges from small-scale Q&A systems (Zhang et al. 2007) to very large-scale "human flesh search" (a Chinese translation, in which "human flesh" refers to human empowerment; it has another name as crowd-powered search) (Wang et al. 2010; Zhang 2012; Zhang et al. 2012), in which a large number of voluntary Web users formed groups to collaborate on a single task. In 2006, Howe coined the term crowdsourcing and gave a definition of crowdsourcing as "the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in an open call" (Howe 2006). This definition covers most collective intelligence applications (particularly those for business), but it is a little too narrow to cover the large, voluntary, and loose organized crowd behaviors, like human flesh search, or crowdfunding sites like Crowdfunder and Kickstarter. Tarrell et al. reviewed 135 crowdsourcing-related articles from January 2006 to January 2013 (Tarrell et al. 2013). They found that research on crowdsourcing has been growing steadily. Researchers from computer science (CS) and information systems (IS) are the major contributors to this field. There are different focuses of CS and IS researchers. Generally speaking, CS researchers are mainly interested in modeling the collaboration of crowdsourcing and the design of better crowdsourcing systems (Zhang et al. 2007, 2010; Jurca and Faltings 2009; Pickard et al. 2011; Bozzon et al. 2013; Difallah et al. 2013), while IS researchers are more interested in the topics related to traditional IS research, like knowledge management, knowledge sharing, and incentives of contribution (Moon and Sproull 2008; Olivera et al. 2008; Mannes 2009; Bothner et al. 2011; Boudreau et al. 2011; Bayus 2013). There is also a trend of researchers from both sides are joining together to collaborate on crowdsourcing studies.

Since 2009 (Isolated Chinese SNSs)

As mentioned previously, as the country with the largest number of Internet users, China has blocked the access to some major SNSs including Facebook and Twitter (for consistency, we use "China" to refer Chinese Mainland only. Facebook and Twitter are popular in Hong Kong, Macau, and Taiwan, though a portion of people from these regions also use Chinese SNSs). However, Renren, Weibo, and other regional SNSs successfully took over the roles. In fact, these blocked SNSs did not perform very well in Chinese Mainland (as compared to their popularity in Taiwan) before they were blocked.

Although there is censorship upon Chinese cyberspace, and the "real-name policy" was recently applied to SNSs, SNSs are still among the freest platforms for Chinese Internet users to express their opinions (http://www.theatlantic.com/china/archive/2013/03/why-chin as-real-name-internet-policy-doesnt-work/274373/). Sometimes, the topics and keywords of users' discussions are seen censored and deleted automatically by SNSs, but users could generally find an alternative way to express the same meaning. There are countless "juicy stories" being generated by SNS users, in particular, Weibo users. The topics of their discussion are not quite the same as Twitter. Business people and brands have unique ways of marketing on Weibo and WeChat. For example, they create WeChat groups and push multimedia advertisements to users and communicate with users directly using the WeChat account. Rumors abound in the community. People collaborate to conduct "human flesh search" (Lu and Qiu 2013). The topics of users' arguments and fights can range from a tiny statement made by a celebrity or a TV program to serious economic or political issues.

Here, we present one example of the human flesh search (HFS) against corruption that aroused on Weibo. In 2012, a government official was photographed smirking after a tragic traffic accident. It enraged Chinese Internet users and this official's life quickly became under scrutiny. The HFS against him was started right away. Several photos of him started spreading in Weibo next day, and people quickly discovered 11 pricey watches he was wearing from these photos. Weibo users thought that there was no way that he could afford these watches on an honest government official's salary. The discoveries from HFS made the government start to investigate whether he was a corrupt official. Eventually, he lost his job and political career and is now under further investigation by judicial departments. The above story is an illustration of the Chinese Internet users using SNSs to do HFS for anticorruption purposes. However, there are also some other examples, in which people violated the personal privacy of others. A complete analysis, in English, of HFS can be found in Zhang (2012).

There are also "Internet water armies" (paid Internet commentators) on Weibo (Zheng et al. 2011), for example, groups advertising for a brand and attacking other brands, groups doing HFS and being HFSed, and groups criticizing or defending the government (there are mainly two major groups: (a) those who mainly criticize the government and would like a change and (b) those who defend the government and prefer a more stable society rather than a radical change). It was reported that there are government-/institute-/organization-funded Internet commentators trying to steer the public opinions towards the policies of governments (both within China and overseas) (http://news.bbc.co.uk/2/hi/asia-pacific/7783640.stm). Internet users called those defending government as "(RMB) 50 Cent Party" and those attacking the government (sometimes with fake rumors) as "(USD) 50 Cent Party," because the two "parties" are paid RMB 50 cent by the local government or USD 50 cent by a foreign government or institutes. These groups have been fighting each other on Weibo, and some groups were making up fault rumors to attack others and try to attract more Weibo users to support them (Fossato 2009; Bremmer 2010).

To regulate people's fights and control the existence of rumors, Sina Weibo proposed a credit system. In this system, each user has a credit score, people can sue others if they intentionally spread fake rumors, insult others, violate others' personal privacy, etc. If a user's score is low, he or she will be marked as a "low-credit user." A lot of such interesting things are happening in Chinese SNSs. However, most research on Chinese SNSs repeated the study of Twitter. How to distill interesting and unique research questions based on Weibo data and to properly answer them is a strongly promising and needed research.

Future Directions

In this entry, we briefly review the history of SNSs worldwide. In particular, we describe the use of SNS in China, which has not been well covered by the literature in the West. SNS is still a rapidly evolving area, with new types of SNSs emerging and novel research directions being explored. Despite numerous powerful quantitative analysis methodologies developed, there are still a large number of unanswered research questions from theoretical social sciences. The link between computational sciences and social sciences could be much stronger with solid research, which answered key research questions derived from social theories. Another future research topic that we anticipate is the cross-cultural analytics of SNSs. Most researches to date have been focused on popular SNSs in the West, with datasets that mostly came from one SNS and a single country or language. What are the differences across different SNSs? How were multiple SNSs linked together? Are there any cultural differences in the behavior of people using SNSs? These research topics are expected to not only fill the holes of current literature, but also to shed light on an in-depth understanding of the use in different cultures. We hope that our review and discussions can help researchers and practitioners to get a brief overview of SNS to date and gain an outlook of future research directions on SNS.

Cross-References

- ▶ Classical Algorithms for Social Network Analysis: Future and Current Trends
- ▶ Collective Intelligence: Overview
- ▶ Community Detection: Current and Future Research Trends
- ▶ Crowdsourcing and Human Computation: Introduction
- ▶ Crowdsourcing and Social Networks
- ▶ Dark Side of Online Social Networks: Technical, Managerial, and Behavioral Perspectives
- ▶ Data Mining
- ▶ East Asian Social Networks
- ▶ Futures of Social Networks: Where Are Trends Heading?
- ▶ Online Communities
- ▶ Privacy Issues for SNS and Mobile SNS
- ▶ Social Computing
- ▶ Social Media, Definition, and History
- ▶ Social Networking Sites
- ▶ Structural Holes
- ▶ Topic Modeling in Online Social Media, User Features and Social Networks for
- ▶ Web Communities Versus Physical Communities

Acknowledgments This work was supported in part by a grant from the US Defense Advanced Research Project Agency Social Media in Communication Project (SMISC) and the US Army Research Laboratory Network Science CTA.

References

Barabási A-L (2013) Network science. Philos Trans R Soc A 371:20120375. http://www.barabasilab.com/pubs//CCNR-ALB_Publications/201302-18_RoyalSoc-NetworkScience/201302-18_RoyalSoc-NetworkScience.pdf

Bayus BL (2013) Crowdsourcing new product ideas over time: an analysis of the Dell IdeaStorm community. Manag Sci 59(1):226–244

Bishop CM, Nasrabadi NM (2006) Pattern recognition and machine learning. Springer, New York

Bonabeau E (2009) Decisions 2.0: the power of collective intelligence. MIT Sloan Manag Rev 50(2):45–52

Bothner MS, Podolny JM, Smith EB (2011) Organizing contests for status: the Matthew effect vs. the Mark effect. Manag Sci 57(3):439–457

Boudreau KJ, Lacetera N, Lakhani KR (2011) Incentives and problem uncertainty in innovation contests: an empirical analysis. Manag Sci 57(5):843–863

Boyd DM, Ellison NB (2007) Social network sites: definition, history, and scholarship. J Comput-Mediat Commun 13(1):210–230

Bozzon A, Brambilla M, Ceri S, Mauri A (2013) Reactive crowdsourcing. In: Proceedings of the 22nd international conference on World Wide Web, International World Wide Web Conferences Steering Committee, Rio de Janeiro, 13–17 May 2013

Bremmer I (2010) Democracy in cyberspace – what information technology can and cannot do. Foreign Aff 89:86

CNNIC (2013) 32nd Statistical report on Internet Development in China. China Internet Network Information Center. https://cnnic.com.cn/IDR/ReportDownloads/201310/P020131029430558704972.pdf

Difallah DE, Demartini G, Cudré-Mauroux P (2013) Pick-a-crowd: tell me what you like, and i'll tell you what to do. In: Proceedings of the 22nd international conference on World Wide Web, International World Wide Web Conferences Steering Committee, Rio de Janeiro

Fossato F (2009) Web captives. Index Censorsh 38 (3):132–138

Glott R, Schmidt P, Ghosh R (2010) Wikipedia survey – overview of results. United Nations University: Colleaborative Creativity Group. http://www.ris.org/uploadi/editor/1305050082Wikipedia_Overview_15March2010-FINAL.pdf

Gonzalez R, Cuevas R, Motamedi R, Rejaie R, Cuevas A (2013) Google+ or Google−?: dissecting the evolution of the new OSN in its first year. In: Proceedings of the 22nd international conference on World Wide Web, International World Wide Web Conferences Steering Committee, Rio de Janeiro

Han J, Kamber M, Pei J (2006) Data mining: concepts and techniques. Morgan Kaufmann, San Francisco

Howe J (2006) The rise of crowdsourcing. Wired Mag 14 (6):1–4

ITU (2013) Key ICT indicators for developed and developing countries and the world (totals and penetration rates). International Telecommunications Unions. https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx

Java A, Song X, Finin T, Tseng B (2007) Why we Twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on web mining and social network analysis, ACM, San Jose

Jurca R, Faltings B (2009) Mechanisms for making crowds truthful. J Artif Intell Res 34(1):209

Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on World Wide Web, ACM, Raleigh

Lazer D, Pentland AS, Adamic L, Aral S, Barabasi AL, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M (2009) Life in the network: the coming age of computational social science. Science 323 (5915):721

Lévy P, Bonomo R (1999) Collective intelligence: mankind's emerging world in cyberspace. Perseus Publishing, Cambridge, MA

Lu J, Qiu Y (2013) Microblogging and social change in China. Asian Perspect 37(3):305–331

Mannes AE (2009) Are we wise about the wisdom of crowds? The use of group judgments in belief revision. Manag Sci 55(8):1267–1279

Manning CD, Schütze H (1999) Foundations of statistical natural language processing. MIT, Cambridge, MA

Moon JY, Sproull LS (2008) The role of feedback in managing the Internet-based volunteer work force. Inf Syst Res 19(4):494–515

Murthy D (2013) Twitter: social communication in the Twitter age. Polity Press, Cambridge

Olivera F, Goodman PS, Tan SS-L (2008) Contribution behaviors in distributed environments. Manag Inf Syst Q 32(1):23

Pickard G, Pan W, Rahwan I, Cebrian M, Crane R, Madan A, Pentland A (2011) Time-critical social mobilization. Science 334(6055):509–512

Schuler D (1994) Social computing. Commun ACM 37 (1):28–29

Shadbolt N, Hall W, Hendler JA, Dutton WH (2013) Web science: a new frontier. Philos Trans R Soc A Math Phys Eng Sci 371(1987):20120512

Tarrell A, Tahmasbi N, Kocsis D, Tripathi A, Pedersen J, Xiong J, Oh O, de Vreede G-J (2013) Crowdsourcing: a snapshot of published research. In: Proceedings of the nineteenth Americas conference on information systems, Chicago

Wang F-Y, Carley KM, Zeng D, Mao W (2007) Social computing: from social informatics to social intelligence. IEEE Intell Syst 22(2):79–83

Wang F-Y, Zeng D, Hendler JA, Zhang Q, Feng Z, Gao Y, Wang H, Lai G (2010) A study of the human flesh search engine: crowd-powered expansion of online knowledge. Computer 43(8):45–53

Zhang Q (2012) Analyzing cyber-enabled social movement organizations: a case study with crowd-powered search. PhD, The University of Arizona, Tucson

Zhang J, Ackerman MS, Adamic L (2007) Expertise networks in online communities: structure and algorithms. In: Proceedings of the 16th international conference on World Wide Web, ACM, Banff

Zhang Q, Feng Z, Wang F-Y, Zeng D (2010) Modeling cyber-enabled crowd-powered search. In: The second Chinese conference on social computing, Beijing

Zhang Q, Wang F-Y, Zeng D, Wang T (2012) Understanding Crowd-Powered Search Groups: A Social Network Perspective. PLoS ONE 7(6): e39749. https://doi.org/10.1371/journal.pone.0039749

Zheng X-L, Zhong Y-G, Wang F-Y, Zeng D-J, Zhang Q-P, Gui K-N (2011) Social dynamics research based on web information. Complex Syst Complex Sci 8 (3):1–12

Recommended Reading

Adamic LA, Glance N (2005) The political blogosphere and the 2004 US election: divided they blog. In: Proceedings of the 3rd international workshop on link discovery, ACM, Chicago

Albert R, Barabasi A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97

Antal T, Krapivsky P, Redner S (2005) Dynamics of social balance on networks. Phys Rev E 72(3):036121

Bakshy E, Rosenn I, Marlow C, Adamic L (2012) The role of social networks in information diffusion. In: Proceedings of the 21st international conference on World Wide Web, ACM, Lyon

Barabási A-L (2002) Linked: the new science of networks. Basic Books, New York

Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Bond RM, Fariss CJ, Jones JJ, Kramer AD, Marlow C, Settle JE, Fowler JH (2012) A 61-million-person experiment in social influence and political mobilization. Nature 489(7415):295–298

Burt RS (2009) Structural holes: the social structure of competition. Harvard University Press, Cambridge, MA

Cha M, Mislove A, Gummadi KP (2009) A measurement-driven analysis of information propagation in the Flickr social network. In: Proceedings of the 18th international conference on World Wide Web, ACM, Madrid

Chaney AJB, Blei DM (2012) Visualizing topic models. In: Proceedings of the sixth international AAAI conference on weblogs and social media, ICWSM, Dublin, 4–7 June 2012

Chang J, Boyd-Graber J, Blei DM (2009) Connections between the lines: augmenting social networks with text. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, Paris

Cheng X, Dale C, Liu J (2008) Statistics and social network of YouTube videos. In: 16th international workshop on quality of service, 2008 (IWQoS 2008), IEEE, Enskede

Conover M, Ratkiewicz J, Francisco M, Gonçalves B, Menczer F, Flammini A (2011) Political polarization on Twitter. In: Proceedings of the fifth international AAAI conference on weblogs and social media, ICWSM, Barcelona

Cook J, Kenthapadi K, Mishra N (2013) Group chats on Twitter. In: Proceedings of the 22nd international conference on World Wide Web, International World Wide Web Conferences Steering Committee, Rio de Janeiro

De Choudhury M, Sundaram H, John A, Seligmann DD (2009) What makes conversations interesting?: themes, participants and consequences of conversations in online social media. In: Proceedings of the 18th international conference on World Wide Web, ACM, Madrid

Diakopoulos NA, Shamma DA (2010) Characterizing debate performance via aggregated Twitter sentiment. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, Atlanta

Dwyer C, Hiltz SR, Passerini K (2007) Trust and privacy concern within social networking sites: a comparison of Facebook and MySpace. In: Proceedings of the thirteenth Americas conference on information systems, AMCIS, Keystone

Ellison NB, Steinfield C, Lampe C (2007) The benefits of Facebook "friends": social capital and college students' use of online social network sites. J Comput-Mediat Commun 12(4):1143–1168

Ghannam J (2011) Social media in the Arab world: leading up to the uprisings of 2011. Center for International Media Assistance/National Endowment for Democracy 3. http://www.cima.ned.org/wp-content/uploads/2015/02/CIMA-Arab_Social_Media-Report-10-25-11.pdf

Goetz M, Leskovec J, McGlohon M, Faloutsos C (2009) Modeling blog dynamics. In: Proceedings of the international conference on weblogs and social media, ICWSM, San Jose

Harper FM, Moy D, Konstan JA (2009) Facts or friends?: distinguishing informational and conversational questions in social Q&A sites. In: Proceedings of the 27th international conference on human factors in computing systems, ACM, Boston

Hinds D, Lee RM (2008) Social network structure as a critical success condition for virtual communities. In: Proceedings of the 41st annual Hawaii international conference on system sciences, IEEE, Hawaii

Huberman B, Romero D, Wu F (2008) Social networks that matter: Twitter under the microscope. First Monday 14 (1). http://firstmonday.org/article/view/2317/2063

Jiang J, Wilson C, Wang X, Huang P, Sha W, Dai Y, Zhao BY (2010) Understanding latent interactions in online social networks. In: Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, ACM, Melbourne

Kairam S, Brzozowski M, Huffaker D, Chi E (2012) Talking in circles: selective sharing in Google+. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, Austin

Khondker HH (2011) Role of the new media in the Arab Spring. Globalizations 8(5):675–679

Kumar R, Novak J, Raghavan P, Tomkins A (2004) Structure and evolution of blogspace. Commun ACM 47 (12):35–39

Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. Phys Rev E 78(4):046110

Larsson AO, Moe H (2012) Studying political microblogging: Twitter users in the 2010 Swedish election campaign. New Media Soc 14(5):729–747

Leskovec J, Huttenlocher D, Kleinberg J (2010a) Predicting positive and negative links in online social networks. In: Proceedings of the 19th international conference on World Wide Web, ACM, Raleigh

Leskovec J, Huttenlocher D, Kleinberg J (2010b) Signed networks in social media. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, Atlanta

Lewis K, Kaufman J, Gonzalez M, Wimmer A, Christakis N (2008) Tastes, ties, and time: a new social network dataset using Facebook.com. Soc Networks 30(4):330–342

Lotan G, Graeff E, Ananny M, Gaffney D, Pearce I, Boyd D (2011) The revolutions were tweeted: information flows during the 2011 Tunisian and Egyptian revolutions. Int J Commun 5:1375–1405

Morris MR, Teevan J, Panovich K (2010) What do people ask their social networks, and why?: a survey study of status message Q&A behavior. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, Atlanta

Myers SA, Zhu C, Leskovec J (2012) Information diffusion and external influence in networks. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, Beijing

Newman ME (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256

Newman ME (2006) Modularity and community structure in networks. Proc Natl Acad Sci U S A 103 (23):8577–8582

Qu Y, Huang C, Zhang P, Zhang J (2011) Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake. In: Proceedings of the ACM 2011 conference on computer supported cooperative work, ACM, Hangzhou

Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World Wide Web, ACM, Raleigh

Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with Twitter: what 140 characters reveal about political sentiment. In: Proceedings of ICWSM'10, Washington, DC, pp 178–185

Wang T, Zhang Q, Liu Z, Liu W, Wen D (2012) On social computing research collaboration patterns: a social network perspective. Front Comput Sci China 6(1):122–130

Watts DJ (1999) Small worlds: the dynamics of networks between order and randomness. Princeton University Press, Princeton

Watts D, Strogatz S (1998) Collective dynamics of small-world networks. Nature 393:440–442

Yang Z, Wilson C, Wang X, Gao T, Zhao BY, Dai Y (2011) Uncovering social network Sybils in the wild. In: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, ACM, Berlin

Zink M, Suh K, Gu Y, Kurose J (2008) Watch global, cache local: YouTube network traffic at a campus network: measurements and implications. In: Electronic imaging 2008, International Society for Optics and Photonics, San Jose

Social Networking Services

▶ Social Networking on the World Wide Web

Social Networking Sites

▶ Ethics of Social Networks and Mining
▶ Friends Recommendations in Dynamic Social Networks
▶ Online Privacy Paradox and Social Networks
▶ Social Networking on the World Wide Web

Social Networks

▶ Counterterrorism, Social Network Analysis in
▶ Detecting and Identifying Communities in Dynamic and Complex Networks: Definition and Survey
▶ Game-Theoretic Framework for Community Detection
▶ Game Theory and Social Networks
▶ Network Actors Within Entrepreneurial Networks: The Current State of Research
▶ Network Analysis in French Sociology and Anthropology
▶ Networks in Rural Sociology
▶ Sampling Effects in Social Network Analysis
▶ Social Capital
▶ Social History of Computing and Online Social Communities
▶ Social Networks in Emergency Response
▶ Spatiotemporal Topic Detection from Social Media
▶ Telecommunications Fraud Detection, Using Social Networks for
▶ Top Management Team Networks
▶ Tracking Dynamic Community Evolution and Events Mobility in Social Networks

Social Networks Analysis

▶ Stability and Evolution of Scientific Networks

Social Networks and Politics

Elena Pavan
Institute of Humanities and Social Sciences,
Scuola Normale Superiore, Florence, Italy

Synonyms

Collective action; Digital media; Globalization; Governance; Networked politics

Glossary

Social Networks	A specific form of social organization based on patterns of communication and exchange amongst actors involved
Politics	A wide variety of dynamics aimed at the production of public purpose, from the expression of visions, to the production of policies and regulations to norms consolidation

Definition

The more time goes by, the more the nexus between social networks and politics becomes a paramount and yet complex site for developing critical reflections and innovative research practices at the crossroads between social and computational sciences. In comparison to only a few years ago, when the conceptual and empirical effort of this *Encyclopedia* begun, what we framed as “future perspectives” of research – essentially, the necessity to understand the role of digital communication networks within the vast realm of politics – is today something we deal with on a regular basis and within research teams that have never been as multidisciplinary as they are today. Indeed, the boundaries between *social networks*, broadly understood as complex social organizational structures defined at the intersections between multiple relations established amongst social actors (Powell 1990), and *the networked digital infrastructure* that sustains and contributes to their formation are increasingly blurred. At the crossroads between social and technological opportunities and agencies, major mobilizations such as the Arab Springs, the spread of Occupy! as well as of the Indignados, but also the emergence of Podemos and Syriza as major political alternatives, have provided eloquent examples of the potentialities of digital media-assisted collective networks for the renewal of political and governance arrangements all over the world (Bennett and Segerberg 2013).

Precisely because we are proceeding so enthusiastically and rapidly into the future of research on how social networks and politics intersect, we must be equipped with a proper toolkit, one that contains the knowledge resources that can help us exploring new frontiers while avoid making false steps. Certainly, the diffusion of digital communications and the progressive hybridization between online and offline spaces of action are conveying an unprecedented prominence upon communicative and symbolic interactions amongst individuals. Thus, the overall invisibility of these online communication networks does not make them less real or less relevant as spaces where power dynamics are played out (see Padovani and Pavan 2016). But for how relevant online communication networks may have become in these days, we shall not forget that *social networks* have *always* existed, way before the Internet, and that they have always been relevant within and beyond the domain of politics. In the same way, we shall not forget that *a myriad of social networks*, not only those established upon digital communication platforms, remain relevant to determine the dynamics that prelude to the production of public purpose – i.e., the "expression of vision, values, plans, policies and regulations that are valid for and directed towards the general public" (Sørensen and Torfing 2007, p. 10) – and that here we summarize under the label of "politics."

In what follows, we first reconstruct the premises to networked politics to then illustrate the main forms it can take. We continue with an overview of current research trends on digital communication networks as political networks – a key application of network approaches to the study of politics. We conclude by looking at future perspectives on the study of the nexus between social networks and politics, reflecting in particular upon the implications of "big data" approaches to investigation.

Introduction

Addressing the nexus between social networks and politics is a complex task which requires, in the first place, an exploration of the background scenario which led to the consolidation of networks as both metaphors to depict contemporary complex political arrangements and as a true organizational mode for the conducts of politics at all levels. Also, this task requires some systematization effort aimed at clarifying the main traits of networked politics in its most popular declinations: as forms of collaboration between institutions (*government networks*), as multiactor arrangements for the definition and the implementation of policies (*policy networks*), and as the overall perspective to depict different instances of collective action (*collective action networks*).

Key Points

Even if today the nexus between social networks and politics is increasingly defined by digital communication exchanges, its fuller and more genuine understanding cannot transcend from contextualizing the rise of *networks as a mode for organizing and conducting politics* at all levels, from the national to the global, and for the production of different types of public purposes, from formal policies (Knoke et al. 1996); to norms, i.e., cognitive frameworks that guide the action of institutional and noninstitutional actors (Finnemore and Sikkink 1998); to the "demand and/or provision of a collective good" (Baldassarri 2009: 321), which remains the prerequisite for collective actions and social movement dynamics to develop. Within such a broader context, ICTs and digital media are not the sole space for political dynamics to unfold but, more realistically, one of the most crucial intervening elements shaping the patterns of political dynamics as they foster the construction of communication-based social relations and, in the end, provide “the means of political debate: the arena, the communication links, the agenda” (Bijker 2006).

Historical Background: Globalization, the Overcoming of Nation-State Politics, and the Role of Networks

One of the main features of the contemporary world, perhaps the most emphasized, is *interconnectedness*. Societies and economies today are linked in complex webs of interactions, influencing each other in non-trivial ways, both enhancing possibilities (as in the case of solidarity networks after the attacks to Paris in 2015) as well as augmenting the reach of negative dynamics (as it is in the case of the global financial crisis).

Whether it is considered its cause or its consequence, interconnectedness is often related to the concept of *globalization*, which can be defined as a set of processes impacting the spatial organization of social relations and transactions “generating transcontinental or inter-regional flows and networks of activity, interaction and the exercise of power” (Held et al. 1999, p. 16). Globalization processes have taken place in a variety of fields (economy, politics, culture, environmental concerns) and have contributed to the transformation of the world into a “shared social space” (Held et al. 1999, p. 1), where traditional boundaries (territorial, thematic, or based on competences) are now blurred.

At the same time, globalization is presenting us with a number of challenging aspects. Looking in particular at the domain of politics, many issues (as the sustainable use of energetic resource, the control of financial markets or even the definition of national labor policies, etc.) are not any longer managed by homogeneous societies or economies embodied by the nation-state. Rather, these issues are now spanning a wide range of geographically distant and socioculturally heterogeneous constituencies and come to represent global *societal challenges* characterized by their global scale and by features of *diversity* (of actors and perspectives), *dynamics* (the continuous evolution of issues at stake as well of perspectives upon them), and *complexity* (of the webs of interaction).

Hence, governments and institutions are pressured to intervene in a complex scenario where the distinction between domestic and foreign affairs is blurred and where multiple and diversified knowledge is required to keep under control all the facets of global problems. Shortcomings in facing these challenges have translated into a threefold deficit of *legitimacy*, *knowledge*, and *access* (Hockings 2006), which questions the traditional hierarchical nation-state model as the preferred governance mechanism. Thus, the recent global financial crisis has highlighted the limits of a regulation model based on free market assets.

Furthermore, the increased level of interconnectedness fosters the proliferation of non-traditional political actors (e.g., civil society organizations and coalitions, social movements, subcultures, single committed individuals, loose platforms for action), which constitute a plurality of different *publics*, all exerting control on the management of public affairs, possessing the required knowledge for the management of global challenges, proposing alternative solutions to current mechanisms deficiencies, and willing to take part into reformative dynamics and governance experiments to increase the effectiveness and the democratic features of political mechanisms.

Contemporary global settings call then for a “decentralized concept of social organization and governance [for which] society is no longer exclusively controlled by a central intelligence (e.g., the state); rather controlling devices are dispersed and intelligence is distributed amongst a multiplicity of action (or ‘processing’) units” (Kenis and Schneider 1991, p. 26).

Here, networks enter as a powerful image to depict the growing complexity, but they also represent a truly new social morphology (Castells 2011): one for which policy outcomes and outputs are “generated within multiple-actors-set in which actions are interrelated in a more or less systematic way” (Kenis and Schneider 1991, p. 32). Within an overall context of uncertainty, due to the shortcoming of conventional political mechanisms and to the difficulties of reorganizing steering activities so to include all actors and stakeholders (Börzel 1998), networks emerge to *incorporate*, *supply*, and *challenge* market and hierarchies as governance mechanisms for the production of public purpose (Kahler 2009).

Networks emerge then in response to the lack of a central authority able to set the widely accepted benchmarks for the conduct of public affairs. As a mode of (re)organizing political dynamics, they are based on cooperation (and yet allow for the development and management of conflicts), foster mutual learning and the spread of knowledge, allow a fast translation of knowledge into action, and, hence, are flexible enough to compensate the variability and the overall uncertainty of the future (Powell 1990). For their peculiarities, networks become then the preferred arrangement for sustaining contemporary governance efforts, i.e., for producing rules, norms, and, more broadly, the conditions for ensuring order through new strategies of problem-solving based on relationships between private and public actors that augment governing capacities.

Networked Politics as Forms of Communication Networks

Generally speaking, (social) networks in politics have been used to study both the emergence of coalitions within states, with a specific accent on resource mobilization and power redistribution, and the creation of interdependencies between states (Wellman 2002). Over time, labels have multiplied as to depict a variety of situations in which interdependency between political actors is experienced and managed. However, the application of a relational view for studying political transformations has not happened in a consistent way: similar situations have been labeled differently, the same label has been applied to different occurrences, and the underlying assumptions leading to the choice of a specific network concept over the other are seldom made explicit (Börzel 1998). The heterogeneity of uses somehow jeopardizes the heuristic potential of the network idea itself for the study of politics, and despite studies adopting a network point of view have multiplied also in the study of political dynamics, an overall consensus on what networks mean for politics (a mere metaphor, a method, an analytic tool, or a proper theory) is still missing (Börzel 1998).

However, the heterogeneity of labels and uses is not a total impediment to a systematic overview of different conceptualizations of networked politics. In fact, all applications of the (social) network concept in the study of political dynamics share the initial assumption that both the hierarchical nation-state and the market models present major shortcomings that hinder the achievement of satisfactory results. Because they are not self-sufficient, states need to collaborate with other actors and to internalize the knowledge coming from these collaborations within policy-finding and policy-making processes. This creates an overall situation of interdependency between institutional and noninstitutional actors that is managed first and foremost through the establishment of communication flows from one actor to others. In this sense, all applications of the network concept to politics can be summarized through the idea of "communication networks" that join together actors mainly through the exchange of messages across time and space in the attempt to stabilize structures of interaction out of the chaos provided by the globalized context (Monge and Contractor 2003).

On these bases, we can distinguish between different types of political communication networks leaning on the elements that define networks as specific forms of social organizations, i.e., *actors* and *relations*. Looking then at which are political actors involved, how heterogeneous they are, and at why they interact, we can then make sense of different applications of the network concept in the study of politics – in particular between *government networks*, *policy networks* and *collective action networks*.

Government networks are composed by national governmental and intergovernmental organizations officials with the overall aim of providing traditional political actors with the necessary global reach they miss in the contemporary globalized political milieu through their engagement and exploitation of flexible arrangements for collaboration (Slaughter 2004). Examples of such networks are the G-7 or the G-8 and the G-20 as well as the Asia-Pacific Economic Cooperation (APEC) or the Organization for Economic Cooperation and Development (OECD). Actually, these networks are not completely new phenomena, but at the present stage, their scale, scopes, and type of ties are undergoing an unprecedented growth.

Government networks are composed of *homogeneous nodes*, i.e., governmental and intergovernmental actors, who can be further differentiated on the bases of the interests they carry (Slaughter 2004). Furthermore, government networks can be *horizontal* (aimed at exchanging information and best practices) or *vertical* (in which authority is delegated to a higher-level organization, e.g., in the field of justice with international courts). Slaughter points out that, in mobilizing "traditional" political actors, government networks respond to the "governance trilemma" for which (a) contemporary political settings see the need for official regulatory activity at global level yet without centralization of power and ensuring accountability across different policy mechanisms, (b) governmental actors can and should interact with a multiplicity of nongovernmental organizations that have emerged as important actors but (c) their role in governance bears distinct and different responsibilities (Slaughter 2004, p. 12–15). In this context, government networks offer "a flexible and relatively fast way to conduct the business of global governance, coordinating and even harmonizing national government action while initiating and monitoring different solutions to global problems. Yet they are decentralized and dispersed, incapable of exercising centralized coercive authority. Further (. . .) they can interact with a wide range of NGOs, civic and corporate, but their responsibilities and constituencies are far broader" (Slaughter 2004, p. 11).

That of ***policy networks*** is probably the most widely used label to describe a whole set of very different processes revolving around transformations of policy-making processes. In their seminal work, Marin and Mayntz argue that policy networks "are explicitly defined not only by their structure as interorganizational arrangements, but also by their function – the formulation and the implementation of policy" (Marin and Mayntz 1991, p. 16). Actors involved in collective decision processes might be of *different nature*, but their ability to enter the network varies depending on the porosity of the policy domain under discussion (i.e., the more uncertain the domain, the wider the constituency of actors involved).

In being the most widespread label for depicting the nexus between social networks and politics, policy networks have been reviewed and classified in several ways (see Börzel 1998; Adam and Kriesi 2007). Overall, existing literature points out the use of this concept to identify, depending on concrete case studies, structures for interest intermediation among actors; alternative governance structures challenging markets and hierarchies; multiactor arrangements for policy implementation; or a "formalized, quantitative approach of social network analysis (. . .) that focuses on the relations between actors and not on actors' characteristics" (Adam and Kriesi 2007, p. 130).

As a specific approach for studying policymaking activities through network analysis techniques, *social network analysis of policy networks* (e.g., Knoke et al. 1996) is mostly concerned with the redistribution of power along network ties, where the degree of power is proportional to the degree at which interests held by different actors involved are reflected through policy outcomes and not in relation to innate qualities. Concrete operationalizations of this relational view of power have translated into two types of studies: positional, which are primarily concerned with actors' positions within the network, and relational, concerned with characters and effects of relations existing between actors in a system (Lotan et al. 2011).

More recently, the idea of *governance networks* has been proposed to expand the reach of the policy network approach also to the production of non-binding policy outcomes, e.g., norms. In this sense, as a sort of "second generation" of policy network studies, governance networks studies are not so much focused on the actual existence of networks as distinct and legitimate forms of governance (Sørensen and Torfing 2007). Rather, they start from an explicit recognition of networks existence and political meaning to model interactions thus keeping into account structural, processual, and cognitive elements. In this sense, governance networks can be defined as "(1) a horizontal articulation of interdependent, but operationally autonomous actors; (2) who interact through negotiations; (3) transpiring within a regulative, normative, cognitive and imaginary framework; (4) that to a certain extent is self-regulating; and (5) which contributes to the production of public purpose within a particular area" (Sørensen and Torfing 2007, p. 9).

Although they are often studied in the context of policies production and coordination, the potential of governance networks as analytical tools goes beyond conventional policy making to include "decision finding rather than decision making processes" (Hemmati 2002, p. 19). In this sense, governance networks are the preferred label to study those political dynamics that are not necessarily finalized to the formulation of binding provisions but, rather, are aimed at the production of shared norms and knowledge – e.g., the United Nation World Summit on the Information Society or the Internet Governance Forum for the creation of a common vision between governments, private sector, and civil society on how to manage and regulate contemporary societies and their technological infrastructures (see Pavan 2012).

Finally, that of ***collective action networks*** is a conceptual perspective based on social networks that has been pushed forward within the study of political participations and contentious politics to differentiate and underline the specificities of the diverse collective action instances: social movements, coalitions, organizational action, and communities/subcultures (Diani 2003; Diani and Bison 2004; Diani 2008). According to this specific perspective, and consistently with the premises of the structural approach to the study of politics, the accent is put on actors' interactions rather than on actors' features (e.g., the level of formalization of organizational assets, the sociodemographic characteristics of citizens who mobilize or participate politically). Thus, this perspective was elaborated in the first place to specify social movements in comparison to other forms of contentious politics or political participation (Diani 2008), but it can be adopted to study of all forms of collective political participation.

It is the combination of three different network characteristics that allows to distinguish between different realizations of collective action: (i) the presence or absence of conflictual orientations towards clearly identified opponents, (ii) network density (sparse vs. dense networks), and (iii) presence of a strong or weak network collective identity. While the presence of conflict refers more to specific repertoires of actions adopted within all types of collective actions, levels of density and of identity sharing are the two main dimensions along which instances of action can be distinguished.

The "intensity" of network identity divides dense networks into *social movements*, characterized by a strong identity, and *coalitions*, where collective identity is weak. Collective identity is important for social movements as it entails the presence of shared visions and values that sustain a long-term involvement over time and, in this sense, is what bonds different individuals and organizations, each of which with its own agenda, modes of behaviors, and perceptions, within the same mobilization effort over time (Melucci 1996). Thus, although social movements can be based on consensual repertoires, they are often coupled by a marked attitude toward conflict, as they rise as explicit expressions of social dissent towards identified opponents (Diani and Bison 2004). Conversely, when network identity is lower, dense networks of exchange between actors respond to instrumental and more short-term goals. Instrumentality of action is what characterizes coalitions in general (Gamson 1961), but it is worth noting that it is not tantamount to the lack of values or solidarity. In fact, although coalitions lack a long-term vision, in their attempt to pursue a specific goal they can repeat over time, as it happens for example in the case of the campaign Take Back the Tech! against ICTs-assisted violence on women, which runs every year from November 25 to December 10 (Pavan 2013). Moreover, especially when coalitions are transnational and the goal they pursue is linked to a reform of social arrangements, there is the need to supply instrumentality with shared views and values.

If looser networks are coupled with weak collective identity, the focus shifts from collective dynamics to single *organizations*, while if they are associated with strong identities, they generate *communities*. Within specific organizations, such as Greenpeace, Sea Shepherd, or Oxfam, action is carried on very much following the agenda and the modus operandi of the single organization, i.e., under an organizational (rather than collective) identity (Diani 2008; Diani and Bison 2004). Thus, participation to action is consequential to the ownership of established membership criteria (e.g., all sorts of eligibility conditions from having paid a fee to possessing some specific skills or competences). Differently from social movements and communities, which join together a plurality of organizations under widely shared frames and beliefs systems, organizational collective action is characterized by a specific entrenchment within the boundaries of the organization itself, which is responsible for determining how the mobilization is carried on. Conversely, social movements, as well as communities, are "multicentric" as none can claim to represent the totality of the network (Diani 2008). Communities instead carry on collective activity through networks which are sparse and yet are characterized by a shared sense of belonging diffused among members. Here, the idea of community can be detached from that of territoriality and should be rooted in the shared practices and views thus blending the networked structure of mobilization within daily activities, which are conducted following the very values and ideals that jointly define the collective identity (Diani 2008).

Key Applications

Digital Communication Networks as Loci of Networked Politics

The global diffusion of digital media stands today as one of the most relevant factors contributing to redefine the nexus between social networks and politics. In particular, as seen above, digital communications via social media platforms enable the participation of nonconventional and nongovernmental actors and, hence, foster the circulation and the contamination of ideas and political agendas (Padovani and Pavan 2012).

Quite interestingly, one of the more relevant transformations triggered by this diffusion is to be found not at the practical level of networked politics but, rather, at the conceptual one. In fact, beside few exceptions related to the abovementioned collective action networks approach, social networks have often been considered as part of the context motivating individuals to mobilization and, far more rarely, as a true space for contention to occur (Diani 2015). Researchers and practitioners who engage today in the study of current forms of activism have instead to become more welcoming with the idea that *networks are true spaces of collective action* and not just an environmental factor influencing it. Indeed, by virtue of their materiality (i.e., their inherent networked, communicative and participated nature), digital media do not simply foster but, rather, *impose* a relational logic to the deployment and the organization of social relations (Pavan 2014). In this sense, networks cease to be only a privileged entry point to unveil and tackle the inherent diversity, dynamics, and complexity of social interactions that sustain collective projects. More importantly, *online communication networks become an actual component of organized collective efforts*, one where the collective construction of discourses summarizes the essence of both political activity and social conflict (Pavan 2012). Moreover, whereas so far networks have been considered as a possible and fruitful outcome of collective efforts (see Bosi et al. 2016), in a context where digital media materiality regulates the modalities through which individuals and groups relate one another, networks become a *prerequisite* to collective action.

In this context, while research questions linked to the forms and the efficacy of government, policy and "offline" collective action networks still remain open, a new set of inquiry lines are opening up. Main research activities are today oriented to uncover the relationship that exists between online network features and their capacity to promote and, possibly, achieve social change; how contents travel along network ties and thus become the fuel to enhance democracy or, quite the opposite, to exacerbate discriminations; what is the role played by technologies themselves in the construction of online networks of political action; how are power and leadership dynamics transformed, if at all, in the online space; and what is the interplay between online and offline courses of (collective) action.

In only a few years, the systematic exploration of online structures has yielded to some consistent results, although the dynamic nature of online communications pushes for continuous innovation in knowledge practices. What emerges quite consistently from the examination of contemporary forms of networked politics, in particular when it comes to collective action networks, is that online forms of engagement are more flexible, whether they refer to people's engagement with causes, ideas, or political organizations (Bennett and Segerberg 2013). In a communicative context dominated by social media and participatory communications, mobilization processes and commitment to collective efforts transcend the traditional mediation of social movement organizations and are instead fostered by social media platforms, which become true hubs favoring the creation of relationships and the contamination of agendas and action strategies (Bennett and Segerberg 2013). What motivates individuals to action are the commonalities of perspectives, needs, visions, and motivations that are shared within complex networks of mass self-communication (Castells 2012), to the point that the "collective" is experimented and experienced through the "individual" and the group becomes the means, rather than the end, of collective participation (Milan 2015).

This argument resonates with broader scope analyses of our societies, which emphasize the tendency to organize in a networked fashion and to engage selectively and temporarily within networks to provide ad-hoc contributions in view of the collective achievement of a common goal (see Benkler and Nissembaum 2006; Rainie and Wellman 2012). Along these lines, several labels have been proposed to portray flexible modes of political engagement. Bennett and Segerberg (2013), for example, coined the now consolidated expression "connective action logic," to stress the role of technology-enabled and enforced relationships created online by individual users. Analogously, Milan (2013) speaks about "cloud protesting," a dynamic of protest for which individuals act on their personal capacity, on the bases of their needs and agendas, and often within the spaces opened up by digital technologies rather than in those created by traditional movement organizations. Thus, observers have identified a trend toward the "hybridization of action repertoires" through which political organizations (whether civil society association or political parties), adopt mixed action strategies that smooth their hierarchical features, and are more open to the inclusion of flexible individual, sometimes occasional, participation (Bimber et al. 2012; Chadwick 2013). In turn, this increased flexibility seems to be conductive of a more enlarged, although not always sustained, civic engagement not only in relation to unconventional forms of participation but also in relation to more traditional political parties activities (Vaccari and Valeriani 2016).

Particularly in response to more skeptical observers, who have pointed out that the flexibility of online political networks is often used as a cover to hide their inherent weakness (e.g., Gladwell 2010; Morozov 2009; Diani 2011), another strand of research has developed to build a more solid bridge between the extemporary use of social media platforms and the transformation of collective action structuring dynamics. For example, Pavan (2013) and Tremayne (2014) analyzed how the circulation of tweets results into the construction of collective action frames. Other observers connected informational cascades occurring via digital media communications to recruitment mechanisms, leadership definition and the emergence of coalitional patterns (González-Bailón et al. 2011; Lotan et al. 2011). To some extent, these explorations have not been limited only to "good" forms of collective action, like antiausterity mobilizations gathered under the Occupy! umbrella, but have also been conducted looking at morally contested communities such as that of far-right organizations (see Caiani and Parenti 2013). A great deal of effort has also been put in analyzing the interplay between online and offline dynamics and how increases in digital communication volumes (in particular upon social media platforms) prelude to the realization of massive offline protests (see Howard and Hussain 2013; Steinert-Threlkeld et al. 2015).

Finally, reflections have multiplied on how power dynamics are played out within online networks. In this regard, some observers have stressed that "the exercise of power relationships is decisively transformed in the new organizational and technological context derived from the rise of global digital networks of communication" (Castells 2009). Power within digital networks has hence been associated with actors' capacity to direct and control the construction of meanings (Castells 2011; Bennett and Segerberg 2014; Pavan 2012) as actors' positions within online networks facilitate (or prevent) their gatekeeping function in controlling flows of information and ideas. Starting from actors structural positions within online networks, several observers have provided possible classification of "power roles." For example, looking at Twitter networks González-Bailón and colleagues (González-Bailón et al. 2013) distinguish between "common users"; traditional "influentials" (i.e., celebrities); "broadecasters" (i.e., incumbents on more than average central positions and are more active in sending than receiving contents); and "hidden influentials" (i.e., incumbents on more than average positions that are also often targeted through protest messages). Analogously, but looking at hyperlink networks, Padovani and Pavan (2016) proposed to distinguish between "programmers" (i.e., incumbents of central positions often targeted by other network members); "mobilizers" (i.e., nodes that are particularly active in sending hyperlinks); and "switchers" (i.e., nodes bridging different networks subgroups).

Future Directions

With much of current research on the nexus between social networks and politics oriented toward the study of online dynamics, ongoing developments in the study of digital and "big" data will certainly open new avenues and perspectives of research. In this sense, what Wagner-Pacifici and colleagues (2015) identify as the challenges for big data research do constitute by all means challenges for the study of networked politics. Thus, the authors identify three "analytic binaries" representing the critical aspects that demand our attention.

The first is the binary "Life/Data" and concerns the extent to which big and digital data can neutrally "stand for social life itself." Here, the authors endorse current critical reflections that underline how big and digital data are far from being neutral (boyd and Crawford 2012). Quite the opposite, not only the adoption of empirical approaches that lean on big and digital data are the actual reflection of a legitimate, and yet discretional, epistemological choice. More importantly, as much as any other type of data, big and digital data are influenced and shaped by the very materiality of the devices that have produced them. When it comes to the study of the nexus between social network and politics, this binary "Life/Data" assumes a specific meaning. It stands for the necessity to recognize acts of digital communications as true forms of political participation. This entails not only recognizing that the employment of digital media within political dynamics can changes the *action repertoires*, that is, the way in which politics is enacted. More deeply than that, accepting digital politics requires recognizing the necessary transformation of traditional policy making and policy finding logics, shifting the focus from "the usual suspects" (i.e., governments, intergovernmental agencies, structured civil society non-governmental organizations) to the myriad of individual, informal but nonetheless organized forms of political participation, expressed first and foremost through extemporary communicative acts. Moreover, a genuine focus on digital politics requires operating a conceptual and empirical shift from social to multidimensional networks, where not only social agency is at play but also a true technological agency

(Pavan 2014). Indeed, as much as digital media enhance the spaces of politics, they also contribute to shape them by imposing to human creativity what could be called a "material boundary", that is, defining the types of actions to be performed and the "code of conduct" to be followed. The extent to which the production of public purpose is materially enhanced or constrained cannot be assessed a-priori: it is a matter than can be only empirically addressed by inserting technologies within political networks and identifying the position they occupy within the system.

The second binary identified by the authors is "Mind/Machine" and concerns the important matter of interpretation. It did not take too much time to the whole research community to understand that "numbers do not speak alone," even when these numbers are truly "big" (boyd and Crawford 2012, p. 666). Numeric representations produced by machines always require interpretations performed through human minds, hence, nourished by systems of values, norms, and expectations that are proper of the social and political environment to which we belong. When it comes to the nexus between social networks and politics, this binary invites systematic reflections on the role played by digital political networks in relation to other forms of networked politics. No digital network can be claimed to represent the totality of a political process as digital networks enrich, rather than substitute, other forms of networked politics. In the same way, we cannot assume that, only because we are looking at digital forms of politics we are witnessing *a new form* of politics. In this sense, when investigating digital political networks, we should seek to unveil what is substantially new or different in digital online government, policy and collective action networks in comparison to those that have been developed in the pre-digital age; and, what are, conversely, the traits of continuity that we can discern. In spite of the many progresses made in mapping and analyzing new forms of networked politics, we remain quite far from having fully understood how digital behaviors "scale down" (Breiger 2015) to the modes of doing politics that have been developed over centuries.

Finally, the last binary couples "Induction/Deduction" and concerns the epistemological approach that researchers adopt when they engage in the production of knowledge (Wagner-Pacifici et al. 2015). Not differently from what has happened so far in every other scientific domain, hypothesis testing has been the principal mode of producing knowledge in the social sciences. The study of the nexus between social networks and politics has not escaped the moral imperative to recognize and proceed according to existing theoretical accounts, either to confirm or confute them (Goldberg 2015). However, current technological advancements endow us with a twofold and unprecedented form of creativity: on the one hand, the possibility to research social and political dynamics in spaces of action where we never thought we could access as observers (i.e., the bulk of daily individual practices of interactions via communication); on the other, the possibility to refine the theories we have so far followed to produce better explanations and, possibly predictions, of social and political trends. This twofold creativity is particularly relevant to the study of the nexus between social networks and politics. On the one hand, with all the caveats induced by other binaries, we can search and unveil dynamics of public purpose production even in the most extemporary communication act, even in a simple "like" attached to a post of a petition we have not signed (yet). On the other hand, we can investigate how democracy is progressing (or threatened) in a hyperconnected (not simply interconnected) context like the one we live now by looking at data that have not generated following any a-priori theoretical approach but, rather, are true traces of technology-mediated social behaviors. In this situation, we do not derive any conclusion from an already settled theoretical mindset, nor we do infer any new theory. Rather, we put data and theory in a dynamic relationship and progressively tune one on the other, to proceed dynamically in the generation of knowledge and insights on our social, and political, ways of being.

Cross-References

- ▶ Community Detection: Current and Future Research Trends
- ▶ Community Evolution
- ▶ Desire Lines in Big Data
- ▶ E-Government
- ▶ International Hyperlink Networks
- ▶ Mapping Online Social Media Networks
- ▶ Network Social Capital
- ▶ Online Communities
- ▶ Policy Networks: History
- ▶ Political Networks
- ▶ Social Capital
- ▶ Social Media Analysis for Monitoring Political Sentiment
- ▶ Social Media and Social Networking in Political Campaigns/Movements
- ▶ Social Network Analysis in an Age of Digital Information
- ▶ Web Communities Versus Physical Communities

References

Adam S, Kriesi H (2007) The network approach. In: Sabatier PA (ed) Theories of the policy process. Westview, Boulder, pp 129–154

Baldassarri D (2009) Collective action. In: Hedstrom P, Bearman P (eds) Oxford handbook of analytic sociology. Oxford University Press, Oxford, pp 391–417

Benkler Y, Nissembaum H (2006) Commons-based peer production and virtue. J Political Philos 14(4):394–419

Bennett LW, Segerberg A (2014) Three patterns of power in technology-enabled contention. Mobilization: An International Journal 19(4):421–439

Bennett LW, Segerberg A (2013) The logic of connective action. Cambridge University Press, New York

Bosi L, Giugni M, Uba K (2016) The consequences of social movements. Cambridge University Press, Cambridge (UK)

boyd D, Crawford K (2012) Critical questions for big data. Inform Commun Soc 15(5):662–679

Bijker WB (2006) Why and how technology matters. In: Goodin RE, Tilly C (eds) The Oxford handbook of contextual political analysis. Oxford University Press, Oxford, pp 681–706

Breiger RL (2015) Scaling down. Big Data Soc 2(2):1–4

Börzel T (1998) Organizing Babylon: on the different concepts of social networks. Public Adm 76:253–273

Caiani M, Parenti L (2013) European and American extreme right groups and the Internet. Ashgate, Farnham

Castells M (2012) Networks of outrage and hope. Polity Press, Cambridge

Castells M (2011) A network theory of power. Int J Commun 5:773–787

Castells M (2009) Communication power. Oxford University Press, Oxford

Chadwick A (2013) The hybrid media system. Oxford University Press, Oxford

Diani M (2015) The cement of civil society. Cambridge University Press, New York (NY)

Diani M (2011) Networks and internet into perspective. Swiss Political Sci Rev 17(4):469–474

Diani M (2008) Modelli di azione collettiva: quale specificità per i movimenti sociali? Partecipazione e Conflitto 1:43–66

Diani M (2003) Networks and social movements: a research programme. In: Diani M, McAdam D (eds) Social movements and networks: relational approaches to collective action. Oxford University Press, Oxford, pp 299–319

Diani M, Bison I (2004) Organizations, coalitions and movements. Theor Soc 33:281–309

Finnemore M, Sikkink K (1998) International norm dynamics and political change. Int Organ 52(4):887–917

Gamson WA (1961) A theory of coalition formation. Am Sociol Rev 26(3):373–382

Gladwell M (2010) Small change: why the revolution will not be tweeted. *New Yorker*. Retrieved at http://www.newyorker.com/magazine/2010/10/04/small-change-malcolm-gladwell. Accessed Oct 2015

Goldberg A (2015) In defense of forensic social science. Big Data Soc 2(2):1–3

González-Bailón S, Borge-Holthoefer J, Moreno Y (2013) Broadcasters and hidden influentials in online protest diffusion. Am Behav Sci 57(7):943–965

González-Bailón S, Borge-Holthoefer J, Rivero A, Moreno Y (2011) The dynamics of protest recruitment through an online network. Scientific Reports 1(197):1–7

Held D, McGrew A, Goldblatt D, Perraton J (1999) Global transformations: politics, economics, and culture. Polity, Cambridge

Hemmati M (2002) Multi-stakeholder processes for governance and sustainability: beyond deadlock and conflict. Earthscan, London

Hockings B (2006) Multistakeholder diplomacy: forms, functions and frustrations. In: Kurbaljia J, Katrandjiev V (eds) Multistakeholder diplomacy: challenges and opportunities. Diplo Foundation, La Valletta, pp 13–32

Howard PN, Hussain MM (2013) Democracy's fourth wave. Oxford University Press, Oxford

Kahler M (2009) Introduction: networked politics, agency, power and governance. In: Kahler M (ed) Networked politics. Cornell University Press, Ithaca/London, pp 1–20

Kenis P, Schneider V (1991) Policy network and policy analysis: scrutinizing a new analytical toolbox. In: Marin B, Mayntz R (eds) Policy networks. Empirical evidence and theoretical considerations. Westview, Boulder, pp 25–62

Knoke D, Pappi FU, Broadbent J, Tsujinaka Y (1996) Comparing policy networks. Cambridge University Press, Cambridge

Lotan G, Graeff E, Ananny M, Gaffney D, Pearce I, Boyd D (2011) The revolutions were tweeted: information flows during the Tunisian and Egyptian revolutions. Int J Commun 5:1375–1405

Marin B, Mayntz R (1991) Introduction: studying policy networks. In: Marin B, Mayntz R (eds) Policy networks: empirical evidence and theoretical considerations. Westview, Boulder, pp 1–24

Melucci A (1996) Challenging codes: collective action in the information age. Cambridge University Press, Cambridge

Milan S (2015) From social movements to cloud protesting: the evolution of collective identity. Inform Commun Soc 18(8):887–900

Milan S (2013) WikiLeaks, anonymous, and the exercise of individuality: protesting in the cloud. In: Brevini B, Hintz A, McCurdy P (eds) Beyond Wikileaks: implications for the future of communications, journalism and society. Palgrave Macmillan, Basingstoke (UK), pp 191–208

Monge P, Contractor N (2003) Theories of communication networks. Oxford University Press, Oxford

Morozov E (2009) Why promoting democracy via the Internet is often not a good idea. *Foreign Pol*, 24 Apr http://neteffect.foreignpolicy.com/posts/2009/04/24/why_promoting_democracy_via_the_internet_is_often_not_a_good_idea

Padovani C, Pavan E (2016) Global governance and ICTs: exploring online governance networks around gender and media. Global Networks 16(3):350–371

Padovani C, Pavan E (2012) International norms and socio-technical systems: connecting institutional and technological infrastructures in governance processes. In: ICT critical infrastructures and society. IFIP advances in information and communication technology – Proceeding of the 10th IFIP Human Choice and Computers International Conference, Amsterdam

Pavan E (2014) Embedding digital communications within collective action networks: a multidimensional network perspective. Mobilization 19(4):441–455

Pavan E (2013) Collective action and web 2.0: an exploratory network analysis of Twitter use during campaigns. Sociologica 7(3):1–29

Pavan E (2012) Frames and connections in the governance of global communications: a network study of the internet governance forum. Lexington, Lanham

Powell W (1990) Neither market nor hierarchy: network forms of organization. Res Organ Behav 12:295–336

Rainie L, Wellman B (2012) Networked: the new social operating system. The MIT Press, Cambridge

Slaughter A (2004) A new world order. Princeton University Press, Princeton

Sørensen E, Torfing J (2007) Introduction governance network research: towards a second generation. In: Sörensen E, Torfing J (eds) Democratic network theories. Palgrave Macmillan, London, pp 1–24

Steinert-Threlkeld ZC, Mocanu D, Vespignani A, Fowler J (2015) Online social networks and offline protest. EPJ Data Sci 4:19. https://doi.org/10.1140/epjds/s13688-015-0056-y

Tremayne M (2014) Anatomy of protest in the digital era: a network analysis of Twitter and Occupy Wall Street. Soc Mov Stud 13(1):110–126

Vaccari C, Valeriani A (2016) Party campaigners or citizen campaigners? How social media deepen and broaden party-related engagement. Int J Press/Polit. https://doi.org/10.1177/1940161216642152

Wagner-Pacifici R, colleagues (2015) Ontologies, methodologies, and new uses of big data in the social and cultural sciences. Big Data Soc 2015:1–11

Wellman B (2002) Structural analysis: from method and metaphor to theory and substance. In: Scott J (ed) Social networks: critical concepts in sociology, Vol I. Routledge, London/New York, pp 81–122

Social Networks for Quantified Self

Ted Vickey and John Breslin
Digital Enterprise Research Institute, National University of Ireland at Galway, Galway, Ireland

Glossary

Connected health	Health care through the use of technology
mHealth	Mobile health
Mobile fitness apps	Mobile fitness applications used from a smartphone or website

Definition

Over three quarters of US health care spending goes to the care of people with chronic conditions, including heart disease, diabetes, and asthma, while in 2004, nearly half of the Americans were diagnosed with one or more chronic conditions, a number expected to increase dramatically as the baby boomer generation rapidly approaches their retirement age (Accenture 2009). The new reality, dubbed “Connected Health,” incorporates a broad range of health and fitness applications that are always on, always active, and always aware (Accenture 2009).

Since many aspects of health promotion professionals involve interdependent actors, social networks are of increasing interest to health services researchers (O'Malley and Marsden 2008). The creation of a social network map of a person's social network can help visualize and thus better understand the strengths of the social ties of the network (Christakis and Fowler 2009).

Technology Will Transform the Future of Chronic Care

In a 1995 editorial in the American Journal of Public Health, former US Surgeon General C. Everett Koop stated, "Cutting-edge technology, especially in communication and information transfer, will enable the greatest advances yet in public health. Eventually, we will have access to health information 24 hours a day, 7 days a week, encouraging personal wellness and prevention, and leading to better informed decisions about health care" (Koop 1995). Technologies like miniaturized health sensors, broadband networks, and mobile devices are enhancing and creating new health-care capabilities such as remote monitoring and online care (Accenture 2009).

In 2009, management and technology consultant Accenture released a report on how technology will transform the future of chronic care. Cited in the report is the anticipated crisis in care that will be further challenged as the baby boomer generation begins to retire.

According to the US Census Bureau, the world's population of people age 65 and older is projected to triple by mid-century, from 516 million in 2009 to 1.53 billion in 2050. This growing trend places a tremendous economic burden on governments, private employers and individual consumers alike. It also puts strain on the capacity of skilled care professionals and nursing homes (Accenture 2009).

In addition to the inexpensive cost of computers and Internet connectivity, the report identifies three technological advancements that are paramount to the future of chronic care:

- Seamless capture and sharing of patient information in real-world settings
- Improvements in ways to combine and interpret data about an individual's health and wellness so that appropriate interventions can be made before an acute situation occurs
- Innovative tools including user modeling, advanced visualization, decision support, and collaboration

Health and Social Networking

One aspect of "Connected Health" is via the power of a person's social network. Research suggests that people interact with their social network with regard to their health. Christakis and Fowler (2009) concluded that ". . . a person with more friends and social contacts generally has better health than a person with fewer friends, and a person at the center of a network is more susceptible to both the benefits and risks of social connection than those at the periphery of a network." This would suggest that a person is not only affected by their location in a social network but also influenced by the behaviors of those who are "close" to them in the network. Perceived social support and physical activity are directly associated with a person's perceived health status (Almeida 2008).

As technology continues to impact humanity, the understanding of one's social network may be one key to better health. The basic element of a person's social network is simple: a social network starts with a central person (called an ego) and other people (called nodes) that are interconnected by links (called ties). As the numbers of nodes and links increase, the number of possible connections grows exponentially – known as the network effect (Christakis and Fowler 2009).

Christakis and Fowler (2009) suggest that "people are inter-connected and so their health is inter-connected. Inter-personal health effects in social networks provide a new foundation for public health." As online connections between people become ever more interweaved with offline real-world interests, social networking methods are moving towards simulating real-life social interactions, including physical activity, health, and disease management: rather than randomly approaching each other, people meet through things they have in common (Breslin and Decker 2007).

Technology and Health Behavior Modification

By using Mobile Health technology (mHealth), health providers can practice a more “personalized medicine” and potentially reach more individuals with effective health-related advice and information at a very low cost (Strecher 2007). Griffiths et al. (2006) suggest a number of reasons for delivering web-based health, wellness, and fitness interventions including reduced delivery costs, convenience to users, timeliness, reduction of stigma, and reduction of time-based isolation barriers.

Technologies can play three roles with regard to behavior modification: as tools, as media, and as social actors.

- As a tool, interactive technologies can be persuasive by making target behavior easier, leading people through a process, or performing calculations/measurements that motivate.
- As a medium, interactive technologies can be persuasive by allowing people to explore cause-and-effect relationships, providing people with experiences that motivate, or helping people to rehearse a behavior.
- As a social actor, interactive technologies can be persuasive by rewarding people with positive feedback, modeling a target behavior or attitude, and providing a social network of support (Fogg 2002).

Within the health-care field, interactive technologies can be effectively deployed to take on multiple roles at the same time. For example, a simple persuasive tool can measure calories while at the same time giving a reward upon attainment of a personal goal. This type of self-monitoring is a key ingredient in successful behavioral modification. In addition, if several people are connected through the Internet, then social support can be leveraged, which has been shown to impact motivation and behavior change (Chatterjee and Price 2009).

The Quantified Self

The idea of measuring things relative to a business or personal goal is common in today’s society. The same measurement tools can be used within the self-tracking of a person’s health and fitness. Commonly known as the Quantified Self movement, this is eclectic mix of early adopters, fitness fanatics, technology evangelists, personal development junkies, hackers, and patients suffering from a wide range of health challenges (The Quantified Self – Counting Every Moment 2012). Some measure their hourly mood swings, while others the stages of their nightly sleep habits. Some track every meal, snack, or drink, while others share on Twitter and Facebook their workout routine complete with heart rate, time, distance, calories burned, and musical preferences.

Ongoing research aims to classify and understand why a person shares their workouts within their social network via Twitter and the associated benefits. While there are various personal devices that monitor/track a person’s exercise characteristics (e.g., Body Media, Fitbit, MapMyFitness, and Nike+), the effectiveness of online sharing via social networks of one’s physical activity is limited in scientific research. Studies have indicated that “lack of motivation” is a key factor in why a person does not exercise.

One factor to address is the relationship between participant and provider (i.e., personal trainer) and/or participant and social network, including their influence. People join gyms not only for health and fitness but also for the social atmosphere. To fully understand the power of combining social networking and exercise adherence, the physical barrier of the four walls of an exercise facility is removed, and technology is used that enables a measurable improvement towards one’s fitness goals.

Conclusion

With the move towards making machine-understandable data available for computers, allowing exercise data to become accessible/exchangeable between trusted peers is quite important. However, one's historical exercise records are often locked in to proprietary systems. By publishing selected aspects of these profiles using semantic terms, it will become easier for people to search for and discover relevant exercise regimes.

Early prevention and healthy lifestyles may be the least expensive and best ways to combat the growing prevalence of avoidable diseases associated with a lack of physical activity including obesity (Almeida 2008). If people who lead sedentary lives would adopt a more active lifestyle, there would be enormous benefit to the public's health and to individual well-being. An active lifestyle does not require a regimented, vigorous exercise program. Instead, small changes that increase daily physical activity will enable individuals to reduce their risk of chronic disease and may contribute to enhanced quality of life (Pate et al. 1995).

Cross-References

- ▶ Actionable Information in Social Networks, Diffusion of
- ▶ Data Mining
- ▶ Twitter Microblog Sentiment Analysis

References

Accenture (2009) Always on, always connected: how technology will transform the future of chronic care. Accenture, New York

Almeida F (2008) The relationship between social networks, social support, physical activity and self-rated health: an exploratory study. University of Denver, Boulder

Breslin J, Decker S (2007) The future of social the need for semantics. IEEE Internet Comput 5:86–90

Chatterjee S, Price A (2009) Healthy living with persuasive technologies: framework, issues, and challenges. J Am Med Inform Assoc 16(2):171–178. https://doi.org/10.1197/jamia.M2859

Christakis NA, Fowler JH (2009) Social network visualization in epidemiology. Health Care 19(1):5–16

Fogg B (2002) Persuasive technology: using computers to change what we think and do. Ubiquity, 5 Dec 2002. https://doi.org/10.1145/763955.763957

Griffiths F, Lindenmeyer A, Powell J, Lowe P, Thorogood M (2006) Why are health care interventions delivered over the internet? A systematic review of the published literature. J Med Internet Res 8(2):e10. https://doi.org/10.2196/jmir.8.2.e10

Koop CE (1995) A personal role in health care reform. Am J Public Health 85(6):759–760. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1615490&tool=pmcentrez&rendertype=abstract. Accessed 13 June 2012

O'Malley A, Marsden P (2008) The analysis of social networks. Health Serv Outcome Res Methodol 8(4):222–269

Pate RR, Pratt M, Blair SN, Haskell WL, Macera CA, Bouchard C, Buchner D et al (1995) Physical activity and public health: a recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine. JAMA 273(5):402–407. https://doi.org/10.1001/jama.1995.03520290054029

Strecher V (2007) Internet methods for delivering behavioral and health-related interventions (eHealth). Annu Rev Clin Psychol 3:53–76. https://doi.org/10.1146/annurev.clinpsy.3.022806.091428

The Quantified Self — Counting Every Moment (2012) Economist. Retrieved from http://www.economist.com/node/21548493. Accessed 15 Sept 2012

Social Networks in Emergency Response

Dashun Wang[1,2], Yu-Ru Lin[3] and
James P. Bagrow[4]
[1]Kellogg School of Management, Northwestern University, Evanston, IL, USA
[2]Northwestern Institute on Complex Systems (NICO), Northwestern University, Evanston, IL, USA
[3]School of Information Sciences, University of Pittsburgh, Pittsburgh, PA, USA
[4]Mathematics and Statistics and Vermont Complex Systems Center, University of Vermont, Burlington, VT, USA

Synonyms

Collective response; Communication Network; Data mining; Disaster; Emergency; Event detection; Spatiotemporal analysis; Social networks.

Glossary

Emergency	An unexpected and often dangerous situation, typically affecting multiple individuals and requiring immediate action
Social and Communication Networks	Networks of people interacting with each other through web-based (e.g., Twitter) and mobile-based (e.g., mobile phone) technologies
Social Media	Web-based tools that enable people to communicate and interact with each other in various media forms including text and multimedia. Examples of these tools include emails, instant messengers (IM), blogs, microblogs (e.g., Twitter), vlogs (e.g., YouTube), podcasts, forum, wikis, social news (e.g., Digg), social bookmarking (e.g., Delicious), and social networks (e.g., Facebook, MySpace, and LinkedIn)

Definition

Modern datasets derived from telecommunication technologies such as online social media and mobile phone systems offer a great potential to understand the behaviors of large populations during emergencies and disasters. This entry reviews recent studies using large-scale, modern data to understand emergency and disaster response, covering work focused on social network activity during earthquakes and disease outbreaks and mobile phone communications following bombing and other emergency events. The key techniques and research trends are also discussed.

Introduction

Large-scale emergencies and disasters are an ever-present threat to human society. With growing populations and looming threat of global climate change, the numbers of people at risk will continue to grow. Thus there is a great need to optimize response efforts from search and rescue to food and resource disbursement. Human dynamics research offers a promising avenue to understand the behaviors of large populations, and modern datasets derived from cutting-edge telecommunications such as online social media and pervasive mobile phone systems bring a wealth of potential new information. Such massive data offers a promising complement to existing research efforts in disaster sociology, which primarily focus on eyewitness interviews, surveys, and other in-depth but small-scale data (Rodríguez et al. 2006).

Yet most current human dynamics research is focused primarily on data collected under normal circumstances, capturing baseline activity patterns. Here we review a number of studies pushing the envelope of modern data into the realm of unexpected deviations in these population behaviors. We discuss research focused on massive datasets from social network activity during earthquakes and disease outbreaks to mobile phone communications following bombings, power outages, and more.

We review a number of recent studies using large-scale, modern data to understand emergency and disaster response. We begin with a review quantifying how expectations of communication in today's world may influence our perception of the severity of an emergency. We then cover works focused on social media and mobile phones. These works use Twitter, a prominent online social media service, to understand more about disease outbreaks and the impact of earthquakes. The mobile phone studies feature a number of emergencies, including earthquakes, bombings, and a plane crash. The results of these studies have the potential to revolutionize disaster response in the future, with the critical goal of saving lives.

Historical Background

Connectivity and information access through global telecommunications have become increasingly pervasive due to modern technologies such as mobile phones and the Internet. People are becoming increasingly reliant on these communication modes and so an important question asked by Sheetz et al. (2010) is as follows: what do people expect about their access to these communication channels when an emergency occurs? They explored how the expectation of the availability of these communication technologies may influence their perceptions of how they would use these technologies during and after a crisis.

To answer this question, the authors conducted online surveys and follow-up interviews with Virginia Tech students, faculty, and staff (participants). This university suffered a tragic attack on April 16, 2007, and the authors reported that local cellular networks were overwhelmed by traffic. Surveying witnesses and survivors at the university allows the authors to study how the perceptions of information access meshed with the unfortunate events that occurred.

Through these surveys and interviews, they found that participants have a range of expectations for connectedness in normal activities. Most participants did not expect to be able to immediately contact someone. This held even for strong social ties, for example, a student trying to reach his or her parents. Most importantly, the authors discovered that participants who do have high expectations of connectivity (and also tend to be more extroverted individuals) were more likely to report problems with connectivity than users with lesser expectations. These problems can lead these people to form overestimate of the severity of the crisis, compared with individuals who have lower expectations for their communication and are thus less likely to find communication loss a cause for concern. This means that an individual's personal traits may directly influence how he or she estimates the severity of a crisis.

While the authors admitted that they had a small sample size and that their interview methods may not be perfect, this study is an important step towards further understanding the interplay between modern telecommunications and emergency events.

Emergencies and Social Media

Today, social media such as Twitter and Facebook have been popularly used as everyday communication tools. Millions of people use "tweets" or Facebook "statuses" to inform family, friends, colleagues, or any others about information, opinion, and emotions about events just happening, leading to the great potential of using social media for monitoring and rescue purposes. Twitter allows users to send and receive tweets (140-character messages) via text messages and Internet-enabled devices, providing the public with detailed anecdotal information about their surroundings. Given the real-time nature of Twitter and the emerging social networking technologies, social media has the potential to fundamentally alter our discussions of emergencies. We briefly review some of the recent work on detecting disease outbreaks and earthquake response with Twitter.

Twitter and Disease Outbreaks

Various studies have shown the potential of using Twitter data to monitor the current public health status of a population, as people often tweet when they feel ill or recognize disease symptoms. Quincey and Kostkova (2010) collected tweets that contained instances of the keyword "flu" in a week during the swine flu pandemic. Their study suggests that the copresence of other words in tweets can be used by public health authorities to gather information regarding disease activity, early warning, and infectious disease outbreak. For example, in the majority of the collected tweets, the word "swine" was present along with "flu"; the words "have flu" and "has flu" may indicate that the tweet contains information about the users or someone else having flu. The words "confirmed" and "case(s)" perhaps indicate a number of tweets that are publicizing "confirmed cases of swine flu." Culotta (2010) collected over 500,000 influenza-related tweets during 10 weeks and analyzed the correlation between these messages and the Centers for Disease Control and Prevention (CDC) statistics. The paper reported a correlation of 0.78 by leveraging a document classifier. Chew and Eysenbach (2010) collected over 2 million tweets containing the keywords "H1N1," "swine flu," and "swinflu" within 8 months in 2009. Using manual and automated content coding, they found temporal correlation of Twitter activity with major news stories and H1N1 incidence data. In addition, they found that the majority of these tweets contained resource-related posts (e.g., links to news websites). Gomide et al. (2011) analyzed how the dengue outbreaks in 2009 were mentioned on Twitter. Using a linear regression model, they showed promising results to predict the number of dengue cases by leveraging tweet content and spatiotemporal information. Signorini et al. (2011) tracked time-evolving public sentiments about H1N1 or swine flu and studied the probability of using Twitter stream for real-time estimation of weekly influenza-like illness (ILI) statistics generated by CDC.

There has also been work addressing the technical challenges of collecting tweets that are related to health or disease. Zamite et al. (2011) proposed a system architecture for collecting and integrating epidemiological data based on the principles of interoperability and modularity. Prier et al. (2011) proposed using a Latent Dirichlet Allocation (LDA) model to effectively identify health-related topics in Twitter. Paul and Dredze (2011) collected two billion tweets related to illness, disease symptoms, and treatment from May 2009 to October 2010. They proposed a probabilistic aspect model to separate tweets related to health from unrelated tweets. Aramaki et al. (2011) collected 300 million tweets from 2008 to 2010. They applied the Support Vector Machines (SVMs) to find tweets related to influenza with a correlation of 0.89% compared with Google Flu Trends (Ginsberg et al. 2008). These tools offer the means to transform the overwhelming flood of big data into more manageable information.

Besides social media, there are also other solutions to estimate a population's health from Internet activity, most notably Google Flu Trends service, which correlates search term frequency with influenza statistics reported by the CDC (Ginsberg et al. 2008).

Twitter and Earthquakes

In recent years, tremendous effort has been made towards leveraging Twitter to study earthquakes, mainly falling into two lines of research: real-time detection (Sakaki et al. 2010; Guy et al. 2010; Earle et al. 2012) and crisis management (Hughes and Palen 2009; Caragea et al. 2011; Li and Rao 2010; Mendoza et al. 2010).

Early earthquake detection and the delivery of timely alerts is an extremely challenging task. Depending on peculiarities of the earthquake, from size to location, alerts may take between 2 and 20 min to publish, owing to the propagation time of seismic energy from the epicenter to seismometers and the latencies in data collection and validation. Therefore, it has been practically impossible for affected populations to know about an earthquake before it arrives. This situation is changing, however, thanks to the pervasive use of Twitter. Users submit their tweets via text messages and Internet-enabled devices, and these messages are available to their followers and the public within seconds, making Twitter an ideal environment for the dissemination of breaking news to large populations. Therefore, by using populations as social sensors, Twitter may be a viable tool for rapid assessment, reporting, and potentially real-time detection of a hazard event. Sakaki et al. (2010) investigated events such as earthquakes and typhoons in Twitter and proposed an algorithm to monitor tweets and to detect earthquakes. They extracted features such as keywords in a tweet by semantic analysis and used Support Vector Machines (SVMs) to classify a tweet into a positive or negative class. By regarding a tweet as a social sensor associated with location information, the authors transformed the earthquake detection problem into an object detection problem in ubiquitous and pervasive computing. They derived a probabilistic model by applying Kalman filtering and particle filtering to estimate the epicenter of an earthquake and the trajectories of a typhoon. They then deployed an earthquake reporting system in Japan, which delivers earthquake notifications to their users faster than the announcementsbroadcast by Japan Meteorological Agency. Meanwhile, researchers from the US Geological Survey (USGS) reported an earthquake detection system that adopts social network technologies, called Twitter Earthquake Detector (TED) (Guy et al. 2010; Earle et al. 2012). They downloaded tweets that contain the words "earthquake," "gempa," "temblor," "terremoto," or "sismo" from August to the end of November 2009. Based on tweet-frequency time series, they used a short-term-average, long-term-average algorithm to identify earthquakes, finding 48 earthquakes around the globe with only 2 false triggers in 5 months of data. The detections are faster than seismographic detections, with 75% occurring within 2 min. These results demonstrate the efficiency of using Twitter as a detection tool, potentially achieving better and more accurate results when combined with existing systems.

The rich semantics of tweets and Twitter's broadcasting nature also hint at the potential of using Twitter for rapid emergency response tools to assist in intervention and crisis management. Caragea et al. developed a reusable information technology infrastructure, called Enhanced Messaging for the Emergency Response Sector (EMERSE) Caragea et al. (2011). The system is aimed at classifying tweets and text messages automatically, together with the ability to deliver relevant information to relief workers. EMERSE has four components, including an iPhone application, a Twitter crawler, machine translation, and automatic message classification. The system analyzed the information about the Haiti earthquake relief and provided their output to NGOs, relief workers, and victims and their friends and relatives in Haiti. To use Twitter as an emergency response tool, it is important to assess the information quality of tweets during an emergency situation. Li and Rao (2010) studied Twitter usage following the Sichuan earthquake in China in 2008. They focused on five information quality dimensions: timeliness, accessibility, accuracy, completeness, and collective intelligence, arguing that Twitter is an effective tool for information dissemination in critical moments following earthquake and its broadcasting nature plays an important role in emergency response. Mendoza et al. (2010) studied the dissemination of false rumors and confirmed news following 2010 Chile earthquake, finding that false rumors tend to be questioned much more than confirmed news.

Their study indicates the possibility of using Twitter to detect rumors after an earthquake to make the rescue efforts more efficient.

Emergencies and Mobile Phones

In addition to social media websites, the pervasive adoption of mobile phones provides another potentially even more detailed avenue to monitor large populations. Mobile phone records usually include fine-grained longitudinal mobility traces and communication logs. The data allows greater opportunity to study personal social networks through their relationship with physical space, compared to the online social networks (e.g., "friends" and "followers" on Twitter). Mobile phones are well established in many areas, even in third world countries such as Rwanda (Kapoor et al. 2010). Leveraging their presence to assist in emergency response has great potential to save lives. Here we review two recent papers focused on mobile phones and emergencies. The first studied an earthquake that occurred in central Africa (Kapoor et al. 2010). The second analyzed a corpus of events including non-emergency controls such as music festivals occurring in Western Europe (Bagrow et al. 2011).

An Earthquake in Central Africa

To understand how effective mobile phones are at understanding emergency situations, a number of studies have been conducted. Kapoor et al. (2010) studied a 5.9 magnitude earthquake that occurred February 3, 2008, in Lake Kivu region of the

Democratic Republic of Congo Kapoor et al. (2010). The dataset is the cellular activity patterns of mobile phone users in Rwanda. They used daily call volume on a per tower basis, and they also had the geographic coordinates of the towers. Their goal was to determine the location of the epicenter algorithmically using only the cellular data and to assess or predict what areas of the country are most in need of aid due to the earthquake.

To study these problems, they assumed that (i) cell tower traffic deviates in a statistically significant manner from normal activity levels and trends when an event occurs, (ii) areas that are more disturbed by the event will display traffic deviations for longer periods of time, and (iii) disruptions are inversely proportional to the distance from the catastrophe.

To detect an event they assumed the typical daily traffic on a tower obeys a gaussian distribution and they used a negative log-likelihood score to compare the current traffic with this distribution. The higher this score, the more likely there was an anomalous event on that day. They demonstrated that this score spikes on the day of the event, although they did not discuss a specific algorithm to automatically flag scores (e.g., introducing a threshold score such that an event is anomalous when its score exceeds that threshold).

To estimate the location of the event, they assumed the activity levels at a tower during the event follow a normal or gaussian distribution but that the mean of this tower's distribution is now a function of the distance from the epicenter. Specifically they used for tower *i* a distance-dependent mean $m\,i + \alpha D\,i.\,e\,x;(\,e\,y\,)^{-1}$, where $m\,i$ is the normal mean traffic for *i*, α is some configurable scaling parameter, and $D\ i\ (\ e\ x\ ,\ e\ y\ )$ is the geographic distance of tower *i* from an epicenter located at coordinates $(\,e\,x\,,e\,y\,)$. They determined this epicenter $(\ e\ x,\ e\ y\ )$ (and also α) using well-established maximum likelihood estimates, that is, they found the epicenter and scaling parameters that maximize the sum of the log's of all the tower's probabilities.

The other problem they wish to address is to predict what areas are most in need of emergency aid. To do this, they want to predict whether a particular tower will experience a significant increase in traffic some number of days after the event. They accomplished this by building a classifier which allows them to estimate this persistence probability. Since it is reasonable to assume that areas with higher populations are likely to require more aid, they built an "assistance opportunity score" for a location by taking the product of the persistence probability estimate for that location and the population at that location. Such a score allows emergency responders to potentially prioritize aid efforts.

The authors also pointed out an important issue when using mobile phone data to study these problems: the density of towers, and therefore information, is not uniform. Cities have many more towers than rural regions, and this leads to far greater granularity in areas of high population and greater information uncertainty in areas with fewer towers. They exploited this fact to estimate what areas are most valuable to survey manually for information after an event, by prioritizing surveys towards areas with more uncertainty. They did this by devising a simple mechanism to drive down the entropy in the information that may be gained from the system, and they even incorporated geographic distances since it is more expensive in terms of time and effort to survey more remote regions.

All of their methods were validated by comparison with the February 3 earthquake and were shown to work rather well. For future work they discussed a number of interesting advancements such as incorporating richer models of geographic terrain.

Mobile Phones and Disasters

Bagrow et al. (2011) performed a data-driven analysis of a number of emergencies, including bombings, a plane crash, and another earthquake. This work reported a number of empirical discoveries regarding the response of populations in the wake of emergencies (and non-emergency control events such as festivals), as measured from the country-wide data of a single mobile phone provider in Western Europe. The assumptions made by Kapoor et al. (2010) are further justified by their work.

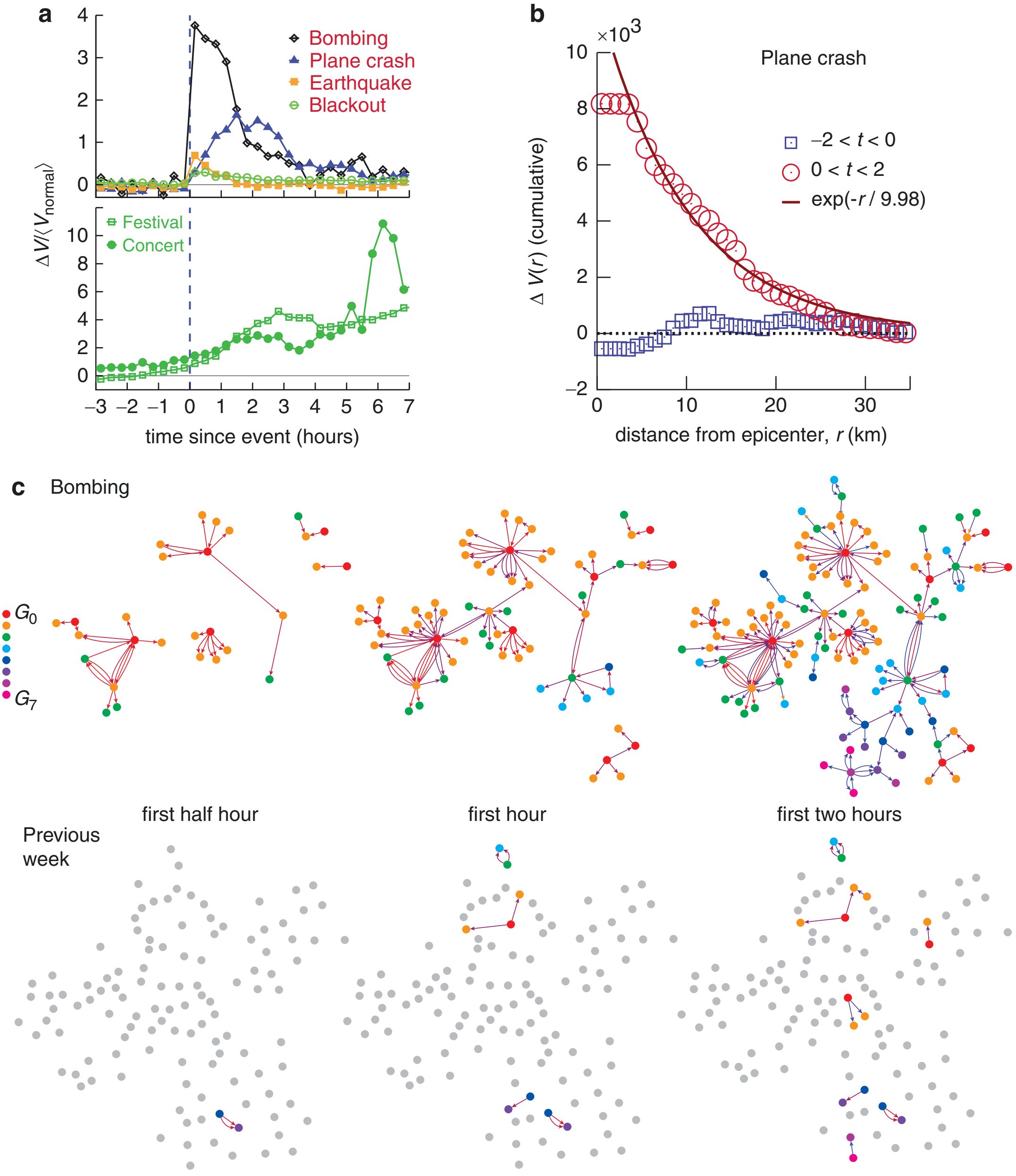
They found that emergencies trigger a sharp spike in call activity (number of outgoing calls and text messages) in the physical proximity of the event, confirming that mobile phones act as sensitive local "sociometers" to external societal perturbations. In Fig. 1a, we plot the relative call volume $\Delta\ V/\langle V_{\text{normal}}\rangle$ as a function of time, where $\Delta\ V= V_{\text{event}} - \langle V_{\text{normal}}\rangle$, V_{event} is the number of calls made from nearby towers during the event, and $\langle V_{\text{norma}}\rangle$ is the average call volume during the same time period of the week (Figure adapted from Bagrow et al. (2011)).

The anomalous traffic starts to decay immediately after the emergency occurs, suggesting that the urge to communicate is strongest right at the onset of the event. There was virtually no delay between the onset of the event and the jump in call volume for events that were directly witnessed by the local population, such as the bombing, the earthquake, and the blackout. Brief delay was observed only for the plane crash, which took place in an unpopulated area and thus lacked eyewitnesses. In contrast, non-emergency events, like the festival and the concert, displayed a gradual increase in call activity.

The temporally localized spikes in call activity (Fig. 1a) raise an important question: is information about an event limited to the immediate vicinity of the emergency or do emergencies, often immediately covered by national media, lead to spatially extended changes in call activity (Petrescu-Prahova and Butts 2008)? To investigate this, Bagrow et al. inspected the change in call activity in the vicinity of each event's epicenter, finding that for the bombing, for example, the change in call volume is strongest near the event and drops rapidly with the distance r from the epicenter. To quantify this effect across all emergencies, they integrated the call volume over time in concentric shells of radius r centered on the epicenter. The observed decay in anomalous traffic was approximately exponential, $\Lambda\ V(\ r) \sim \exp.(-\ r/r_{\ c})$, allowing one to characterize the spatial extent of the reaction with a decay rate $r\ c$ (we present their results for the plane crash in Fig. 1b). The observed decay rates ranged from 2 km (bombing) to 10 km (plane crash), indicating that the anomalous call activity is limited to the event's vicinity. An extended spatial range ($r\ c$ ≈110 km) was seen only for the earthquake. Meanwhile, nonemergencies are highly localized: they possess decay rates less than 2 km. This systematic split in $r\ c$ between the spatially extended emergencies and well-localized non-emergencies persisted for all explored events.

Despite the clear temporal and spatial localization of anomalous call activity during emergencies, one expects some degree of information propagation beyond the eyewitness population. To study how emergency information diffuses through a social network, Bagrow et al. used mobile phone records to identify those individuals located within the event region, forming a population called G_0 as well as a group called G_1 consisting of individuals outside the event region but who receive calls from the G_0 group during the event, a G_2 group that receive calls from G_1, and so on. They reveal that the G_0 individuals typically engage their social network within minutes and that the G_1, G_2, and occasionally even the G_3 group show an anomalous call pattern immediately after an emergency. We present their illustration of a segment of this contact network for the bombing in Fig. 1c. The authors proceeded to further quantify and control for this social propagation and showed that the bombing and plane crash have significant propagation up to the third and second neighbors of $G\ _0$, respectively. They found that other emergencies, the earthquake and blackout, displayed relatively little propagation. This seems reasonable given the less severe nature of those events (the earthquake was relatively minor).

Finally, we also presented a breakdown of a number of measurable features for each emergency and non-emergency and showed that these features may be used to distinguish anomalous call activity due to benign events such as music festivals from spikes in call volume that indicate a dangerous event has occurred. Using such factors may allow first responders to more accurately understand rapidly unfolding events and may even allow them to actively solicit information from mobile phone users likely to be near the event.



Social Networks in Emergency Response, Fig. 1 Social networks in emergency response Temporal, spatial, and social response during emergencies. (**a**) The time dependence of call volume *V*(*t*) after four emergencies and two non-emergencies. We plot the relative change in call volume $\Delta V/\langle V_{\text{normal}}\rangle$, where $\Delta V = V_{\text{event}} - \langle V_{\text{normal}}\rangle$, V_{event} is the call volume on the day of the event, and $\langle V_{\text{normal}}\rangle$ is the average call volume during the same period of the week. (**b**) The total change in call volume during 2-h periods before and after the plane crash, as a function of distance *r* from the epicenter of the crash. Following the event, we see an approximately exponential decay $\Delta V \sim\sim \exp r/r\ c$ characterized by decay rate *r c*. (**c**) Part of the contact network formed between mobile phone users in the wake of the bombing. Nodes are colored by group, with G_0 representing phone users calling from the event region, G_1 the recipients of those calls, etc. As time goes by more users are contacted as information propagates. Those same users make little contact during a corresponding time period the week before (Figure adapted from Bagrow et al. (2011))

Key Applications

We summarize the key techniques that have been used in the above-mentioned studies.

Event Detection The first challenge in large-scale emergency studies is to determine and collect a subset of data relevant to emergencies under consideration. With Twitter or other social media data where the communication content is available in text format, most studies begin with a simple keyword matching, that is, collecting data that contained instances of the relevant keywords such as “flu,” “H1N1,” and “earthquake.” The initial collections could be refined by manual and automated classification process. Classification techniques such as Support Vector Machines (SVMs) have been employed (Aramaki et al. 2011; Sakaki et al. 2010), and topic clustering methods such as Latent Dirichlet Allocation (LDA) can be used to improve the classification (Paul and Dredze 2011; Prier et al. 2011). Validation of this body of work is often conducted based on authority reports such as Centers for Disease Control and Prevention (CDC) statistics (Signorini et al. 2011) (for disease outbreaks) or US Geological Survey (USGS) reports (Guy et al. 2010; Earle et al. 2012) (for earthquakes). While the messages disseminated in social media might be inaccurate, there has been work on determining the quality of information sources (Li and Rao 2010; Mendoza et al. 2010). Further, by applying time-series analysis and spatiotemporal pattern analysis (e.g., Kalman filtering and particle filtering in Sakaki et al. (2010)), researchers have developed powerful earthquake detectors with performance comparative to existing earthquake detection systems.

Event Prediction and Forecasting The development of event prediction and forecasting is still in its early stage. Gomide et al. (2011) used a linear regression model to predict the number of dengue cases. The earthquake detectors (Sakaki et al. 2010; Guy et al. 2010; Earle et al. 2012) that reported earthquakes faster than the seismo-graphic detection can be used as early warning system. There has been work on developing information infrastructure which has the ability to deliver relevant information to users once events are detected (Caragea et al. 2011).

Spatiotemporal Pattern Recognition of Events Unlike social media data, the content of communication is often unavailable in mobile phone data, and hence the identification of emergency events in mobile phone data relies on analyses of spatial and temporal anomalies of call logs. The main challenge of this research is to construct reasonable null model in order to recognize anomaly events. Bagrow et al. (2011) proposed using pre-emergency normal activities as well as the activities during non-emergency events to contrast the activities of emergency events. Based on this approach the epicenter of an emergency event can be identified. Kapoor et al. (2010) used a similar methodology to identify event epicenters as well as to predict the locations in need of emergency aid.

Future Directions

Foundational work understanding the sociology of disaster was limited in scale by available data but surveys and interviews can ask a number of in-depth follow-up questions. To understanding population response from, for example, mobile phone call volume alone is potentially more challenging as such data, while perhaps being more objective, is also far shallower. This begs the question: can more depth be found in communications data? The wealth of textual information available within social media such as Twitter can be leveraged to learn more context about how populations respond to emergencies, and advances in data mining and natural language processing techniques offer the promise of even greater information. This may allow researchers to separate relevant information from spurious activity, improving the accuracy and precision of information available to rescuers.

One can reasonably expect a degree of noise from any communication system, as users will be focused on diverse topics. Yet when something of overwhelming importance occurs, such as an emergency, it seems reasonable to expect that event to capture the majority of user attention. This may lead to a communication system that is less noisy and more focused as the severity of the event increases, in the sense that an increasing fraction of the system' communication will be about that event. Given this, it may be worth trying to develop (rigorous) bounds on how much useful information can be successfully extracted from such a system during and immediately following an event. This could allow quantitative benchmarking of algorithms designed to assist rescuers by comparing, for example, how much emergency information was extracted by an algorithm with the maximum amount possible.

Meanwhile, it will be crucial going forward to develop algorithms that combine and help understand multiple data sources–such as cell phone call volume, twitter messages, and perhaps even security cameras, all from a given geographic locale. This trend towards greater data availability and unification will only continue as more advanced and entirely new forms of telecommunication come into widespread use. Without methods to handle the increased diversity and volume of communication, rescuers may be unable to capitalize on the extra information provided by future telecommunications.

Conclusion

We have reviewed a number of works focused on the use of communications data, from social media to mobile phones, to understand how people react to emergencies and disasters. This problem is of critical importance: in many areas of the world, more people than ever are at risk, as both human populations and threats due to climate change continue to grow. Hopefully tools derived from social media and other communication datasets will help rescuers improve their emergency and disaster response by providing accurate, useful, and timely information in the wake of such events.

Cross-References

- ▶ Actionable Information in Social Networks, Diffusion of
- ▶ Assessing Individual and Group Behavior from Mobility Data: Technological Advances and Emerging Applications
- ▶ Counterterrorism, Social Network Analysis in
- ▶ Disaster Response and Relief, VGI Volunteer Motivation in
- ▶ Modeling and Analysis of Spatiotemporal Social Networks
- ▶ Social Network Datasets
- ▶ Social Networking in the Telecom Industry
- ▶ Spatiotemporal Proximity and Social Distance
- ▶ Temporal Networks

References

Aramaki E, Maskawa S, Morita M (2011) Twitter catches the flu: detecting influenza epidemics using twitter. In: Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, Edinburgh, pp 1568–1576

Bagrow JP, Wang D, Barabsi A-L (2011) Collective response of human populations to large-scale emergencies. PLoS One 6(3):e17680

Caragea C, McNeese N, Jaiswal A, Traylor G, Kim H, Mitra P, Wu D, Tapia A, Giles L, Jansen B, et al. (2011) Classifying text messages for the Haiti earthquake. In: Proceedings of the 8th international ISCRAM conference, ISCRAM, Harbin, vol 11

Chew C, Eysenbach G (2010) Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. PLoS One 5(11):e14118

Culotta A (2010) Towards detecting influenza epidemics by analyzing twitter messages. In: Proceedings of the first workshop on social media analytics. ACM, Washington, DC, pp 115–122

Earle P, Bowden D, Guy M (2012) Twitter earthquake detection: earthquake monitoring in a social world. Ann Geophys 54(6):708–715

Ginsberg J, Mohebbi M, Patel R, Brammer L, Smolinski M, Brilliant L (2008) Detecting influenza epidemics using search engine query data. Nature 457 (7232):1012–1014

Gomide J, Veloso A, Meira W, Almeida V, Benevenuto F, Ferraz F, Teixeira M (2011) Dengue surveillance based on a computational model of spatio-temporal locality of twitter

Guy M, Earle P, Ostrum C, Gruchalla K, Horvath S (2010) Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies. In: Proceedings of the 9th international conference on advances in intelligent data analysis, Tucson, pp 42–53

Hughes A, Palen L (2009) Twitter adoption and use in mass convergence and emergency events. Int J Emerg Manag 6(3):248–260

Kapoor A, Eagle N, Horvitz E (2010) People, quakes, and communications: inferences from call dynamics about a seismic event and its influences on a population. In: Proceedings of AAAI artificial intelligence for development (AI-D'10), Stanford

Li J, Rao H (2010) Twitter as a rapid response news service: an exploration in the context of the 2008 china earthquake. Elec J Inf Syst Dev Ctries 42(0)

Mendoza M, Poblete B, Castillo C (2010) Twitter under crisis: can we trust what we rt? In: Proceedings of the first workshop on social media analytics. ACM, Washington, DC, pp 71–79

Paul M, Dredze M (2011) You are what you tweet: analyzing twitter for public health. In: Proceedings of the 5th international AAAI conference on weblogs and social media (ICWSM), Barcelona

Petrescu-Prahova M, Butts CT (2008) Emergent coordinators in the World Trade Center disaster. Int J Mass Emerg Dis 28(3):133–168

Prier K, Smith M, Giraud-Carrier C, Hanson C (2011) Identifying health-related topics on twitter. In: Proceedings of the 4th international conference social computing, behavioral-cultural modeling and prediction, College Park, pp 18–25

Quincey E, Kostkova P (2010) Early warning and outbreak detection using social networking websites: the potential of twitter, Electronic Healthcare. Springer, Berlin/Heidelberg, pp. 21–24

Rodríguez H, Quarantelli E, Dynes R (2006) Handbook of disaster research. Springer, New York

Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on world wide web. ACM, Raleigh, pp 851–860

Sheetz S, Kavanaugh AL, Quek F, Kim BJ, Lu S-C (2010) The expectation of connectedness and cell phone use in crises. J Emerg Manag 7(2):124–136

Signorini A, Segre A, Polgreen P (2011) The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. PLoS One 6(5):e19467

Zamite J, Silva F, Couto F, Silva M (2011) Medcollector: multisource epidemic data collector. In: Transactions on large-scale data-and knowledge-centered systems IV. Springer, Berlin/Heidelberg, 7(2):124–136

Social Networks in Healthcare: Case Study

Fei Wang
Healthcare Analytics Research Group, IBM T. J. Watson Research Center, Yorktown Heights, NY, USA

Synonyms

Electronic health record; Patient similarity

Patient similarity	The clinical similarity score between pairwise patients derived from their records
Patient network	A network with nodes representing patient entities, edges representing pairwise patient similarities

Definition

Constructing an undirected patient network with patients as nodes and pairwise clinical similarities as edge weights can enable many applications in modern medical informatics such as physician decision support, risk stratification, and comparative effectiveness research, because similar patients have similar clinical characteristics and thus the treatment on one patient might be helpful to his/her similar patients. Therefore, constructing such a patient network is very important to data-driven analytics for healthcare, and effective patient similarity evaluation is the key to construct the patient network.

Introduction

Healthcare has undergone a tremendous growth in the use of electronic health records (EHR) systems to capture patient disease and treatment histories. However, these systems store the data in a manner that makes it difficult for clinicians to extract what is necessary to make clinical decisions at the point-of-care. Most of the EHR systems are primarily used to record clinical events for bookkeeping and claim purposes as opposed to be used as a decision support tool for better diagnosis and treatment. Constructing a patient network with nodes representing patients and edges connecting clinically similar patients might be very helpful to such a clinical decision support system, as the physician can look at the treatments and disease condition evolutions of the similar patients to come up with a better care plan for the current patient.

Actually besides decision support systems, there are also other areas in medical informatics where such patient network could be very helpful, for example, comparative effectiveness research (CER), which is the direct comparison of existing healthcare interventions to determine which work best for which patients and which pose the greatest benefits and harms (http://en.wikipedia.org/wiki/Comparative_effectiveness_research 2013). In such a case, if we can first stratify the patients into different cohorts according to their clinical similarity, then CER can be performed on the patients within the same cohort. Under a similar setting, patient risk stratification aims to stratify the patients according to their disease condition risks. This is a crucial step for effective management of patients, because for patients with different risks, we may have different treatment plans. One step forward, if we can construct an undirected patient network using such patient similarity, we can expect to discover some disease and their evolution patterns, as well as the care/treatment patterns, which would be clinically very useful.

Current Technologies

There have already been quite a few patient similarity evaluation techniques. Before giving an overview of them, we first need to introduce the vector space representation of the patient clinical characteristics, which is an enabling technique to invoke the similarity learning and computations.

Patient Profiling

Patient EHRs contain lots of heterogeneous information, such as demographic information, diagnosis, medication, and lab tests. We call these different information source features. To facilitate the process of similarity learning, some researchers proposed to construct a profile for each patient, which is a feature vector with the dimensionality equal to the number of different features. Before constructing such a vector, we first define a time period of interest, within which we will aggregate the features to get the entries in the patient profile (e.g., the average value of a specific lab test or the count of a specific diagnosis code). In this way, after profiling, each patient is represented as a feature vector (Wang et al. 2011a, b, 2012).

Locally Supervised Metric Learning

Locally Supervised Metric Learning (LSML) is a supervised metric learning approach that has been proved to be useful in patient similarity evaluation (Sun et al. 2010a,b; Ebadollahi et al. 2010). This algorithm was initially proposed in Wang et al. (2009) for measuring text similarity. In the following, we use $\mathbf{X}$ D$[\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{\mathrm{d} \times \mathrm{n}}$ to represent a data matrix from a single specific party, and $\mathbf{y}$ D$[y_1, \ldots, y_n]^T \in \mathbb{R}^n$ is the corresponding label vector with $y_i \in \{1, 2, \ldots, C\}$ denoting the label of $\mathbf{x}_i$, and C is the number of classes. Some examples of the labels here can be diagnosis, for example, the patient has diagnosis or not, or hospitalization, meaning the patient is hospitalized or not, etc.

Our goal is to learn a Mahalanobis distance as follows:

$$d_\Sigma(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \Sigma (\mathbf{x}_i - \mathbf{x}_j)} \qquad (1)$$

where $\sum \in \mathbb{R}^{d \times d}$ is a symmetric positive semidefinite (SPSD) matrix. Following Wang et al. (2009), we define the homogeneous neighborhood and heterogeneous neighborhood around each data point as:

- (Homogeneous neighborhood). The homogeneous neighborhood of $\mathrm{x_i}$, denoted as N_i^o, is the $|N_i^o|$-nearest data points of $\mathrm{x_i}$ with the same label.
- (Heterogeneous neighborhood). The heterogeneous neighborhood of $\mathrm{x_i}$, denoted as N_i^e, is the $|N_i^e|$-nearest data points of $\mathrm{x_i}$ with different labels.

In the above two definitions, we use $|\cdot|$ to denote set cardinality. In order to define the individual distance metric on this party, we need to first construct the neighborhood N_i^o and N_i^e. Then we can define the local compactness and scatterness around point $\mathrm{x_i}$ as

$$C_i = \sum\nolimits_{j:\mathbf{x}_j \in N_i^o} d_\Sigma^2(\mathbf{x}_i, \mathbf{x}_j) \tag{2}$$

$$S_i = \sum\nolimits_{k:\mathbf{x}_k \in N_i^e} d_\Sigma^2(\mathbf{x}_i, \mathbf{x}_k) \tag{3}$$

Then we can learn an optimal distance metric by minimizing the following discrimination criterion

$$J = \sum\nolimits_{i=1}^{n} (C_i - S_i) \tag{4}$$

which makes the data in the same class compact while data in different class diverse. As $\sum$ is SPSD, we can factorize it using incomplete Cholesky decomposition as

$$\Sigma = \mathrm{W}\mathrm{W}^\top \tag{5}$$

Then, J can be expanded as

$$J = tr\left(\mathrm{W}^\top(\Sigma_C - \Sigma_S)\mathrm{W}\right) \tag{6}$$

where tr(·) is the matrix trace, and

$$\Sigma_C = \sum\nolimits_i \sum\nolimits_{j:\mathbf{x}_j \in N_i^o} \left(\mathbf{x}_i - \mathbf{x}_j\right)\left(\mathbf{x}_i - \mathbf{x}_j\right)^\top \tag{7}$$

$$\Sigma_S = \sum\nolimits_i \sum\nolimits_{k:\mathbf{x}_k \in N_i^e} (\mathbf{x}_i - \mathbf{x}_k)(\mathbf{x}_i - \mathbf{x}_k)^\top \tag{8}$$

are the local compactness and scatterness matrices. Hence, the distance metric learning problem can be formulated as

$$\min_{\mathrm{W}:\mathrm{W}^\top \mathrm{W}=\mathrm{I}} tr\left(\mathrm{W}^\top(\Sigma_C - \Sigma_S)\mathrm{W}\right) \tag{9}$$

Note that the orthogonality constraint $\mathbf{W} \top \mathbf{W} = \mathbf{I}$ is imposed to reduce the information redundancy among different dimensions of W, as well as control the scale of W to avoid some arbitrary scaling. The optimal solution of W can be obtained by doing eigenvalue decomposition to $\sum_c - \sum_s$ with the largest eigenvectors.

In summary, the individual distance metric, which is parameterized by a projection matrix W, can finally be learned from local neighborhood information. Next, we will show how to combine these neighborhoods from different base metrics into a single optimal distance metric.

Efficient Metric Updating: Interactive Metric Learning

One issue for applying the above LSML technique in patient similarity evaluation for physician decision support is that the physician may give some feedback after he/she sees the results. Therefore, it is important for LSML to be capable of efficiently incorporating those feedbacks. The feedbacks in general can be regarded as in the form of label changes of y, which consequently leads to changes to $\sum_c$ and $\sum_s$, and the key is to efficiently updating the eigenvalue and eigenvectors of $\sum_c - \sum_s$. In the following, we will briefly describe how the authors in Wang et al. (2011b) solve this problem.

Definition and Setup

To facilitate the discussion, we define the following matrix:

$$\Sigma = \Sigma_C - \Sigma_S \tag{10}$$

Next we introduce an efficient technique based on matrix perturbation (Stewart and Sun 1990) to adjust the learned distance metric according to changes of $\sum$. Suppose that after adjustment L becomes

$$\widetilde{\Sigma} = \Sigma + \Delta\Sigma \tag{11}$$

We define (λ_i, w_i) as one eigenvalue-eigenvector pair of matrix $\sum$. Similarly, we define $\left(\widetilde{6}_i, \tilde{w}_i\right)$ as one eigenvalue-eigenvector pair of $\widetilde{\Sigma}$.

Then we can rewrite $\left(\widetilde{6}_i, \tilde{w}_i\right)$ as

$$\widetilde{6}_i = 6_i + \Delta 6_i \tag{12}$$

$$\tilde{w}_i = w_i + \Delta w_i \tag{13}$$

Next we can obtain

$$\begin{aligned}(\Sigma + \Delta\Sigma)(w_i + \Delta w_i)\\ = (6_i + \Delta 6_i)(w_i + \Delta w_i)\end{aligned} \tag{14}$$

Now the key questions are how to compute changes to the eigenvalue $\Delta\lambda_i$ and eigenvector $\Delta\mathbf{w}_i$, respectively.

Eigenvalue Update

Expanding Eq. 14 and using the fact that $\sum\mathbf{w}_i = \lambda_i\mathbf{w}_i$, we can obtain the following equation:

$$(\Sigma + \Delta\Sigma)w_i = 6_i w_i + \Delta_i w_i \tag{15}$$

Now multiplying both sides of Eq. 15 with $w_i^\top$ and because of the symmetry of $\sum$, we get

$$\Delta 6_i = w_i^\top \Delta\Sigma w_i \tag{16}$$

Eigenvector Update

Since the eigenvectors are orthogonal to each other, we assume that the change of the eigenvector $\Delta\mathbf{w}_i$ is in the subspace spanned by those original eigenvectors, i.e.,

$$\Delta w_i \approx \sum\nolimits_{j=1}^{d} \alpha_{ij} w_j \tag{17}$$

where $\{\alpha_{ij}\}$ are small constants to be determined. Bringing Eq. 17 into Eq. 15, we obtain

$$\Sigma \sum_{j=1}^{d} \alpha_{ij} w_j + \Delta\Sigma w_i = 6_i \sum_{j=1}^{d} \alpha_{ij} w_j + \Delta 6_i w_i$$

Multiplying $w_k^\top (k \neq i)$ on both side of the above equation and discarding the high-order term and bringing in Eq. 17, we get

$$\Delta w_i = -\sum_{j \neq i} \frac{w_j^\top \Delta\Sigma w_i}{6_i - 6_j} w_j \tag{18}$$

Collective Intelligence: Composite Distance Integration

Another challenge in patient similarity is that different physicians have different opinions, then how to integrate all of them to come up with an objective patient similarity? Wang et al. (2011a, 2012) presented an approach on integrating neighborhood information from multiple parties (physicians) when performing LSML. Next, we will briefly review this technique.

Objective Function

The goal here is still learning a Mahalanobis distance as in Eq. 1 but integrating the neighborhood information from all parties. Here, the qth party constructs homogeneous neighborhood $N_i^o(q)$ and heterogeneous neighborhood $N_i^e(q)$ for the ith data point in it. Correspondingly, the compactness matrix Σ_C^q and the scatterness matrix Σ_S^q are computed and shared by the qth party:

$$\begin{aligned}\Sigma_C^q &= \sum\nolimits_{i \in X_q} \sum\nolimits_{j:\mathbf{x}_j \in N_i^o(q)} \left(\mathbf{x}_i - \mathbf{x}_j\right)\left(\mathbf{x}_i - \mathbf{x}_j\right)^\top\\ \Sigma_S^q &= \sum\nolimits_{i \in X_q} \sum\nolimits_{j:\mathbf{x}_k \in N_i^e(q)} \left(\mathbf{x}_i - \mathbf{x}_k\right)\left(\mathbf{x}_i - \mathbf{x}_k\right)^\top\end{aligned}$$

Similar to one party case presented in Eq. 6, we generalize the optimization objective as

$$\begin{aligned}J &= \sum\nolimits_{q=1}^{m} \alpha_q J^q\\ &= \sum\nolimits_{q=1}^{m} \alpha_q tr\left(W^\top\left(\Sigma_C^q - \Sigma_S^q\right)W\right)\end{aligned} \tag{19}$$

where α_q is the importance for the qth party and $\alpha = .\alpha_1, \alpha_2, \ldots, \alpha_m{}^\top$ is constrained to be in a simplex as $\alpha_q \geq 0$, $\sum^q \alpha^q$D 1, and m is the number of parties. Note that by minimizing Eq. 19, the proposed approach actually leverages the local neighborhoods of all parties to get a more powerful discriminative distance metric. Thus, it aims at solving the following optimization problem:

$$\begin{aligned}&\min_{\alpha,\mathrm{W}} \sum_{q=1}^{m} \alpha_q tr\left(\mathrm{W}^\top \left(\Sigma_{\mathcal{C}}^{q} - \Sigma_{\mathcal{S}}^{q}\right)\mathrm{W}\right)\\ &\quad + 6\Omega(\alpha)\\ &\text{s.t.}\quad \alpha \geq 0, \alpha^\top \mathbf{e} = 1\\ &\quad \mathrm{W}^\top \mathrm{W} = \mathrm{I}\end{aligned} \tag{20}$$

Here $\Omega(\alpha)$ is some regularization term used to avoid trivial solutions, and $\lambda \geq 0$ is the trade-off parameter. In particular, when $\lambda = 0$, i.e., without any regularization, only $\alpha_q = 1$ for the best party, while all the others have zero weight. The best λ can be selected through cross-validation.

Problem (20) can be solved by alternating optimization and the procedure is guaranteed to converge to a local optimum.

Future Trends

Although LSML is a powerful methodology and it has been proved to be useful on some real-world clinical data (Ebadollahi et al. 2010; Sun et al. 2010b; Wang et al. 2012), there are still some limitations which include: (1) It is a supervised approach, meaning, for all the training data, we need to have their supervision information (either in terms of labels or pairwise constraints) – this is difficult in medical scenario as the supervision information is expensive and time-consuming to obtain – and (2) it needs to construct different types of neighborhoods; this could be time-consuming when the data set scale is large. Therefore, the future research toward effective patient similarity evaluation should be the following: (1) Use less supervisions and more unsupervised data. Semisupervised learning techniques (Zhu and Goldberg 2009) could be helpful in this scenario. (2) Improve the scalability of the algorithm and make it fit in the scenario when we have millions of patients.

Conclusion

This chapter reviews the state-of-the-art technology for patient similarity evaluation, which can be used for constructing a patient network. Specifically, we introduced the Locally Supervised Metric Learning (LSML) algorithm as well as its two variants on how to make real-time updates and integrate multiple experts' opinions. We finally point out that the future research directions of this research topic.

Cross-References

- ▶ Data Mining
- ▶ Disease Surveillance: Case Study
- ▶ Distance and Similarity Measures
- ▶ Online Healthcare Management

References

Ebadollahi S, Sun J, Gotz D, Hu J, Sow D, Neti C (2010) Predicting patient's trajectory of physiological data using temporal trends in similar patients: a system for near-term prognostics. In: AMIA annual symposium proceedings, pp 192–196

Stewart G, Sun JG (1990) Matrix perturbation theory. Academic, Boston

Sun J, Sow DM, Hu J, Ebadollahi S (2010a) Localized supervised metric learning on temporal physiological data. In: International conference on pattern recognition (ICPR), pp 4149–4152

Sun J, Sow DM, Hu J, Ebadollahi S (2010b) A system for mining temporal physiological data streams for advanced prognostic decision support. In: IEEE international conference on data mining, pp 1061–1066

Wang F, Sun J, Li T, Anerousis N (2009) Two heads better than one: metric+active learning and its applications for it service classification. In: IEEE international conference on data mining, pp 1022–1027

Wang F, Sun J, Ebadollahi S (2011a) Integrating distance metrics learned from multiple experts and its application in inter-patient similarity assessment. In: SIAM data mining conference, pp 59–70

Wang F, Sun J, Hu J, Ebadollahi S (2011b) IMET: interactive metric learning in healthcare applications. In: SIAM data mining conference, pp 944–955

Wang F, Sun J, Ebadollahi S (2012) Composite distance metric integration by leveraging multiple experts' inputs and its application in patient similarity assessment. Stat Anal Data Min 5(1):54–69

Zhu X, Goldberg AB (2009) Introduction to semi-supervised learning. Synthesis lectures on artificial intelligence and machine learning. Morgan and Claypool, San Rafael

Social Networks Members

► Social Influence Analysis

Social Networks Users

► Social Influence Analysis

Social Order in Online Social Networks

Tina Eliassi-Rad
College of Computer and Information Science,
Northeastern University, Boston, MA, USA

Glossary

Online Social Network	A social network on the World Wide Web
Social Network	A set of individuals connected by a set of dyadic ties
Tie	A relationship between two individuals

Definition

Social order, a technical term from social sciences (Frank 1944), is the study of how social creatures (such as human beings) are both individual and social (Hechter and Horne 2003). As Hechter and Horne (2003) point out, social order occurs when individuals coordinate and cooperate with each other.

Social order in online social networks and the coordination and cooperation that give rise to them appear in many different structural forms. Examples include *homophily, communities* (a.k.a. groups), *weak ties, structural holes*, and *social capital*.

Homophily The notion of homophily (i.e., "of like attracting like") has been around since the ancient Greeks. It is often quoted that Plato said, "Similarity begets friendship." Previous research (McPherson et al. 2001) has shown that homophily is a major criterion governing the formation of ties in social networks. Many social networks have high levels of homophily (Easley and Kleinberg 2010, pp. 79–81). Coordination and cooperation is often more successful between people who are similar to each other – either in terms of status or value (McPherson et al. 2001).

Communities Generally speaking, communities are defined as groups of individuals that are well connected to each other. The existing literature contains many objective functions and algorithms that formalize the aforementioned definition and produce communities (Leskovec et al. 2010). The one pertinent to social order is where a community has low conductance, i.e., where the ratio of ties crossing the community boundary to ties within the community is low (Leskovec et al. 2010). Members of such communities are often tightly connected. These highly connected structures, in turn, promote trust among their members – an important property for social order.

Leskovec et al. (2008) found that the sizes of communities in large online networks roughly follow the Dunbar number (~150) (Dunbar 1998) and that large well-defined communities are absent in online networks. These findings make intuitive sense since maintaining relationships besides the trivial ones requires substantial investment in terms of our neocortex processing capabilities (Dunbar 1998).

Moreover, Leskovec et al. (2008) describe large social networks as having a nested core-periphery structure, where the network is composed of layers of large cores and a small number of dense communities loosely connected to the core. This result indicates the presence of a hierarchy or nested social order in online social networks. In other words, the levels of coordination and cooperation vary depending on where in the nested core-periphery structure a person resides.

Weak Ties Granovetter (2003) was the first to distinguish between weak and strong ties in social networks. He informally defined *tie strength* as the "amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie" (Granovetter 2003, p. 1361). Weak ties correspond to "local bridges" (Easley and Kleinberg 2010), where two people have zero common friends. The lack of common friends can make coordination and cooperation difficult and reduce social order.

Structural Holes Burt (2004) defined structural holes as the empty spaces (i.e., no connections) between groups in the social network. People who fill these structural holes bring social order to the network because they control the information flow and are rewarded with power and wealth.

Social Capital Being members of a community has many advantages (Portes 1998). For example, belonging to a community with high *triadic closure* (where friend of a friend is a friend) and *embeddedness* (where two people share many of the same friends) enforces norms and maintains reputational effects. In other words, this "closure" of friends promotes trust. The counterbalance to closure is *brokerage*. People who are "brokers" interact at the boundary of various communities, i.e., they fill the structural holes. As mentioned above, such people have more social capital compared to others in the community.

Social order, in terms of closures and brokerages, is essential in the preservation of social networks. Closures give rise to communities, while brokerages give rise to connections across various communities.

Cross-References

- ▶ Centrality Measures
- ▶ Community Detection: Current and Future Research Trends
- ▶ Human Behavior and Social Networks
- ▶ Role Discovery
- ▶ Role Identification of Social Networkers
- ▶ Social Capital
- ▶ Structural Holes
- ▶ Trust in Social Networks

References

Burt RS (2004) Structural holes and good ideas. Am J Sociol 110(2):349–399

Dunbar R (1998) Grooming, Gossip, and the evolution of language. Harvard University Press, Cambridge, MA

Easley D, Kleinberg J (2010) Networks, crowds, and markets. Cambridge University Press, New York

Frank LK (1944) What is social order? Am J Sociol 49 (5):470–477

Granovetter M (2003) The strength of ties. In: Hechter M, Horne C (eds) Theories of social order: a reader. Stanford University Press, Stanford, pp 323–332

Hechter M, Horne C (2003) Theories of social order: a reader. Stanford University Press, Stanford

Leskovec J, Lang K, Dasgupta A, Mahoney M (2008) Statistical properties of community structure in large social and information networks. In: The 17th international conference on World Wide Web, Beijing, pp 695–704

Leskovec J, Lang KJ, Mahoney MW (2010) Empirical comparison of algorithms for network community detection. In: The 19th international conference on World Wide Web, Raleigh, pp 631–640

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Ann Rev Sociol 27:415–444

Portes A (1998) Social capital: its origins and applications in modern sociology. Ann Rev Sociol 24:1–24

Recommended Reading

Blau P, Schwartz J (1997) Crosscutting social circles: testing a macro-structural theory of intergroup relations. Transaction Publishers, Piscataway

Burt RS (2005) Brokerage and closure: an introduction to social capital. Oxford University Press, Oxford, UK

Gellner E (2003) Trust, cohesion, and the social order. In: Hechter M, Horne C (eds) Theories of social order: a reader. Stanford University Press, Stanford, pp 300–305

Henderon K, Eliassi-Rad T, Papadimitriou S, Faloutsos C (2010) HCDF: a hybrid community discovery framework. In: The 10th SIAM international conference on data mining, Columbus, pp 754–765

Lazarsfeld P, Merton RK (1954) Friendship as a social process: a substantive and methodological analysis. In: Berger M, Abel T, Page CH (eds) Freedom and control in modern society. Van Nostrand, Toronto, pp 18–66

Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci USA 103: 8577–8582

Simmel G (2003) The web of group-affiliations. In: Hechter M, Horne C (eds) Theories of social order: a reader. Stanford University Press, Stanford, pp 316–322

Watts D (2004) Six degrees: the science of a connected age. W.W. Norton & Company, New York

Social Phishing

Yada Zhu[1] and Jingrui He[2]
[1]IBM Research, Yorktown Heights, NY, USA
[2]Computer Science and Engineering, Arizona State University, Tempe, AZ, USA

Synonyms

Email; Fraud; Information; Internet; Social network; Suspicious

Glossary

Anti-phishing	Efforts taken from multiple perspectives to combat phishing crimes
Email spam	Unsolicited emails for the purpose of advertisement or committing fraud
Machine learning	The design and development of algorithms that takes as input empirical data, and outputs patterns and predictions for future data
Phisher	Fraudsters who commit phishing crimes
Phishing	Electronic fraud based on social engineering
Phishing site	Web sites created by phishers to steal sensitive information from users

Definition

Nowadays, phishing has gradually become a popular type of electronic fraud that makes use of social engineering to steal sensitive information from users such as user name, password, bank account number, credit card details, etc. (https://kb.iu.edu/d/arsf, http://en.wikipedia.org/wiki/Phishing#Early_phishing_on_AOL). Phishing can be carried out via emails, instant messages, phone calls, text messages, etc. (https://kb.iu.edu/d/arsf, http://en.wikipedia.org/wiki/Phishing#Early_phishing_on_AOL), where phishers pretend to be a trustworthy party in an attempt to lead the users to disclose the above sensitive information. Based on the collected information, the phishers can withdraw money from the accounts, causing significant financial loss.

To combat phishing crimes, people are making efforts from various aspects. For example, there have been constant efforts toward raising public awareness of this rapidly proliferating cyber crime, so that users are not easily spoofed into giving up sensitive information; researchers from academia and industry have been tracking the recent developments of phishing techniques with the hope of catching them in time; there has also been efforts from the government by filing law suits against phishers and proposing laws to fight this crime.

The purpose of this entry is twofold. The first is to introduce the evolution process of the phishing techniques, with an emphasis on its current status; and the second is to look into the techniques for anti-phishing, which shed lights on the future generation of phishing methods.

Introduction

According to http://en.wikipedia.org/wiki/Phishing#Early_phishing_on_AOL, the word "phishing" came into use as a variant of "fishing" in mid-1990s, which is connected to "baits" used therein to induce the users into disclosing sensitive information. Also, the "ph" spelling was used to link phishing scams with some underlying communities, such as the hackers known as "phreaks" (http://www.phishing.org). A typical example of phishing is a fake email masqueraded to come from a bank, which asks the user to follow an embedded link to a phishing site (which often highly mimics the authentic website) and give up his/her bank account information. Another example of phishing takes place in an online chat session, where the phisher pretends to be an agent from the online vendor and requires the sensitive information from the user. In both cases, the baits are the masqueraded identity of the email sender, the embedded link, and the online agent. If the user is tricked into believing this identity and reveals his/her account information, he/she will suffer significant financial loss.

Due to the severe challenge posed by phishing, recent years have seen rapidly growing efforts in anti-phishing. To be specific, in academic, many universities have set up groups devoted to anti-phishing research, such as the Anti-Fishing Group at Indiana University (https://www.sice.indiana.edu/graduate/degrees/informatics/security/research.html), the Anti-Phishing Group at City University of Hong Kong (https://www.cs.cityu.edu.hk/research/research_areas/Information_Security.html), the Center for Education and Research in Information Assurance and Security at Purdue University (http://www.cerias.purdue.edu), the Stanford Security Laboratory (http://theory.stanford.edu/seclab/), the Cylab Usable Privacy and Security Laboratory (http://cups.cs.cmu.edu), to name a few. In industry, a variety of anti-phishing solutions have been proposed, such as the suite of solutions provided by the Anti-Phishing Working Group, the phishing protection services from Dell SecureWorks and RSA, anti-phishing toolbars from eBay, Netcraft, and EarthLink, as well as the anti-phishing filters in Firefox, Internet Explorer, Google Chrome, etc.

Key Points

In the rest of this entry, we first review the history and the current status of phishing, followed by a discussion of anti-phishing techniques.

Historical Background

According to http://www.phishing.org, the very first phishing attacks happened on American Online (AOL) on January 2, 1996. At that time, phishers sent messages to users through AOL instant messengers and email systems, requesting the users to verify their accounts or to confirm their billing information. Many users gave up the account information upon such requests and later experienced financial loss. In response to such phishing crimes, AOL and later many banks and online payment systems include warnings in their emails and instant messenger chat windows preventing the users to disclose the sensitive information in such scenarios.

The phishing crimes quickly ramped up since late 2003, with the registration of domains suggesting legitimate sites such as eBay and PayPal, which were used as phishing sites. Phishers then sent out emails to the users, leading them to these phishing sites and asking them to update their credit card information. Later the phishing techniques evolved into using popup windows of online banks to gather account information from the users, which was proven to be very effective in 2004.

According to the surveys by Gartner between 2005 and 2007, there was continuous increase in the percentage of phished web users in the USA (Herley and Florencio 2008): 0.5% in 2005, 1.05% in 2006, and 2.18% in 2007, resulting in huge financial loss of these victims. Similar as in the USA, the losses in the United Kingdom from web banking fraud, most of which are phishing fraud, almost doubled from 2004 to 2005 (http://en.wikipedia.org/wiki/Phishing#Early_phishing_on_AOL).

The most up-to-date phishing techniques have been summarized in a variety of websites, such as http://en.wikipedia.org/wiki/Phishing#Early_phishing_on_AOL, http://www.phishing.org, and https://www.cs.cityu.edu.hk/research/research_areas/Information_Security.html. These include email spamming, web based delivery (a.k.a., man-in-the-middle), instant messaging, Trojan hosts, link manipulation, key loggers, session hacking, system reconfiguration, content injection, phishing through search engines, phone phishing, malware phishing, etc. (http://www.phishing.org). In addition, according to http://en.wikipedia.org/wiki/Phishing#Early_phishing_on_AOL and http://www.csionsite.com/2012/phishing/, phishing with specific targets are sometimes referred to as spear phishing and whaling. Furthermore, some smart phishers make use of advanced techniques to get around anti-phishing software (e.g., by making use of images instead of text (http://en.wikipedia.org/wiki/Phishing#Early_phishing_on_AOL) or to gain trust of the potential victims (e.g., by including personal information obtained from social networks (Jagatic et al. 2007)). All the above evidence highlights the urgency of effective anti-phishing techniques.

Key Techniques

There are several different ways to combat phishing, including end user education, legislation, and technology developed specially to fight against phishing. This section discusses the use of technology to protect against phishing website and email.

Many technical solutions have been used to identify a web page as a phishing site, including blacklists (fraudulent sites), heuristics, page analysis, ratings, and their combinations. Blacklisting is a widely used approach in phishing detection mechanisms which maintains a list of known phishing websites and check websites against the list. This method has been implemented in numerous browser-integrated anti-phishing tools, such as Internet Explorer (IE), Google Safe Browsing (Schneider et al. 2007), NetCraft toolbar (NetCraft 2007), Firefox, and eBay tool bar (eBay 2007). The IE browser queries lists of blacklisted and whitelisted domains from Microsoft servers and makes sure that the user is not accessing any phishing sites. The Google Safe Browsing uses blacklists of phishing URLs to identify phishing sites. The users are warned before they attempt to navigate to a known phishing site. Blacklists can be created using a set of classification rules based on previous phishing patterns, manually classified by the user or crowd sourced by users of a given service (Wilson and Argles 2011).

The effectiveness of a blacklist is determined by the coverage and quality of the list, and the time it takes to include a phishing site. The quality shows the number of safe sites is falsely included into the list. Study shows that the URLs that have been verified by users tend to be classified with lower false positive rate (Sun et al. 2010). Timeliness may be a challenge for blacklisting because the average lifetime of phishing sites is only a few days or maybe a few hours for the low cost of creating a phishing site. Ludl et al. use 10,000 phishing URLs to test the effectiveness of the blacklists maintained by Google and Microsoft (Ludl et al. 2007). They demonstrate that blacklists provided by Google can recognize almost 90% of live phishing sites, while IE contained only 67% of them. They also find that on average it takes Microsoft 6.4 h to add an initially not blacklisted entry with a standard deviation of 6.2 h. For Google, it takes somewhat longer, 9.3 h on average with a standard deviation of 7.2 h. Sheng et al. (2009) use 191 fresh phish that are less than 30 min old to conduct two tests on eight blacklists based anti-phishing toolbars. By hour 2, 63% of phishing campaigns in their dataset are finished, but only 7.9% of those phish are taken down. On average, 33% of the websites are taken down within 12 h, around half are taken down after 24 h, and 27.7% are still alive after 48 h. They conclude that blacklists are not effective when protecting users initially, as most of the tools catch less than 20% of phish at hour zero. In addition, they show that blacklists are updated at different speeds and vary in coverage, as 47–83% of phish appear on blacklists 12 h from the initial test. They also demonstrate that two tools use heuristics to complement their blacklists trigger catch significantly more phish initially than those using only blacklists. However, it takes a long time for phish detected by heuristics to appear on blacklists. Ramachandran et al. measure the effectiveness of eight spam blacklists in real time by analyzing a 17-month trace of over 10 million spam messages collected at an Internet "spam sinkhole," and by correlating this data with the results of IP-based blacklist lookups (Ramachandran and Feamster 2006). In their study, whenever a host spammed their domain, they examine whether that host IP is listed in a set of Domain Name Service-based Blackhole Lists (DNSBLs) in real time. Their study indicates that about 80% of the received spams are listed in at least one of eight blacklists, but even the most comprehensive blacklist has a false negative rate of about 50%.

Heuristic techniques analyze whether a page possesses suspicious behavior, e.g., examining the characters of the URLs and site's hostname.

Since a phishing site is usually a mimicry of a legitimate site, page analysis or content-based method detects phishing by examining their similarity in terms of page properties, such as the number of password fields, the number of links, or the organization's logo. Using a search with the extracted keywords, it retrieves candidates for the legitimate site. If the page on the user's browser and the one of the candidate sites have the same domain name, the target site is judged legitimate; otherwise, a phishing site. Rating methods determine phish sites based on user ratings. Each site's rating is computed by aggregating all rates given for that site, with each user's rating of a site weighted according to that user's record of correctly identifying phishing sites. Heuristic, content-analysis and rating are employed by numerous anti-phishing products, for example, Spoof Guard is based on heuristic and ratings; Calling ID toolbar is based on heuristic; Cloudmark Anti-Fraud toolbar is based on ratings; and EarthLink toolbar is based on the combination of heuristic and user rating.

Heuristics can detect attacks as soon as they are launched, without the need to wait for blacklists to be updated. However, attackers can design their attacks to avoid heuristic detection. In addition, heuristic approach may produce false positives, incorrectly labeling a legitimate site as phishing. On the other hand, page analysis techniques also have high false positive rates due to the similarity between the phishing pages and the legitimate ones (Wilson and Argles 2011). User ratings might become meaningless if URLs of legitimate sites are too complex to be known or recognized by users. In response to this challenge, Ludl et al. (2007) analyze a large number of phishing pages and explore the page properties that can be used to identify phishing pages. These features from the HTML source of a page include: the number of forms, input fields (e.g., the number of input fields, text fields, password fields and hidden fields), links (e.g., the number of internals links to internal links to resources located in the page's domain as well as external links to resources stored on other sites), whitelist references, and script tags. Zhang et al. (2011) introduce a content-analysis based large scale anti-phishing gateway. When the HTTP(s) traffic is intercepted by the gateway, the system fetches and filters the target URL. If the URL is not prefiltered by the black and white hash repository, the system fetches the web page content and extracts features. They build a phishing page template database as a repository. After feature extraction, the system calculates the similarity scores between the evaluated web page and each template in the database. They evaluate the performance of the detection system based on 118,165 positive URLs and 92,970 negative URLs. The maximum false positive rate is below 0.1%, and the average false positive rate and false negative rate is 0.05% and 1.78%, respectively. The system demonstrates better performance than several other approaches. Whittaker et al. (2010) present a logistic regression classifier based on features that describe the composition of the web page's URL, the hosting of the page, and the page's HMTL content as collected by a crawler. The evaluation of the classifier is based on two data sets. The first one contains 446,152,060 URLs and the second contains 74,816,740 URLs. The phishing pages make up 1.1% of each data set. The study shows that the classifier can maintain a false positive rate well below 0.1%.

Due to inevitable false positives, directly blocking users' connections to suspected phishing sites is unacceptable. Therefore, phishing site warning mechanisms become mandatory in popular browsers including Firefox and IE. If a web page is correctly identified as a phishing site, a user is directed to a warning page and not allowed to proceed without interacting with the warning page. If the user chooses to ignore the link, the warning page disappears and the user is exposed to the risk of phishing. Otherwise, the user is directed to a default page. A hybrid solution Anti-Phish (Kirda and Kruegel 2006) integrates phishing warning and page analysis for phishing identification. It keeps track of where sensitive information is being submitted. If it detects that confidential information such as a password is being entered into a form on a suspicious web site, a warning is generated and the pending operation is canceled.

However, users tend to ignore the warnings or have learned to bypass the warnings. Wu et al. (2006) conduct a study of three simulation anti-phishing toolbars to determine how effective they are at preventing users from visiting web sites that the tools have determined to be fraudulent. They find that many participants do not notice warning signals or assume the warnings are invalid. In a follow-up study, the authors test anti-phishing toolbars that produce pop-up warnings and block access to fraudulent websites until overridden by the user. These pop-up warnings reduce the rate at which users fall for fraudulent sites, but do not completely prevent all users from failing for these sites. Egelman et al. (2008) compare the effectiveness of active and passive phishing warning. They designed two phishing websites to mimic the login pages of Amazon and eBay, the most commonly phished nonbank websites. They divide the 60 participants into four groups: Firefox warning, active IE warning, passive IE warning, and no warning at all. The results show that over 45% and 90% of participants ignore the strong warning or the passive warning, respectively. Similarly, Schneider et al. (2007) demonstrate that over 50% participants of a warning usability test ignore the warning and enter their credentials, despite the strong wording of the warning page.

Traditional phishing begins with e-mail spam. SMTP (Simple Mail Transfer Protocol) (Jonathan 1982) is the protocol to deliver e-mails in the Internet. It is a simple protocol which lacks necessary authentication mechanisms. Information related to sender, such as the name and email address of the sender, etc., can be counterfeited in SMTP. Therefore, attackers can send out spoofed emails that are seems from a friend, relative, or a reputable business where victims might have an account. A number of solutions have been proposed to solve the anti-phishing problem at the e-mail level. Since the phishing email usually contains some socially engineered message asking users to submit information or to visit the phishing website, filters and content analysis are used to prevent phishing e-mails from reaching their addresses' inbox. For example, MailScanner (Field 2017) is an anti-spam package for e-mail gateway systems in attempts to combat email fraud by examining email contents. ClamAV (ClamAV 2016. http://www.clamav.net) is another toolkit for e-mail scanning making use of blacklisting and phishing signature, such as the use of a specific phrase or looking for the PayPal (https://www.paypal.com/home) logo that many phishing e-mails contain.

The effectiveness of such techniques relies on critical factors, such as natural language processing, filter training using machine learning approaches, and the availability of anti-phishing tools in the e-mail system. Chandrasekaran et al. (2006) use the distinct structural features present in e-mail to classify phishing emails. A total of 25 features consisting of a mixture of style marker (e.g., account, risk, bank, risk, and vocabulary richness) and structural attributes (e.g., the structure of the greeting in the body and the structure of the subject line of the email) are considered. Features are ranked based on their relevance to email classification. Total 400 emails out of which 200 are phishing emails are used in training and evaluating the model. The results demonstrated a detection rate of 95%. Similarly, based on structural features of the phishing emails, Abu-Nimeh et al. (2008) investigate phishing detection in a mobile environment utilizing modified Bayesian Additive Regression Trees (BART). The algorithm modification intends to reduce the computation time and memory overhead of MCMC simulations. About 6561 raw emails are used in building the dataset, from which 1409 emails are phishing. The legitimate emails are collected from financial institutions such as Bank of America, eBay, Chase, etc., and regular communication emails. The dataset constitutes of 60 style marker features and 10 structural attribute features, respectively. The results show a detection rate of 97% and a false positive rate of 3%. However, no matter how effective, some phishing emails can still successfully get through the filters and reach potential victims.

Future Directions

The battle between phishing and anti-phishing is far from over. With the advancement of anti-phishing techniques, phishers constantly come up with new ways of stealing sensitive information from users, by pretending to come from their close friends, by including fake US Airways itineraries, by quietly changing the content in one of the browser tabs, etc. Therefore, it is necessary to raise the awareness of phishing crimes among the general public, to keep the anti-phishing tools up-to-date regarding the newly developed crime patterns, and to even predict the emergence of novel phishing patterns.

Conclusion

In this entry, we focus on social phishing, which is a common social engineering technique for conducting fraud. Ever since its first appearance in the mid-1990s, it has evolved into a variety of sophisticated forms. In the future, to effectively combat phishing, coordinated efforts have to be made from multiple aspects, e.g., education, legislation, and improved anti-phishing techniques.

Acknowledgments

Research was sponsored by the US Defense Advanced Research Projects Agency (DARPA) under the Social Media in Strategic Communication (SMISC) program, Agreement Number W911NF-12-C0028. The views and conclusions contained in this document are those of the author (s) and should not be interpreted as representing the official policies, either expressed or implied, of the US Defense Advanced Research Projects Agency or the US government. The US government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation hereon.

Cross-References

- ► Spam Detection on Social Networks
- ► Spam Detection: E-mail/Social Network

References

Abu-nimeh S, Nappa D, Wang X, Nair S (2008) A distributed architecture for phishing detection using Bayesian additive regression trees. In: eCrime Researchers Summit. Atalanta, Georgia

Chandrasekaran M, Narayanan K, Upadhyaya S (2006) Phishing email detection based on structural properties. In: Proceedings of the NYS cyber security conference. Albany, NY

eBay (2007) eBay tool bar. http://anywhere.ebay.com/browser/firefox/

Egelman S, Cranor LF, Hong J (2008) You've been warned: an empirical study of the effectiveness of web browser phishing warnings, CHI, Florence

Field J (2017) MailScanner. http://www.mailscanner.info

Herley C, Florencio D (2008) A profitless endeavor: phishing as tragedy of the commons. In NSPW'08: Proceedings of the 2008 workshop on new security paradigms

ClamAV (2016) http://www.clamav.net

Jagatic T, Johnson N, Jakobsson M, Menczer F (2007) Social phishing. Commun ACM 50(10):94–100

Jonathan BP (1982) Simple mail transfer protocol. RFC821. https://tools.ietf.org/html/rfc821

Kirda E, Kruegel C (2006) Protecting users against phishing attacks. Comput J. https://doi.org/10.1093/comjnl/bxh169

Ludl C, McAllister S, Kirda E, Kruegel C (2007) On the effectiveness of techniques to detect phishing sites. In: DIMVA '07: Proceedings of the 4th international conference on detection of intrusions and Malware, and vulnerability assessment. Springer, Berlin/Heidelberg, pp 20–39

NetCraft (2007) Netcraft anti-phishing tool bar. http://toolbar.netcraft.com/

Ramachandran A, Feamster N (2006) Understanding the network-level behavior of spammers. In: SIGCOMM '06: proceedings of the 2006 conference on applications, technologies, architectures, and protocols for computer communications. ACM, New York, pp 291–302

Schneider F, Provos N, Moll R, Chew M, Rakowski B (2007) Phishing protection design documentation. https://wiki.mozilla.org/Phishing_Protection:_Design_Documentation

Sheng S, Wardman B, Warner G, Cranor LF, Hong J, Zhang C (2009) An empirical analysis of phishing blacklists. In: CEAS 2009: sixth conference on email and anti-spam, July 2009

Sun B, Wen Q, Liang X (2010) A DNS based anti-phishing approach. In: Second international conference on networks security. Wireless Communications and Trusted Computing, Beijing

Whittaker C, Ryner B, Nazif M (2010) Large-scale automatic classification of phishing pages. In: NDSS'10. San Diego, California

Wilson C, Argles D (2011) The fight against phishing: technology, the end user and legislation. In: The international conference on information society (i-Society), London

Wu M, Miller RC, Garfinkel SL (2006) Do security toolbars actually prevent phishing attacks? In: Proceedings of the SIGCHI conference on human factors in computing systems, Montreal

Zhang J, Wu C Guan H, Wang Q, Zhang L, Ou Y, Xin Y, Chen L (2011) An content-analysis based large scale anti-phishing gateway. In: 12th IEEE international conference on communication technology, Nanjing

Social Provenance

Zhuo Feng[1], Pritam Gundecha[2] and Huan Liu[3]
[1]AI+R, Microsoft, Sunnyvale, CA, USA
[2]IBM Research, Almaden, San Jose, CA, USA
[3]Data Mining and Machine Learning Lab, School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA

Synonyms

Data mining; Disinformation; Information provenance; Misinformation; Provenance paths; Social computing; Social media; Social network

Glossary

Information Provenance	Sources of a piece of information
Social Computing	An area of computer science that is concerned with the intersection of social behavior and computational systems (Social Computing)
Social Media	A group of Internet-based applications that build on the ideological and technological foundations of Web 2.0 and that allow the creation and exchanges of user-generated content (Kaplan and Haenlein 2010)
Data Mining	The computational process of discovering patterns in large datasets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems (Data Mining)
Social Network	A social network is a social structure made up of a set of social actors (such as individuals or organizations), sets of dyadic ties, and other social interactions between actors (Social Network)
Misinformation	Misinformation is false or incorrect information that is spread intentionally or unintentionally (without realizing it is untrue) (Misinformation)
Disinformation	Disinformation is intentionally false or misleading information that is spread in a calculated way to deceive target audiences (Disinformation)
Provenance Paths	Paths of information propagation from sources to terminals

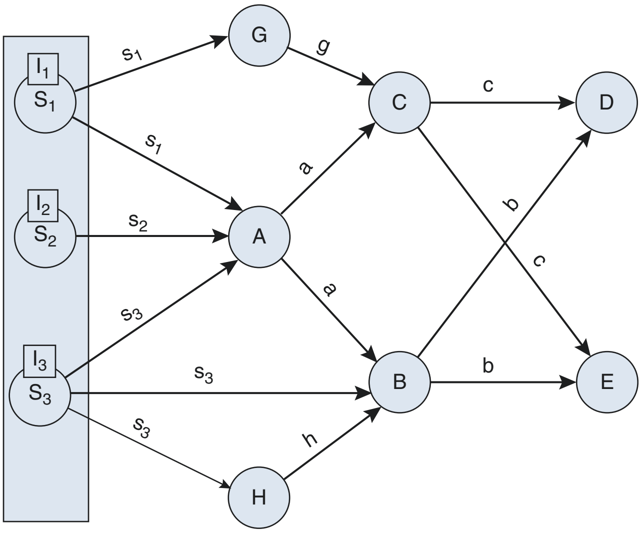
Definition

Social Provenance

An information propagation network can be represented as a directed graph G (V, E), where V is the node set and E is the edge set. Each node in the graph represents an entity, which publishes a piece of information on social media. The entity may refer to an individual user or a webpage. A directed edge between nodes represents the direction of information flow. For a given piece of information propagating through the social media, the *social provenance* informs a user about the *sources* of a given piece of information. Sources refer to the nodes that first publish the concerned messages.

Figure 1 shows an information propagation graph indicating the flow of information $I = \{I_1; I_2; I_3\}$ which is about the same event. S_1, S_2, and S_3 are the source nodes, or the

Social Provenance, Fig. 1 Information propagation in social media



originators of I_1, I_2 and I_3, respectively. The information is transmitted through different nodes in social media or by their recipients. These nodes propagate information; some may retransmit it with modifications. Each edge is labeled with the information indicating where it comes from, e.g., “a” on edge “A-C” means that it is from “A.” A social provenance problem is to help a recipient (say, node D) to answer what are possible information sources in social media for a given piece of information. A provenance path delineates how information spreads from a source to a recipient, including those responsible for retransmitting the information from the sources through intermediaries. If the provenance paths are known, the sources of information can be determined. More often than not, however, provenance paths of a known piece of information are unknown.

Provenance has been studied in the data management field. In data management, provenance represents the creator of the data and how data has been modified and transferred. Provenance information is used to determine the authenticity and trustworthiness of information. Provenance is the key to solve the data conflict problem (Moreau 2009). Unlike social media, data propagation can be captured in the data management systems. Social provenance has been introduced in the book (Barbier et al. 2013) and received some attention in recent years (Gundecha et al. 2013a; Ranganath et al. 2013; Feng et al. 2013; Gundecha et al. 2013b; Wu et al. 2016). Shah and Zaman (2011) proposed a centrality-based method to determine the single information source among all known recipients on an undirected network. It assumes that information spread on a network follows the susceptible infected (SI) model. Since this method requires the knowledge of all recipients, it is not practical for social provenance. Also, the source computed using this method is more biased toward higher-degree nodes.

Barbier, in his dissertation (Barbier 2012), proposed a method to collect metadata about the received information. Such metadata is referred as *provenance attributes*. Provenance attributes can play a vital role in obtaining social provenance. As shown in the dissertation (Barbier 2012), some attribute values are easier to obtain than others and some attribute values may be more valuable to a recipient than others. For example, a political statement published by a political candidate might be assessed with some bias if the recipient knows information about the political candidate, such as political party affiliation or special interest associations. An interesting example of the value of provenance attribute would be to reveal the political affiliation and special interests of an unfamiliar social media user propagating political statements, which may help understand latent motivations for propagating a statement in social media. Fake news and its propagation on social media sites have widely been reported during recent US election (http://www.nytimes.com/2016/11/18/technology/fake-news-on-facebook-in-foreign-elections-thats-not-new.html?_r=0, http://www.vox.com/new-money/2016/11/16/13659840/facebook-fake-news-chart). Provenance attributes, including political affiliations and special interest group associations, education, occupation, and demographic attributes, of the nodes involved in the propagation of news articles would have helped recipients to decide quickly the fake news from the real ones. Barbier et. al. (2013) reviews the current research on social provenance and explores exciting research opportunities to address pressing needs. Papers (Gundecha et al. 2013a, b; Ranganath et al. 2013; Feng et al. 2013) show how data mining can enable a social media user to make informed judgments about statements published in social media. The chapter (Wu et al. 2016) proposes few benchmark datasets and evaluation metrics to study the social information problem further.

Key Research Issues in Social Provenance

Social media can help in solving the problem of social provenance due to its unique features: user-generated content (e.g., tweets, blog posts, news articles, etc.), users' profiles, user interactions (e.g., links between friends, hyperlinks on the blog or news articles), and spatial or temporal information. These features can help reconstruct an information propagation network of a given message, and the network is essential for social provenance.

The *social provenance problem* answers which nodes are the possible sources of some particular information, say a text message. The *provenance path problem* seeks to identify the paths that allow us to trace back possible sources. Solving the social provenance problem entails solving the provenance path problem. We present some key research issues in this burgeoning area below:

(a) What are the characteristics of sources such that we can identify a source when we encounter one? It is a challenging task because source nodes are not necessarily those without incoming links in social media networks.
(b) How can we use different parts of social media data for inferring provenance paths? Content, user profiles, and interaction patterns can play complementary roles in backtracking information propagation. As a popular source can lead to a shallow cascade (Leskovec et al. 2009), the study of node centrality measures can be of help.
(c) How can we infer missing links in reconstructing a provenance path with partial information? By the nature of social media, most information is informal and partial. Links can expand the network (i.e., new nodes can be added), and data associated with a node provides more information, though still partial.
(d) How can we limit the search space in the vast land of social media? It is incumbent to develop a scalable solution for the social provenance problem.
(e) What are effective and objective ways of verifying and comparing different approaches to social provenance and provenance path problems? Lack of ground truth constitutes one of the foremost difficulties.

Illustrative Examples and Impact

One of the important applications of social provenance is to find the rumormongers or misinformation centers in social media (Wu et al. 2016). As mentioned in several news recently, misinformation has helped unnecessary fears and conspiracies spread through social media. One such example is related to the Ebola outbreak (http://time.com/3479254/ebola-social-media/). As some potential cases are found in Miami and Washington, DC, some tweets sounded as if Ebola is rampant and some kept tweeting even after government issued a statement to dispel the rumor. The "Assam Exodus" is an another example that illustrates the importance of social provenance. Assam is a large state in the North-East of India and a series of riots broke out in July and August 2012. Following the riots, virulent messages along with misinformation were spread in other parts of India via social media. Bulk text messages (short message services, SMS) and social media sites were extensively used to spread information, aiming to incite certain Indian population against the North-East Indian population. For example, a Wall Street journalist reported that a twitter user used a gory video clip on riots in Indonesia as that of Assam riots (Twitter 2012). Violent messages were also spread on Facebook that incite hatred and vengeance against the North-East Indian population (Facebook). The misinformation as well as virulent messages resulted in deep fear among North-East Indian population, which ultimately led to their exodus from some major metropolitan cities across India, which includes Bangalore, Mumbai, Hyderabad, Chennai, and Pune (Wikipedia 2012). In all of these cases, social provenance might be able to help to find the rumormongers or misinformation sources early and to help stop the viral spread of misinformation.

Knowing the social provenance of a piece of information published in social media – how the piece of information was modified as it was propagated through social media and how an owner of the piece of information is connected to the transmission of the statement – provides additional context to the piece of information. A social media user can use this context to help assess how much value, trust, and validity should be placed on the information.

In early 2010, it was rumored that the Chief Justice of the US Supreme Court was going to retire due to medical reasons. In fact, the Justice had no plans to retire. The statement originated from a Georgetown University Law School class and was meant only to be a teaching point. However, with the availability of the Internet, before the Law professor revealed the falsehood, students in the class had transmitted the statement, which was subsequently published on a news blog (http://www.npr.org/templates/story/story.php?storyId=124371570, http://nymag.com/daily/intelligencer/2010/03/heres_how_the_rumor_that_john.html). Had the social provenance information been made available, recipient users might not have considered the statement credible. In another case, a US Department of Agriculture employee was erroneously fired after information about her appearing in social media was published out of context (https://en.wikipedia.org/wiki/Firing_of_Shirley_Sherrod). Had social provenance information been available, sought out, or examined, it might have prevented an injustice to the employee and embarrassment for the Department of Agriculture. Fake news and its impact on recent US election have widely been reported (http://www.nytimes.com/2016/11/18/technology/fake-news-on-facebook-in-foreign-elections-thats-not-new.html?_r=0, http://www.vox.com/new-money/2016/11/16/13659840/facebook-fake-news-chart). Social provenance, if available, would have informed users its credibility.

The social provenance problem presents an unprecedented challenge, and its research progress can pave way for many equally challenging and important issues such as source trustworthiness, information reliability, and user credibility.

Cross-References

- ▶ Mathematical Model for Propagation of Influence in a Social Network
- ▶ Social Media, Definition, and History
- ▶ Trust in Social Networks
- ▶ User Behavior in Online Social Networks: Influencing Factors

References

Barbier G (2012) Finding provenance data in social media. Doctoral dissertation

Barbier G, Feng Z, Gundecha P, Liu H (2013) Provenance data in social media. Synth Lect Data Min Knowl Discov 4(1):1–84

Data Mining. https://en.wikipedia.org/wiki/Data_mining

Disinformation. https://en.wikipedia.org/wiki/Disinformation

Facebook. https://www.facebook.com/photo.php?fbid=268506716591158&set=a.247241168 71771349889.

247222755386221&type=3&theater. Accessed 17 Dec 2012

Feng Z, Gundecha P, Liu H (2013) Recovering information recipients in social media via provenance. Short paper, the IEEE/ACM international conference on advances in social networks analysis and mining

Gundecha P, Feng Z, and Liu H (2013a) Seeking provenance of information in social media. Short paper, the 22nd ACM international conference on information and knowledge management

Gundecha P, Ranganath S, Feng Z, and Liu H (2013b) A tool for collecting provenance data in social media, Demonstration paper, the 19th ACM SIGKDD international conference on knowledge discovery and data mining

Kaplan AM, Haenlein M (2010) Users of the world, unite! The challenges and opportunities of social media. Bus Horiz 53(1):59–68

Leskovec J, Backstrom L, Kleinberg J (2009) Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 497–506

Misinformation. https://en.wikipedia.org/wiki/Misinformation

Moreau L (2009) The foundations for provenance on the web. Found Trends Web Sci 2:99–241

Ranganath S, Gundecha P, and Liu H (2013) A tool for assisting provenance search in social media. Demonstration paper, the 22nd ACM international conference on information and knowledge management

Shah D, Zaman T (2011) Rumors in a network: who's the culprit? IEEE Trans Inf Theory 57:5163–5181

Social Computing. https://en.wikipedia.org/wiki/Social_computing

Social Network. https://en.wikipedia.org/wiki/Social_network

Twitter (2012) https://twitter.com/dhume01/status/236321660184178688. Accessed 17 Dec 2012

Wikipedia (2012) http://en.wikipedia.org/wiki/2012_Assam_violence#Attacks_on_people_from_North_East_Exodus. Accessed 17 Dec 2012

Wu L, Morstatter F, Hu X, Liu H (2016) Mining Misinformation in Social Media, Big Data in Complex and Social Networks, CRC Press, pp 123–152

http://time.com/3479254/ebola-social-media/. Accessed Oct 2014

http://www.nytimes.com/2016/11/18/technology/fake-news-on-facebook-in-foreign-elections-thats-not-new.html?_r=0. Accessed Dec 2016

http://www.vox.com/new-money/2016/11/16/13659840/facebook-fake-news-chart. Accessed Dec 2016

http://www.npr.org/templates/story/story.php?storyId=124371570. Accessed Dec 2016

http://nymag.com/daily/intelligencer/2010/03/heres_how_the_rumor_that_john.html. Accessed Dec 2016

https://en.wikipedia.org/wiki/Firing_of_Shirley_Sherrod. Accessed Dec 2016

Social Recommendation in Dynamic Networks

Hao Ma[1], Irwin King[2] and Michael R. Lyu[2]
[1]Microsoft Research, Redmond, WA, USA
[2]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong

Synonyms

Collaborative filtering; Matrix factorization; Social network analysis; Social recommender system

Glossary

Collaborative filtering	A type of recommendation technique
Matrix factorization	Factorizing the user-item matrix into user latent matrix and item latent matrix
Recommender system	A system that provides recommendations for users
Social relations	Various social relationships between users, like social trust relationships

Definition

The research of social recommendation aims at modeling recommender systems more accurately and realistically. The characteristic of social recommendation that is different from the tradition recommender system is the availability of social network, i.e., relational information among the users. Social recommendation focuses on how to utilize user social information to effectively and efficiently compute recommendation results.

Introduction

As the exponential growth of information generated on the World Wide Web, the Information Filtering techniques like recommender systems have become more and more important and popular. Recommender systems form a specific type of information filtering technique that attempts to suggest information items (movies, books, music, news, Web pages, images, etc.) that are likely to interest the users. Typically, recommender systems are based on collaborative filtering, which is a technique that automatically predicts the interest of an active user by collecting rating information from other similar users or items. The underlying assumption of collaborative filtering is that the active user will prefer those items which other similar users prefer (Ma et al. 2007). Based on this simple but effective intuition, collaborative filtering has been widely employed in some large, well-known commercial systems, including product recommendation at Amazon and movie recommendation at Netflix.

Due to the potential commercial values and the great research challenges, recommendation techniques have drawn much attention in data mining, information retrieval, and machine learning communities. Recommendation algorithms suggesting personalized recommendations greatly increase the likelihood of customers making their purchases online.

Traditional recommender systems assume that users are independent and identically distributed. This assumption ignores the social relationships among the users. But the fact is, offline, social recommendation is an everyday occurrence. For example, when you ask a trusted friend for a recommendation of a movie to watch or a good restaurant to dine, you are essentially soliciting a verbal social recommendation. In (2001), Sinha and Swearingen (2001) have demonstrated that, given a choice between recommendations from trusted friends and those from recommender systems, in terms of quality and usefulness, trusted friends' recommendations are preferred, even though the recommendations given by the recommender systems have a high novelty factor. Trusted friends are seen as more qualified to make good and useful recommendations compared to traditional recommender systems (Bedi et al. 2007). From this point of view, the traditional recommender systems that ignore the social network structure of the users may no longer be suitable.

Thanks to the popularity of the Web 2.0 applications, recommender systems are now associated with various kinds of social information. This kind of information contains abundant additional information about users, hence providing a huge opportunity to improve the recommendation quality. For example, in users' social trust network, users tend to share their similar interests with the friends they trust. In reality, we always turn to friends we trust for movie, music, or book recommendations, and our tastes and characters can be easily affected by the company we keep. Hence, how to incorporate social information into the recommendation algorithms becomes a trend in the research of recommender systems.

Historical Background

As mentioned in Huang et al. (2004), one of the most commonly used and successfully deployed recommendation approaches is collaborative filtering. In the field of collaborative filtering, two types of methods are widely studied: neighborhood-based approaches and model-based approaches.

Neighborhood-based methods mainly focus on finding the similar users (Breese et al. 1998; Jin et al. 2004) or items (Deshpande and Karypis 2004; Linden et al. 2003; Sarwar et al. 2001) for recommendations. User-based approaches predict the ratings of active users based on the ratings of similar users found, while item-based approaches predict the ratings of active users based on the computed information of items similar to those chosen by the active user. User-based and item-based approaches often use Pearson Correlation Coefficient (PCC) algorithm (Resnick et al. 1994) and Vector Space Similarity (VSS) algorithm (Breese et al. 1998) as the similarity computation methods. PCC method can generally achieve higher performance than VSS approach, since the former considers the differences of user rating style.

In contrast to the neighborhood-based approaches, the model-based approaches to collaborative filtering use the observed user-item ratings to train a compact model that explains the given data, so that ratings could be predicted via the model instead of directly manipulating the original rating database as the neighborhood-based approaches do (Liu and Yang 2008). Algorithms in this category include the clustering model (Kohrs and Merialdo 1999), the aspect models (Hofmann 2003, 2004; Si and Jin 2003), the latent factor model (Canny 2002), the Bayesian hierarchical model (Zhang and Koren 2007), and the ranking model (Liu and Yang 2008). Kohrs and Merialdo (1999) presented an algorithm for collaborative filtering based on hierarchical clustering, which tried to balance both robustness and accuracy of predictions, especially when few data were available. Hofmann (2003) proposed an algorithm based on a generalization of probabilistic latent semantic analysis to continuous-valued response variables.

Recently, due to the efficiency in dealing with large datasets, several low-dimensional matrix approximation methods (Rennie and Srebro 2005; Salakhutdinov and Mnih 2008a, b; Srebro and Jaakkola 2003) have been proposed for collaborative filtering. These methods all focus on fitting the user-item rating matrix using low-rank approximations and employ the matrix to make further predictions. The Low-rank matrix factorization methods are very efficient in training since they assume that in the user-item rating matrix, only a small number of factors influence preferences and that a user's preference vector is determined by how each factor applies to that user. Low-rank matrix approximations based on minimizing the sum-squared errors can be easily solved using Singular Value Decomposition (SVD), and a simple and efficient Expectation Maximization (EM) algorithm for solving weighted low-rank approximation is proposed in Srebro and Jaakkola (2003). In 2004, Srebro et al. (2004) proposed a matrix factorization method to constrain the norms of U and V instead of their dimensionality. Salakhutdinov and Mnih presented a probabilistic linear model with Gaussian observation noise in (2008b). In Salakhutdinov and Mnih (2008a), the Gaussian–Wishart priors are placed on the user and item hyperparameters.

Traditional recommender systems have been well studied and developed both in academia and in industry, but they are all based on the assumption that users are independent and identically distributed, and ignore the relationships among users. Based on this intuition, many researchers have recently started to analyze trust-based recommender systems (Bedi et al. 2007; Massa and Avesani 2004, 2007; O'Donovan and Smyth 2005).

Bedi et al. (2007) proposed a trust-based recommender system for the Semantic Web; this system runs on a server with the knowledge distributed over the network in the form of ontologies and employs the Web of trust to generate the recommendations. In Massa and Avesani (2004), a trust-aware method for recommender system is proposed. In this work, the collaborative filtering process is informed by the reputation of users, which is computed by propagating trust. Trust values are computed in addition to similarity measures between users. The experiments on a large real dataset show that this work increases the coverage (number of ratings that are predictable) while not reducing the accuracy (the error of predictions). In O'Donovan and Smyth (2005), two trust-aware methods are proposed to improve standard collaborative filtering methods. The experimental analysis shows that these trust information can help increase recommendation accuracy.

Previously proposed trust-aware methods are all neighborhood-based methods which employ only heuristic algorithms to generate recommendations. There are several problems with this approach, however. The relationship between the trust network and the user-item matrix has not been studied systematically. Moreover, these methods are not scalable to very large datasets since they may need to calculate the pairwise user similarities and pairwise user trust scores.

Social Recommendation Using Matrix Factorization

Matrix Factorization

In this subsection, we review one popular matrix factorization method that is widely studied in the literature.

Considering an m × n matrix R describing m users' ratings on n items, a low-rank matrix factorization approach seeks to approximate the frequency matrix R by a multiplication of d-rank factors $R \approx U_T V$, where $U \in \mathbb{R}^{d \times m}$ and $V \in \mathbb{R}^{d \times n}$ with $d \ll \min.(m, n)$. The matrix R in the real world is usually very sparse since most of the users only visited a few Web sites.

Traditionally, the Singular Value Decomposition (SVD) method is employed to estimate a matrix R by minimizing

$$\min_{U,V} \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} I_{ij} \left(r_{ij} - \mathbf{u}_i^T \mathbf{v}_j \right)^2, \qquad (1)$$

where u_i and v_j are column vectors with d values and I_{ij} is the indicator function that is equal to 1 if user i rated item j and equal to 0 otherwise.

In order to avoid overfitting, two regularization terms are added into Eq. 1. Hence we have the following Regularized SVD equation:

$$\begin{aligned} \min_{U,V} \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} I_{ij} \left(r_{ij} - \mathbf{u}_i^T \mathbf{v}_j \right)^2 n \\ + \frac{\lambda_1}{2} \|U\|_F^2 + \frac{\lambda_2}{2} \|V\|_F^2, \end{aligned} \qquad (2)$$

where λ_1; $\lambda_2 > 0$. The optimization problem in Eq. 2 minimizes the sum-of-squared-errors objective function with quadratic regularization terms. Gradient-based approaches can be applied to find a local minimum. It also contains a nice probabilistic interpretation with Gaussian observation noise, which is detailed in Salakhutdinov and Mnih (2008b). In Salakhutdinov and Mnih (2008b), the conditional distribution over the observed data is defined as

$$\begin{aligned} p\left(R \mid U, V, \sigma_R^2\right) &= \prod_{i=1}^{m} \prod_{j=1}^{n} \\ &\times \left[\mathcal{N}\left(r_{ij} \mid \mathbf{u}_i^T \mathbf{v}_j, \sigma_R^2 \right) \right]^{I_{ij}}, \end{aligned} \qquad (3)$$

where $\mathcal{N}(x \mid \mu, \sigma^2)$ is the probability density function of the Gaussian distribution with mean μ and variance σ^2. The zero-mean spherical Gaussian priors are also placed on user and item feature vectors:

$$\begin{aligned} p\left(U \mid \sigma_U^2\right) &= \prod_{i=1}^{m} \mathcal{N}\left(\mathbf{u}_i \mid 0, \sigma_U^2 \mathbf{I} \right), \\ p\left(V \mid \sigma_V^2\right) &= \prod_{j=1}^{n} \mathcal{N}\left(\mathbf{v}_j \mid 0, \sigma_V^2 \mathbf{I} \right). \end{aligned} \qquad (4)$$

Through a Bayesian inference, we can easily obtain the objective function in Eq. 2.

By adopting a simple stochastic gradient descent technique, for each observed rating r_{ij}, we have the following efficient updating rules to learn latent variables u_i, v_j:

$$\begin{aligned} \mathbf{u}_i &\leftarrow \mathbf{u}_i + \gamma_1 \left(\Delta_{ij} \mathbf{v}_j - \lambda_1 \mathbf{u}_i \right), \\ \mathbf{v}_j &\leftarrow \mathbf{v}_j + \gamma_2 \left(\Delta_{ij} \mathbf{u}_i - \lambda_2 \mathbf{v}_j \right), \end{aligned} \qquad (5)$$

where $\Delta_{ij} = r_{ij} - \mathbf{u}_i^T \mathbf{v}_j$, and γ_1; γ_2 are the learning rates.

The Regularized SVD algorithm introduced in this section is both effective and efficient in solving the collaborative filtering problem, and it is perhaps one of the most popular methods in collaborative filtering.

Social Trust Ensemble

However, the above algorithm does not consider any information from users' social network. In order to better model the recommendation problem, in Ma et al. (2009), Ma et al. proposed a matrix factorization-based Social Trust Ensemble (STE) method upon the following intuitions:

- Users have their own tastes.
- Users can also be easily influenced by the trusted friends they have.

- A user's final rating is composed of the combination of this user's own taste and this user's friends' tastes.

Based on the above interpretations, the objective function can be formulated as

$$\begin{aligned} L &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}\left(r_{ij} - \left(\alpha \mathbf{u}_i^T \mathbf{v}_j + (1-\alpha)\sum_{k\in\mathcal{T}(i)} w_{ik}\mathbf{u}_k^T\mathbf{v}_j\right)\right)^2 \\ &\quad + \frac{\lambda_1}{2}\|U\|_F^2 + \frac{\lambda_2}{2}\|V\|_F^2, \end{aligned} \tag{6}$$

where α is a parameter to balance the impact of user's own taste and user's friends' tastes, $\mathcal{T}(i)$ represents a list of user i's trusted friends, and w_{ik} is a normalized weight that equals to $1/| \mathcal{T}(i) \|$.

We can see that in this approach, a user's latent factor is smoothly integrated with this user's trusted friends' tastes. This equation also coincides with the real-world observation that we always ask our friends for movies, books, or music recommendations.

For each observed rating r_{ij}, the stochastic gradient decent learning rules for this method are

$$\begin{aligned} \mathbf{u}_i &\leftarrow \mathbf{u}_i \\ &+\gamma_1\left(\Delta_{ij}\left(\alpha + (1-\alpha)\sum_{p\in\mathcal{B}(i)} w_{pi}\right)\mathbf{v}_j - \lambda_1\mathbf{u}_i\right), \\ \mathbf{v}_j &\leftarrow \mathbf{v}_j \\ &+\gamma_2\left(\Delta_{ij}\left(\alpha\mathbf{u}_i + (1-\alpha)\sum_{k\in\mathcal{T}(i)} w_{ik}\mathbf{u}_k\right) - \lambda_2\mathbf{v}_j\right), \end{aligned} \tag{7}$$

where

$$\begin{aligned} \Delta_{ij} &= r_{ij} \\ &- \left(\alpha\mathbf{u}_i^T\mathbf{v}_j + (1-\alpha)\sum_{k\in\mathcal{T}(i)} w_{ik}\mathbf{u}_k^T\mathbf{v}_j\right), \end{aligned} \tag{8}$$

and $\mathcal{B}(i)$ is the set that includes all the users who trust user i.

Social Regularization

The STE method mentioned above is originally designed for trust-aware recommender systems. In trust-aware recommender systems, we can always assume that users have similar tastes with other users they trust. Unlike trust relationships among users, the tastes among social friend relationships are more diverse. User k is a friend of user i does not necessarily indicate that user k has similar taste with user i. Hence, in order to model the social recommendation problems more accurately, another more general social recommendation approach, Social Regularization (SR), is proposed in Ma et al. (2011).

The objective function of this approach is formulated as

$$\begin{aligned} L = \frac{1}{2}&\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}\left(r_{ij} - \mathbf{u}_i^T\mathbf{v}_j\right)^2 \\ &+ \frac{\alpha}{2}\sum_{i=1}^{m}\sum_{f\in\mathcal{F}^+(i)} s_{if}\left\|\mathbf{u}_i - \mathbf{u}_f\right\|_F^2 \\ &+ \frac{\lambda_1}{2}\|U\|_F^2 + \frac{\lambda_2}{2}\|V\|_F^2, \end{aligned} \tag{9}$$

where s_{if} indicates the similarity between user i and user f and $\mathcal{F}^+$. (i) represents user i's outlink friends.

In this method, the social network information is employed in designing the social regularization term to constrain the matrix factorization objective function. The social regularization term also indirectly models the propagation of tastes. More specifically, if user i has a friend f and user f has a friend user g, this regularization term actually indirectly minimizes the distance between latent vectors u_i and u_g. The propagation of tastes will reach a harmonic status once the learning is converged.

Similarly, for each observed rating r_{ij}, we have the following stochastic gradient descent updating rules to learn the latent parameters:

$$\begin{aligned} &\mathbf{u}_i \leftarrow \mathbf{u}_i \\ &+\gamma_1\left(\Delta_{ij}\mathbf{v}_j - \alpha\sum_{f\in\mathcal{F}^+(i)} s_{if}\left(\mathbf{u}_i - \mathbf{u}_f\right) - \alpha\sum_{g\in\mathcal{F}^-(i)} s_{ig}\left(\mathbf{u}_i - \mathbf{u}_g\right) - \lambda_1\mathbf{u}_i\right), \\ &\times\mathbf{v}_j \leftarrow \mathbf{v}_j + \gamma_2\left(\Delta_{ij}\mathbf{u}_i - \lambda_2\mathbf{v}_j\right), \end{aligned} \tag{10}$$

where $\Delta_{ij} = r_{ij} - \mathbf{u}_i^T \mathbf{v}_j$, and $\mathcal{F}^-(i)$ represents user i's inlink friends.

The experiments conducted in Ma et al. (2009, 2011) suggest that social recommendation algorithms outperform traditional recommendation algorithms, especially when the user-item matrix is sparse. This indicates that using social information is a promising direction in the research of recommender systems.

Future Directions

The methods mentioned above can be solved efficiently by using simple gradient descent or stochastic gradient descent algorithms. However, for statistical machine learning's point of view, the methods themselves are not full Bayesian methods. Hence, learning those methods can easily have the overfitting problem. How to apply full Bayesian method on these models hence becomes worth of studying.

We already demonstrate how to recommend by incorporating users' social trust and friend information. Actually, sometimes there are more data sources available on Web 2.0 sites, such as tags issued by users to items and temporal information. These sources are also valuable information to improve recommender systems.

Cross-References

- ▶ Human Behavior and Social Networks
- ▶ Inferring Social Ties
- ▶ Matrix Decomposition
- ▶ Probabilistic Graphical Models
- ▶ Recommender Systems Using Social Network Analysis: Challenges and Future Trends
- ▶ Social Recommender System

Acknowledgments The work described in this article is supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project No.: CUHK 413212 and CUHK 415212).

References

Bedi P, Kaur H, Marwaha S (2007) Trust based recommender system for semantic web. In: Proceedings of IJCAI'07, Hyderabad, pp 2677–2682

Breese JS, Heckerman D, Kadie C (1998) Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of UAI'98, Madison

Canny J (2002) Collaborative filtering with privacy via factor analysis. In: Proceedings of SIGIR'02, Tampere, pp 238–245

Deshpande M, Karypis G (2004) Item-based top-n recommendation. ACM Trans Inf Syst 22(1):143–177

Hofmann T (2003) Collaborative filtering via Gaussian probabilistic latent semantic analysis. In: Proceedings of SIGIR'03, Toronto, pp 259–266

Hofmann T (2004) Latent semantic models for collaborative filtering. ACM Trans Inf Syst 22(1):89–115. https://doi.org/10.1145/963770.963774

Huang Z, Chen H, Zeng D (2004) Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. ACM Trans Inf Syst 22 (1):116–142

Jin R, Chai JY, Si L (2004) An automatic weighting scheme for collaborative filtering. In: Proceedings of SIGIR'04, Sheffield, pp 337–344

Kohrs A, Merialdo B (1999) Clustering for collaborative filtering applications. In: Proceedings of CIMCA, Gold Coast

Linden G, Smith B, York J (2003) Amazon.com recommendations: item-to-item collaborative filtering. IEEE Intern Comput 7:76–80

Liu NN, Yang Q (2008) Eigenrank: a ranking-oriented approach to collaborative filtering. In: Proceedings of SIGIR'08, Singapore, pp 83–90

Ma H, King I, Lyu MR (2007) Effective missing data prediction for collaborative filtering. In: Proceedings of SIGIR'07, Amsterdam, pp 39–46

Ma H, King I, Lyu MR (2009) Learning to recommend with social trust ensemble. In: Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval, SIGIR'09, Boston, pp 203–210

Ma H, Zhou D, Liu C, Lyu MR, King I (2011) Recommender systems with social regularization. In: Proceedings of the fourth ACM international conference on web search and data mining, WSDM'11, Hong Kong, pp 287–296

Massa P, Avesani P (2004) Trust-aware collaborative filtering for recommender systems. In: Proceedings of CoopIS/DOA/ODBASE, Irvine, pp 492–508

Massa P, Avesani P (2007) Trust-aware recommender systems. In: Proceedings of RecSys'07, Minneapolis, pp 17–24

O'Donovan J, Smyth B (2005) Trust in recommender systems. In: Proceedings of IUI'05, San Diego, pp 167–174

Rennie JDM, Srebro N (2005) Fast maximum margin matrix factorization for collaborative prediction. In: Proceedings of ICML'05, Bonn

Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J (1994) Grouplens: an open architecture for collaborative filtering of netnews. In: Proceedings of CSCW'94, Chapel Hill

Salakhutdinov R, Mnih A (2008a) Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In: Proceedings of ICML'08, Helsinki

Salakhutdinov R, Mnih A (2008b) Probabilistic matrix factorization. In: Proceedings of NIPS'08, vol 20, Vancouver

Sarwar B, Karypis G, Konstan J, Reidl J (2001) Item-based collaborative filtering recommendation algorithms. In: Proceedings of WWW'01, Hong Kong, pp 285–295

Si L, Jin R (2003) Flexible mixture model for collaborative filtering. In: Proceedings of ICML'03, Washington, DC

Sinha RR, Swearingen K (2001) Comparing recommendations made by online systems and friends. In: DELOS workshop: personalisation and recommender systems in digital libraries, Dublin

Srebro N, Jaakkola T (2003) Weighted low-rank approximations. In: Proceedings of ICML'03, Washington, DC, pp 720–727

Srebro N, Rennie JDM, Jaakkola T (2004) Maximum-margin matrix factorization. In: Proceedings of NIPS'04, Vancouver

Zhang Y, Koren J (2007) Efficient Bayesian hierarchical user modeling for recommendation system. In: Proceedings of SIGIR'07, Amsterdam, pp 47–54

Social Recommender System

▶ Recommender Systems, Semantic-Based

▶ Social Recommendation in Dynamic Networks

Social Recommender Systems

▶ Community Detection and Recommender Systems

Social Reconnaissance

▶ Reconnaissance and Social Engineering Risks as Effects of Social Networking

Social Relationships

▶ Actionable Information in Social Networks, Diffusion of

▶ Inferring Social Ties

Social Search and Querying

Georgia Koloniari[1], Panagiotis Lionakis[2] and Kostas Stefanidis[3]
[1]Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece
[2]Department of Computer Science, University of Crete, Heraklion, Greece
[3]School of Information Sciences, University of Tampere, Tampere, Finland

Synonyms

Logical algebra; Query model; Search and querying social data; Time-aware social search

Glossary

Social content — The content that appears in social networks due to user activities, reflecting their relationships with other users and the content they shared

Social graph — A graph representing a social network. The nodes of the graph correspond to the entities of the social network, while the edges capture the social relationships between the entities. Typically, there are two types of entities: the type user, representing the social network users or participants; and the type object, including all other entities other than users, e.g., images, videos, events, and applications

Social network — A structure, typically nowadays expressed as an online application, that enables social interactions and personal relationships between users, by allowing them to post information, comments, messages, images, and videos

Definition

Web search and querying is the precursor of social search, and typically offers searching functionalities on the web. Results are usually a mix of web pages, images, and various types of files. Also results may be mined from available databases or open directories, or even they can be derived by running algorithms on web crawlers. Recently, web search is enhanced with social data, due to the increasing popularity of social networks and the vast amount of information contained in them. The goal of social search is mainly to retrieve user-generated content such as images, videos, or news appearing in social media like Facebook, Twitter, and Instagram. Thus, social search combines traditional searching algorithms with online community filtering to produce personalized results, based on the intuition that the network of a user might be more relevant to her particular needs.

Introduction

Due to the increasing popularity of social networks and the vast amount of information in them, recently there have been many efforts in enhancing web search based on social data. Specifically, social networks contain different types of data, namely data about the network, i.e., data for which users visit the network, data about the users and their social connections, and data about the social activities users perform. The richness of data that social networks know about users comes through the rich information users voluntarily provide that reflects their interests. Therefore, an effective and efficient way to query social data is essential for satisfying the users search needs. This has lead to the emergence of social search and querying that utilizes the underlying graph structure and the content of a social network to provide both more personalized and expressive search features for the users.

The way in which social networks help users explore and discover data, has recently evolved from the traditional methods of keyword-based search. Actually, there is a gap between typical information retrieval approaches and approaches that target social-related information recommendation. That is, information retrieval focuses on identifying data semantically relevant to the users queries, while social-related information recommendation focuses on locating data that users may like, based on their social profile activities and the activities of their social connections. The principle behind the speedily growing social networks is that the query results coming from these user-oriented networks could be more meaningful and relevant for the users, instead of having algorithms deciding the results for specific queries. Results are still presented in a ranked-based manner from the most useful to the less useful ones, where usefulness defines the value of results for the users in the query.

Besides the different types of data that social networks contain, forming a structure dictated by the relationships among the entities, i.e., users and objects, of the network, another important dimension of social networks is their dynamic nature. New content is added through user activities and updates occur both in the structure of the graph and the content shared, representing respective changes in the users' interests. This temporal aspect of the information should influence social search either explicitly by enabling users to query for particular time points or periods (Koloniari et al. 2012; Ren et al. 2011; Koloniari and Pitoura 2013) or implicitly by providing the most recent results and higher ranking of fresher content (Joho et al. 2013; Stefanidis et al. 2013; Huo and Tsotras 2014).

Key Points

By exploiting the content of social networks and their temporal aspects, we focus in this entry on a complete social search framework that tries to satisfy the varying user search needs. Specifically, we present a time-independent and a time-dependent query model and their corresponding logical algebras that both can be applied to social networks.

- To encompass different querying needs, we use a query model that exploits the underlying social graph representation. The model defines queries for the entities of the network, i.e., users and objects. It supports two querying perspectives. *User-centric queries* offer a personalized search feature by exploiting the social relationships of a user, whereas *system-centric queries* provide a global search feature with many applications in online-shopping and target-advertising, so as to select the best group for a new product or the best products to promote to a given user. This model supports queries independent of time. In addition, it is extended to enables time-awareness by allowing time-dependent queries that exploit time explicitly.
- To deploy the querying model over any social network, we use a logical algebra that provides the set of basic operators required to evaluate both the time-independent and time-dependent queries specified by the corresponding query model. Besides the basic operators, the algebra includes a set of operators for supporting a ranking functionality. The ranking mechanism enables the implicit use of time to enhance the results of a query by providing a time-dependent ranking, so that more recent or fresher results are returned first.

Historical Background

In the classic problem of web search, the user supplies a keyword query which expresses her information needs. Given a collection of web documents that are indexed by a web search engine, the documents matching the user query are determined by using information retrieval techniques. The final results are presented to the user ranked based on a score that quantifies their relevance to the input query.

As the web scaled up and bigger portions of it became commercialized, the need to improve web search and prevent results from being manipulated for profit emerged. Thus, from early on, researchers proposed to exploit besides the content of the web, also its structure as dictated by the links between web pages. Thus, algorithms such as PageRank (Brin and Page 1998) and HITS (Kleinberg 1999) were proposed to identify important documents on the web, determining their importance based on their in- and outgoing links in the web graph. Thus, query results are not only ranked based on their relevance to the user query but also on their importance in the web graph.

Personalized web search focused more on the user perspective, trying to better fulfil the information needs of users by providing results tailored to their preferences and characteristics (Jeh and Widom 2003). A user's web search history is exploited in this case to provide context for her future queries.

Therefore, the continuous trend is to enrich the web search process with all the available information exploiting diverse sources of data such as the web content, the web structure, and the web usage data. The first attempts on exploiting social content to enhance web search is through social annotations (Bao et al. 2007). In this case, tags provided by users are considered as appropriate web page summaries that are more descriptive of the actual contents of a page. Queries are evaluated against these tagbased summaries to calculate their relevance to each page.

Data Model

The typical entities of a social network represent users and objects. An entity n_i is described by a set of predicates a_{i_j} of the form (a_{i_j}.attribute = a_{i_j}.value). For example, an attribute for a user can be "*name* = *Alice*" and an attribute for an object, e.g., an event, can be "*location* = *Shanghai*."

Time-Independent Data Model: A social network can be modeled as an undirected graph, $G = (V, E)$. The set of nodes V corresponds to the entities that belong to the social network, i.e., $V = U \cup O$, where U is the set of users and O the set of objects. The set of edges E captures the relationships between the entities that belong to V; user-to-user edges capture the friendship between the corresponding users, while user-to-object edges declare that a user uses or participates in some way in an object o_j.

Time-Dependent Data Model: The typical graph model can be extended with temporal information toward making social search time-dependent. We exploit the valid time of data items and consider an element, node, or edge, of a graph G as *valid* for the period for which the corresponding element of the network it represents is also valid. To do this, each element e_i in the social graph is annotated with a label with the time intervals for which the element is valid. To cope with the dynamic nature of the social network that causes elements to become valid (a user joins the network), invalid (it leaves), and then valid again (rejoins the network), for each element $e_i \in G$, its label is defined as a set of disjoint intervals $l(e_i) = \left\{ \left(t^j_{\text{start}}, t^j_{\text{end}} \right) | j \geq 1 \right\}$, which implies that element e_i is valid for the time intervals $\left\{ \left[t^1_{\text{start}}, t^1_{\text{end}} \right), \left[t^2_{\text{start}}, t^2_{\text{end}} \right), \ldots \right\}$.

Query Model

The focus in this section is on how to support queries for the social graph that exploit the graph structure, as well as its time dimension. Specifically, one can discern between four structural parts that compose a query:

1. The set of result nodes V', which are the nodes to be retrieved and form the query result.
2. The set of qualifying nodes QV, which are the intermediate nodes through which the different entities in the query are connected, and are specified by a set of predicates P.
3. The reference node u_i, which is the node around which the query is centered, and a distance d from it.
4. The time constraint T that checks for the validity of nodes and edges involved in the query.

More formally, a query Q can be defined as follows:

Definition 1 Given a graph $G = (V, E)$, $V = U \cup O$, predicates P, a user node u_i, a distance d, and time constraints T, we define a query Q as a query that retrieves a set of nodes $V' \subseteq V$, such that, $v_k \in V'$, if and only if, $\forall v_l \in V$ for which $P(v_l) = true$ and (v_l is valid according to T and $\exists (v_j, v_l) \in E$), or ($\exists (v_j, v_l) \in E$ that is valid according to T), and

- if $v_l \in U$, $\exists$ path between v_i and v_l with length $= d$, or
- if $v_l \in O$, $\exists$ path between v_i and v_j with length $= d$.

For clarifying this, let us consider an example of a popular query, $Q1$: "*find all of Ross' friends that attend sports events in July 2016*." This query expresses constraints both on the structure of the graph and on time. We use $Q1$ as our illustrative example to demonstrate the different possibilities one has for querying the social graph, so as to deduce a complete query model that covers our social search needs.

So, let us begin with the result of $Q1$. The query retrieves a subset of the friends of the user *Ross*, i.e., the result set, V', is a set of user nodes, $V' \subseteq U \subseteq V$. Similarly, one could query for object instead of user nodes, i.e., $Q2$: "*find all events in July 2016 that Ross' friends attend*," where it holds: $V' \subseteq O \subseteq V$. Thus, we can discern, according to the type of nodes that form our result set, between *queries for friends*, such as $Q1$, and *queries for objects*, such as $Q2$. The first type gives emphasis on the company of the user, while the second focuses on the objects to be consumed.

We ignore the time constraint for now, as we treat time as a separate dimension. As our goal is to exploit the graph structure, our queries include constraints on the connectivity between the different entities that are referenced in the query context. In particular, $Q1$ requires the result nodes to be directly connected to a set of intermediate object nodes specified through a set of predicates, i.e., $Q1$ requires the friends of Ross that are connected with objects of type sports events. Let P be a set of predicates concerning the attributes in the nodes descriptions of the form (*attribute* θ *value*). For numerical attributes, $\theta \in \{=, <, >, \leq, \geq, \neq\}$, and, for nonnumerical attributes, $\theta \in \{=, \neq, \mathit{prefix}\}$. We say that $P(v_i) = \mathit{true}$, if and only if, node v_i satisfies all predicates in P. We refer to nodes, for which P is *true*, as the *qualifying* nodes in the query, denoted as QV. The retrieved nodes (the ones forming the result set) need to be connected to all the qualifying nodes, i.e., we want the friends of Ross that attend all sports events in July 2016. Since the *all* operator is a very strict requirement, it can be relaxed to require that the retrieved nodes are at least connected to one qualifying node, i.e., the friends of Ross that attend *any* sports event in July 2016.

Similarly to the result nodes, the predicates can also refer to either user or object nodes. To illustrate this, for $Q2$, the result nodes, i.e., events, need to be connected to all Ross' friends. In this example, P is empty as we have no predicates specified for these user nodes. Note that as our graph model does not define connections between objects, for a query for objects the predicates are only applied on user nodes. However, we may also define queries for users on which the predicates are also applied on other users. For instance, $Q3$: "*find all of Ross' friends that are friends with users with occupation athlete*." In $Q3$, the result user nodes need to be connected (be friends) with other users for which the predicate *occupation = athlete* is true.

Furthermore, $Q1$ includes a reference node, i.e., the user node that corresponds to user *Ross*. Our query is centered around this reference node, u_i, and we call such queries *user-centric queries*. $Q1$ requires the user nodes that form the result set to be friends with *Ross*, i.e., it requires for them to be directly connected to the reference node. We can consider extensions, such as $Q4$: "*find all the friends of the friends of Ross that attend sports events in July 2014*." In this example, the result nodes are at distance $d = 2$ from the reference node, which means that there is a path of two edges connecting them. Although $d > 2$ is also plausible, in practice it is rarely used.

$Q2$ is also a user-centric query, but in this case, the reference node is not directly connected to the result nodes. Instead, it is required that the qualifying user nodes of the query are directly connected to *Ross*.

On the other hand, one may also want to pose queries where no reference node is specified at all, i.e., *system-centric* queries. For instance, consider query $Q5$: "*find all users that attend sports events in July 2016*." Such queries capture the system perspective and their goal is to identify either sets of users that share some common interests and, for instance, may be interested in a particular product or event the system wants to promote, or similarly, sets of objects that may be of interest to some particular users so that they can choose to promote these objects to them. System-centric queries can be extended to allow for connectivity constraints to be specified on the returned results. For instance, $Q6$: "*find all groups of users that are friends and attend sports events in July*" requires not a set of user nodes as a result, but groups of such users that are directly connected within each group.

The last part of our query is the time constraint. Thus, for $Q1$, T specifies that the sports events that Ross' friends attend are valid in July 2016. In $Q1$, T is applied on the qualifying nodes QV and therefore can be treated as another predicate that however does not concern the object descriptions, but rather their labels. Another possibility is to specify time constraints on the result nodes, i.e., $Q7$: "*find all the friends of Ross that are valid in July 2016 and have attended sports events*." If a time constraint is specified for any part of the query, then we are able to capture *time-dependent* queries.

T can be also applied on the labels of the edges of the social graph, introducing a different type of queries. For instance, $Q8$: "*find all the friends of Ross that express that they will attend VLDB 2016 in August 2016*," requires a set of users that have established their connections with an event during a specific time period.

Given that T is defined as a time period $[b, c)$, we discuss next how we handle *validity*. Specifically, to determine whether a node v_i is valid for T, one needs to compare T against $l(v_i)$. If one of the time intervals included in $l(v_i)$ is included in the interval specified by T, i.e., $\exists \left(t^j_{\text{start}}, t^j_{\text{end}}\right) \in l(v_i)$, such that, $t^j_{\text{start}} \geq b$ and $t^j_{\text{end}} < c$, then v_i is valid for T. A reverse interpretation is also possible. That is, one could require the valid time of the node to include T, i.e., $\exists \left(t^j_{\text{start}}, t^j_{\text{end}}\right) \in l(v_i)$, such that, $t^j_{\text{start}} < b$ and $t^j_{\text{end}} > c$. From a different perspective, we can consider *before* and *after* semantics, requiring $t^j_{\text{start}} < b$ or $t^j_{\text{start}} > c$, respectively. In a more relaxed interpretation, it suffices for the valid time of a node to simply intersect with T, for the node to be considered valid. Finally, for $b = c$, time point queries are supported. Similarly, we can define the validity of the edges appearing in the social graph.

A Logical Algebra

Developing a flexible and expressive mechanism to manipulate data in social graphs is an important challenge. Our focus here is on integrating in a principled way the search process into the context of social graphs. Toward discovering information that covers the needs of users in a flexible manner, we exploit the following logical algebraic framework. By using this logical algebra, we can express sophisticated tasks for retrieving data relevant to the user queries both at a semantic (e.g., by querying for specific predicates) and social (e.g., by querying for friendships) level. We start with the core operators of the algebra that can be used for evaluating the system- and user-centric queries introduced above, and then define time-dependent and ranking operators.

Time-Independent Operators: The *select node* operator takes as input a set of nodes V, a set of predicates P, and a parameter Z that defines the retrieving focus of the operator, i.e., users U or objects O. The operator outputs the nodes that satisfy the predicates in P. Formally:

Definition 2 (Select Node Operator) $\sigma_{P,Z}(V) = \{v | v \in Z \wedge P(v) = true\}$.

This operator is appropriate for implementing system-centric queries.

For user-centric queries, we have additional parameters, i.e., the reference node and a distance from it. Thus, we define the *select node from graph* operator that takes as input a social graph G, a node v_i corresponding to a user $u_i \in U$, a distance d, a set of predicates P, and a parameter Z that defines the retrieving focus of the operator, i.e., users U or objects O. The operator outputs the nodes v_j from G that satisfy the predicates in P for which there exists at least one path with length at most d between v_i and v_j. Formally:

Definition 3 (Select Node from Graph Operator) $\sigma_{P,Z,d}\ (v_i,G) = \{v_j | v_j \in Z \wedge P\ (v_j) = true \wedge \exists$ path between v_i and v_j with $length \leq d\}$.

In the main framework, we define also the *difference operator* for excluding from a set, nodes that are not directly connected with the nodes of a different set.

Definition 4 (Difference Operator) $dif\ (V_i,V_j, G) = \{v_x | v_x \in V_i \wedge \exists v_y \in V_j$, such that, $\exists\ (v_x, v_y)$ in $G\}$.

Given $G = (V,E)$, the query $Q9$: "*find all Ross' friends that have attended sports events*" can be accomplished as follows. Direct friends of Ross, say V_1, are captured by $\sigma_{\{\},friends,1}(Ross,G)$. We retrieve the objects, say V_2, with $topic = sports\ events$, by $\sigma_{topic=sports\ events,objects}\ (V)$, and $dif\,(V_1, V_2,G)$ locates the subset of Ross' friends that have attended sports events.

Time-Dependent Operators: For handling time-dependent queries, the above operators are extended in order to take into account time constraints T for users and/or objects.

Definition 5 (Temporal Select Node Operator) $\sigma_{P,Z,T}(V) = \{v|v \in Z \wedge P(v) = true \wedge v\ is\ valid\ for\ T\}$.

Definition 6 (Temporal Select Node from Graph Operator) $\sigma_{P,Z,T,d}(v_i,G) = \{v_j|v_j \in Z \wedge P(v_j) = true \wedge \exists$ path between v_i and v_j with $length \leq d \wedge v_j$ is valid for $T\}$.

Then, for evaluating $Q1$ that augments $Q8$ with the temporal constraint [1/12/2013, 31/12/2012) for events, we locate set $V2$ as:

$$\sigma_{\text{topic}=\text{sports events},\ \ \text{objects},\ [1/12/2013,\,31/12/2012)}(V).$$

Operators for locating valid edges are defined in a similar manner.

Ranking Operators: To provide more meaningful results than a simple set of returned nodes and enable the implicit use of time, we exploit a ranking functionality. Ranking is time-dependent, in the sense that it ranks more recent nodes higher. To determine how recent a node is, we rely on the time of the user activities rather than the actual valid time of the nodes themselves, as we expect that more recent activities tend to better reflect the current trends in the network. This information is captured in the labels of the edges of the network that connect the result and qualifying nodes.

Thus, given a node v_j, we use the notion of *freshness* of a node v_i (*fresh*(v_i)) as the maximum t_{start} value in the labels of the edges that connect v_i and v_j. In particular, to support the ranking functionality, we use the *temporal social winner* and *rank* operators.

A node v_i belongs to the winner if there is no node v_l with age greater than the age of v_i.

Definition 7 (Temporal Social Winner Operator) winner $(V) = \{v_i \mid v_i \in V \wedge \nexists v_l \in V$, such that, fresh $(v_l) >$ fresh $(v_i)\}$.

Ranking all nodes in V can be achieved by repetitive applications of this operator.

Definition 8 (Temporal Social Rank Operator)

$$\text{rank}^i(V) = \begin{cases} \text{rank}^1(V) = \text{winner}(V) \\ \text{rank}^{i+1}(V) = \text{winner}\left(V - \cup_{k=1}^{i}\text{rank}^k(V)\right) \end{cases}$$

In general, there are different ways for handling operators. Firstly, operators can be implemented on-top of a DBMS either as standalone programs or as user-defined functions. Alternatively, operators may be translated into other, existing relational algebra operators during a preprocessing step. Finally, operators can be implemented inside the database engine using specific physical operators and algorithms.

Key Applications

Social search applications can be divided into two categories: either applications that enable querying and searching information derived solely from social networks or applications that by exploiting this social information enhance web search, usually by providing alternative ranking mechanisms.

In the first category, each social network (e.g., Facebook, Twitter) provides its own search functionality. Facebook also introduced a search functionality, Facebook Graph Search, that took into account the structural information of the social graph. In particular, it allowed users to pose queries for users or objects directly connected to other nodes in the network, and it also provided a "near" operator for supporting geographical locality in queries. Google presented a search engine, Social Searcher, that enables users to select which social network they want to query by posing keyword queries. The engine makes implicit use of time by ranking results based either on relevance or time. SQTime (Lionakis et al. 2014) is also a system designed specifically to enable time-aware social search by supporting both historical queries and ranking of query results based on time.

Searching for people or person search is an alternative application that falls in this first category of social search. Instead of searching for results matching a query in a social network, this type of search aims at locating the appropriate person or people that can provide the best answers to a user query or better match a description specified by the query. Aardvark (Horowitz and Kamvar 2010) is a social search engine that searches the social graph to locate the most relevant users to another user input query. The engine does not rely on its own social network, but derives its social graph by exploiting the social networks that a user belongs to. It orders a query's result, i.e., the users likely to provide the best answers to the query, according to the strength of their connection with the user who posed the query. In (Hsieh et al. 2015), person search aims at locating in a social network the people that better match a set of social labels, such as interests, home town, and others. The ranking mechanism exploits the structure of the social graph ranking higher either users that are more well-connected in the global social graph, and thus, assumed as more important, or users that are in closer proximity with the user that posed the query, similarly to Aardvark.

In the second category, based on the assumption that user needs are often varied and require more than one result to their queries, the premise is that web search combined with information derived from social content might yield the most promising results (Amer-Yahia et al. 2009).

Personalized search is an application in this category (for a survey on personalized data management, see (Stefanidis et al. 2011)). To offer results that better reflect a specific user's information needs, instead of relying on user web history, personalized social search makes use of knowledge derived from a user's social network, such as her profile and her connections to other users in the social graph structure. In (Carmel et al. 2009), social information is used to provide a reranking of the web results for a user query. The basic idea is similar to (Horowitz and Kamvar 2010) and (Hsieh et al. 2015), where query results are ranked based on the connection strength in the graph of the node representing the user that posed the query with the nodes containing the query results. In (Bender et al. 2008), the goal is not only to provide a better ranking of the web results, but to retrieve results that better match the user information needs. To this end, tags are incorporated in the social graph and used for query expansion and rewriting so as to attain more relevant results.

Future Directions

In the future, going beyond harvesting the information from a single social network, we envision a framework that enables open access to a social graph that as a whole brings together different social networks. Time-awareness is also essential in a scenario such as this, as it is essential to answer the query needs of users who are looking for information obtained by integrating numerous and heterogeneous sources.

Cross-References

- ▶ Geotemporal Querying of Social Networks and Summarization
- ▶ Query Answering in the Semantic Social Web: An Argumentation-Based Approach
- ▶ Querying Volatile and Dynamic Networks
- ▶ Social Web Search
- ▶ Stream Querying and Reasoning on Social Data

References

Amer-Yahia S, Lakshmanan LVS, Yu C (2009) Socialscope: enabling information discovery on social content sites. In: CIDR 2009, Fourth biennial conference on innovative data systems research

Bao S, Xue G, Wu X, Yu Y, Fei B, Su Z (2007) Optimizing web search using social annotations. In: Proceedings of the 16th international conference on World Wide Web, WWW '07, pp 501–510

Bender M, Crecelius T, Kacimi M, Michel S, Neumann T, Parreira JX, Schenkel R, Weikum G (2008) Exploiting social relations for query expansion and result ranking. In: IEEE 24th international conference on data engineering workshop, ICDEW, pp 501–506

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. In: Proceedings of the seventh international conference on World Wide Web 7, WWW7, pp 107–117

Carmel D, Zwerdling N, Guy I, Ofek-Koifman S, Har'el N, Ronen I, Uziel E, Yogev S, Chernov S (2009) Personalized social search based on the user's social network. In: Proceedings of the 18th ACM conference on information and knowledge management, CIKM '09, pp 1227–1236

Horowitz D, Kamvar SD (2010) The anatomy of a large-scale social search engine. In: Proceedings of the 19th international conference on World Wide Web, WWW '10, pp 431–440

Hsieh H-P, Li C-T, Yan R (2015) I see you: person-of-interest search in social networks. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, SIGIR '15, pp 839–842

Huo W, Tsotras VJ (2014) Efficient temporal shortest path queries on evolving social graphs. In: Conference on scientific and statistical database management, SSDBM '14, Aalborg, 30 June–02 July 2014, pp 38:1–38:4

Jeh G, Widom J (2003) Scaling personalized web search. In: Proceedings of the 12th international conference on World Wide Web, WWW '03, pp 271–279

Joho H, Jatowt A, Roi B (2013) A survey of temporal web search experience. In: Proceedings of the 22nd international conference on World Wide Web, WWW '13 Companion, pp 1101–1108

Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632

Koloniari G, Pitoura E (2013) Partial view selection for evolving social graphs. In: First international workshop on graph data management experiences and systems, GRADES 2013, co-loated with SIGMOD/PODS

Koloniari G, Souravlias D, Pitoura E (2012) On graph deltas for historical queries. First International Workshop on Online Social Systems (WOSS), in conjunction with VLDB

Lionakis P, Stefanidis K, Koloniari G (2014) Sqtime: time-enhanced social search querying. In: Advances in conceptual modeling – ER Demos. Atlanta, GA, USA, pp 303–307

Ren C, Lo E, Kao B, Zhu X, Cheng R (2011) On querying historical evolving graph sequences. PVLDB 4(11):726–737

Stefanidis K, Koutrika G, Pitoura E (2011) A survey on representation, composition and application of preferences in database systems. ACM Trans Database Syst 36(3):19

Stefanidis K, Ntoutsi E, Petropoulos M, Nørvåg K, Kriegel H (2013) A framework for modeling, computing and presenting time-aware recommendations. In: Transactions on large-scale data- and knowledge-centered systems, vol 10. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 146–172

Social Selection

▶ Stability and Evolution of Scientific Networks

Social Spam

▶ Spam Detection: E-mail/Social Network

Social Status

▶ Role Discovery

Social Structural Analysis

▶ Origins of Social Network Analysis

Social Tagging

▶ Folksonomies

Social Tags

▶ Social Bookmarking or Tagging

Social Theory

▶ Web Communities Versus Physical Communities

Social Trends Discovery

▶ Semantic Social Networks Analysis

Social Trust

► Social Interaction Analysis for Team Collaboration

Social Virtual Objects

► Social Internet of Mobile Things and Decision Support Tools

Social Web Search

Maryam Shoaran[1], Alex Thomo[2] and Jens Weber[2]
[1]Department of Mechatronics, School of Engineering-Emerging Technologies, University of Tabriz, Tabriz, Iran
[2]Department of Computer Science, University of Victoria, Victoria, BC, Canada

Synonyms

Recommender systems; Social analysis; Social content search; Social navigation

Glossary

Social Media	Online systems with high public participation and interaction rate
User Metadata	Data created as the result of user interactions in an information space
Socially Enhanced Search	Quality-improved search resulting from employing user metadata
Personalization	Adjustment of a system or process to fit user preferences
Blogosphere	Collection of interconnected Web logs
Facebook Graph Search	Information lookup in the Facebook graph-structured data
Collaborative Filtering	Discovery of new knowledge and patterns through filtering data produced by collaboration between different individuals

Definition

Social search is an online search process that employs user-generated data and user-user relationships produced by social systems including bookmarking sites, Web forums, social networks, and blogs to discover the best matching content to user queries in an information space. This is different from the methods used in traditional Web search engines in the sense that search techniques in the latter are mostly based on page-author-generated data such as page content, anchor text, and link connections. User-created data forms a rich source of metadata that expresses single-user or community preferences, ideas, and needs. User tags and queries can be considered as new descriptions of Web page content. Social search utilizes this new and fast expanding source of information to establish a fine-grained and more personalized or community-based online search.

A variety of information systems ranging from the World Wide Web to special-purpose social systems, such as social networks, bookmarking sites, document-or media-sharing communities, and e-commerce, benefit from the capabilities of social search. In the literature, social search also refers to the process of the analysis and discovery of new knowledge from social media.

Introduction

The objective of an online search system is to locate the relevant objects (e.g., Web pages) to a user-generated query from the Web or a community-based collection. Over the decades, Web search engines have improved their quality of search by inventing new techniques to retrieve query-relevant documents and rank them based on their quality. The characteristic of almost all of these techniques is that they are based on the data created by the Web page builders or document authors. Two types of ranking methods are used in search engines: first, query-dependent or similarity measures that use document content, title, and anchor text to find similar documents and second, query-independent or static measures that use page connectivity (link structure) as a quality measure to rank similar documents. The prominent static metrics are PageRank (Page et al. 1999) and HITS (Kleinberg 1999).

Recently, with the ever-increasing activity and popularity of social media, a new type of information – user-created metadata – is available that can be used to enhance the quality of search.

User-generated content can be categorized as *explicit* or *implicit*. Explicit user data is created by visitors of Web sites in the form of annotations and viewpoints in order to describe, organize, and share their favorite entities (URLs, movies, songs, books, articles, etc.) online. Social systems capture explicit user annotations and viewpoints (feedback) in different forms. For example, book, article, and movie review sites collect user reviews and ratings as text and star points. Social bookmarking sites store user tags and favorite URLs, and social networks capture user comments and their likes. User annotations and viewpoints constitute a precious information source that can be utilized to extract, for example, Web page descriptions (using tags and comments), page or media popularities (using bookmarks and ratings), and user preferences (using ratings and likes).

Monitoring user online behavior builds another valuable source of information. Implicit user data is automatically extracted from system logs containing user search queries, browsing history (clickthrough data), and amount of time spent by users on different pages. This data can help improve the quality of search in different ways. For instance, user queries can be considered as “URL tags” describing the content of pages. User browsing history is an indication of user interest and can be used to resolve the ambiguity that often exists in user queries. The amount of time spent by users reading the content of Web sites might be an indication of the importance of sites and can be used to improve the ranking process, especially in community-based search environments.

Key Points

In social search, user-created data is the prominent resource of analysis systems. User metadata consists of a wide range of data either directly provided by users like user annotations and viewpoints on websites, or automatically extracted from user behaviors. New techniques employ user data to improve the quality of query systems or extract useful information from the web and social media.

Historical Background

For decades, Web search engines have used ranking methods like similarity-based measures and famous PageRank (Page et al. 1999) and HITS (Kleinberg 1999) metrics to rank similar documents.

Integrating user data (annotations) by introducing new metrics such as SocialSimRank (SSR) and SocialPageRank (SPR) has improved the quality of Web search (Bao et al. 2007; Yanbe et al. 2007).

Special navigational systems like Knowledge Sea II (Brusilovsky et al. 2004) have been built to assist users (students of a class) searching for special documents (course resources) based on the behavior of the past users. Recommender systems (Resnick and Varian 1997) improve user experience in online shopping, movie, and music Web sites by making appropriate recommendations for new items using sophisticated techniques such as *collaborative filtering* (Koren et al. 2009; Koren and Bell 2011) *Latent factor* (Koren et al. 2009; Koren and Bell 2011). Regarding the differences in the type of the data shared in and interconnections provided by social media sites such as social networks and the blogosphere, special-purpose search and information discovery systems have been created on the content of each site (Facebook graph search; Bansal and Koudas 2007; Mathioudakis and Koudas 2010; analysis of blog data; Gruhl et al. 2005; Bansal and Koudas 2007).

Social Web Search and Analysis

Online social systems holding a rich public participation have been able to accumulate valuable and heterogeneous collections of user metadata. Social search and analysis is focused on taking advantage of such data sources by new techniques that either help to improve the functionality of already existing systems or devise novel analysis and knowledge discovery schemes. Social search and analysis is active in different areas as follows: (1) socially enhanced Web search, (2) social navigation, (3) social analysis, (4) recommender systems, and (5) social content search.

Socially Enhanced Web Search

Social Web search aims at improving the quality of Web search by combining traditional search methods, e.g., query-document similarity and PageRank, with new techniques that employ social content. For instance, *SocialSimRank* (SSR) and *SocialPageRank* (SPR) (Bao et al. 2007) are two new methods that integrate social annotations available in social bookmarking sites, e.g., del.icio.us, into the page ranking process.

SocialSimRank (SSR) is a similarity ranking algorithm for queries and social annotations. The algorithm is based on the assumption that social annotations provide good summaries of Web pages from various user perspectives. Based on this observation, similarities for every pair of annotations and similarities for every pair of pages are iteratively computed. The similarities are recursively defined as follows. The more similar the pages are, the more similar their corresponding annotations are. Conversely, the more similar annotations are, the more similar their associated pages are. These similarities are integrated into each other's computation. That is, in the equation that calculates the similarity between two annotations, one of the parameters is the similarity between two pages to which these annotations are assigned, and vice versa. After several iterations, this process typically converges, and the system is ready to answer queries. Each query term is considered to be a page annotation. The similarity of a query q to a Web page p is computed as the sum of the similarities of each term in q to each annotation associated with p.

SocialPageRank (SPR) computes the page quality (popularity) with the intuition that the number of annotations assigned to a Web page indicates the quality of the page in some sense. SPR uses an iterative algorithm to compute page popularities based on user and annotation popularities. Integrating SSR and SPR into a ranking function that also uses traditional document similarity metric and PageRank improves the quality of Web search (Bao et al. 2007).

Other enhanced search methods are also proposed that benefit from different aspects of user annotations. A hybrid search technique is presented in Yanbe et al. (2007) that combines a link-based ranking method with a new metric that is based on user-generated data in social bookmarking sites (e.g., del.icio.us). The new metric utilizes SBRank (Social Bookmarking Rank) which is the number of user bookmarks on a page, the sentiment-based and temporal information extracted from user annotations, as well as general statistics derived from user interactions with Web pages.

Community-based search systems improve the quality of search by incorporating user search behaviors within the community, e.g., user queries and result selections, into the ranking method. The underlying intuition is that among the users of similar mind, e.g., social network or enterprise intranet users, the context of queries is similar, the query repetition is high, and also there rarely exist malicious behaviors that can negatively affect popularity metrics (Freyne et al. 2007). This type of social search is also called *collaborative Web search* (CWS) (Morris and Teevan 2009), and I-SPY (Smyth et al. 2004) is an implementation of it. Such systems record the queries and result selection of the community searchers, and upon exposure to a new query, the information of search sessions of a similar pattern is retrieved. The system re-ranks the result returned by the underlying search engines to reflect the implicit preferences of the community. Each item in the result list is also augmented by a set of past related queries that can be used to start new searches.

Social Navigation

The goal of social navigation is to enhance the quality of user browsing by providing various types of navigational assistance based on the visiting behavior of similar-minded users in the past. Social navigation systems benefit from different implicit and explicit user-generated data. They keep track of the browsing behavior of the users by collecting user queries and browsing paths (personal footprints). The time spent reading a page is also taken into account as an indication of user intention. Such systems also benefit from user annotations that can provide useful information about the importance of visited pages. When a user clicks on a source or on a (page) link in the search result list, the system provides a visual guide containing different navigational cues, for example, the source or page visit frequency (browsing popularity), the number of associated annotations (annotation popularity), and a list of queries leading to this source or page (search popularity).

Knowledge Sea II (Brusilovsky et al. 2004) is an example of a social browsing system that was developed to help students in a class to find the most useful sources for a particular course. This system organizes sources in a table with each cell associated with one source. The available navigational clues include the background color of cells indicating visit frequency, a sticky note for the presence of annotations, and a thermometer representing the number of positive annotations.

Another interesting system is the one presented in Freyne et al. (2007) that facilitates community-based access to the Communications of the ACM (CACM) magazine. This system integrates social search and social navigation in both the interface level and its internal mechanisms. When a new search query is initiated, the search component of the system retrieves similar queries and their associated search results. Then, the results are scored based on their relevance to the new query, and finally the top-*k* results are placed ahead of the other results returned by the ACM search engine. Each result item is appended by complementary information presented as icons. Five icons with different levels of filling indicate, respectively, (1) the relevance of the result to the query (the percentage of times the result has been selected for the query by community users), (2) a list of other queries that have led to the selection of this result by community users, (3) the last time the result was encountered by the users (a view of the freshness), (4) the browsing popularity of the result (footprints), and (5) the user annotations. When a result is selected, the browsing component also augments the opened pages with social assistance icons.

Social Quality Analysis

The quality of user-generated content in online social systems varies from excellent to spam due to the participation of individuals with different intentions and levels of expertise. This is especially important in knowledge-based social systems such as question/answering portals, online forums, and networks of email exchangers. Social quality analysis aims at identifying knowledge experts and high-quality user-created content in order to improve the quality of information-retrieval tasks (cf. Zhang et al. 2007; Campbell et al. 2003; Agichtein et al. 2008; Yang et al. 2011). Various analysis methods are used ranging from link-based ranking algorithms, e.g., PageRank (Page et al. 1999) and HITS (Kleinberg 1999), to text classification techniques and user clickthrough information.

Since 2006, some interesting systems have been presented that automatically evaluate the quality of questions and answers in question/answering domains (cf. Jeon et al. 2006; Agichtein et al. 2008). The framework presented in Agichtein et al. (2008) first identifies a collection of quality-indicating features of social media and associated interactions. Then, these features are used as input to a classifier (a stochastic gradient boosted tree), in order to extract high-quality content. A wide range of information sources are used to extract features of the following categories:

(1) Content based: textual features of questions and answers, such as word *n*-grams, punctuation and typos, syntactic and semantic complexity measures, and grammaticality measures
(2) Connectivity based: link-based metrics (authority scores and PageRank) in user-item and user-user relationship graphs, where an item is a query or an answer
(3) Usage based: temporal statistics, number of clicks on items, and time spent on reading

Recommender Systems

In online shopping, movie, and music Web sites, the goal is to improve the user experience by providing appropriate recommendations about new items that match user interests, ideas, and needs. These Web sites collect different types of user-produced data, ranging from explicit user ratings to implicit purchase history, browsing, and search activities. Recommender systems (Resnick and Varian 1997), using sophisticated algorithms, combine data from independent contributors to discover new knowledge about relations between users and items.

There are two major approaches in recommender systems: *content filtering* and *collaborative filtering* (Koren et al. 2009; Koren and Bell 2011). Content filtering discovers matching users and items based on their individual characteristics. Items (products) are profiled by domain experts and user profiles are created by users' explicit answers to specific questions, e.g., demographic questions. A problem with content-based filtering is the difficulty of gathering relevant information.

Collaborative filtering, on the other hand, is based on user behavior in the past, for example, user transactions and product ratings. By analyzing the relationships among users and among items, collaborative filtering predicts new relations between particular users and items. Suppose that in a movie rental site, user *u* has not watched and rated movie *x* yet, and the system would like to know whether it should recommend *x* to *u*. In the user-centered collaborative approach, first the similarity between *u* and all other users who have rated *x* is computed using some similarity measure, e.g., Euclidean distance or Pearson correlation coefficient. Then, the system predicts how *u* would rate *x* by computing a weighted average of the ratings for *x* by the most similar users to *u*. If the predicted rating is above a certain threshold, the system recommends *x* to *u*. In the item-centered approach the prediction is made based instead on the similarity between items.

Motivated by the Netflix prize contest, significant improvements have been made in the quality of recommender systems. *Latent factor* models are another approach in collaborative filtering that maps the users and items to a common multidimensional space based on the past rating patterns. Latent factor models are based on *sparse matrix factorization*, and they are among the most popular and the best performing approaches (Koren et al. 2009; Koren and Bell 2011).

Social Content Search

Despite the similarities between social media sites such as social networks, the blogosphere, and microblogging systems like Twitter, they differ in the type of data predominantly posted and shared by users, as well as in the form of user interconnections they offer. Based on these characteristics, special-purpose search and information discovery efforts are applicable on the content of each site (Facebook graph search; Bansal and Koudas 2007; Mathioudakis and Koudas 2010).

Facebook has recently launched *Graph Search* (Facebook Graph Search) as a new feature to benefit from its massive storage of data and relationships. Using this tool people can search for real-world objects in Facebook's knowledge graph, which is comprised of objects such as people, places, and things and inter-object connections, for example, Friendship and Likes. An important advantage of Facebook's search is that it has access to the collective knowledge of its vast community of users (more than one billion) to answer questions involving different layers of searching. Appealing examples are as follows: "What to read that is liked by my friends in college," "Where to eat in Toronto that my friends living there like," "Where to go in Asia that my friends and friends of friends of my age found interesting," "What iPhone app to download that my friends use to track their jogging and cycling," etc. People's experience with Facebook Graph Search will highly depend on the level of their connectedness and participation in the system.

Blogging is another online social activity that has received an increasing popularity in recent years. The free context of blogs makes the blogosphere (the collection of connected blogs) a rich source of heterogeneous information including personal experiences and opinions about a variety of subjects. Mining and analysis of blog data can capture public insight in different topics (Gruhl et al. 2005; Bansal and Koudas 2007). For instance, BlogScope (Bansal and Koudas 2007) is one of the systems designed to analyze the textual content of blogs and to provide information such as *when, where*, and *why* about interesting topics. When the user selects one of the daily hot keywords provided by the system or poses a query, all the relevant blog posts are retrieved and the result of various analysis on their content is presented. For example, the system can display the following information: (1) a popularity curve for a keyword as a function of time, (2) a list of the most closely related keywords in blog posts, (3) a distribution of the related posts on the map, and (4) a synopsis set which is the maximal set of keywords correlated with query that exhibits a bursty behavior in the associated popularity curve.

Key Applications

Social search and analysis techniques are used in different areas such as *search engines* to enhance the quality of search, *special-purpose browsing systems* to assist users to navigate through related Web sites, and *recommender systems* to improve user experience by recommending appropriate new items.

Future Directions

Whereas the usefulness of the annotations in small-scaled information communities has been demonstrated by several works, social annotations and bookmarks lack yet the sufficient size and quality to significantly influence the performance of search engines in a large scale (cf. Heymann et al. 2008; Bao et al. 2007). For example, the number of unique URLs in del.icio.us is relatively small in comparison with the indexes of the major search engines that include hundreds of billions of pages.

Augmenting more Web sites with improved tagging systems and also aggregating the data from various bookmarking and tagging Web sites would significantly help to improve the situation. The design of appealing and structured user interfaces that could provide a list of possible tags that do not appear in the page content and title can enhance the quality of tagging. Providing incentives, such as site access privileges or special offers to stimulate user tagging activities, could be another approach to create enriched user metadata.

Cross-References

- ▶ Microtext Processing
- ▶ Recommender Systems: Models and Techniques
- ▶ Social Bookmarking or Tagging

References

Agichtein E, Castillo C, Donato D, Gionis A, Mishne G (2008) Finding high-quality content in social media. In: WSDM, Stanford, pp 183–194

Bansal N, Koudas N (2007) Blogscope: spatio-temporal analysis of the blogosphere. In: WWW, Banff, pp 1269–1270

Bao S, Xue G-R, Wu X, Yu Y, Fei B, Su Z (2007) Optimizing web search using social annotations. In: WWW, Banff, pp 501–510

Brusilovsky P, Chavan G, Farzan R (2004) Social adaptive navigation support for open corpus electronic textbooks. In: AH, Eindhoven, pp 24–33

Campbell CS, Maglio PP, Cozzi A, Dom B (2003) Expertise identification using email communications. In: CIKM, New Orleans, pp 528–531

del.icio.us. https://delicious.com/

Facebook graph search. https://www.facebook.com/about/graphsearch

Freyne J, Farzan R, Brusilovsky P, Smyth B, Coyle M (2007) Collecting community wisdom: integrating social search & social navigation. In: IUI, Honolulu, pp 52–61

Gruhl D, Guha RV, Kumar R, Novak J, Tomkins A (2005) The predictive power of online chatter. In: KDD, Chicago, pp 78–87

Heymann P, Koutrika G, Garcia-Molina H (2008) Can social bookmarking improve web search? In: WSDM, Stanford, pp 195–206

Jeon J, Croft WB, Lee JH, Park S (2006) A framework to predict the quality of answers with non-textual features. In: SIGIR, Seattle, pp 228–235

Kleinberg JM (1999) Hubs, authorities, and communities. ACM Comput Surv 31(4es):5

Koren Y, Bell RM (2011) Advances in collaborative filtering. In: Ricci F et al (eds) Recommender systems handbook. Springer, London/New York, pp 145–186

Koren Y, Bell RM, Volinsky C (2009) Matrix factorization techniques for recommender systems. IEEE Comput 42 (8):30–37

Mathioudakis M, Koudas N (2010) Twittermonitor: trend detection over the twitter stream. In: SIGMOD Conference, Indianapolis, pp 1155–1158

Morris MR, Teevan J (2009) Collaborative web search: who, what, where, when, and why. Synthesis lectures on information concepts, retrieval, and services, Morgan & Claypool, San Rafael

Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking: bringing order to the web. Technical report 1999–66, Stanford InfoLab, Nov 1999. Previous number = SIDL-WP-1999-0120

Resnick P, Varian HR (1997) Recommender systems – introduction to the special section. Commun ACM 40 (3):56–58

Smyth B, Balfe E, Freyne J, Briggs P, Coyle M, Boydell O (2004) Exploiting query repetition and regularity in an adaptive community-based web search engine. User Model User-Adapt Inter 14(5):383–423

Yanbe Y, Jatowt A, Nakamura S, Tanaka K (2007) Can social bookmarking enhance search in the web? In: JCDL, Vancouver, pp 107–116

Yang L, Bao S, Lin Q, Wu X, Han D, Su Z, Yu Y (2011) Analyzing and predicting not-answered questions in community-based question answering services. In: AAAI, San Francisco

Zhang J, Ackerman MS, Adamic LA (2007) Expertise networks in online communities: structure and algorithms. In: WWW, Banff, pp 221–230

Social Weight

▶ User Behavior in Online Social Networks: Influencing Factors

Social Yield

▶ Link Dynamics and Community Formation in Social Networks

Social-Based Collaborative Filtering

Kostas Stefanidis[1], Eirini Ntoutsi[2,5], Haridimos Kondylakis[2,3] and Yannis Velegrakis[4]
[1]School of Information Sciences, University of Tampere, Tampere, Finland
[2]Department of Electrical Engineering and Computer Science, Leibniz University Hannover, Hannover, Germany
[3]Foundation for Research and Technology Hellas, Institute of Computer Science (ICS), Hellas, Greece
[4]Department of Information Engineering and Computer Science, University of Trento, Trento, Italy
[5]Ludwig Maximilian Univ. Munich, Munich, Germany

Synonyms

Collaborative filtering using social data; Social-Based Recommendations

Glossary

Collaborative filtering	Given a ratings matrix R, representing the preferences of users U for items I, recommend to each user a list of items in descending order of their relevance for

the user. The relevance scores are estimated based on ratings of similar users

Ratings matrix — Assume a set of users U and a set of items I in the recommender system. A user $u \in U$ might provide her preference for an item $i \in I$ in form of a rating denoted by *rating* (*u, i*), which typically takes values in (Adomavicius et al. 2011; Blei et al. 2003). The preferences of users for individual items are represented by a ratings matrix R, where the $R_{u,i}$ entry corresponds to *rating* (*u, i*)

Recommendation — A suggestion or proposal to a user for an item, e.g., book, movie, video, news article, that is potentially interesting for the user

Recommender system — A system or engine that produces recommendations by predicting the preferences of users for certain items

Social content — The content that is generated in a social network as a result of activities and interactions between users

Social network — A structure, nowadays typically implemented as an online platform, which enables content sharing by allowing users to post information and interactions between users by allowing them to comment on each other's posts, exchange messages, etc

Definition

Collaborative filtering is a special category of recommender systems that generate recommendations to the users about certain items, e.g., products or services, by relying on preferences of similar users. Social-based collaborative filtering exploits the vast amount of data generated in the social networks to improve the quality of recommendations. Such data are utilized for different purposes, e.g., to deal with traditional recommendation problems, such as the cold start problem, which is caused due to lack of user-related data in the recommender, to derive a better-quality user neighborhood by integrating user-related information from the social networks, and to enrich the recommendations by capturing different aspects of the users as they are manifested through the content they share and their interactions in the social networks.

Introduction

With the growing complexity of the Web, users find themselves overwhelmed by the mass of choices available. To facilitate the user selection process, recommender systems provide suggestions on data items of potential interest to the users. The interest of a user for an item is inferred from the user history, e.g., user purchases or browsing history. A key challenge in recommenders is the so-called sparsity data problem; typically, users rate only a few items, due to the huge amount of offered items and the low engagement of the users with the applications.

Nowadays, stunning opportunities are offered for dealing with the lack of data in recommender systems, as users publicly share their preferences and provide reviews and other information on certain items in several social networks. This data though is not explicit for a recommender; rather, it is implicitly given in the context of the social networks. As such, except for its volume and velocity, the information is blur, unstructured, diverse, uncertain, and incomplete, and, therefore, exploiting this data for recommendations is a big challenge.

This entry focuses on how the collaborative filtering recommender systems are reshaped by looking out-of-the-recommender box and hunting for relevant information in the social networks.

This way, from the traditional collaborative filtering approaches that utilize within-the-recommender data, i.e., the ratings matrix mainly, we are moving toward an open system that would potentially result in high-quality recommendations by enriching information for users, items, and ratings through the available social network data. We illustrate such a scenario in the following example.

Example 1 Assume a user, say Sophia, and a recommender system, say MinoS, that produces recommendations for travel destinations by using a collaborative filtering approach that exploits Sophia's social data. For doing so, summarily, MinoS locates Sophia and the set of destinations she has visited in the social networks she used. Then, it focuses on finding users, other than Sophia, and the places they have visited. Having collected the visiting history of the users, MinoS computes the similarity between Sophia and every other user; the users' visiting history contains the necessary information for computing similarities between them. MinoS keeps only the users with similarity greater than a threshold value to Sophia, namely, the highly similar users to Sophia. Intuitively, at the next step, it performs a composition between Sophia and her highly similar network users and users and the places they have visited. This way, for each of Sophia's similar user, who has visited a destination, we make a connection between Sophia and that destination. In overall, the relevance score of each such destination is calculated by taking into account the ratings and reviews of the users that have visited the destination, as well as their similarity with Sophia. On the basis of these scores, MinoS recommends the best destinations to Sophia.

Key Points

Recommender systems have become indispensable for several Web sites, such as Amazon, Netflix, and Google News, helping users to navigate through the infinite number of available choices, like products, movies, and news articles, respectively. Recently, social networks offer opportunities for generating better recommendations, as more than ever before, users publicly share their preferences and provide information for certain items. In this entry, focusing on social-based collaborative filtering, we target the following issues:

- How can we enrich the three main entities in a recommender system, namely, users, items, and ratings, by integrating content- and structure-related data from social networks?
- How can we construct social-enhanced user profiles, by exploiting the integrated data?
- How is the collaborative filtering engine reshaped due to the volume and the variety of the available data?

Historical Background

Collaborative filtering predicts user preferences for data items by keeping track of their likes and dislikes. Based on the assumption that similar items will be of interest for similar users, i.e., users with tastes in common, via collaborative filtering, users can help each other to choose products.

More specifically, assume a recommender system where I is the set of items to be rated and U is the set of users in the system. A user $u \in U$ might rate an item $i \in I$ with a score *rating* (u, i) in Adomavicius et al. (2011) and Blei et al. (2003). Typically, the cardinality of the item set I is high and users rate only a few items. For the items unrated by the users, recommender systems estimate a relevance score, denoted as *relevance* (u, i), $u \in U, i \in I$. In general, there are different ways to estimate the relevance score of an item for a user. In the content-based approach (e.g., Mooney and Roy 2000), the estimation of the rating of an item is based on the ratings that the user has assigned to similar items, whereas in collaborative filtering systems (e.g., Konstan et al. 1997), this rating is predicted using previous ratings of the item by similar users. Typically, similar users are located via a *similarity function* simU (u, u') that evaluates the proximity between $u, u' \in U$ by considering their shared dimensions. We use N_u to denote the set of the most similar users to u, hereafter, referred to as the *neighbors* of u.

Given a user u and his neighbors N_u, if u has expressed no preference for an item i, the relevance of i for u is estimated as:

$$\text{relevance}_{\,N_u(u,i)} = \frac{\sum_{u' \in N_u} \text{simU}(u,u')\text{rating}(u',i)}{\sum_{u' \in N_u} \text{simU}(u,u')}$$

Typically, after estimating the relevance scores of all unrated items by a user, the top-k rated items are recommended to the user.

Using Social Content for Collaborative Filtering

Most recommender systems nowadays are based, in practice, on large datasets. As a result, the rating matrix used for collaborative filtering could be extremely large and sparse, which affects the quality of the produced recommendations. Social-based collaborative filtering exploits data that users publicly share in social networks in order to deal with the sparsity problem. This way, from the traditional collaborative filtering approaches that utilize mainly the ratings matrix, we are moving to an open system that would potentially result in high-quality recommendations by enriching information for users, items, and ratings through the available social data, appearing out-of-the-recommender box.

Data Enrichment and Integration

The main entities involved in a recommendation application, i.e., users, items, and ratings, can be semantically enhanced with information from social networks.

With respect to the users, instead of using the plain information that a user gives for himself to a recommender system, we can exploit information available in numerous external sources, such as Facebook, LinkedIn, Google+, Foursquare, and Amazon. The motivation behind this is that a user describes himself differently in different networks, depending on the domain, so we can identify different interests, user activities, information about places he visited, and so forth. A challenge toward this direction is to integrate the user's social profile, as well as to integrate, or expand, the social graph to bring together different social networks. Approaches like Tang et al. (2010) try to construct a complete user profile by finding, extracting, and fusing the semantic-based user profile from the Web, whereas approaches like Amer-Yahia et al. (2009) collect, integrate, and discover profile information based on social content sites. In Wischenbart et al. (2012), the schema is first extracted out of each social site, transformed to a common representation, and instantiated using possibly overlapping data. Another popular technique for collecting and understanding user data is the analysis and creation of topic models within the user-generated content (Pennacchiotti and Gurumurthy 2011), especially Twitter data (Zhao et al. 2011). Such solutions can be used for encountering the cold start problem as well.

At item level, one can consider that information about items can be enhanced with semantic information. In addition to the item descriptions, in which temporal characteristics of the items, such as popularity and freshness (Stefanidis et al. 2013), can be maintained in real time, we can exploit information retrieved from the Web, such as published results and reports, Web pages, thesauri, or ontologies. The plethora of well-organized information over the Web in collectively maintained knowledge repositories, such as Wikipedia and LibraryThing, can be used for correlating and computing similarities between data items (Christophides et al. 2015). In addition, items can be annotated using terms from ontologies and other semantic resources to enhance the description quality of specific items (Kondylakis et al. 2015).

Regarding user preferences for certain items, we can rely on users' online activities that involve such items. For instance, a user may tweet about a movie, might rate its trailer in YouTube and share the link, and comment in Facebook. Except for filling missing ratings, existing ratings could be also enhanced in terms of their context (e.g., time or place) and rating criteria (e.g., director vs actors). *Contextual recommendations* have been already studied (Adomavicius et al. 2005; Stefanidis et al. 2012); however, they rely on user explicit feedback. An indirect inference of the context though is an interest area for further exploration. *Multi-criteria ratings* (Adomavicius et al. 2011) rely also on user explicit subratings for different aspects of the items and are suffering from lack of data as the item space is further expanded due to the subcriteria. With the abundance of free text reviews nowadays, we can implicitly extract both the aspects of the items that are of interest to a user and their associated ratings/sentiment, using NLP and sentiment analysis techniques (Zimmermann et al. 2015). In an *active learning* manner (Fu et al. 2013), we can explicitly ask users to provide more ratings, by giving them incentives. The challenge here is the selection of a few informative items for which the user should be asked.

Even if we are able to identify additional information for users, items, and preferences, appearing outside the recommender system, it is important to understand which pieces of information refer to the same entities, so as to integrate them, in order to manage and further process them. The problem of entity resolution aims to identify different descriptions that refer to the same entity and emerges as a central data-processing task for an entity-centric organization of Web data (Christophides et al. 2015). It is needed to enrich interlinking of entity descriptions, so that Web data can be accessed by machines as a global data space allowing the use of standard languages. Although entity resolution has attracted significant attention in information systems, database, and machine-learning communities, there are new challenges stemming from the Web openness in describing a multitude of entity types across domains. The scale and diversity of descriptions challenge the core entity resolution tasks, namely, (i) how descriptions can be effectively compared for similarity and (ii) how resolution algorithms can efficiently filter the candidate pairs of descriptions that need to be compared.

User Profiling

A fundamental ingredient of every successful recommendation is the ability to accurately model the user preferences and habits. This model is typically known as *profile*. A profile is actually a structure that represents the principal characteristics of a user which are turned into preferences.

One way of building profiles is to have them explicitly provided by the end user, but this brings along a number of limitations. First of all, users may not be aware of all the characteristics they have or may not be willing to provide all of them. Thus, it is better if the profile is built automatically by observing the user actions, which is exactly what traditional recommendation techniques are doing, i.e., they are monitoring the previous user interactions with the system, and based on them, they are building a user model. Unfortunately, profiles generated in this way may end up highly dependent, restricted, and very sensitive to the user interactions with the system. This means that unless a user exposes one of her characteristics with some specific action, that characteristic can never become known to the system. Furthermore, an action performed by a user, e.g., a purchase, may be recorded and leave its trace in the user preferences for a very long time even if it has been only an occasional action or an action performed for someone else, for instance, the purchase of a gift for a friend. All these mean that the user interaction history with the system, although a valuable resource, may not provide the complete information that a system would like to know about its user.

Social media constitute a great additional source of information for building user profiles since they may expose user characteristics that are not recorded in the user interaction history with the system. The scale and spectrum of activities for which the social media are used nowadays are unprecedented. Through these activities the users leave their footprint that can be used to build a rich user description, i.e., an accurate and complete profile. There are two types of information that can be exploited in social media. One is the content and the other is the social network. Exploitation of the content means that the profiling of the users is based on the content that the users publish in the social media, e.g., the tweets they send or the posts they make. Exploitation of the social network means that the profiling is constructed by using information on the way users are connected to each other to establish friendships and communications or to follow the activity of others. Based on the above, one can distinguish two main profiling methods, the *content* based and the *network* based.

Generating profiles from social media is a challenging. The content is user generated and is done through some new nontraditional forms of communication. As a consequence, it has no controlled vocabulary, no restrictive syntax, and no specific rules and is full of shorthands and jargon. For what concerns the social network information, among the challenging issues is the fact that not all the connections are of equal value. Users may follow actively only a selective set of their connections or connections may exist with central non-real users like news channels or group accounts. Finally, users may use different accounts for different purposes which make hard the recognition of the actions with user provenience.

The idea behind the content-based profiling is to see the content as a vector of terms. These terms can play the role of features for which a classifier can be trained or some discriminating score can be computed (Hecht et al. 2011). Of course, not all the words are equally important. Techniques like LDA (Blei et al. 2003) can be used to summarize the content or to identify those words that actually matter and then use only these words (Cheng et al. 2010) for the profiling task. Alternatively, an information theoretic approach can be employed to identify the amount of information that each word communicates in relationship with the rest of the content and use that information of deciding the importance of each word (Han et al. 2012). An important characteristic of social data affecting significantly the results is sparsity. To overcome sparsity, smoothing has to often be applied, with Laplace smoothing (Herrmann 1976) being the most prevalent technique.

The intuition behind the network-based profiling, on the other hand, is similar to that of the collaborative filtering. A user is likely to like things that are similar to what her friends like. However, instead of the collaborative filtering idea in which similarity is used to propagate features of one user directly to another, in the network-based profiling, the social network is the main carrier. For instance, considering a user as a node in the network graph, a simple counting and aggregation of the features of the reachable nodes are required. The features that end up to be more popular are used to form the user profile (Davis et al. 2011). Instead of a simple counting and aggregation, more complex probabilistic models can be used that take into consideration the distance on the social graph to adjust the weight of each feature appearing on the connected nodes (Backstrom et al. 2010; Sadilek et al. 2012).

Unfortunately, neither the content-based nor the network-based profiling techniques work perfectly all the time. Naturally, there have been techniques that combine them to improve the quality of the results. One approach is using a single model for both the network and the content (Li et al. 2012). However, these two parameters can be also considered independently and their results be combined (Li et al. 2014) using some mathematical model like those used to combine different lists (Fagin 2002) or considered orthogonal factors studied together using some bi-clustering technique (Prelic et al. 2006).

Social-Based Collaborative Filtering

Traditionally recommenders rely upon the rating matrix, i.e., the explicitly given (numerical) ratings of users to certain items. In collaborating filtering, for a query user $u \in U$ her similar users in U are located using the rating matrix and some appropriate similarity function, like cosine similarity, Pearson correlation, or Spearman rank correlation (Desrosiers and Karypis 2011). Recently though except for the rating matrix, other sources of information, such us the network information and reviews accompanying the user ratings, are employed in order to improve the quality of recommendations.

In particular, the network information, i.e., explicit connections between the users like friends in Facebook, following/followers in Twitter, etc., can be employed in order to select a better user neighborhood for a given user and also for dealing with the cold start problem. For example, Jamali and Ester (2009) replace the traditional notion of user neighborhood, consisting of users with similar ratings to the query user, by that of the trust neighborhood derived from the trust network where the nodes correspond to the users and the edges to trust statements.

In a different direction, textual reviews which typically accompany user ratings nowadays have been recently explored for recommendations. User reviews comprise a rich source of information as they justify user decisions on certain items, and moreover they reveal which aspects of the items the users liked/disliked. Combining numerical ratings with textual reviews for recommendations was first introduced in McAuley and Leskovec (2013a). The authors propose the hidden factor model which aligns hidden factors in product ratings with hidden factors in product reviews (discovered through LDA). The key idea is to link latent factors in ratings to hidden factors in review texts, where topics discussed in review texts for a certain product correspond to products having a certain property represented in the latent factor model. In a follow-up work, the same authors also used reviews to model personal evolution or experience for recommendations (McAuley and Leskovec 2013b).

Textual reviews have been also employed for extracting context, such as location and accompanying people, which allows for contextual recommendations. Explicit user-defined context is hard to acquire, but usually such information is contained in the reviews, which are typically freely available. For example, Chen and Chen (2014) implicitly extract such sort of information, by employing online reviews.

Employing heterogeneous data for recommendations, such as the network and the reviews, definitely alters the recommendation process. In a different direction, the process is also altered due to the long-term tracking of users, items, and their preferences, which call for online methods that are able to identify drifts in user preferences and periodicity in their habits. Data aging is a typical way to deal with drifts in user profiles by downgrading historical obsolete data and paying more attention to recent ones that reflect the current user profile best (e.g., Ding and Li 2005; Stefanidis et al. 2013). However, approaches that discard past instances have been criticized as loosing too much signal, and although more elaborate methods exist, e.g., Koren (2009), which separate transient factors from lasting ones, what to forget and what to remember are still a challenge. Moreover, a long-term user monitoring implies an extensive knowledge about user tastes and preferences, which might result in privacy risks for the user. This is especially critical for mobile app recommenders.

Key Applications

Traditionally, recommendations are produced within a domain, i.e., when asking for movies, the suggestions consist only of movies. Examples of domain-specific social network recommenders include Flixster (https://video.flixster.com/) for movie recommendations and Epinions (http://www.epinions.com/) for a wide range of product recommendations. This paradigm can be extended so as to support cross domain recommendations. For example, packet recommendations produce composite items consisting of a central item, possibly in the main domain of interest for a user, and a set of satellite items from different domains compatible though with the central item. Compatibility can be assumed either as soft (e.g., other books that are often purchased together with the movie being browsed) or hard (e.g., a travel destination that must be within a certain distance from the main destination). The notion of cross domain recommendations can be extended, so as to support data items outside of the data repository of the recommender. There are already such sort of aggregators in the Web, which act as wrappers over items from different stores. For example, users in Polyvore (http://www.polyvore.com/) mix and match fashion items from different brands.

Well-established applications of social-based collaborative filtering appear in the domain of social media. Unlike traditional media in which few editors set the guidelines, in the era of social media, we may have a very big number of editors, and content data improves its quality as the number of contributors increases. Typical examples include YouTube (https://www.youtube.com/), Last.fm (http://www.last.fm/), and Reddit (https://www.reddit.com/).

In a different scenario, given a social community, a collaborative filtering application is to suggest compelling data items as judged by the community. For example, consider the news aggregator Digg (http://digg.com/) that in its front page shows stories as they are rated positively by the community. Larger and more diverse communities offer stories that better reflect the average interest of the community participants.

The well-used social networks offer recommendations as well. Either general purpose social networks, e.g., Facebook (https://www.facebook.com/) and Twitter (https://twitter.com/), or domain-specific ones, e.g., Linkedin (https://www.linkedin. com/), offer suggestions about friends, people to follow, and jobs you may be interested in.

Finally, from a different point of view, there are social networks that provide recommendations by exploiting review text to uncover user's implicit tastes and item's properties. The reviews comprise a rich source of information as they justify a user's decision on a certain choice. Such piece of information is used, for example, by Booking.com (http://www.booking.com/), for making hotel suggestions.

Future Directions

We consider that the next day of recommenders is to put the users in the foreground and try to exploit their social interactions to fulfil their needs, as opposed to approaches focusing more on the companies' viewpoints. Next, we highlight services toward this direction.

Interactive exploration: New forms of data exploration and interaction become increasingly more attractive to aid users navigate through the information space and overcome the challenges of information overload. The interaction between users and recommenders can be driven directly by the interpretation of users' needs. Users have to peruse the suggested results, and systems have to be able to react to the on-the-fly changes in the users' demands. Although long challenged by works, such as the berry picking model, common systems still assume that the user has static needs, which remain unchanged during the seeking process.

Visualization: Techniques for visualization contribute toward helping users perceive an overview of the data items included in the suggestions produced for them. Explanations can be used as a means for visualization to assist users identify the what, where, when, how, and who of a data item. That is, explanations target at telling the story that the data has to say, aiming at minimizing the browsing effort of the users.

Seeking your past: As data and knowledge bases get larger and accessible to a more diverse and less technically oriented audience, new forms of data seeking become increasingly more attractive. Refinding is a different form of exploring data, aiming to locate suggestions seen in the past; here we face the task of recovery. Explicit (given by a user) or implicit (extracted, for instance, by his online traces, e.g., via Foursquare) feedback on suggestions, content, and users can significantly increase the quality of recommendations and searching features of a system.

Guessing the future: Modern systems use the past as a mean to guess where the user aims at, so that the system can make the suggestions that will drive toward the fulfillment of the goal(s) as fast as possible. Existing techniques are based mainly on agents and libraries. There is a great deal of opportunities in adapting these techniques into a recommendation model that adjusts dynamically the suggestions as the user provides more feedback and the goals become more clear.

Cross-References

- ▶ Recommender Systems Based on Linked Open Data
- ▶ Recommender Systems Based on Social Networks
- ▶ Recommender Systems: Models and Techniques
- ▶ Social Search and Querying

References

Adomavicius G, Sankaranarayanan R, Sen S, Tuzhilin A (2005) Incorporating contextual information in recommender systems using a multidimensional approach. TIS, 23(1):103–145

Adomavicius G, Manouselis N, Kwon Y (2011) Multicriteria recommender systems. In: Recommender systems handbook. Springer US, Boston, MA, pp 769–803

Amer-Yahia S, Lakshmanan LVS, Yu C (2009) Socialscope: enabling information discovery on social content sites. In: CIDR, Asilomar, CA, USA

Backstrom L, Sun E, Marlow C (2010) Find me if you can: improving geographical prediction with social and spatial proximity. In: Proceedings of the 19th international conference on world wide web, WWW 2010, Raleigh, 26–30 Apr 2010. pp 61–70

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3(4–5):9931022

Chen G, Chen L (2014) Recommendation based on contextual opinions. In: UMAP, Aalborg, Denmark

Cheng Z, Caverlee J, Lee K (2010) You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM conference on information and knowledge management, CIKM 2010, Toronto, 26–30 Oct 2010, pp 759–768

Christophides V, Efthymiou V, Stefanidis K (2015) Entity resolution in the web of data. Synthesis lectures on the semantic web: theory and technology. Morgan & Claypool Publishers, California, USA

Davis CA Jr, Pappa GL, de Oliveira DRR, de Lima Arcanjo F (2011) Inferring the location of twitter messages based on user relationships. Trans GIS 15(6):735–751

Desrosiers C, Karypis G (2011) A comprehensive survey of neighborhood-based recommendation methods. In: Recommender systems handbook. Springer US, Boston, MA, pp 107–144

Ding Y, Li X (2005) Time weight collaborative filtering. In: CIKM, ACM, New York, NY, USA

Fagin R (2002) Combining fuzzy information: an overview. SIGMOD Rec 31(2):109–118

Fu Y, Zhu X, Li B (2013) A survey on instance selection for active learning. Knowl Inf Syst 35(2):249–283

Han B, Cook P, Baldwin T (2012) Geolocation prediction in social media data by finding location indicative words. In: COLING 2012, 24th international conference on computational linguistics, proceedings of the conference: technical papers, 8–15 Dec 2012, Mumbai, pp 1045–1062

Hecht B, Hong L, Suh B, Chi EH (2011) Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In: Proceedings of the international conference on human factors in computing systems, CHI 2011, Vancouver, 7–12 May 2011, pp 237–246

Herrmann LR (1976) Laplacian-isoparametric grid generation scheme. J Eng Mech Div 5(102):749756

Jamali M, Ester M (2009) Trustwalker: a random walk model for combining trust-based and item-based recommendation. In: SIGKDD, KDD. ACM, Paris, France, pp 397–406

Kondylakis H, Koumakis L, Kazantzaki E, Chatzimina M, Psaraki M, Marias K, Tsiknakis M (2015) Patient empowerment through personal medical recommendations. In: MEDINFO, Sao Paulo, Brazil

Konstan JA, Miller BN, Maltz D, Herlocker JL, Gordon LR, Riedl J (1997) Grouplens: applying collaborative filtering to use net news. Commun ACM 40(3):77–87

Koren Y (2009) Collaborative filtering with temporal dynamics. In: SIGKDD, ACM, New York, NY, USA

Li R, Wang S, Deng H, Wang R, Chang KC-C (2012) Towards social user profiling: unified and discriminative influence model for inferring home locations. In: The 18th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '12, Beijing, 12–16 Aug 2012, pp 1023–1031

Li R, Wang C, Chang KC (2014) User profiling in an ego network: co-profiling attributes and relationships. In: 23rd international world wide web conference, WWW '14, Seoul, 7–11 Apr 2014, pp 819–830

McAuley J, Leskovec J (2013a) Hidden factors and hidden topics: understanding rating dimensions with review text. In: RecSys, ACM, New York, NY, USA

McAuley JJ, Leskovec J (2013b) From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In: Proceedings of the 22nd international conference on world wide web. ACM, New York, NY, USA, pp 897–908

Mooney RJ, Roy L (2000) Content-based book recommending using learning for text categorization. In: ACM DL, ACM, New York, NY, USA

Pennacchiotti M, Gurumurthy S (2011) Investigating topic models for social media user recommendation. In: WWW, ACM, New York, NY, USA, pp 101–102

Prelic A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics 22 (9):1122–1129

Sadilek A, Kautz HA, Bigham JP (2012) Finding your friends and following them to where you are. In: Proceedings of the fifth international conference on web search and web data mining, WSDM 2012, Seattle, 8–12 Feb 2012, pp 723–732

Stefanidis K, Shabib N, Nørvåg K, Krogstie J (2012) Contextual recommendations for groups. In: Advances in conceptual modeling ER 2012 workshops, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 89–97

Stefanidis K, Ntoutsi E, Petropoulos M, Nørvåg K, Kriegel H (2013) A framework for modeling, computing and presenting time-aware recommendations. Large Scale Data Know Centered Syst 10:146–172

Tang J, Yao L, Zhang D, Zhang J (2010) A combination approach to web user profiling. TKDD 5(1):2

Wischenbart M, Mitsch S, Kapsammer E, Kusel A, Pröll B, Retschitzegger W, Schwinger W, Schönböck J, Wimmer M, Lechner S (2012) User profile integration made easy: model-driven extraction and transformation of social network schemas. In: WWW, ACM, New York, NY, USA

Zhao WX, Jiang J, Weng J, He J, Lim E, Yan H, Li X (2011) Comparing twitter and traditional media using topic models. In: ECIR, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 338–349

Zimmermann M, Ntoutsi E, Spiliopoulou M (2015) Discovering and monitoring product features and the opinions on them with OPINSTREAM. Neurocomputing 150:318–330

Recommended Reading

Shi Y, Larson M, Hanjalic A (2014) Collaborative filtering beyond the user-item matrix: a survey of the state of the art and future challenges. ACM Comput Surv 47 (1):3:1–3:45

Stefanidis K, Ntoutsi E, Kondylakis H Information hunting: the many faces of recommendations for data exploration. ACM SIGMOD Blog. http://wp.sigmod.org/?p=1580

Social-Based Recommendations

► Social-Based Collaborative Filtering

Socialbots

Abigail Paradise[1], Rami Puzis[2] and Asaf Shabtai[1]
[1]Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel
[2]University of Maryland Institute for Advanced Computer Studies (UMIACS), College Park, MD, USA

Synonyms

Fake profiles; Infiltration; Social network security

Glossary

Advanced persistent threat (APT)	A class of sophisticated cyber-attacks that target organizations
Infiltration	A means of compromising the social network graph by connecting with a large number of users; socialbots can be executed to infiltrate social networks
Influence bots	A bot that tries to influence conversation on a specific topic
Socialbot	An artificial, machine-operated profile in a social network that mimics human users, looks genuine, and behaves in a sophisticated manner
Spambot	A computer program designed to help send spam
Sybil attack	A type of attack in which a malicious user creates multiple fake identities (Sybils) in order to unfairly increase power and influence within a target community

Definition

In recent years, online social networks (OSNs) are becoming an essential part of our lives. However, OSNs have also been abuses by cyber criminals that exploit the platform for malicious purposes including spam and malware distribution, harvesting personal information, infiltration of organizations, and spreading rumors.

Socialbots are artificial, machine-operated profiles controlled by malicious users that are using deceptive techniques to act and look like human accounts. This allows socialbots to avoid detection and cause prolonged harm.

It is important to understand the security risks posed by socialbots in order to design effective mechanisms to identify and detect them.

Introduction

Online social networks (OSNs) are a popular, important, and powerful tool. OSNs play a central role in modern social life, as a helpful means of communication, sharing opinions and information, finding and disseminating information (Kwak et al. 2010), expanding connections (Gilbert and Karahalios 2009), and promoting businesses.

OSNs also have become the target of exploitation and abuse and have begun to attract unethical and illegal activities. Spreading spam, rumors, and malicious content (Burghouwt et al. 2011), invasion of user privacy (Jurgens 2013), political astroturfing (Ratkiewicz et al. 2011), and attaining an influential position and spreading misinformation or propaganda (Ferrara et al. 2014) are examples of today's misuse of OSNs. Between 8% and 10% of all social media profiles are malicious in nature (Ahmad 2015). This enormous number emphasizes the acute problem we are facing and demonstrates the need for new solutions (Abokhodair et al. 2015), particularly since OSN providers have repeatedly failed to mitigate such abuse and the threats they pose.

Attackers utilize artificial, machine-operated, OSN profiles called socialbots in order to execute their attacks. Unlike a regular bot (Boshmaf et al. 2012; Ferrara et al. 2014; Ji et al. 2016), a socialbot mimics human users by simulating the actions of a real OSN user. Socialbots have the ability to conduct social activities online, such as posting a message or sending a friendship request (Boshmaf et al. 2012).

These days, socialbots have become very sophisticated, and therefore their detection has become more difficult. Socialbots can be executed to infiltrate an OSN (Boshmaf et al. 2013). In addition, socialbots can be used to infiltrate communities and organizations in order to obtain sensitive information and gain a foothold in the organization by utilizing connections with an organization's employees through an OSN.

Several studies have focused on designing socialbots and attack methods for different purposes: infiltrating an organization by maintaining friendships with profiles (Elyashar et al. 2013), targeting specific users in organization (Elyashar et al. 2014), gathering personal information (Bokobza et al. 2015), or simply gaining influence (Aiello et al. 2012; Messias et al. 2013; Ferrara et al. 2014).

Due to the evidence of the increased use of OSNs by malicious users, one of the greatest challenges is to detect socialbots. Protecting OSNs from socialbots is important to both users and OSN providers. Recently the academic community has become interested in the detection and identification of socialbots and the development of advanced automatic solutions.

The solutions for detecting socialbots controlled by malicious users is primarily focused on feature extraction to distinguish between socialbots and real users using machine learning, crowdsourcing-based detection, honeypots and monitoring-based detection, and graph-based detection.

Key Points

In this work we focus on the creation, design, and development of socialbots in social networks. We present the socialbots' goals and the strategies and actions they perform to achieve these goals. In addition, we discuss detection methods used to identify and detect socialbots.

Historical Background

A bot is a software that runs automated tasks. Bots are designed to maintain communication structures and distribute commands and data through a command and control (C&C) channel (Puri 2003).

The existence of bots has been known for some time; for example, bot algorithms designed to hold a conversation with a human were reported over 50 years ago (Turing 1950).

Socialbots are different from regular bots, since socialbots are designed to be more sophisticated and stealthy (Boshmaf et al. 2011). In recent years socialbots have become increasingly sophisticated and difficult to detect.

The first social botnet "koobface" was revealed in 2008, and this botnet targeted OSNs using clever social engineering attacks and the link opening behavior of social media users (Wuest 2010).

In 2009, another social botnet called Naz bot was discovered on Twitter (Nazario 2009), and more recently, there have been several incidents regarding socialbots. In 2013, the "Pony" botnet was discovered on Facebook, Twitter, and Yahoo. This botnet has stolen two million passwords

(Finkle 2014). In the next chapter, we mention additional sociabot incidents according their goals.

Goals of Socialbots

Socialbots can be used to obtain an influential position, to mislead users, harvest useful and sensitive information from infiltrated users, infiltrate communities and organizations, and distribute malicious content, rumors, spam, and misinformation (Ferrara et al. 2014; Ferrara 2015).

Gaining Influence

Socialbots can be used to gain influence in the OSN and subsequently achieve influence outside the network (Aiello et al. 2012).

Dickerson et al. (2014) named this type of socialbot "Influence Bots" – bots that try to influence conversations on a specific topic in the OSN. Additionally, socialbots are used in political campaigns for propaganda and recruitment using different manipulation strategies depending on the targets of their campaigns. Socialbots can affect public opinion by distributing misinformation with political astroturfing (Berger and Morgan 2015).

As a case in point, during the 2010 US midterm elections, socialbots injected thousands of tweets directing users to websites with fake news reports supporting specific candidates (Ratkiewicz et al. 2011). Another example is an attack with 25,860 socialbots that spread 440,793 tweets in order to disrupt conversations about the Russian election (Thomas et al. 2012). Moreover, in the 2009 Massachusetts election, nine socialbots attempted to cause a specific URL to rise in influence via Twitter. These socialbots produced 929 tweets in 138 minutes, all of which included a link to the website of a candidate. The tweets were used to expose the politician to a large audience (Mustafaraj and Metaxas 2010).

In addition, socialbots can be used to change the stability of markets; in 2013, for example, a group of Syrian hackers claimed responsibility for tweeting a false tweet about explosions at the White House that injured Obama that caused the Dow Jones stock exchange to fall by 1%, and $200 billion dollars was erased from the entire market (Ferrara et al. 2014).

Malicious Content and Spam Distribution

Socialbots can distribute malicious content through OSNs by sending spam, spreading phishing messages, spreading malware, propagating malicious URLs, and launching distributed denial of service (DDoS) attacks (Hwang et al. 2012; Ji et al. 2016).

Recent research (Osterman Research Consultants 2016) confirmed that one out of five businesses are infected by malware through OSN's and an even larger proportion simply wasn't aware how the malware entered.

Incidents of malware distribution are frequently reported; for example, a Trojan attack that infected an estimated 110,000 Facebook users' machines over a two-day period (Trend Micro 2015) and a W32.Koobface worm (Wuest 2010).

Information Gathering

User profiles provide large amounts of private information, including photos, locations, postings, opinions, comments, beliefs, political views, attitudes, and connections. Moreover, additional researching of the profile makes it is easy to determine family relationships, circles of friends, main interests, and hobbies (Sulick 2016).

Attackers can make use of socialbots in order to harvest this useful and sensitive information from infiltrated users in many ways. Socialbots can extract personally identifiable information such as email addresses (Boshmaf et al. 2012). Polakis et al. (2010) demonstrated how just the names from the profiles within OSNs can be used to harvest email addresses as a first step for personalized phishing campaigns.

Karlinsky (2014) mentioned that attackers use the profile information to obtain answers to security questions used to verify the user's identity when attempting to log in to such services. Additionally, Karlinsky (2014) described that today's fraudsters can easily find suppliers that offer personal identifiable information harvesting and sell complete user profiles.

Infiltration

Socialbots can be used to infiltrate communities and organizations in the OSN to pursue a variety of goals, including harvesting information about an organization and perfroming industrial espionage (Sulick 2016), harvesting private employees' information before launching an advanced attack (Molok et al. 2011; Elyashar et al. 2013; Paradise et al. 2014), selecting employees that can be exploited as an entry point into the organization using social engineering methods (an email message with a malicious URL or payload).

Reconnaissance is the first phase and an essential component of a successful advanced persistent threat (APT). This phase involves collecting information, an important preparatory step required before the subsequent more aggressive steps of APT attack (Kim et al. 2014; Ask et al. 2013). In this phase, attackers identify and study the targeted organization and collect information about the technical environment and key personnel in the organization using open-source intelligence (OSINT) tools and social engineering techniques (Wuest 2010). OSINT is a form of intelligence collection from publicly available sources, and nowadays it typically refers to aggregating information about a subject via free sources on the Internet (Chen et al. 2014).

Information extracted from OSNs may include positions and roles within the organization, contact information, etc. Attackers can use the collected information to construct the organization's structure, identify leaders, location, and specialized branch offices (Fire and Puzis 2012). In addition, the attacker can select organization members that can be exploited to penetrate the organization and serve as potential entry points into the organization (Boshmaf et al. 2011; Elyashar et al. 2013). Once an attacker is a friend of an employee, he or she may trick them to download infected emails with malicious attachment or URL to provide access to important assets in the organization.

Attackers can also perform an attack through news, status messages, or job postings that lead the user to a subverted Internet resource (Section 9 lab 2014).

Recently, there have been an increasing number of incidents reported in the media regarding cyber-attacks using OSNs; in 2015, Russian intelligence used socialbots to penetrate the Pentagon. In this case a Russian intelligence officer identified targets in the OSN and fabricated a profile that would be appealing (based on common interest) to these targets. He established a connection with the targets directly and/or developed a relationship with a friend or follower of the target who shared the same interest. Once the relationship with the target matured, the Russian intelligence officer sent the target a phishing message with a link or attachment, thereby gaining access to the target's computer holdings. (Sulick 2016).

In 2011–2014 an attack originating in Iran took place, primarily targeting senior US military. This attack used artificial profiles on social networking sites to build relationships and trust that were later exploited to gain access to sensitive information and deliver malware (Ahmad 2015). Another recent example is a case involving friend requests that were sent through Viadeo (a professional social media network based in France) to the French offices of Trend Micro. The requests targeted several specific employees, and the profile which sent the requests pretended to be an IT manager from the Trend Micro Australia office who had been with the company for 18 years. Checking the company directory confirmed that there was no employee with that name (Pernet 2015).

Methods Used by Socialbots

Socialbots adopt different methods to achieve their goals. In this section we present the methods, strategies, and actions performed by socialbots according to their goals.

Methods for Gaining Influence

Several studies have presented strategies for connecting profiles using socialbots and gaining influence on the Twitter OSN (Aiello et al. 2012; Messias et al. 2013). Researchers have reached several interesting conclusions. Freitas et al. (2014) found that higher Twitter activity is the most important factor for successful infiltration. Aiello et al. (2012) and Messias et al. (2013) demonstrated that socialbots can become influential, like celebrities, in Twitter.

Additional research on the Twitter OSN was a competition associated with "The Web Ecology Project" (2011); in this project the goal was to explore different ways in which a socialbot could influence a target network of 500 Twitter users. The results of this competition showed that socialbots were able to influence user behavior. The socialbots' strategy, used to persuade targets to interact with them, was based on replying to targets' tweets, mentioning targets in their tweets, retweeting tweets shared by the targets, and following the targets in Twitter.

Mitter and Strohmaier (2013) analyzed the data from the "The Web Ecology Project" to explore the manner in which socialbot attacks can influence the links created within OSNs between targeted real OSN users. They found that socialbots may have the ability to shape and influence the social graph in OSNs.

The Robin Sage Experiment (Ryan and Mauch 2010) is another study that emphasized the influence that can be achieved using socialbots. They created a socialbot that aimed to influence users. The influence was reflected in the ability to gain the trust of other users. The experiment proved that socialbots can manage to attract, interact, and influence victims. During the experiment, the profile interacted and elicited information from senior-level US government and industry personnel in sensitive roles. Robin Sage was offered free conference tickets, she was asked to speak at security conference, and she received multiple job offers and gifts.

The following observation can be made based on the studies presented: to gain influence, Twitter socialbots primarily apply simple strategies, such as only following users that followed the socialbots, and posting tweets about popular and focused topics (Messias et al. 2013). The socialbots were simple, with predictive behavior and without sophisticated strategies, yet they were able to successfully infiltrate the OSN and become popular and influential.

The Sybil attack is also a central attack method used to gain influence in OSNs. This attack refers to the situation in which an attacker creates multiple fake identities (Sybils) in order to unfairly increase power and influence within a target community; the attacker controls the set of the identities and joins a targeted system multiple times with these Sybil identities (Douceur 2002). The attacker can mount many follow-up attacks in order to disrupt the targeted system using the Sybil identities.

Methods for Malicious Content and Spam Distribution

Mitter et al. (2014) classified different attack methods adopted by socialbots within the OSN functionality. The attacks include: (1) Abusive usage of topics – changing the initial meaning of a topic to a specific new topic, (2) Unsolicited Communication – sending messages and communicating in an unsolicited way, (3) Clickjacking attacks – trying to trick users into clicking on links embedded in unobtrusive context, (4) Affiliation Attacks – trying to make a user buy something on a specific website, (5) Spoofing – impersonating a specific user to perform an attack.

Ji et al. (2014) identified the following main phases in the socialbot attack lifecycle. The first phase is infection, in which socialbots use infection mechanisms, such as the use of malicious URLs in an email, unwanted malware downloading, and installation of cracked software. After the infection phase, socialbots perform predefined host behaviors, such as modifying the bootstrap list of a system and checking Internet cookies. After that, socialbots work to build a C&C (Command and Control) connection in order to receive commands from the attacker (botmaster). Finally, in the last phase the bots execute the commands received from the botmaster.

Methods for Information Gathering

Several researchers have designed socialbots that attempted to connect OSN users in order to obtain their personal information (Sophos Press Release 2007; Boshmaf et al. 2011; Magdon-Ismail and Orecchio 2012); in each of these studies, the researchers presented socialbots that infiltrate random user profiles with the mission of achieving as many connections and as much information as possible. The socialbots were able to gather sensitive and personal details such as email addresses, dates of birth, phone numbers, and photos. Sophos Press Release (2007) created a fake profile on Facebook that sent random friend requests to 200 users. The profile obtained a 41% rate of acceptance. Boshmaf et al. (2011) also created socialbots on Facebook that infiltrated random user profiles. They concluded that OSNs are vulnerable to large-scale infiltration, and that most OSN users are not careful enough when accepting friend requests from strangers, especially when they have mutual connections. Additionally, Magdon-Ismail and Orecchio (2012) developed a model for the infiltration of users based on two assumptions: users would like to have as many connections with others as possible, and users are more likely to connect to trusted nodes. Their results showed that random friend requests are much less successful than even simple greedy strategies that select a profile that is a connection of the user in the neighborhood (second level connection).

A number of studies have demonstrated attacks using existing features in the OSN, including the mutual friends feature (Jin et al. 2013) and the people you may know feature (Krombholz).

Jin et al. (2013) defined three types of attacks that use the mutual friends feature when the user's privacy setting does not authorize the attacker to see the user's friend list: (1) Friend exposure attack – an attacker tries to identify many of a target's friends, (2) Distant neighbor exposure attack – an attacker's goal is to identify many of the target's distant neighbors, (3) Hybrid attack – an attacker's goal is to identify both the target's friends and distant neighbors. Their results showed that attackers are able to identify more than 60% of a targeted user's friends and subsequently can harvest their information. Krombholz et al. (2012) simulated a data harvesting attack based solely on the use of the people you may know Facebook feature.

Bilge et al. (2009) suggested a two-stage cross-site profile cloning attack. The first stage was based on identifying a victim and using it to create a new identical profile. In the second stage of the attack, a cross-site profile cloning attack was launched. This attack included the automatic creation of profiles in networks where the victim was not registered. These profiles connected to the victim's friends that had profiles on both networks. The authors were able to conclude the feasibility and effectiveness of this type of socialbot attack.

Methods for Infiltration

Recent research has also focused on methods to infiltrate an organization using socialbots to connect to employees through an OSN.

Elyashar et al. (2013) showed that socialbots can be used to infiltrate an organization by maintaining friendships with profiles in the OSNs. Their method aimed at establishing a foothold in an organization by gaining friends among the organization's employees. The method includes first sending friend requests to the most connected members of the organization and then reaching out to members that have the highest number of friends in common with the socialbot. The researchers focused on two organizations and tested their method on Facebook. They were able to disclose up to 13.55% more employees and up to 18.29% more informal links compared to crawling with a public profile that has no friends. These results demonstrate how easily attackers can infiltrate user's OSN profiles and obtain access to valuable information.

Paradise et al. (2014) expanded the socialbot strategy presented by Elyashar et al. (2013) and proposed a method for acquiring friends in an OSN that are employees of a certain organization by sending a friend request to employees with the highest probability of accepting the friend request. The socialbot can estimate the probabilities of an OSN user to accept its friend requests given the total number of friends that the user has and the number of mutual friends the user shares with the socialbot. The authors found that the probability that a target will accept a friend request from the socialbot can be as high as 80% if they share more than 11 mutual friends (Boshmaf et al. 2011).

Other research conducted by Elyashar et al. (2014) showed that socialbots can be used to infiltrate a specific user from an organization. This method is based on sending friend requests to the friends of a specific user and then sending friend requests directly to the specific user. Their results on two organizations within Facebook showed that they were able to infiltrate 50% and 70% of the targeted users they attempted to infiltrate.

Bokobza et al. (2015) investigated wiring strategies an attacker may employ in order to connect with employees' profiles and harvest leaked information using socialbots. The evaluation was performed using real information (diffusion data) on Twitter and Flickr. Their results emphasize the need to raise employees' awareness to the threats of accepting friend requests from strangers and exposing information on OSNs. Additionally, results demonstrate that the most effective socialbot wiring strategy for harvesting information was PageRank.

Detection of Socialbots

In this section, we discuss various socialbot detection methods. Studies have suggested solutions based on the use of machine-learning techniques, monitoring and honeypots, crowdsourcing, graph-based detection, and other techniques.

Graph-Based Detection

Several studies have presented techniques to detect Sybil attacks using graph-based detection (Yu et al. 2006, 2008; Danezis and Mittal 2009; Cao et al. 2012; Wei et al. 2012; Xie et al. 2012; Xue et al. 2013; Pham et al. 2015). Graph-based detection examines the structure of the social network graph. There are a number of studies utilizing Sybil detection that are based on the probability of a short random walk in the non-Sybil region (Yu et al. 2006, 2008; Danezis and Mittal 2009; Cao et al. 2012; Wei et al. 2012). SybilGuard was among the first Sybil detection approaches (Yu et al. 2006), and SybilLimit improved on SybilGuard, using multiple walks. SybilInfer (Danezis and Mittal 2009), however, does not provide any assumption on the number of Sybil identities accepted per attack edge. SybilDefender (Wei et al. 2012) also utilizes a community detection approach.

Sybil detection relies on social graph structures for detection, as well as the assumption that socialbots cannot send many requests to benign users, and therefore there is a sparse cut between Sybil and non-Sybil regions. Sybil detection also assumes that the honest region is time mixing, meaning that socialbots connect to only a few tightly-knit communities; this assumption was not found true on online social networks, where Sybil profiles do not form tight-knit communities. Instead, they slowly gain access and trust within a close-knit social network and integrate into the honest region as legitimate users (Mohaisen et al. 2010; Yang et al. 2011).

Xie et al. (2012) presented a system that recognizes legitimate users based on the connections and interactions. This study relied on the assumption that legitimate users refuse to interact with unknown profiles, an assumption that was proven to be inaccurate when dealing with advanced attackers (Boshmaf et al. 2011; Elyashar et al. 2013).

Xue et al. (2013) presented two observations: Sybils receive few incoming requests from real users and Sybils are more likely to receive rejections than real users. They proposed new techniques to classify Sybils; their method is based on global vote aggregation and local community expansion. A profile is considered a Sybil if its global acceptance rate is below a certain threshold. They deployed the VoteTrust system at Renren and showed that VoteTrust can accurately detect real, large-scale Sybil collusion.

A recent study by Pham et al. (2015) presented a solution to detect infiltration of a specific user in the organizational OSN. They built a target function (based on distances in the OSN), and when any user sends a friend request to users belonging to the organization, the target function is calculated; scores below a certain threshold indicate that the profile is suspicious. This solution was tested utilizing the attack strategy that was presented in Elyashar et al. (2014).

Honeypots and Monitoring-Based Detection

Several previous studies have used honeypots to detect spambots in OSNs (Webb et al. 2008; Lee et al. 2010; Stringhini et al. 2010; Lee et al. 2011). This research has focused on identification of unique behaviors of spammers using honeypots in order to distinguish between social spammers and legitimate users.

Webb et al. (2008) introduced social honeypots to inspect spam. Their results show that the behavior of spam profiles is followed by distinct temporal patterns: 57.2% of the spam profiles have "About me" content that is not original, i.e., from another profile.

Lee et al. (2010) also attempted to expose social spammers in OSNs. They found that their honeypots were able to identify social spammers with low false positive rates in an effective way. Lee et al. (2011) found that the social honeypots identified polluters much earlier than traditional Twitter spam detection methods.

The possibility of using a monitoring approach has been explored by Paradise et al. (2014, 2015); they presented a method to detect socialbots during the reconnaissance phase of a sophisticated attack in which attackers attempt to infiltrate an organization by intelligently selecting organization member profiles and monitoring their activity. The results showed that they can limit the strength of sophisticated friend request strategies by reducing their effectiveness to a level below that of random spraying.

A number of works have used the monitoring activity of users to detect socialbot (Burghouwt et al. 2011; Wang et al. 2013). In these studies the researchers analyzed the aggregate behavioral patterns of OSN profiles to distinguish between malicious and legitimate users. Beutel et al. (2013) presented "CopyCatch" to detect lockstep page like patterns on Facebook, and they showed that malicious profiles tend to post fake likes to several fraudulent pages at the same time.

Wang et al. (2013) developed a detection approach that uses user clickstreams to identify fake profiles. A clickstream is the sequence of HTTP requests made by a user to a website. Experiments using ground truth data show that their system generates 1% false positives and 4% false negative.

Cao et al. (2014) presented "SynchroTrap," a system used to uncover large groups of malicious profiles. They observed that malicious profiles usually perform loosely synchronized actions, and they group profiles with similar action sequences into clusters, and designate large profile clusters as suspicious.

Other research conducted by Burghouwt et al. (2011) suggested using monitoring activity to detect socialbot communication. They presented a method to detect social media-based C&C traffic by monitoring user activity. They measure causality between user activity and network traffic. The presence or absence of certain key strokes and mouse clicks is used to determine if network traffic is legitimate or associated with a socialbot.

Egele et al. (2015) designed "COMPA" to detect compromised OSN profiles, and they built a behavioral profile for OSN profiles, based on the past messages sent by the profile. This research showed that COMPA can reliably detect compromised OSN profiles.

Machine Learning-Based Detection

Detection using machine learning methods is aimed at distinguishing between real users and fake profiles. Several previous studies have used machine learning to detect spambots in online social networks (Benevenuto et al. 2010; Gee and Teh 2010; Wang 2010; Jin et al. 2011; Mccord and Chuah 2011; Song et al. 2011; Wang et al. 2011; Zhang et al. 2012) in which the identification includes the analysis of collected data, feature extraction, and machine-learning methods for classification.

Wang (2010) used machine learning to identify the spambots in Twitter; they showed that graph-based features and content-based features are efficient and accurate in identifying spambots. Zhang et al. (2012) proposed a framework to detect spammers in OSNs. In addition to feature extraction, their framework included a URL-driven estimation method to measure the similarity between two profiles; they also integrated a graph-based approach in their framework in order to extract dense subgraphs as candidate campaigns. Results on a Twitter dataset showed that they were able to extract the actual campaigns with high precision and recall. Wang et al. (2011) presented a framework for spam detection that can be used across all OSNs. Song et al. (2011) were focused on detecting spam messages in Twitter. They used distance and connectivity between messages to determine whether a message was spam or not; their results indicate that most spam comes from profiles that are not well connected (fewer relations) with the receiver.

The limitation of the presented studies is based on their assumption that legitimate users have many legitimate friends while spammers have a small number of friends, a lower friend request acceptance rate, and almost never reply to comments. As socialbots become more sophisticated and human-like, detection methods based on these assumptions are simply not effective enough.

Machine learning has also been used to detect more advanced socialbots (Chu et al. 2010; Yang et al. 2011). Yang et al. (2011) revealed that existing Sybil defenses are unlikely to succeed in today's OSNs as the Facebook Immune System (Stein et al. 2011), and therefore there is a need for new techniques. Using features, such as the frequency of friend requests and the fraction of accepted requests, the authors were able to train a classifier with a 99% true positive rate (TPR).

Wagner et al. (2012) developed predictive models according to three different feature groups (network, behavioral, and linguistic) in order to identify users who are more susceptible to social infiltration in Twitter. They found that susceptible users (potential victims) tend to use Twitter for conversational purposes and are more open and social, since they communicate with many different users.

Boshmaf et al. (2015) designed and evaluated Íntegro, a defense system that leverages victim classification to rank most real profiles higher than fakes. Íntegro starts by identifying potential victims from user level activities, using supervised machine learning and based on the landing probability of a short random walk that starts from a known real profile. The limitation of this system is that it is intended to complement existing detection systems and is designed to detect automated fake profiles that befriend many victims for subsequent attacks and therefore it is able to detect socialbots with these behaviors.

Dickerson et al. (2014) proposed "SentiBot," sentiment-aware architecture for identifying socialbots on Twitter using tweet sentiment to differentiate between human and nonhuman users on Twitter. They concluded that the use of sentiment-aware features improves accuracy where fielded algorithms currently fail. Davis et al. (2016) also used sentiment features in the classification of socialbots in Twitter; they presented a publicly available service called "BotOrNot" which computes a bot likelihood score. This system generates more than 1,000 features, and using machine learning techniques learns the signature of human-like and bot-like behaviors.

Subrahmanian et al. (2016) described a competition in 2015 in which six teams tried to identify influence bots in Twitter. Their overall framework included: (1) machine learning; (2) clustering, outliers, and network analysis (i.e., finding bots that are distant from all clusters, using local ego networks of known socialbots to obtain insight about the structural connectivity pattern of socialbots); and (3) classification and outlier analysis.

A problem associated with this type of detection is that cyber criminals have begun to sell legitimate profiles that have been compromised in Twitter (Stringhini et al. 2013). Attackers can buy friends, or even profiles, and use them for malicious purposes, making the identification of sophisticated socialbots very difficult since they look like legitimate users.

Table 1 provides a summary of the types of features employed in socialbot identification.

Table 2 lists the related research and the features used by each to identify socialbots.

Crowdsourcing-Based Detection

Wang et al. (2012) suggested the use of humans to detect Sybil profiles. They created an online social turing test platform using data from Facebook and Renren in which "experts" and "turkers" were classified profiles based on the profiles information. The authors observed that experts

Socialbots, Table 1 Feature types used for the identification of socialbots

Features	Description	Examples
Content based	Features that are extracted from content the user exposed using methods as text analysis, natural language processing	Number of links, number of replies/ mentions
Graph based/ network topology based	Extraction of features related to the profile network – connections, retweets, mentions, hashtag cooccurrence	Average clustering coefficient of retweet and mention network, number of followers
User based	Features based on profile information	Age, marital status, gender
Timing based	Features related to timing patterns in activities of the profile	Maximum idle duration between posts, average time between posts
Image content based	Meta data related to the images in the profile	Color histogram, color correlogram
Behavioral/ activity based	Features related to a profile's activity	Number of friend requests a user has sent, incoming requests accepted
Semantic based	Extracted features from the content that are based on sentiment analysis algorithms	Emotion score

Socialbots, Table 2 Researches with regard to features they offered

Features versus research	Content based	Graph based/ network topology based	User based	Timing	Image content based	Behavioral/ activity based	Semantic based
Benevenuto et al. (2010)	•	•					
Boshmaf et al. (2015)		•	•	•		•	
Chu et al. (2010)	•	•	•				
Davis et al. (2016)	•	•	•	•		•	•
Dickerson et al. (2014)	•	•					•
Gee and Teh (2010)		•		•			
Jin et al. (2011)	•	•			•		
Lee et al. (2011)	•	•	•	•			
Mccord and Chuah (2011)	•	•					
Stringhini et al. (2010)	•	•					
Subrahmanian et al. (2016)	•	•	•	•		•	•
Song et al. (2011)		•					
Wagner et al. (2012)	•	•	•			•	•
Wang (2010)	•	•					
Yang et al. (2011)		•				•	
Zhang et al. (2012)	•	•	•				

consistently produce near-optimal results. The limitations of this method are the fact that it might not be cost-effective for an OSN with a large number of users, and the fact that sophisticated socialbots may appear as real as human profiles, so crowdsourcing might not be able to distinguish between a real user and a fake profile (in this study only profile information was analyzed).

Table 3 summarizes defense solutions (detection methods) and attacks in a matrix, as a means of mapping current defense solutions to relevant attacks.

Post-Detection

Once socialbot is detected, a number of actions need to be taken: the socialbot needs to be examined by the research community (Mahmoud et al. 2015), the socialbot needs to be analyzed to understand its behavior, the profile features, connections, content that it has been exposed to or sent .In order to analyze the socialbot it is possible to use methods similar to the analysis of data collected from honeypots (Holz et al. 2008). The analyzing of the socialbot may help in developing new proactive defenses against this threat (Vogt et al. 2007).

In order to assess the damage caused by the socialbot to individual, group or organization, it is necessary to carefully examine its connections and friends and even contact them to understand if there were further actions by the socialbot, for example, by sending a malicious e-mail.

Eventually there is a need to report the socialbot to the OSN providers so they could be tracked and the profile brought down.

Key Applications

Socialbots avoid detection and cause prolonged harm to users, communities, and OSN providers.

Socialbots, Table 3 Defense solutions versus attack matrix

Attacks/Defense	Spambot	Socialbots that infiltrate an organization	Sybil attack	Socialbots	Influence bot
Machine learning-based detection	Benevenuto et al. 2010; Gee and Teh 2010; Wang 2010; Jin et al. 2011; Mccord and Chuah 2011; Song et al. 2011; Wang et al. 2011; Zhang et al. 2012			Chu et al. 2010; Yang et al. 2011; Boshmaf et al. 2015; Davis et al. 2016; Dickerson et al. 2014	Wagner et al. 2012; Subrahmanian et al. 2016
Honeypots and monitoring-based detection	Webb et al. 2008; Lee et al. 2010; Stringhini et al. 2010; Lee et al. 2011	Paradise et al. 2014; Paradise et al. 2015		Burghouwt et al. 2011; Cao et al. 2014; Wang et al. 2013; Egele et al. 2015	
Crowdsourcing-based detection				Wang et al. 2012	
Graph-based detection		Pham et al. 2015	Yu et al. 2006; Yu et al. 2008; Danezis and Mittal 2009; Cao et al. 2012; Wei et al. 2012	Xie et al. 2012; Xue et al. 2013	

As we mentioned before, socialbots applications may include promoting agenda and campaign, obtaining an influential position, harvesting useful and sensitive information from infiltrated users, infiltrating communities and organizations, and distributing malicious content, rumors, spam, and misinformations.

Future Directions

In this article we presented the socialbots' goals and described several strategies and actions that socialbots perform to achieve these goals.

In order to provide the best detection mechanism, one must understand the motives, purposes, and strategies behind these fake profiles.

In general, as socialbots have become more sophisticated and deceptive, most of the detection methods have become less effective, and thus there is a need for new solutions. Future work focuses on developing new algorithms for detecting sophisticated socialbots and improving existing detection mechanisms. Attackers that make use of social networks to infiltrate an organization are largely unaddressed and undetected by traditional mechanisms. Therefore, there is a growing need for tools that can be used to detect reconnaissance and initial penetration performed with the help of social networks.

Cross-References

- ▶ Dark Side of Online Social Networks: Technical, Managerial, and Behavioral Perspectives
- ▶ Privacy in Social Networks, Current and Future Research Trends on
- ▶ Spam Detection on Social Networks

References

Abokhodair N, Yoo D, McDonald DW (2015) Dissecting a social botnet: growth, content and influence in Twitter. In: Proceedings of the 18th ACM conference on computer supported cooperative work & social computing, pp 839–851

Ahmad I (2015) How many internet and #SocialMedia users are fake? http://www.digitalinformationworld.com/2015/04/infographic-how-many-internets-users-are-fake.html. Accessed 2 Apr 2015

Aiello LM, Deplano M, Schifanella R, Ruffo G (2012) People are strange when you're a stranger: impact and influence of bots on social networks. Links 697 (483,151):1–566

Ask M, Bondarenko P, Rekdal JE, Nordbø A, Bloemerus P, Piatkivskyi D (2013) Advanced persistent threat (APT) beyond the hype. Project report in IMT4582 Network security at GjoviN University College, Springer

Benevenuto F, Magno G, Rodrigues T, Almeida V (2010) Detecting spammers on twitter. In: CEAS, The seventh annual collaboration, electronic messaging, anti-abuse and spam conference, July 2010, vol 6, p 12

Berger JM, Morgan J (2015) The ISIS Twitter census: defining and describing the population of ISIS supporters on Twitter. The Brookings project on US relations with the Islamic World 3(20)

Beutel A, Xu W, Guruswami V, Palow C, Faloutsos C (2013) Copycatch: stopping group attacks by spotting lockstep behavior in social networks. In: Proceedings of the 22nd international conference on World Wide Web, pp 119–130

Bilge L, Strufe T, Balzarotti D, Kirda E (2009) All your contacts are belong to us: automated identity theft attacks on social networks. In: Proceedings of the 18th international conference on World wide web, pp 551–560

Bokobza Y, Paradise A, Rapaport G, Puzis R, Shapira B, Shabtai A (2015) Leak sinks: the threat of targeted social eavesdropping. In: 2015 IEEE/ACM international conference on advances in social networks analysis and mining, pp 375–382

Boshmaf Y, Muslukhov I, Beznosov K, Ripeanu M (2011) The socialbot network: when bots socialize for fame and money. In: Proceedings of the 27th annual computer security applications conference, pp 93–102

Boshmaf Y, Muslukhov I, Beznosov K, Ripeanu M (2012) Key challenges in defending against malicious socialbots. In: Presented as part of the 5th USENIX workshop on large-scale exploits and emergent threats

Boshmaf Y, Muslukhov I, Beznosov K, Ripeanu M (2013) Design and analysis of a social botnet. Comput Netw 57(2):556–578

Boshmaf Y, Ripeanu M, Beznosov K, Santos-Neto E (2015) Thwarting fake OSN profiles by predicting their victims. In: Proceedings of the 8th ACM workshop on artificial intelligence and security, pp 81–89

Burghouwt P, Spruit M, Sips H (2011) Towards detection of botnet communication through social media by monitoring user activity. In: International conference on information systems security. Springer, Berlin/Heidelberg, pp 131–143

Cao Q, Sirivianos M, Yang X, Pregueiro T (2012) Aiding the detection of fake profiles in large scale social online services. In: Proceedings of the 9th USENIX conference on networked systems design and implementation, pp 15–15

Cao Q, Yang X, Yu J, Palow C (2014) Uncovering large groups of active malicious profiles in online social networks. In: Proceedings of the 2014 ACM SIGSAC conference on computer and communications security, pp 477–488

Chen P, Desmet L, Huygens C (2014) A study on advanced persistent threats. In: Communications and multimedia security, pp 63–72

Chu Z, Gianvecchio S, Wang H, Jajodia S (2010) Who is tweeting on Twitter: human, bot, or cyborg? In: Proceedings of the 26th annual computer security applications conference, pp 21–30

Danezis G, Mittal P (2009) SybilInfer: detecting sybil nodes using social networks. In: NDSS, presented at NDSS, California, 8–11 Feb 2009

Davis CA, Varol O, Ferrara E, Flammini A, Menczer F (2016) Botornot: a system to evaluate social bots. In: Proceedings of the 25th international conference companion on World Wide Web, pp 273–274

Dickerson JP, Kagan V, Subrahmanian VS (2014) Using sentiment to detect bots on Twitter: are humans more opinionated than bots? In: Advances in social networks analysis and mining (ASONAM), 2014 IEEE/ACM international conference on, pp 620–627

Douceur JR (2002) The sybil attack. In: International workshop on peer-to-peer systems, pp 251–260

Egele M, Stringhini G, Kruegel C, Vigna G (2015) Towards detecting compromised profiles on social networks. IEEE Trans Dependable Secure Comput

Elyashar A, Fire M, Kagan D, Elovici Y (2013) Homing socialbots: intrusion on a specific organization's employee using Socialbots. In: Proceedings of the 2013 IEEE/ACM international conference on ASONAM, pp 1358–1365

Elyashar A, Fire M, Kagan D, Elovici Y (2014) Guided socialbots: infiltrating the social networks of specific organizations' employees. AI Commun 29(1):87–106

Ferrara E (2015) Manipulation and abuse on social media by Emilio Ferrara with Ching-man Au Yeung as coordinator. ACM SIGWEB Newsletter, (Spring):4

Ferrara E, Varol O, Davis C, Menczer F, Flammini A (2014) The rise of social bots. arXiv preprint arXiv:1407.5225

Finkle J (2014) "Pony" botnet steals bitcoins, digital currencies: Trustwave. http://www.reuters.com/article/us-bitcoin-security-idUSBREA1N1JO20140224. Accessed 1 Jan 2014

Fire M, Puzis R (2012) Organization mining using online. Netw Spatial Econ 16(2):545–578

Freitas CA, Benevenuto F, Ghosh S, Veloso A (2014) Reverse engineering socialbot infiltration strategies in twitter. arXiv preprint arXiv:1405.4927

Gee G, Teh H (2010) Twitter spammer profile detection. Unpublished

Gilbert E, Karahalios K (2009) Predicting tie strength with social media. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 211–220

Holz T, Steiner M, Dahl F, Biersack E, Freiling FC (2008) Measurements and mitigation of peer-to-peer-based botnets: a case study on storm worm. LEET 8(1):1–9

Hwang T, Pearce I, Nanis M (2012) Socialbots: voices from the fronts. Interactions 19(2):38–45

Ji Y, He Y, Jiang X, Li Q (2014) Towards social botnet behavior detecting in the end host. In: 2014 20th IEEE international conference on parallel and distributed systems (ICPADS), pp 320–327

Ji Y, He Y, Jiang X, Cao J, Li Q (2016) Combating the evasion mechanisms of social bots. Computers & Security 58:230–249

Jin X, Lin C, Luo J, Han J (2011) A data mining-based spam detection system for social media networks. Proceedings VLDB Endowment 4(12):1458–1461

Jin L, Joshi JB, Anwar M (2013) Mutual-friend based attacks in social network systems. Comput Secur 37:15–30

Jurgens D (2013) That's what friends are for: inferring location in online social media platforms based on social relationships. ICWSM 13:273–282

Karlinsky A (2014) How cybercriminals monetize information obtained from social networks. https://securityintelligence.com/how-cybercriminals-monetize-information-obtained-from-social-networks/. Accessed 3 Sep 2014

Kim Y, Kim I, Park N (2014) Analysis of cyber attacks and security intelligence. In: Mobile, ubiquitous, and intelligent computing, pp 489–494

Krombholz K, Merkl D, Weippl E (2012) Fake identities in social media: a case study on the sustainability of the Facebook business model. J Serv Sci Res 4(2):175–212

Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web, pp 591–600

Lee K, Caverlee J, Webb S (2010) Uncovering social spammers: social honeypots+ machine learning. In: Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval, pp 435–442

Lee K, Eoff BD, Caverlee J (2011) Seven months with the devils: a long-term study of content polluters on Twitter. Paper presented at the ICWSM, Barcelona, 17–21 Jul 2011

Magdon-Ismail M, Orecchio B (2012) Guard your connections: infiltration of a trust/reputation based network. In: Proceedings of the 4th annual ACM web science conference, pp 195–204

Mahmoud M, Nir M, Matrawy A (2015) A survey on botnet architectures, detection and defences. Int J Network Security 17(3):264–281

Mccord M, Chuah M (2011) Spam detection on twitter using traditional classifiers. In: Autonomic and trusted computing, pp 175–186

Messias J, Schmidt L, Oliveira R, Benevenuto F (2013) You followed my bot! Transforming robots into influential users in Twitter. First Monday, 18, 7–1 July 2013

Mitter CW, Strohmaier M (2013) Understanding the impact of socialbot attacks in online social networks. arXiv preprint arXiv:1402.6289

Mitter S, Wagner C, Strohmaier M (2014) A categorization scheme for socialbot attacks in online social networks. arXiv preprint arXiv:1402.6288

Mohaisen A, Yun A, Kim Y (2010) Measuring the mixing time of social graphs. In: Proceedings of the 10th ACM SIGCOMM conference on internet measurement, pp 383–389

Molok NA, Ahmad A, Chang S (2011) Information leakage through online social networking: opening the doorway for advanced persistence threats. J Aust Ins Profess Intellig Officer 19(2):38

Mustafaraj E, Metaxas PT (2010) From obscurity to prominence in minutes: political speech and real-time search. Unpublished

Nazario J (2009) Twitter-based Botnet Command Channel. https://www.arbornetworks.com/blog/asert/twitter-based-botnet-command-channel/. Accessed 13 Aug 2009

Osterman Research Consultants (2016) The need to manage social media properly. http://ostermanresearch.com/wordpress/?p=138. Accessed 17 Mar 2016

Paradise A, Puzis R, Shabtai A (2014) Anti-reconnaissance tools: detecting targeted socialbots. IEEE Internet Comput 18(5):11–19

Paradise A, Shabtai A, Puzis R (2015) Hunting organization-targeted socialbots. In: Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015, pp 537–540

Pernet C (2015) Reconnaissance via professional social networks. http://blog.trendmicro.com/trendlabs-security-intelligence/reconnaissance-via-professional-social-networks/. Accessed 2 Jun 2015

Pham CV, Hoang HX, Vu MM (2015) Preventing and detecting infiltration on online social networks. In: Computational social networks, pp 60–73

Polakis I, Kontaxis G, Antonatos S, Gessiou E, Petsas T, Markatos EP (2010) Using social networks to harvest email addresses. In: Proceedings of the 9th annual ACM workshop on privacy in the electronic society, pp 11–20

Puri R (2003) Bots & botnet: an overview. SANS Institute, 3:58

Ratkiewicz J, Conover M, Meiss M, Gonçalves B, Patil S, Flammini A, Menczer F (2011) Truthy: mapping the spread of astroturf in microblog streams. In: Proceedings of the 20th international conference companion on World wide web, pp 249–252

Ryan T, Mauch G (2010) Getting in bed with Robin Sage. Presented at Black Hat conference, Las Vegas, 24–27 Jul 2010

Section 9 lab (2014) Automated linkedIn social engineering attacks. https://medium.com/section-9-lab/automated-linkedin-social-engineering-attacks-1c88573c577e. Accessed 1 Sep 2014

Song J, Lee S, Kim J (2011) Spam filtering in twitter using sender-receiver relationship. In: International workshop on recent advances in intrusion detection, pp 301–317

Sophos Press Release (2007) Sophos Facebook ID probe shows 41% of users happy to reveal all to potential identity thieves. http://www.sophos.com/en-us/press-office/press-releases/2007/08/facebook.aspx. Accessed 14 Aug 2007

Stein T, Chen E, Mangla K (2011) Facebook immune system. In: Proceedings of the 4th workshop on social network systems, p 8

Stringhini G, Kruegel C, Vigna G (2010) Detecting spammers on social networks. In: Proceedings of the 26th annual computer security applications conference, pp 1–9

Stringhini G, Wang G, Egele M, Kruegel C, Vigna G, Zheng H, Zhao BY (2013) Follow the green: growth and dynamics in twitter follower markets. In: Proceedings of the 2013 conference on internet measurement conference, pp 163–176, ACM

Subrahmanian VS, Azaria A, Durst S, Kagan V, Galstyan A, Lerman K, Waltzman R (2016) The darpa twitter bot challenge. arXiv preprint arXiv:1601.05140

Sulick M (2016) Espionage and social media. https://www.thecipherbrief.com/article/espionage-and-social-media. Accessed 30 Jan 2016

The Web Ecology Project (2011) The 2011 socialbots competition. http://www.webecologyproject.org/category/competition

Thomas K, Grier C, Paxson V (2012) Adapting social spam infrastructure for political censorship. In: Presented as part of the 5th USENIX workshop on large-scale exploits

Trend Micro (2015) Social media malware on the rise. http://blog.trendmicro.com/social-media-malware-on-the-rise/. Accessed 24 Feb 2015

Turing AM (1950) Computing machinery and intelligence. Mind 59(236):433–460

Vogt R, Aycock J Jacobson MJ Jr (2007) Army of botnets. Presented at NDSS, California, 28 Feb–2 Mar 2007

Wagner C, Mitter S, Körner C, Strohmaier M (2012) When social bots attack: modeling susceptibility of users in online social networks. In: Proceedings of the 2nd workshop on making sense of microposts (#MSM2012), pp 46–48

Wang AH (2010) Detecting spam bots in online social networking sites: a machine learning approach. In: IFIP annual conference on data and applications security and privacy, pp 335–342

Wang D, Irani D, Pu C (2011) A social-spam detection framework. In: 8th annual conference on collaboration, Electronic messaging, Anti-Abuse and Spam, pp 46–54

Wang G, Mohanlal M, Wilson C, Wang X, Metzger M, Zheng H, Zhao BY (2012) Social turing tests: crowdsourcing sybil detection. arXiv preprint arXiv:1205.3856

Wang G, Konolige T, Wilson C, Wang X, Zheng H, Zhao BY (2013) You are how you click: clickstream analysis for sybil detection. In: Presented as part of the 22nd USENIX security symposium (USENIX security 13), pp 241–256

Webb S, Caverlee J, Pu C (2008) Social honeypots: making friends with a spammer near You. Presented at the *CEAS*, California

Wei W, Xu F, Tan CC, Li Q (2012) Sybildefender: defend against sybil attacks in large social networks. In: INFOCOM, 2012 proceedings IEEE, pp 1951–1959

Wuest C (2010) The risks of social networking. https://www.symantec.com/content

Xie Y, Yu F, Ke Q, Abadi M, Gillum E, Vitaldevaria K, Mao ZM (2012) Innocent by association: early recognition of legitimate users. In: Proceedings of the 2012 ACM conference on computer and communications security, pp 353–364

Xue J, Yang, Z, Yang X, Wang X, Chen L, Dai Y (2013) Votetrust: leveraging friend invitation graph to defend against social network sybils. In: INFOCOM, 2013 proceedings IEEE, pp 2400–2408

Yang Z, Wilson C, Wang X, Gao T, Zhao B, Dai Y (2011) Uncovering social network sybils in the wild. arXiv preprint arXiv:1106.5321

Yu H, Kaminsky M, Gibbons PB, Flaxman AD (2006) Sybilguard: defending against sybil attacks via social networks. IEEE/ACM Trans Networking 16(3):576–589

Yu H et al (2008) Sybillimit: a near-optimal social network defense against sybil attacks. IEEE symposium on security and privacy

Zhang X, Zhu S, Liang W (2012) Detecting spam and promoting campaigns in the twitter social network. In: 2012 I.E. 12th international conference on data mining, pp 1194–1199

SocialMediaLab

► R Packages for Social Network Analysis

Social–Spatiotemporal Analysis of Topical and Polarized Communities in Online Social Networks

Mauro Coletto[1] and Claudio Lucchese[2]
[1]IMT School for Advanced Studies, Ca' Foscari University of Venice, Pisa, Italy
[2]Ca' Foscari University of Venice, Pisa, Italy

Synonyms

Communities in online social networks; Computational social science; Groups and communities discovery; Polarization in online social networks; Social media analysis and mining

Glossary

Computer science (CS)	Discipline based on a scientific and practical approach to computation and its applications
Computational social science (CSS)	New discipline based on interdisciplinary investigation of the social universe on many scales, ranging from individual actors to the largest groupings, through the medium of computation (Cioffi-Revilla 2014)
Dunbar number	Value of the cognitive limit to the number of people with whom a person can maintain stable social relationships (≈150)
Echo chamber	"Enclosed" system in which information, ideas, or beliefs are amplified or reinforced by internal transmission and repetition
Ego network	Focal node ("ego") and the nodes to which the ego is directly connected (friends or alters) plus the ties, if any, among the alters
Group or community	Set of two or more people who interact with one another, share similar traits, and collectively have a sense of belonging
Machine learning (ML)	Subfield of computer science, which evolved from the study of pattern recognition and computational learning theory in artificial intelligence (AI)
Online social network (OSN)	Platform to build social relations among people who share similar personal and career interests, activities, backgrounds, or real-life connections (Buettner 2016). Alternatively, they are called social network sites (SNS)

Polarizing subgroup	Set of people sharing similar points of view about a specific discussed topic
Social group	Bond-based group characterized by personal social relations among members
Social network	Structure made up of a set of actors, sets of dyadic ties, and interactions between individuals
Social sciences (SS)	Set of academic disciplines concerning society and the relationships among individuals within it
Topical group	Identity-based group whose members share a common interest (topic)
Virtual world	Computer-based simulated environment populated by users who simultaneously and independently explore the setting, participate in its activities, and interact with others. An online social network is an example of a virtual world, as is the web itself

Definition

The analysis of communities in online social networks (OSNs) refers to the task of investigating, from both microscopic and macroscopic points of view, the organization of individuals into groups, their relationships, and interactions with other groups, and the behavior and the development of such groups within the whole OSN. In this work, we analyze *topical* communities, with a focus on *polarizing* topics, and we discuss a general analytical framework to study online communities along the social, spatial, and temporal dimensions.

Introduction

The understanding of user communities and their interactions in OSNs is a crucial task in many application fields, such as sociology, psychology, computer science, and business. In the past, sociological studies were mainly conducted through a modest number of surveys and questionnaires. Nowadays, OSNs allow scientists to investigate large volumes of very detailed data about millions of users. Indeed, the analysis of the social interactions in OSNs is interesting to shed light on human behavior within and beyond such virtual environments.

We present different approaches in studying groups and communities in OSNs, and we give an overview of the results achieved so far. We describe the characterizing features of OSNs such as topical groups, social drivers, and homophily. Furthermore, we look at polarizing communities and how users' opinions shape the structure of these communities. Finally, we focus on the analytical dimensions to be taken into consideration when describing communities in OSNs, and we propose a framework to blend together the social, spatial, and temporal dimensions.

Need for Aggregation and Interaction

> Man is a social animal. (Aristotle, 384–322 BC)

The Greek philosopher Aristotle, more than 2300 years ago, highlighted that the social nature of human beings urges us to self-organize into groups of different scales: family, tribe, and society. Our lives constantly depend on other people.

Human activities and social behaviors have always been a top area of investigation for anthropologists, sociologists, and psychologists. Animals in general exhibit social behaviors, embedded for instance in the concepts of territory and dominance. However, some social traits are exclusive to the human species, which is organized into a social network without analogous cases in the animal realm, mainly the prevalence of rationality-based decision making and use of a complex communication language (Barrett et al. 2007).

In fact, most human desires are based on social life. In developed countries, people have fulfilled psychological and safety needs—as classified by Maslow (1943)—such as food, water, sleep, and security. Beyond these needs, and possibly with an even stronger desire, humans look for a sense of belonging and love, esteem, and finally self-actualization. These all involve social interactions and they can hardly be obtained in isolation. Humans then, by their nature, organize themselves into social structures.

From the "Real World" to the "Virtual World"

Social Networks

Groups and social networks have been studied for decades. The small-world effect is one of the main findings (de Sola Pool and Kochen 1979), with several implications—for instance, the maximum distance between any two users in the social graph was quantified by Milgram in 6 hops according to his famous experiment (Milgram 1967). Moreover, by looking at interactions among people in social groups, researchers have pointed out the presence of *strong* and *weak* ties, which structure the network into tightly clustered communities, with different roles in information spreading (Granovetter 1973).

Recently, the focus of social studies has extended to the digital world, resulting in a "marriage" between social sciences and computer science. Novel network analysis techniques and large-scale computational approaches have been developed to analyze the behavior of individuals and communities in massive virtual contexts (Wasserman and Faust 1994).

Online Social Networks

The birth of OSNs in the late 1990s and their increasing popularity in the early 2000s have answered the human need for belonging even in the virtual world.

The success of OSNs was anticipated by the diffusion of virtual environments and the development of the web. In particular, virtual games have been precursors of OSNs. Messinger et al. (2008) described the historical progression of virtual worlds starting from arcade games, which started in 1972 with the Pong game by Atari Interactive. After that, the path toward OSNs was marked by the introduction of console systems (1986), followed by LAN games, which created the concept of digital communities through internet connectivity. Gaming environments progressively integrated additional social features with unstructured games and player-generated content (e.g., The Sims). Social networking sites are a further evolution in the development of open virtual worlds, with properties that make them equivalent, or at least comparable, to real-world environments.

In OSNs, an individual creates his own profile, publishes content, and interacts with other users through discussions and actions (resharing content, liking, disliking). Users can also set friendship or subscription (following) links with other users.

The world population is ≈7.4 billion people; among them, ≈3.4 billion (46%) are internet users and ≈2.3 billion (31%) are active social media users (Global Web Index data, January 2016). These statistics suggest that there is a large interest in joining OSNs. Oh et al. (2014) have shown positive associations among the number of friends in OSNs, supportive interactions, affect, perceived social support, sense of community, and life satisfaction. On average, the time spent by users in these networking platforms is significant: almost 2 h per day according to the Global Web Index data.

To researchers willing to investigate human behaviors in social environments, such a huge volume of information is a gold mine, with no equivalent in the "real world."

Structure of Online Social Networks

Today, OSNs represent a significant portion of web traffic, and the pervasive use of these platforms, together with the possibility of tracking any user action, has attracted scientists interested in investigating OSNs' properties.

The first studies on OSNs explored the topology and the structure of these large networks. From a topological point of view, an OSN can be considered as a graph where the nodes are users and the edges are connections (friendships or subscription/follower relations). Many works have analyzed OSNs from a structural point of view, showing again a small-world effect (Buyukkokten et al. 2005), i.e., a high clustering coefficient and a short average path length (average degree of separation from 4 to 5) in different OSNs: Flickr, LiveJournal, Orkut, YouTube (Mislove et al. 2007), Twitter (Kwak et al. 2010), Facebook (Backstrom et al. 2012; Ugander et al. 2011; Wilson et al. 2012), and Google+ (Magno et al. 2012). Other studies have evaluated the node in/out degree distribution (which typically is a power law) (Ugander et al. 2011; Wilson et al. 2012) and the degree correlation, thus detecting the presence of a large, strongly connected component (Kumar et al. 2010). Finally, interesting works have investigated the evolution of social graphs over time (Wilson et al. 2012).

Moreover, online social microblogging platforms and social networks have proven to be a rich source of information to track and monitor the behavior of users over time. Interactions in OSNs have been studied by weighting the social graph through quantitative considerations on the strength of social ties. These graphs, called *interaction graphs*, differ from social graphs since they include quantifying mechanisms about the intensity of the connections. The interaction strength in a social network is a mix of the amount of time spent together, intimacy, emotional intensity, and reciprocal services (Granovetter 1973), but in most cases it has been quantified in real OSN applications only in terms of the duration and frequency of contacts (e.g., in Wilson et al. 2012), even though there have been theoretical studies, starting with Marsden and Campbell (1984), which have tried to translate qualities such as intensity and intimacy into quantity values. Interaction graphs in OSNs have been studied, showing microproperties related to ego networks (looking, for instance, at close friends, inactive relationships, homophily, turnover of friendships) and macroproperties related to the whole network (diameters, degree distributions, clustering coefficients), which are generally more stable (e.g., in Twitter (Arnaboldi et al. 2013) and Facebook (Wilson et al. 2012)).

The following sections present the main results achieved in analyzing communities in OSNs and, in particular, they focus on community definition, considering both topical and social groups. A further layer of analysis is added by looking at user interactions. We then present some results aimed at understanding opinions and tracking polarized communities over time. Finally, we look at the main dimensions that have to be taken into consideration in the analysis of groups in the digital world, and we discuss how these dimensions can be used in a unified approach.

Key Points

We analyze communities in OSNs by looking at them through different perspectives, ranging from a computer science point of view to a sociological point of view. We discriminate between social and topical groups to focus in particular on the latter, and we show how to further characterize subcommunities on the basis of user opinions. We look at polarization of users when discussing around a topic and how opinion-based groups can be tracked over time in accordance with the topic's evolution. Finally, we give an overview of some analytical dimensions to be taken into consideration in order to characterize groups in OSNs.

In particular:

- We propose an additional topical group classification based on polarization.
- We describe a method to track polarizing communities and topics for OSNs.
- We analyze topical communities and, in particular, we focus on alternative and isolated groups; we show how deviant community analysis can be extended to take into consideration the relationship with the whole social network.
- We propose an analytical framework to describe communities in OSNs by looking at different dimensions: temporal, social, and spatial.

Historical Background

In social sciences, a *community*, or a group, is defined as a set of two or more people who interact with one another, share similar traits, and collectively have a sense of belonging. This definition implies three main concepts, which have been extensively debated: *interdependence*, *homophily*, and *social identity*. These characteristics shift the definition of community beyond the simplistic idea of a group as an aggregation of individuals and entail a degree of subjectivity, which makes the task of identifying communities hard.

Indeed, interdependence and homophily can be measured, and they have been studied in a quantitative way (Aiello et al. 2012; Bisgin et al. 2010). On the other hand, the concept of social identity, which has been extensively studied—first, by Tajfel (1982)—is hard to frame and has been an object of investigation. The subjectivity related to this concept is hardly treatable within the computer science framework. The concept of group membership as a matter of shared self-definition is predominant (Turner 1981), but it is hardly captured by computational studies that base their findings on cohesive interpersonal relationships by looking at interaction patterns. The matching between sociological findings and computational approaches to quantify them is still a challenging research area.

In the following sections, we discuss the computer science perspective in defining and analyzing communities, and we explore how social sciences and computational approaches can be matched.

Computer Science and Social Sciences

Social Sciences

The analysis of social networks is an interdisciplinary academic field, which emerged from social psychology, sociology, statistics, and graph theory. Social structures such as groups and dyadic ties are analyzed to study human behavior and social interactions. Social studies have successfully defined several theoretical models able to explain the patterns observed in these structures (Wasserman and Faust 1994). In fact, social network and community analysis is currently one of the major cores in contemporary sociology and is also employed in a number of other formal sciences.

Computer Science

In computer science, the term *community* is more frequently used than the term *group,* which has been widely adopted in social sciences. In this work, we use both terms without distinction. According to computer science terminology, the discovery of communities is related to the task of clustering the nodes of the graph used to represent the social network. People are mapped to nodes of a graph, and edges are created according to their interactions. Borrowing tools from clustering and theoretical graph analysis, a number of techniques have thus been used to detect communities in social networks (e.g., the Girvan–Newman method (Girvan and Newman 2002) and the modularity-based method (Newman and Girvan 2004)).

Such “algorithmic communities” emerge from a *data-driven* approach, which is now considered a “paradigm shift” in the machine learning (ML) field. This data-driven phenomenon was described by Kuhn (1962) as a phenomenon in which an abrupt shift in the values, goals, and methods of the scientific community occurs (Cristianini 2014). Some success stories (from spelling correction to face recognition, including question answering, machine translation, and information retrieval) have shown how the data-driven approach centered on machine learning technologies is the winning one in many applications (Cristianini 2014).

Community detection techniques have been largely employed in recent years to describe the structure of complex social systems. However, these algorithmic communities are totally defined on the basis of some graph properties (e.g., density) and discard the subjective concept of the *identity* of the community members. This leads to detection of groups of users who are not always aware of being members of them. Groups detected algorithmically (detected groups) do not correspond to user-generated groups (declared groups), which are considered in social sciences. Attempts to evaluate this mismatch have been made (Aiello 2015).

Computational Social Science

Recently a new discipline based on quantitative understanding of complex social systems

(Cioffi-Revilla 2010) has been born. Computational social science (CSS) is the bridge between social sciences (SS) and computer science (CS), based on the study of what is proper in social studies through computational techniques and approaches developed often in the computer science community. CSS can benefit from the presence of huge volumes of data on society's everyday behavior due to the significant integration of technology into people's lives (Conte et al. 2012).

Social–Spatiotemporal Analysis of Topical and Polarized Communities in Online Social Networks

Users spend a considerable amount of time in OSNs, creating original content (posting), sharing multimedia content from other users (sharing), discussing (messages and comments), and reinforcing external content (liking). Communities emerge around different topics of interaction, and analysis of the social aggregations in a virtual context is interesting to shed light on human behavior.

In the following sections, after reviewing some basic concepts related to the communities in OSNs, we discuss recent research results along three important analytical dimensions: *social*, *spatial*, and *temporal*. Furthermore, we include a novel orthogonal dimension: *polarization*.

Communities in Online Social Networks

Homophily and Diversity

Homophily is a main driver that characterizes communities in both real and digital contexts. Homophily induces similarity between members of communities: "birds of a feather flock together" (McPherson et al. 2001). This is due to two cofounding principles: (i) selection mechanisms (preferences are connected to similar users' traits); and (ii) social contagion (how much linked people influence each other) (Leenders 1997). Homophily has been widely studied in OSNs showing correlation between friendships and interests (Aiello et al. 2012) or between profile information and communication patterns (Leskovec and Horvitz 2008). Local proximity and age are another example of homophily factors in OSNs (Kumar et al. 2005).

On the other hand, it has been shown that diversity in the discussed topics or in the shared content favors the stability of a community as group members continue being stimulated by new input (Ludford et al. 2004). Models of growth and longevity of groups in digital contexts have been also investigated (Backstrom et al. 2006).

Size, Membership, and Barriers

The group size affects the dynamics of interactions. The phenomenon has been deeply studied in real-world social networks by Robin Dunbar (1992). He correlated the volume of the neocortex in primates with the number of social stable relationships they have. He adapted the same theory to humans (Dunbar 1993), concluding that the number of people with whom a person can maintain stable social relationships is about 150. Similar results have been found in the Facebook friendship network, showing similarities between ego network structures in OSNs and in real life (Arnaboldi et al. 2012). Ego networks are graphs where the central node is the studied user and all of the other nodes connected to him/her represent his/her friends. Similarly, Goncalves et al. (2011) performed comparable experiments on Twitter, measuring the average interaction strength.

Topical and Social Groups

Two main processes can be identified in the development of communities in social networks: users create ties based on common interests or based on personal social relationships. The resulting kinds of group have been referred to as *common identity* groups or *common bond* groups (Prentice et al. 1994). We also adopt the lexicon proposed by Martin-Borregon et al. (2014) and refer to those kinds of group as *topical* and *social*.

Members of topical groups discuss a specific topic or a specific area of interest, and they do not usually have personal relationships with each other. Conversely, members of social groups tend to be reciprocal in their interactions with other members, and discussions are focused on multiple topics. One implication is that social groups are vulnerable to turnover, since personal relationships are present and they can influence user departure. Topical groups, on the other hand, are robust to departures and they are open to accepting new members (Aiello 2015).

To distinguish between the different kinds of group, Aiello suggested quantifying the reciprocity of interactions and the topical width of the discussions (Aiello 2015). Typically, greater reciprocity indicates a higher probability that the group is social, while a small topical width indicates topical groups. These variables integrate social and content-based aspects. In practice, groups can be both *topical* and *social*.

From a computer science point of view, instead, when we refer to topics, we mean a multinomial distribution of words that represents a coherent concept in a set of text documents. To extract the most important topics from a piece of text (a topic selection task), different techniques have been developed: the most frequently applied method is the unsupervised latent Dirichlet allocation (LDA) (Blei et al. 2003). Novel methods still based on LDA have been proposed recently (Blei and Lafferty 2007; McAuliffe and Blei 2008; Wang et al. 2009).

Moreover, researchers have explored the relationship between diffusion of a topic and network structure (Barbieri et al. 2013), focusing on the structural and dynamic properties of specific topical communities such as groups supporting political parties (Conover et al. 2011) or groups discussing various conspiracy theories (Bessi et al. 2015), rumors and hoaxes (Ratkiewicz et al. 2010), deviant behaviors (Coletto et al. 2016a), or more ordinary topics such as fashion or sport. In the section "Social Dimension and Polarization" we describe in more detail how we can study a topical group in an OSN. In particular, we discuss the case of deviant communities, which are highly topical.

Topical Groups and Polarization

An interesting fine-grained investigation can be obtained by focusing on user opinion. When users discuss a specific topic, they cluster into subgroups. If the topic is controversial, such groups are strongly polarized. The relation between the topic and polarization is dynamic: users generally start discussing a topic and, around that, different opinions emerge.

Understanding opinion and polarization is a challenging task, and it has recently received a lot of attention in the information retrieval and data mining communities. Analysis of polarization is useful to investigate the evolution of groups, and sometimes it can be used to predict the behavior of users or their activities, e.g., predicting vote intention among Twitter users (Coletto et al. 2015) or understanding product preferences for marketing aims (Leskovec et al. 2007).

In Coletto et al. (2016c), we proposed an iterative procedure to detect polarized users and to monitor topic evolution over time. We focused on the frequent scenario where users interact and produce content according to a set of *polarization classes*. By polarization classes, we mean subjects that require the user to side exclusively with one part. Political parties are typical examples of these classes: users discuss several parties, and their opinion changes over time, but they can eventually vote for only one. Other examples include brand analysis, product comparison, and opinion mining in general. Topic detection and tracking (TDT) (Allan 2012) has been widely explored within the scope of news stream analysis (Walls et al. 1999). We are interested in content and user tracking for polarized users. This notion is connected with the concept of controversy in social media, which has been studied, mostly in political contexts, using data coming from different sources (Adamic and Glance 2005; Coletto et al. 2017b; Garimella et al. 2016; Gysel et al. 2015; Makazhanov et al. 2014).

In these scenarios, the polarization classes are known, and some limited information may also be available, e.g., a set of relevant keywords. By leveraging such limited knowledge, several challenging tasks can be tackled:

1. How to identify the users being polarized (or not) according to those classes
2. How to identify the most relevant subtopics being discussed among such users
3. How to monitor the evolution of such user communities and their online discussions over time

Those tasks are all very challenging, as the available knowledge may be approximate or insufficient, and it may also become obsolete over time. Therefore, the classification into polarization classes should be able to self-update continuously by catching upcoming relevant users and discussion topics.

We have described PTR (Polarization TRacker) (Coletto et al. 2016c) for the discovery of *polarized users* in a Twitter stream. While there exist several works about community detection and trending topic tracking, we have proposed a novel setting where the number of communities is known, but very little information is provided (a keyword per class only) and those communities are competing with each other.

The PTR algorithm is illustrated in Algorithm 1. As input the algorithm receives an initial set of polarized keywords $\{H_c^0\}$ (initial *seed*) for each polarized class *c*. For instance, in Twitter the initial seeds can be selected by analyzing the most frequent hashtags and manually selecting a few per class. One of the benefits of PTR is that after only a few iterations the results

Social–Spatiotemporal Analysis of Topical and Polarized Communities in Online Social Networks, Algorithm. 1 PTR Algorithm

```
Require: a set of users $\mathscr{U}$, their messages $\mathscr{T}$ with keywords $\mathscr{H}$, a set of keywords $H_c^0$ for
         each class $c \in \mathscr{C}$

Ensure: Classification of users $U_c$ and keywords $H_c$

procedure PTR ($\{H_c^0\}_{c \in \mathscr{C}}$)
    $\tau \leftarrow 0$
    repeat
                                        ▷ Classify messages on the basis of the used keywords
        $\{T_c^\tau\}_{c \in \mathscr{C}} \leftarrow$ TweetClassify($\{H_c^\tau\}_{c \in \mathscr{C}}, \mathscr{T}$)
                                        ▷ Classify users on the basis of the written messages
        $\{U_c^\tau\}_{c \in \mathscr{C}} \leftarrow$ UserClassify($\{T_c^\tau\}_{c \in \mathscr{C}}, \mathscr{U}$)
                                        ▷ Find better keywords on the basis of $U_c^\tau$
        $\{H_c^{\tau+1}\}_{c \in \mathscr{C}} \leftarrow$ HashTagsClassify($\{U_c^\tau\}_{c \in \mathscr{C}}$)
        $\tau \leftarrow \tau + 1$
    until convergence
    return $\{\mathscr{U}_c^\tau\}_{c \in \mathscr{C}}, \{\mathscr{T}_c^\tau\}_{c \in \mathscr{C}}, \{H_c^\tau\}_{c \in \mathscr{C}}$
end procedure
```

are less dependent on the size of the original seed, since new relevant keywords are continuously discovered.

The algorithm iterates the following classification steps:

- TweetClassify: a message/post is said to be polarized to class c if it does not mention any keyword from classes other than c, which are denoted with $\{H_c\}_{c \in \mathscr{C}}$.
- UserClassify: a user is polarized to one class c only if his/her polarized tweets of class c are at least twice the number of the polarized tweets of any other class.
- HashTagsClassify: a keyword h is assigned to one class c if $S_c(h) > \beta \cdot S_{c'}(h), \forall c' \neq c$, where $S_c(h)$ is the joint probability of observing h in messages polarized to class c, and not observing h in messages polarized to other classes.

The procedure iterates classifying tweets, users, and keywords (e.g., Twitter hashtags) until convergence. Note that the whole content stream is taken into consideration at each iteration, including those users/posts/keywords that were not previously labeled with a polarization class. In Coletto et al. (2016c), the accuracy of the method is greater in comparison with a k-means baseline in terms of F-measure by a percentage from 7% to 71% in relation to the specific dataset considered.

The PTR algorithm can be used to perform polarization or sentiment analysis, to discover polarized communities, and to study their structural evolution over time in different contexts. In Coletto et al. (2016b), for example, it was used to study the opinions of users about the social phenomenon of migration and the refugee crisis.

A temporal version of the proposed algorithm, TPTR (temporal PTR), is also described in the previously mentioned paper (Coletto et al. 2016c). TPTR is able to track users and topics over time.

A similar approach, aimed at dynamically tracking polarization in OSNs, was proposed by Lu et al. (2015). They presented an efficient optimization-based opinion bias propagation method over the social/information network.

Analytical Dimensions

So far we have pointed out different ingredients that compose the concept of communities in the digital context. Moreover, to analyze a community in an OSN, we might look at different dimensions: temporal, spatial, and social (Martin-Borregon et al. 2014). We extend the three-dimensional characterization of groups proposed by Martin-Borregon et al. (2014) with the concept of polarization of users around a topic.

Temporal Dimension

The temporal dimension is crucial in studying dynamics of groups in a social network. Aiello (2015) has proposed to classify groups into three classes:

- *Short-lived*: The groups in this category show a low level of activity after being created, soon becoming inactive.
- *Evergreen*: The evergreen cluster is characterized by groups created at a certain point in the past, which have been growing in number of users and content produced.
- *Bursty*: These are groups with the lowest skewness and big burstiness, especially in the number of users joining. Usually the highest activity is registered at the beginning of their life and from time to time they experience content production. Some of these groups are related to recurring (e.g., yearly) events that regularly attract the attention of users.

Moreover, we have already described the importance of time in dynamically tracking polarized communities (through TPTR) (Coletto et al. 2016c). To develop realistic predictive models, for instance, we have to take into consideration in a proper way dynamical changes in the network (Tatemura 2000).

From a microscopic point of view, temporal analyses of interactions among users in OSNs are fundamental. Miritello et al. have worked extensively on temporal patterns of human communications and their influence on the spreading of information in social networks (Miritello et al. 2011; Miritello 2013a, b). In OSNs, study on the evolution in time of interactions has been performed in recent research works: Viswanath et al. (2009) showed a dynamical study for the case of Facebook. Many studies have led to insights on how an interaction graph is structurally different from the social network itself, and the temporal component is important to track the strength of the relationships. Over time, social links can grow stronger or weaker, and this knowledge is crucial to characterize communities and their evolution.

Furthermore, understanding dynamical phenomena such as cascades and flames in conversations or linking communication patterns with events though time or content is equally important. The task, however, is challenging, since the variables to consider are innumerable and the data processing is not trivial. In addition to that, integrating the dynamical characteristic of the interactions and external data related to events is a very complex task.

Spatial Dimension

There is a significant increase of interest in collecting and analyzing geo-located data from OSNs. Typically, OSNs enable different location information for users and for actions. Usually there are two classes of geographical information: the locations of the users (GPS, user description) or places mentioned in the interactions (Coletto et al. 2016b).

Several works have studied different aspects of the geographical dimension of OSNs: a broad study on this argument was reported in Scellato et al. (2010). The authors proposed a framework to compare social networks based on two new measures: one capturing the geographical closeness of a node with its network neighborhood and a clustering coefficient weighted on the geographical distance between nodes.

Liben-Nowell et al. (Kumar et al. 2005) have found a strong correlation between location and friendship. Twitter geo-located posts have been studied by Takhteyev et al. (2012) to understand how Twitter social ties are affected by distance. Linked users are identified as “egos” and “alters” and the distance between them is analyzed by considering the correlation with the air travel connection distance and with national borders and languages. An analogous objective was the focus of Kulshrestha et al. (2012), who inferred the locations of 12 million Twitter users in a worldwide dataset. In contrast to the previous paper, they studied the correlation between the Twitter population and the socioeconomic status of a country, suggesting that highly developed countries are characterized by greater Twitter usage.

Finally, the geographical properties of an OSN have also been shown to be useful to study migration phenomena (Coletto et al. 2016b; Hawelka et al. 2014; Zagheni et al. 2014).

Social Dimension and Polarization

Even though it is computationally hard to model the concept of social identity, this is the base driver for human association. Groups give us a sense of belonging to the social world, which is a source of pride and self-esteem. In order to increase our self-image, we enhance the status of the group to which we belong, or we discriminate and hold prejudices against the groups to which we do not belong (Turner 1981). The notion of social closeness and sense of belonging is tightly linked to cultural proximity and sharing opinions and values. Communities emerge in digital contexts, as in real life, following specific drivers classified as *common identity* and *common bond* (Prentice et al. 1994).

A further level of differentiation is the polarization process, which creates subcommunities based on sharing values and opinions around a specific topic. Polarization can be studied by looking at the interactions among these different subcommunities and by evaluating the controversy in the conversations among users belonging to different subgroups. On the web, and in particular in the context of OSNs, controversy has been studied from different perspectives (Coletto et al. 2017b; Dori-Hacohen 2015; Dori-Hacohen and Allan 2013; Garimella et al. 2016).

Moreover, it is interesting to associate polarization with topics. The problem has been partially studied (Jo and Oh 2011), proposing some models that associate topic modeling with sentiment analysis: the Topic Sentiment Mixture (TSM) model (Mei et al. 2007), Multi-Aspect Sentiment (MAS) model (Titov and McDonald 2008), and Joint Sentiment/Topic (JST) model (Lin and He 2009). Briefly, in TSM, each word is connected with a specific topic and independently with a sentiment; in MAS, the set of aspects to be evaluated is fixed and sentiment is modeled as a probability distribution for each aspect; in JST, sentiment is integrated with a topic in a single language model (Jo and Oh 2011). Alternative recent interesting approaches proposed in this field are a semisupervised model based on conditional random fields (Marcheggiani et al. 2014) and a hierarchical text classification method (Esuli et al. 2008). The models still are not flexible enough to catch evolution of topics and opinions in conversation in OSNs, but this is still an open issue.

Deviant Behaviors Online social media are also favorable ecosystems for the formation of topical communities centered on matters that are not commonly taken up by the general public because of the embarrassment, discomfort, or shock they may cause (Coletto et al. 2016a).

Those are communities that depict or discuss what are usually referred to as *deviant behaviors* (Clinard and Meier 2015)—conducts that are commonly considered inappropriate because they are somehow violative of society's norms or moral standards that are shared among the majority of the members of society. Pornography consumption, drug use, excessive drinking, illegal hunting, eating disorders, or any self-harming or addictive practice are all examples of deviant behaviors. Many of them are represented, to different extents, on social media (De Choudhury 2015; Haas et al. 2010; Morgan et al. 2010). However, since all of these topics touch upon different societal taboos, the commonsense assumption is that they are embodied either in niche, isolated social groups, or in communities that might be quite numerous but whose activity runs separately from the mainstream social media life. In line with this belief, researchers have mostly considered those groups in isolation, focusing predominantly on the communication patterns among community members (Tyson et al. 2015) or, from a sociological perspective, on the motivations of their members and on the impact of the groups' activities on their lives and perceptions (Attwood 2005).

In reality, people who are involved in deviant practices are not segregated outcasts but part of the fabric of the global society. As such, they can be members of multiple communities and interact with very diverse sets of people, possibly exposing their deviant behavior to the public. In Coletto et al. (2016a), we aimed to go beyond previous studies that looked at deviant groups in isolation by observing them in context. In particular, we wanted to shed light on a matter that is relevant to both network science and social sciences: how much deviant groups are structurally secluded from the rest of the social network, and what the characteristics of their subgroups who build ties with the external world are (Coletto et al. 2016a).

In Coletto et al. (2016a), we focused on the behavior of *adult content* consumption. Public depiction of pornographic material is considered inappropriate in most cultures, yet the number of consumers is strikingly high (Sabina et al. 2008). Despite that, we were not aware of any study about online communities that produced that type of content interfacing with the rest of the social network. We studied this phenomenon in a large dataset from Tumblr, considering a big sample of the follow and reblog networks for a total of more than 130 million nodes and almost 7 billion dyadic interactions. To spot the community that generated adult content, we also recurred to a large sample of 146 million queries from a 7-month query log from a very popular search engine, out of which we built an extensive dictionary of terms related to adult content that we made publicly available.

The results showed that:

- The deviant network is a tightly connected community structured into subgroups, but it is linked with the rest of the network with a very large number of ties.

- The vastest amount of information originating in the deviant network is produced from a very small core of nodes but spreads widely across the whole social graph, potentially reaching a large audience of people who might unwillingly see that type of content.

Echo Chambers Different polarized clusters usually interact in exchanging points of view, but often it is the case of *echo chambers*, which represent a situation in which information, ideas, and opinions are amplified or reinforced by transmission and repetition inside an "isolated" system, where alternative views are censored or underrepresented.

This phenomenon has been studied in particular in the context of misinformation, by analyzing interactions between scientific communities and groups focused on conspiracy theories. Bessi et al. (2015) studied the scientific community and the conspiracy community in Facebook, and the results showed that polarized communities emerge around distinct types of content and usual consumers of conspiracy news become more self-contained and focused on their specific content.

Key Applications

Toward a Unified Approach

Analytical Framework

The social, spatial, and temporal dimensions, as well as the topic and polarization dimensions, which we have discussed so far, can be used to analyze communities in OSNs in a general and comprehensive way.

In Coletto et al. (2017a), we proposed an analytical framework to investigate social trends in large tweet collections by extracting and crossing information about the following three dimensions: time, location, and polarization. The methodology described how to: (1) extract relevant spatial information; (2) enrich data with the sentiment of the message and of the user (retrieved in an automatic iterative way through machine learning); and (3) perform multidimensional analyses considering content and locations in time.

The approach is general and can be easily adapted to any topic of interest involving multiple dimensions.

Micromarriage and Macromarriage

To study social processes, it is furthermore useful to merge dyadic interaction analysis with the study of multiuser interactions.

Social networks are complex systems and the relationships between atomic components create emergent behaviors that can hardly be modeled directly from the composition of the individual parts (Aiello 2015).

Group analysis then is very important to understand the properties of the collectivity in a macroperspective. The challenge is to merge the analysis of dyadic interactions and one-to-one communication patterns (microanalysis) with more general comprehension of multiuser aggregations (macroanalysis) to fully understand social behaviors.

Research on communities in OSNs through the abovementioned dimensions could create the bridge between individual characterization of the users and their interactions and the understanding of collective behaviors that are detected on a higher level in the social network.

Isolation and Spreading

Finally, it is important to study not only the community and its properties but also its relation to the rest of the network, as was done, for instance, in Coletto et al. (2016a). The authors studied a community in an OSN in relation to all of the network. Usually this is not done because of the idea that some communities are isolated or for the lack of broader data.

Future Directions

We have analyzed communities in online social networks from different angles. We have discussed:

- Why people organize into groups
- Some reasons for the success of online social networks and the digital environment
- How people interact in online social networks
- How communities are detected in online social networks
- The computer science and social sciences perspectives in defining groups
- The success of computational social science and the data-driven approach
- The typology of communities (topical and social)
- The structure of communities in polarized subgroups
- How to track polarization in time and topic evolution
- Some examples of topical communities (in particular, we proposed the study of a deviant community and its relationships with the whole network)
- The importance of time, place, social links, and polarization in analyzing communities in online social networks

This work summarizes the recent advances in this topic, but still many challenges are open. We have contributed to exploring this area by proposing new methods and analyses, but there are many steps that must be taken by the research community to create a robust framework in order to achieve comprehensive studies of interactions among groups in online social networks and their evolution.

In fact, this research area may contribute to creating bridges between the computer science and social science communities in order to develop new meaningful research through computational tools by merging the skills and the knowledge of the two domains.

The study of people and how they interact, both in the real-world case and in the digital environment, is a complicated task because humans are heterogeneous systems and, through their interactions, each system is linked to the others in an even more complex relation. We believe that a successful approach should foster the integration of the different tools and methodologies made available by the research conducted so far in the fields of both social and computational sciences, thus supporting multidimensional analyses as discussed before.

The presence of massive datasets and the continuous increase of new social traces on the one side, and the evolution of methods and technology on the other side, will provide more keys to better deal with the complexity that is the foundational pillar of social phenomena.

Cross-References

▶ Analysis and Mining of Tags, (Micro)Blogs, and Virtual Communities
▶ Assessing Individual and Group Behavior from Mobility Data: Technological Advances and Emerging Applications
▶ Behavior Analysis in Social Networks
▶ Behavior Modeling in Social Networks
▶ Clustering Algorithms
▶ Cognitive Strategic Groups
▶ Community Detection and Analysis on Attributed Social Networks
▶ Community Detection in Social Network: An Experience with Directed Graphs
▶ Community Detection: Current and Future Research Trends
▶ Community Evolution
▶ Community Structure Characterization
▶ Connecting Communities
▶ Data Mining
▶ Detecting and Identifying Communities in Dynamic and Complex Networks: Definition and Survey
▶ Extracting and Inferring Communities via Link Analysis
▶ Geography and Web Communities
▶ Human Behavior and Social Networks
▶ Inferring Social Ties
▶ Mobile Community
▶ Modeling Social Behavior
▶ Modeling Social Preferences Based on Social Interactions
▶ Models for Community Dynamics
▶ Modularity
▶ Modularity Measures

- ▶ Network Analysis
- ▶ Network Mining
- ▶ Networks
- ▶ Online Communities
- ▶ Pornography Online
- ▶ Relationship Extraction
- ▶ Relationship Formation
- ▶ Relationship Mining
- ▶ Relationships
- ▶ Sentiment Analysis in Social Media
- ▶ Sentiment Analysis of Microblogging Data
- ▶ Small Communities in Social Networks
- ▶ Social Interaction Analysis for Team Collaboration
- ▶ Social Media Analysis for Monitoring Political Sentiment
- ▶ Social Media and Social Networking in Political Campaigns/Movements
- ▶ Social Networks and Politics
- ▶ Topology of Online Social Networks
- ▶ Tracking Dynamic Community Evolution and Events Mobility in Social Networks
- ▶ User Behavior in Online Social Networks: Influencing Factors
- ▶ User Sentiment and Opinion Analysis

References

Adamic LA, Glance N (2005) The political blogosphere and the 2004 US election: divided they blog. In: Proceedings of the 3rd international workshop on link discovery. ACM, New York, NY, USA, pp 36–43

Aiello LM (2015) Group types in social media. In: Paliouras G, Papadopoulos S, Vogiatzis D, Kompatsiaris Y (eds) User community discovery, Human-computer interaction series. Springer International Publishing, Switzerland, pp 97–134

Aiello LM, Barrat A, Schifanella R, Cattuto C, Markines B, Menczer F (2012) Friendship prediction and homophily in social media. ACM Trans Web (TWEB) 6(2):9

Allan J (2012) Topic detection and tracking: event-based information organization, vol 12. Springer Science & Business Media

Arnaboldi V, Conti M, Passarella A, Pezzoni F (2012) Analysis of ego network structure in online social networks. In: Privacy, security, risk and trust (PASSAT), 2012 international conference on and 2012 international conference on social computing (SocialCom), pp 31–40. IEEE

Arnaboldi V, Conti M, Passarella A, Dunbar R (2013) Dynamics of personal social relationships in online social networks: a study on Twitter. In: Proceedings of the first ACM conference on online social networks, COSN ’13. ACM, New York, pp 15–26

Attwood F (2005) What do people do with porn? Qualitative research into the comsumption, use, and experience of pornography and other sexually explicit media. Sex Cult 9(2)

Backstrom L, Huttenlocher D, Kleinberg J, Lan X (2006) Group formation in large social networks: membership, growth, and evolution. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, Ithaca, NY, pp 44–54

Backstrom L, Boldi P, Rosa M, Ugander J, Vigna S (2012) Four degrees of separation. In: Proceedings of the 3rd annual ACM web science conference. ACM, Ithaca, NY, pp 33–42

Barbieri N, Bonchi F, Manco G (2013) Cascade-based community detection. In: WSDM. ACM, New York, NY, USA

Barrett L, Henzi P, Rendall D (2007) Social brains, simple minds: does social complexity really require cognitive complexity? Philos Trans R Soc Lond B: Biol Sci 362 (1480):561–575

Bessi A, Coletto M, Davidescu GA, Scala A, Caldarelli G, Quattrociocchi W (2015) Science vs conspiracy: collective narratives in the age of misinformation. PLoS One 10(2)

Bisgin H, Agarwal N, Xu X (2010) Investigating homophily in online social networks. In: Web intelligence and intelligent agent technology (WI-IAT), 2010 IEEE/WIC/ACM international conference on, vol 1. IEEE, Toronto, ON, Canada, pp 533–536

Blei DM, Lafferty JD (2007) A correlated topic model of science. Ann Appl Stat:17–35

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Buettner, R. (2016) Getting a job via career-oriented social networking sites: the weakness of ties. In: 2016 49th Hawaii international conference on system sciences (HICSS). IEEE, Koloa, HI, USA, pp 2156–2165

Buyukkokten O, Adar E, Adamic L (2005) A social network caught in the web. First Monday 8(6):15–40

Cioffi-Revilla C (2010) Computational social science. Wiley Interdiscip Rev: Comput Stat 2(3):259–271

Cioffi-Revilla C (2014) Introduction to computational social science: principles and applications. Berlin/New York: Springer 10 (2014): 978–1

Clinard M, Meier R (2015) Sociology of deviant behavior. Cengage Learning, Wadsworth

Coletto M, Lucchese C, Orlando S, Perego R (2015) Electoral predictions with Twitter: a machine-learning approach. IIR

Coletto M, Aiello LM, Lucchese C, Silvestri F (2016a) On the behaviour of deviant communities in online social networks. In: Tenth international AAAI conference on web and social media (ICWSM), pp 72–81

Coletto M, Lucchese C, Muntean CI, Nardini FM, Esuli A., Renso C, Perego R (2016b) Sentiment-enhanced multidimensional analysis of online social networks:

perception of the Mediterranean refugees crisis. In: Advances in social networks analysis and mining (ASONAM), 2016 IEEE/ACM international conference on, pp 1270–1277. IEEE

Coletto M, Lucchese C, Orlando S, Perego R (2016c) Polarized user and topic tracking in Twitter. In: Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval. ACM, Pisa, ITALY, pp 945–948

Coletto M, Esuli A, Lucchese C, Muntean CI, Nardini FM, Perego R, Renso C (2017a) Perception of social phenomena through the multidimensional analysis of online social networks. Online Soc Netw Media 1:14–32

Coletto M, Garimella K, Gionis A, Lucchese C (2017b) A motif-based approach for identifying controversy. In: Proceedings of the eleventh international conference on web and social media, ICWSM 2017, Montréal, 15–18 May 2017, pp 496–499. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15653

Conover M, Ratkiewicz J, Francisco M, Gonçalves B, Menczer F, Flammini A (2011) Political polarization on Twitter. In: ICWSM, vol 133, pp 89–96

Conte R, Gilbert N, Bonelli G, Cioffi-Revilla C, Deffuant G, Kertesz J, Loreto V, Moat S, Nadal JP, Sanchez A et al (2012) Manifesto of computational social science. Eur Phys J Spec Top 214(1):325–346

Cristianini N (2014) On the current paradigm in artificial intelligence. AI Commun 27(1):37–43

De Choudhury M (2015) Anorexia on Tumblr: a characterization study. In: Florence, Italy, Digital health. ACM

de Sola Pool I, Kochen M (1979) Contacts and influence. Soc Netw 1(1):5–51

Dori-Hacohen S (2015) Controversy detection and stance analysis. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. ACM, New York, NY, pp 1057–1057

Dori-Hacohen S, Allan J (2013) Detecting controversy on the web. In: Proceedings of the 22nd ACM international conference on information & knowledge management. ACM, New York, NY, pp 1845–1848

Dunbar RI (1992) Neocortex size as a constraint on group size in primates. J Hum Evol 22(6):469–493

Dunbar RI (1993) Coevolution of neocortical size, group size and language in humans. Behav Brain Sci 16 (04):681–694

Esuli A, Fagni T, Sebastiani F (2008) Boosting multi-label hierarchical text categorization. Inf Retr 11(4):287–313

Garimella K, De Francisci Morales G, Gionis A, Mathioudakis M (2016) Quantifying controversy in social media. In: Proceedings of the ninth ACM international conference on web search and data mining (WSDM). ACM, pp 33–42

Girvan M, Newman ME (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99 (12):7821–7826

Gonçalves B, Perra N, Vespignani A (2011) Modeling users' activity on Twitter networks: validation of Dunbar's number. PLoS One 6(8):e22,656

Granovetter MS (1973) The strength of weak ties. Am J Sociol:1360–1380

Haas SM, Irr ME, Jennings NA, Wagner LM (2010) Online negative enabling support groups. New Media Soc

Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C (2014) Geo-located Twitter as proxy for global mobility patterns. Cartogr Geogr Inf Sci 41(3):260–271

Jo Y, Oh AH (2011) Aspect and sentiment unification model for online review analysis. In: Proceedings of the fourth ACM international conference on web search and data mining. ACM, New York, NY, pp 815–824

Kuhn TS (1962) The structure of scientific revolutions. University of Chicago Press

Kulshrestha J, Kooti F, Nikravesh A, Gummadi KP (2012) Geographic dissection of the Twitter network. In: Proceedings of the sixth international AAAI conference on weblogs and social media (ICWSM)

Kumar R, Liben-Nowell D, Novak J, Raghavan P, Tomkins A (2005) Theoretical analysis of geographic routing in social networks. CSAIL Technical Reports, MIT Massachusetts, USA

Kumar R, Novak J, Tomkins A (2010) Structure and evolution of online social networks. In: Link mining: models, algorithms, and applications. Springer, New York, pp 337–357

Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on world wide web. ACM, New York, NY, USA, pp 591–600

Leenders, R.: Longitudinal behavior of network structure and actor attributes: modeling interdependence of contagion and selection. Evolution of social networks 1 (1997). Evolution of social networks, 1997, 1: 165–184.

Leskovec J, Horvitz E (2008) Planetary-scale views on a large instant-messaging network. In: Proceedings of the 17th international conference on World Wide Web. ACM, New York, NY, USA, pp 915–924

Leskovec J, Adamic LA, Huberman BA (2007) The dynamics of viral marketing. ACM Trans Web (TWEB) 1(1):5

Lin C, He Y (2009) Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM conference on information and knowledge management. ACM, New York, NY, USA, pp 375–384

Lu H, Caverlee J, Niu W (2015) Biaswatch: a lightweight system for discovering and tracking topic-sensitive opinion bias in social media. In: Proceedings of the 24th ACM international on conference on information and knowledge management. ACM, New York, NY, USA, pp 213–222

Ludford PJ, Cosley D, Frankowski D, Terveen L (2004) Think different: increasing online community participation using uniqueness and group dissimilarity. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, NY, USA, pp 631–638

Magno G, Comarela G, Saez-Trumper D, Cha M, Almeida V (2012) New kid on the block: exploring the Google+ social graph. In: Proceedings of the 2012 ACM conference on internet measurement conference. ACM, New York, NY, USA, pp 159–170

Makazhanov A, Rafiei D, Waqar M (2014) Predicting political preference of Twitter users. Soc Netw Anal Min 4(1):1–15

Marcheggiani D, Täckström O, Esuli A, Sebastiani F (2014) Hierarchical multi-label conditional random fields for aspect-oriented opinion mining. In: Advances in information retrieval. Springer, pp 273–285

Marsden PV, Campbell KE (1984) Measuring tie strength. Soc Forces 63(2):482–501

Martin-Borregon D, Aiello LM, Grabowicz P, Jaimes A, Baeza-Yates R (2014) Characterization of online groups along space, time, and social dimensions. EPJ Data Sci 3(1):8

Maslow AH (1943) A theory of human motivation. Psychol Rev 50(4):370

McAuliffe JD, Blei DM (2008) Supervised topic models. In: Advances in neural information processing systems, pp 121–128

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Annu Rev Sociol 27:415–444

Mei Q, Ling X, Wondra M, Su H, Zhai C (2007) Topic sentiment mixture: modeling facets and opinions in weblogs. In: Proceedings of the 16th international conference on World Wide Web. ACM, New York, NY, USA, pp 171–180

Messinger PR, Stroulia E, Lyons K (2008) A typology of virtual worlds: historical overview and future directions. J Virtual Worlds Res 1(1)

Milgram S (1967) The small world problem. Psychol Today 2(1):60–67

Miritello G (2013a) Information spreading on communication networks. In: Temporal patterns of communication in social networks. Springer, Switzerland, pp 107–130

Miritello G (2013b) Temporal patterns of communication in social networks. Springer

Miritello G, Moro E, Lara R (2011) Dynamical strength of social ties in information spreading. Phys Rev E 83 (4):045102

Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattacharjee B (2007) Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM conference on internet measurement. ACM, New York, NY, USA, pp 29–42

Morgan EM, Snelson C, Elison-Bowers P (2010) Image and video disclosure of substance use on social media websites. Comput Hum Behav 26(6):1405–1411

Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69 (2):026113

Oh HJ, Ozkaya E, LaRose R (2014) How does online social networking enhance life satisfaction? The relationships among online supportive interaction, affect, perceived social support, sense of community, and life satisfaction. Comput Hum Behav 30:69–78

Prentice DA, Miller DT, Lightdale JR (1994) Asymmetries in attachments to groups and to their members: distinguishing between common-identity and common-bond groups. Key Readings Soc Psychol 20(5):484–493

Ratkiewicz J, Conover M, Meiss M, Gonçalves B, Patil S, Flammini A, Menczer F (2010) Detecting and tracking the spread of astroturf memes in microblog streams, Palo Alto, California. arXiv preprint arXiv:1011.3768

Sabina C, Wolak J, Finkelhor D (2008) The nature and dynamics of internet pornography exposure for youth. CyberPshychol Behav 11(6)

Scellato S, Mascolo C, Musolesi M, Latora V (2010) Distance matters: geo-social metrics for online social networks. In: Conference on online social networks, WOSN'10

Tajfel H (1982) Social psychology of intergroup relations. Annu Rev Psychol 33(1):1–39

Takhteyev Y, Gruzd A, Wellman B (2012) Geography of Twitter networks. Soc Netw 34(1):73–81

Tatemura J (2000) Virtual reviewers for collaborative exploration of movie reviews. In: Proceedings of the 5th international conference on intelligent user interfaces. ACM, New York, NY, USA, pp 272–275

Titov I, McDonald RT (2008) A joint model of text and aspect ratings for sentiment summarization. In: ACL, vol 8. Citeseer, pp 308–316

Turner JC (1981) Towards a cognitive redefinition of the social group. Cahiers de Psychologie Cognitive/Current Psychol Cognition, pp 15–40

Tyson G, Elkhatib Y, Sastry N, Uhlig S (2015) Are people really social in porn 2.0? In: ICWSM. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10511

Ugander J, Karrer B, Backstrom L, Marlow C (2011) The anatomy of the Facebook social graph. arXiv preprint arXiv:1111.4503

Van Gysel C, Goethals B, de Rijke M (2015) Determining the presence of political parties in social circles. In: ICWSM, pp 690–693

Viswanath B, Mislove A, Cha M, Gummadi KP (2009) On the evolution of user interaction in Facebook. In: Proceedings of the 2nd ACM workshop on online social networks. ACM, New York, NY, USA, pp 37–42

Walls F, Jin H, Sista S, Schwartz R (1999) Topic detection in broadcast news. In: Proceedings of the DARPA broadcast news workshop, Morgan Kaufmann Publishers, Inc., pp 193–198

Wang Y, Bai H, Stanton M, Chen WY, Chang EY (2009) Plda: Parallel latent Dirichlet allocation for large-scale applications. In: Algorithmic aspects in information and management. Springer, pp 301–314

Wasserman S, Faust K (1994) Social network analysis: methods and applications, vol 8. Cambridge University Press, Cambridge

Wilson C, Sala A, Puttaswamy KP, Zhao BY (2012) Beyond social graphs: user interactions in online social networks and their implications. ACM Trans Web (TWEB) 6(4):17

Zagheni E, Garimella VRK, Weber I, State B (2014) Inferring international and internal migration patterns from Twitter data. In: WWW conference, WWW'14 Companion, April 7–11, 2014, Seoul, Korea.

Socioeconomic Stratification

▶ Demographic, Ethnic, and Socioeconomic Community Structure in Social Networks

Sociograms

▶ History and Evolution of Social Network Visualization

Sociograph Representations, Concepts, Data, and Analysis

Elie Raad[1] and Richard Chbeir[2]
[1]Faculty of Business, Memorial University of Newfoundland, St. John's, NL, Canada
[2]University Pau and Pays Adour, Laboratoire LIUPPA, Anglet, France

Synonyms

Centrality measures; Graph representation; Online social networks' concepts; Social network data; Social network representation

Glossary

A graph	Is usually used to represent networks and consist of nodes to represent actors, and edges to represent relationship
A network	A structure that consists of a set of actors
Social network data	Data available on online social networks
Sociograph analysis	Graph-based analysis of social networks, their concepts, and their data

Introduction

With the proliferation of online social networks, information sharing on these networks is gaining an ever-increasing importance. Obviously, online social networks have found ingenious ways to collect data as users socialize. Not surprisingly, when socializing users communicate, interact, and tend to freely reveal personal information in line with their perceptions and preferences. Understanding the characteristics of social networks is of considerable importance. Namely the structure of the networks, the user-generated content, the level of interaction, as well as other dimensions, can be used to analyze users' behaviors and understand their needs. In this work, we detail the most common representations of social networks, define their fundamental concepts, describe their social network data, and provide an overview of its most common analysis measures.

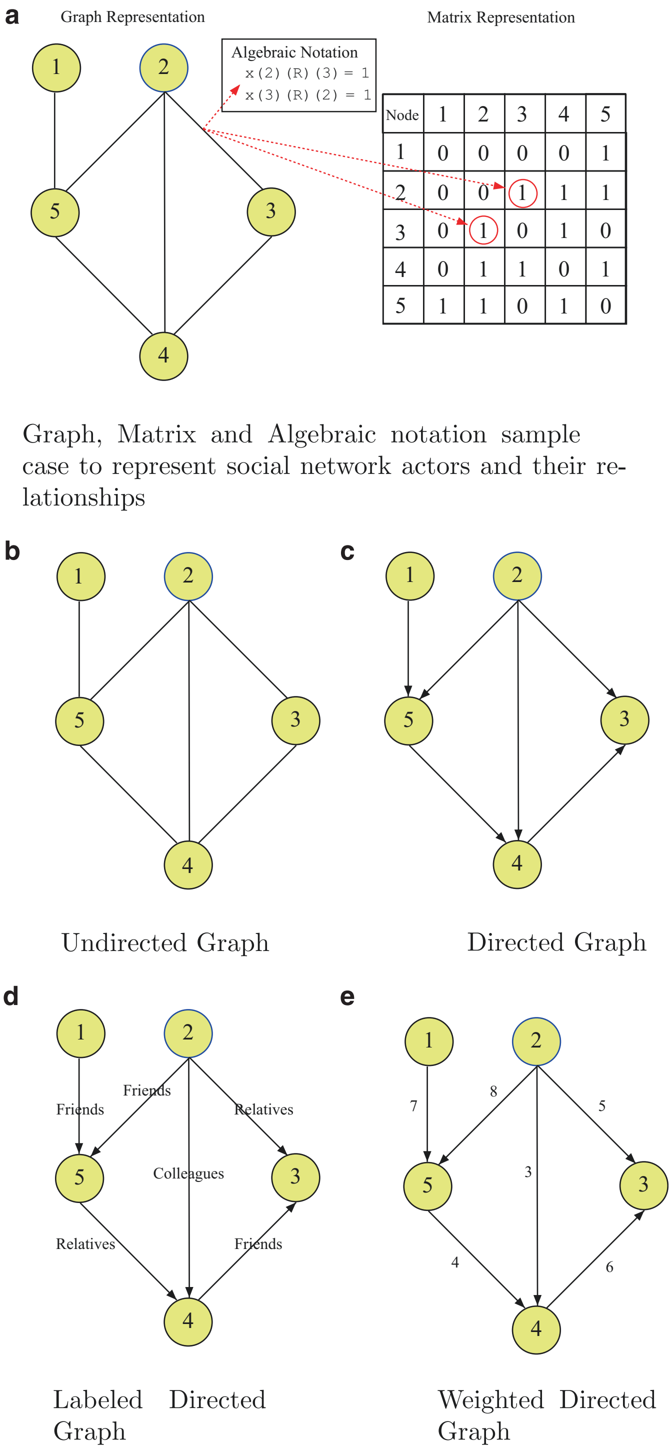
Representation of Social Networks

Finding an appropriate representation that can facilitate efficient and accurate interpretation of network data is an important step in social network studies. Just as graphs are a set of interconnected nodes, social networks are built on the foundation of actors interconnected through relationships. The use of graphs is a powerful visual tool and a formal means to represent social networks as detailed in this section.

Various Notations

There are many notations to represent social networks: algebraic notations, matrices, and graphs. A sample algebraic notation, a matrix representation, and a graph are illustrated in Fig. 1a. Depending on the data to be processed, the notation whose representation best fits the social network to describe is typically selected. But, there are well-known limits to the extent to which social networks can be formalized using matrices or algebraic notations to be recalled here. First, social networks hold valued relations and user-related attributes that algebraic notations cannot

Sociograph Representations, Concepts, Data, and Analysis, Fig. 1 A social network representation using a graph, its related matrix, and a sample algebraic notation (**a**), an undirected graph (**b**), a directed graph (**c**), a labeled graph (**d**), and a weighted graph (**e**) with $n = 5$ nodes and $m = 6$ links



Algebraic Notation

x(2)(R)(3) = 1

x(3)(R)(2) = 1

Node	1	2	3	4	5
1	0	0	0	0	1
2	0	0	1	1	1
3	0	1	0	1	0
4	0	1	1	0	1
5	1	1	0	1	0

Graph, Matrix and Algebraic notation sample case to represent social network actors and their relationships

Undirected Graph

Directed Graph

Labeled Directed Graph

Weighted Directed Graph

handle. Second, matrices are mostly efficient for small networks. Consequently, due to the large size of social networks, matrices are not the most appropriate way to represent these networks. Note that to represent a social network using matrices, a two-way matrix, also called sociomatrix, can be used. A sociomatrix consists of rows and columns that denote social actors, and numbers or symbols in cells that denote existing relationships. Thus, graph-based representations are by far the most common form for modeling social networks (Wasserman and Faust 1994; Newman 2003; Boccaletti et al. 2006). Graphically representing social networks facilitates the understanding, labeling, and modeling of many properties of these networks (e.g., friendships networks with labeled actors and relationships). Hence, graphs can represent various social data properties and their attributes while handling large real-world networks. Besides an adequate vocabulary to denote structural properties, graph-based representations have shown their mathematical reliability as well as their capacity to prove theorems for different social structural properties (Wasserman and Faust 1994). More details about the advantages and drawbacks of each representation are provided in Table 1.

Graph Representation

Graphs are usually used to represent networks in different fields such as biology, sociology, and computer science (Fortunato 2010). Graphs consist of nodes to represent actors, and edges to represent relationships. The terms *nodes* and *objects* are usually used to denote *actors*. Likewise, *edges* may also be called *links*, or *relationships*. Nodes with multiple edges are used to represent *ties* related pairs of actors with more than one relationship.

More formally, a graph, $G = (V, E)$, consists of a set of nodes, V, and a set of edges, E. The number of elements in V and E are respectively denoted as $n = \|V\|$, the number of nodes, and $m = \|E\|$, the number of edges. The ith node, v_i, is usually referred to by its order i in the set V. Note that E consists of a finite set of relationships that is built from all relationships $R_i, R_{i+1}, \ldots, R_k$, where k is the total number of relationships linking the pairs of actors. A subgraph $G' = (V', E')$ of $G = (V, E)$ is a graph such that $V' \subseteq V$ and $E' \subseteq E$. To represent different forms of data and to model the structural properties of social networks, graphs can have their edges and nodes labeled or unlabeled, directed or undirected, weighted or unweighted as explained in what follows.

Directed and Undirected Graphs

In an undirected graph, the order of the connected vertices of an edge is not important. We refer to each link by a couple of nodes i and j such as $e(i, j)$ or e_{ij}, i and j are the end-nodes of the link. A directed graph is defined by a set of nodes and a set of directed edges. The order of the two nodes is important: e_{ij} denotes the link from i to j, and $e_{ij} \neq e_{ji}$. To graphically indicate the direction of the links, directed edges are depicted by arrows. Depending on the nature of the relationship (asymmetric or symmetric), social network graphs can be undirected or directed. In fact, social networks can be modeled as undirected graphs when relationships between actors are mutual

Sociograph Representations, Concepts, Data, and Analysis, Table 1 Social network representations: advantages and drawbacks

Representation	Advantages	Drawbacks
Algebraic notations	Useful for multirelational networks as they can easily denote the combination of relations	Cannot handle valued relations and user-related attributes
Matrices	Efficient for small networks Easy to denotes ties between a set of actors (a matrix for each relationship)	Not a best choice for large social net-works Difficult to use when network data contain information on attributes
Graphs	Handle large social networks Provide a rich vocabulary to easily model social networks (labels, values, weights, etc.) Provide mathematical operations that can be used to quantify structural properties and prove graph-based theorems	Scalable visualization techniques are needed Signed and valued graphs have to be used to represent valued relations

(e.g., symmetric relationships on Facebook (http://www.facebook.com) where e_{ij} or e_{ji} both denote a *friendship* link between user *i* and user *j*). Social networks can also be modeled as directed graphs when relationships are not bidirectional (e.g., asymmetric relationships on Twitter (http://www.twitter.com) where e_{ij} stands for user *i* is *following* user *j*). Figure 1b, c shows respectively a representation of an undirected and a directed graph, both with $n = 5$ and $m = 6$. Directed links are important to evaluate the role of actors in a social network. They are key factors in measuring the centrality of actors in a social network. An interesting research work conducted by Brams et al. (2006) described how to transform undirected graphs to directed ones in order to explore additional information about the networks' structure. This transformation is an important step in understanding the flow of influence in the context of terrorist networks. In another study, Morselli et al. (2007) investigated and compared the structure of criminal and terrorist networks. The authors used links to compute a number of measures such as degree, betweenness, and centrality measures (see section for more details). These measures are used in order to discover the organizational hierarchy and to identify central and powerful criminal and terrorist actors.

Labeled and Unlabeled Graphs

Labels are important since they can identify the type of relationships between social network actors. When graphs are labeled, this means that a label is used to indicate the type of link that characterizes the relationship between the connected labeled nodes. Note that labeled graphs are considered to be signed graphs whenever their edges are labeled with either a + or a −. For example, a signed graph can be used to model the inferred trust or distrust relationships in online social networks (Bachi et al. 2012). Figure 1d shows a labeled graph where the relationship type between linked actors is indicated. On social networks, relationship can be used to organize contacts based on their relationship types. This is useful in different situations such as improving face clustering and annotation of personal photo collections (Zhang et al. 2011), organizing friends into social circles (Raad et al. 2013; McAuley and Leskovec 2012), and enforcing access control (Carminati et al. 2009). Relationship-based access control is highly interesting in order to enable users to manage and fine-tune their privacy settings.

Weighted and Unweighted Graphs

Weights represent the strength of relationships between social network actors. When graphs are weighted, this means that their edges are assigned with a numerical weight, *w*, that can provide various indications such as link capacity, link strength, level of interaction, or similarity between the connected nodes (e.g., the number of messages that actors have exchanged, the number of common friends, etc.). Figure 1e shows a weighted graph (on a scale of 0–10) where the numeric values are assigned to the links and indicate the level of interaction between social network's actors. One way to characterize relationships is by computing their strength. On social networks, link strength is highly correlated with the level of interaction between users. Link strength can be used to model different levels of friendship where high weights represent "close friends" and low weights represent "acquaintances." Xiang et al. (2010) estimated the link strength from interaction activities (e.g., communication, tagging, etc.) and user similarities. More recently, another research explored a more specific aspect related to the predictive capacity of link strength to generalize from one social network to another (Gilbert 2012). Typically, link strength is primarily used to build intelligent systems that can favor interactions with strong ties without missing interesting activities derived by weak ties. Specifically, this interesting study showed that the link strength model captured in one social network can be generalized to another network, one in which it did not train.

Social Networks' Concepts

Networks have been used to model many systems of interest such as the World Wide Web, computer networks, biochemical networks, diffusion networks, and social networks. Each of these networks is a structure that consists of a set of actors representing, for instance, web pages on the World Wide Web or persons in a social network, connected together by relations, representing links between web pages or friendships between persons. Besides these structural properties (actors and relations), Wasserman and Faust (1994) identified a number of fundamental concepts like ties, dyads, triads, subgroups, and groups, that characterize networks. In the following, we detail the concepts of actors, relations, and ties, the building blocks of social networks, before illustrating their use in online social networks.

General Concepts

The following defined concepts (actors, relations, and ties) are particularly important to understand and to study social networks.

Actor

An actor is a social entity that interacts with other entities not only to maintain existing relations but also to establish new ones. On social networks, the concept of actors can refer to various types of entities such as persons, groups, and organizations. Actors interact with each other through a variety of meaningful relations that denote different patterns of communication. Relations like friendship, collaboration, and alliance can vary across time, applications, or in terms of the involved actors (Wilson et al. 2012). Consequently, there are two main categories of networks that can be identified based on the type of actors, one-mode networks and two-mode networks. While one-mode networks have a single type of actors, two-mode networks, also called bipartite, are networks with two types of actors. For instance, social networks modeling friendship between actors are an example of one-mode networks whereas those concerned with group memberships or attendance at events are two-mode networks.

Relation

A relation represents a connection from one actor to another one. A relation, also called relationship, plays an important role when studying the structure of social networks and the interactions among their actors. A relationship is characterized by various features such as its content, direction, and strength. The relationship types have been addressed in several studies. Borgatti et al. (2009) distinguished between four basic types of relationships: similarities, social relations, interactions, and flows. For instance, these relationships can express memberships (e.g., same club), kinships (e.g., mother of), affections (e.g., likes), interactions (e.g., talked to), and flows (e.g., flow of information), among others. Relationships on social networks can be directed or undirected. Depending on their content, relationships may (or may not) have a specific direction. While relationships such as "marriage" and "friendship" are undirected, other relationships such as "parent of" or "fan of" are directed. Social network relationships can also differ in strength. Usually, the strength can be estimated in a variety of ways using information about the actors, their interaction activities, or the correlation between them as the most common indicators (Wilson et al. 2012; Gilbert 2012).

Tie

A tie is the set of all relationships that exist between two actors. It is tightly connected to the concept of relationship as it aggregates the different types of relationships that exist between two actors. Just like relationships, ties also vary in terms of their content, direction, and strength. Actors can be connected either with one relationship exclusively (e.g., employees of the same company) or with many relationships (e.g., employees of the same company and members of a sport club at the same time). Consequently, pairs of actors who maintain more than a single relationship are said to have a tie (Haythornthwaite 1996; Musial and Kazienko 2013). While each individual relationship within a tie carries its own content and direction, the strength of a tie depends on many factors such as the number of relationships that actors maintain, the reciprocity of these relationships, and their duration. Granovetter (1973) distinguished between strong and weak ties on the basis of the time actors spend together, their intimacy, and the emotional intensity of the existing relationships. Generally, weak ties are infrequently maintained with little interactions among actors (e.g., between distant acquaintances). Strong ties link similar actors, such as close friends, whose social circles tightly overlap with each other. Often, actors with strong ties that maintain many kinds of relations tend to communicate frequently with each other and use different channels of communication.

Online Social Networks' Concepts

Social networks and content-sharing sites with social networking functionalities have become an important part of the online activities on the web and one of the most influencing media. Facebook, Twitter, LinkedIn (http://www.linkedin.com), Google + (http://plus.google.com), Instagram (http://www.instagram.com), Pinterest (http: //www.pinterest.com), Flickr (http://www.flickr.com), and Youtube (http://www.youtube.com) are among the most popular online social networks. These networks are attracting an ever-increasing number of users, many of whom are interested in establishing new connections, maintaining existing relations, and using the various social networks' services. The impact of social-based technologies on users, and particularly the influence of online social networks, is becoming the major source of contemporary fascination and controversy (Musial and Kazienko 2013; Heidemann et al. 2012). A number of studies shed the light on different research directions like the implications of online social networks on individual connectivity (Hua and Wellman 2010), the capacity of technology to override cognitive limits in order to socialize with larger groups (Dunbar 2012), and the challenge to maintain a balance between security, privacy, usability, and sociability on online social networks (Zhang et al. 2010; Zheleva et al. 2012).

Social Network User

While many definitions exist for the term social network user (Adamic and Adar 2005; Boyd and Ellison 2007; Schneider et al. 2009), all of them are centered around social network users. First, these users create a personal profile which usually contains identifying information (e.g., name, age, photos, etc.) and captures users' interests (e.g., joining groups, liking brands, etc.). Afterward, users start to socialize by interacting with other network members using a wide variety of communication tools offered by different social networks. In reality, each social network offers particular services and functionalities to target a well-defined community in the real world. Many of these available services are designed to help foster information sharing, bridge online and offline connections to enforce interactions, provide instant information help, and enable users to derive a variety of uses and gratifications from these sites. To make use of the provided functionalities and to stay tuned with their related members, users create several accounts on various social networks where they disclose personal information with varying degrees of sensitivity (Raad et al. 2010). Personal information available on these networks commonly describes users and their interactions, along with their published data.

User Profile

Information about each social network user is maintained in a user profile which contains a number of attributes related to the demographics of users, their personal and professional addresses, their interests and preferences, as well as different types of user-generated contents (e.g., posts, photos, videos, etc.) (Thelwall 2008). Prior studies have noted the importance of user profiles to shape users' personalities, identities, and behaviors on social networks (Ryan and Xenos 2011; Gentile et al. 2012). These studies showed that among the disclosed attributes such as personal information and user-generated contents, photos and status updates have higher preferences for users.

Social Relationship

While myriad social networks' services assist users to find new contacts and establish new connections (e.g., friend suggestion systems through locations (Cranshaw et al. 2010), based on interactions (Wilson et al. 2012), etc.), users get connected to different types of contacts such as friends, relatives, colleagues, and strangers. Nevertheless, social relationship types between users and their contacts are rarely identified neither by the users nor by the existing social network sites (Raad et al. 2013; Tang and Liu 2009; McAuley and Leskovec 2012). This diversity, yet the different levels of social closeness between users and their contacts, entails an increasing need to analyze social interactions for better relationship (and consequently privacy) management. Currently, users are often provided with an exclusive and default relationship type connecting them to each of their contacts within a single social network site. However, it is common that social network users initiate connections with other contacts without any prior offline connection (Ellison et al. 2011). On Facebook, for instance, these contacts are known as *friends* even though social network users do not particularly know or trust them. Consequently, many privacy-related concerns are raised in terms of identity disclosure, information sharing, access control, etc. (Zhang et al. 2010). The default social relationship(s) among the users of a number of famous social networks, along with other information, can be found in Table 2.

Social Network Data

Besides the fact that social networks are made of several components and can have various representations, online social networks can also hold different types of data as detailed in the following. There are many types of social network data that can be collected from various sources on the web (i.e., different social network sites) and extracted from the daily activities and interactions between users. In this context, Schneier (2010) proposed a taxonomy of social data that we further develop into two main categories:

1. **Explicit data:** is the set of explicit information that is provided by social network users or the data that is embedded in the provided information, i.e., metadata embedded in photos. Explicit information may include different forms of data such as text messages, photos, or videos. In this category, social network users actively participate in the creation of information.
   (a) **Service data:** is the set of data that a user provides to the social network to create her account such as the user's name, date of birth, country, etc.
   (b) **Disclosed data:** is what the user posts on her social network profile. This might include comments, posted photos, posted entries, captions, shared links, etc.
   (c) **Entrusted data:** is what the user posts on other users' profiles. This might include comments, captions, shared links, etc.
   (d) **Incidental data:** is what other social network users post about the user. It might include posted photos, comments, notes, etc.
2. **Implicit data:** is the set of information that is not explicitly provided by social network

Sociograph Representations, Concepts, Data, and Analysis, Table 2 Famous social networks with their main focus, default relationship(s), and the relationship's direction

Social network	Focus	Default relationship(s)	Relationship direction
Facebook	General use	Friendship	Symmetrical
Flickr	Photo-sharing	Contact and optionally Friend or family	Symmetrical
Google+	General use	Friends, family, acquaintances, and following	Symmetrical
LinkedIn	Professional	Business	Symmetrical
MySpace	General use	Friendship	Symmetrical
Twitter	Microblogging	Follower–followee	Asymmetrical
Youtube	Video sharing	Subscribed-to	Asymmetrical

users. However, social networks or third parties can use the set of explicit data to infer more information about the user. Inferring implicit data is founded on the analysis of the users' behaviors or derived from one or more user-provided information. For instance, it is possible to predict the characteristics of relationships between a number of users by examining the different aspects related to the patterns of communication between users (e.g., text messages, published photos, number of common friends, etc.) (Diesner et al. 2005; Raad et al. 2013). Consequently, in this category social network users are considered to be passive since the inferred information is extracted from prior activities or previously posted data.

(a) **Behavioral data:** is the data inferred from the user's behaviors. Social networks can collect information about the user's habits by tracking the patterns of activities of the user and consequently analyzing the user's behavior. Inferred behavioral data can reveal various information such as what the user usually do on the social networks, with whom the user usually interacts, and in what news topics the user is interested. Social networks collect such information by analyzing the articles that the user reads, the posts that the user publishes, the game that the user plays on social networks, etc.

(b) **Derived data:** is the data about the user that can be inferred from all other data. It is not related to the habit of the user. For example, the IP address can be used to infer the users' actual location. The derived data can also be inferred from the combination of two (or more) information. For example, if a significant number of contacts live in one city, one can say that the social network user might live there as well. In this case, social networks or third parties must have access to two information in order to infer the derived data (the contacts of a user as the first information and their corresponding hometown as the second information).

Sociograph Analysis

Concerned with the structural analysis of social interactions, research in social network analysis developed new models to study the fundamental properties of diverse theoretical and real-world networks (Luke and Harris 2007). Social network analysis has been used in different application domains such as email communication networks, learning networks, epidemiology networks, terrorist networks, and online social networks. These works tried to answer a handful of questions such as how highly an actor is connected within a network? Who are the most influential actors in a network? How central is an actor within a network?

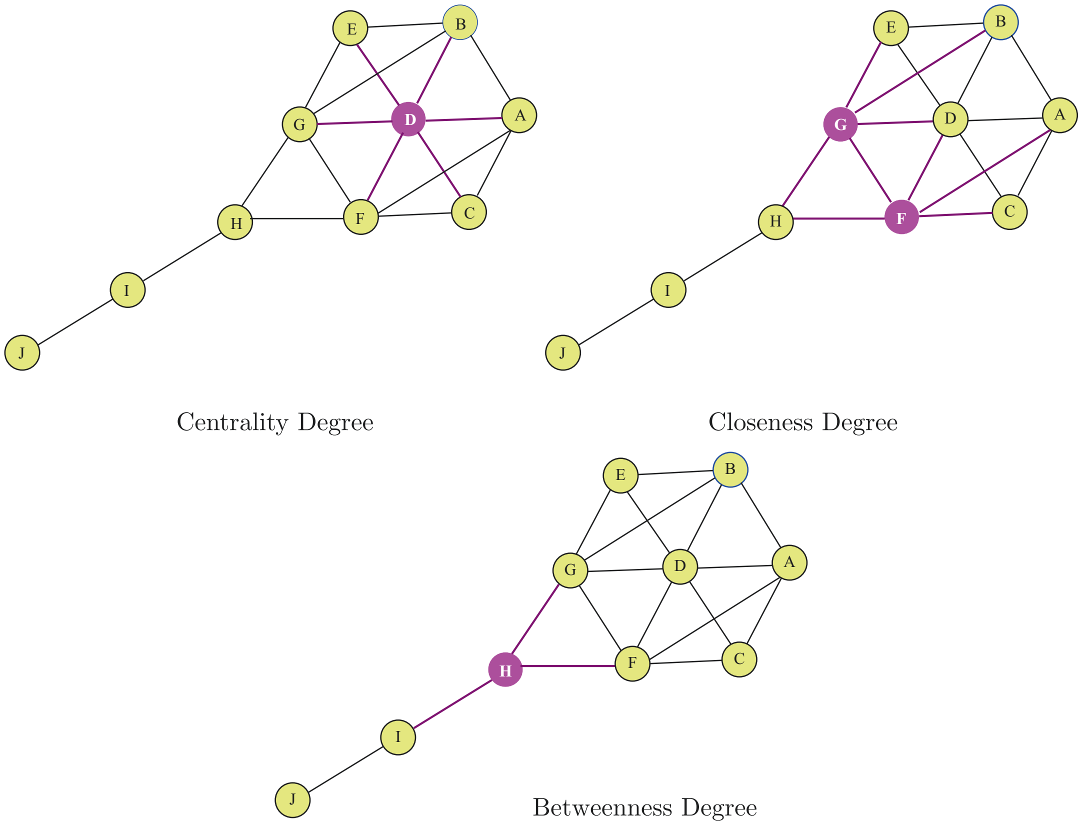
To capture the importance of actors within a network, a number of measures have been proposed in the literature (Koschtzki et al. 2005). A commonly accepted measure is the centrality measure (Faust 1997). Centrality consists of giving an importance order to the actors of a graph by using their connectivity within the network. Several structure-based metrics have been proposed to compute the centrality of an actor within a network, such as degree, closeness, and betweenness centrality (Freeman 1978). In what follows, we explain each of these metrics in details. Table 3 summarizes the characteristics of these structure-based centrality measures. As shown in Fig. 2, different central actor(s) in a network can be identified using each of these structural measures (degree, closeness, and betweenness).

Degree Centrality

Degree centrality measures how much an actor is highly connected to other actors within a network.

Sociograph Representations, Concepts, Data, and Analysis, Table 3 Main centrality measures and their characteristics

Centrality measure	Characteristic
Degree	Measures how much an actor is highly connected to other actors within a network
Closeness	Computes the length of paths from an actor to other actors in the network
Betweenness	Measures the extent to which an actor lies on the paths between other actors



Centrality Degree

Closeness Degree

Betweenness Degree

Sociograph Representations, Concepts, Data, and Analysis, Fig. 2 A network shaped as a kite graph where each centrality measure yields a different central actor: degree centrality (*D*), closeness centrality (*F* and *G*), and betweenness centrality (*H*)

Degree centrality is a local measure since its value is computed by considering the number of links of an actor to other actors directly adjacent to it. A high degree centrality denotes the importance of an actor and gives an indication about (Granovetter 1973) potentially influential actors in the network. With a high degree of centrality, actors in social networks serve as hubs and as major channels of information in a network. Degree centrality, C_D, of an actor, v_i, can be computed as follows (Freeman 1978):

$$C_D(v_i) = \sum_{i=1}^{n} a(v_i, v_j) \tag{1}$$

where n is the total number of actors in the social network, $a(v_i, v_j) = 1$ if and only if v_i and an actor, v_j, are connected by an edge; otherwise $a(v_i, v_j) = 0$.

Closeness Centrality

Closeness centrality computes the length of paths from an actor to other actors in the network. By measuring how close an actor is to all other actors, closeness centrality is also known as the median problem or the service facility location problem. Actors with small length path are considered more important in the network than those with high length path. Closeness centrality, C_C, of an actor, v_i, can be computed as follows (Freeman 1978):

$$C_C(v_i) = \frac{n-1}{\sum_{i=1}^{n} d(v_i, v_j)} \tag{2}$$

where n is the total number of actors in the social network, $d(v_i, v_j)$ is the geodesic distance from actor v_i to another actor v_j.

Betweenness Centrality

Betweenness centrality measures the extent to which an actor lies on the paths between other actors. It denotes the number of times an actor needs to pass via a given actor to reach another one, and thus represents the probability that an actor is involved into any communication between two other actors. Actors with high betweenness centrality facilitate the flow of information as they form critical bridges between other actors or groups of actors. Such central actors control the spread of information between groups of nonadjacent actors. Betweenness centrality, C_B, of an actor, v_i, can be computed as follows (Freeman 1978):

$$C_B(v_i) = \sum_{j<} \sum_{k} \frac{g_{jk}(n_i)}{g_{jk}} \quad i \neq j \neq k \qquad (3)$$

where n is the total number of actors in the social network, $C_B(v_i)$ is the betweenness centrality for actor v_i, and g_{jk} is the number of geodesics linking actors v_j and v_k that also pass through actor v_i.

To sum up, structural characteristics of a graph are a key aspect for social networks as they can be used to analyze the activity and to understand the behaviors of social network users. In most cases, the networks of interconnected users are mainly represented by graphs, while graphs resulting from users' activity are usually referred to as the activity graphs. The activity captured within social networks is between users (the nodes) sharing various social data, connected with directed or undirected relationships (the links), and having different levels of interactions (strong and weak ties). In this regard, these characteristics can be used to identify well-connected, central, and influential users. This would give more visibility and understanding for the network analyzer but at the same time this can possibly reveal additional and sensitive information about the users, thus raising privacy concerns.

Cross-References

- ► Centrality Measures
- ► Classical Algorithms for Social Network Analysis: Future and Current Trends
- ► Graph Classification in Heterogeneous Networks
- ► Network Data Collected Via Web
- ► Social Networking Sites
- ► Spectral Evolution of Social Networks

References

Adamic L, Adar E (2005) How to search a social network. Soc Networks 27(3):187–203

Bachi G, Coscia M, Monreale A, Giannotti F (2012) Classifying trust/distrust relationships in online social networks. In: Privacy, security, risk and trust (PASSAT), 2012 international conference on and 2012 international conference on social computing (SocialCom), Amsterdam, pp 552–557

Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U (2006) Complex networks: structure and dynamics. Phys Rep 424(4–5):175–308

Borgatti S, Mehra A, Brass D, Labianca G (2009) Network analysis in the social sciences. Science 323 (5916):892–895

Boyd D, Ellison N (2007) Social network sites: definition, history, and scholarship. J Comput-Mediat Commun 13 (1):210–230

Brams S, Mutlu H, Ramirez S (2006) Influence in terrorist networks: from undirected to directed graphs. Stud Conflict Terrorism 29(7):703–718

Carminati B, Ferrari E, Perego A (2009) Enforcing access control in web-based social networks. ACM Trans Inf Syst Secur 13(1):1–38

Cranshaw J, Toch E, Hong J, Kittur A, Sadeh N (2010) Bridging the gap between physical location and online social networks. In: UbiComp'10 – Proceedings of the 2010 ACM conference on ubiquitous computing. ACM, New York, pp 119–128

Diesner J, Frantz T, Carley K (2005) Communication networks from the enron email corpus: "it's always about the people. Enron is no different". Comput Math Organ Theory 11(3):201–228

Dunbar RIM (2012) Social cognition on the internet: testing constraints on social network size. Philos Trans R Soc B Biol Sc 367(1599):2192–2201

Ellison N, Steinfield C, Lampe C (2011) Connection strategies: social capital implications of facebook-enabled communication practices. New Media Soc 13 (6):873–892

Faust K (1997) Centrality in affiliation networks. Soc Networks 19(2):157–191

Fortunato S (2010) Community detection in graphs. Phys Rep 486(3–5):75–174

Freeman L (1978) Centrality in social networks conceptual clarification. Soc Networks 1(3):215–239

Gentile B, Twenge J, Freeman E, Campbell W (2012) The effect of social networking websites on positive self-views: an experimental investigation. Comput Hum Behav 28(5):1929–1933

Gilbert E (2012) Predicting tie strength in a new medium. In: Proceedings of the ACM 2012 conference on computer supported cooperative work, CSCW'12. ACM, New York, pp 1047–1056
Granovetter MS (1973) The strength of weak ties. Am J Sociol 78(6):1360–1380
Haythornthwaite C (1996) Social network analysis: an approach and technique for the study of information exchange. Libr Inf Sci Res 18(4):323–342
Heidemann J, Klier M, Probst F (2012) Online social networks: a survey of a global phenomenon. Comput Netw 56(18):3866–3878
Hua W, Wellman B (2010) Social connectivity in America: changes in adult friendship network size from 2002 to 2007. Am Behav Sci 53(8):1148–1169
Koschtzki D, Lehmann K, Peeters L, Richter S, Tenfelde-Podehl D, Zlotowski O (2005) Centrality indices. In: Brandes U Erlebach T (eds) Network analysis. Lecture notes in computer science, 3418. Springer, Berlin/Heidelberg, pp 16–61
Luke D, Harris J (2007) Network analysis in public health: history, methods, and applications. Annu Rev Public Health 28:69–93
McAuley J, Leskovec J (2012) Learning to discover social circles in ego networks. Adv Neural Inf Proces Syst 25:548–556
Morselli C, Gigure C, Petit K (2007) The efficiency/security trade-off in criminal networks. Soc Networks 29 (1):143–153
Musial K, Kazienko P (2013) Social networks on the internet. World Wide Web 16(1):31–72
Newman M (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256
Raad E, Chbeir R, Dipanda A (2010) User profile matching in social networks. In: Proceedings – 13th international conference on network-based information systems, NBiS 2010, Takayama, pp 297–304
Raad E, Chbeir R, Dipanda A (2013) Discovering relationship types between users using profiles and shared photos in a social network. Multimed Tools Appl 64 (1):141–170
Ryan T, Xenos S (2011) Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage. Comput Hum Behav 27(5):1658–1664
Schneider F, Feldmann A, Krishnamurthy B, Willinger W (2009) Understanding online social network usage from a network perspective. In: Proceedings of the 9th ACM SIGCOMM conference on internet measurement conference, IMC'09, Chicago. ACM, New York, pp 35–48
Schneier B (2010) A taxonomy of social networking data. IEEE Secur Priv 8(4):88
Tang L, Liu H (2009) Scalable learning of collective behavior based on sparse social dimensions. In: Proceedings of the 18th ACM conference on information and knowledge management, CIKM'09. ACM, New York, pp 1107–1116
Thelwall M (2008) Social networks, gender, and friending: an analysis of mySpace member profiles. J Am Soc Inf Sci Technol 59(8):1321–1330
Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge
Wilson C, Sala A, Puttaswamy KPN, Zhao BY (2012) Beyond social graphs: user interactions in online social networks and their implications. ACM Trans Web 6(4):17:1–17:31
Xiang R, Neville J, Rogati M (2010) Modeling relationship strength in online social networks. In: Proceedings of the 19th international conference on World wide web, WWW'10. ACM, New York, pp 981–990
Zhang C, Sun J, Zhu X, Fang Y (2010) Privacy and security for online social networks: challenges and opportunities. IEEE Netw 24(4):13–18
Zhang T, Chao H, Tretter D (2011) Dynamic estimation of family relations from photos. In: Lee K-T, Tsai W-H, Liao H-Y, Chen T, Hsieh J-W, Tseng C-C (eds) Advances in multimedia modeling. Lecture notes in computer science, 6524. Springer, Berlin/Heidelberg, pp 65–76
Zheleva E, Terzi E, Getoor L (2012) Privacy in social networks. Synth Lect Data Min Knowl Disc 3(1):1–85

Sociology of the Web

► Web Science

Sociomatrix

► History and Evolution of Social Network Visualization

Sociometric Diagram

► History and Evolution of Social Network Visualization

Sociopsychological Theories

► Friends Recommendations in Dynamic Social Networks

Sociotechnical Systems

▶ Futures of Social Networks: Where Are Trends Heading?

Socio-technical Systems

▶ Social Interaction Analysis for Team Collaboration

Software

▶ NetMiner

Software Development

▶ Social Collaborative Media in Software Development

Sources of Network Data

Monika Cerinšek[1] and Vladimir Batagelj[2,3,4]
[1]Abelium d.o.o, Research and Development, Ljubljana, Slovenia
[2]Faculty of Mathematics and Physics, Department of Mathematics, University of Ljubljana, Ljubljana, Slovenia
[3]Department of Theoretical Computer Science, Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia
[4]University of Primorska, Andrej Marušič Institute, Koper, Slovenia

Synonyms

Almost network data; Archives; Boundary problem; Copyrights; Databases; Ego-centered networks; Ethics; Networks; Observation; Random networks; Semantic web; Surveys

Glossary

Cloud Technology	A use of hardware and software that are delivered as a service over a network (usually the Internet)
Computer-Assisted Text Analysis – CATA	Techniques that model and structure the information content of textual sources on a computer
Genealogy	A study of families and tracing of their lineages
Network Analysis	A study of networks as a representation of relations between discrete objects
Social Network	A social structure based on a set of actors (individuals or organizations) and the ties between these actors
Web Crawler	An Internet bot that automatically browses the World Wide Web

Definition

In network data different entities are linked through their relations. They can be found in many forms and obtained from observations, surveys, archives, databases, etc. Network data can also be generated from other types of data, semantic web, or even be randomly generated.

Introduction

We can find the network data almost everywhere in our lives:

- Cities are linked with roads.
- People in a group are linked by exchange of messages (mail, phone).
- Works from a field of research are linked with citations.
- Researchers are linked through their collaborations.
- Atoms in molecules are linked with their chemical bonds.

- Words are linked according to their co-appearances in sentences of some text.
- In genealogies people are linked by marriage and parent-child ties.

A ***graph*** $\mathcal{G}$ is an ordered pair of sets $(\mathcal{V}, \mathcal{L})$, with the set of ***nodes*** $\mathcal{V}$ and the set of ***links*** $\mathcal{L}$. Every link has two end-nodes. Every link is either directed, an ***arc***, or undirected, an ***edge***. A ***network*** $\mathcal{N} = (\mathcal{V}, \mathcal{L}, \wp, \mathcal{W})$ consists of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{L})$, describing the structure of a network, and additional data: ***properties*** $\wp$ of nodes and ***weights*** $\mathcal{W}$ on links.

There are different types of networks in addition to the ordinary networks.

A ***two-mode*** network is a network $\mathcal{N} = ((\mathcal{I}, \mathcal{J}), \mathcal{L}, \wp, \mathcal{W})$, where the set of nodes $\mathcal{V} = \mathcal{I} \bigcup \mathcal{J}$ is split into two disjoint subsets of nodes $\mathcal{I}$ and $\mathcal{J}$ and each link from $\mathcal{L}$ has one end-node in $\mathcal{I}$ and the other end-node in $\mathcal{J}$.

A ***multirelational*** network $\mathcal{N} = (\mathcal{V}, \mathcal{L}, \wp, \mathcal{W})$ allows multiple relations to exist in the network $\mathcal{L} = (\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_r)$.

In a ***temporal*** network $\mathcal{N} = (\mathcal{V}, \mathcal{L}, \wp, \mathcal{W}, \mathcal{T})$ the time $\mathcal{T}$ is attached to a network. For all nodes and links we have to specify the time intervals in which the element is active (present) in the network. Also properties and weights can change through time – we describe their values as temporal quantities (Batagelj and Praprotnik 2016).

Sometimes a given system is described with a ***collection*** of (often two-mode) ***linked*** networks. For example, bibliographic networks (Batagelj and Cerinšek 2013) and MetaMatrix (Carley 2003).

When constructing a network we must first specify what the nodes are and which relation is linking them – the ***network boundary*** problem (Wasserman and Faust 1994; Marsden 2011). According to the plan of network analyses, we need to bound the set of nodes to those that we need. Along with nodes and links, we also select their properties. We have to decide whether the network is one-mode or two-mode and which node properties are important for our intended analyses. We have to answer several questions about the links: Are the links directed? Are there different types of links (relations)? Can a pair of nodes be linked with multiple links? What are the weights on the links? Is the network static, or is it changing through time?

Sometimes the list of nodes is known in advance (e.g., students in the class). But often the set of nodes is constructed during the network data collection process. In this case we have to specify the membership criteria determining for each potential node whether it belongs to the network or not.

For collecting the network data, the ***snowball*** procedure is often used. We first select a (small) set of nodes as initial candidates. Then we collect the data about each candidate and determine its neighboring nodes. The new ones among them we add to the list of candidates. The inclusion of the new nodes can also be determined by some other criteria, for example, by the distance from the closest initial node. We end this process when the list of candidates is exhausted, or the limit to the number of inspected nodes is reached.

Another problem that often occurs when defining the set of nodes is the ***identification*** (entity resolution, disambiguation) of nodes. The unit corresponding to a node can have different names (***synonymy***), or the same name can denote different units (***homonymy*** or ***ambiguity***). For example, in a bibliography on mathematics from Zentralblatt MATH, the names Borštnik, N. S. Mankoč; Mankoč Borštnik, N.; Mankoč-Borštnik, Norma; Mankoč Borštnik, Norma Susana; Mankoc-Borstnik, N.S.; and Mankoč Borštnik, N.S. belong to the same author. On the other hand, in Zentralblatt MATH at least two different Smith, John W. are recorded, because publications of the author(s) with this name spanned from 1868 to 2007. There are at least 623 different mathematicians with the name Zhang, Li in the MathSciNet Database. Its editors are trying hard, from the year 1985, to resolve the author’s identification problem (Martin et al. 2013) during the data entry phase. In the future the problem could be eliminated by general adoption of initiatives such as ResearcherID or ORCID.

The identification problem also appears when the units are extracted from the plain text, for example, “the President of the USA” and “Barack Obama.” To resolve it we have to provide lists of equivalent terms. Another source of identification problems is the grammar rules of the language used in text. For example, the action “go” can appear in the text in different forms: “go,” “goes,” “gone,” “going,” “went”. To resolve these problems we apply the stemming or lemmatization procedures from natural language processing toolkits such as NLTK or MontyLingua.

A special approach of collecting data for a network analysis is by forming ***ego-centered*** networks (Lozar Manfreda et al. 2004). This approach is used when the population of our interest is too large. From the population we select a sample of units (***egos***) and collect the data about them and their neighbors (***alters***) and links among them. An example of ego-centered networks is the friendship networks of selected persons from Facebook.

Collecting the network data we have to respect legal (copyright, privacy) and ethical constraints (Borgatti and Molina 2003; Eynon et al. 2008; Charlesworth 2008; Breiger 2005).

The network data can be obtained in many ways:

- By observation
- With surveys or interviews
- From archives and databases
- From data organized in a network form
- Derived from the data
- From semantic web
- With generating random networks

Each of the above methods for gathering the network data is described in more details in the following subsections. For details additional references are provided.

Observation

To form a network we must first obtain the data. The ways of obtaining the data have been changing through history following the technological developments. A basic approach is the observation (Mitchell 1969). The observation is a human activity consisting of receiving information about the outside world through the senses, or the recording of data using scientific instruments and includes also any data collected during this activity. Scientific instruments were developed to amplify human powers of observation, such as weighing scales, clocks, telescopes, microscopes, thermometers, cameras, and tape recorders, and also to translate into perceptible form events that are unobservable by human senses, such as voltmeters, spectrometers, infrared cameras, oscilloscopes, interferometers, Geiger counters, x-ray machines, and radio receivers (Shipman et al. 2009).

Making direct measurements is the most accurate method for many variables but can be limited by the technology available. The main alternative to direct observation is to require others to report their activities.

An example of the observational network data collection is described in the PhD thesis of Sampson (1968). He did an ethnographic study of community structure in a New England monastery – he divided 18 novices into 4 groups at 5 time points based on his observations and analyses. Another example is the detection of molecular structure of organic molecules.

Surveys

Survey is a data-gathering method that actively includes the observed people (Marsden 1990). They allow us to study attitudes, beliefs, behaviors, and other characteristics. With carefully prepared questionnaires one can collect vast amount of quality data. A ***questionnaire*** is a list of questions. Answers can be ***closed*** – selected from a given list. They are easier to analyze. But the ***open*** answers, that are not given in advance, allow the analysts to get a wider amount of information. A survey can take different forms: face-to-face, paper and pencil, telephone, e-mail, or online. Nowadays questionnaires are mostly digital (online surveys) that allow them to be adaptable, immediate checking of the entered data, and also collecting some contextual (observational) data.

The use of direct observation in combination with surveys can provide additional information. It can confirm or negate information gained from surveys. As observation itself also the observation in combination with surveys must be prepared. The observant might use appropriate scales, checklists, and other observation materials that are selected in accordance with the questions and possible closed answers in the survey.

An interesting network obtained by interviewing is the *Edinburgh Associative Thesaurus*.

Surveys are the most commonly used methods to gather social network data. They are also used to study interorganizational relations (Mizruchi and Galaskiewicz 1993). For details on surveys and questionnaires, see the entry ▶ "Questionnaires for Measuring Social Network Contacts."

Archives and Databases

An archive is a collection of historical data, or the physical place where they are located (Schmidt 2011). Archives have a historical, cultural, and evidentiary value. Archives exist everywhere, where data has been stored. Every organization has an archive of past activities; universities have archives of past students' achievements and research; backup on a personal computer is an archive of past usage of the computer, etc. With the transition of office work to computers and the spread of Internet, many archives became digital. A database is an organized collection of data, mostly in digital form. Database is organized in records and for each record it has some properties stored (Ullman and Widom 2008). Because data is organized, it is very easy to transform it in a collection of (often two-mode) networks which are then used in the network analysis. Smaller amounts of data can be presented in a tabular form as spreadsheets.

For example, there exist many bibliographic databases (Web of Science, Scopus, Zentralblatt MATH, etc.) that are keeping data about published papers and books. Even the World Wide Web is being partially collected and preserved as an archive for future researchers, historians, and the public.

As a source of data, the archives of various kinds are inexpensive and advantageous for studying especially social networks in the past (Marsden 1990). The network data can be derived from archived data. For example, relations between corporations can be studied based on information about persons on the boards of directors of the corporations.

Historical archives help researchers to gain knowledge about the development of some field – economics, scholar, military, etc. For example, with data from World War II one can study the military movements through the war, the transfer of refugees or prisoners, the transfer of weapons, etc. Another example is the analysis of alliances between the most powerful countries over a selected time period.

Archived data about the inhabitants of a city or an area can be used for genealogical analysis. In genealogy we can search for typical marriage patterns and their irregularities. For example, marriages among relatives to keep the family's wealth, or on the other hand, marriages outside the family to increase its influence. The genealogical data are often available in the GEDCOM format. Large collection of family genealogies is available at the *Genealogy Forum*. For "scientific" genealogies used in anthropological research, see the site *KinSource*.

Activities on the Internet, such as e-mail, chat, and forums, leave traces that can be used as sources for network data. A notorious example is the *Enron e-mail data*.

Especially interesting for network analysis is the World Wide Web as an archive. The web crawlers visit the page with URL from the list of URLs, identify all hyperlinks in it, and add the URLs of these hyperlinks to the list. The largest web archiving organization based on crawling approach is the *Internet Archive*, but also national libraries, national archives, and other organizations are involved in archiving mostly culturally important web content (*Web Archiving Service*).

Enormous archives are being formed by different social networking services such as Facebook, Twitter, LinkedIn, and Google+. These organizations are collecting the data about users, their posts, or tweets. Data about users are not publicly available. The user can download only the data about his past activity and the data that other users declared visible to him.

A large amount of data is stored in Internet Movie Database (*IMDb*) and services such as Amazon, lastFM, *Pandora*, or *Netflix*. Converting data into multiple two-mode networks and combining them in network analysis allows us to obtain information about collaboration between actors, producers and composers, similarity of the movies according to different measures, etc.

With the development of technology, different types of databases occurred, where the type of the database is defined with the way the data is stored in a database. With growth of available data the data warehouses were developed. Data warehouses archive data directly from the source. It is a central source of data for use by managers for creating statistical dashboards and reports about it. The other very popular type of database is cloud database that relies on the cloud technology (Voorsluys et al. 2011).

A graph database (Angles and Gutierrez 2008) is also useful in the network analysis and it is interesting because of the way the data is stored in it. It uses graph structure to represent and store information. Specialized graph database uses a network model, which is conceived as a flexible way of representing objects and their relationships. See for example the Neo4j support of the Panama Papers (ICIJ 2016).

Every day large amounts of data are being collected. Big data (White 2012) is considered to be a collection of large data sets. These data sets are so large and complex that it is very difficult to process them using traditional data processing applications. Also suitable technologies are required such as cluster analysis, machine learning, neural networks, pattern recognition, and anomaly detection.

Many repositories of networks and datasets of other types are available: *Repositories of Datasets*, *KDnuggets Datasets for Data Mining*, *Data Surfing on the World Wide Web*, *Public Data Sets on Amazon Web Services*, *TunedIT*, the *Internet2 Observatory Data Collections*, *Infochimps*, the Cooperative Association for Internet Data Analysis (*CAIDA Data*), *Network Data Sources on Pajek*'s *web page*, *KONECT*, *SNAP*, *GDELT*, *NYC Taxi and Uber Trips*, *Bike sharing*, and *MAG*.

Different activities are traced by their logs. Mobile network operators record the usage of the phones by their users, the data from weather stations is collected, online social network providers collect the data about their users (Abdesslem et al. 2012), different sensor networks are being established, peer-to-peer (P2P) networks are more and more interesting, using the radio-frequency identification (RFID) tags we can follow the movement of their owners or collect the transportation statistics, etc. Such data can be used for prediction or just for the behavioral analysis of the users.

Almost Network Data

Some data are already organized in a network form. A transportation network is a network of roads, pipes, streets, or any other similar structure that allows transportation of some kind. They are represented as links, and crossings are presented as nodes. Another area that deals with a lot of data in a network form is chemistry. The structure of every molecule is a network with atoms as nodes and covalent chemical bonds as links between them. The most interesting for network analysis are organic molecules such as proteins, lipids, hydrocarbons, and DNA. A lot of chemical and biological data is available at *Ensembl*, *GO Database*, *KEGG*, and *Protein Data Bank*.

To analyze such data using the selected network analysis tool, we usually have to transform them into the corresponding input network data format. These issues are elaborated in details in the entry ▶ "Network Data File Formats."

Sometimes special programming solutions should be developed to perform the required transformation. For example, the transformation of the ESRI shape file describing the map of borders between the country's administrative units (states, counties) into the neighborhood relation of the administrative units can be done with a short program in R using the function `poly2nb` from the package `spdep`.

Networks Derived from Data

Some data sources require more sophisticated procedures to transform them into corresponding networks.

Very intriguing data sources are also the daily news archives of the news agencies (Agence France-Presse, Reuters, United Press International, American Press Agency, Xinhua, ITAR-TASS, etc.). A single news is essentially a (tagged) plain text that can be analyzed with ***computer-assisted text analysis*** (Popping 2000). One of the main approaches to this type of text analysis is the semantic text analysis. The units of the text are encoded according to the Chomsky's ***subject-verb-object*** model which can be directly transformed into temporal multirelational networks with subjects and objects as nodes and verbs as relations. Examples of applications of this approach are the *Kansas Event Data System*, *Paul Hensel's International Relations Data Site*, or *Correlates of War*. An elaboration of this approach is given in the Franzosi's book *From Words to Numbers* (Franzosi 2004). See also the *Centering Resonance Analysis approach* proposed by Steve Corman.

Another example is the neighbors' networks. Let $\mathcal{V}$ be a set of (multivariate) units and $d(u,v)$ a ***dissimilarity*** on it. They determine two types of networks:

the ***k-nearest neighbors*** network: $\mathcal{N}(k) = (\mathcal{V}, \mathcal{A}, w)$.
$(u, v) \in \mathcal{A}$ iff v is among k nearest neighbors of u, $w(u,v) = d(u,v)$.
and the ***r-neighbors*** network: $\mathcal{N}(r) = (\mathcal{V}, \mathcal{E}, w)$.
$(u, v) \in \mathcal{E}$ iff $d(u,v) \leq r$ and $w(u,v) = w(v, u) = d(u,v)$.

These networks provide a link between (multivariate) data analysis and network analysis. For larger sets of units a problem of an efficient algorithm for determining the nearest neighbors arises. David M. Mount wrote the *Approximate Nearest Neighbor Library* with fast algorithms for the (approximate) nearest neighbor search. In R these algorithms are available through the function `ann` in package `yaImpute`.

Semantic Web

Semantic web (Berners-Lee et al. 2001) is an upgrade and an extension of the ordinary web. It provides a data layer in the World Wide Web to be used by web services. The basis for semantic web is the semantic description of the web content with the use of metadata and ontologies. The aim is to convert web of unstructured documents into a web of data. This would also make easier to analyze this data, because it would be already in a network form.

Semantic web is based on Uniform Resource Identifier (URI), Resource Description Framework (RDF), and Web Ontology Language (OWL). The URI is a string used to identify a name or a resource and enables interaction with representations of the resource over a network using specific protocols. RDF is a W3C standard for encoding knowledge. It is used for conceptual description or modeling of information from web resources and by computers to seek the knowledge. RDF is actually a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web. The OWL is a family of knowledge representation languages for authoring ontologies.

A piece of knowledge is in RDF represented as a triple subject-predicate-object. A subject denotes the resource; the predicate denotes aspects of the resource and expresses a relationship between the subject and the object. The resources are always named by URIs plus optional anchor IDs (URL and URN are its subsets). The triples form a multirelational network with subjects and objects represented as nodes and predicates determining types of ties – relations. There are large collections of RDF triples: *Linked Data* – Connect Distributed Data across the Web, *Freebase*, and *DBpedia*.

Different syntax formats exist and are quite varying in their complexity: N3, N-Triples, TRiG, TRiX, Turtle, RDF/XML, RDFa, and JSON-LD. The purpose of RDF is to provide an encoding and interpretation mechanism so that resources can be described in a way that a compatible software can understand it. Some formats are not human friendly but more machine friendly. See also SPARQL – an RDF query language.

Generating Random Networks

Generation of random networks (Batagelj and Brandes 2005; van der Hofstad 2011) has become important for studies of complex systems such as electrical power grid, social relations, the World Wide Web and Internet, and collaboration and citation networks of scientists. Random networks are used for modeling classes of graphs.

Paul Erdős and Alfréd Rényi proposed in Erdős and Rényi (1959) an approach to formalize the notion of a random graph. The ***Erdős-Rényi*** model, denoted by $\mathcal{G}(n,m)$, where n is the number of nodes and m is the number of edges, generates a random graph on n nodes and m edges (uniformly) randomly selected among the $n(n-1)/2$ potential edges.

Another, closely related to Erdős-Rényi model, is the ***Gilbert's*** model $\mathcal{G}(n,p)$ (Gilbert 1959), where n is the number of nodes and p is the probability that an edge is included in the random graph. In this model the $n(n-1)/2$ potential edges of a simple undirected graph $\mathrm{G}(n,p) \in \mathcal{G}(n,p)$ are included independently with the probability p.

A model called ***small worlds*** was introduced by Watts and Strogatz (1998). This class of random graphs depends on two structural features: the clustering coefficient is high and the average distance between pairs of nodes is short. Networks such as social networks, the Internet, and gene networks all exhibit small world network characteristics.

The degree distribution of random graph from Erdős-Rényi's or Gilbert's model is sharply concentrated around its average degree. In most real-life networks, it roughly follows the powerlaw. Such networks are called ***scale-free***. Barabási and Albert (1999) described a process of ***preferential attachment*** that generates graphs with this property. The preferential attachment process creates one node at a time and each newly created node is attached to a fixed number of already existing nodes. The probability of selecting a specific node for a neighbor is proportional to its current degree.

Different classes of random graphs can be described also as ***probabilistic inductive classes*** of graphs (Kejžar et al. 2008).

Future Directions

As mentioned in the Introduction one of the basic problems in a network construction is the identification (entity resolution) problem. We expect a further development of methods and tools for solving this problem, and development and proliferation of standardized ids in different fields.

Cross-References

- ▶ Collection and Analysis of Relational Data in Organizational and Market Settings
- ▶ Ethical Issues Surrounding Data Collection in Online Social Networks
- ▶ Ethics of Social Networks and Mining
- ▶ Network Data Collected via Web
- ▶ Network Data File Formats
- ▶ Quality of Social Network Data
- ▶ Questionnaires for Measuring Social Network Contacts

Acknowledgments This work was supported in part by Slovenian Research Agency (ARRS) – project Z7-7614 (B) and grant P1-0294, as well as by grant N1-0011 within the EUROCORES Programme EUROGIGA (project GReGAS) of the European Science Foundation. The first author was financed in part by the European Union, European Social Fund.

References

Abdesslem FB, Parris I, Henderson T (2012) Reliable online social network data collection. Computational social networks. Springer, London

Angles R, Gutierrez C (2008) Survey of graph database models. ACM Comput Surv 40(1):1–39

Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Batagelj V, Brandes U (2005) Efficient generation of large random networks. Phys Rev E 71(3):036113

Batagelj, V, Cerinšek, M: On bibliographic networks. Scientometrics 96 (2013) 3, 845–864. https://doi.org/10.1007/s11192-012-0940-1

Batagelj V, Praprotnik S (2016) An algebraic approach to temporal network analysis based on temporal quantities. Soc Netw Anal Min 6(1):1–22

Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Sci Am 284(5):28–37

Borgatti SP, Molina JL (2003) Ethical and strategic issues in organizational network analysis. J Appl Behav Sci 39(3):337–349

Breiger RL (2005) Ethical dilemmas in social network research: introduction to special issue. Soc Netw 27 (2):88–93

Carley KM (2003) Dynamic network analysis. CASOS/CMU, Pittsburgh

Charlesworth A (2008) Understanding and managing legal issues in internet research. In: Fielding NG, Lee RM, Blank G (eds) The SAGE handbook of online research methods. SAGE, London

Erdős P, Rényi A (1959) On random graphs. Publ Math Debr 6:290–297

Eynon R, Fry J, Schroeder R (2008) The ethics of internet research. In: Fielding NG, Lee RM, Blank G (eds) The SAGE handbook of online research methods. SAGE, London

Franzosi R (2004) From words to numbers: narrative, data, and social science. Cambridge University Press, Cambridge

Gilbert EN (1959) Random graphs. Ann Math Stat 30:1141–1144

Kejžar N, Nikoloski Z, Batagelj V (2008) Probabilistic inductive classes of graphs. J Math Sociol 32 (2):85–109

Lozar Manfreda K, Vehovar V, Hlebec V (2004) Collecting ego-centred network data via the web. Metodološki zvezki 1(2):295–321

Marsden PV (1990) Network data and measurement. Ann Rev Sociol 16:435–463

Marsden PV (2011) Survey methods for network data. In: Scott J, Carrington PJ (eds) The SAGE handbook of social network analysis. SAGE, London

Martin T, Ball B, Karrer B, Newman MEJ (2013) Coauthorship and citation in scientific publishing. Arxiv: http://arxiv.org/abs/1304.0473. Accessed 26 Aug 2016

Mitchell JC (1969) The concept and use of social networks. In: Mitchell JC (ed) Social networks in urban situations. Manchester University Press, Manchester

Mizruchi MS, Galaskiewicz J (1993) Networks of interorganizational relations. Soc Methods Res 22 (1):46–70

Popping R (2000) Computer-assisted text analysis. SAGE, London

Sampson SF (1968) A novitiate in a period of change. An experimental and case study of social relationships. PhD thesis, Cornell University

Schmidt L (2011) Using archives. A guide to effective research. Society of American Archivists, Wheaton

Shipman J, Wilson JD, Todd A (2009) Introduction to physical science, 12th edn. Cengage Learning, Boston

Ullman J, Widom J (2008) First course in database systems, 3rd edn. Prentice-Hall, Upper Saddle River

van der Hofstad R (2011) Random graphs and complex networks. http://www.win.tue.nl/~rhofstad/NotesRGCN.pdf. Accessed 23 Aug 2016

Voorsluys W, Broberg J, Buyya R (2011) Introduction to cloud computing. In: Buyya R, Broberg J, Goscinski A (eds) Cloud computing: principles and paradigms. Wiley, New York

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge

Watts DJ, Strogatz SH (1998) Collective dynamics of ‘small-world’ networks. Nature 393(6684):440–442

White T (2012) Hadoop: the definite guide, 3rd edn. O’Reilly Media, Sebastopol

Web References

Approximate Nearest Neighbor Library. http://www.cs.umd.edu/~mount/ANN

CAIDA (The Cooperative Association for Internet Data Analysis) Data. http://www.caida.org/data/

Centering Resonance Analysis approach proposed by Steve Corman. http://www.crawdadtech.com/

Correlates of War. http://www.correlatesofwar.org/

Data Surfing on the World Wide Web. http://it.stlawu.edu/~rlock/datasurf.html

DBpedia. http://en.wikipedia.org/wiki/DBpedia

Edinburgh Associative Thesaurus. http://www.eat.rl.ac.uk/

Enron E-mail Data. http://www.isi.edu/~adibi/Enron/Enron.htm

Ensembl. http://www.ensembl.org/index.html

Freebase. http://www.freebase.com/

GDELT. http://blog.gdeltproject.org/mapping-media-geographic-networks-the-news-co-occurrence-globe/

Genealogy Forum. http://www.genealogyforum.com/gedcom/

GO Database. http://geneontology.org/page/go-database

ICIJ – The International Consortium of Investigative Journalists (2016) Offshore leaks database. https://offshoreleaks.icij.org/pages/database

Infochimps. http://infochimps.com/

Internet Archive. http://archive.org/index.php

Internet Movie Database. http://www.imdb.com/

KDnuggets Datasets for Data Mining. http://www.kdnuggets.com/datasets/index.html

KEGG: Kyoto Encyclopedia of Genes and Genomes. http://www.genome.jp/kegg/

KinSource. http://kinsource.net/csac/wiki/kinsrc/KinSources/

KONECT – The Koblenz Network Collection. http://konect.uni-koblenz.de/

Linked Data – Connect Distributed Data across the Web. http://linkeddata.org/

MAG – Microsoft Academic Graph. https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

Netflix. https://www.netflix.com/
Network Data Sources on Pajek's web page. http://vladowiki.fmf.uni-lj.si/doku.php?id=pajek:data:index
NYC Taxi and Uber Trips. http://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/
Pandora. http://www.pandora.com/
Paul Hensel's International Relations Data Site. http://www.paulhensel.org/data.html
Protein Data Bank. http://www.rcsb.org/pdb/home/home.do
Public Data Sets on Amazon Web Services. http://aws.amazon.com/publicdatasets/
Repositories of Datasets. http://www.trustlet.org/wiki/Repositories_of_datasets
Revolutions: Bike sharing in 100 cities. http://blog.revolutionanalytics.com/2013/07/bike-sharing-in-100-cities.html
SNAP. http://snap.stanford.edu/data/index.html
The Internet2 Observatory Data Collections. http://www.internet2.edu/observatory/archive/data-collections.html
The Kansas Event Data System. http://web.ku.edu/keds/
TunedIT. http://tunedit.org/repo
Web Archiving Service. https://archive-it.org/

Space-Embedded Networks

▶ Spatial Networks

Spam Detection

▶ Identifying Spam in Reviews

Spam Detection on Social Networks

Ninghao Liu and Xia Hu
Department of Computer Science and Engineering, Texas A&M University, College Station, TX, USA

Synonyms

Anomaly detection; Misinformation detection; Suspicious behavior detection

Glossary

Blacklist	A list of URLs that point to malicious contents or websites
Features	An object's original attributes, or manually extracted attributes based on predefined measures
Labeled Dataset	A dataset consisting of examples where we already know whether each of them belongs to spams/spammers or not
Reflexive Reciprocity	The phenomenon that a user is more likely to follow back those who have followed him/her, in social networks where the connections are unidirectional (Hu et al. 2013
Spam Detector	An automated tool for detecting spams or spammers, with the purpose of eliminating their influences
Spam	Unwanted, malicious, unsolicited content or behavior that affects normal social network users, directly or indirectly
Spammer	Spam originator
Sybils	A large number of fake entities created and controlled by a few malevolent users, with the purpose of dominating the online social environment and disrupting its key functions

Definition

Spam in social networks refers to the unwanted, malicious, unsolicited content or behavior, manifested in various ways including microblogs, messages, malicious links, fake friends, fraudulent reviews, etc. Spams have already pervaded in many traditional information systems such as instant messaging and email services. As social network services are increasingly popular recently, they have also become the targets of spammers. The spams have been observed in different platforms including social networking services (e.g., Facebook, Twitter, LinkedIn), e-commerce systems (e.g., Amazon, eBay), review websites (IMDb, GameRankings), as well as discussion forums.

The overwhelming of spams brings in significantly negative impact to the communication quality and user experience in social networks. The goal of spam detection is to develop effective and efficient techniques to automatically identify spams and their originators, in order to prevent legitimate users from being affected.

Introduction

Social networks such as Facebook, Twitter, WeChat, and Amazon have become important platforms in satisfying people's needs of social interactions, information acquiring, and dissemination. However, such unprecedented convenience also facilitates the activities of malevolent entities, such as spreading viruses, scamming, fabricating product reviews, and link farming. Spams are usually in the form of microblogs, messages, and website links, while they can also be hidden in synchronized activities such as sybil attacks. The motivation of spamming is the pursuit for fame and gain. Spammers can either directly receive economic returns (e.g., write fake reviews to promote their commodities) or gain popularity (e.g., build connections to many other users). According to a recent study (Stringhini et al. 2010) that, in 2008, around 85% of the social network users have been affected by spams including unwanted friend requests, advertisements, as well as links pointing to phishing websites or viruses. Although most spams can be recognized by normal users (Kanich et al. 2008), spammers can still survive and be profitable from such small portion of successes. It motivates us to investigate effective detection techniques to fight against them.

The spams deteriorate the quality of communication provided by social networks. They pollute the network environment and distort people's perception of online information. If someone gets exposed to too much unwanted information, the user experience will dramatically decreases, which will cause customer losses for the social service provider. Therefore, it is necessary for social platforms to develop algorithms to detect spams/spammers, sometimes even in real time.

The form of spams changes according to the specific type and nature of a social platform. The tactics taken by spammers may also vary along with the infrastructure and population composition of social networks. As the response, spam detectors should also adapt themselves to different scenarios. There will inevitably be some abnormal attributes that sell spams/spammers out and serve as the breakthrough for detectors to discover them, either in individuals or in groups. In the cases where labeled datasets are available, we can apply supervised learning methods to learn the patterns of spams/spammers and fight against them. Unfortunately, in some other cases, it may be very difficult to collect labeled data. Therefore, we have to resort to other approaches like unsupervised learning models and graph-based models which do not rely on such prior knowledge.

Key Points

The structure of this entry is organized as follows. First, we discuss four typical types of spams in social networks, including social spams, rumors, spam reviews, and link farming. Such categorization is based on the consequences that spams may bring to social network users and the system. Some of them may appear simultaneously in the same social media platform. Second, each type of spams, along with the originators, would indicate different features. Some features can be selected directly from attributes available in social media, while others are extracted from manually designed measurements. Here we analyze three types of features, including user profiles, content information, and behavioral patterns. Third, we briefly introduce the approaches for detecting spams and spammers, including blacklists, supervise models, unsupervised models, and graph-based models. Blacklists have already been widely applied in traditional email spam detection. When labeled datasets are available, we can use supervised learning methods, which include various classification models such as the Naïve Bayes model, Probabilistic Graphical Model, Logistic Regression, SVM, Decision Tree, as well as ensemble learning that combines a collection of individual classifiers for a more powerful one. Otherwise, we can apply unsupervised learning models or graph-based models. The key idea of the former is detecting anomalies after clustering objects, while the latter utilize the topological structure of the networks.

Historical Background

The prevalence of spam cannot be separated from the development of computer and communication techniques. With the emergence of Internet and search engine, people began to combat "spamdexing" (a compound of "spam" and "indexing") where spammers took improper means to promote the ranking of websites returned by the search engine. After email and blog services became popular, email spam and blog spam started to pervade information systems. Nowadays, as social media becomes an indispensable part of people's daily lives, spams also pose new challenges to the networking environment. According to Statista (http://www.statista.com/topics/1164/social-networks/), in 2013 the number of social network users has reached 1.61 billion, and it is estimated to be 2.34 billion by 2016. However, in Nexgates report (http://nexgate.com/tag/social-media-spam/), in the first half of 2013, the growth of social spams has been 355%, which is much faster than the growing rate of social network itself. Spamming has become not only a technical challenge but also a security crisis and social issue.

Spamming Scenarios

The spam in social networks, depending on the scenarios where it is generated, has different characteristics that lead to specific detection techniques. In this section, we divide spams into four categories: social spams, rumors, fake reviews, and link farming.

Social spam: It refers to the unwanted or malicious content posted on the social network platform. It includes scam malware, phishing website, commercial advertisement, and promotion information, etc. This type of spam may appear in various ways. For examples, it could be advertisements, the disguised posts that look normal but contain malevolent URLs, reply spams, as well as reposting spams that append junk information behind original posts (Benevenuto et al. 2010; Grier et al. 2010).

Rumors: Social networks enable real-time information dissemination to a wide range of users, which makes it an ideal platform for propagating news from the geographical origin. However, such nice characteristics also provide a desirable environment for spreading rumors, i.e., inaccurate or baseless news and information. Rumors in social network environment can be produced in the form of slander, political abuse, and misinformation in massive events (Mendoza et al. 2010; Ratkiewicz et al. 2011). For the last one, it could be the baseless information fabricated during major events like earthquake, sports game, and election, in order to arouse mass panic or to defame certain targets.

Spam reviews: Honest customer reviews and ratings can benefit e-commerce social media. However, driven by profit and fame, some malicious users put fake reviews to promote or demote certain products (Jindal and Liu 2008). Spam reviews distort the perception of potential customers on the products. In some other cases, spammers post irrelevant advertisements under the product, which impedes others from getting useful information.

Link farming: The popularity of a social network account largely depends on the size of its social circle. Link farming is defined as deliberately and aggressively establishing a large number of social links that are not supposed to naturally exist. Typical link farmers include sybils and social capitalists. Intuitively, the updates posted by celebrities are more easily to be circulated in social networks. Moreover, nowadays the content in social networks becomes accessible via search engines. High-profile news, as well as the posts created by famous people, are more likely to appear earlier in search engines. Therefore, in major social network platforms like Twitter and Facebook, some users are willing to pay fraudsters for a large number of followers or friends, in order to gain more popularity (Jiang et al. 2014).

Generally, these fake followers (or friends) are very socially inactive, i.e., except following (or making friend with) others, they do not show other behaviors. That is why they are also called "zombies." Sybil accounts boost spurious popularity, so it is a severe problem if they are utilized to support criminals. Those who are highly aggressive in promoting their influences (e.g., sending out many friend requests, following many people) are called social capitalists (Ghosh et al. 2012).

Feature Extraction

Spam and its generators will inevitably possess some abnormal attributes or features which manifest themselves in front of the detector. The process of constructing features can have a significant impact on the final detection performance. Details of how these features can be extracted and applied into spam/spammer detection are discussed below.

User Profiles

The profile of an account summarizes the basic information of its owner. Some representative attribute items include: (1) demographics: name, gender, age, geographical location; (2) user-contributed info: such as self-introduction, hobbies, education history, religious belief; and (3) social information: number of followers and followees, number of posts (Castillo et al. 2011). The roles of these attributes in spotting suspicious entities are illustrated as follows.

Checking demographics information is one of the most direct ways to discover dubious accounts. For instance, a simple but surprising fact is that sybil accounts may apply similar but anomalous naming rules, such as "Buy_BT27", "Buy_BT68," and "Buy_BT68," since they are created in batch by spammers (Jiang et al. 2015). Moreover, during a significant event, the geographical information of an account can help judge whether the posts generated by it can be trusted. Users who positioned at the site of an event are more likely to share first-hand information that is not available elsewhere (Starbird et al. 2012).

In social networks, links serve as the pathway for information propagation. In Twitter, by utilizing the reciprocity rule, spammers will strive to follow many users and hope they will follow back, which leads to a low follower/followee ratio (Benevenuto et al. 2010). In Facebook which only allows bidirectional links, spammers will send out a huge number of friend requests to try their luck that some incautious users will accept them. However, it still results in a low request/accept ratio (Stringhini et al. 2010). Moreover, different from zombie accounts that are totally silent, social spammers may consistently post content in the timeline. However, since these posts are usually generated by a central controller, sybils do not interact with other users. Therefore, the number of replies or mentions of a spammer is likely to be small.

Content Information

Content information refers to the characteristics of the posts and messages on social networks. It could be social platform artifacts like "#" and "@," textual words or implicit information like sentiment and syntax. The Vector Space Model (VSM) can be applied where each entry corresponds to a symbol, a word or a piece of information. In this case, a text can be represented as a vector $\mathbf{x} = [x_1, x_2, \ldots, x_M]^T$ where the length M is a predefined parameter.

Artifacts in social platforms are designed to facilitate information propagation in the network. However, they also provide convenience for spammers to spread rumors and advertisements (Lee et al. 2012). It is a common tactic that spammers use "#" hashtag in the posts to mark certain topics and attract those who are interested, but attach totally unrelated URLs to lead the viewers to malicious websites. Spammer may also add many "@"s in the posts to direct unwanted information to other users.

It is found that posts containing clear sentiment words are more likely to be credible. Tools such as "Stanford Parser" for obtaining Part-Of-Speech (POS) tags and Linguistic Inquiry and Word Count (LIWC) are useful for extracting sentiment features from the posted text. We may also use the lexicons constructed by experts for sentiment analysis (Hu et al. 2014a; Ratkiewicz et al. 2011).

Meanwhile, similar to how we filter out traditional junk emails, spam words such as “reverses aging” and “great deal” are important indicators of suspicious content in social networks (Rubin 2012).

In some situations, we are not able to judge the credibility of individual posts. For example, to detect rumors, we may want to collect multiple mutually correlated microbloggings and messages, evaluate the consistency across them, and then find out the baseless ones. This is due to the fact that questionable information is very likely to encounter questionings.

Behavioral Patterns

Statistical patterns of an account’s behavior are powerful in spotting suspicious entities in social network. Here “behaviors” refer to all kinds of online activities taken by users. The behavior pattern is different from the content information discussed before, since the former focuses on the user side, while the latter one studies user-generated content.

Behavioral patterns are useful in many complicated detection scenarios. For examples, in e-commercial systems, the rating deviation (i.e., the deviation of a user’s ratings from those of other users) can expose spammers since they usually promote or demote products on purpose (Jindal and Liu 2008). The burst of reviews from a single customer is also a good indicator of suspiciousness, since it is easier for a spammer to write many fake reviews in a short period of time (Fei et al. 2013). Besides spotting individual spammers, behavior information can also help capture spammer groups. For examples, sybil accounts usually take similar actions with high synchronicity (Jiang et al. 2015). Members from fake reviewer groups are more likely to give similar ratings on certain products with close timestamps. The behavior patterns can help the detector design features and algorithms to reveal suspicious entities.

Detection Approaches

Besides providing the interface for users to report spams by themselves, it is also necessary to automatically detect spams/spammers existed in social network platforms. In the adversarial settings between spammers and detectors, the detection methods have been consistently evolving in the last few years.

Blacklists

In the early stage, the most direct way for discovering spam content is to construct a blacklist recording malicious URLs and detect posts with these URLs as spams. This simple method suffers from several disadvantages. First, there is a time delay before a URL is added to the blacklist. Second, the URLs shortening service may provide convenience for spammers to obfuscate their domains. According to Thomas et al. (2011), however, it was found that only 8% of tweet spam URLs were captured by blacklists.

Supervised Learning Methods

Spams in social networks are of various forms and they may change over time, so some rules must be made in order to filter them out. However, manually designing and refining these rules are very time-consuming and error-prone. With the development of machine learning (ML) methods, we are able to build classifiers that work automatically to detect spams/spammers (Castillo et al. 2011). In the scheme of supervised learning, a labeled training dataset is applied to build the classifier. One entry in the training dataset consists of features representing an object and a label indicating if the object is spam/spammer or not. The way of extracting features is discussed in the previous section. Different types of features can be combined to jointly make decisions (Hu et al. 2014b). After training, the classifier can be used to tell whether new entries represent spam objects. To evaluate the performance of the classifier, we can use another testing dataset. Some widely used machine learning classifiers include Bayesian classifier, logistic regression, support-vector machine, decision trees, etc.

Suppose the network object i is represented as $\mathbf{x}^i$ using VSM and there are k classes to label it, so that the label $0 \le c \le k - 1$. In the scenario of spam classification, k is set to 2 where $c = 1$ for being spam and $c = 0$ for normal. The probability $P(C^i = c|\mathbf{x}^i)$ is calculated according to the Bayes rule,

$$P\left(C^i = c \middle| \mathbf{x}^i\right) = \frac{P\left(\mathbf{x}^i \middle| C^i = c\right) P\left(C^i = c\right)}{P\left(\mathbf{x}^i\right)} \quad (1)$$

The object is labeled into class $c = \text{argmax}_c\, P(C^i = c|\mathbf{x}^i)$. Since $P(\mathbf{x}^i)$ does not affect the relative magnitude of $P(C^i = c|\mathbf{x}^i)$ across different c, it can be ignored when making the decision. However, the number of all cases of $\mathbf{x}^i$ could be prohibitively large. In practice, a widely adopted assumption is that features are conditionally independent with each other given the class label c, so that:

$$P\left(\mathbf{x}^i \middle| C^i = c\right) = \prod_{m=1}^{M} P\left(x_m^i \middle| C^i = c\right). \quad (2)$$

Naïve Bayes classifiers have been shown to be powerful in many spam detection problems. However, spammers may adopt the Bayesian poisoning technique, i.e., fabricating features that looks normal, to degrade the effectiveness of the classifier. Naïve Bayes model is one of the simplest examples of Probabilistic Graphical Model (PGM). Complex PGMs are able to represent more intricate relationships among features and entities. They have already been applied into various spam detection problems such as credibility assessment in health communities (Mukherjee et al. 2014) and review spam discovery (Rayana and Akoglu 2015). The logistic classifier is another model for spam detection (Jindal and Liu 2008). Its parametric form can be implied by that of the Naïve Bayes classifier, but assuming the Gaussian distribution for $P\left(x_m^i \middle| C^i = c\right)$.

Another class of machine learning method that can be used for spam detection is SVM. To start with, it is a binary linear classifier.

$$f\left(\mathbf{x}^i\right) = \mathbf{w}^T \mathbf{x}^i + b \quad (3)$$

where $C^i = 1$ if $f(\mathbf{x}^i) > 0$ and $C^i = 0$ otherwise. The parameters $\mathbf{w}$ and b together determine the decision boundary. SVM learns the optimal classifier in terms that it maximizes the "margin," which is the minimum distance between the decision boundary and any training data samples. Incorporating "kernels" to SVM can adapt the model to nonlinear classification problems (Benevenuto et al. 2010; Ratkiewicz et al. 2011).

Spam detection involves a bunch of decision-making processes based on the features of entities. Decision tree, with the flowchart-like structure, is a good match to the problem (Castillo et al. 2011). Each internal node in the tree represents a decision-making step on a feature, each branch directs to the outcome of the decision, and the leaf node tells the label assigned. Given the training dataset, we can build decision trees to embody the spam judgement procedures. When fed with a new data example $\mathbf{x}$, the final leaf node to which it goes through would decide if $\mathbf{x}$ represents a spam.

Besides using a single classifier, assembling multiple classifiers together usually produces a stronger one. Several classifiers can vote on the final label of the new data example, which is more reliable than individual ones. For example, random forest, where each component classifier is a decision tree, has been shown successful in various applications (Lee et al. 2010).

Unsupervised Learning Methods

There are some scenarios where it is difficult to obtain enough representative labeled data for building classifiers. For example, in the sybils detection problem, it is hard for us to know in advance what accounts are sybils who do not represent real persons. It is true that sybil accounts do not provide detailed profile information, but so do some normal users. Moreover, these social bots usually do not leave any post in the timeline, so that it impossible to extract features from this source. We may create sybils by ourselves and treat them as labeled data. However, it is not guaranteed that our rule for creating these virtual bots is the same as the one adopted by those real spammers. One solution to get rid of the dilemma is to apply unsupervised learning methods, such as clustering models that only utilize the correlation among objects. First, social network objects are mapped into clusters. Then, those individuals or small clusters that deviate from the major distribution are regarded as spams/spammers. Sometimes this is also called anomaly detection.

Features related with social connections and behavioral patterns are very useful for this kind of tasks. In order to enlarge the impact, a fraudster usually controls a large number of social accounts, towards some common goals. These fake accounts will inevitably manifest similar activity patterns. Features that capture the synchronization of actions taken by online users are good indicators revealing the orchestration of these accounts. Suppose we define K features to form the new space, then each user n can be represented by a feature vector $u_n \in \mathbf{R}^K$ to be mapped into the space. Effective features will map spammers to separate regions separate from those of normal users (Jiang et al. 2014, 2015). After that, we can simply apply anomaly detection techniques to pick them out.

Another class of clustering algorithm is spectral method. Given the adjacency matrix $A \in \mathbf{R}^{N\times N}$ of a network, we apply singular value decomposition as $A = U\Sigma_K V^T$, where $U,V \in \mathbf{R}^{N\times K}$, $\Sigma_K \in \mathbf{R}^{K\times K}$. Here U and V are called left and right singular vectors. The row vector $(u_{n,1}, u_{n,2},\ldots,u_{n,K})$ represents the coordinate of node n in the left space, while $(v_{n,1},v_{n,2},\ldots,v_{n,K})$ corresponds to the coordinates in the right space, where $u_{n,i} = U(n,i)$ and $v_{n,i} = V(n,i)$. Then we are able to create a bunch of scatter plots $(u_{n,i},u_{n,j})$ or $(v_{n,i},v_{n,j})$ in 2-dimensional space once we have picked i and j. It is found that abnormal users are located far away from normal ones in the plots, so that the problem of spammer detection is also turned into the density-based anomaly detection (Ying et al. 2011).

Graph-Based Algorithms

To utilize the interdependency to detect spamming groups, we can also resort to graph-based algorithms leveraging the topology of social network graphs. The assumption behind this class of models is that the community of spammers is separate from that of normal users. Normal users are less likely to build connections with unsolicited users, while spammers themselves may connect to support each other. The reason for spammers to be collaborative is that if they do not connect with each other, then the number of their followers or friends will be very small, which is a strong sign of being abnormal.

An important intuition behind graph-based algorithms is “like attracts like,” i.e., normal users tend to connect with each other and so do spammers. This is also the key idea of the ranking algorithms such as PageRank and HITS, which have shown to be successful in building search engines. Some representative examples of the algorithms are discussed as following. Start from a small set of predetected seed spammers, the malicious-relevance scores of all network members are deduced though iterative propagation (Yang et al. 2012). From another perspective, the loose interaction between spammer group S and the rest of the graph members $\overline{S}$ will lead to a low “cut” value $\varphi(S)$. Given a graph $G = (V, E)$ and a subgraph S, the cut $\varphi(S)$ is defined as:

$$\varphi(S) = \frac{\sum_{i\in s, j\in \overline{S}} a_{i,j}}{\min\left(a(S), a\left(\overline{S}\right)\right)} \tag{4}$$

$$a(S) = \sum_{i\in S, j\in V} a_{i,j} \tag{5}$$

where $a_{i,j} = 1$ if node i and j are connected, and otherwise $a_{i,j} = 0$. A consequence of low $\varphi(S)$ is that it is more likely to take long times for the random walks starting from $i \in S$ to converge to a steady state. The seminal work for discovering such low-cut groups can be found in (Yu et al. 2006).

Besides spotting spammer groups, some algorithms target at discovering unreliable events in social media. Gupta et al. proposed graph-based approaches to automatically assess the credibility of events in Twitter, based on the idea that multiple reliable sources of information regarding to the same event should be coherent (Gupta et al. 2012).

Key Applications

The techniques introduced above can be applied in detecting spam content and originators. The originators may be discovered in individuals or in groups, depending on the specific scenario and the detection method. Furthermore, the techniques can also be used to solve traditional web-spam or email-spam problems or be extended to computer/network security problems. Some basic ideas behind can be borrowed to the anomaly detection, which may be seen as a superset of spam detection.

Future Directions

The volume of data collected from the real-world social networks is becoming larger and larger. The availability of big data puts new demands on the scalability of the detection model. At the same time, it is more common for network objects to be described from multiple aspects of different natures. The resultant dataset is usually more expressive and contains more information due to the potential interaction between several aspects. The increasing availability of heterogeneous data may indicate another direction for designing detectors.

The adversarial relationship between spammers and detectors results in the fact that the competition between them will continue. The ever evolving nature of spammers will post new challenges for the detector, requiring it to adapt quickly to the varying spamming tactics. Meanwhile, the social network itself is not a static system. Building new models whenever external condition changes would be costly and time-consuming, so designing dynamic detection models is one of the future directions to be done (Hu et al. 2014c). Online learning techniques are highly correlated with this type of models. Moreover, in many cases the data come in the form of streams, so the existing model can update itself to fit the new data, which further justifies the need to develop dynamic models.

Cross-References

- ▶ Dark Side of Online Social Networks: Technical, Managerial, and Behavioral Perspectives
- ▶ Identifying Spam in Reviews
- ▶ Online Social Network Phishing Attack
- ▶ Spam Detection: E-mail/Social Network
- ▶ Trust in Social Networks

References

Benevenuto F, Magno G, Rodrigues T, Almeida V (2010) Detecting spammers on twitter. In: Collaboration, electronic messaging, anti-abuse and spam conference (CEAS), vol 6, pp 12

Castillo C, Mendoza M, Poblete B (2011) Information credibility on twitter. In: Proceedings of the 20th international conference on world wide web. ACM, New York, pp 675–684

Fei G, Mukherjee A, Liu B, Hsu M, Castellanos M, Ghosh R (2013) Exploiting burstiness in reviews for review spammer detection. ICWSM 13:175–184

Ghosh S, Viswanath B, Kooti F, Sharma NK, Korlam G, Benevenuto F, Ganguly N, Gummadi KP (2012) Understanding and combating link farming in the twitter social network. In: Proceedings of the 21st international conference on world wide web. ACM, New York, pp 61–70

Grier C, Thomas K, Paxson V, Zhang M (2010) @ spam: the underground on 140 characters or less. In: Proceedings of the 17th ACM conference on computer and communications security. ACM, New York, pp 27–37

Gupta M, Zhao P, Han J (2012) Evaluating event credibility on twitter. In: SDM. SIAM, Anaheim, pp 153–164

Hu X, Tang J, Gao H, Liu H (2014a) Social spammer detection with sentiment information. In: 2014 I. E. international conference on data mining (ICDM). IEEE, Washington, DC, pp 180–189

Hu X, Tang J, Liu H (2014b) Leveraging knowledge across media for spammer detection in microblogging. In: Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval. ACM, New York, pp 547–556

Hu X, Tang J, Liu H (2014c) Online social spammer detection. AAAI, In, pp 59–65

Hu X, Tang J, Zhang Y, Liu H (2013) Social spammer detection in microblogging. In: IJCAI, vol 13. Citeseer, pp 2633–2639

Jiang M, Beutel A, Cui P, Hooi B, Yang S, Faloutsos C (2015) A general suspiciousness metric for dense blocks in multimodal data. In: 2015 I.E. international conference on data mining (ICDM). IEEE, Washington, DC, pp 781–786

Jiang M, Cui P, Beutel A, Faloutsos C, Yang S (2014) Catchsync: catching synchronized behavior in large directed graphs. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 941–950

Jindal N, Liu B (2008) Opinion spam and analysis. In: Proceedings of the 2008 international conference on web search and data mining. ACM, New York, pp 219–230

Kanich C, Kreibich C, Levchenko K, Enright B, Voelker GM, Paxson V, Savage S (2008) Spamalytics: an empirical analysis of spam marketing conversion. In: Proceedings of the 15th ACM conference on Computer and communications security. ACM, New York, pp 3–14

Lee K, Caverlee J, Kamath KY, Cheng Z (2012) Detecting collective attention spam. In: Proceedings of the 2nd joint WICOW/AIRWeb workshop on web quality. ACM, New York, pp 48–55

Lee K, Caverlee J, Webb S (2010) Uncovering social spammers: social honeypots+ machine learning. In: Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval. ACM, New York, pp 435–442

Mendoza M, Poblete B, Castillo C (2010) Twitter under crisis: can we trust what we rt? In: Proceedings of the first workshop on social media analytics. ACM, New York, pp 71–79

Mukherjee S, Weikum G, Danescu-Niculescu-Mizil C (2014) People on drugs: credibility of user statements in health communities. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 65–74

Ratkiewicz J, Conover M, Meiss M, Gonçalves B, Flammini A, Menczer F (2011) Detecting and tracking political abuse in social media. ICWSM 11:297–304

Rayana S, Akoglu L (2015) Collective opinion spam detection: bridging review networks and metadata. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 985–994

Rubin K (2012) The ultimate list of email SPAM trigger words

Starbird K, Muzny G, Palen L (2012) Learning from the crowd: collaborative filtering techniques for identifying on-the-ground twitterers during mass disruptions. In: Proceedings of 9th international conference on information systems for crisis response and management (ISCRAM)

Stringhini G, Kruegel C, Vigna G (2010) Detecting spammers on social networks. In: Proceedings of the 26th annual computer security applications conference. ACM, New York, pp 1–9

Thomas K, Grier C, Song D, Paxson V (2011) Suspended accounts in retrospect: an analysis of twitter spam. In: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. ACM, New York, pp 243–258

Yang C, Harkreader R, Zhang J, Shin S, Gu G (2012) Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In: Proceedings of the 21st international conference on world wide web. ACM, New York, pp 71–80

Ying X, Wu X, Barbará D (2011) Spectrum based fraud detection in social networks. In: 2011 I.E. 27th international conference on data engineering (ICDE). IEEE, Washington, DC, pp 912–923

Yu H, Kaminsky M, Gibbons PB, Flaxman A (2006) Sybilguard: defending against sybil attacks via social networks. ACM SIGCOMM Comput Commun Rev 36 (4):267–278

Spam Detection: E-mail/Social Network

Cailing Dong and Bin Zhou
Department of Information Systems, University of Maryland, Baltimore County, Baltimore, MD, USA

Synonyms

Junk e-mail; Social spam; Unsolicited bulk e-mail

Glossary

Spam	Unsolicited, unwanted message intended to be delivered to an indiscriminate target, directly or indirectly, notwithstanding measures to prevent its delivery
Spammer	Originator of spam message
Spam Filter	An automated tool that is built to detect spam message with the purpose of preventing its delivery
Whitelist	A list of contacts whose e-mails should be delivered
Blacklist	A list of contacts whose e-mails are deemed to be spam
Classifier	A model that identifies which of a set of categories an object belongs to

Definition

Spam generally refers to "unsolicited, unwanted message intended to be delivered to an indiscriminate target, directly or indirectly, notwithstanding measures to prevent its delivery" (Cormack 2008). While e-mail spam is the mostly widely recognized form of spam, spam actually pervades many existing information systems and social media, including instant messaging (Paulson 2004), blogs (Abu-Nimeh and Chen 2010), newsgroups and forums (Shin et al. 2011), and online social media (Jin et al. 2011). Spam also exists in web search, where search engines are used as a delivery mechanism for web spam (Gyöngyi and Garcia-Molina 2005).

The overwhelming of spam messages in existing information systems and social media severely deteriorates the quality of communication. The objective of spam detection is to develop effective and efficient anti-spam techniques with the purpose of preventing the delivery of spam messages.

Introduction

Regardless of various forms of spam in reality, sending spam messages is essentially profit-driven activity. Spammers, the originators of spam message, intend to deliver the information to a large volume of recipients. Spam messages often contain advertising for commercial products, URL links to promoted websites which may serve as means of adult content dissemination and phishing attacks, or even computer malwares that are specifically designed to hijack the recipient's computers. Although some forms of spam can be identified whenever the message is delivered and viewed by the receipts, spammers can still be profitable even only a very small fraction of recipients take responses to spam messages. Due to the fact that the operating cost of sending spam messages is substantially cheap and the barrier to entry is quite low, the volume of spam has been consistently increasing in the past several years. According to the 2009 report by Ferris Research (Jennings 2009), the worldwide financial losses caused by spam were estimated to be $130 billion in 2009, a 30% increase over the 2007 estimates. Comparing to the estimated figures in 2005, the total losses were increased by 160% in 2009.

Spam dramatically deteriorates the quality of communication. From the users' point of view, users are victims affected directly by spam. Not only users' material wealth but also their personal information could be under risk to spammers. From the system providers' point of view, they are forced to waste a significant amount of computational and storage resources for spam messages. Moreover, if a user receives many spam messages, his/her trust in the system can be drastically weakened, which inevitably makes the user switch from the current system provider to another competitor. Therefore, both users and system providers have strong incentive to wipe out spam thoroughly.

Historical Background

The term "spam" is named for Spam luncheon meat by way of a Monty Python sketch where Spam is depicted as ubiquitous and unavoidable. Among various forms of spam in the literature, e-mail spam is the most common one. Along with the vigorous development of the Internet since the mid-1990s, e-mail becomes a popular communication and information exchanging method. E-mail spam started to be a serious problem since then, and it grew exponentially over the following years. Nowadays, spam comprises the vast majority of e-mail messages sent daily. It is reported that 78% of the e-mails are spam (Fletcher 2009). Due to such large impacts, e-mail spam becomes not only a technical challenge but also a legal crisis and political issue. Myriad technological and legal-based attempts have been developed to combat e-mail spam.

Meanwhile, the user-centered design feature of Web 2.0 further involves people in a rapid information sharing and propagation era, which brings the prosperity of many online social websites. Social websites are designed to support and foster various social interactions, which on the other hand heavily rely on users for content contribution and distribution. However, such interactive and dynamic features also provide fertile soil for spam. Spammers in social websites disguise their spam messages as links, content, video, audio, and executable files. Unlike traditional e-mail spam which usually comes from strangers, this new form spam, also named as "social" spam, often appears to be from a "friend" in social websites. Spammers find social websites alluring because they can broadcast spam messages through a chain of trusted sources and target at a large number of users. According to Fowler et al. (2012), 4% of the content shared on Facebook is spam. Making matters even worse, a statistical report from Facebook (Fowler et al. 2012) indicates that the volume of social spam is growing much faster than its user base.

Foundations

The majority of existing efforts to combat spam are based on filtering spam messages using spam filters. In this section, we first describe the general framework of spam detection using spam filters and then discuss some representative spam detection solutions for e-mail spam and social spam, respectively.

A General Framework

The success of spam detection relies on spam filters, which are automated tools that are built to detect spam with the purpose of preventing its delivery. For every effective spam filter, the core component is a spam classifier which categorizes whether a specific message is spam or not. The decision of spam classification using a spam filter is often made based on different pieces of information. For example, in order to construct an effective e-mail spam filter, the content of the e-mail messages and the characteristics of the e-mail sender and the e-mail receiver are usually considered. In addition, collaborative knowledge such as the feedback of other receipts receiving similar e-mails is also valuable for building the spam filter. In many cases, some external pieces of information (e.g., spam repositories) provided by some third parties are also considered.

Once a spam classifier is constructed, a specific message can be categorized as spam or not. A simple way to obtain the spam categorization results is using a binary classifier. In a binary classifier, a message is either labeled as spam or non-spam. Although this solution is simple, it suffers from the lack of adaptability. Users have little control in the spam detection process. A more common way is to use probabilistic classifiers which can provide more informative indications on how likely the spam classifier considers the message to be spam. For example, some spam classifiers can calculate a spamicity score for each message. The spamicity score is in the range [0,1]. The larger the spamicity score, the more likely the spam classifier considers the message to be spam. In practice, a spamicity threshold value is often configured to filter spam messages. Different from the binary spam classifier, users are able to adjust the threshold value so as to capture different spam detection scenarios.

Spam filters automatically filter those messages that are labeled as spam. For example, spam e-mails are automatically placed into the junk folder in each user's e-mail account; spam web pages are automatically removed from the web search results; spam information in online social media is automatically flagged and is prohibited to be propagated through social networks.

E-mail Spam Detection

The methods for e-mail spam detection have been evolved continuously in the past years. In the early stage, spam filter for e-mail spam detection is mainly based on receiver's judgements. For example, users have the opportunity to handcraft several logical rules and guidelines to filter spam e-mails. A common practice is to maintain a whitelist and a blacklist in each user's e-mail account. A whitelist refers to a list of contacts whose e-mails should be delivered. Oppositely, a blacklist refers to a list of contacts whose e-mails are deemed to be spam. This solution is acceptable when the volume of contacts and e-mails is low. However, this solution becomes problematic when the volume gets larger and larger. In addition, the success of manually handcrafted spam filter relies on user's judgements. The assumption that users are savvy enough to construct robust spam filtering rules is questionable. To make matters worse, when spam e-mails change over time, it becomes an even time-consuming and error-prone process for users to constantly turn and refine those spam filtering rules.

To address the problems with the manual construction of spam filtering rules, latest spam filters for e-mail spam detection are able to automatically adapt to the changing characteristics of spam e-mails over time. Machine learning algorithms have been widely adopted to build robust and adaptive spam filters. Since those spam filters are built directly from user's e-mail repository, they are able to be personalized to meet particular characteristics of each user. In other words, these spam filters are tailored specifically to meet each individual's requirements on spam judgements.

Among many machine learning-based spam filters, the one based on Bayesian classifier is most popular and widely adopted (Sahami et al. 1998). A Bayesian classifier is a probabilistic classifier. It uses Bayesian inference to calculate a probability which indicates how likely an e-mail message is spam. The motivation of using Bayesian classifier for e-mail spam detection is that particular information (e.g., words in the e-mail) has particular probabilities to occur in either spam or legitimate e-mails. If all such probabilities are calculated, they can be used to compute the overall probability that a specific e-mail message with a particular set of information in it belongs to either spam category or non-spam category.

In practice, the Vector Space model in which each dimension corresponds to a given word in the entire corpus is often adopted. Thus, each individual message can be represented as a binary vector denoting which words are present or absent. Some other pieces of information, for example, domain-specific properties, can also be incorporated into the representation. Specifically, an e-mail message is represented as a vector of n features x = .(x_1, x_2,. . ., x_n). Each feature $X_i(1 \le i \le n)$ represents a specific piece of information, and $x_i(1 \le i \le n)$ is the value pertaining to feature X_i.

Assume that there are m different classes to categorize e-mail messages, each is denoted as $c_j(1 \le j \le m)$. Consider a specific e-mail message x, the Bayesian classifier can calculate the probability $P(C = c_j \mid X = x)$ for each possible class $c_j(1 \le j \le m)$. The calculation is achieved according to the well-known Bayes' theorem, that is,

$$P(C=c_j|\mathbf{X}=\mathbf{x}) = \frac{P(\mathbf{X}=\mathbf{x}|C=c_j)P(C=c_j)}{P(\mathbf{X}=\mathbf{x})}. \tag{1}$$

In Eq. 1, the calculation $P(X = x|C = c_j)$ is often impractical without imposing some independence assumptions. Thus, the Naive Bayesian classifier is often adopted for the calculation. Given the class variable C, the Naive Bayesian classifier assumes that each feature X $_i$ is conditionally independent of every other feature. As a result, we have

$$P(\mathbf{X} = \mathbf{x}| C = c_j) = \prod_i P(X_i = x_i| C = c_j). \tag{2}$$

Spam filter using Bayesian classifier has been shown a very powerful technique for dealing with e-mail spam. It can tailor itself to the specific needs of individual users and provides low false positive spam detection rates. However, due to the intrinsic problem of Bayesian classifier, the constructed spam filter may also be susceptible to Bayesian poisoning, a technique used by spammers in an attempt to degrade the effectiveness of spam filters.

In the recent decade, many efforts have been devoted to constructing effective and efficient machine learning-based spam filters. Many existing studies mainly focus on two critical issues in building a spam filter: one is feature selection, that is, selecting a subset of relevant features for building robust spam filters; the other is classifier construction, that is, using different machine learning methods to "learn" robust spam classifiers from training data.

As many pieces of information can be extracted from e-mail messages but not all of them are useful for building spam filters, the feature selection problem needs to find the best subset of the available features. The concept of "best" may rely on different factors including the number of selected features, the effectiveness of the trained spam classifier, and the tractability of the algorithm to perform the selection process. In general, given n features of e-mail messages, a straightforward but prohibitively inefficient solution is to examine all 2^n subsets of n features and choose the one which achieves the highest spam detection result. In practice, some greedy heuristics are adopted for ranking features in decreasing order based on some characterizations of their usefulness for spam detection. For example, Sebastiani (2002) analyzed several heuristics relying on statistics such as term frequency and information gain for ranking and selecting features. Regardless of particular statistics of different features, selecting the optimal feature set is always a challenge.

The process of feature selection should not be considered separately from the process of classifier construction. Different machine learning algorithms have been considered for constructing spam filters. For example, the k-Nearest Neighbors Classifier (Firte et al. 2010) builds on top of the k-NN algorithm and classifies a message according to the classes of its nearest neighbors in the training data. Artificial neural networks are also applicable for constructing spam filters. In particular, the algorithms such as perceptron and multilayer perceptron (Tran et al. 2008) have been shown quite successful for filtering e-mail spams. Several recent studies also considered the application of SVM for spam detection. The motivation of the SVM classification (Zhang et al. 2004) is to find a separation boundary which can correctly classify training samples. Different from the perceptron algorithm, the SVM-based approach tries to find a special maximal margin separating hyperplane such that the distance to the closest training sample is maximal.

In practice, there exist some types of e-mail messages that cannot be clearly categorized as either spam or non-spam. Such examples include newsletters and legitimate advertisements. These types of e-mail messages are usually regarded as gray mail (Chang et al. 2008). Detecting gray mail introduces more challenges. Even an optimal spam filter could inevitably perform unsatisfactorily on gray mail. Chang et al. (2008) systematically studied the problem of gray mail detection and concluded that user preferences are needed to be considered. The experimental results in Chang et al. (2008) indicate that e-mail messages which are labeled differently in the training data are the most reliable source for learning a gray mail detector.

Social Networks Spam Detection

Online social websites have different characteristics comparing to e-mail systems. Some spam filtering techniques for e-mail spam detection may be useful to detect spam in social networks as well; however, some particular requirements of social spam detection need to be considered. Following are the four most important features of online social websites, which to some extent differentiate the characteristics of social spam detection compared to traditional e-mail spam detection:

- Existence of one managing entity. In online social websites, there exists an entity who manages and maintains the system, defines the system policy, and determines the privileges of participated users.
- Well-defined social interactions. Users have very close interactions with the social websites to contribute social contents. Meanwhile, social websites also provide some functionalities to share and distribute users' contents. However, the available interactions of participated users are constrained in the system.
- Unique identifier. In online social websites, each user has to maintain a unique identifier or a personal profile. This unique identifier is associated with each user's interaction in the websites.
- Multiple views of information access. Users in online social websites have multiple views to get access to the available contents.

Users are a key component in social spam detection. Social spam detection has several unique challenges. First, unlike many e-mail systems, the managing entity and restricted interactions in online social websites provide the opportunity to prevent spam effectively even before its emergence. For example, by defining appropriate terms of service and adjusting the trade off between users' privileges and the information flow rate, social websites are able to keep spam in the prevention stage. Second, due to the unique identifier in social websites, the origins of social spam can be controlled since users' interactions are tied to a specific identifier. Third, multiple views of information access lead to different snapshots of available contents in the websites. Therefore, social spam detection should consider all the possible spam tricks and various relations among them. Last but not the least, social websites contain large population and their social interactions, which makes information propagation much faster. This results in increasing and dynamic evolution of social spam. Inevitably, scalability and timely detection requirements become key issues in social spam detection.

Several popular anti-spam strategies for online social websites, named as detection, demotion, and prevention, are analyzed thoroughly in Heymann et al. (2007). Detection is made based on the predefined discriminative features extracted from given spam and non-spam instances. This is similar to e-mail spam detection. The features used for building spam filters are mainly extracted from contents and topological structure of social networks. In addition, the analysis of users' social behavior and domain-specific features (e.g., features extracted from figures or videos) are often largely considered. Different from e-mail spam detection, demotion and prevention are also considered in social spam detection. The demotion strategy adopts rank-based methods to downgrade the prominence of contents that are deemed to be spam. In some situations, due to the fast propagation and evolution capabilities of social websites, decreasing the rankings of spam messages might not be enough.

Many current techniques of social spam detection largely depend on the set of features extracted from user behaviors and social interactions. Although such features are useful for social spam detection, there is always a considerable time delay until the spam is successfully identified. The fast information exchanging rate in social websites requires a real-time framework to combat spam. To achieve this goal, prevention-based strategy to identify social spam becomes quite useful (Irani et al. 2010). Once user profiles are created in social websites, some features are directly extracted from the static profile contents. The motivation is to identify the potential spammers in the early stage, even before the creation and propagation of spam messages in social networks. A popular solution is to treat social spam detection as an adversarial classification problem (Dalvi et al. 2004). However, this prevention-based solution may be vulnerable. In practice, this technique is often used as a filter even before many sophisticated spam detection techniques are employed. For example, user profiles that are deemed to be spammers are treated as gray profiles. These user profiles need particular attentions for further analysis to support spam detection.

Some recent studies (Boykin and Roychowdhury 2005) proposed an integrated framework of social spam detection and e-mail spam detection and applied social network analysis for e-mail spam detection. The algorithm proposed in Boykin and Roychowdhury (2005) analyzes "From," "To," Cc" and "Bcc" fields of the e-mail headers so as to construct a network representing social relations of different users. The foundation is based on the fact that the underlying e-mail social networks are useful for judging the trustworthiness of users. For example, the trust can be measured based not only on how well a user knows a specific person but also on how well the other users in the e-mail network know that person. Once the social network of e-mail communications is built, an automated anti-spam tool can exploit the properties of social networks to distinguish spam messages from non-spam ones.

Conclusion

There is an adversarial relationship between spam and anti-spam techniques. In recent years, machine learning-based spam detection approaches, the probabilistic classifier-based spam filter in particular, have been widely applied to detect various forms of spam. However, the performance of anti-spam techniques is still far from perfect. The creativity and efforts of spammers who manage to violate laws and social norms to deliver spam messages will provide a continuing challenge for developing anti-spam techniques. Developing robust and adaptive anti-spam techniques is a long-term strategy to combat spam.

Cross-References

- ▶ Dark Side of Online Social Networks: Technical, Managerial, and Behavioral Perspectives
- ▶ Ethics of Social Networks and Mining
- ▶ Online Social Network Phishing Attack
- ▶ Social Engineering
- ▶ Trust in Social Networks

References

Abu-Nimeh S, Chen T (2010) Proliferation and detection of blog spam. IEEE Secur Priv 8(5):42–47

Boykin PO, Roychowdhury VP (2005) Leveraging social networks to fight spam. Computer 38(4):61–68. https://doi.org/10.1109/MC.2005.132

Chang M, Yih W, McCann R (2008) Personalized spam filtering for gray mail. In: Proceedings of the fifth conference on email and anti-spam (CEAS '08), Mountain View

Cormack GV (2008) Email spam filtering: a systematic review. Found Trends Inf Retr 1(4):335–455

Dalvi N, Domingos P, Mausam SS, Verma D (2004) Adversarial classification. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, KDD &apos;04. ACM, New York, pp 99–108. http://doi.acm.org/10.1145/1014052.1014066

Firte L, Lemnaru C, Potolea R (2010) Spam detection filter using KNN algorithm and resampling. In: Proceedings of the 6th international conference on intelligent computer communication and processing (ICCP '10), Cluj-Napoca

Fletcher D (2009) A brief history of spam. Time. http://www.time.com/time/business/article/0,8599,193 3796,00.html

Fowler GA, Raice S, Efrati A (2012) Spam finds new target: Facebook and Twitter build up their defenses as hackers attack social networks. The Wall Street Journal

Gyöngyi Z, Garcia-Molina H (2005) Web spam taxonomy. In: First international workshop on adversarial information retrieval on the Web (AIRWeb '05), Chiba

Heymann P, Koutrika G, Garcia-Molina H (2007) Fighting spam on social web sites: a survey of approaches and future challenges. IEEE Internet Comput 11(6):36–45

Irani D, Webb S, Pu C (2010) Study of static classification of social spam profiles in MySpace. In: Proceedings of the fourth international conference on weblogs and social media, Washington, DC

Jennings R (2009) Cost of spam is flattening: our 2009 predictions. Ferris research. http://email-museum.com/2009/01/28/cost-of-spam-is-flattening-our-2009-predictions/

Jin X, Lin CX, Luo J, Han J (2011) Socialspamguard: a data mining-based spam detection system for social media networks. PVLDB 4(12):1458–1461

Paulson LD (2004) Spam hits instant messaging. Computer 37(4). IEEE Computer Society Press, Los Alamitos

Phuoc TT, Po-Hsiang T, Tony J (2008) An adjustable combination of linear regression and modified probabilistic neural network for anti-spam filtering. In: Proceedings of the 19th international conference on pattern recognition, Florida

Sahami M, Dumais S, Heckerman D, Horvitz E (1998) A bayesian approach to filtering junk E-mail. In: Learning for text categorization: papers from the 1998 workshop, AAAI Technical Report WS-98-05, Madison. citeseer.ist.psu.edu/sahami98bayesian.html

Sebastiani F (2002) Machine learning in automated text categorization. ACM Comput Surv 34(1):1–47. http://doi.acm.org/10.1145/505282.505283

Shin Y, Gupta M, Myers SA (2011) Prevalence and mitigation of forum spamming. In: Proceedings of the 30th IEEE international conference on computer communications, Shanghai

Zhang L, Zhu J, Yao T (2004) An evaluation of statistical spam filtering techniques. ACM Trans Asian Lang Inf Process 3(4):243–269

Spam Review

▶ Identifying Spam in Reviews

SPARQL

Axel Polleres
Siemens AG Österreich, Vienna, Austria

Synonyms

SPARQL 1.1; W3C Standard RDF Query Language

Glossary

BGP	Basic Graph Pattern, a set of RDF triple "templates" where variables are allowed in either subject predicate or object position, which can be read as a conjunctive query
HTTP	Hypertext Transfer Protocol
OWL	Web Ontology Language, a schema language on top of RDF, rooted in Description Logics
RDF Graph	A set of RDF triples
RDF	Resource Description Framework
RDFS	RDF Schema, a lightweight ontology language on top of RDF
RIF	Rule Interchange Format, a standard to encode and exchange rules
SPARQL Endpoint	The URI at which a SPARQL service listens for requests from clients
SPARQL Protocol	Defines how to invoke SPARQL queries and updates via a SPARQL endpoint and how results should be returned via HTTP
SPARQL Service	Any implementation conforming to the SPARQL Protocol
SPARQL	Initially "Simple Protocol and RDF Query Language," or nowadays more often referred to by the recursive acronym "SPARQL Protocol and RDF Query Language"; in its 1.1 version, SPARQL comprises not only a query language and a protocol but also a data manipulation language and other features
Triple	An atomic statement of the form (subject predicate object) in RDF
URI	Universal Resource Identifiers, a generalization of URLs (cf. IETF RFC1630)
W3C Recommendation	Standards published by the W3C
W3C	World Wide Web Consortium

Definition

SPARQL, the "Simple Protocol and RDF Query Language," is the W3C's standard query language for RDF (the Resource Description Framework, an emerging data format on the growing Web of Data). However, the SPARQL standard does not only comprise a query language but a family of W3C standards to access and manipulate RDF data; in its current version SPARQL 1.1, the standard comprises:

- A query language (SPARQL 1.1 Query Language)
- A data manipulation language (SPARQL 1.1 Update)
- A mechanism to describe and discover SPARQL endpoints (SPARQL 1.1 Service Description)
- An extension to delegate parts of a query to a remote SPARQL endpoint (SPARQL 1.1 Federated Query)
- Various result formats (SPARQL 1.1 Query Results JSON Format, SPARQL 1.1 Query Results CSV and TSV Formats, SPARQL Query Results XML Format)
- A normative way to return additional results entailed by schema and rules languages such as RDFS, OWL, and RIF (SPARQL 1.1 Entailment Regimes)

- A protocol to invoke SPARQL queries and updates via HTTP (SPARQL 1.1 Protocol)
- An extension to the SPARQL Protocol, to perform certain operations to manage collections of graphs directly via HTTP (SPARQL 1.1 Graph Store HTTP Protocol)

Introduction

The Semantic Web is in principle a family of standards to enable a Web of Data, with the final goal of enabling nothing less than the vision of the Web as a database (Berners-Lee 1999). The architecture of these standards comprises of (i) a simple graph-based data model, RDF; (ii) schema languages, RDFS and OWL; (iii) rules languages, RIF; and last but not least, (iv) a query language, sparql. The existence of such a standard query language has significantly contributed to the increasing uptake of RDF as a basic data format on the Web over the past years. After SPARQL's first edition has become a W3C recommendation in 2008, the community and implementers have requested a variety of additional features that the SPARQL 1.1 working group took as a starting point in 2009 for re-shaping the next version of the standard. In March 2013, the group concluded its work by publishing 11 specification documents (listed above) as a W3C recommendation.

Methodology

In this section, we introduce various parts of the SPARQL specification by a short example.

We will illustrate the use of SPARQL's languages, protocols, and related specifications with a small example RDF graph published on the Web at the URL "http://example.org/alice" which contains personal information about Alice and her social contacts. We use Turtle (Beckett et al. 2013) syntax here for illustration:

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<http://example.org/alice#me> a foaf:Person .
<http://example.org/alice#me> foaf:name "Alice" .
<http://example.org/alice#me> foaf:mbox
<mailto:alice@example.org> .
<http://example.org/alice#me> foaf:knows
<http://example.org/bob#me> .
<http://example.org/bob#me> foaf:knows
<http://example.org/alice#me> .
<http://example.org/bob#me> foaf:name "Bob" .
<http://example.org/alice#me> foaf:knows
<http://example.org/charlie#me> .
<http://example.org/charlie#me> foaf:knows
<http://example.org/alice#me> .
<http://example.org/charlie#me> foaf:name "Charlie" .
<http://example.org/alice#me> foaf:knows
<http://example.org/snoopy> .
<http://example.org/snoopy> foaf:name "Snoopy"@en .
```

With SPARQL 1.1, one can query such graphs, load them into RDF stores, and manipulate them in various ways.

Firstly, the *SPARQL 1.1 Query Language* (Harris and Seaborne 2013) can be used to formulate queries against RDF ranging from simple graph pattern matching to complex queries. For instance, one can ask using a SPARQL SELECT query for names of persons and the number of their friends:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name (COUNT(?friend) AS ?count)
WHERE{
 ?person foaf:name ?name .
 ?person foaf:knows ?friend .
 }GROUP BY ?person ?name
```

Complex queries may include union, optional query parts, and filters; new features like value aggregation, path expressions, and nested queries have been added in SPARQL 1.1. Apart from SELECT queries – which return variable bindings – SPARQL supports ASK queries, i.e., Boolean "yes/no" queries, and CONSTRUCT queries, by which new RDF graphs can be constructed from a query result; all the new query language features of SPARQL 1.1 are likewise usable in ASK and CONSTRUCT queries.

Results of SELECT queries in SPARQL comprise bags of mappings from variables to RDF terms, often conveniently represented in tabular form. For instance, the query from section 2 has the following results:

In order to exchange these results in machine-readable form, SPARQL supports four standard formats to exchange results, namely, the Extensible Markup Language (XML) (Hawke 2013), the JavaScript Object Notation (JSON) (Seaborne 2013a), as well as the Comma-Separated Values (CSV) and Tab-Separated Values (TSV) (Seaborne 2013b).

?name	?count
"Alice"	3
"Bob"	1
"Charlie"	1

The *SPARQL 1.1 Federated Query* (Prud'hommeaux and Buil-Aranda 2013) extension allows to explicitly delegate certain subqueries to different SPARQL endpoints. For instance, in our example, one may want to know whether there is anyone among Alice's friends with the same name as the resource identified by the IRI <http://dbpedia.org/resource/Snoopy> at DBpedia. This can be done by combining a query for the names of friends with a remote call to the SPARQL endpoint at http://dbpedia.org/sparql finding out the name of <http://dbpedia.org/resource/Snoopy> using the SERVICE keyword as follows:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name WHERE {
<http://example.org/alice#me> foaf:knows [ foaf:name ?name] .
   SERVICE <http://dbpedia.org/sparql>
   { <http://dbpedia.org/resource/Snoopy> foaf:name ?name } }
```

Here, the first part of the pattern in the WHERE part is still matched against the local SPARQL service, whereas the evaluation of the pattern following the SERVICE keyword is delegated to the respective remote SPARQL service.

SPARQL can be used together with *entailment regimes* (Glimm and Ogbuji 2013), that is, exploiting ontological information in the form of, for example, RDF Schema (RDFS) or OWL axioms. For instance, let us assume that – apart from the data about Alice – some ontological information in the form of RDFS (Brickley and Guha 2004) and OWL (2012) constructs defining the FOAF vocabulary is loaded into our example SPARQL service.

The FOAF ontology (cf. http://xmlns.com/foaf/spec/, retrieved April2013), of which we only give a small excerpt here, contains, for instance, the following RDFS axiom:

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
...
foaf:name rdfs:subPropertyOf rdfs:label .
...
```

The following query asks for labels of persons:

- `SELECT? label`
- `WHERE {?person rdfs:label? label}`

A SPARQL engine that does not consider any special entailment regimes (on top of standard simple entailment) would not return any results for this query, whereas an RDF Schema aware query engine will return

Since foaf:name is a sub-property of rdfs:label.

?label
"Alice"
"Bob"
"Charlie"
"Snoopy"@en

The *SPARQL 1.1 Update* (Gearon et al. 2013) specification defines the syntax and semantics of SPARQL 1.1 Update requests. Update operations can consist of several sequential requests and are performed on a collection of graphs in a Graph Store. Operations are provided to update, create, and remove RDF graphs in a Graph Store. For instance, the following request inserts a new friend of Alice named Dorothy into the default graph of our example SPARQL service and thereafter deletes all names of Alice's friends with an English language tag.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/> .

INSERT DATA { <http://www.example.org/alice#me> foaf:knows
[foaf:name "Dorothy" ]. } ;
DELETE { ?person foaf:name ?mbox }
WHERE {
   <http://www.example.org/alice#me> foaf:knows ?person .
   ?person foaf:name ?name . FILTER ( lang(?name) = "EN" ) .}
```

As the second operation shows, insertions and deletions can be dependent on the results of queries to the Graph Store; the respective syntax used in the WHERE part is derived from the SPARQL 1.1 *Query Language*.

The *SPARQL 1.1 Protocol for RDF* (Feigenbaum et al. 2013) defines how to transfer SPARQL 1.1 queries and update requests to a SPARQL service via HTTP. It also defines how to map requests to HTTP GET and POST operations and what respective HTTP responses to such requests should look like. Additionally, the *SPARQL 1.1 Service Description* (Williams 2013) document describes a method for discovering and an RDF vocabulary for describing SPARQL services made available via the SPARQL 1.1 Protocol. According to this specification, a service endpoint, when accessed via an HTTP GET operation without further (query or update request) parameters, should return an RDF description of the service provided.

For many applications and services that deal with RDF data, the full SPARQL 1.1 Update language might not be required. To this end, the *SPARQL 1.1 Graph Store HTTP Protocol* (Ogbuji 2013) provides means to perform certain operations to manage collections of graphs directly via HTTP operations.

For instance, the first part of the update request in above is a simple insertion of triples into an RDF graph. On a service supporting this protocol, such insertion can – instead of via a SPARQL 1.1 Update request – directly be performed via an HTTP POST operation taking the RDF triples to be inserted as payload.

Implementations

A list of SPARQL 1.0 implementations is available at http://www.w3.org/wiki/SparqlImplementations (retrieved April 2013), whereas a list of implementations of the new features of SPARQL 1.1 along with reports on test coverage is available at http://www.w3.org/2009/sparql/implementations/ (retrieved April 2013). As for performance evaluations, a list of benchmarks is available at http://www.w3.org/wiki/RdfStoreBenchmarking; The Europeana report (Haslhofer et al. 2011) compares and describes various curre nt SPARQL implementations, not yet mentioning SPARQL 1.1 implementations, though.

SPARQL in Academia

The formal semantics of the SPARQL Query Language in its original recommendation in 2008 has been very much inspired by academic results, such as by the papers of Pérez et al. (2006, 2009). Angles and Gutierrez (2008) later showed that SPARQL – as defined in those papers -has exactly the expressive power of non-recursive safe Datalog with negation. Another translation from SPARQL to Datalog has been presented in Polleres (2007).

Extensions that were now standardized in SPARQL 1.1 such as subqueries (Angles and Gutierrez 2011), path expressions (Alkhateeb et al. 2009; Pérez et al. 2010), or aggregates (Polleres et al. 2007) have also been discussed or proposed in some variants in the academic literature. Details about the differences of the semantics as defined in the official W3C specification and in most of these academic papers are discussed in Polleres (2012). Query optimization and particularly equivalence of SPARQL queries have been discussed to some extent already in Pérez et al. (2009). These results were refined and extended in Schmidt et al. (2008), Letelier et al. (2012), and Chekol et al. (2012). The semantics of SPARQL entailment regimes has been discussed in Kollia et al. (2011); foundational aspects of federated queries are discussed in Buil-Aranda et al. (2013). The semantics of path expressions in SPARQL 1.1 has been discussed in Arenas et al. (2012) and Losemann and Martens (2012), and it should be noted that these papers to some extent influenced the definition of the semantics of path expressions in the final specification. More practical proposals for query optimizations are discussed in Stocker et al. (2008) and Vidal et al. (2010). Overall, SPARQL is a source of ongoing research and inspired various academic works on its foundations, optimization, and extensions, a full account of which would be beyond the scope of this article.

Future Directions

Various additional features requested to the query language could not yet be taken into account for SPARQL 1.1, and the working group has collected a list of open work items on its wiki page (http://www.w3.org/2009/sparql/wiki/Future_Work_Items, retrieved April 2013) which comprises features that had to be left out either for reasons of priorities or missing implementation experience to be standardized already. It may be expected that – just like in the transition from SPARQL 1.0 to SPARQL 1.1 -upon implementation experience and community feedback from implementers, a new working group by W3C will be formed in the future to add additional features. As already mentioned in the previous section, academic research can potentially impact these future directions; for instance, extensions of regular path queries (a small subset of which is now incorporated into SPARQL 1.1) which are currently being investigated in academia (e.g., Barceló et al. 2010) might be viewed as very valuable additions to query graph data in RDF.

Cross-References

- ▶ RDF
- ▶ RIF: The Rule Interchange Format
- ▶ Web Ontology Language (OWL)
- ▶ Xpath/XQuery

Acknowledgments The author would like to thank all members of the W3C SPARQL working group as well as various people who helped to improve the standard by their comments to public-rdf-dawg-comments@w3.org. A more detailed version of the examples provided herein can be found in the SPARQL 1.1 Overview document (SPARQL 1.1 Overview 2013).

References

Alkhateeb F, Baget J-F, Euzenat J (2009) Extending SPARQL with regular expression patterns (for querying RDF). J Web Semant 7(2):57–73

Angles R, Gutierrez C (2008) The expressive power of SPARQL. In: International semantic web conference, Karlsruhe, pp 114–129

Angles R, Gutierrez C (2011) Subqueries in SPARQL. In: Alberto Mendelzon international workshop on foundations of data management, Santiago

Arenas M, Conca S, Pérez J (2012) Counting beyond a Yottabyte, or how SPARQL 1.1 property paths will prevent adoption of the standard. In: WWW 2012, Lyon, pp 629–638

Barceló P, Hurtado CA, Libkin L, Wood PT (2010) Expressive languages for path queries over graph-structured data. In: PODS 2010, Indianapolis, pp 3–14

Beckett D, Berners-Lee T, Prud'hommeaux E, Carothers G (eds) (2013) Terse RDF triple language. W3C Candidate Recommendation, 19 Feb 2013

Berners-Lee T (1999) Weaving the web. Harper, San Francisco

Brickley D, Guha RV (eds) (2004) RDF vocabulary description language 1.0: RDF schema. W3C Recommendation, 10 Feb 2004

Buil-Aranda C, Arenas M, Corcho O, Polleres A (2013) Federating queries in SPARQL 1.1: syntax, semantics and evaluation. J Web Semant 18(1):1–17

Chekol MW, Euzenat J, Genevès P, Layaïda N (2012) SPARQL query containment under SHI axioms. In: AAAI 2012, Toronto

Feigenbaum L, Williams GT, Clark KG, Torres E (eds) (2013) SPARQL 1.1 Protocol. W3C Recommendation, 21 Mar 2013

Gearon P, Passant A, Polleres A (eds) (2013) SPARQL 1.1 Update. W3C Recommendation, 21 Mar 2013

Glimm B, Ogbuji C (eds) (2013) SPARQL 1.1 entailment regimes. W3C Recommendation, 21 Mar 2013

Harris S, Seaborne A (eds) (2013) SPARQL1.1 query language. W3C Recommendation, 21 Mar 2013

Haslhofer B, Roochi EM, Schandl B, Zander S (2011) Europeana RDF store report. Technical report, University of Vienna, Vienna

Hawke S (ed) SPARQL query results XML format, 2nd edn. W3C Recommendation, 21 Mar 2013

Kollia I, Glimm B, Horrocks I (2011) SPARQL query answering over OWL ontologies. In: ESWC 2011 (1), Heraklion, pp 382–396

Letelier A, Pérez J, Pichler R, Skritek S (2012) Static analysis and optimization of semantic web queries. In: PODS 2012, Scottsdale, pp 89–100

Losemann K, Martens W (2012) The complexity of evaluating path expressions in SPARQL. In: PODS 2012, Scottsdale, pp 101–112

Ogbuji C (ed) (2013) SPARQL 1.1 graph store HTTP protocol. W3C Recommendation, 21 Mar 2013

OWL (2012) OWL 2 web ontology language document overview (2nd edn). W3C Recommendation, 11 Dec 2012

Pérez J, Arenas M, Gutierrez C (2006) Semantics and complexity of SPARQL. In: International semantic web conference, Athens, pp 30–43

Pérez J, Arenas M, Gutierrez C (2009) Semantics and complexity of SPARQL. ACM Trans Database Syst 34(3):6

Pérez J, Arenas M, Gutierrez C (2010) nSPARQL: a navigational language for RDF. J Web Semant 8(4):255–270

Polleres A (2007) From SPARQL to rules (and back). In: WWW, Banff, pp 787–796

Polleres A (2012) How (well) do datalog, SPARQL and RIF interplay? In: Datalog 2012 workshop, Vienna, pp 27–30

Polleres A, Scharffe F, Schindlauer R (2007) SPARQL++ for mapping between RDF vocabularies. In: OTM conferences (1), Vilamoura, pp 878–896

Prud'hommeaux E, Buil-Aranda C (eds) (2013) SPARQL 1.1 federated query. W3C Recommendation, 21 Mar 2013

Schmidt M, Meier M, Lausen G (2008) Foundations of SPARQL query optimization. CoRR abs/0812.3788

Seaborne A (ed) (2013a) SPARQL 1.1 query results JSON format. W3C Recommendation, 21 Mar 2013

Seaborne A (ed) (2013b) SPARQL 1.1 query results CSV and TSV formats. W3C Recommendation, 21 Mar 2013

SPARQL 1.1 Overview (2013) SPARQL 1.1 overview. W3C Recommendation, 21 Mar 2013

Stocker M, Seaborne A, Bernstein A, Kiefer C, Reynolds D (2008) SPARQL basic graph pattern optimization using selectivity estimation. In: WWW 2008, Beijing, pp 595–604

Vidal ME, Ruckhaus E, Lampo T, Martínez A, Sierra J, Polleres A (2010) Efficiently joining group patterns in SPARQL queries. In: ESWC 2010 (1), Heraklion, pp 228–242

Williams GT (ed) (2013) SPARQL 1.1 service description. W3C Recommendation, 21 March 2013

SPARQL 1.1

▶ SPARQL

Spatial Analysis

▶ Spatial Statistics

Spatial Interaction

▶ Spatiotemporal Footprints in Social Networks

Spatial Item Recommendation

▶ Spatiotemporal Recommendation in Geo-Social Networks

Spatial Networks

Marc Barthelemy
Institut de Physique Théorique, CEA,
Gif-sur-Yvette, France

Synonyms

Network geography; Space-embedded networks; Transportation systems; Urban networks

Glossary

Graph (or Network) A set of vertices connected by edges

Adjacency Matrix A matrix A which represents the structure of a graph. The element A_{ij} is either 0 if i and j are not connected or $A_{ij} = 1$ if there is an edge from i to j. For a spatial network, the position of the nodes $\{x_i\}$ is needed in order to completely characterize the network

Betweenness Centrality The betweenness centrality of a vertex (or an edge) x is defined as $BC(x) = \sum_{s,t \in V} \frac{\sigma_{st}(x)}{\sigma_{st}}$ where $\sigma_{st}(x)$ is the number of shortest paths between s and t using x and σ_{st} is the number of all shortest paths between s and t

Betweenness Centrality Impact Measures how a new link affects the average betweenness centrality of a graph. This quantity can help in characterizing the different types of new links during the evolution of a (spatial) network

Cell Also called face for planar network is a region bounded by edges. The Euler formula relates the number of nodes, edges, and cells (faces)

Diameter The diameter of a graph is defined as the maximum value of all $\ell(i, j)$, is the distance between i and j, and is used to measure the "size" of it. For most real-world spatial network, the diameter scales as the number of nodes to the power $1/d$ where d is the dimension of the embedding space

Planar Graph A planar graph can be drawn in 2-D such that none of its edges are crossing

Organic Ratio Measures the proportion of degree 1 ("dead ends") and degree 3 nodes (“T-shaped intersections”). If the organic ratio is small, the corresponding spatial network is very close to a regular rectangular lattice

Alpha Index Also called the meshedness, it measures the ratio of observed circuits to the maximum number of elementary circuits which can exist in the network

Gamma Index Ratio of the number of edges to the maximum number possible for a planar graph with the same number of nodes

Shape Factor Ratio of the area of a cell to the area of the circumscribed circle

Route Distance Distance between two nodes measured by the length of the shortest path connecting them

Detour Index Ratio of the route distance between two nodes and the Euclidean distance between them

Network Cost Ratio of the total length of the network to the total length of the minimum spanning tree constructed on the same set of nodes

Network Performance Ratio of the average shortest path of the network to the average shortest path of the minimum spanning tree constructed on the same set of nodes

Definition

More generally, the term “spatial network” has come to be used to describe any network in which the nodes are located in a space equipped with a metric (Barthelemy 2011). For most practical applications, the space is the two-dimensional space and the metric is the usual Euclidean distance. For these networks we thus need both the topological information about the graph (given by the adjacency matrix) and the spatial information about the nodes (given by the position of the nodes).

Transportation and mobility networks, Internet, mobile phone networks, power grids, social and contact networks, and neural networks are all examples where space is relevant and where topology alone does not contain all the information.

Characterizing and understanding the structure and the evolution of spatial networks is crucial for many different fields ranging from urban-ism to epidemiology. An important consequence of space on networks is that there is usually a cost associated to the length of edges which in turn has dramatic effects on the topological structure of these networks. Indeed, a long link will be very costly and can exist if this cost if balanced with another good reason (economical or connection to a hub, . . .). For most real-world spatial networks, we indeed observe that the probability of finding a link between two nodes will decrease with the distance. Spatial constraints affect not only the structure and properties of these networks but also processes which take place on these networks such as phase transitions, random walks, synchronization, navigation, resilience, and disease spread.

All planar graphs can be embedded in a two-dimensional space and can be represented as spatial networks, but the converse is not necessarily true: there are some spatial and nonplanar graphs. In general, however, most spatial networks are, to a good approximation, planar graphs (Clark and Holton 1991), such as road or railway networks, but there are some important exceptions such as the airline network (Barrat et al. 2004): in this case the nodes are airports and there is a link connecting two nodes if there is at least one direct connection. For many infrastructure networks, however, planarity is unavoidable. Power grids, roads, rail, and other transportation networks are to a very good accuracy planar networks. For many applications, planar spatial networks are the most important and most studies have focused on these examples.

Also, the above definition does not imply that the links are necessarily embedded in space. Indeed, in social networks, individuals are connected through a friendship relation which is a virtual network of relations. There is however a strong spatial component in these networks as the probability that individuals located in space are friends generally decreases with the distance between them (Liben-Nowell et al. 2005).

Introduction

For many critical infrastructures, communication or biological networks, space is relevant: most of the people have their friends and relatives in their neighborhood, power grids and transportation networks depend obviously on distance, many communication network devices have short radio range, the length of axons in a brain has a cost, and the spread of contagious diseases is not uniform across territories. In particular, in the important case of the brain, regions that are spatially closer have a larger probability of being connected than remote regions as longer axons are more costly in terms of material and energy (Bullmore and Sporns 2009). Wiring costs depending on distance are thus certainly an important aspect of brain networks, and we can probably expect spatial networks to be very relevant in this rapidly evolving topic. Another particularly important example of such a spatial network is the Internet which is defined as the set of routers linked by physical cables with different lengths and latency times. More generally, the distance could be another parameter such as a social distance measured by salary, socio-professional category differences, or any quantity which measures the cost associated with the formation of a link.

All these examples show that these networks have nodes and edges which are constrained by some geometry and are usually embedded in a two- or three-dimensional space, and this has important effects on their topological properties and consequently on processes which take place on them. If there is a cost associated to the edge length, longer links must be compensated by some advantage, for example, being connected to a well-connected node - that is, a hub. The topological aspects of the network are then correlated to spatial aspects such as the location of the nodes and the length of edges.

Tools for Characterizing Spatial Networks

Graphs are usually characterized by the adjacency matrix A where the elements are $A_{ij} = 1$ if nodes i and j are connected (see, e.g., a graph textbook Clark and Holton 1991). This matrix completely characterizes the topology of the graph and is enough for most applications. This is however not the case for spatial networks where the spatial information is contained in the location of the nodes x_i. Two topologically identical graphs can then have completely different spatial properties, and this is at the heart of the richness and complexity of spatial networks.

In this section, we will discuss some tools which can be helpful to characterize some aspects of spatial networks.

Degree Distribution, Clustering, and Average Shortest Path Length

Degree Distribution

In complex networks, the degree distribution, the clustering spectrum, and the average shortest distance are of utmost importance (Albert and Barabasi 2002). Their knowledge already gives a useful picture of the graph under study. In contrast, in spatial networks, physical constraints impose some of the properties. In particular, there is usually a sharp cutoff on the degree distribution $P(k)$ which is therefore not broad. This is true for most spatial and planar networks such as power grids or transportation networks, for example. For a spatial, nonplanar network such as the airline network, the cutoff can be large enough and the degree distribution could be characterized as broad.

Clustering

The clustering coefficient of a node counts how its neighbors are connected with each other. For spatial networks, the dominant mechanism is usually to minimize cost associated with length, and nodes have a tendency to connect to their nearest neighbors, independently from their degree. This in general implies that the clustering spectrum $C(k)$ is relatively flat for spatial networks.

The same argument can be used to show that the assortativity “spectrum” defined as the function $k_{nn}(k)$ is also approximately constant in general when spatial constraints are very strong (see Barthelemy 2011 for more details).

Average Shortest Distance

Usually, there are many paths between two nodes in a connected network, and the shortest one defines a distance on the network:

$$\ell(i,j) = \min_{\text{paths}(i \to j)} |\text{path}| \tag{1}$$

where the length |path| of the path is defined as its number of edges. This quantity is infinity when there are no paths between the nodes and is equal to one for the complete graph (for which $\ell(i, j) = 1$). For weighted graphs, we assign to each link e a weight w_e and the length of a path is given by $|\text{path}| = \sum_{e\ \in\ \text{Path}} w_e$.

In most complex networks, one observes a small-world behavior (Watts and Strogatz 1998) of the form

$$\langle \ell \rangle \sim \log N \tag{2}$$

In contrast, for a real-world spatial network embedded in a d-dimensional space, we usually observe the very different behavior:

$$\langle \ell \rangle \sim N^{1/d} \tag{3}$$

which also means that to go from one node to another one, one has to cross a path of length of the order of the diameter (which is not the case when shortcuts exist). The measure of the average shortest path length could thus be a first indication whether a network is close to a lattice or if long-range links are important.

Organic Ratio

We note that more recently, other interesting indices were proposed in order to characterize specifically road networks (Xie and Levinson 2007). Indeed, the degree distribution is very peaked around 3–4, and an interesting information is given by the ratio

$$r_N = \frac{N(1) + N(3)}{\sum_{k \neq 2} N(k)} \tag{4}$$

where N(k) is the number of nodes of degree k. If this ratio is small, the number of dead ends and of “unfinished” crossing ($k = 3$) is small compared to regular crossing with $k = 4$, signalling a more organized city. In the opposite case of large $r_N \simeq 1$, there is dominance of $k = 1$ and $k = 3$ nodes which signals a more “organic” city.

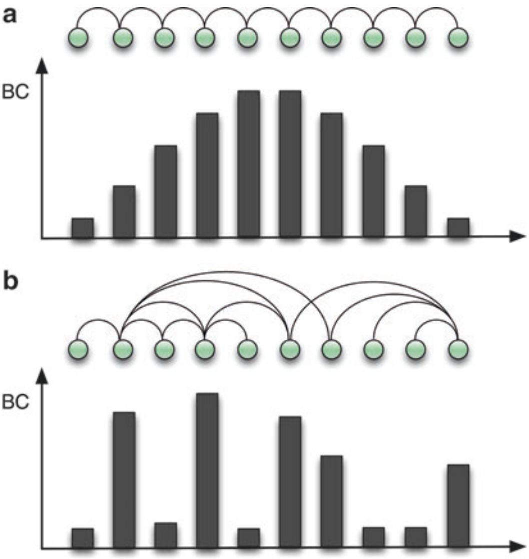
Betweenness Centrality

Anomalies

The betweenness centrality (BC) of a vertex (Freeman 1977) is determined by its ability to provide a path between separated regions of the network. Hubs are natural crossroads for paths, and it is natural to observe a marked correlation between the average $\left(g(k) = \sum_{i/k_i=k} g(i)/N(k)\right.$ and k as expressed in the following relation:

$$g(k) \sim k^{\eta} \tag{5}$$

where η depends on the characteristics of the network. We expect this relation to be altered when spatial constraints become important, and in order to understand this effect, we consider a one-dimensional lattice which is the simplest case of a spatially ordered network. For this lattice the shortest path between two nodes is simply the Euclidean geodesic, and for two points lying far from each other, the probability that the shortest path passes near the barycenter of the network is very large. In other words, the barycenter (and its neighbors) will have a large centrality as illustrated in Fig. 1a. In contrast, in a purely topological network with no underlying geography, this consideration does not apply anymore, and if we rewire more and more links (as illustrated in Fig. 1b), we observe a progressive decorrelation of centrality and space while the correlation with degree increases. In a lattice, it is easy to show that the BC depends on space and is maximum at the barycenter, while in a



Spatial Networks, Fig. 1 (**a**) Betweenness centrality for the (one-dimensional) lattice case. The central nodes are close to the barycenter. (**b**) For a general graph, the central nodes are usually the ones with large degree

network the BC of a node depends on its degree. When the network is constituted of long links superimposed on a lattice, we then expect the appearance of "anomalies" characterized by large deviations around the behavior $g \sim k^{\eta}$.

Betweenness Centrality Impact

When studying the time evolution of networks, it is important to be able to characterize quantitatively new links. This is particularly true for spatial networks, but what follows could also be applied to general, complex networks.

We consider a time-evolving graph G_t described by a set of nodes V_t and edges E_t at time t. In order to evaluate the impact of a new link on the overall distribution of the betweenness centrality in the graph at time t, we first compute the average betweenness centrality of all the links of G_t as

$$\bar{b}(G_t) = \frac{1}{(N(t)-1)(N(t)-2)} \sum_{e \in E_t} b(e) \qquad (6)$$

where $b(e)$ is the betweenness centrality of the edge e in the graph $\mathrm{G_t}$. Then, for each new link e^* added in the time window $[t-1, t]$, we consider the new graph obtained by removing the link e^* from G_t, denoted by $\mathrm{G_t \backslash \{e^*\}}$. The impact $\delta_b(e^*)$ of edge e^* on the betweenness centrality of the network at time t is then defined as (Strano et al. 2012)

$$\delta_b(e^*) = \frac{\left[\bar{b}(G_t) - \bar{b}(G_t \backslash \{e^*\})\right]}{\bar{b}(G_t)} \qquad (7)$$

The betweenness centrality impact is thus the relative variation of the graph average betweenness due to the removal of the link e^* and can thus help to characterize quantitatively the various mechanisms at play during the evolution of the network (Strano et al. 2012).

Mixing Space and Topology

All the previous indicators describe essentially the topology of the network, but are not specifically designed to characterize spatial networks. We will here briefly review other indicators which provide useful information about the spatial structure of networks. Different indices were defined a long time ago mainly by scientists working in quantitative geography since the 1960s and can be found in Haggett and Chorley (1969) (see also the more recent paper by Xie and Levinson (2007)). Most of these indices are relatively simple but still give important information about the structure of the network in particular if we are interested in planar networks. These indices were used so far to characterize transportation networks such as highways or railway systems.

Alpha and Gamma Indices

The most important indices are called the "alpha" and the "gamma" indices. The simplest index is called the gamma index and is simply defined by

$$\gamma = \frac{E}{E_{\max}} \qquad (8)$$

where E is the number of edges and $E_{\max}$ is the maximal number of edges (for a given number of nodes N). For nonplanar networks, $E_{\max}$ is given by $N(N - 1)/2$ for nondirected graphs and for planar graphs $E_{\max} = 3\ N - 6$ leading to

$$\gamma_P = \frac{E}{3N - 6} \tag{9}$$

The gamma index is a simple measure of the density of the network, but one can define a similar quantity by counting not the edges but the number of elementary cycles. The number of elementary cycle for a network is known as the cyclomatic number (see, e.g., Clark and Holton 1991) and is equal to

$$\Gamma = E - \mathrm{N} + 1 \tag{10}$$

For a planar graph this number is always less or equal to $2\ N - 5$ which leads naturally to the definition of the alpha index (also coined as meshedness in Buhl et al. 2006)

$$\alpha = \frac{E - N + 1}{2N - 5} \tag{11}$$

This index belongs to [0,1] and is equal to 0 for a tree and equal to 1 for a maximal planar graph.

Cell Area and Shape

For planar spatial networks, we have faces or cells which have a certain area and shape. In certain conditions, it can be interesting to characterize statistically these shapes, and various indicators were developed in this perspective (see Haggett and Chorley 1969 for a list of these indicators).

The first, simple important information is the distribution of the area $P(A)$ which for many cases follows a power law (Lammer et al. 2006; Barthelemy and Flammini 2008):

$$P(A)\tilde{}A^{-\tau} \tag{12}$$

where $\tau \approx 2$. We can note here that a simple argument on node density fluctuation leads indeed to this value $\tau = 2$ and further empirical analysis is needed to test the universality of this result.

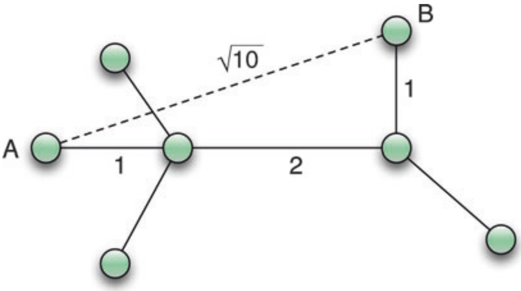
In addition to the area of the cell, its shape distribution is also interesting and contains a large part of the information about the structure of the network. A simple way to characterize the shape is given by the form factor ϕ. If we denote by L the major axis, the shape ratio is defined as A/L^2 (or equivalently, we can define the elongation ratio $\sqrt{A}/L$). In the paper (Lammer et al. 2006) on the road network structure, Lämmer et al. use another definition of the form factor and define it as

$$\varphi = \frac{4A}{\pi D^2} \tag{13}$$

where $\pi\ D^{\mathrm{N}}$ is the area of the circumscribed circle. If this ratio is small, the cell is very anisotropic, while on the contrary if ϕ is closer to one, the corresponding cell is almost circular. In many cases where rectangles and squares predominate (Lammer et al. 2006; Strano et al. 2012), we have $\phi \approx 0{:}5$–$0{:}6$.

Detour Index

When the network is embedded in a two-dimensional space, we can define at least two distances between the pairs of nodes. There is of course the natural Euclidean distance $d^E\ (i, j)$ which can also be seen as the "as crow flies" distance. There is also the total "route" distance $d^R\ (i, j)$ from i to j by computing the sum of lengths of segments belonging to the shortest path between i and j. The detour index – also called the route factor – for this pair of nodes (i, j) is then given by (see Fig. 2 for an example)



Spatial Networks, Fig. 2 Example of detour index calculation. The "as crow flies" distance between the nodes A and B is $d_E\,(A, B) = \sqrt{10}$, while the route distance over the network is $d_R(A, B) = 4$ leading to a detour index equal to $Q(A, B)$ D $4/\sqrt{10} \simeq 1{:}265$

$$Q(i,j) = \frac{d_R(i,j)}{d_E(i,j)} \tag{14}$$

This ratio is always larger than one, and the closer to one, the more efficient the network. From this quantity, we can derive another one for a single node defined by

$$\langle Q(i)\rangle = \frac{1}{N-1}\sum_j Q(i,j) \tag{15}$$

which measures the "accessibility" for this specific node i. Indeed the smaller it is, the easier it is to reach the node *i*. This quantity is related to the quantity called "straightness centrality" (Crucitti et al. 2006)*:*

$$C^S(i) = \frac{1}{N-1}\sum_{j\neq i}\frac{d_E(i,j)}{d_R(i,j)} \tag{16}$$

And if one is interested in assessing the global efficiency of the network, one can compute the average over all pairs of nodes:

$$\langle Q\rangle = \frac{1}{N(N-1)}\sum_{i\neq j} Q(i,j) \tag{17}$$

The average $\langle Q\rangle$ or the maximum $Q_{\max}$, and more generally the statistics of $Q(i,j)$, is important and contains a lot of information about the spatial network under consideration (see Aldous and Shun 2010 for a discussion on this quantity for various networks). For example, one can define the interesting quantity Aldous and Shun (2010)

$$\rho(d) = \frac{1}{N_d}\sum_{ij/d_E(i,j)=d} Q(i,j) \tag{18}$$

(where N_d is the number of nodes such that $d_E(i,j) = d$) whose shape can help in characterizing combined spatial and topological properties.

Cost and Efficiency

The minimum number of links to connect N nodes is $E = N - 1$ and the corresponding network is then a tree. We can also look for the tree which minimizes the total length given by the sum of the lengths of all links:

$$\ell_T = \sum_{e\in E} d_E(e) \tag{19}$$

where $d_E(e)$ denotes the length of the link e. This procedure leads to the minimum spanning tree (MST) which has a total length ℓ_T^{MST} (see, e.g., Clark and Holton 1991). Obviously the tree is not a very efficient network (e.g., from the point of view of transportation), and usually more edges are added to the network, leading to an increase of accessibility but also of ℓ T. A natural measure of the "cost" of the network is then given by

$$C = \frac{\ell_T}{\ell_T^{\mathrm{MST}}} \tag{20}$$

We note here that we easily estimate the total length if the segment length distribution is peaked around its average ℓ_1, and if the node distribution is uniform, $\ell_1 \sim 1 = \surd\rho$ where $\rho = N/A$ is the average node density (A is the area of the system). In this case, the total length is given by $\ell_T = E \mp {}_1$ leading to

$$\ell_T = \frac{\langle k\rangle}{2}\sqrt{AN} \tag{21}$$

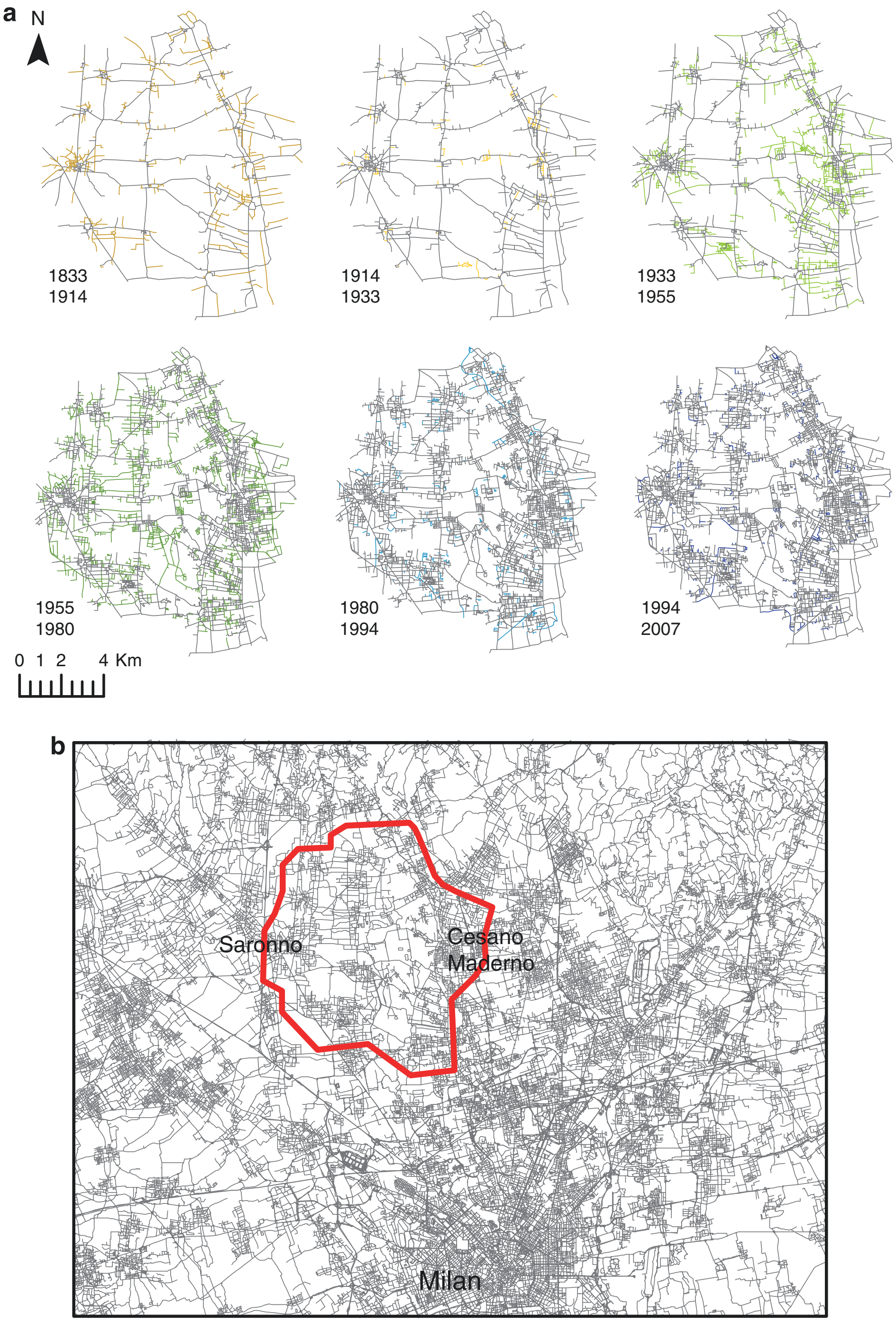
where $\langle k\rangle$ is the average degree of the graph. Adding links thus increases the cost but improves accessibility or the transport performance P of the network which can be measured as the minimum distance between all pairs of nodes, normalized by the same quantity computed for the minimum spanning tree*:*

$$P = \frac{\langle \ell\rangle}{\langle \ell_{\mathrm{MST}}\rangle} \tag{22}$$

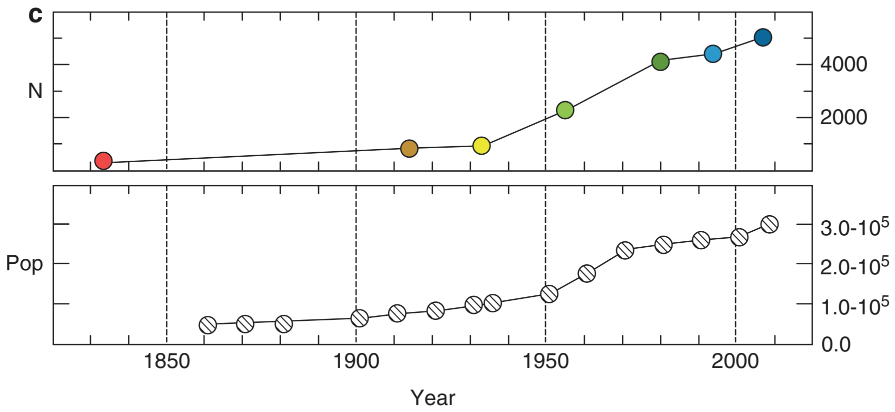
Another measure of efficiency was also proposed in Latora and Marchiori (2001) and is defined as

$$E = \frac{1}{N(N-1)}\sum_{i\neq j}\frac{1}{\ell(i,j)} \tag{23}$$

where $\ell(i,\ j)$ is the shortest path distance from *i* to *j*. Combination of these different indicators



Spatial Networks, Fig. 3 (continued)



Spatial Networks, Fig. 3 (**a**) Evolution of the road network from 1833 to 2007 (for each map we show in *grey* all the nodes and links already existing in the previous snapshot of the network and in colors the new links added in the time window under consideration). (**b**) Map showing the location of the studied area (Groane area in the metropolitan region of Milan). (**c**) Time evolution of the total number of nodes N in the network and of the total population in the area (obtained from census data) (Figure taken from Strano et al. 2012)

and comparisons with the MST or the maximal planar network can be constructed in order to characterize various aspects of the networks under consideration (see, e.g., Buhl et al. 2006).

Finally, adding links improves the resilience of the network to attacks or dysfunctions. A way to quantify this is by using *fault tolerance* (*FT*) (see, e.g., Tero et al. 2010) measured as the probability of disconnecting parts of the network with the failure of a single link. The benefit/cost ratio could then be estimated by the quantity FT/ℓ_T^{MST} which is a quantitative characterization of the trade-off between cost and efficiency (Tero et al. 2010).

Future Directions

In this final section, we discuss briefly two directions for future research which seem very promising. Both directions come from the fact that ever more data are available, opening the path for new measures, new models, and new understanding of the formation and evolution of spatial networks.

Measuring and Modeling the Time Evolution of Spatial Networks

Thanks to the efforts of GIS scientists (Batty 2005), we now have digitalized maps, combined with data from remote sensing, which allows for studying the time evolution of spatial networks such as roads and streets over long periods. Understanding the evolution of transportation networks (Xie and Levinson 2009) is important from a fundamental point of view but also sheds some light on the crucial problem of understanding the time evolution of a city. Recent studies (Xie and Levinson 2009; Strano et al. 2012; Barthelemy et al. 2013) started to quantify the evolution of spatial networks, and more empirical results are certainly to come (Fig. 3).

At the time this article is written, we are still in the process of collecting data, processing them, and extracting stylized facts. The next important step will be the modeling of the evolution of these systems. There are already some simplified models, but we will now be able to confront theoretical models with stylized fact and hopefully converge to simple realistic models of spatial network evolution. In particular, all these studies will have to address the issue of self-organization versus centralized planning for different time scales, a crucial problem in the modeling of urban systems.

Connecting Spatial Networks with Socioeconomical Indicators

Revealing the relationships of network topology to socioeconomical features is not a new project. There is indeed a wealth of papers in quantitative geography of the 1960s–1970s (see, e.g., Haggett and Chorley 1969; Radke 1977 and references therein). In 1969, for example, (Kissling 1969) concludes that the analysis of the network structure is "likely to reveal probable growth points in the system." However, the recent availability of spatial data on networks and on socioeconomical indicators reinvigorates this direction of research. This can even be done at various scales. At large scales, for example, one can try to understand the relation between population, activity densities, and the structure of transportation networks. At a smaller scale, one can try to understand crime rates and activity density fluctuation in terms of topological properties of the transportation network.

This problem will also require a lot of efforts from the modeling side. In particular, we know that there is strong coupling between the population density and the network structure, but we still need a modeling framework for describing such a coupling and coevolution. From a longer time scale perspective, these studies on spatial networks belong to the more general problem of understanding the time evolution of a city. So far, modeling a city has mostly been done in the field of spatial economics (Fujita et al. 1999). However most of these studies consider monocentric structures and static properties, and their predictions are not compared with empirical data. Gathering various data, proposing simple dynamical models integrating the most relevant economical ingredients, and confronting their prediction to data will certainly lead in some future to a wealth of new and original results about this very complex system that is a city.

Cross-References

- ▶ Analysis and Planning of Urban Networks
- ▶ Detecting and Identifying Communities in Dynamic and Complex Networks: Definition and Survey
- ▶ Location-Based Social Networks
- ▶ Networks in Geography
- ▶ Spatial Statistics
- ▶ Temporal Networks
- ▶ Tools for Networks

References

Albert R, Barabasi AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47

Aldous DJ, Shun J (2010) Connected spatial networks over random points and a route-length statistic. Stat Sci 25:275–288

Barrat A, Barthelemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. Proc Natl Acad Sci USA 101:3747

Barthelemy M (2011) Spatial networks. Phys Rep 499:1

Barthelemy M, Flammini A (2008) Modelling urban street patterns. Phys Rev Lett 100:138702

Barthelemy M, Bordin P, Berestycki H, Gribaudi M (2013) Self-organization versus top-down planning in the evolution of a city. Nat Sci Rep 3:2153

Batty M (2005) Network geography: relations, interactions, scaling and spatial processes in GIS. In: Fisher PF, Unwin DJ (eds) Re-presenting GIS. Wiley, Chich-ester, pp 149–170

Buhl J, Gautrais J, Reeves N, Solé RV, Valverde S, Kuntz P, Theraulaz G (2006) Topological patterns in street networks of self-organized urban settlements. Eur Phys J B-Condens Matter Complex Syst 49(4):513–522

Bullmore E, Sporns O (2009) Complex brain networks: graph theoretical analysis of structural and functional systems. Nat Rev Neurosci 10(3):186–198

Clark J, Holton DA (1991) A first look at graph theory, vol 6. World Scientific, Teaneck

Crucitti P, Latora V, Porta S (2006) Centrality in networks of urban streets. Chaos Interdiscip J Nonlinear Sci 16 (1):015113–015113

Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40:35–41

Fujita M, Krugman PR, Venables AJ (1999) The spatial economy: cities, regions and international trade, vol 213. MIT, Cambridge

Haggett P, Chorley RJ (1969) Network analysis in geography. Edward Arnold, London

Kissling CC (1969) Linkage importance in a regional highway network. Can Geogr 13:113–129

Lammer S, Gehlsen B, Helbing D (2006) Scaling laws in the spatial structure of urban road networks. Phys A Stat Mech Appl 363(1):89–95

Latora V, Marchiori M (2001) Efficient behavior of small-world networks. Phys Rev Lett 87:198701

Liben-Nowell D, Novak J, Kumar R, Raghavan P, Tomkins A (2005) Geographic routing in social networks. Proc Natl Acad Sci USA 102:11623–11628

Radke JD (1977) Stochastic models in circuit network growth. Thesis and dissertations (Comprehensive). Paper 1450, Wilfrid Laurier University

Strano E, Nicosia V, Latora V, Porta S, Barthelemy M (2012) Elementary processes governing the evolution of road networks. Nat Sci Rep 2:296

Tero A, Takagi S, Saigusa T, Ito K, Bebber DP, Fricker MD, Yumiki K, Kobayashi R, Nakagaki T (2010) Rules for biologically inspired adaptive network design. Sci Signal 327:439

Watts D, Strogatz S (1998) Collective dynamics of small-world networks. Nature 393:440–442

Xie F, Levinson D (2007) Measuring the structure of road networks. Geogr Anal 39:336–356

Xie F, Levinson D (2009) Topological evolution of surface transportation networks. Comput Environ Urban Syst 33:211–223

Spatial Scan Statistic

▶ Disease Surveillance: Case Study

Spatial Statistics

Victor Oliveira[1] and A. Alexandre Trindade[2]
[1]Department of Management Science and Statistics, The University of Texas at San Antonio, San Antonio, TX, USA
[2]Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX, USA

Synonyms

Geocomputation; Geostatistics; Spatial analysis

Glossary

Correlation/ covariance	Measures of similarity between observations
Geostatistics	A branch of spatial statistics
Isotropy	Property of covariance and variogram functions that make them is invariant under rotation of locations
Kriging	Method for linear unbiased prediction
Random field	A collection of random variables indexed by location
Stationarity	Property of random fields in which their mean and covariance functions are invariant under translation of locations
Variogram/ semivariogram	Measures of dissimilarity between observations

Definition

Spatial statistics is a branch of statistics that studies methods to make inference based on data observed over spatial regions. In typical applications these regions are either 2- or 3-dimensional. The methodology is mostly aimed at accounting and modeling aspects of the so-called First Law of Geography: attributes from locations that are closer together are more closely related than attributes from locations that are farther apart. This is accomplished through appropriate measures of spatial association. An overview of models and methods is given for the three main types of spatial data: geostatistical, lattice, and point pattern.

Introduction

Spatial data refer to measurements of phenomena that vary over a region of space $D \subset \mathbb{R}^d$, $d \geq 1$, which would be called the region of interest. Each datum is associated to a subset of D that indicates where it was collected, often called the datum's support. This may be a single point or a larger subset, depending on the context.

There are three basic types of spatial data: geostatistical (or point referenced), lattice (or areal), and point pattern. The three types may be viewed as pairs $\{(s_i, z_i) : i = 1, \ldots, n\}$ where the interpretation and characteristics of the data components vary from type to type. For geostatistical data $z_1,\ldots,z_n$ are measurements or observations of a phenomenon of interest taken at sampling locations $s_1 ; \ldots , s_n \in D$, which are single points. In the models to be described later, the z_i s are random, while n (the sample size) and the s_i s are known and fixed. For lattice data $s_1,\ldots,s_n$ are subregions that form a partition of D, such as counties or postal codes, and $z_1,\ldots,z_n$ are averages or summaries of the phenomenon of interest over these subregions. For this type of data, it also holds that the z_i s are random, while n and the s_i s are known and fixed. For point pattern data $s_1,\ldots,s_n$ are points where a certain event of interest occurs, such as the presence of a type of tree or the epicenter of an earthquake, and $z_1,\ldots,z_n$ are a

Spatial Statistics, Table 1 Summary and overview of concepts, models, and examples in the three types of spatial data

	Geostatistical	Lattice	Point pattern
Domain D	Fixed, continuous	Fixed, discrete	Random, continuous
Observation sites $\{s_i : i = 1, \ldots, n\}$	s_i fixed n fixed	s_i fixed n fixed	s_i random n random
Inference for	Z.(s) only	Z.(s) only	Both Z.(s) and D
Main models for Z.(s)	Sum of regression trend and stationary random field	Simultaneous Autoregressive Model (SAR)	Poisson process (homogeneous and inhomogeneous)
Key aims, concepts	Kriging (minimum MSE prediction)	Spatial proximity matrix (W)	Assess tendency for clustering
Examples	Meteorological and geological variables	Geographic and demographic variables	Location and intensity of events

feature of the aforementioned events, such as the diameter of the tree at breast height or the magnitude of the earthquake. In the models for point pattern data to be described later, all components n, s_i and, z_i are random. Often the z_i s are absent when interest centers only on the pattern of occurrences. For all three types of spatial data, additional variables could also be available, that serve as explanatory variables. Comprehensive treatments of statistical models and methods for all three types of data appear in Cressie (1993), Schabenberger and Gotway (2005), and the recent edited volume by Gelfand et al. (2010). Table 1 summarizes the key concepts and gives an overview of models and examples.

Key Points

Random Fields

A random field $\{Z(s) : s \in D\}$ on the region $D \subset \mathbb{R}^d$ is a collection of random variables indexed by the elements of D, where D can be finite or infinite. These random variables are often nonidentically distributed and dependent, so modeling these aspects is a key starting point. The simplest way to do this is through the mean and covariance functions of the random field, defined as

$$\mu(s) := \mathbb{E}\{Z(s)\} \quad \texttt{and}$$
$$C(s,u) := \texttt{cov}\{Z(s), Z(u)\}, \quad s, u \in D.$$

The former determines the spatial trend, a measure of variation over large distances, while the latter determines the spatial association, a measure of variation over small distances.

Other features of a random field also related with spatial association are the correlation function and semivariogram function, defined respectively as

$$\begin{aligned} K(s,u) &:= \operatorname{corr}\Big\{Z(s), \{Z(u)\} = \frac{C(s,u)}{\sigma(s)\sigma(u)} \\ \gamma(s,u) &:= \frac{1}{2}\operatorname{var}\{Z(s) - Z(u)\} \\ &= \frac{1}{2}\left(\sigma^2(s) + \sigma^2(u) - 2C(s,u)\right), \end{aligned}$$

where $\sigma^2 . (s) = \operatorname{var}\{Z . (s)\}$ is the variance function. The functions C.(s, u) and γ.(s, u) provide similar information about the spatial association of the random field, with the former being a measure of similarity between Z.(s) and Z(u), while the latter is a measure of dissimilarity. When choosing the aforementioned functions, it is important to note that any function can be used as a mean function, but not any function can be used as a covariance function. The latter needs to be positive semi-definite, meaning that for any $m \in \mathbb{N}, s_1, \ldots, s_m \in D$ *and* $a_1, \ldots, a_m \in \mathbb{R}$ it holds that

$$\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j C(s_i, s_j) \geq 0.$$

This is a difficult condition to verify, but fortunately the literature provides many functions known to be positive semi-definite; see Cressie (1993) and Chilès and Delfiner (1999) for examples. These references also provide an intermediate treatment on the theory and methods of random fields and their application to spatial statistics, while Matérn (1986), Yaglom (1987), and Stein (1999) provide more mathematical treatments.

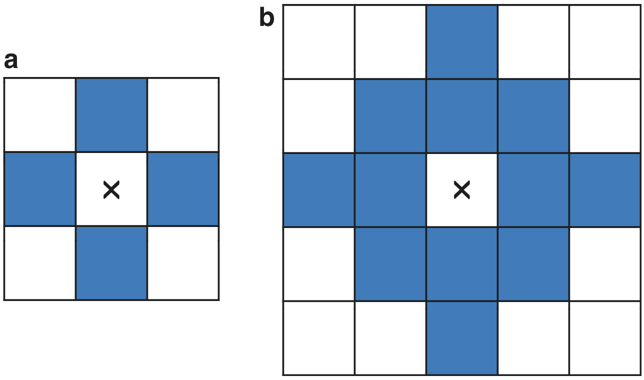
Lattice data usually represent averages or summaries of a quantity of interest over subregions, so covariance and semivariogram functions are not the most suitable to quantify spatial association among this type of data. Instead, neighborhood relations and weight matrices are used. In this case the collection of subregions $\{s_1,\ldots,s_n\}$ is endowed with a neighborhood system $\{N_i : i = 1, \ldots, n\}$, where N_i denotes the subregions that are, in a precisely defined way, neighbors of subregion s_i. For rectangular regular lattices where the subregions may be thought of as pixels, it is common to use first-order neighborhood systems, where the neighbors of a pixel are the pixels adjacent to the north, south, east, and west; see Fig. 1a. Second-order neighborhood systems are also used, where the neighbors of a pixel are its first-order neighbors and their first-order neighbors; see Fig. 1b. In these cases all pixels have the same number of neighbors, except for pixels at (or near) the boundary of D. For regions divided in unequally shaped subregions (like counties in a state), a commonly used neighborhood system is defined in terms of geographic adjacency, $N_i = \{s_j$: subregions s_i and s_j share a boundary$\}$; other examples not based on geographic adjacency are also possible. In these cases the number of neighbors for each subregion usually differs.

In addition a weight (or neighborhood) matrix $W = (w_{ij})$ is specified, where w_{ij} measures the strength of direct association between sites s_i and s_j. It must satisfy that $w_{ij} \geq 0$, $w_{ii} = 0$ and $w_{ij} > 0$ if and only if s_i and s_j are neighbors (i.e., $s_j \in N_i$). The most common example of weight matrix is $w_{ij} = 1$ if s_i and s_j are neighbors and $w_{ij} = 0$ otherwise, but other more refined specifications are also possible, e.g., based on distance between subregions' centroids. Anselin (1988), Cressie (1993), Rue and Held (2005), and LeSage and Pace (2009) provide ample treatments of models and methods for the analysis of lattice data.

For geostatistical and lattice data, the sampling locations $s_1,\ldots,s_n$ are fixed and known, so these types of data are usually written as $z = (z_1;\ldots,z_n)^T$ (^{T} denotes transpose of a vector or matrix). The stochastic approach for modeling and inference assumes the data are a part of a realization of a random field $Z(\cdot)$, so datum z_i is the realized value of the random variable $Z(s_i)$.

Stationarity and Ergodicity

Spatial data typically contain no replicates as usually a single observation is available at each location, so some assumptions on the random field are needed to make statistical inference feasible. To illustrate this point consider the conceptual decomposition $Z(s_i) = \mu(s_i) + \varepsilon(s_i)$, with $\varepsilon(\cdot)$a random field with mean zero and co-variance function $C(s, u)$. Without some extra assumptions it is not possible to identify both $\mu(s_i)$ and $\varepsilon(s_i)$ with a single observation at s_i. This is so because a term can be added to $\mu(s_i)$ and subtracted from $\varepsilon(s_i)$ in infinitely many ways, any of which will



Spatial Statistics, Fig. 1 Examples of first-order (**a**) and second-order (**b**) neighborhood systems. Pixels in *blue* are the neighbors of the pixel marked with an "x" not change the datum Z(s$_i$) but will change the components that seek to be identified.

The assumptions alluded above are those of stationarity and ergodicity. A random field Z (s) is said to be (second-order or weakly) stationary if

$$\mu(s) = \mu(\text{constant})\text{and}$$
$$C(s,u) = \tilde{C}(s-u), s, u \in D,$$

where $\tilde{C}(\cdot)$ is a function of a single spatial variable. The above means that the mean and covariance functions are invariant under translations of the spatial locations. From these follow that the variance, correlation, and semivariogram functions are also invariant under translations of the spatial locations, and we have

$$\sigma^2(s) = \sigma^2, C(s,u) = \sigma^2\tilde{K}\ (s-u),$$
$$\gamma(s,u) = \sigma^2\left(1-\tilde{K}(s-u)\right).$$

An important and commonly used special case of stationarity is called isotropy, meaning that $C(s,u) = \overline{C}(||s-u||)$, where $||h|| := \left(h_1^2 + \cdots + h_d^2\right)^{1/2}$ is the Euclidean norm of $h \in \mathbb{R}^d$ and $\overline{C}\ (\cdot)$ is a function of a single real variable. In this case the covariance function is also invariant under rotations of the spatial locations, so the nature of spatial association is the same in all directions; see Ripley (1981), Cressie (1993), and Schabenberger and Gotway (2005) for further discussion on stationarity.

A precise definition of ergodicity is somewhat technical (see Cressie 1993, pp. 53–58), but this assumption is key to make statistical inference based on spatial data feasible. This is so because the meaning and interpretation of many features of a random field, such as the mean function, are based on ensemble (i.e., population) averages, namely, averages over the possible realizations of the random field. Ergodicity requires that spatial averages computed from a single realization converge to their respective ensemble averages as the sample size increases to infinity.

A complete description of a random field requires specifying its family of finite-dimensional distributions, namely, the family of joint distributions

$$F_{s_1,\ldots,s_m}(x_1,\ldots,x_m) = P\{Z(s_1) \\ \leq x_1,\ldots,Z(s_m) \leq s_m\},$$

$\forall m \in \mathbb{N}$ *and* $s_1, \ \ldots \ , s_m \in D$. The simplest and most commonly used of such specification is that of Gaussian random fields, meaning that all the aforementioned distributions are multi-variate normal. Gaussian random fields are completely specified by their mean and covariance functions, and when they are stationary, a sufficient condition for them to be ergodic is that $\lim_{||h||\to\infty}\tilde{C}(h) = 0$. Gaussian random fields are the most commonly used models because of their convenient mathematical properties and wide applicability, as well as their use as "building blocks" for more complex random fields models. Examples of the latter are hierarchical models used to describe discrete spatial data; see Banerjee et al. (2004) and Diggle and Ribeiro (2007).

Models and Inference

Geostatistical Data Models

The basic geostatistical model is based on the conceptual decomposition of the random field of interest as

$$Z(s) = \mu(s) + \varepsilon(s), \quad s \in D,$$

where μ(s) is the mean function (spatial trend) and ε(·) is a zero-mean random field that describes the short-range variation, with the same covariance function as Z(·). The usual model for the spatial trend is similar to that used in linear regression models

$$\mu(s) = \sum_{j=1}^{p} f_j(s)\beta_j = \boldsymbol{f}(s)^{\mathrm{T}}\boldsymbol{\beta},$$

where β = (β$_1$,. . .,β$_p$)^T are unknown regression parameters and f(s) = (f$_1$(s),. . .,f$_p$(s))^T are known location-dependent covariates. The latter may include related spatially varying processes. For instance, if Z(s) = rainfall amount that fell over a period of time at locations s, then f.s/ = altitude at location s may be a useful explanatory variable. More often a spatial trend is described in terms of a polynomial in the spatial coordinates. For the case when d = 2 and s = (x, y), this would be

$$\mu(s) = \sum_{0 \le i+j \le p} \beta_{ij} x^i y^j, \text{for some } p \ge 1 \text{ known.}$$

Many examples of stationary covariance models have been proposed in the literature (see Cressie 1993; Chilès and Delfiner 1999). An example of a flexible family of isotropic covariance functions is the so-called Matérn family (Matérn 1986; Stein 1999)

$$\overline{C}(t) = \frac{2\sigma^2}{\Gamma(v)} \left(\frac{t}{2\phi}\right)^v \mathrm{K}_v\left(\frac{t}{\phi}\right), \quad t \ge 0,$$

where $\Gamma(\cdot)$ is the gamma function and $\mathrm{K}_v(\cdot)$ is the modified Bessel function of the second kind and order v. For such model $\phi > 0$ (mainly) controls how fast the correlation decreases with distance, and $v > 0$ controls the smoothness of the realizations of the random field. The commonly used exponential and Gaussian covariance functions are special cases obtained, respectively, by setting $v = 1/2$ and $v \rightarrow \infty$.

The above description assumes the process of interest is measured exactly (or nearly so), but more often the data contain measurement error; see Le and Zidek (2006) for an extensive discussion. In this case the simplest model for the observed data is

$$Z_{i,\,\text{obs}} = Z(s_i) + \acute{o}_i, \quad i = 1, \ldots, n,$$

where $\acute{o}_1,\ldots,\acute{o}_n$ are assumed i.i.d with mean 0, variance $\tau^2 > 0$ and independent of $Z(\cdot)$. Under the above model the data $Z_{\text{obs}} = (Z_{1\text{ , obs}}, \ldots, Z_{n\text{ , obs}})^T$ follow the general linear model

$$\boldsymbol{Z}_{\text{obs}} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where X is the n by p matrix with entries $(X)_{ij} = f_j(s_i)$ and ε is a random vector with $\mathbb{E}\{\varepsilon\} = 0$ and $\text{var}\{\varepsilon\} = \text{var}\{Z_{\text{obs}}\} = \Sigma_\theta$, with the n by n matrix Σ_θ having entries $(\Sigma_\theta)_{ij} = \sigma^2(2/\tau(\gamma))(t_{ij}/2\phi)^\gamma \mathrm{K}_\gamma(t_{ij} = \phi)$ and $t_{ij} = \| s_i - s_j 1 \;; 1 \quad (A)$ denotes the indicator function of A. This basic specification of a geostatistical model depends on unknown regression parameters β and covariance parameters $\theta = (\sigma^2; \phi, v; \tau^2)$.

Parameter Estimation

The classical geostatistical method of estimation uses a distribution-free approach (Journel and Huijbregts 1978; Cressie 1993; Chilès and Delfiner 1999). First, the regression parameters are estimated by least squares, resulting in

$$\widehat{\boldsymbol{\beta}} = (X'QX)^{-1} X^T Q \boldsymbol{Z}_{\text{obs}},$$

where $Q = I_n$ (ordinary least squares) or $Q = \Sigma_{\boldsymbol{\theta}}^{-1}$ (generalized least squares); the latter requires an estimate of Σ_θ. In both cases X is assumed to have full rank. The second choice of Q results in a more efficient estimator, but often in practice there is little difference between them. The resulting trend surface estimate is $\widehat{\mu}(s) = \boldsymbol{f}(s)^{\mathbf{T}} \widehat{\boldsymbol{\beta}}$.

Second, when the mean function is constant, the covariance parameters are estimated by the following two-stage approach: For selected distances $t_1 < \ldots < t_k$, the (model-free) semi-variogram estimates are first computed

$$\widehat{\gamma}(t_j) = \frac{1}{2|N(t_j)|} \sum_{N(t_j)} \left(z_{i,\,\text{obs}} - z_{j,\,\text{obs}}\right)^2,$$

where $N(t) = \{(i, j) : t - \Delta t < \| s_i - s_j \| < t + \Delta t\}$, with $\Delta t > 0$ fixed and $|N(t)|$ the number of elements in N(t). A proposed semi-variogram model, say $\gamma(t;\, \theta)$, is then fitted to the above semivariogram estimates $\widehat{\gamma}(t_1), \ldots, \widehat{\gamma}(t_k)$ using (nonlinear) least squares, so the covariance parameter estimates are

$$\widehat{\boldsymbol{\theta}} = \arg\min \sum_{j=1}^{k} \left(\widehat{\gamma}(t_j) - \gamma(t_j;\, \boldsymbol{\theta})\right)^2.$$

The resulting semivariogram function estimate is $\gamma\left(\widehat{\boldsymbol{\theta}}\right)$. When μ(s) is not constant a similar procedure is done using the residuals $\mathbf{e} = \mathbf{z}_{\text{obs}} - X\widehat{\boldsymbol{\beta}}$, rather than the observed data. This estimation method is popular among practitioners, but the statistical properties of the resulting estimators are not well understood.

When the random field Z(·) is Gaussian, all the parameters can be jointly estimated by maximum likelihood (Cressie 1993; Stein 1999), resulting in the estimators

$$\left(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\theta}}\right) = \arg\max L(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{z}_{\text{obs}}), \qquad (1)$$

where

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{z}_{\text{obs}}) = \left(\tfrac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} |\Sigma_{\boldsymbol{\theta}}|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{z}_{\text{obs}} - X\boldsymbol{\beta})^{\top}\Sigma_{\boldsymbol{\theta}}^{-1}(\mathbf{z}_{\text{obs}} - X\boldsymbol{\beta})\right\}. \qquad (2)$$

This method is more statistically satisfactory than the two-stage approach described above but is also more computationally demanding, to the point of not being feasible for very large datasets (n very large) due the need of storing and numerically inverting the n by n matrix Σ_θ; see Cressie (1993), Schabenberger and Gotway (2005), and chapter 4 in Gelfand et al. (2010) for other methods of estimation.

Spatial Prediction (Kriging)

The primary task in the analysis of geostatistical data is often spatial prediction, also known as kriging, which consists of making inference about $Z(s_0)$ where $s_0 \in D$ is an unsampled location. The classical approach uses optimal linear unbiased prediction and only requires knowledge of the mean and covariance (or semivariogram) functions. Specifically, the method seeks to minimize the mean squared prediction error

$$MSPE\left(\widehat{Z}(s_0)\right) = \mathbb{E}\left\{\left(Z(s_0) - \widehat{Z}(s_0)\right)^2\right\},$$

over the class of linear unbiased predictors, that is, predictors of the form $\widehat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i(s_0) z_{i,\text{obs}}$ that satisfy $\mathbb{E}\left\{\widehat{Z}(s_0)\right\} = \mathbb{E}\{Z(s_0)\}$. Under the aforementioned linear model, the optimal coefficients are obtained as the solution of a linear system of equations, and the resulting optimal predictor is

$$\widehat{Z}^K(s_0) = \left(\boldsymbol{\sigma}_0 + X\left(X^T\Sigma_{\boldsymbol{\theta}}^{-1}X\right)^{-1}\left(\boldsymbol{f}(s_0) - X^T\Sigma_{\boldsymbol{\theta}}^{-1}\boldsymbol{\sigma}_0\right)\right)^T \mathbf{Z}_{\text{obs}},$$

where $\sigma_0 = \text{cov}\{Z_{\text{obs}}, Z(s_0)\}$; this is called the best linear unbiased predictor (BLUP) or kriging predictor of $Z(s_0)$. The usual uncertainty measure associated with the kriging predictor is $\text{MSPE}\left(\widehat{Z}^K(s_0)\right)$, which is given by

$$\begin{aligned}\boldsymbol{\sigma}^{2K}(s_0) &= C(0) - \boldsymbol{\sigma}_0^{\top}\sum\nolimits_{\boldsymbol{\theta}}^{-1}\boldsymbol{\sigma}_0 + (\boldsymbol{f}(s_0) \\ &\quad - X^{\top}\sum\nolimits_{\boldsymbol{\theta}}^{-1}\boldsymbol{\sigma}_0\Big)^{\top}\left(X^{\mathrm{T}}\sum\nolimits_{\boldsymbol{\theta}}^{-1}X\right)^{-1}(\boldsymbol{f}(s_0) \\ &\quad - X^{\top}\sum\nolimits_{\boldsymbol{\theta}}^{-1}\boldsymbol{\sigma}_0\Big).\end{aligned}$$

When the random field Z(·) is Gaussian, then $\widehat{Z}^K(s_0)$ is also the best unbiased predictor (it minimizes MSPE(·) over the class of all unbiased predictors), and a 95% prediction interval for $Z(s_0)$ is $\widehat{Z}^K(s_0) \pm 1.96\boldsymbol{\sigma}^K(s_0)$; see Cressie (1993), Chilès and Delfiner (1999), and Schabenberger and Gotway (2005) for methodological details and Stein (1999) for theoretical underpinnings.

The computation of kriging predictors and the validity of their optimality properties require the covariance parameters θ to be known, which is certainly not the case in practice. The simplest and most commonly used practical solution is to use empirical or plug-in predictors and mean squared prediction errors obtained by replacing in the above formulas unknown co-variance parameters with their estimates. But the properties of the resulting plug-in predictors and mean squared prediction errors differ from those of their known covariance parameters counterparts since the former do not take into account the sampling variability of parameter estimators. As a result plug-in mean square prediction errors tend to underestimate the true mean square prediction errors of plug-in predictors, and the true coverage probability of plug-in prediction intervals tends to be smaller than nominal. Possible approaches to account for parameter uncertainty when performing predictive inference include using bootstrap (Sjöstedt-De Luna and Young 2003) and the Bayesian approach (Banerjee et al. 2004; Diggle and Ribeiro 2007), where the latter approach appears to be the most effective.

Lattice Data Models

The starting point in the construction of models for lattice data is to empirically assess the existence of spatial association, which as mentioned in a previous section is usually specified in terms of neighborhood systems and weight matrices. The two most common statistics to diagnose spatial association among lattice data are Moran's I (an analogue of the lagged autocorrelation used in time series) and Geary's c (an analogue of the Durbin-Watson statistic used in time series). For random fields with constant mean, these statistics are defined as

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(Z(s_i)-\overline{Z}\right)\left(Z(s_j)-\overline{Z}\right)}{S_0\sum_{i=1}^{n}\left(Z(s_i)-\overline{Z}\right)^2}$$

$$c = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(Z(s_i)-Z(s_j)\right)^2}{2S_0\sum_{i=1}^{n}\left(Z(s_i)-\overline{Z}\right)^2},$$

where $\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z(s_i)$ and $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$. For Gaussian processes, $\mathbb{E}\{I\} = -(n-1)^{-1}$ and $\mathbb{E}\{c\} = 1$ when observations are independent. Hence, observed values of I substantially below/above $-(n-1)^{-1}$ indicate negative/positive spatial association, while for Geary's c the interpretation is reversed, with observed values of c substantially above/below 1 indicating negative/positive association. When the random field has a nonconstant mean, the above statistics are computed using residuals; see Cressie (1993) and Cliff and Ord (1981) for further details.

A large number of models for lattice data have been proposed in the literature (see Cressie 1993 and LeSage and Pace 2009), where most of them involve the specification of a neighborhood system $\{N_i\}$ and weight matrix W. One of the most common models for lattice data is the Simultaneous Auto-regressive (SAR) model specified by a set of autoregressions

$$Z(s_i) = \boldsymbol{f}(s_i)^{\top}\boldsymbol{\beta} + \rho\sum_{j=1}^{n} w_{ij}\left(Z(s_j) - \boldsymbol{f}(s_j)^{\top}\boldsymbol{\beta}\right) + \acute{o}_i, \quad i = 1, \ldots, n,$$

where $f(s_j)$ and β have the same interpretation as in models for geostatistical data and $\acute{o}_i \sim N(0, \xi_i)$ are independent errors. This is a spatial analogue of autoregressive time series models, but unlike the latter the response and error vectors are correlated. Provided $I_n - \rho W$ is non-singular, it follows that $Z \sim N_n(X\beta, (I_n - \rho W)^{-1}M(I_n - \rho W^T)^{-1})$, where $M = \mathrm{diag}(\xi_1, \ldots, \xi_n)$. It is common to assume $\xi_i = \xi$for all i, in which case the model parameters are β and $\theta = (\xi, \rho)$.

Another large class of models for lattice data is that of Markov random fields (Rue and Held 2005; Li 2009). These models construct the joint distribution for the data by specifying the set of all full conditional distributions, namely, the conditional distributions of $Z(s_i)$ given $Z_{(i)}$, $i = 1, \ldots, n$, where $Z_{(i)} = (Z(s_j) : j \neq i)$. In addition, these models assume a Markov property stating that the distribution of each datum depends on the rest only through its neighbors. An example of this is the class of Conditional Autoregressive (CAR) models with full conditional distributions

$$\left(Z(s_i)|Z(s_j), j \neq i\right)\tilde{}N\left(\boldsymbol{f}(s_i)^{\top}\boldsymbol{\beta} + \rho\sum_{j=1}^{n} w_{ij}\left(Z(s_j) - \boldsymbol{f}(s_j)^{\top}\boldsymbol{\beta}\right), \sigma_i^2\right), i = 1, \ldots, n.$$

To guarantee the above set of full conditional distributions determines a unique joint distribution, it is required that $\sigma_j^2 w_{ij} = \sigma_i^2 w_{ji}$ for all i,j, and $M^{-1}(I_n - \rho W)$ be positive definite, with $M = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right)$, in which case $Z \sim N_n(X\beta, (I - \rho W)^{-1}M)$. It is common to assume that $\sigma_i^2 = \sigma^2$ for all i, in which case the model parameters are β and $\theta = (\sigma^2, \rho)$. An extensive comparison between the SAR and CAR models is given in Cressie (1993, chapter 6).

The most commonly used method for parameter estimation in these models is maximum likelihood. As for geostatistical models, the resulting estimators are given by Eq. 1 where in the likelihood Eq. 2 $\Sigma_{\boldsymbol{\theta}}^{-1} = \left(I_n - \rho W^T\right)M^{-1}(I_n - \rho W)$ for

SAR models and $\Sigma_{\boldsymbol{\theta}}^{-1} = M^{-1}(I - \rho W)$ for CAR models. For both models (as for geostatistical models) the computation of these estimators requires the use of numerical iterative methods.

A point worth noting is that, unlike in geostatistical models, in SAR and CAR models the spatial association structure is specified in terms of the inverse covariance matrix, rather than the covariance matrix, so the interpretation of parameters controlling spatial association is less straightforward than that in geostatistical models.

Point Process Models

A point process on $D \subset \mathbb{R}^d$ is a random field whose realizations are sets of points in D, called point patterns (events). In the most general case attributes may also be observed along with the location of the events, resulting in a marked point process. For any A ⊂ D, let N(A) denote the number of events in A and v(A) the size of A $(=\int_A ds)$. The intensity function of a point process is the function $\lambda : D \rightarrow [0; \infty)$ with the property that $\mathbb{E}\{N(A)\} = \int_A \lambda(s)ds$. Alternatively, using an "infinitesimal disc" ds centered at s the intensity function can be defined as the ratio of the expected number of points in ds to its size, that is,

$$\lambda(s) = \lim_{v(ds) \to 0} \frac{\mathbb{E}\{N(\mathrm{ds})\}}{v(\mathrm{ds})}.$$

The most fundamental point process model is the Poisson process with intensity function λ(s), which satisfies the following: For any $n \in \mathbb{N}$ and $A_1, \ldots, A_n$ disjoint subsets of D, it holds that (i) $N(A_i)$ has Poisson distribution with mean $\int_{A_i} \lambda(s)ds$, and (ii) $N(A_1),\ldots,N(A_n)$ are independent random variables. When the intensity function is constant, λ(s) = λ, the above is called a homogeneous Poisson process (HPP), and otherwise it is called an inhomogeneous Poisson process (IPP). Point patterns from HPP have the property of complete spatial randomness (CSR): given the number of events in a set A, these events are independently and identically distributed over A, so there is no "interaction" between events. Poisson processes are often used on their own for the analysis of point patterns, or as "building blocks" for more complex models; see Diggle (2003) and Illian et al. (2008) for introductory treatments and Cressie (1993) and Daley and Vere-Jones (2003, 2007) for more mathematical treatments.

A basic question in the analysis of point patterns is to assess whether the events have the CSR property. Departures from this comprise either clustering (events tend to aggregate) or regularity (events tend not to aggregate). The standard model by which to assess the CSR property is the HPP. Testing for CSR is based on either counts of events in regions (quadrants) or distance-based measures using the event locations. Focusing on the former, the distributions of some test statistics are known (usually only asymptotically), which allows for closed-form tests. The default is the chi-square test, whereby the region D is bounded by a rectangle and divided into r rows and c columns. If n_{ij} denotes the number of events in the quadrant corresponding to the i-th row and j-th column, and $\bar{n}$ is the expected number of events in any quadrant, then under CSR the statistic

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \bar{n})^2}{\bar{n}},$$

follows a χ^2_{rc-1} distribution, asymptotically. Tests based on more complex nonstandard statistics can be carried out by resorting to Monte Carlo simulation.

Rejection of CSR may lead one to consider modeling a possibly nonconstant intensity function. This can be done either parametrically, by proposing a specific function for the intensity whose parameters are then estimated via maximum likelihood, or nonparametrically by means of kernel smoothing. For example, under an IPP λ(s) can be estimated as a function of coordinates or covariates by fitting a log-linear model of the form

$$\log(\lambda(s)) = \beta_0 + \sum_{j=1}^{p} \beta_j f_j(s),$$

which provides a way to accommodate departures from CSR based on changes in the mean structure.

Alternatively, rejection of CSR may lead one to consider modeling interactions between events, when for non-overlapping regions A and B, N (A), and N(B) are correlated. The second-order intensity function, λ_2(s, u), extends the definition of λ(s) to measure the covariance between points at s and u, defined as

$$\lambda_2(s,u) = \lim_{v(ds),v(du)\to 0} \frac{\mathbb{E}\{N(ds)N(du)\}}{v(ds)v(du)}.$$

For stationary and isotropic processes, where λ(s) ≡ λ and λ_2(s, u) = λ_2(‖s − u‖) ≡ λ_2(t), the K-function is a more informative tool for assessing dependence defined, when d = 2, as

$$K(t) = \frac{2\pi}{\lambda^2}\int_0^t x\lambda_2(x)\mathrm{d}x.$$

Then, λ K.(t) represents the expected number of extra events within a distance t from the origin, given that there is an event at the origin. For a HPP one has K (t) = π t^2; values larger (smaller) than this being indicative of clustering (regularity) on that distance scale. Plotting the estimated K(t) versus t, or the closely related L-function, $L(t) = \sqrt{K(t)/\pi}$, enables one to glean the degree of dependence with reference to the HPP for which L(t) = t; see Diggle (2003) and Illian et al. (2008) for further details.

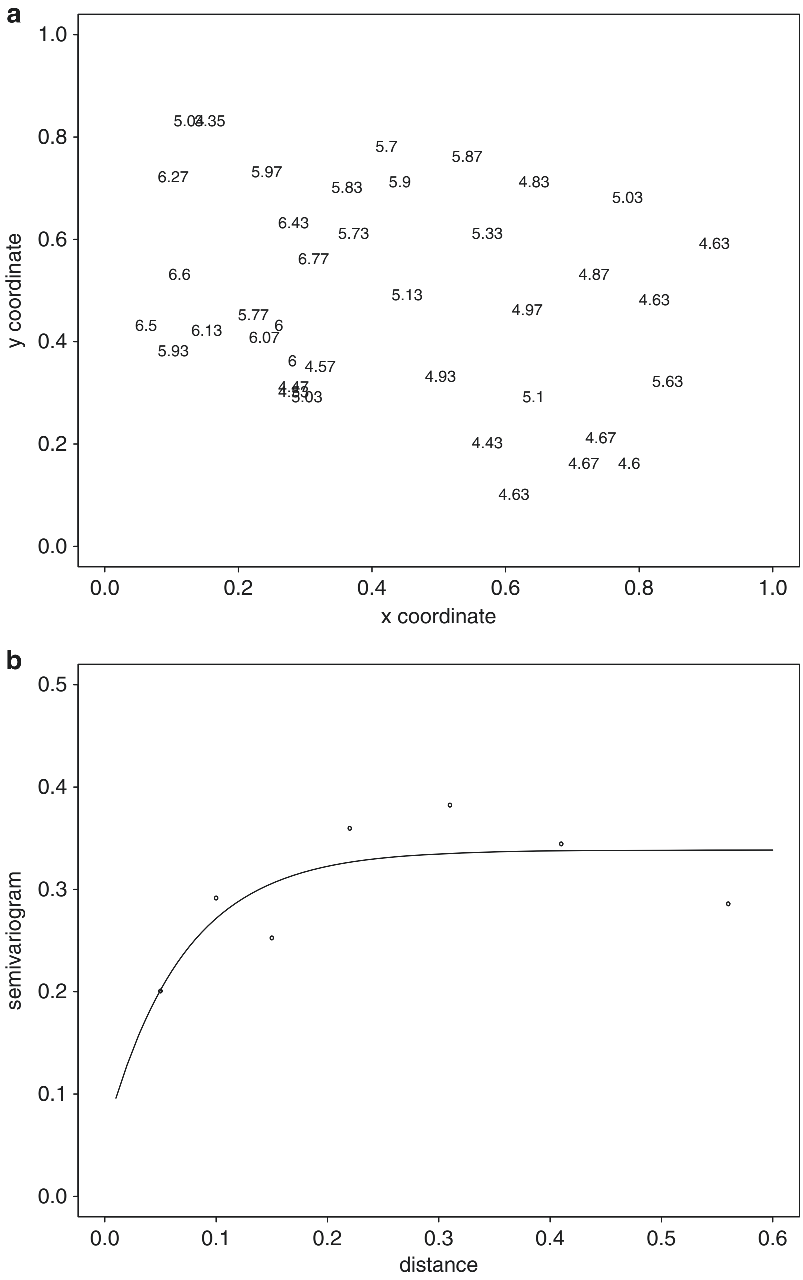
Key Applications

Example 1 As an illustration of a geostatistical dataset Fig. 2a displays pH measurements of wet deposition (acid rain) at 39 rainfall stations taken in April 1987 over the Lower Saxony state in northwest Germany (Berke 1999). Each datum is associated with the sampling location where the pH measurement was taken. For instance, a pH value of 4.63 was observed at the sampling location s = (0.61,0.1) (the southernmost station). For this dataset the coordinates of the sampling locations were provided without units and are all between 0 and 1, which (presumably) mean they were scaled by the maximum distance between stations. A key characteristic of this phenomenon is that a pH value is associated with each location. A typical goal in the analysis of such datasets is the prediction of pH values over a dense grid of prediction locations, which together provide an estimated map of pH over the entire region.

By plotting the pH values against the spatial coordinates, it can be seen that the pH values tend to decrease in the eastward direction and increase in the northward direction. We use a model with μ(s) = β_1 + β_2 x + β_3 y, with s = (x, y), for which the OLS estimates are $\left(\widehat{\beta}_1,\widehat{\beta}_2,\widehat{\beta}_3\right) = (5.627,\ -1.440, 0.761)$. The second-order specification is completed by assuming the covariance function of the true pH process is isotropic and exponential. Figure 2b shows empirical semivariogram estimates at a few selected distances (dots) based on the OLS residuals. It displays an apparent discontinuity at the origin, suggesting the data contain measurement error, so the covariance function of the pH data is C(h) = σ^2 exp. (−h = ϕ) + $r^2$1{h = 0}. The estimated semivariogram function is also displayed in Fig. 2b (line), obtained using the parameters $\left(\widehat{\sigma}^2,\widehat{\phi},\widehat{\tau}^2\right) = (0.270, 0.070, 0.059)$, estimated by least squares.

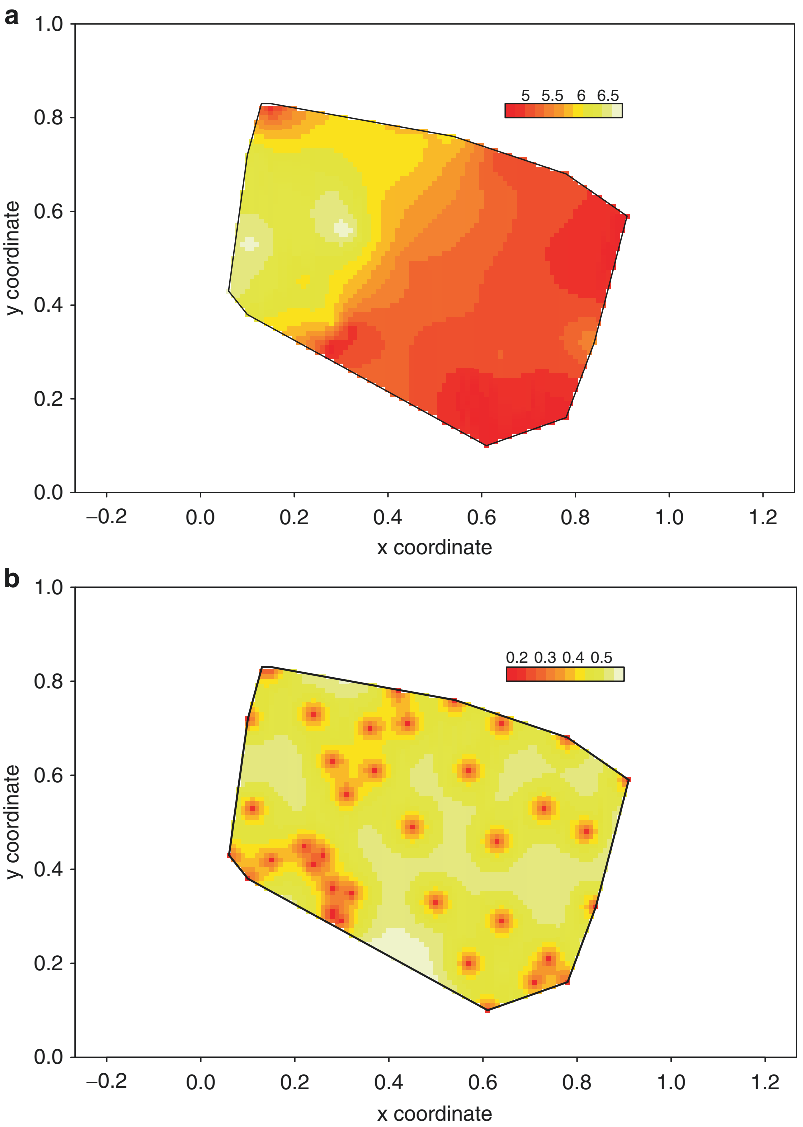
Figure 3a shows a map of estimated pH values obtained by computing the kriging predictor with estimated parameters at about 4,200 prediction locations located inside the convex hull of the sampling locations. Except for the northwest corner of the prediction region that correspond to a group of islands, the pH values are high in the northwest of the state and decrease toward the south and east. Figure 3b shows a map of the square root of the kriging variance at the prediction locations, displaying the typical behavior of having small values at prediction locations close to some sampling location and larger values away from sampling locations.

Example 2 As an illustration of a lattice dataset, we study the relation between poverty level (POV) and total population (POP) at the county



Spatial Statistics, Fig. 2 (**a**) pH measurements and sampling locations and (**b**) empirical and fitted semivariogram function of pH measurements

level in 2009 in the US state of Texas, using data obtained from the US Census Bureau. Figure 4 displays the state of Texas, composed of 254 counties color-coded by the 2009 logarithm of poverty levels. By plotting the data it can be seen that the logarithm of poverty level is closely linearly related with the logarithm of total population, where the least squares fit is $\hat{e}\{\log(\text{POV})|\text{POP}\} = -1 : 741 + 0 : 992 \cdot \log O\ \text{POP}/$. Based on the residuals from this fit, we have that Moran's and Geary's

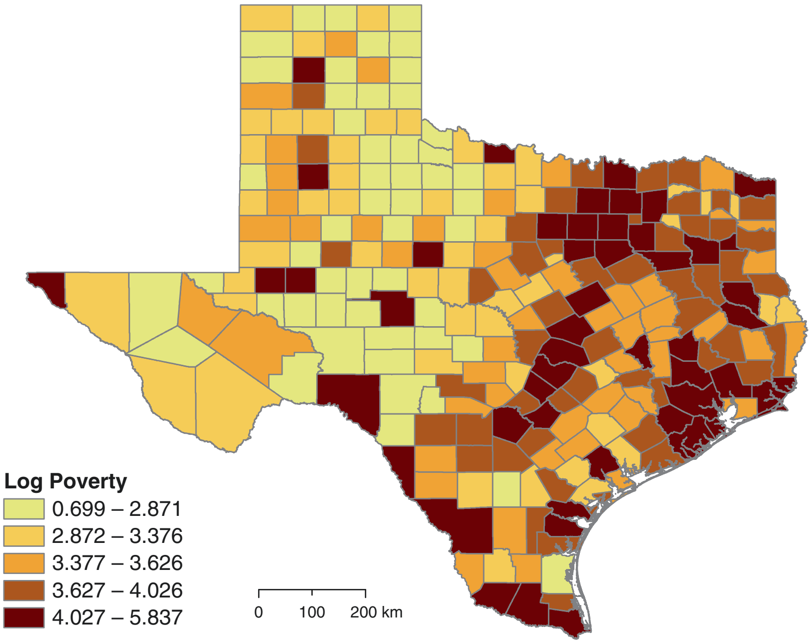


Spatial Statistics, Fig. 3 (**a**) Map of kriging predictor of pH and (**b**) map of square root of kriging variance of pH

statistics are I = 0:391 and c = 0.568, respectively, which are both highly significant for the hypothesis of no spatial association (p-values $<10^{-15}$). Hence, there is substantial spatial association among county log poverty levels, even after accounting for log total population.

We fitted both CAR and SAR models using log poverty level as the response and log total population as the explanatory variable, and the neighborhood system based on geographic adjacency: two counties are neighbors if and only if their boundaries intersect. As for the weights we assume that $w_{ij} = 1$ for any two neighbors s_i and s_j. The SAR model is fit by maximum likelihood, resulting in the estimates $\hat{e}\{\log(\mathrm{POV})|\mathrm{POP}\} = -2.123 + 1:034\cdot \log(\mathrm{POP})$, and $(\widehat{\sigma}^2, \widehat{\rho}) = (0.067, 0.116)$. The estimated mean for the CAR model is similar, but the fit is slightly inferior.

Example 3 As an illustration of a point pattern data, we consider earthquakes (with magnitude



Spatial Statistics, Fig. 4 Choropleth map of county log poverty level in the US state of Texas in 2009

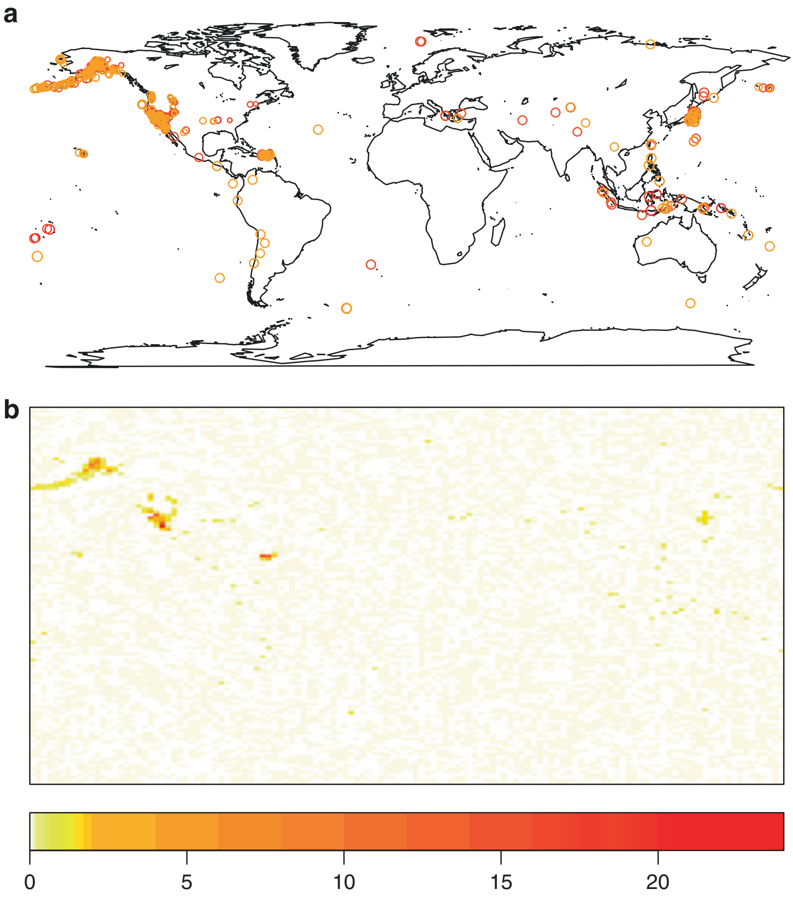
1.0 or more on the Richter scale) that occurred worldwide in 2011 over the 8 consecutive days beginning at 00:00 h UTC on May 20. Figure 5a displays the locations of the 981 events as a "bubble map" with respect to magnitude (size of bubble is proportional to square root of earthquake's magnitude) and so provides a fair visual comparison of the relative sizes (magnitudes) among events. The color-coding scheme renders earlier events in lighter shades of orange and later events in darker shades of red. Since magnitude is an attribute recorded along with each event's location, this is a marked point pattern.

We focus merely on assessing tendency for clustering and disregard magnitude. It is obvious in the current context that there is clustering as geology informs us that this tends to occur at the junction of tectonic plates and fault zones. The Aleutian Islands/Bering Strait and southern Alaska are prominent "hot spots." In fact, a chi-square test strongly rejects CSR (p-value $\approx 10^{-16}$). Assuming (for the sake of illustration) stationarity and isotropy, the estimated L-function reveals a pronounced upward bow that falls well outside the 95% confidence envelopes for a HPP, thus further confirming the strong tendency for clustering on this spatial scale.

Since a constant intensity function is an inadequate hypothesis, we continue the analysis by producing an estimate of the intensity function in the context of an IPP. The result is displayed in Fig. 5b which shows a kernel smoothing estimate (with bandwidth selected by cross-validation; see Diggle 2003). Since intensity is the expected number of (random) points per unit area, the units are "earthquakes per unit area." The two Alaskan hot spots alluded to earlier are clearly visible. Interestingly, the central Caribbean emerges as a third hot spot.

Historical Background and Final Remarks

Early pioneers of statistical inference (e.g., Fisher, Gossett, Pearson) alluded to issues arising from the correlation of observations due to spatial proximity in designed experiments and proposed methods to account for it. Some of the history and early developments in spatial statistics is reviewed in chapter 1 of Gelfand et al. (2010).



Spatial Statistics, Fig. 5 (**a**) Worldwide earthquakes, May 20–27, 2011, and (**b**) Corresponding estimated intensity function

Since some areas of spatial statistics have not been included in this brief overview, we end with some additional pointers to the literature. A review of non-stationary spatial processes is given in chapter 9 of Gelfand et al. (2010). The problems of spatial sampling and design (how and where to collect the data) are treated in Cressie (1993), Le and Zidek (2006), Müller (2007), and chapter 10 of Gelfand et al. (2010). Multivariate methods in spatial statistics are treated in Banerjee et al. (2004), Le and Zidek (2006), Wackernagel (2010), and chapter 21 of Gelfand et al. (2010). Hierarchical models for the modeling of non-Gaussian spatial data, specially models for discrete spatial data, are discussed in Banerjee et al. (2004) and Diggle and Ribeiro (2007), where the Bayesian approach is featured prominently. Models for more complex types of spatial random objects are treated in Matheron (1975), Cressie (1993), and Nguyen (2006). Finally, an extensive discussion of available software written in R that implements the methods described here for the statistical analysis of the three types of spatial data appears in Bivand et al. (2008).

Cross-References

- ▶ Distance and Similarity Measures
- ▶ Least Squares
- ▶ Modeling and Analysis of Spatiotemporal Social Networks
- ▶ Regression Analysis
- ▶ Spatial Networks

▶ Theory of Probability: Basics and Fundamentals
▶ Theory of Statistics: Basics and Fundamentals
▶ Univariate Descriptive Statistics

Acknowledgments The authors thank Edgar Muñoz for producing Fig. 4. The first author was partially supported by National Science Foundation Grant HRD-0932339.

References

Anselin L (1988) Spatial econometrics: methods and models. Kluwer, Dordrecht
Banerjee S, Carlin BP, Gelfand AE (2004) Hierarchical modeling and analysis for spatial data. Chapman & Hall/CRC, Boca Raton
Berke O (1999) Estimation and prediction in the spatial linear model. Water Air Soil Pollut 110:215–237
Bivand RS, Pebesba EJ, Gómez-Rubio V (2008) Applied spatial data analysis with R. Springer, New York
Chilès J-P, Delfiner P (1999) Geostatistics: modeling spatial uncertainty. Wiley, New York
Cliff AD, Ord JK (1981) Spatial processes: models and applications. Pion, London
Cressie NAC (1993) Statistics for spatial data. Wiley, New York
Daley D, Vere-Jones DJ (2003) Introduction to the theory of point processes, volume I: elementary theory and methods, 2nd edn. Springer, New York
Daley D, Vere-Jones DJ (2007) Introduction to the theory of point processes, volume II: general theory and structure, 2nd edn. Springer, New York
Diggle PJ (2003) Statistical analysis of spatial point patterns, 2nd edn. Arnold, New York
Diggle PJ, Ribeiro PJ (2007) Model-based geostatistics. Springer, New York
Gelfand AE, Diggle PJ, Guttorp P, Fuentes M (eds) (2010) Handbook of spatial statistics. Chapman & Hall/CRC, Boca Raton
Illian J, Penttinen A, Stoyan H, Stoyan D (2008) Statistical analysis and modelling of spatial point patterns. Wiley, Chichester
Journel AG, Huijbregts CJ (1978) Mining geostatistics. Academic, London
Le ND, Zidek JV (2006) Statistical analysis of environmental space-time processes. Springer, New York
LeSage JP, Pace RK (2009) Introduction to spatial econometrics. Chapman & Hall/CRC, Boca Raton
Li SZ (2009) Markov random field modeling in image analysis, 3rd edn. Springer, London
Matérn B (1986) Spatial variation. Lecture notes in statistics, 2nd edn. Springer, Berlin
Matheron G (1975) Random sets and integral geometry. Wiley, New York
Müller WG (2007) Collecting spatial data: optimum design of experiments for random fields, 3rd edn. Springer, Heidelberg
Nguyen HT (2006) An introduction to random sets. Chapman & Hall/CRC, Boca Raton
Ripley BD (1981) Spatial statistics. Wiley, New York
Rue H, Held L (2005) Gaussian Markov random fields: theory and applications. Chapman & Hall/CRC, Boca Raton
Schabenberger O, Gotway CA (2005) Statistical methods for spatial data analysis. Chapman & Hall/CRC, Boca Raton
Sjöstedt-De Luna S, Young A (2003) The bootstrap and kriging prediction intervals. Scand J Stat 30: 175–192
Stein ML (1999) Interpolation of spatial data: some theory for kriging. Springer, New York
Wackernagel H (2010) Multivariate geostatistics: an introduction with applications, 3rd edn. Springer, Berlin
Yaglom AM (1987) Correlation theory of stationary and related random function I: basic results. Springer, New York

Spatial–Temporal Data Analysis

▶ Location-Based Social Network Analysis

Spatial-Textual Web Search

▶ Spatiotemporal Information for the Web

Spatiotemporal Analysis

▶ Modeling and Analysis of Spatiotemporal Social Networks
▶ Social Networks in Emergency Response
▶ Tensor-Based Analysis for Urban Networks

Spatiotemporal Collaborative Filtering

▶ Spatiotemporal Personalized Recommendation of Social Media Content

Spatiotemporal Context-Aware Recommendation

► Spatiotemporal Recommendation in Geo-Social Networks

Spatiotemporal Data

► Spatiotemporal Topic Detection from Social Media

Spatiotemporal Footprints in Social Networks

Linna Li[1] and Michael F. Goodchild[2]
[1]Department of Geography, California State University Long Beach, Long Beach, CA, USA
[2]Department of Geography, University of California Santa Barbara, Santa Barbara, CA, USA

Synonyms

Check-in; Location; Movement; Place; Spatial interaction; Time; Trajectory

Glossary

Flickr	A popular photo-sharing website allowing people to upload and share photos that may be tagged with location
GPS	Global Positioning System, a satellite-based navigation system that provides location and time almost anywhere near the Earth's surface
Spatial interaction models	Models describing interaction between two locations as a variable dependent on distance
Twitter	A popular microblogging and social networking service that supports sending text messages (which are called tweets) of less than 140 characters. Tweets may be associated with location
VGI	Volunteered geographic information, a type of user-generated content with a spatial component (Goodchild 2007)

Definition

Spatiotemporal footprints discussed in this chapter are locational and temporal information regarding people's activities that are digitally recorded. Spatial location may be automatically captured as latitude and longitude by a GPS receiver or provided by a user as a place name, e.g., a city or neighborhood. Along with spatial information, time is usually automatically recorded, too. Spatiotemporal footprints may be intentionally collected as a series of point locations when people move around, such as GPS tracks. They may also be attached to instant messages, associated with telephone calls, tagged to photos, or linked with other forms of human activities. Footprints are basically location and time that indicate where a person went and when. Several other terms are also partially synonymous with spatiotemporal footprints. Check-in is used by social networking services, such as Foursquare, Google+, and Facebook, to allow users to instantly share their physical locations. Concatenation of spatial footprints in the order of time generates trajectories of human movements, which have been a long-lasting research topic in the social sciences, where they are studied in the context of migration, travel, commuting, etc.

Introduction

Society is comprised of many different types of social networks on various levels. Social networks play a critical role in achieving goals and solving problems. While traditional social networks only existed within a very limited geographical distance (e.g., villages) constrained by temporal factors, modern technologies – especially the growth of the Internet, the wide adoption of smartphones, and the support of Web 2.0 technologies – have greatly reduced spatiotemporal limitations on human communication. People who live on different continents in different time zones can interact with each other using phones, emails, and websites. Particularly, online social networking services provide an effective channel to enhance existing social networks and to initiate new ones. Facebook, for example, offers services to create profiles, add friends, and exchange information. Twitter, as another example, provides a platform to share and discover "what is happening right now (at where)?" These services offer an alternative and complementary form of social networks with a growing number of users.

Human activities take place in particular locations at specific times; activities in online social networks may reflect activities in the physical world. For instance, people may comment on an event that they are currently experiencing, such as a football game or a fire. On the other hand, activities in the physical world may generate a virtual social network. To facilitate organization of regular gatherings, local friends may create a virtual group on Facebook to publish information and to share photos. One of the major reasons for people to record spatiotemporal footprints of activities is to share them with family and friends, basically with people within their social networks; such sharing has generated large amounts of locational and temporal data. This phenomenon has attracted increasing attention from both academia and industry, because the prevalence of such information provides a great potential to study human mobility, human activities, and the composition of large-scale social networks, using vast volumes of geospatial data on large samples of people, for the first time in history.

Key Points

- Digitally recorded spatiotemporal footprints in social networks provide an unprecedented opportunity to study the physical environments and human activities.
- Place names ubiquitous in documents, photos, and videos should be incorporated into studying spatiotemporal footprints.
- Various degrees of uncertainty associated with spatiotemporal footprints are a critical consideration in assessing the accuracy and scope of the conclusion.
- Spatiotemporal footprints have been utilized to study place and people. Key applications focus on the characteristics of geographic features and events, people's behaviors, and human interactions within social networks.
- Research challenges and opportunities exist in many areas, including data access and synthesis, representativeness of available spatiotemporal footprints, social relations and group behaviors inferred from footprints in social networks, and privacy issues in such studies.

Historical Background

Spatiotemporal footprints have been used to study social phenomena for a long time. As early as the mid-1800s, an English physician John Snow found a water pump on Broad Street as the source of cholera by plotting footprints of people who died of the epidemic in London. One major research area that relies on spatiotemporal footprints is travel behavior. Accurate records of footprints are critical in understanding human mobility patterns. In the past few decades, travel diaries collected as part of a travel survey conducted once in a while were a major data source for such studies. These travel diaries record location and time of different trips in order to understand travel in relation to the choice, location, and scheduling of daily activities. From face-to-face interviews to mail and telephone surveys, footprints collected in travel diaries and geocoded in geographic information systems have enabled us to study people's activity patterns spatially and temporally. Such work dates back to 1944 when Liepmann obtained and analyzed the travel data on workers in England in the

1930s (Liepmann 1944). One important current data source is the National Household Travel Survey that has been conducted periodically since 1969. This traditional data collection method using travel diaries has one obvious shortcoming: it is very expensive and time-consuming. With the advancement of modern geospatial and communication technologies in the twenty-first century, it becomes increasingly easier to record location and time with little monetary or labor cost. As a result, large amounts of spatiotemporal footprints have been generated every day to complement the data collected in travel diaries. In addition, the applications of this new data source of footprints have been extended to many other areas, including the studies on the Earth's surface, the discovery of important events, and the detection of human relationships within social networks based on co-occurrences in space and time.

Place Names as Footprints

Place names are an important form of footprints although spatiotemporal footprints usually denote exact point locations or linear trajectories. An accurate location, or even a highly inaccurate location, is no longer indispensable in representing footprints. It is not uncommon that people travel to a place without knowing its exact location on the Earth's surface. An interesting example was provided by Gould and White (1986), who recalled an overheard conversation between two young women in London. "Where did you go for a holiday this year?" "Oh, we went to Majorca." "Was it nice?" "Absolutely smashing!" "Where is Majorca?" "I don't know exactly, I flew." This woman seemed to have a radiant image of Majorca without knowing its location on the Earth. To accommodate informal information like this, spatial footprints can simply be represented as place names with relevant attributes. For example, the name "Majorca" could be associated with numerous properties in this conversation, such as weather, restaurants, and hotels, without any locational reference, and the woman definitely left her footprint there.

If we incorporate place names in the representation and analyses of footprints, it is possible to create a much richer and more complete picture. Although a large number of spatiotemporal footprints are constantly being recorded by GPS units in mobile devices, voluminous footprints are mentioned and documented as place names in travel blogs, social networking websites, photo-sharing platforms, and search engines. Eventually, very few people know the latitude and longitude of their homes or the city they live in; but when the name of the neighborhood or the city is mentioned, they can quickly recognize the reference and recall various properties associated with the place. A sequence of footprints do not have to be represented as connected dots on the Earth's surface but instead as a chain of place names with relevant attributes. This also enables us to reconstruct a track of footprints from places with vague and ambiguous boundaries. Such information has been underutilized because many places are geographic objects with indeterminate boundaries or changes through time that are difficult to define as the exact polygonal extents traditionally required by GIS (Burrough and Frank 1996). In such cases, schematic maps without geometric or locational information may be appropriate. A trajectory of footprints may be represented as sequential place names with no or distorted geometries due to lack of or unnecessary locational information. As demonstrated in the London Tube map, stations on transit lines are represented in their relative positions. Distance in central London is expanded, while distance in the periphery is shrunk. Lines are straightened and directions are adjusted to horizontals, verticals, and diagonals. As a result, geometry is reduced for the sake of clear representation and reduction of cognitive load in comprehending essential information. Considering place names as footprints can also facilitate integration of geographic information from disparate sources. Linked by names, one study showed the promise of conflating different aspects of the same place described and discussed as fragmented knowledge in documents, photos, and videos scattered on the web (Gao et al. 2013).

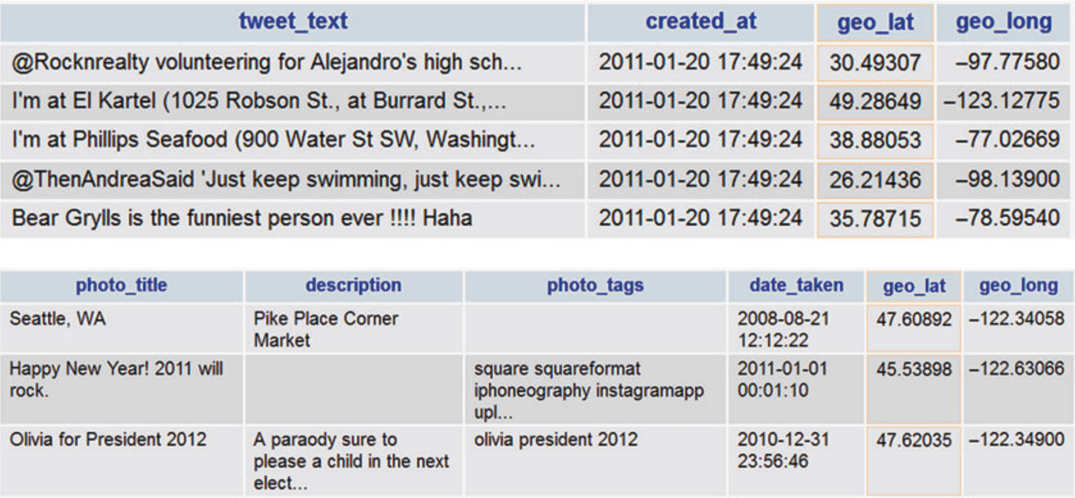
Uncertainty in Footprints

Spatiotemporal footprints are voluminous with very rich information; however, we need to be aware of the inherent uncertainty associated with them in order to validate conclusions based on these data. Spatial uncertainty is the difference between a recorded position and the corresponding position in reality. It is critical to understand uncertainty when dealing with locational information, so it has been studied extensively in the field of GIScience. For example, Goodchild and Gopal (1989) compiled a set of papers that address accuracy of spatial data from a wide range of applications, including both physical and social phenomena. Zhang and Goodchild (2002) systematically examined spatial uncertainty modeling in continuous and categorical variables. Footprints generated in social media and social networks are a type of VGI. Unlike location recorded in a scientific database that aims to minimize spatial uncertainty and inaccuracy as much as possible with standard quality control procedures, uncertainty associated with footprints may vary from case to case due to the nature of VGI (Goodchild and Li 2012). Users may choose to disclose or hide footprints and select the level of disclosure. For example, they can reveal their exact point location as coordinates, or they can show only the city name as the location.

When users of social networking services or social media share their locations, they usually have several options. If GPS is enabled in a mobile device such as a smartphone, location is recorded automatically as latitude and longitude. Otherwise, users can select a city or neighborhood name from a set of provided place names that are usually reverse geocoded based on an estimate of their device location. The degree of uncertainty in footprints depends on the mechanism used to record location. There are two major categories of footprints: those recorded by a digital system automatically and those provided by a user manually. Automatic footprints may be generated by GPS, relative location of cell phone towers, or IP address. Location produced by the same method has a similar level of uncertainty, but it varies from one method to another. Spatial footprints captured by GPS (either a GPS unit or built-in GPS in mobile phones or cameras) in the form of latitude and longitude are supposed to be the most accurate means to record location. Uncertainty of footprints recorded by GPS is usually several meters, depending on the particular device and the surrounding conditions (e.g., satellite visibility). However, the number of decimal digits of stored coordinates may not reflect their actual degree of uncertainty. For example, latitude and longitude associated with tweets in the Twitter database and photos in the Flickr database both have five decimal places (Fig. 1), indicating that spatial precision should be around 1 m, which is not the correct expectation of GPS uncertainty in most devices. In addition, approximate location may be determined by the relative position of a user's equipment in a cellular network, leading to uncertainty as large as a cell area, ranging from 150 m to 30,000 m (Zhao 2000). Physical location inferred from an IP address may be at the level of ZIP code, city, state, or even country and thus show varying accuracy. Databases have been established to map the correspondence between IP address and physical address (e.g., http://whatismyipaddress.com/ip-lookup), and efforts have been made to increase accuracy of IP address locators (Guo et al. 2009). Uncertainty of temporal footprints is less complex. Time of users' interaction with the web or mobile services is always automatically recorded with good accuracy. However, temporal information provided by a user may be arbitrary with an uncertainty that is difficult to estimate. For instance, there are various degrees of uncertainty in the times associated with photos in Flickr.

Key Applications

Numerous questions that were not answerable before due to the lack of data can now be investigated using spatiotemporal footprints. In general, footprints have been utilized to study place and people. Place refers to geographic features that are present at particular locations on the Earth's surface. A place could be a simple feature such as a road or a restaurant with a clear boundary or a



tweet_text	created_at	geo_lat	geo_long
@Rocknrealty volunteering for Alejandro's high sch...	2011-01-20 17:49:24	30.49307	–97.77580
I'm at El Kartel (1025 Robson St., at Burrard St.,...	2011-01-20 17:49:24	49.28649	–123.12775
I'm at Phillips Seafood (900 Water St SW, Washingt...	2011-01-20 17:49:24	38.88053	–77.02669
@ThenAndreaSaid 'Just keep swimming, just keep swi...	2011-01-20 17:49:24	26.21436	–98.13900
Bear Grylls is the funniest person ever !!!! Haha	2011-01-20 17:49:24	35.78715	–78.59540

photo_title	description	photo_tags	date_taken	geo_lat	geo_long
Seattle, WA	Pike Place Corner Market		2008-08-21 12:12:22	47.60892	–122.34058
Happy New Year! 2011 will rock.		square squareformat iphoneography instagramapp upl...	2011-01-01 00:01:10	45.53898	–122.63066
Olivia for President 2012	A paraody sure to please a child in the next elect...	olivia president 2012	2010-12-31 23:56:46	47.62035	–122.34900

Spatiotemporal Footprints in Social Networks, Fig. 1 Latitude and longitude associated with tweets in Twitter and photos in Flickr

vague feature without an exact agreed-upon location, such as “downtown.” The focus of this type of research is on the geographic landscape associated with footprints without explicit consideration of the people who provide them. The second category of research on footprints emphasizes the people who record them, their behavior, and the relationships that may be inferred from the pattern of their whereabouts. Here are some example research questions: How do people move around a city? What percentage of their trips is captured by footprints recorded in social networking services? What relationships can be extracted from the pattern of spatiotemporal footprints of two people or a group of people in a social network?

Locational information in footprints can be used to characterize the position and shape of geographic features. This type of footprint is usually called VGI – a special form of user-generated content with a geographic component (Goodchild 2007). A typical example is OpenStreetMap (http://www.openstreetmap.org/), with a goal to create a free editable map of the world. Originally, map data were contributed by volunteers using a handheld GPS to record their walking or biking paths. People purposefully collect continuous footprints to produce geographic infrastructure data, mostly of roads and points of interest (POIs). In this case, spatial footprints are used to identify the location of geographic features: what is available at that particular location on the Earth’s surface? People may choose to map whatever features that are interesting to them. For instance, SeeClickFix, a web service supported in both web browsers and mobile apps, enables citizens to report the location of nonemergency issues within their communities (e.g., a broken traffic light), while governments use this information to respond more promptly to the problems and take actions to fix them. In addition to geographic features created by individual users, places may also be inferred by aggregating spatial footprints generated by multiple people. Clusters of spatial footprints suggest popular places, and clusters of footprints in both space and time may indicate events. For example, location and spatial extent of specific places can be constructed as a probability-density surface by extracting and summarizing footprints of photos tagged with the same place name (Li and Goodchild 2012). Spatial boundaries of city cores (e.g., downtown, CBD) may also be defined based on photo footprints obtained from Flickr (Hollenstein and Purves 2010). Lee and Sumiya (2010) proposed a method to detect unusual geo-social events (e.g., local festivals) by comparing spatiotemporal patterns of footprints in the study area with the distribution of footprints in normal times.

Spatiotemporal footprints can also be used in combination with other information to gain knowledge about the Earth. Together with visual information in photos, representative scenes at different locations were automatically selected from vast volumes of geotagged photos in Flickr (Crandall et al. 2009). In another type of application, location is attached to collected data about some natural phenomenon as a systematic spatial sampling strategy to facilitate scientific data analysis. One famous example is the Audubon Society's Christmas Bird Count project that started in 1900. Volunteer birdwatchers are divided into small groups to follow assigned routes and to count birds they see along the routes. Moreover, locational information with regard to disaster status contributed by citizens has been proved very helpful in emergency response (Li and Goodchild 2010), such as wild fires (Goodchild and Glennon 2010). A spatial model was proposed to estimate the location of an earthquake and the trajectory of a typhoon in Japan based on georeferenced tweets (Sakaki et al. 2010). Recently, Panteras et al. (2015) used both Twitter and Flickr to study the impact area of a natural disaster.

Furthermore, footprints are also used to study the people who generate them. Concatenation of spatiotemporal footprints of a single user provides a trajectory of the places he or she has visited at specific time of the day, which may shed light on people's daily activities. We may collect data automatically on places people have been to or how long they stay at a particular place. For example, georeferenced tweets may provide a real-time record of people's activity episodes that is even more accurate than a travel diary recorded from memory recall. Besides, this data collection is nonintrusive, so subjects do not need to write down what they are doing and at what time, because the time of a tweet is recorded automatically by the online service and the tweet content may suggest their activities. Traditionally, travel behavior was studied using travel diaries that were a part of a travel survey, which is very expensive in terms of time and labor. Research has already been done on methods to use recorded footprints as complementary to traditional self-reported travel surveys in studying travel behavior. Murakami and Wagner (1999) discussed the use of GPS to automatically collect date, start time, end time, and vehicle position in trips at frequent intervals. A comparison between GPS-recorded trips and self-reported trips shows that self-reported distances are much longer than the actually traveled distances. In addition to locational and temporal information stored by GPS in trips, even purposes of trips may be inferred from footprints with auxiliary land-use data (Wolf et al. 2001). Crandall et al. (2009) reconstructed the pathways of people who visit Manhattan and the San Francisco Bay area based on footprints associated with photos uploaded to Flickr. Also relying on geotagged photos in Flickr, a routing recommendation system was developed to suggest popular landmarks for travel planning (Sun et al. 2015). Although footprints recorded by GPS have been used to study travel behavior and trajectories may be extracted from photo footprints, not much research has been done to investigate detailed travel behavior at the level of traditional travel surveys using footprints collected in social networks and social media.

Comparison of spatiotemporal footprints between different users or groups of people may indicate the relationship between them in social networks. According to the first law of geography, "Everything is related to everything else, but near things are more related than distant things" (Tobler 1970). If this is true for social phenomena, we can infer the strength of social interactions between people at two places based on their spatial footprints. Distance between people may signify the probability of them being friends or acquaintances. Spatial interaction models have been developed to describe this relationship (Isard 1960). The distance decay effect is characterized this way: as the distance between two locations increases, the interaction between them decreases. A typical example is a gravity model (Abler et al. 1971):

$$I_{ij} = a\frac{M_i M_j}{d^b{}_{ij}} \tag{1}$$

where I_{ij} is the interaction between i and j, a is a constant, M_i and M_j are properties associated with i and j, d_{ij} is the distance between i and j, and *b* is another constant, dependent on the phenomenon.

Researchers have started to investigate the role of distance derived from footprints in studying social relations. A collection of maps were produced to represent cyberspace, including online communications and connections between people located in different places (Dodge and Kitchin 2001). Strong connections may exist between people who visit the same place at the same time regularly. Using geotagged photos from Flickr, the probability of a social tie between people is calculated based on the co-occurrences of their spatial and temporal footprints (Crandall et al. 2010). Cho et al. (2011) studied the relationship between social ties and people's movement patterns using data collected from public check-ins in online social networks and cell phone location trace data. A study on mobile phone data demonstrates that distance decay is present in the number of calls and the number of co-locations, defined as people sharing the same location at the same time (Calabrese et al. 2011). Hardy et al. (2012) applied a gravity model to describe the decrease of the likelihood of a person to contribute to a georeferenced article in Wikipedia when the distance between the user and the subject place in the article increases.

Future Directions

Although voluminous amounts of footprints are generated in social networks every day, data discovery and data access are still two important considerations. How can we find relevant geographic data in social media? How can we collect more spatial and temporal footprints of social networks from spatially embedded populations, both online and offline? How can geography promote the "human as sensor" paradigm in spatial data generation? How can we harvest spatiotemporal footprints about involved people from existing sources?

Another critical question is the synthesis of footprints with various accuracies generated in different contexts. How can we quantify the uncertainty in a particular footprint? Can we apply mathematical models of uncertainty developed in the GIScience literature to study uncertainty in footprints? How can we use footprints that are available as both coordinates and place names to do cross validation, so as to increase spatial accuracy? Currently, research has been done with only a single source of footprint data (i.e., footprints created in one social networking service). It would be valuable to investigate the potential for using footprints collected from different sources to improve the data quality and quantity. Geographic data conflation has been applied to merge spatial data from multiple sources (Saalfeld 1988; Li 2010). Can these techniques be used in integration of spatiotemporal footprints generated in various social networks? What type of new methodologies might be required in footprint data synthesis?

Furthermore, representativeness of the available footprints is an interesting yet challenging research area. Since systematic sampling strategies are not applied in the collection of footprints, how representative are these data compared to the total population under study? What are the major types of motivation of people who join online social networks? The usual users of social media are undoubtedly self-selected. What characteristics cause them to join online social networks and to leave spatiotemporal footprints? Is there a way to measure the bias in this type of data source?

More analyses could be performed using footprints generated in social networks. How can we identify social relations and mobile patterns from heterogeneous footprint data? Social networks are embedded in space and time. However, that embedding may not always be relevant to specific analysis. How can we incorporate spatial interaction functions into different types of network, particularly when space and time vary significantly? For example, spatial dependence and distance decay are valid in many processes in geographic space but are less relevant when all people in the social network are in the same room. While it is known that social network links decay with distance and that new technologies do not completely overcome this decay, what is not known is the circumstances where the technologies overcome the decay (e.g., where in a task cycle). Moreover, the predictive value of knowing that network links fall off with distance seems low. What can be predicted with a better understanding of the relationship between space and networks? What are the types of activities in social networks that are strongly constrained by space and time? For what type of groups is group maintenance and persistence dependent on spatial locations? How can we use spatiotemporal footprints to infer missing network data, select specialized social network subgroups, and forecast change? How are network-mediated processes (e.g., information diffusion, VGI) influenced by spatial and temporal relations (e.g., nearness in space or time)?

Finally, privacy in social network studies has attracted much attention. When is it appropriate to collect information on people's footprints and to study them without their knowledge? Revelation of locations may lead to crimes, such as stalking and burglary. Is there a way to preserve spatiotemporal patterns of social networks and to protect privacy simultaneously? What types of generalization and aggregation from statistics and cartography can be adapted to achieve the two objectives? What would be an appropriate level of generalization of locational data for a particular application? How does the level of abstraction limit the types of network questions that can be answered?

Cross-References

- ▶ Detection of Spatiotemporal Outlier Events in Social Networks
- ▶ Spatiotemporal Information for the Web
- ▶ Spatiotemporal Proximity and Social Distance

References

Abler R, Adams J, Gould P (1971) Spatial organization—the geographer's view of the world. Prentice-Hall, Englewood Cliffs

Burrough PA, Frank AU (eds) (1996) Geographic objects with indeterminate boundaries. Taylor and Francis, London

Calabrese F, Smoreda Z, Blondel VD, Ratti C (2011) Interplay between telecommunications and face-to-face interactions: a study using mobile phone data. PLoS One 6(7):e20814

Cho E, Myers S, Leskovec J (2011) Friendship and mobility: user movement in location-based social networks. In: KDD, pp 1082–1090. ACM.

Crandall DJ, Backstrom L, Huttenlocher D, Kleinberg J (2009) Mapping the world's photos, Proceedings of the 18th international conference on World wide web, April 20–24, 2009, Madrid

Crandall DJ, Backstrom L, Cosley D, Suri S, Huttenlocher D, Kleinberg J (2010) Inferring social ties from geographic coincidences. Proc Natl Acad Sci 107(52):22436–22441

Dodge M, Kitchin R (2001) Mapping cyberspace. Routledge, New York

Gao, S., Janowicz, K., McKenzie, G., & Li, L. (2013). Towards platial joins and buffers in place-based GIS. In Proceedings of the 1st ACM SIGSPATIAL international workshop on computational models of place (COMP' 2013) , November 5, 2013. Orlando, FL, USA. ACM

Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. Geo J 69:211–221

Goodchild MF, Glennon JA (2010) Crowdsourcing geographic information for disaster response: a research frontier. Int J Digit Earth 3(3):231–241

Goodchild MF, Gopal S (1989) Accuracy of spatial databases. Taylor and Francis, New York

Goodchild MF, Li L (2012) Assuring the quality of volunteered geographic information. Spat Stat 1:110–120

Gould P, White R (1986) Mental maps, 2nd edn. Routledge, London

Guo, C., Liu, Y., Shen, W., Wang, H. J., Yu, Q., & Zhang, Y. (2009, April). Mining the web and the internet for accurate ip address geolocations. In INFOCOM 2009, IEEE (pp. 2841–2845). 19–25 April 2009, Rio de Janeiro, Brazil, IEEE

Hardy D, Frew J, Goodchild M (2012) Volunteered geographic information production as a spatial process. Int J Geogr Inf Sci 26:1191–1212

Hollenstein L, Purves R (2010) Exploring place through user-generated content: using Flickr to describe city cores. J Spat Inf Sci 1(1):21–48

Isard W (1960) Methods of regional analysis. MIT Press, Cambridge

Lee R, Sumiya K (2010) Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In: Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks (LBSN 2010), pp 1–10

Li L (2010) Design of a conceptual framework and approaches for geo-object data conflation. PhD dissertation, Department of Geography, University of California, Santa Barbara

Li L, Goodchild MF (2010) The role of social networks in emergency management: a research agenda. Int J Inf Sys Crisis Response Manag (IJISCRAM) 2(4):49–59

Li L, Goodchild MF (2012) Constructing places from spatial footprints, In: Proceedings of ACM SIGSPATIAL GEOCROWD'12, November 6, 2012. Redondo Beach

Liepmann K (1944) The journey to work. K. Paul, Trench, Trubner, London

Murakami E, Wagner D (1999) Can using global positioning system (GPS) improve trip reporting? Transportation Res C 7:149–165

Panteras G, Wise S, Lu X, Croitoru A, Crooks A, Stefanidis A (2015) Triangulating social multimedia content for event localization using flickr and twitter. Transactions in GIS 19(5):694–715

Saalfeld A (1988) Conflation Automated map compilation. Int J Geogr Inf Sys 2(3):217–228

Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes Twitter users: real-time event detection by social sensors, In: Proceedings of the 19th international conference on World wide web, April 26–30, 2010, Raleigh, North Carolina

Sun Y, Fan H, Bakillah M, Zipf A (2015) Road-based travel recommendation using geo-tagged images. Comput Environ Urban Sys 53:110–122

Tobler W (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46 (2):234–240

Wolf J, Guensler R, Bachman W (2001) Elimination of the travel diary: an experiment to derive trip purpose from GPS travel data. In: Proceedings from the transportation research board 80th annual meeting, Washington DC

Zhang J-X, Goodchild MF (2002) Uncertainty in geographical information. Taylor and Francis, New York

Zhao Y (2000) Mobile phone location determination and its impact on intelligent transportation systems. Intelligent Trans Sys IEEE Trans On 1(1):55–64

Spatiotemporal Information for the Web

Peiquan Jin, Sheng Lin and Qingqing Zhang
School of Computer Science and Technology,
University of Science and Technology of China,
Hefei, China

Synonyms

Spatial-textual Web search; Spatiotemporal Web; Temporal-textual Web search.

Glossary

AD	Attribute descriptor
GEO/GEO Ambiguity	It refers that many locations can share a single place name
GEO/NON-GEO Ambiguity	It refers that a location name can be used as other types of names
GRT	Global reference time
LD	Location descriptor
Location	A site name or geographic scope mentioned in Web pages
LRT	Local reference time
NER	Named entity recognition
OID	Object identifier
Primary Location	The most appropriate location associated with a Web page
Primary Time	The most appropriate time associated with a Web page
Search Engine	Search engine is a popular tool to find information in the Web
TD	Time descriptor
Time	One or more units of chronons. It can be a time instant or a time period

Definition

This subject is mainly towards the spatiotemporal information involved in the Web, particularly in Web pages. Typical spatiotemporal information in the Web includes the locations and time mentioned in Web pages, the update date of Web pages, and the Web server locations. As we know, location and time are the essential dimensions of information including Web information. However, they are usually ignored in traditional keyword-based Web search engines.

Traditional search engines are basically based on keyword-based approaches or content-based methods. Though many contributions have been presented in both directions, in some cases users are still difficult to express their search needs. For example, more than 70% Web queries are related with time and locations

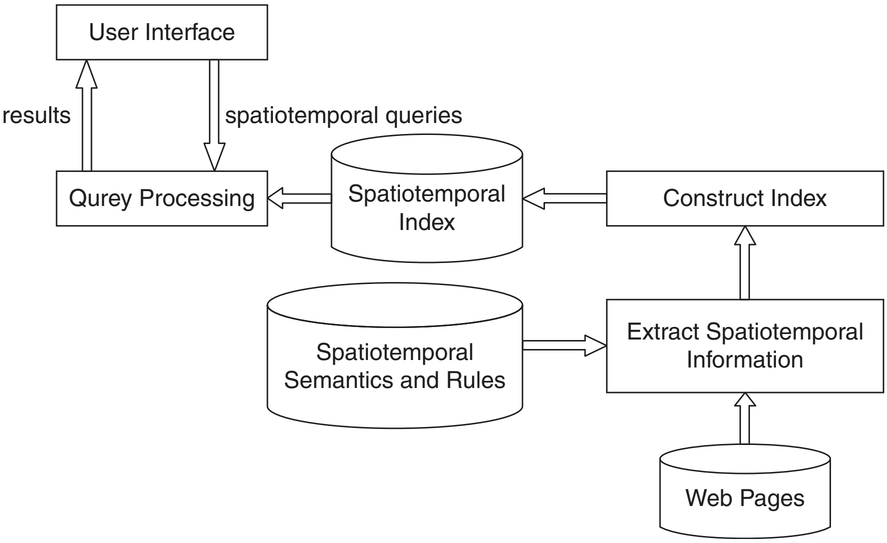
(Setzer and Gaizauskas 2002; Sanderson and Kohler 2004), but spatiotemporal Web queries such as “to get the news about Olympic Beijing in recent three days” or “to get the sales information about Nike in Beijing in this week” are often with bad results in traditional search engines. One reason is that such queries are difficult to express in keyword-based search engines. Moreover, traditional search engines also lack of the ability to process such spatiotemporal queries.

Aiming at improving the effectiveness and efficiency of spatiotemporal queries in search engines, many researchers began to study the spatiotemporal information in the Web. However, most of previous researches focused on time-based Web search (Nunes et al. 2008) and location-based Web search (Wang et al. 2005; Ding et al. 2000; Zhou et al. 2005; Markowetz et al. 2005) separately. And few works considered the temporal information of the content in Web pages. In this entry, we will describe the semantics of spatiotemporal information in the Web and try to present a framework for spatiotemporal information extraction under the Web context.

Introduction

The main goal of incorporating spatiotemporal semantics into search engines is to develop a search engine that is able to express and process spatiotemporal queries. Figure 1 shows the framework of a spatiotemporal Web search engine. In this framework, spatiotemporal information is first extracted from Web pages based on spatiotemporal semantics and rules, and then we use them to construct a spatiotemporal index for Web pages, using the extracted spatiotemporal information. Users can input location- or time-related queries through the user interface, and the query processing engine will interpret the queries and perform an index-based search on archived Web pages. Finally, the resulted Web pages are returned to users according to an improved ranking algorithm, which combines text ranking techniques with new temporal and spatial ranking mechanisms.

Spatiotemporal information has been deeply studied in spatiotemporal database area, in which moving geographic objects are concentrated. However, they are not popular in Web context. So at present, the main focus on spatiotemporal information in the Web is to integrate location and



Spatiotemporal Information for the Web, Fig. 1 The framework of Web search system based on spatiotemporal semantics time information into search process, such as information extraction, indexing, querying, ranking, and visualization.

In this entry, we focus on the spatiotempo-ral semantics of Web information, mainly of Web pages, and present a framework to represent and extract the spatiotemporal information in the Web.

Key Points

The spatiotemporal semantics of Web pages refer to the ontological meaning of the time and location information of Web pages. The Web can be regarded as a database of Web pages, so a Web page can be looked as an object in the Web. According to the object-oriented theory, an object consists of a unique identifier and other attributes that describe the properties of the object. Based on this view, a Web page is a spatiotemporal object which contains the following parts:

- Identifier: The identifier of a Web page is usually the URL.
- Locations: The spatial information of a Web page may consist of two types of locations, which are provider location and content locations (Wang et al. 2005). The provider location refers to the physical location of the provider who owns the Web resource. The content locations are the geographic locations that are described in the content of a Web page.
- Time: The temporal information of a Web page has two types: update time and content time. The update time is the latest modified time of a Web page. The content time is the time that the content of a Web page indicates. The content time may contain implicit time such as "Today" and "Three Days Ago".
- Non-spatiotemporal attributes: The non-spatiotemporal attributes of a Web page refer to the traditional keywords set of the Web page.

Historical Background

Most Web pages contain location and time information. Previous works regard the locations in Web pages as geographic scope (Ding et al. 2000), which can be determined by analyzing the content and links in the Web page. The locations in a Web page usually have spatial containment relationships. For example, "China" contains "Beijing." In the literatures (Zhou et al. 2005), a classification framework for Web locations is presented, and an algorithm to extract the locations in Web pages is further proposed. In order to support spatial computation, they use MBRs (Minimal Bounding Rectangle) to represent the geographic scope of Web pages. There are also some other methods proposed to represent geographic scopes, such as raster-based representation (Markowetz et al. 2005). Generally, the MBR-based method is widely used (Lee et al. 2003; Ma and Tanaka 2004). One problem of those previous works is that they treat the geographic scope of a Web page as exact one MBR, which is not very precise for many Web pages.

Temporal information is also very common in Web pages, especially in news pages. Temporal information extraction first appeared in MUC-5 whose task was to extract from business news when a joint venture took place. In MUC-6 some research was done on extracting absolute time information as part of general tasks of named entity recognition (Sundheim and Chinchor 1995). In MUC-7, the notion of temporal information extraction was expanded to include relative time in named entities (Chinchor 1998). MUC is practically the pioneer and prime driver of temporal information extraction research.

The temporal information of a Web page refers to the time related with it, e.g., the created date of the Web page and the date of an event reported in the Web page. There are many representation forms for the temporal information in Web pages, such as yesterday, Christmas, and August 15, 2012. Besides, many Web queries are time sensitive. Fresh Web pages have more important roles when users are searching news or sales information.

Proposed Solution and Methodology

In this section, we introduce the approach in capturing spatiotemporal information in Web search. The proposed approach consists of three main components: (1) Semantic Modeling for Spatiotemporal Information in the Web, (2) Extracting Primary Location from the Web, and (3) Extracting Primary Time for Web Pages.

Semantic Modeling for Spatiotemporal Information in the Web

From an object-oriented perspective, a Web page can be defined as follows:

Definition 1 A Web page is a quintuple O = <OID, LD, TD, AD>, where OID (Object IDentifier) is the identifier of the Web page, LD (Location Descriptor) is the location descriptor describing the location information of the Web page, TD (Time Descriptor) is the time descriptor describing the temporal information of the Web page, and AD is the attribute descriptor which describes the non-spatiotemporal properties of the Web page.

Figure 2 shows the spatiotemporal semantic model of Web pages.

Location Descriptor

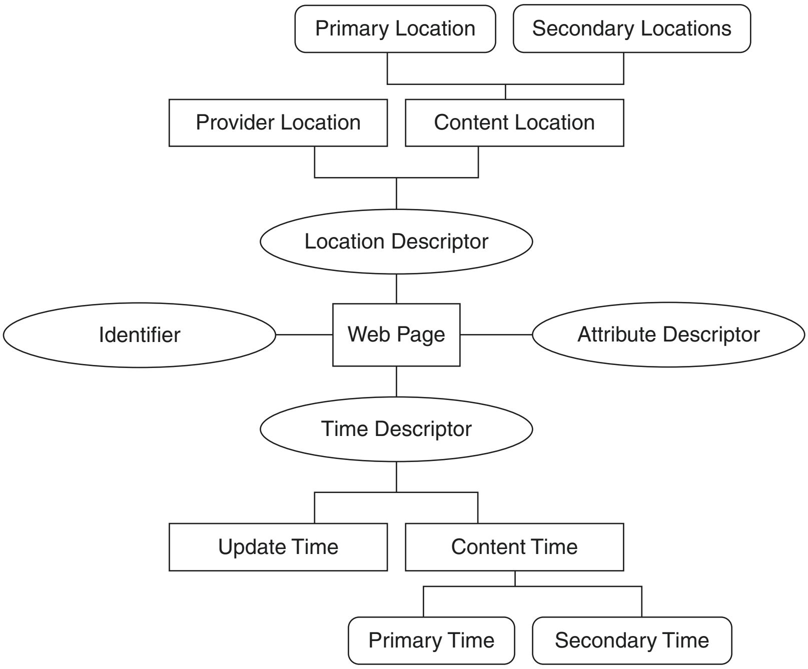
Location descriptor represents the location information of a Web page. A Web page has a unique provider location which is the geographic location of the Web server containing the Web page. The location information that described in the content of a Web page is called content locations. For example, in a company's homepage, the provider location may be "Beijing," since the Web server containing the homepage is located in Beijing, while the content locations may include the address of the company and other locations. As many locations may be involved in the content of a Web page, we should define a primary location for the content of a Web page. The primary location is the most appropriate location that describes the location information of a Web page. In the previous example, the primary location of the Web page could be the address of the company. However, how to compute the primary location of a Web page is an unrevealed issue in location-based Web search area.

Time Descriptor

Time descriptor represents the temporal information of a Web page. There are two types of temporal information related to Web pages:

- Update Time. This refers to the update time of the corresponding file of a Web page. For a

Spatiotemporal Information for the Web, Fig. 2 Spatiotemporal semantic modeling for a Web page



given Web page, the update time is unique and can be regarded as the timestamp of the Web page. Whenever a Web page is updated, the update time is also renewed.

- Content Time. This refers to the involved temporal information in the text content of a Web page. Compared with update time, which is unique and explicit for a specific Web page, the content time is a set of time instant or time period which may be explicit or implicit. For example, a news page may contain the explicit published time “2008-1-24” of the news in the title. Meanwhile, in the news body, there may have some temporal keywords such as “three days ago” and “today.” The implicit content time should be translated into calendar time. Among the many time instants and periods described in the content of a Web page, we also need to define and compute the primary time of the Web page. The primary time of a Web page is the most appropriate time related to the Web page. In time-based Web search engine, primary time and secondary time should be treated and searched in different ways.

  The above classification on the Web page time mainly considers the role of time in Web pages. Upon another view on time structure, there are two types of time: instant and period.

- Instant. Instant is a specific point in the timeline. An instant may be a second, e.g., “2008-04-01 11:59:59.” It also can be a time point related to current time, e.g., “one hour ago” means the time instant which is 1 h before current time.
- Period. Period is time duration. It contains a pair of instants and represents the time duration between the instants. For example, “[2000-09-01 00:00:00, 2003-02-01 00:00:00]” represents the time duration from “2000-09-01 00:00:00” to “2003-02-01 00:00:00,” and “[2002-09-01 00:00:00, NOW]” indicates the time duration since “2002-09-01 00:00:00.”

Another issue when considering the temporal semantics of Web pages is the granularity of the time. Different events in Web pages will have different granularities, e.g., the foundation event of a company may use “day” as the granularity, while a news report about earthquake may use “second.” How to set up a unified referential framework for the temporal granularity is a critical issue in the spatiotemporal information modeling of Web pages.

Attribute Descriptor

Attribute descriptor describes the text keywords that mostly depict the content of a Web page. Generally, it consists of a set of keywords which are extracted from the Web page. Many traditional technologies can be used to construct the attribute descriptor of a Web page, such as word segment and keyword extraction in commercial search engines.

Extracting Primary Locations from the Web

Most Web pages are associated with certain locations, e.g., news report and retailer promotion. Therefore, how to extract locations for Web pages and then use them in Web search process has been a hot and critical issue in current Web search.

As a Web page usually contains two or more location words, it is necessary to find the primary locations of the Web page. The primary locations represent the most appropriate locations associated with contents of a Web page. Generally, we assume that each Web page has several primary locations. The most difficult issue in determining primary locations is that there are GEO/GEO and GEO/NON-GEO ambiguities existing in Web pages. The GEO/GEO ambiguity refers that many locations can share a single place name. For example, Washington can be 41 cities and communities in the USA and 11 locations outside. The GEO/NON-GEO ambiguity refers that a location name can be used as other types of names, such as person names. For example, Washington can be regarded as a person name as George Washington and as a location name as Washington, D.C. Mark Sanderson et al.’s work (2000) shows that 20–30% extent of error rate in location name disambiguation was enough to worsen the performance of the information retrieval methods. Due to those ambiguities in Web pages, previous research failed to reach a satisfied performance in primary location extraction.

On the other side, it is hard to resolve the GEO/GEO and GEO/NON-GEO ambiguities as well as to determine the primary locations of Web pages through the widely studied named entity recognition (NER) approaches. Current NER tools in Web area aim at annotating named entities including place names from Web pages. However, although some of the GEO/NON-GEO ambiguities can be removed by NER tools, the GEO/GEO disambiguation is still a problem. Furthermore, NER tools have no consideration on the extraction of the primary locations of Web pages. Basically, the NER tools are able to extract place names from Web pages, which can be further processed to resolve the GEO/GEO ambiguities as well as the GEO/NON-GEO ones. Thus, we will not concentrate on the NER approaches but on the following disambiguation and primary location determination. Those works differ a lot from traditional NER approaches.

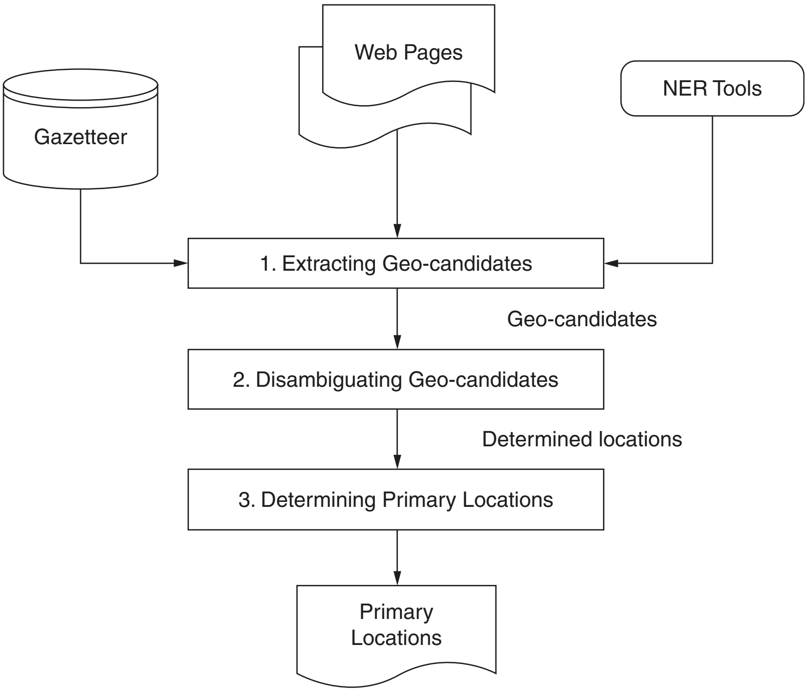
The General Framework

Figure 3 shows the general process to extract primary locations from Web pages, in which we first extract geo-candidates based on Gazetteer and NER (named entity recognition) techniques. After this procedure, we get a set of geo-candidates. In this set, the relative order of candidates is the same as that in the text. Here, geo-candidates are just possible place names, e.g., “Washington.” Then, we run the disambiguation procedure to assign a location for each GEO/GEO ambiguous geo-candidate and remove GEO/NON-GEO ambiguous geo-candidates. A location means a concrete geographic place in the world, e.g., USA/Washington, D.C. As a geo-candidate may refer to many locations in the world, the GEO/GEO disambiguation will decide which is the exact location that the geo-candidate refers to, and the GEO/NON-GEO disambiguation is going to determine whether it is a location or not. Finally, we present an effective algorithm to determine the primary locations among the resolved locations.

Geo-candidates Disambiguation

As Fig. 3 shows, we get a set of geo-candidates before the disambiguation procedure. We assume that all geo-candidates are associated with the locations in the Web page.

Spatiotemporal Information for the Web, Fig. 3 The general process to extract primary locations from Web pages


Web Pages
Gazetteer
NER Tools
1. Extracting Geo-candidates
Geo-candidates
2. Disambiguating Geo-candidates
Determined locations
3. Determining Primary Locations
Primary Locations

Basically, we assume there are n geo-candidates in a Web page and totally N locations that those geo-candidates may refer to. Then the GEO/GEO disambiguation problem can be formalized as follows:

Given a specific geo-candidate G, determining the most appropriate location among its possible locations.

We use a basic idea similar to PageRank (Brin and Page 1998) to resolve the GEO/GEO ambiguity, which is named GeoRank. The PageRank algorithm introduced an iterated voting process to determine the ranking of a Web page. We also regard the GEO/GEO disambiguation in a Web page as a voting process. Figure 4 shows the similar problem definition between PageRank and our GeoRank algorithm. Specially, in GeoRank, nodes are the possible locations corresponding to geo-candidates, and a linkage from one node A to another node B is marked with a score of evidence which represents A's voting for B to be the right location for the given geo-candidate.

In detail, as a geo-candidate can give more evidence to the one near to it in a Web page (text contribution) and a location can give more evidence to the one near to it in the geographic context (geographic contribution), we first construct a matrix M involving all locations (with each location occupies one row and one column), whose values are scores of each location of each geo-candidate voted by other ones that belong to different geo-candidates. This procedure is much like the voting process in PageRank, except that the items in M are locations but not Web pages and the scoring policy is based on text contribution and geographic contribution but not based on Web links.

Named entity recognition tools usually can remove some types of the GEO/NON-GEO ambiguities in a Web page. In order to get an improved performance, we propose two additional heuristics to further resolve GEO/NON-GEO ambiguities.

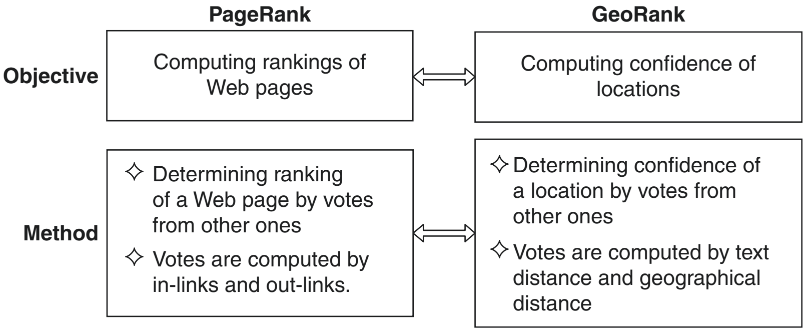
Rule 1: In the matrix *M*, if a location of a geo-candidate gets score averagely from all locations of other geo-candidates, it is not considered as a location, because none of any possible location of any other geo-candidate can give evidence to locations of this geo-candidate.

Rule 2: After removing the GEO/GEO ambiguity, if a non-country location does not have the same country with any other location, it is considered not a location. Here we get the rule from our observation that a Web page is unlikely to mention a non-country location that does not share a same country with any other locations.

Determining Primary Locations

In this stage, we calculate the scores of all the locations after disambiguation and then return the focused ones for the Web page. We consider three aspects when computing the scores of a location, namely, the term frequency, position, and geographic contributions (the contributions from locations geographically contained by the location). The motivation of the geographic contribution is that if there are many states of USA in a Web page, the location "USA" will receive

Spatiotemporal Information for the Web, Fig. 4 PageRank vs. GeoRank


PageRank
GeoRank
Objective
Computing rankings of Web pages
Computing confidence of locations
Method
✧ Determining ranking of a Web page by votes from other ones
✧ Votes are computed by in-links and out-links.
✧ Determining confidence of a location by votes from other ones
✧ Votes are computed by text distance and geographical distance

contributions from those states, as those states are all geographically contained in the USA. As a result, we use an explicit score to represent the term frequency of a location name and an implicit score for the geographic contribution. The score of a location is determined by its explicit score and implicit score.

For a location D_i its explicit score, denoted as ES (D_i), is defined as the term frequency of D_i in the Web page.

Then we use the following heuristics to modify ES (D_i):

1. If D_i follows on the heels of the other location D_j and D_i has some relationship with D_j, suppose D_j is contained in D_i, then we think the appearance of D_i in the page will emphasize D_j, so we take 0.5 away from D_i and add it to D_j, i.e., ES $(D_i) = ES(D_i) - 0{:}5$, $ES(D_j) = ES(D_j) + 0.5$.
2. If D_i appears in the title of a Web page, then we add half of SUM to D_i to emphasize this appearance, where SUM is the sum of all the ES values, as defined in the formula (1):

$$\mathrm{SUM} = \sum_{i=1}^{n} ES(D_i) \tag{1}$$

For the implicit scores, since many locations that appear in one Web page usually have some geographic relationships, we take this feature when computing the implicit score of a location. In particular, we add some contributions from those locations contained by the given location into the score. Suppose a location D_i contains n sub-locations in the Gazetteer: S_1, S_2; … S_n, and the former m sub-locations appear along with D_i in the Web page, then those m sub-locations will provide geographic contributions to D_i. The implicit score of D_i is defined in the formula (2) and (3)

$$IS(D_i) = \sum_{k=1}^{m} \left(ES(S_k) + IS(S_k)\right)^{*} \frac{m}{n^{*} diff} \tag{2}$$

$$diff = \frac{\mathrm{avg}(S_1, S_2, \ldots, S_m)}{\max(S_1, S_2, \ldots, S_m)} \tag{3}$$

Here, diff refers to the score difference among S_1, S_2; … S_m. The average value of S_1, S_2; … S_m must be less than or equal the maximum value of them, so diff ≤ 1. If D_i contains no sub-locations, then $IS(D_i) = 0$.

Based on a Gazetteer, we can build a hierarchy location tree. Then we start from the leaf nodes and compute the scores of all locations. After that, we sort all locations according to their scores and partition locations into three groups based on the scores. The first group with the highest scores is determined as the primary locations.

Extracting Primary Time from the Web

The various forms of temporal expressions in Web pages impose some challenging issues to temporal information extraction within the scope of Web search:

1. How to determine the right temporal information for implicit expressions contained in Web pages? Differing from the explicit expressions, which can be directly found in a calendar, the implicit expressions need a transformation process and usually a referential time is required.
2. How to determine the primary time for a Web page? A Web page may contain a lot of temporal information, but which ones are the most appropriate times associated with the Web page? This is very important to temporal-textual Web search engines which support both term-based and time-based queries, as they aim at finding "the Web pages associated with the given terms and under the given temporal predicate." For instance, to answer the query specifying "finding the information about tourism during the National Day," the search engines have to first determine which Web pages are mostly related with "the National Day."

For the first issue, namely, implicit time resolution, the difficult part is to select the referential time which is used to resolve implicit expressions. For example, to determine the exact time of the implicit expression "Yesterday" in a Web page, we must know the date of NOW under the context.

For the second issue, namely, primary time determination, the difficult part is to develop an effective scoring technique to measure the importance and relevance of the extracted temporal information. As there may be some containment relationship among temporal information, the time ranking task has to consider both frequency and the temporal containment. For instance, suppose "April, 2011" and "17 April, 2011" are two extracted time words, and "17 April, 2011" is contained in "April, 2011." Therefore, even "April, 2011" rarely appears in the Web pages, it will still be the primary time for the page in case that there are a great number of extracted time words contained by "April, 2011."

We focus on the above two issues and aim to propose effective solutions to the resolution of implicit expressions and the extraction of the primary time for Web pages. The main ideas can be summarized as follows:

1. We propose a new dynamic approach to resolve the implicit temporal expressions in Web pages. We classify the implicit expressions into global and local temporal expressions and then use different methods to determine the referential time for global expressions and local expressions.
2. We present a score model to determine the primary time for Web pages. Our score model takes into account both the frequency of temporal information in Web pages and the containment relationship among temporal information.

Temporal Expressions Extraction

Temporal expressions in Web pages can be generally classified into two categories.

- Explicit Temporal Expressions. These temporal expressions directly describe entries in some timeline, such as an exact date or year. For example, the token sequences "December 2004" or "September 12, 2005" in a document are explicit temporal expressions and can be mapped directly to chronons in a timeline.
- Implicit Temporal Expressions. These temporal expressions represent temporal entities that can only be anchored in a timeline in reference to another explicit or implicit, already anchored temporal expression. For example, the expression "today" alone cannot be anchored in any timeline. However, it can be anchored if the document is known to have a publication date. This date then can be used as a reference for that expression, which then can be mapped to a chronon. There are many instances of implicit temporal expressions, such as the names of weekdays (e.g., "on Thursday") or months (e.g., "in July") or references to such points in time like "next week" or "last Friday."

The explicit temporal expressions can be recognized by many time annotation tools, such as TempEx and GUTime (GUTime 2012). The temporal expressions in the GUTime output are annotated with TIMEX3 tags, which is an extension of the ACE 2004 TIMEX2 annotation scheme (tern.mitre.org).

For the extraction of implicit temporal expressions, the biggest difference of recognition between the explicit and implicit temporal expressions is that the implicit temporal expressions need to determine a reference time, so choosing the right reference is the key to the identification of the implicit temporal expression. The reference time can either be the publication time or another temporal expression in the document. Although the GUTime has a good performance in the extraction of explicit temporal expressions, it does not perform very well in dealing with the implicit temporal expressions, especially in the case of lacking of the document publication time. To improve the GUTime performance, we need to improve the reference-choosing mechanism of GUTime.

In this entry, we suppose that an implicit time expression consists of a modifier and a temporal noun which is modified by the modifier. For instance, a news report is as follows:

"(Beijing, May 6, 2009) B company took over A company totally on March 8, 2000". After 1 week, B company listed in Hong Kong, and became the first listed company in that industry. However, owing to the decision-making mistakes in the leadership and the company later poor management, B company got into debt for several 100 million dollars, and was forced to announce bankruptcy this Monday.

In this news report, “ten days” is a temporal noun, but “ten days ago” is modified after adding the modifier “ago.” For the two temporal expressions that hold reference relations in this text, “after one week” and “this Monday,” we can achieve the anchor direction easily from the modifiers through some mapping rules. Meanwhile, the offsets are able to understand directly by machine with pattern matching. But for the anchor points (referents), we must build the context-dependent reference reasoning to trace them. The full temporal reference comes from two parts: modifier reference and temporal noun reference. Because the former is inferred from the latter, the temporal noun reference reasoning plays more important roles in normalizations. Actually, we notice that the temporal noun can be classified into two classes according to the reference attributes. One is called Global Time (GT) whose temporal semantics is independent with the current context and takes the report time or publication time as the referent. Another one, Local Time (LT), makes reference to the narrative time in text above on account of depending on the current context.

In our approach, there is a reference time table which is used to hold full reference time for the whole text, and we need to update and maintain it dynamically after each normalizing process. The time table consists of two parts: Global Reference Time and Local Reference Time.

- Global Reference Time: Global Reference Time (GRT) is a type of reference time which is referred to by the Global Time. Specifically, it is the report time or the publication time of the document.
- Local Reference Time: Local Reference Time (LRT) is referenced by the Local Time. It will be updated dynamically after each normalizing.

Different classes of time will dynamically and automatically choose references based on their respective classes rather than doing it using the fixed value or the inconsiderate rule under the static mechanism. And the reference time table is updated in real time finishing each normalizing, which makes the temporal situation compliable with dynamically changeable contexts.

Determining the Primary Time

We proposed a score model to calculate the score of each temporal expression. In detail, we consider two aspects when calculating the score of a temporal expression, namely, the term frequency of the temporal expression and the relevance between temporal expressions. It is easy to understand that the term frequency is related to the score of a temporal expression. Here we focus attention on introducing the relevance between temporal expressions. We make an assumption that there is an article which contains some temporal expressions, and most of them refer to a certain day in March. In this case, we tend to choose March as the primary time rather than any one of them. Based on this view, we think that a temporal expression will make a contribution to its parent temporal expression. For example, the expression March 7, 2012 makes a contribution to its parent expression March 2012, and the expression 1983 contribute to its parent expression 1980s.

Here, we define the score of a temporal expression as a combination of an explicit score and an implicit score. The explicit score is related to the term frequency of a temporal expression, and accordingly, the implicit score is related to the contribution made by all its children expressions. The score of T_i, denoted as $ES(T\neg_i)$, is the sum of its explicit score, denoted as $ES(T_i)$, and its implicit score, denoted as $IS(T_i)$.

The explicit score $ES(T_i)$ is defined as the term frequency of *Ti* in the article. As compared to implicit temporal expressions, the explicit temporal expressions are more accurate in the extraction. In other words, the explicit temporal expressions are more credible, so we add a weighting factor d to the implicit temporal expressions. The explicit score of *Ti* is defined as formula (4):

$$ES(T_i) = TF_{ETE}(T_i) + d^{*}TF_{ITE}(T_i) \tag{4}$$

Here, TF_{ete} (T_i) refers to the term frequency of the explicit temporal expressions which are recognized as T_i. TF_{ite}. (T_i) refers to the term frequency of the implicit temporal expressions which are calculated as T_i. d is the weighting factor; if d is set to 1, it means that the explicit and implicit temporal expression have the same credible level; if d is set to 0, it means that we take no account of implicit temporal expressions.

The implicit score $IS(T_i)$ is related to all the scores of its children, we denoted as C_1, C_2, . . ., C_n, respectively, and we use the symbol N to represent the number of children that T_i contains. For example, if the granularity of T_i is MONTH, then the value of N is 30 because a month contains about 30 days. Likewise, if the granularity of T_i is QUARTER, the value of N should be 3 because a quarter contains 3 months. Here, we use the factor α to represent how much contribution the children of T_i make. So the implicit score $IS(T_i)$ can be defined as formula (5):

$$IS(T_i) = \frac{1}{\alpha \times N} \sum\nolimits_{i=1}^{n} S(C_i) \qquad (5)$$

Finally, we can compute the scores of each time expression based on its explicit score and implicit one and then choose the Top-K time expressions as the primary time of the Web page.

Illustrative Example

In order to show the usability of spatiotemporal information in Web search, we present and implement a prototype system for temporal-sensitive queries called TASE (Time-Aware Search Engine) (Lin et al. 2012). The major features of TASE can be described as follows:

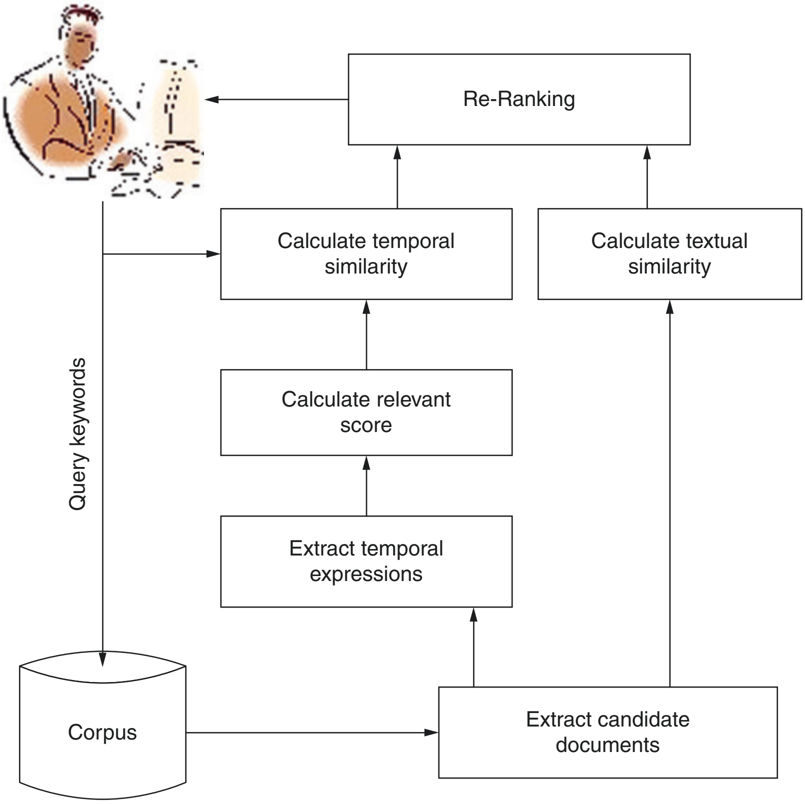
1. TASE extracts the temporal expressions for each Web page and calculates the relevant score between the Web page and each temporal expression. Compared with traditional approaches, TASE uses a new reference time dynamic-choosing approach to extract implicit temporal expressions in Web pages. Besides, it distinguishes the temporal expressions with their relevant score and takes the containment relationship among the temporal expressions into consideration.
2. TASE combines the temporal similarity and the textual similarity to re-rank the search results. Our experiments demonstrate its effectiveness in dealing with temporal-sensitive Web queries.

Figure 5 shows the architecture of TASE, and the interface of TASE is shown in Fig. 6. The four major modules in TASE are described as follows:

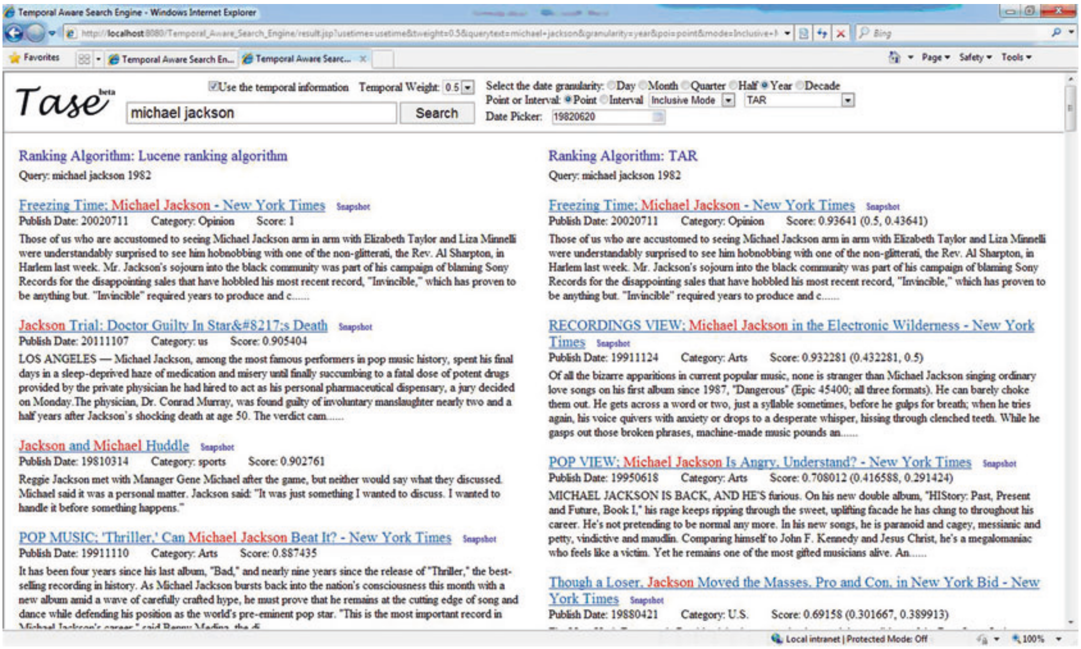
- Extract Candidate Documents. This module extracts the original Top-K documents from the search results which are used as the candidate documents.
- Extract Temporal Expressions. This module extracts all the temporal expressions in each candidate document, including the explicit temporal expressions and the implicit temporal expressions.
- Calculate Relevant Score. The relevant score between a temporal expression and a Web page will be calculated in this module.
- Calculate Temporal Similarity. It calculates the similarity between the temporal expressions in a query and a document.
- Calculate Textual Similarity. TASE is built on Lucene, an open-source search engine. Therefore, we use the textual similarity determined by Lucene as the original textual similarity.
- Re-ranking. In this module, it used the temporal similarity and the original textual similarity to determine the final relevant score of a document.

Key Applications

Spatiotemporal information presented here may be used in many applications. First of all, spatiotemporal information can be used in search engines to improve the quality of results. By designing spatiotemporal indexes and ranking


Re-Ranking
Calculate temporal similarity
Calculate textual similarity
Calculate relevant score
Extract temporal expressions
Query keywords
Corpus
Extract candidate documents

Spatiotemporal Information for the Web, Fig. 5 The architecture of TASE


Temporal Aware Search Engine - Windows Internet Explorer
http://localhost:8080/Temporal_Aware_Search_Engine/result.jsp?usetime=usetime&tweight=0.5&querytext=michael+jackson&granularity=year&poi=point&mode=Inclusive+M
Bing
Favorites
Temporal Aware Search En...
Temporal Aware Searc...
Page
Safety
Tools
Tase beta
Search
Local intranet | Protected Mode: Off
100%

☑Use the temporal information Temporal Weight: 0.5

michael jackson

Select the date granularity: ○Day ○Month ○Quarter ○Half ●Year ○Decade

Point or Interval: ●Point ○Interval Inclusive Mode TAR

Date Picker: 19820620

Ranking Algorithm: Lucene ranking algorithm

Query: michael jackson 1982

Freezing Time: Michael Jackson - New York Times Snapshot

Publish Date: 20020711 Category: Opinion Score: 1

Those of us who are accustomed to seeing Michael Jackson arm in arm with Elizabeth Taylor and Liza Minnelli were understandably surprised to see him hobnobbing with one of the non-glitterati, the Rev. Al Sharpton, in Harlem last week. Mr. Jackson's sojourn into the black community was part of his campaign of blaming Sony Records for the disappointing sales that have hobbled his most recent record, "Invincible," which has proven to be anything but. "Invincible" required years to produce and c......

Jackson Trial: Doctor Guilty In Star&#8217;s Death Snapshot

Publish Date: 20111107 Category: us Score: 0.905404

LOS ANGELES — Michael Jackson, among the most famous performers in pop music history, spent his final days in a sleep-deprived haze of medication and misery until finally succumbing to a fatal dose of potent drugs provided by the private physician he had hired to act as his personal pharmaceutical dispensary, a jury decided on Monday.The physician, Dr. Conrad Murray, was found guilty of involuntary manslaughter nearly two and a half years after Jackson's shocking death at age 50. The verdict cam......

Jackson and Michael Huddle Snapshot

Publish Date: 19810314 Category: sports Score: 0.902761

Reggie Jackson met with Manager Gene Michael after the game, but neither would say what they discussed. Michael said it was a personal matter. Jackson said: "It was just something I wanted to discuss. I wanted to handle it before something happens."

POP MUSIC; 'Thriller,' Can Michael Jackson Beat It? - New York Times Snapshot

Publish Date: 19911110 Category: Arts Score: 0.887435

It has been four years since his last album, "Bad," and nearly nine years since the release of "Thriller," the best-selling recording in history. As Michael Jackson bursts back into the nation's consciousness this month with a new album amid a wave of carefully crafted hype, he must prove that he remains at the cutting edge of song and dance while defending his position as the world's pre-eminent pop star. "This is the most important record in Michael Jackson's career," said Benny Medina, the di

Ranking Algorithm: TAR

Query: michael jackson 1982

Freezing Time: Michael Jackson - New York Times Snapshot

Publish Date: 20020711 Category: Opinion Score: 0.93641 (0.5, 0.43641)

Those of us who are accustomed to seeing Michael Jackson arm in arm with Elizabeth Taylor and Liza Minnelli were understandably surprised to see him hobnobbing with one of the non-glitterati, the Rev. Al Sharpton, in Harlem last week. Mr. Jackson's sojourn into the black community was part of his campaign of blaming Sony Records for the disappointing sales that have hobbled his most recent record, "Invincible," which has proven to be anything but. "Invincible" required years to produce and c......

RECORDINGS VIEW; Michael Jackson in the Electronic Wilderness - New York Times Snapshot

Publish Date: 19911124 Category: Arts Score: 0.932281 (0.432281, 0.5)

Of all the bizarre apparitions in current popular music, none is stranger than Michael Jackson singing ordinary love songs on his first album since 1987, "Dangerous" (Epic 45400; all three formats). He can barely choke them out. He gets across a word or two, just a syllable sometimes, before he gulps for breath; when he tries again, his voice quivers with anxiety or drops to a desperate whisper, hissing through clenched teeth. While he gasps out those broken phrases, machine-made music pounds an......

POP VIEW; Michael Jackson Is Angry, Understand? - New York Times Snapshot

Publish Date: 19950618 Category: Arts Score: 0.708012 (0.416588, 0.291424)

MICHAEL JACKSON IS BACK, AND HE'S furious. On his new double album, "HIStory: Past, Present and Future, Book I," his rage keeps ripping through the sweet, uplifting facade he has clung to throughout his career. He's not pretending to be normal any more. In his new songs, he is paranoid and cagey, messianic and petty, vindictive and maudlin. Comparing himself to John F. Kennedy and Jesus Christ, he's a megalomaniac who feels like a victim. Yet he remains one of the most gifted musicians alive. An......

Though a Loser, Jackson Moved the Masses, Pro and Con, in New York Bid - New York Times Snapshot

Publish Date: 19880421 Category: U.S. Score: 0.69158 (0.301667, 0.389913)

Spatiotemporal Information for the Web, Fig. 6 The interface of TASE algorithms, search engines can be enhanced to process time-and-location-related Web queries effectively and efficiently. Secondly, our method is also useful in focused search engines, such as news search, product search, or stock search. In such applications, time and location information play an important role and our approach can be applied to offer better solutions to the search needs. Thirdly, spatiotemporal information can be utilized in question answering or automatic summarization in the Web. Many questions in the Web are related with time and location, which can be answered if we extract facts as well as their associated time and locations. For the automatic summarization, as events are usually described along a timeline, so it can be well done with the help of the extracted time of the specified topic.

Future Directions

This work can be extended to spatiotemporal analysis and mining in Web data, which may bring values for Web knowledge discovery. As Web has been regarded a major source of competitive intelligence, how to acquire competitive intelligence from the Web has been a hot topic. By using spatiotemporal information, we are able to find some historical information about interested competitors and further detect their future strategic planning in the near future. Spatiotemporal information can also be used to measure the credibility of Web information. With the rapid development of Web 2.0 and social network applications, there are many fakes and false information in the Web, which will introduce a lot of risks in decision making and other applications. Though information credibility involves many aspects of factors, spatiotemporal information can be used as one type of measurement to validate the credibility of specific information. For example, when we want to determine the credibility of a piece of news reporting “Apple iPhone 5 has been released,” we can collect the Web pages or microblogs mentioning the news and perform spatiotemporal clustering process to detect its credibility.

Cross-References

▶ Modeling and Analysis of Spatiotemporal Social Networks
▶ Social Web Search
▶ Spatiotemporal Personalized Recommendation of Social Media Content

Acknowledgments We would like to thank the University of Science and Technology of China for providing the environment where the study described in this entry was completed. The work involved in this article is partially supported by the National Science Foundation of Anhui Province (NO. 1208085MG117) and the USTC Youth Innovation Foundation.

References

Brin S, Page L (1998) The anatomy of a large-scale hyper textual web search engine. In: Proceedings of WWW, Brisbane, pp 107–117

Chinchor N (1998) MUC-7 information extraction task definition, version 5.1. In: Proceedings of the 7th message understanding conference (MUC-7), Fairfax

Ding J, Gravano L, Shivakumar N (2000) Computing geographical scopes of web resources. In: Proceedings of VLDB, Cairo, pp 545–556

GUTime (2012.) http://www.timeml.org/site/tarsqi/modules/gutime/index.html. Accessed Aug 2012

Lee R et al (2003) Optimization of geographic area to a web page for two-dimensional range query processing. In: Proceedings of fourth international conference on web information systems engineering workshops (WISEW 2003), Roma. IEEE Computer Society, pp 9–17

Lin S, Jin P, Zhao X, Yue L (2012) TASE: a time-aware search engine. In: Proceedings Of CIKM‘12, Maui. ACM

Ma Q, Tanaka K (2004) Retrieving regional information from web by contents localness and user location. In: Proceedings of AIRS, Beijing, pp 301–312

Markowetz A, Chen Y, Suel T, Long, X, Seeger B (2005) Design and implementation of a geographic search engine. Technical report TR-CIS-2005-03, Polytechnic University, Brooklyn

Nunes S, Ribeiro C, David G (2008) Use of temporal expressions in web search. In: Proceedings of ECIR&apos; 08, Glasgow, pp 580–584

Sanderson M (2000) Retrieving with good sense. Inf Retr 2 (1):45–65

Sanderson M, Kohler J (2004) Analyzing geographic queries. In: Proceedings of GIR&apos; 04, Sheffield. ACM

Setzer A, Gaizauskas R (2002) On the importance of annotating event-event temporal relations in text. In: Proceedings of LREC&apos; 02, Paris

Sundheim B, Chinchor N (1995) Named entity task definition, version 2.0. In: Proceedings of the 6th message understanding conference (MUC-6), Columbia. Morgan Kaufman, pp 319–332

Wang C, Xie X et al (2005) Web resource geographic location classification and detection. In: Proceedings of WWW&apos; 05, Chiba. ACM

Zhou Y, Xie X, Wang C et al (2005) Hybrid index structures of location-based web search. In: Proceedings of CIKM&apos; 05, Bremen

Spatiotemporal Outlier

► Detection of Spatiotemporal Outlier Events in Social Networks

Spatiotemporal Personalized Recommendation of Social Media Content

Bee-Chung Chen
LinkedIn, Sunnyvale, CA, USA

Synonyms

Location-based recommendation; Positional or layout effect in recommender systems; Spatiotemporal collaborative filtering; Time-sensitive recommendation

Glossary

Context	The situation (which includes time, geographical location, location of a web page, etc.) in which recommendations are made to a user
Feature	Information (about a user, an item, and the context in which the item may be recommended to the user) that can be used to predict the response rate
Graph	A set of nodes connected by a set of edges
Page	A web page on which recommended items are placed
Recommender	A system that recommends items (e.g., news articles, blog posts) to users
Response rate	The probability that a user would respond positively to (e.g., click, share) a recommended item

Definition

Social media sites (like twitter.com, digg.com, blogger.com) complement traditional media by incorporating content generated by regular people and allowing users to interact with content through sharing, commenting, voting, liking, and other actions. Since the number of content items is usually too large for a person to manually examine to find interesting ones, it is important for social media sites to recommend a small set of items that are worth looking at for each user. To satisfy each individual user, recommended items have to match the user's personal interests and be relevant to the user's current spatiotemporal context. For example, a content item about the user's hometown is usually a better choice than an item about an unknown foreign country, and a content item on a fresh trending topic is usually more interesting than an item on a stale topic.

Spatiotemporal personalized recommendation of social media content refers to techniques used to make personalized recommendation based on:

- The geographical location of a user and an item (the location of an item can be the location that the item is about or the location of the author of the item)
- The location of a user in the social space (e.g., the neighborhood of a user in a friendship graph)
- The position of an item placed on a page and the layout of the page
- Temporal evolution of user interests
- Temporal behavior of the popularity of an item
- Identification of trending and/or geo-related topics

Introduction

Social media usually refers to a group of Internet-based applications that allow creation and exchange of user-generated content (Kaplan and Haenlein 2010). For example, weblog sites like blogger.com provide regular people the ability of publishing any article (called blog) on the web, microblogging sites like twitter.com facilitate fast distribution of short messages of any topic posted by any one, and social news sites like digg.com allow their users to vote news articles (and other web content) up or down in order to present popular and interesting news stories based on the wisdom of the crowd (i.e., votes from users), just to name a few. Because of the success of such social media sites, almost all online media sites now provide their users with the functionality of sharing and commenting on content items (e.g., news articles, photos, songs, movies, etc.), no matter whether the content items are generated by regular users. Since sharing and commenting are usually considered as social activities, the distinction between social media and traditional online media blurs. In this article, we discuss recommendation methods suitable for any online media with a special emphasis on spatial, temporal, and social characteristics of users and content items.

The large amount of content generated by social media makes it difficult for users to find personally relevant content. To alleviate such information overload, many social media sites recommend a small set of content items to each user based on what they know about the user and the items. We use the term "item" to refer to any candidate objects to be recommended to users, which include (but are not limited to):

- Publisher-generated items like articles, songs, and movies, which are not generated by regular users, but are voted, shared, liked, or commented on by them
- User-generated items like blogs, tweets (short messages posted on twitter.com), photos, videos, status updates, and comments on other items

Good recommendations help social media sites keep their users engaged and interested.

Key Points

When recommending items to users, it is important to consider whether an item is relevant to a user in the *spatiotemporal context* in which recommendations are to be made. We take a broad view of the spatial aspect that includes locations in geographical space, social space, and positions on a web page. A few key reasons for considering spatiotemporal contexts are listed below.

- Users are likely to be more interested in items about the geographical location of their interests (e.g., their current locations, neighborhoods that they are familiar with, and places that they frequently visit) than items about a random location, which is especially true for mobile applications (see Zheng et al. 2010 for an example).
- In some applications, users tend to have similar preferences to those who are close to them in the social space, which is especially true when closeness is defined based on a trust network (see Jamali and Ester 2010 for an example).
- It is generally true that an item placed at a prominent location (e.g., top) on a page generates more responses from users than the same item placed at a non-prominent location (see Agarwal et al. 2009 for an example).
- Users change their interests in topics over time (see Ahmed et al. 2011 for an example) and their geo-location.
- Popularity of items also changes over time (see Agarwal et al. 2009 for an example).

Many methods have been developed to exploit these spatiotemporal characteristics to improve the performance of recommenders. A comprehensive review of these methods is beyond the scope of this article. Instead, after providing a brief historical background, we illustrate key ideas in spatiotemporal personalized recommendation through a generic supervised learning approach, which handles spatiotemporal characteristics by (1) defining features that capture those characteristics and (2) learning a function that predicts whether a user would respond to an item positively based on these features from a dataset that records users' past responses to items. This approach generally applies to recommendation of any kind of item.

Historical Background

There have been many approaches developed to make personalized recommendations. When an item to be recommended is a text article, which may be represented as a bag of words, an early approach is to also represent a user as a bag of words. The user's bag of words can be constructed by including representative words in the articles that the user likes to read. Then, we can recommend a user the articles which bags of words are most similar to the user's bag of words through Salton's vector space model (Salton et al. 1975). For items that are not easily representable as bags of words, how other users respond to an item may provide a clue as to whether to recommend the item to a user who has not yet responded to the item. Agrawal et al. (1993) proposed that, in a retail store setting, products can be recommended based on customers' co-buying behavior. For example, if the majority of customers who buy product A also buy product B, then we may recommend product B to a customer who only bought product A. This idea was then extended by incorporating a notion of similarity of users or items. For example, when we decide whether to recommend item B to user i, we look at whether users "similar" to user i respond to item B positively. Notice that Agrawal's method is based on the similarity definition that if two customers buy the same product, then they are similar. A different definition of similarity between users leads to a different method. Furthermore, we can also exploit similarity between items in a similar way – when deciding whether to recommend item B to user i, check whether user i liked items that are "similar" to B in the past. Here, similarity between two items can be defined by looking at whether most users responded to the two similarly. Adomavicius and Tuzhilin (2005) provided a good review of such methods. This kind of methods is generally referred to as collaborative filtering because the recommendations that a user receives depend on other users' responses to candidate items – this process can be thought of as a collaboration among users to help one another find interesting items (although users may not be aware of the collaboration).

Conceptually, one can put users' past responses to items into a matrix. Since this matrix-oriented approach is popular in movie recommendation (Koren et al. 2009), we use it as an example in the following discussion. In a movie recommender system, users rate movies. Let y_{ij} denote the rating that user i gives to movie j. For example, y_{ij} may be a numeric value ranging from one to five, representing one star to five stars. Let Y denote the $m \times n$ matrix such that the value in the (i, j) entry is y_{ij}, where m is the number of users and n is the number of movies in the system. Notice that there are many entries with missing (i.e., unknown) values in matrix Y because most users only rate a small number of movies. For user i, if we can predict the missing values in the ith row of matrix Y accurately (where the entries with missing values correspond to movies that have not yet rated by user i and are thus candidate items to be recommended to him/her), then we can recommend user i the movies having the highest predicted rating values. One popular way of making such predictions is through matrix factorization – approximate matrix Y as the product UV' of two low rank matrices U of size $m \times r$ and V of size $n \times r$, where V' denotes the transpose of matrix V and the rank r of matrices U and V is much smaller than the numbers m and n of users and items, respectively. Let $\boldsymbol{u}_i$ denote the ith row of matrix U, $\boldsymbol{v}_j$ denote the jth row of matrix V, and $\Omega = \{(i, j)\text{: user } i \text{ rated movie } j\}$ denote the set of observed entries in matrix Y. This approximation then can be mathematically formulated as the following optimization problem.

Find U and V that minimize

$$\sum_{(i,j)\in\Omega} \left(y_{\mathrm{ij}} - \boldsymbol{u}_i'\boldsymbol{v}_j\right)^2 \tag{1}$$

where $\boldsymbol{u}_i'\boldsymbol{v}_j$ is the inner product of two vectors $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$. Notice that $\boldsymbol{u}_i'\boldsymbol{v}_j$ is the (i, j) entry of matrix (UV') and is also the predicted value of y_{ij}. Thus, the above optimization seeks to minimize the difference between matrix Y and matrix (UV')

over only the set Ω of observed entries of Y. Sum of squared differences is a common choice, while other choices are also available for different problem settings. Recent studies, such as Agarwal and Chen (2009) and Koren et al. (2009) and many others, suggest that matrix factorization usually provides superior recommendations than more traditional methods.

A survey of a wide range of approaches to recommender systems can be found in Jannach et al. (2010) and Ricci et al. (2015). Here, we focus on how to make use of spatial, temporal, and social information to make good recommendations of social media content. In particular, we illustrate key ideas in spatiotemporal personalized recommendation through a general supervised learning (or statistical modeling) approach, which generally applies to recommendation of any kind of item.

Supervised Learning Approach

In general, a recommendation problem can be formulated as follows. A recommender is given:

- A user, who is associated with a vector of *user features*, e.g., age, gender, location (home, workplace, and other frequently visited places)
- A set of candidate items, each of which is associated with a vector of *item features*, e.g., topics, keywords, the time when the item was produced, and the location that the item is about or was produced at
- A context, which is associated with a vector of *context features*, e.g., time of day, day of week when the recommendation is to be made, the age of the item (time since the item was produced), and the distance between the user's current location and the location that the item is about

The goal of the recommender is to rank and pick the top few items from the set of candidate items that best "match" the user's interests and information needed in the context. The supervised learning approach exploits the fact that, in many recommenders, a dataset of users' past responses (e.g., click, share) to items can be collected and defines the degree that an item matches a user as the response rate of the user to the item (e.g., the probability that the user would click the item if he/she sees the item on a web page). Such predictions can be made by using a statistical (regression or machine learning) model, which "learns" the user and item behavior that allows accurate predictions from the dataset, where users' past responses in the dataset "supervise" the learning process via giving desired (e.g., click) and undesired (e.g., no click) examples. When such a model is available, recommendations for a user can be made by picking the top few items having the highest response rates among the set of candidate items. This supervised learning approach applies to recommendation of any kind of item, where spatiotemporal and other characteristics can be incorporated by defining features that capture those characteristics.

To use this supervised learning approach, a developer of a recommender needs to make the following three decisions:

- What response should the model try to predict?
- What features should the model use to capture the characteristics of users, items, and the spatiotemporal context?
- What class of model do we want to use?

After introducing a running example, we discuss how to choose the response, provide a number of useful features, and then introduce two commonly used classes of models, namely, feature-based regression model and latent factor model. See Jannach et al. (2010) and Koren et al. (2009) for other classes of models. See Hastie et al. (2009) for a general introduction to supervised learning.

Example Recommender

For concreteness, we use blog article recommendation as a running example. Consider that we want to develop a recommender for a blog service provider (e.g., blogger.com) that seeks to recommend each user with a set of interesting blog articles posted by other users. To make modeling more interesting, assume that a user can declare friendship with other users, and such friendship connections between users are available to the recommender. In this example, the set of candidate items for each user consists of all of the articles posted within a 1-week time window (to ensure freshness) by any user of this service provider. Notice that the set of candidate items changes over time. For simplicity, we only need to recommend ten articles for each user, once per day, and the recommended articles are displayed in a list on the sidebar of each user's homepage (they are only visible to the owner of the homepage, not the visitors of the homepage, since the recommendations are made to the homepage owner).

Choice of Response

The choice of response depends on the objective that a recommender is developed for and availability of user feedback that the recommender receives. A common objective is to maximize clicks on recommended items because the fact that a user clicks an item indicates that the user is interested in knowing more about the item. Note that clicks are user feedback that can easily be made available to a recommender through logging whether a user clicks a recommended item. In this case, a natural choice of the response is whether a user would click an item if he/she sees the item being recommended. Here, the goal of learning is to predict the probability that a user would click an item based on a dataset that records what items each user clicked and what items each user did not click in the past.

Beyond clicks, a recommender may be developed for other objectives. For example, if the objective of recommendation is to encourage users to make comments on recommended items, then a natural choice of the response would be whether a user would comment on a recommended item or not. On some sites, users can explicitly rate items (e.g., using one star to five stars), and then a natural choice of the response would be the rating that a user would give to an item. For simplicity, we only consider methods that seek to achieve a single objective and model the response rate of a single type of choice (e.g., modeling either click rate or explicit star rating, but not both). See Agarwal et al. (2011a) for an example of multi-objective recommendation, and see Agarwal et al. (2011b) for an example of joint modeling of multiple types of responses.

Let $\boldsymbol{y}_{ijk}$ denote the response that user i gives to item j in context k. For concreteness, assume that we choose to model whether the user would click the item.

Feature Engineering

Having good features is essential to an accurate model, but one usually does not get good features automatically. It requires domain knowledge, good intuition, and experience in the application to define good features. Here, for illustration purposes, we only show a number of example features that can potentially capture different kinds of spatiotemporal characteristics for our example recommender. Real-life recommenders usually need to use much more features than the following ones.

User Features

Let $\boldsymbol{w}_i$ denote the vector of features of user i. For simplicity, we mostly consider binary features, meaning each element in the vector is either 0 or 1. Example features are as follows:

- **Gender**: From the user's registration record when he/she signed up on the site, the recommender obtains the gender of each user. The numeric value of the feature is 1 if the user is a male and 0 if the user if a female.
- **Age**: Also from the user's registration record, the recommender obtains the age of each user. For example, we can group age values into ten age groups, which give ten age features. If the user's age is in an age group, the value of the feature corresponding to that age group is 1, and the rest of age groups get feature value of 0.
- **City**: From the IP address of a user, the recommender can guess the city that the user is in. Here, we use a set of features, one for each city, to represent the user's geographic location. For example, assume the user lives in New York City. Then, the value of the New York City feature is 1 and the values of

the rest city features are all 0 for the user. It is common to only include cities that have at least n users, where n is a threshold that a developer of the recommender can choose to reduce the number of features.

- **Other location features**: Similar to the above city feature, we can also generate location features at different granularities (e.g., region, state, country) for different types of user locations (e.g., home, workplace, other frequently visited places).

Item Features

Let $\boldsymbol{x}_j$ denote the vector of features of item j. Example features are as follows:

- **Bag of words**: It is common to represent the text content of an article as a bag of words, which corresponds to a set of features, one for each keyword. For simplicity, we only consider binary keyword features. The value of a keyword feature is 1 if the article contains the keyword and 0 if the article does not contain the keyword. Since the total number of words in all articles is usually too large, it is also common to reduce the space of all keywords to a relatively small number of important words, e.g., location names or other named entities.
- **Topics**: Another way to reduce the space of words in articles is to group words into topics and then assign topics to articles based on the words in articles. This process can be automated through topic models like latent Dirichlet allocation (Blei et al. 2003). One output from such a model is a vector of topic membership for each article, where each element in the vector represents the probability that the article is about a particular topic.
- **Tags**: In social media, users tag items based on their interests in the items. These tags can be used to generate features in the same way as bag of words. Geo tags and event tags usually capture the spatiotemporal context of the item.
- **Creation time**: This is the time at which the item was produced.
- **Location of the author**: This is the location of the author of the item when producing the item.

Context Features

Let z_{ijk} denote the vector of features of the context in which user i is (to be) recommended with item j in context k (which include time and location). Example features are as follows:

- **Day of the week**: This is the day of the week (weekday vs. weekend) when the recommendation is to be made. User behavior during the weekday can be quite different from that during the weekend. The value of this feature is 1 for weekday and 0 for weekend.
- **Article age**: This is the age of an article (not to be confused with the age of a user), which is the number of days since the article was posted. We put it into the category of context features, instead of item features, because it depends on both the article and time, instead of the article alone. For example, assume the article was posted 2 days ago; then, the value of the feature corresponding to 2 days ago is 1, and the other days get feature value 0. To model finer-grained temporal effect, one may choose a finer time resolution (e.g., hour, instead of day).
- **Position on page**: It is well known that the click rate of an item put on the top of a list on a page is usually higher than that of the same item put in the middle or the bottom of the page. To capture this positional bias, we define a set of features, each of which corresponds to a position in the list. For example, assume the article is put at the third position, the value of the feature corresponding to the third position is 1, and all other positions have feature value 0.
- **Friendship**: This feature is 1 if user i is connected to the author of item j through a friendship connection and is 0 otherwise.
- **Connection strength**: This feature captures the strength of the connection between the user and the author of the item. For example, we can qualify the strength by the number of common friends between the user and the author (and create different strength ranges to generate binary features if desired).
- **Same city**: This feature is 1 if user i is in the same city as the author of item j and is 0 otherwise. Features like the same state and same country can be created in a similar way.

- **Geo-distance**: This is the distance between the user's current location and the location that the item is about (if desired, we can use distance ranges to create binary features).
- **Repeated exposure**: This is the number of times that the user saw this time in the past. This feature can be used to capture the user's fatigue after seeing the same item many times.

Note that the above features are only simple examples. The goal here is to provide concrete examples of features for illustration purposes. Different applications may require different sets of features. Recent work on spatiotemporal topic models (Hu et al. 2013; Yin and Cui 2016; Yuan et al. 2015) can also be used to create features that capture items' spatiotemporal topicality and users' spatiotemporal interests.

Feature-Based Regression Model

After defining the response and features, we have a standard supervised learning problem. When the response is binary (e.g., either click or no click), we can use logistic regression. See Hastie et al. (2009) for an introduction to logistic regression. Let p_{ijk} denote the probability that user i would respond to item j when he/she sees it in context k. There are many ways in which one can define a function that predicts p_{ijk} based on features. A useful prediction function is as follows:

$$p_{\mathrm{ijk}} = \sigma\left(\boldsymbol{w}_i'\boldsymbol{A}\boldsymbol{x}_j + \boldsymbol{\beta}'\boldsymbol{z}_{\mathrm{ijk}}\right), \qquad (2)$$

where $\sigma(a) = \frac{1}{1+\exp(-a)}$ is the sigmoid function that transforms an unbounded value a into a number between 0 and 1 (since p_{ijk} is a probability), $\boldsymbol{A}$ is a regression coefficient matrix, $\boldsymbol{\beta}$ is a regression coefficient vector, and w_i' and $\boldsymbol{\beta}'$ are the row vectors after transposing the two column vectors $\boldsymbol{w}_i$ and $\boldsymbol{\beta}$, respectively. Given a dataset of users' past responses to items, where each record is in the form (y_{ijk}, $\boldsymbol{w}_i$, $\boldsymbol{x}_j$, $\boldsymbol{z}_{ijk}$), off-the-shelf logistic regression packages (e.g., Photon ML at https://github.com/linkedin/photon-ml) can be applied to learn the regression coefficients $\boldsymbol{A}$ and $\boldsymbol{\beta}$.

To better understand this model, we take a closer look at the prediction function. Let A_{mn} denote the (m, n) entry of matrix A, w_{im} denote the mth user feature in vector $\boldsymbol{w}_i$, and x_{jn} denote the nth item feature in vector $\boldsymbol{x}_j$. By definition we have

$$\boldsymbol{w}_i'\boldsymbol{A}\boldsymbol{x}_j = \sum_m \sum_n A_{\mathrm{mn}} w_{\mathrm{im}} x_{\mathrm{jn}}. \qquad (3)$$

For example, assume w_{im} is the feature that indicates whether user i lives in New York City and x_{jn} is the feature that indicates whether article j contains keyword "new york." Then, the regression coefficient A_{mn} would try to capture the propensity that users living in the New York City would click an article that contains keyword "newyork" after adjusting for all other factors. Now, assume that the mth and nth context features in $\boldsymbol{z}_{ijk}$ indicate whether article j is posted 1 day ago, and whether j is posted 5 days ago, respectively. Then, the difference between regression coefficients $\beta_m - \beta_n$ would quantify how much the popularity of an article drops from day 1 to day 5 when all other conditions being equal.

Latent Factor Model

Although feature-based regression models are useful for predicting users' response rates to items, they depend highly on the availability of predictive features, which usually requires a significant feature engineering effort with no guarantee of obtaining predictive features. Also, feature vectors may not be sufficient to capture the differences between users or items. For example, when two users have identical feature vectors, feature-based regression models would be unable to tell the differences between the two. One way of addressing these issues is to add *latent factors* into the prediction function, i.e.,

$$p_{\mathrm{ijk}} = \sigma\left(\boldsymbol{w}_i'\boldsymbol{A}\boldsymbol{x}_j + \boldsymbol{\beta}'\boldsymbol{z}_{\mathrm{ijk}} + \boldsymbol{u}_i'\mathrm{v}_j\right) \qquad (4)$$

where $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ are two r-dimensional vectors both to be learned from data like regression coefficients $\boldsymbol{A}$ and $\boldsymbol{\beta}$, where r is much smaller than the number of users and the number of items. Recall that we have seen $\boldsymbol{u}_i'\boldsymbol{v}_j$ in the matrix factorization method in the historical background section. The difference is that, instead of factorizing the response matrix, here we factorize the residual (i.e., prediction error) matrix of feature-based regression in order to capture the behavior of users and items that the features fail to capture.

Intuitively, one can think of $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ as "latent feature" vectors of user i and item j, respectively. We do not determine the values of these r latent features per user or item before learning the model. Instead, $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ are treated as variables that can be used to reduce the error of predicting the responses in the dataset used for learning. The inner product $\boldsymbol{u}_i'\boldsymbol{v}_j$ then represents the affinity between user i and item j; the larger the inner product value, the higher the probability that user i would click item j. After the learning process, we simultaneously obtain the values of these latent features and also the regression coefficients $\boldsymbol{A}$ and $\boldsymbol{\beta}$. See Agarwal et al. (2010) for an example of such a latent factor model.

Spatiotemporal contexts can also be involved in a latent factor model. For example, assume we want to model a temporal effect through latent factors. Let context index k represent the kth time period (e.g., day). One way of capturing user or item behavioral changes over time is through the following model:

$$p_{\mathrm{ijk}} = \sigma\left(\boldsymbol{w}_i' \mathrm{Ax}_j + \boldsymbol{\beta}' z_{\mathrm{ijk}} + \left\langle \boldsymbol{u}_i, \boldsymbol{v}_j, \boldsymbol{t}_k \right\rangle\right), \qquad (5)$$

where $\langle \boldsymbol{u}_i, \boldsymbol{v}_j, \boldsymbol{t}_k \rangle = \sum_l u_{\mathrm{il}} v_{\mathrm{jl}} t_{\mathrm{kl}}$ is a form of tensor product of three vectors $\boldsymbol{u}_i$, $\boldsymbol{v}_{j,}$ and $\boldsymbol{t}_k$. Note that u_{il} denotes the lth element of vector $\boldsymbol{u}_i$ and so on. Similar to the previous model, $\boldsymbol{u}_i$, $\boldsymbol{v}_j$, and $\boldsymbol{t}_k$ are all latent feature vectors, which values are to be learned from data. Unlike the previous model where the affinity $\boldsymbol{u}_i'\boldsymbol{v}_j$ between user i and item j is fixed over time, now the affinity $\langle \boldsymbol{u}_i, \boldsymbol{v}_j, \boldsymbol{t}_k \rangle$ is a function of time period k, which means this model captures the changing behavior of user-item affinity. Specifically, in this model, the user and item latent feature vectors are fixed over time, but the affinity between the two is a weighted sum of the element-wise product of the two latent feature vectors $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$, where the weight vector $\boldsymbol{t}_k$ changes over time. See Xiong et al. (2010) for an example of such a temporal latent factor model.

We can also use $\boldsymbol{t}_k$ to represent the latent vector for location k if we use k to index locations. We can also use k to index clusters of location-time pairs. Then, $\boldsymbol{t}_k$ represents the latent vector for a spatiotemporal cluster.

Key Applications

Personalized recommendation is an important mechanism for surfacing social media content. The spatiotemporal context in which a recommendation is made provides a key piece of information that helps a recommender to recommend the right item to the right user at the right time. While many methods have been proposed in the literature, the supervised approach is attractive because of its generality, where spatiotemporal characteristics can be incorporated as features or latent factors. In this article, we introduced a number of example features and two example models. They can be applied to personalized recommendations of new articles, blog articles, tweets, shared items (e.g., articles, videos, photos), status updates, and comments on different kinds of items. In practice, many features need to be evaluated, and a number of different models need to be tried, so that a good recommender can be built.

Future Directions

Personalized content recommendation is currently an active research area in data mining, information retrieval, and machine learning. A lot of progress has been made in this area, but challenges remain.

- *Improving response rate prediction accuracy:* Although many models have been proposed to predict response rates and we have seen prediction accuracy improves over time, accurate prediction of the probability that a user would respond to an item is still a challenging problem, especially for users and items that the recommender knows little about. What are the spatial, temporal, social, and other kinds of features that can further improve accuracy? How can a recommender actively collect data to achieve better model learning and evaluation?

- *Multi-objective optimization:* A recommender usually is designed to achieve multiple objectives. For example, many web sites put advertisements on article pages to generate revenue. In addition to recommend articles that users like to click, we may also want to recommend articles that can generate high advertising revenue. How can a recommender optimize multiple objectives in a principled way?
- *Multi-type response modeling:* In social media, users respond to items in multiple ways, e.g., clicks, shares, tweets, emails, likes, etc. How can we jointly model such different types of user responses in order to find out the items that a user truly want to be recommended?
- *Whole page optimization:* On a web page, there can be multiple recommender modules. For example, one recommends news articles, another recommends updates from a user's friends, and yet another recommends online discussions the user may be interested in. How can we jointly optimize multiple recommender modules on a page to leverage the correlation among modules and to ensure consistency, diversity, and serendipity?
- *Collaborative content creation:* Wikipedia demonstrated high-quality content creation through massive collaboration. However, in most recommender systems, items to be recommended are created by a single party (e.g., a publisher or a user). How can we synthesize items at the right level of granularity to recommend to users in a semiautomatic collaborative way?

Cross-References

▶ Data Mining
▶ Friends Recommendations in Dynamic Social Networks
▶ Link Prediction: A Primer
▶ Matrix Decomposition
▶ Mining Trends in the Blogosphere
▶ Probabilistic Graphical Models
▶ Recommender Systems: Models and Techniques
▶ Recommender Systems Using Social Network Analysis: Challenges and Future Trends
▶ Recommender Systems, Semantic-Based
▶ Regression Analysis
▶ Social Media Mining
▶ Social Recommender Systems

References

Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng 17:734–749

Agarwal D, Chen BC (2009) Regression-based latent factor models. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, KDD '09, pp 19–28, URL https://doi.org/10.1145/1557019.1557029

Agarwal D, Chen BC, Elango P (2009) Spatio-temporal models for estimating click-through rate. In: Proceedings of the 18th international conference on World wide web, ACM, New York, NY, USA, WWW '09, pp 21–30, URL https://doi.org/10.1145/1526709.1526713

Agarwal D, Chen BC, Elango P (2010) Fast online learning through offline initialization for time-sensitive recommendation. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, KDD '10, pp 703–712, URL http://doi.acm.org/10.1145/1835804.1835894

Agarwal D, Chen BC, Elango P, Wang X (2011a) Click shaping to optimize multiple objectives. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, KDD '11, pp 132–140, URL http://doi.acm.org/10.1145/2020408.2020435

Agarwal D, Chen BC, Long B (2011b) Localized factor models for multi-context recommendation. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, KDD '11, pp 609–617, URL http://doi.acm.org/10.1145/2020408.2020504

Agrawal R, Imielin'ski T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD international conference on Management of data, ACM, New York, NY, USA, SIGMOD '93, pp 207–216, URL http://doi.acm.org/10.1145/170035.170072

Ahmed A, Low Y, Aly M, Josifovski V, Smola AJ (2011) Scalable distributed inference of dynamic user interests for behavioral targeting. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, KDD '11, pp 114–122, URL http://doi.acm.org/10.1145/2020408.2020433

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022. URL http://dl.acm.org/citation.cfm?id=944919.944937

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, New York

Hu B, Jamali M, Ester M (2013) Spatio-temporal topic modeling in mobile social media for location recommendation. In: 2013 I.E. 13th International Conference on Data Mining, IEEE, pp 1073–1078

Jamali M, Ester M (2010) A matrix factorization technique with trust propagation for recommendation in social networks. In: Proceedings of the fourth ACM conference on Recommender systems, ACM, New York, NY, USA, RecSys '10, pp 135–142, URL http://doi.acm.org/10.1145/1864708.1864736

Jannach D, Zanker M, Felfernig A, Friedrich G (2010) Recommender systems: an introduction. Cambridge University Press, New York. URL http://books.google.com/books?id=eygTJBd U2cC

Kaplan AM, Haenlein M (2010) Users of the world, unite! the challenges and opportunities of social media. Bus Horiz 53(1):59–68. URL http://www.sciencedirect.com/science/article/pii/S0007681309001232

Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. Computer 42(8):30–37

Ricci F, Rokach L, Shapira B, Kantor PB (eds) (2015) Recommender systems handbook. Springer, New York. http://www.springer.com/us/book/9781489976369

Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. Commun ACM 18(11):613–620. URL http://doi.acm.org/10.1145/361219.361220

Xiong L, Chen X, Huang TK, Schneider JG, Carbonell JG (2010) Temporal collaborative filtering with bayesian probabilistic tensor factorization. In: Proceedings of the SIAM International Conference on Data Mining, SDM 2010, April 29–May 1, 2010, Columbus, Ohio, USA, pp 211–222

Yin H, Cui B (2016) Spatio-temporal recommendation in Social Media. Springer, Singapore. http://www.springer.com/us/book/9789811007477

Yuan Q, Cong G, Zhao K, Ma Z, Sun A (2015) Who, where, when, and what: a nonparametric bayesian approach to context-aware recommendation and search for twitter users. ACM Trans Inf Syst (TOIS) 33(1):2

Zheng VW, Cao B, Zheng Y, Xie X, Yang Q (2010) Collaborative filtering meets mobile recommendation: a user-centered approach. In: Proceedings of the 24th AAAI Conference on Artificial Intelligence, pp 236–241

Spatiotemporal Proximity and Social Distance

Christoph Schlieder
University of Bamberg, Faculty for Information Systems and Applied Computer Sciences, Chair of Computing in the Cultural Sciences, Bamberg, Germany

Synonyms

Information filtering; Report confirmation

Glossary

LBSN	Location-based social network
Heuristic principle	An experience-based but fallible problem-solving approach
Information filtering	An algorithm that aims at identifying relevant pieces of information
User-generated content	Text, images, or other media published in a LBSN

Definition

Spatiotemporal proximity and social distance are two heuristic principles for filtering user-generated content produced by the members of a location-based social network (Schlieder and Yanenko 2010; Yap et al. 2012). Information filtering addresses the quality problem which arises when content is created by a large community of voluntary contributors as is the case in Web-based forms of participatory or citizen journalism. While the computational filtering approaches share some basic assumptions with the evaluation approach adopted in classical journalism, there are significant differences with respect to the scale of the problem and the methods for establishing confirmation relationships between reports.

Introduction

The idea of citizen reporters who complement the news coverage provided by professional journalists predates the Web by several decades (Deuze et al. 2007). With the diffusion of smartphones and the mobile access to the Web, however, it became much easier for eyewitnesses of events to report their observations to digital communities. Observation reports are published in location-based social networks (LBSNs), that is, any type of Web-based social medium which provides geolocation metadata about its members and the user-generated content (Symeonidis et al. 2014). Such LBSN also makes available the temporal metadata used by conventional social media services. In other words, each observation reported in the LBSN comes with a time stamp and a place stamp. A prominent example of LBSN technology supporting citizen reporting is the Ushahidi platform which was originally created to collect and visualize reports about incidents of politically motivated violence (Okolloh 2009). Other scenarios include emergency response to natural disasters, the documentation of urban sprawl (Bishr and Mantelas 2008), and reports on wild animal sightings (Schlieder and Yanenko 2010).

Often it is useful to combine automatic filtering as a preprocessing step with manual post-processing by human experts who examine the remaining set of critical cases. Note, however, that in classical journalism, a small number of authors contribute articles each stating a large number of facts, whereas in citizen journalism, a large number of contributors publish reports that mostly state a single elementary fact such as a Tweet about the sighting of a wood fire. Only the automatic approaches scale easily with the number of reports.

Spatiotemporal Proximity and Social Distance

Different heuristic principles are used by the automatic filtering approaches. Bishr and Mantelas (2008) argue for using the spatial proximity of the observer to the object described in the report as a measure of the observer's reputation in contributor status filtering. The rationale behind this heuristic *principle of spatial proximity* is that an eyewitness should have higher reputation than someone who reports from hearsay and that spatial proximity constitutes a necessary – though not sufficient – condition for observing the object. In scenarios such as reporting about natural disasters, however, eyewitnesses are often first-time or infrequent contributors which cannot be handled by reputation models. Report confirmation approaches have been proposed as an alternative by Schlieder and Yanenko (2010) and Yanenko and Schlieder (2012) to handle such scenarios.

Confirmation focuses on events instead of objects, that is, entities extended in time. As a consequence, the principle of spatial proximity needs to be complemented by a related *principle of temporal proximity*. A report of the sighting of a rare bird species, for instance, could be confirmed by a second report one hour later stating the sighting of the same species at a place nearby. Generally, a smaller distance corresponds to better confirmation. Both the spatial and the temporal proximity interact in confirmation and are therefore referred to as a single heuristic *principle of spatiotemporal proximity*.

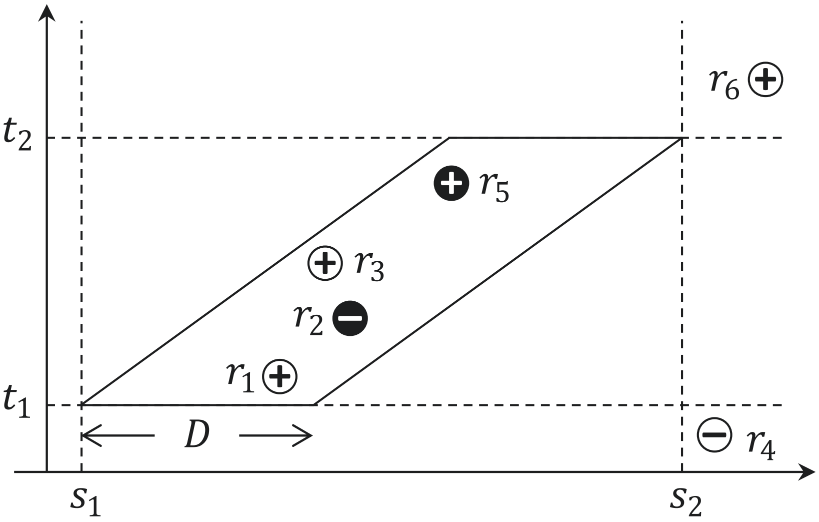
In many application scenarios, observations are informed by the social role of the observer, by his or her affiliation to a subcommunity of the LBSN. An example is the competitive situation in a location-based game. The categorization of a game event as foul play is likely to be affected by which team the observer belongs to or supports (Yanenko and Schlieder 2012). In reporting about political events, such biases become even more important. The confirmation *principle of social distance* addresses such cases. It states that a report from an observer from a subcommunity of the LBSN which takes a different stance on the issue provides better confirmation than a report from an observer from the same subcommunity. According to this principle, a foul play reported by at least one member of both teams is considered having higher confirmation than a foul play reported only by members of the same team. This principle reflects the confirmation approach taken by classical journalism which requires at least two independent sources for each fact reported.

Space-Time Diagram

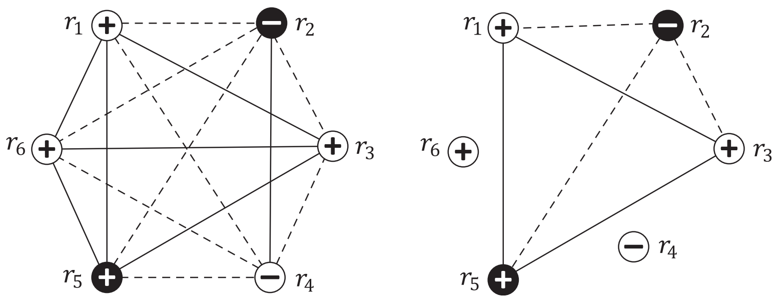
In terms of input and output, report confirmation approaches start from a collection of reports published in a LBSN where each report is represented by a tuple $r = (observer, time, place, observation)$. A confirmation value is computed for each ordered pair of reports as output. Consider the following example: six reports have been published by six different observers who are all positioned on the central avenue of the same city. The observations either affirm or deny that a protest march is moving along the avenue from point s_1 to point s_2 during the time period $[t_1, t_2]$.Since positions and movements refer to a linear geographic object, only a single spatial dimension needs to be represented in the example. In practice, however, confirmation filtering refers to two, sometimes to three spatial dimensions (latitude, longitude, geoid height). Each of the reports $r_1, \ldots, r_6$ corresponds to a point in the space-time diagram (Fig. 1).

Affirmative observations are denoted by (+), negative ones which deny the event by (−). In the example, every observer either supports or opposes the issue of the protest march with social ties being established only within each of the two subcommunities. The space-time diagram distinguishes reports of supporters (white) from those of opposers (black). While reports appear as points in the diagram, events have a spatial and a temporal extension, that is, they cover regions. The spatiotemporal coverage of a protest march of length D moving with constant speed is a parallelogram.

Without spatiotemporal and social context, it is only possible to determine agreement and disagreement between reports. An affirmative report r agrees with every report r_i affirming the same observation. This expressed by the affirmation value $a(r, r_i) = 1$. Similarly, the affirmative report r disagrees with any negative report r_j resulting in $a(r, r_j) = -1$. The graph in Fig. 2 shows agreement by solid lines and disagreement by dashed lines. Agreement, however, is only a necessary, not a sufficient condition for confirmation.



Spatiotemporal Proximity and Social Distance, Fig. 1 Space-time diagram



Spatiotemporal Proximity and Social Distance, Fig. 2 Agreement graph and confirmation graph

Confirmation Graph

Confirmation is modelled by a real-valued function which maps a pair of reports onto a confirmation value $c : (r_i, r_j) \rightarrow v \in [-1, 1]$ where $c(r_i, r_j) = 1$ corresponds to maximal confirmation, $c(r_i, r_j) = 0$ to unrelated reports, and $c(r_i, r_j) = -1$ to maximal conflict. The *spatial proximity* of two reports r_i and r_j is computed by considering the straight line distance or the street distance $d(r_i, r_j)$ between their place stamps and by taking into account additional spatial constraints. In the example, an additional constraint is the maximal length D of a protest march in that city. If the place stamps are farther apart than D, then it is rather unlikely that the two reports refer to the same event: $c_{spa}(r_i, r_j) = a(r_i, r_j)$ when $d(r_i, r_j) \leq D$, and $c_{spa}(r_i, r_j) = 0$ otherwise. Often, a logistic function is used to express the vagueness of the spatial threshold, $c\left(r_i, r_j\right) = a\left(r_i, r_j\right) / \left(1 + e^{\left(d\left(r_i, r_j\right) - D\right)}\right)$.

In a comparable way, temporal constraints such as the maximal duration of an event are exploited to determine *temporal proximity*. Spatial and temporal constraints interact in the movement of the protest march. In such cases spatial distances and temporal distances cannot be considered independently. Measures of *spatiotemporal proximity* take account of the dependency of spatial and temporal constraints such as the assumption of uniform motion of the protest march in the example.

Figure 2 shows the agreement graph (left) and the result of confirmation filtering with the spatial confirmation function $c_{spa}(r_i, r_j)$ and an analogous temporal confirmation function. Generally, confirmation filtering removes edges. Report r_6, for instance, agrees with reports r_1, r_2, r_3, and r_5 but is not considered to confirm those reports because it is spatially farther than D from each of these reports. For the same reason, the negative report r_4 does not confirm the negative report r_2.

A further filtering step evaluates social distance and addresses the issue of (social) independence of sources. In the example, the authors of the reports r_1, r_3, r_4, and r_6 have established links in the social network forming a connected component isomorphic to K_4, while the authors of r_2 and r_5 form a second connected component isomorphic to K_2. There are only two social distances: between different nodes of the same component $d_{soc}(r_i, r_j) = 1$, otherwise $d_{soc}(r_i, r_j) = \infty$. The simplest filtering approach only identifies confirmation edges between socially distant nodes: $c_{soc}(r_i, r_j) = c(r_i, r_j)$ when $d_{soc}(r_i, r_j) > 1$, and $c_{soc}(r_i, r_j) = 0$ otherwise. Applying this filter results in a further pruning of the confirmation graph. Only the edges between r_1 and r_5 , r_1 and r_2, r_2 and r_3, as well as between r_3 and r_5 remain.

In application scenarios, the graphs are much larger, and the distance measures reflect more complex modelling assumptions. Often it is possible to represent a further filtering step which exploits confirmation relations between more than just two reports as a finite domain constraint satisfaction problem which can be solved with algorithmic methods from qualitative spatiotemporal reasoning (Ligozat 2012; Yanenko and Schlieder 2012).

Future Directions

Hybrid approaches combine constraint-based confirmation graph filtering with other methods such as content analysis and environmental contextualization.

Empirical studies provide insights into the type of constraints relevant for spatiotemporal proximity filtering. Truelove et al. (2015) use linguistic content analysis to distinguish reports written by witnesses of an event from accounts that are relayed. Their analysis of a Twitter data set about a wildfire shows that spatiotemporal proximity filtering may need to handle reports, which relate to different types of observations by different filtering rules. In a wildfire, only very few microbloggers are sufficiently close to report seeing flames. This contrasts with reports of seeing or smelling smoke, which can be made from much farther distance. The confirmation graph approach is compatible with multiple filtering rules if the decision on which rule to apply can be based on an additional information source such as a linguistic content analysis that determines the observation type.

Besides the linguistic analysis of observation reports, environmental contextual data may help to improve the confirmation graph filtering. Mehdipoor et al. (2015) have used climate data to identify inconsistent observations in a phenology project, where volunteer contributors report flowering onset dates. A model-based clustering based on the climate data permits to detect outliers. Such a contextual filtering can be integrated with the confirmation graph approach since it does not interfere with the rules for spatiotemporal proximity and social distance filtering.

Cross-References

▶ Location-Based Social Networks

References

Bishr M, Mantelas L (2008) A trust and reputation model for filtering and classifying knowledge about urban growth. GeoJournal 72(3–4):229–237

Deuze M, Bruns A, Neuberger C (2007) Preparing for an age of participatory news. Journal Pract 1:322–338

Ligozat G (2012) Qualitative spatial and temporal reasoning. Wiley, Hoboken

Mehdipoor H, Zurita-Milla R, Rosemartin A, Gerst K, Weltzin J (2015) Developing a workflow to identify inconsistencies in volunteered geographic information: A Phenological Case Study. PLoS One 10(10):e0140811. https://doi.org/10.1371/journal.pone.0140811

Okolloh O (2009). Ushahidi or ‚testimony’: Web 2.0 tools for crowdsourcing crisis information, Participatory Learning and Action, 59, special issue on Web 2.0 for Development, International Institute for Environment and Development. pp 65–70

Schlieder C, Yanenko O (2010) Spatio-temporal proximity and social distance: a confirmation framework for social reporting. In: Zhou X et al (eds) Proceedings of the 2nd international workshop on location based social networks. ACM Press, New York, pp 60–67

Symeonidis P, Ntempos D, Manolopoulos Y (2014) Location-based social networks. In: Recommender systems for location-based social networks. Springer, New York, pp 35–48

Truelove M, Vasardani M, Winter S (2015) Towards credibility of micro-blogs: characterising witness accounts. GeoJournal 80:339–359. https://doi.org/10.1007/s10708-014-9556-8

Yap L, Bessho M, Koshizuka N, Sakamura K (2012) User-generated content for location-based services: a review. In: Lazakidou A (ed) Virtual communities, social networks and collaboration. Springer, Berlin, pp 163–179

Yanenko O, Schlieder C (2012) Enhancing the quality of volunteered geographic information: a constraint-based approach. In: Gensel J et al (eds) Bridging the geographic information sciences. Springer, Berlin, pp 429–446

Spatio-temporal Querying of Big Data and Results Summarization

▶ Geotemporal Querying of Social Networks and Summarization

Spatiotemporal Recommendation in Geo-Social Networks

Hongzhi Yin[1], Bin Cui[2] and Xiaofang Zhou[1]
[1]School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, QLD, Australia
[2]Key Lab of High Confidence Software Technologies (MOE), School of EECS, Peking University, Beijing, China

Synonyms

Location-based recommendation; Mobile recommendation; New city recommendation; Next POI recommendation; Out-of-town recommendation; Place recommendation; POI recommendation; Real-time recombination; Spatial item recommendation; Spatiotemporal context-aware recommendation; Successive POI recommendation

Glossary

Home Location	A user's home location is the place where the user is living. It is a static location instead of a real-time location that is "temporally" related to the user (e.g., the places where the user is visiting)
LBSNs	Location-based social networks
Out-of-Town Recommendation	If the distance between the target user's current location and home location is larger than a threshold d (say 100 km), the spatial item recommendation problem is defined as out-of-town recommendation; otherwise, the problem is hometown recommendation
POI	Point of interest
Spatial Item	A spatial item is an item associated with a geographical location (e.g., a restaurant or a cinema), referring to either a point of interest or an offline activity held in the physical world
Spatiotemporal Recommender	Given a target user u_q with the current time t_q and location l_q (i.e., $q = (u_q, t_q, l_q)$), spatiotemporal recommender is a system that recommends top-k spatial items that u_q would prefer in the given spatiotemporal context

Definition

Unlike traditional virtual social media sites (like digg.com, blogger.com, livejournal.com, delicious.com, pandora.com), the emerging geo-social networks play the role of the proxy of the physical world. Correspondingly, the spatial items are physical items in the real world and have geographical property, which are significantly different from the traditional social media content items (e.g., comments, music, movie, news, and posts). Actually, the geo-social data in the mobile era captures the snapshots of users' everyday lives and provides unprecedented potential for user interest modeling, mobility pattern mining, and user behavior prediction in the physical world. On the other hand, it is crucial to develop spatiotemporal recommendation services for mobile users to explore the new places, attend new events, and find their potentially preferred spatial items from billions of candidate ones.

Spatiotemporal recommendation in geo-social networks refers to techniques used to make personalized recommendation of spatial items, based on the following factors:

- The multimodal content (e.g., user comments, categories, tags and images) and geographical location associated with the spatial item

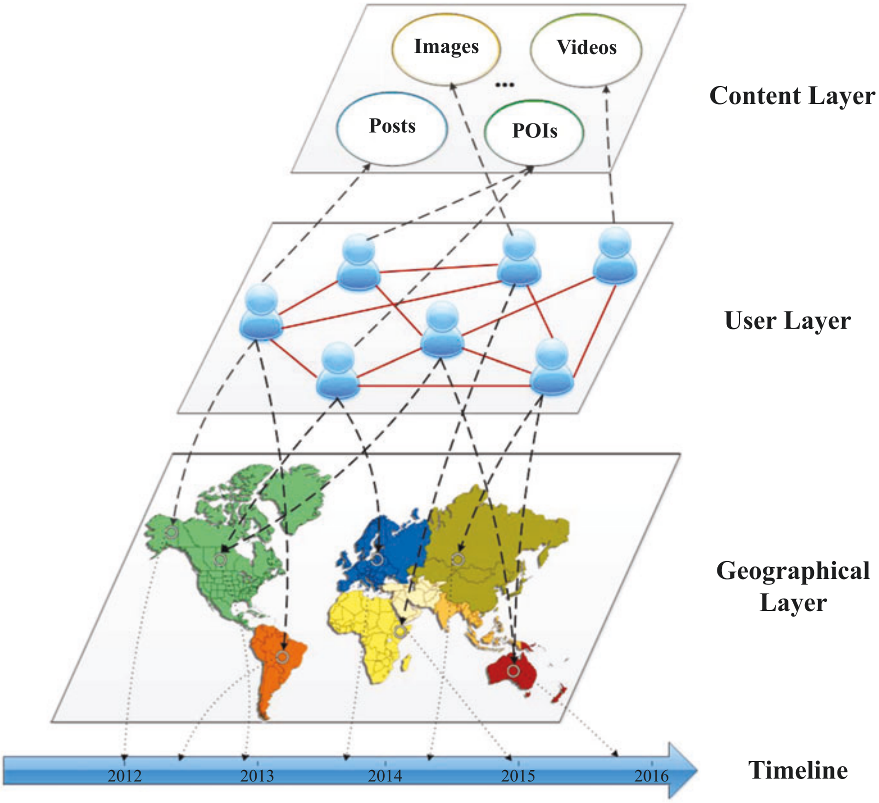
- Both user interests and mobility patterns (e.g., activity ranges)
- The sequential influence from other spatial items recently visited by the same user
- Temporal cyclic patterns of user mobility behaviors
- The wisdom of the crowds who share the same role with the user with respect to the target geographical region
- Social influence from friends who are familiar with the target geographical region
- Temporal dynamic and evolution of user interests

Introduction

The rapid development of Web 2.0, location acquisition, and wireless communication technologies has fostered a profusion of geo-social networks, such as location-based social networks (LBSNs) and event-based social networks (EBSNs). LBSNs (e.g., Foursquare, Yelp, and Google Place) provide users an online platform to share their locations (usually referred to as "check-in"), find interesting points of interest (POIs, e.g., cinemas, galleries, and hotels), generate comments, images, and tips associated with their visited POIs, build connections with friends, and find other friends who are nearby. Moreover, newly emerging EBSNs (Liu et al. 2012) (e.g., Meetup and Plancast) enable users to share more specific activities/events held in the physical world, ranging from informal get-togethers (e.g., movie nights and dining out) to formal activities (e.g., culture salons and business meetings). Compared to traditional GPS trajectories that only track the movement of moving objects, geo-social networking data captures the semantically meaningful snapshots of users' everyday lives and thus provides unprecedented opportunity to better understand users' behaviors. For simplicity, we use the notion of spatial items to denote both points of interests (POIs) and events in a unified way. The use of computational techniques to "best-match" billions of spatial items to billions of users in potentially billions of locations globally will lead to unprecedented multifold benefits for common users, business, content, or service providers. Many recent surveys predict hundreds of billions of dollars in revenue for location-based marketing (Dhar and Varshney 2011). The rapid growth of geo-social networks has led to the availability of large-scale semantically meaningful spatiotemporal data generated by users, which consists of spatiotemporal trajectories, multimodal contents, and social connections, providing both opportunities and challenges for researchers to study users' behaviors and their decision-making for spatial items.

Typical online geo-social networking sites provide location-based services that allow users to "check-in" at physical places and events and automatically include the geographical location into their posts. "Check-in" is an online behavior that posts a user's current geographical position to share personal offline activities in the physical world and tell personal friends where and when s/he is via the social media. Compared with other user online behaviors such as following, rating, voting, tagging, viewing, downloading, and purchasing that interact with the virtual world, "check-in" reflects a user's activity in the real world, residing where the online world and real world interacts. In the geo-social networks, "check-in" not only adds a spatial dimension to the online social networks but also plays an important role in bridging the gap between online social networks and the physical world (Gao and Liu 2014).

Figure 1 shows a "4 W" (i.e., who, when, where, and what) information layout of geo-social networks. The geographical layer contains the historical check-ins of users, which indicates the spatial trajectories of users. The content layer consists of users' feedback to the spatial items, such as comments, tips, images, and videos, which captures the semantic information of users' activities in the physical world. The user layer provides extensive knowledge about users such as their social structure. All these three layers share one timeline, indicating the temporal information of the users' "check-in" behaviors. Thus, the geo-social network data in the mobile era



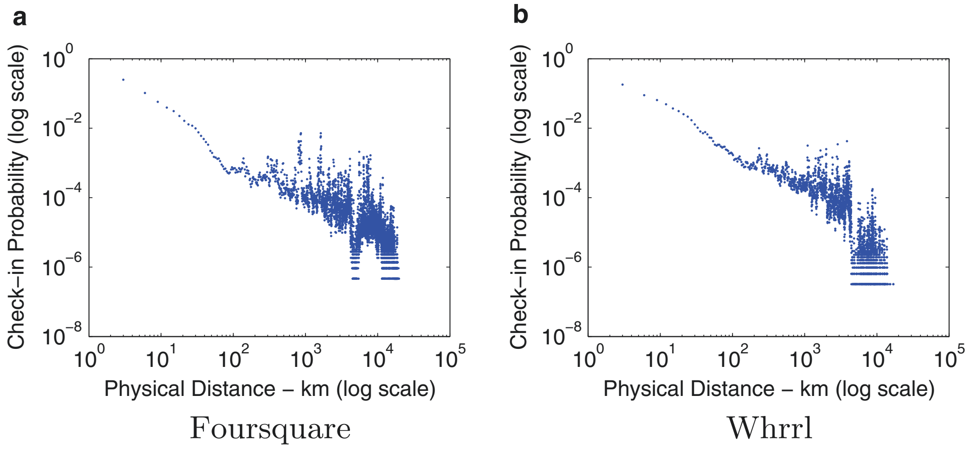
Spatiotemporal Recommendation in Geo-Social Networks, Fig. 1 The information layout of geo-social networks (Gao and Liu 2014)

captures the semantically meaningful snapshots of human everyday activities in the physical world.

Key Points

According to the recent studies, users' visiting behaviors in the real world are influenced by the following factors:

- **Geographical Influence**. Many recent studies show that people tend to explore spatial items near the ones that they have visited before, even if they are far away from home, e.g., a user may explore some restaurants and shops around Time Square when she goes there for a Broadway show (Ye et al. 2011c; Cheng et al. 2012; Lian et al. 2014). To better understand this geographical influence on users, Ye et al. (2011c) performed a spatial analysis on real datasets of user check-in activities collected from two well-known LBSNs, i.e., Foursquare and Whrrl. Specifically, they aimed to study the implication of distance on user check-in behavior by measuring how likely two of a user's check-in POIs are within a given distance. To obtain this measurement, they calculated the distances between all pairs of POIs that a user has checked in and plot a probability density function over the distance of POI pairs checked in by the same user. As shown in Fig. 2, a significant percentage of POIs pairs checked-in by the same user appears to be within short distance, indicating a geographical clustering phenomenon in user check-in activities.
- **Temporal Cyclic Effect**. As suggested in (Ye et al. 2011a; Yuan et al. 2013; Gao et al. 2013), human movement exhibits temporal



Spatiotemporal Recommendation in Geo-Social Networks, Fig. 2 Geographical influence probability distribution

cyclic patterns which are highly relevant to the content of spatial items, and the daily pattern (hours of the day) is one of the most fundamental patterns that reflect a user's mobility behavior. For example, a user may regularly arrive at the office around 9:00 a.m., go to a restaurant for lunch at 12:00 p.m., and watch movies at night around 8:00 p.m. Moreover, a user's preferences change continuously over time, indicating two temporal properties of a user's daily check-in preferences: (1) **nonuniformness**: a user exhibits distinct check-in preferences at different hours of the day; and (2) **consecutiveness**: a user tends to have more similar check-in preferences in consecutive hours than in nonconsecutive hours. Therefore, investigating the features embedded in daily patterns enables us to better understand human mobility behavior, providing a potential opportunity to design more advanced recommender systems on the geo-social networks.

- **Sequential Influence**. In reality, human movement exhibits sequential patterns, which serve as the basis for mobility prediction (Cho et al. 2011). In particular, an analysis has been conducted on three publicly available real-world datasets, Foursquare, Gowalla, and Brightkite in Zhang and Chow (2015), which calculates the probabilities of each of the next spatial items immediately visited by a user after visiting a given spatial item. The results show that each selected spatial item transits to the top hundred items out of several hundred thousand items with a probability greater than 0.5. The nonuniform distribution of transition probabilities between spatial items indicates there are underlying sequential patterns between spatial items visited by users. These sequential patterns result from different factors, such as time in one day (e.g., people tend to go to restaurants at dinner time and then relax in cinemas or bars at night), geographical proximity (e.g., tourists often sequentially visit London Eye, Big Ben, and Downing Street), and the coherence between human preference and the type of places (e.g., people usually check in at a stadium before a restaurant instead of the reverse way because it is not healthy to exercise right after a meal (Wang et al. 2016; Feng et al. 2015)).
- **Social Influence**. Cho (Cho et al. 2011) studied the relation between friendship and user mobility in two geo-social networks, Brightkite and Gowalla. They observed that long-distance travel (i.e., traveling out of town) is more influenced by social network ties, while short-ranged travel (traveling at hometown) is not significantly effected by the social network structure, since people are more likely to visit friends or visit places their friends have visited in the past when they travel far away than at hometown.

- **Semantic Effect**. A recent analysis of the Whrrl dataset shows that the check-in activities of users exhibit a strong semantic regularity (Ye et al. 2011b). In the analysis, Ye et al. studied the diversity of POIs that individual users visit by computing the entropy of semantic categories in their check-ins. The results show that most of the users have very small entropies, indicating that POIs checked-in by an individual user usually have similar semantics (e.g., categories and topics).

Many methods (Yin et al. 2013, 2014b, 2015c, d; Liu et al. 2013; Li et al. 2015; Zhao et al. 2015; Yin et al. 2015a) have been recently developed to exploit the above factors to improve the performance of recommenders in geo-social networks. A comprehensive review of these methods is beyond the scope of this entry. Instead, after providing a brief historical background, we illustrate the key ideas in spatiotemporal recommendation in geo-social networks through a generic and flexible probabilistic graphical model, which mimics user check-in behaviors in a process of decision-making by joint modeling the above factors.

Historical Background

The first commercial location-based social networking service available in the United States is Dodgeball (http://en.wikipedia.org/wiki/Dodgeball), launched in 2000. It allows users to "check in" by broadcasting their current locations through short messages to their friends who are within a 10 block radius; users can also send "shouts" to organize a meeting among friends at a specific place. After acquired by Google in 2005, the original Dodgeball has been replaced by Google Latitude in 2009, while the founder of Dodgeball launched a new location-based social networking service "Foursquare" in the same year. Foursquare utilizes a game mechanism in which users can compete for virtual positions, such as mayor of a city, based on their check-in activities. It has become one of the most successful location-based social networking sites in the United States. Facebook also launched its location-based service, namely Facebook Places in 2010 with its check-in function, and acquired another popular LBSN, Gowalla (http://en.wikipedia.org/wiki/Gowalla), at the end of 2011.

Compared to the traditional spatial trajectories (e.g., GPS trajectories) that only track the time and location information of mobile users, there are multifaceted benefits of leveraging geo-social networking data in recommendation systems. First, the content information in geo-social networking data that refers to categories, tags, images, and user comments associated with spatial items provides a unique opportunity to infer and study users' interests. For example, by observing a spatial item's description as "vegetarian restaurant," one can infer that users who checked-in at this spatial item are interested in vegetarian food and diet. Second, the social network information in geo-social networking data provides extensive knowledge about users' social structures, which are critical to infer and predict users' interests/preferences according to the homophily principle, especially when confronted with the sparsity of user check-in data. For example, two users with a strong social relationship are more likely to enjoy same or similar interests. Third, geo-social networking data provides more fine-grained spatial information which can distinguish between two different spatial items that are located at the same geographical coordinates. This is critical for generating more personalized and accurate recommendations. For example, users can check-in at a cinema or a restaurant at the same shopping mall where both venues share the same geographical coordinates. In contrast, other trajectory data such as cell phone data and GPS trajectory data provides coarse location accuracy and cannot differentiate between users' presence across different floors in the same building.

Recommendation on Trajectory Data Spatial item recommendation was firstly investigated and studied on trajectory data. Due to the lack of mapping relationship between geographical coordinates and specific real-world POIs, a POI is usually defined as the stay points extracted from users' trajectory logs (Zheng et al. 2009; Zheng and Xie 2011). Because of the unavailability of content information associated with spatial items, spatial and temporal patterns are commonly integrated into collaborative filtering methods to make recommendation. Many trajectory-based recommendation algorithms and models have been recently proposed in (Zheng et al. 2009, 2010; Zheng and Xie 2011), which focus on finding users with similar travel trajectories or places with strong correlations (i.e., locations that are frequently co-visited in one trip). This information is then leveraged by collaborative-filtering-based (CF) methods, which make use of the travel histories of a group of similar users (i.e., user-based CF) or a set of similar locations (i.e., item-based CF) to generate location recommendations. However, their performance is largely impeded due to the lack of semantic interpretation of users' activities.

Recommendation on Traditional Social Media Previous research on recommendation has investigated the social and content layers with traditional social media data, and many research efforts (Yin et al. 2014a, 2015b; Xu et al. 2012; Stoyanovich et al. 2008; Liu et al. 2010) have been undertaken to analyze and model online users' behaviors to help them find interesting items on traditional social media platforms such as Twitter, Pandora, and Delicious. Compared with the traditional social networks, the geo-social networks have additional geographical layer. The unique geographical property information presents new challenges for the recommender systems, since traditional recommendation methods would fail to exploit and integrate the new impact factors in geo-social networks, such as geographical influence, temporal cyclic effect, and sequential influence.

Spatial Item Recommendation

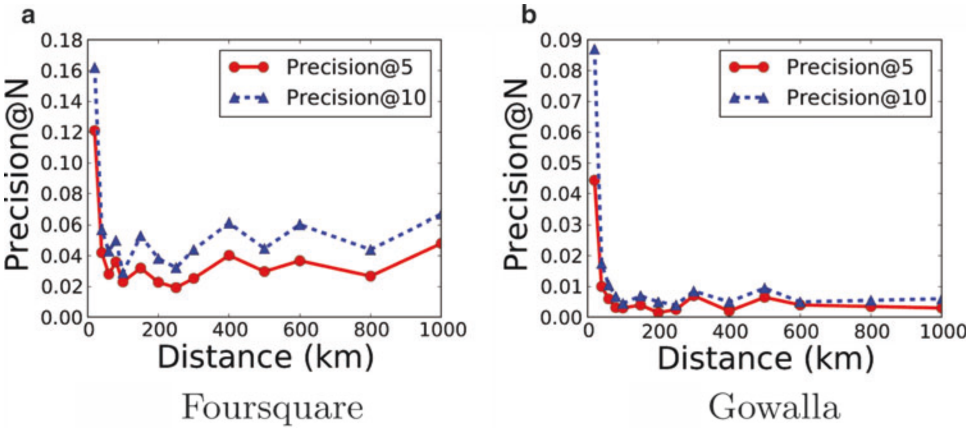
It is crucial to utilize user check-in data to make spatiotemporal recommendation in geo-social networks, which helps users know new places and explore new regions (e.g., cities), facilitate advertisers to launch mobile advertisements to targeted users. This application becomes more important and useful when a user travels to an unfamiliar area, where she has little knowledge about the neighborhood. In this scenario, the recommender system is proposed as recommendation for out-of-town users in (Ference et al. 2013; Yin et al. 2015c, d; Wang et al. 2015), also referred to new city recommendation (Yin et al. 2013, 2014b). Most of the existing location-based recommender models were designed for hometown scenario, while only few work (Ference et al. 2013; Yin et al. 2013, 2014b, 2015c, d; Wang et al. 2015) focused on recommendation for both hometown and out-of-town users. Spatial item recommendation, especially out-of-town recommendation, is a highly challenging problem because of the following three main reasons: *Data Sparsity*, *Travel Locality*, and *Spatial Dynamic of User Preferences*.

Data Sparsity In traditional trajectory data, the geographical location of a moving object is automatically recorded through the GPS, while on geo-social networks, the check-in process is user-driven (Noulas et al. 2011), i.e., the user decides whether to check in at a specific place or not due to certain privacy concerns. In the geo-social networks, the number of spatial items checked-in by an individual user is rather small compared to the total number of spatial items, which results in a very sparse user-item matrix. This issue plagues most of the existing collaborative filtering-based recommender systems.

Travel Locality Moreover, the observation of travel locality exacerbates the issue of data sparsity. The observation of travel locality (Levandoski et al. 2012) shows that most of users' check-ins are generated at their living regions (e.g., home cities), since users tend to travel a limited distance when visiting venues and attending events. An investigation shows that the check-in records generated by users in their nonhome cities are very few and only take up 0.47% of the check-in records they left in their home cities. This observation of travel locality is quite common in the real world (Scellato et al. 2011), aggravating the data sparsity problem with personalized spatial item recommendations (e.g., if we want to suggest spatial items located in Los Angeles to people from New York City) (Ference et al. 2013; Yin et al. 2013). In this case, CF-based and graph-based approaches will be incapable of providing effective recommendations, again, especially when coping with the *out-of-town recommendation* problem, because a target user usually has scarce check-ins in an out-of-town region.

There exists a considerable body of research (Levandoski et al. 2012; Ye et al. 2010; Lian et al. 2014; Gao et al. 2013) which deposited people's check-in history into user-spatial item matrix where each row corresponds to a user's item-visiting history and each column denotes a spatial item. A collaborative filtering-based method is then employed to infer the user's preference regarding each unvisited spatial items. Based on the core idea of collaborative filtering, similar users of the target user (i.e., those who exhibit similar visiting behaviors) are chosen to provide clues for making recommendation. Due to travel locality, most of these similar users are more likely to live in the same region with the target user than other regions. As a recommendation is made by considering spatial items visited by these similar users, most of the recommended spatial items would be located in the target user's hometown.

Let us assume, for example, that target user u is a shopaholic and often visits shopping mall v' in her home city; v is a popular local shopping mall in city r that is new to u. Intuitively, a good recommender system should recommend v to u when she travels to r. However, the pure CF-based methods fail to do so. For the item-based CF (Linden et al. 2003; Sarwar et al. 2001), there are few common users between v and v' according to the property of travel locality, resulting in the low similarity between the two items' user vectors. For the user-based CF (Adomavicius and Tuzhilin 2005), it is most likely that all the k nearest neighbors of user u live in the same city as u and that few of them have visited v according to the property of travel locality. Ference et al. (2013) further conducted experiments using user-based CF on two real geo-social network datasets (Foursquare and Gowalla) to observe how effectively it performs when making recommendations to users located at different distances from their home regions (i.e., the region where the majority of their check-in activities occur). By varying the distance from the home region, Fig. 3 plots *precision@N* for different distance ranges (e.g., 0–20 km, 20–40 km). The user-based CF method performs reasonably well when users travel around their home regions, but the precision degrades when they travel 50–100 km away from their home regions. This phenomenon occurs in both datasets consistently, acknowledging our concern that CF-based methods cannot work well for out-of-town users.



Spatiotemporal Recommendation in Geo-Social Networks, Fig. 3 Recommendation effectiveness of user-based CF

Besides, some recent literatures (Lian et al. 2014; Gao et al. 2013) adopted powerful latent factor models (e.g., matrix factorization) to a user-spatial item matrix to alleviate the data sparsity. However, they do not perform well for the out-of-town scenario because most of the check-ins in the training set are hometown check-ins, and two spatial items located in two different regions, especially when the two regions are far away from each other, have extremely low probability of co-occurring in the same users, resulting in very weak correlation between them. Thus, the predicted rating of a target user to an out-of-town spatial item would be very low, although the user potentially prefers that spatial item.

A promising way for alleviating the issues of data sparsity and travel locality, especially in the out-of-town recommendation scenario, is to exploit and integrate the semantic content information of spatial items. Some recent literatures (Liu and Xiong 2013; Yin et al. 2013, d; Wang et al. 2015) exploited the content information of checked-in spatial items (e.g., categories or tags) to infer the users' interests, which were then used to make POI recommendation. By integrating the content information, these methods can alleviate the data sparsity issue for out-of-town recommendation to some extent, since they can transfer users' interests inferred at their hometowns to out-of-town regions by the medium of contents, i.e., we can recommend spatial items for out-of-town users according to the contents of the ones they have visited at their hometowns.

Spatial Dynamic of User Preferences Another challenge arises from the spatial dynamic of user preferences, also called user interest drift across geographical regions. Users tend to have different preferences when they travel in different regions which have different urban compositions and cultures, as analyzed in (Wang et al. 2015; Yin et al. 2016b). For example, a user u never goes gambling when he lives in Beijing, China, but when he travels in Macao or Las Vegas he is most likely to visit casinos. Through Foursquare API (https://developer.foursquare.com/), we derive top four categories of a group of users' checked-in POIs for three different cities, as shown in Table 1. The group size is 3000 and each member of this group has check-in records in Gold Coast (AU), Las Vegas (USA), and Istanbul (Turkey). Diverse POI types are observed: when traveling in Las Vegas they are interested in casino (80.32%), nightlife (10.61%), and outlet (5.82%), while they prefer beach (71.36%), surf spot (14.82%), and theme park (9.60%) when traveling in Gold Coast.

Spatiotemporal Recommendation in Geo-Social Networks, Table 1 Illustration of spatial dynamic of user preferences

City	Top POI types	Percentage of check-ins (%)
Gold Coast (AU)	Beach	71.36
	Surf spot	14.82
	Theme park	9.60
Las Vegas (USA)	Casino	80.32
	Nightlife	10.61
	Outlet	5.82
Istanbul (Turkey)	Mosque	68.32
	Museum	15.45
	Cafe	7.65

An Example Recommender

To address the above challenges in a unified model framework, we proposed Geo-SAGE, a geographical sparse additive generative model for spatial item recommendation in (Wang et al. 2015). Traditional mixture models, such as (Yin et al. 2013, 2014b), consider multiple factors that influence a user's choice of spatial items by introducing additional latent variables which act as "switches" to control which factor is currently active. Unfortunately, it is not only computationally expensive to learn personalized "switching" variables for individual users but also difficult to learn these variables accurately given sparse datasets. Inspired by the Sparse Additive Generative model (SAGE) (Eisenstein et al. 2011; Hu and Ester 2013; Hong et al. 2012), we have designed our model in a similar way by adding the effect of all the facets in the log space to avoid the inference of the latent "switching" variables, with the aim of achieving improved robustness and predictive accuracy. To model user check-in behaviors, Geo-SAGE takes into account the factors of **geographical influence**, **user interests** (i.e., semantic effect), **spatial dynamic of user interests**, **social influence,** and **temporal cyclic effect**.

In Geo-SAGE, we first exploit the geographical influence to model users' mobility patterns. Different from users' online behaviors in the virtual world, users' check-in activities in the physical world are limited by travel distance. So, it is important to capture users' spatial patterns (or activity ranges) according to the location distributions of their historical checked-in spatial items. The spatial clustering phenomenon indicates that users are most likely to check-in a number of POIs which are usually limited to some specific geographical regions (Ye et al. 2011c) (e.g., "home" and "office" regions). Thus, all spatial items are divided into R regions according to their geographical locations and check-in densities. Following literatures (Liu et al. 2013; Lichman and Smyth 2014), we assume a Gaussian distribution for each region r, and the location for spatial item v is characterized by $l_v \sim \mathrm{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$, as follows:

$$P(l_v | \boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}_r|}} \exp\left(\frac{-(l_v - \boldsymbol{\mu}_r)^T \boldsymbol{\Sigma}_r^{-1}(l_v - \boldsymbol{\mu}_r)}{2}\right) \tag{1}$$

where $\boldsymbol{\mu}_r$ and $\boldsymbol{\Sigma}_r$ denote the mean vector and covariance matrix. We apply a vector of regions ϑ_u^{user} to model u's spatial patterns. We can compute the probability of each region that an individual user is likely to visit according to the location distribution of her/his historical visited spatial items or her/his current location.

Geo-SAGE learns user interests $\boldsymbol{\theta}_u^{user}$ as a vector of latent topics, by mining both the co-occurrence patterns of spatial items and their content information (e.g., tags and categories), inspired by (Yin et al. 2013, 2014b). Exploiting the content information of spatial items not only alleviates the data sparsity issue but also addresses the challenge of travel locality. The content of spatial items serves as the medium for transferring user interests learned from check-ins generated at hometown to unfamiliar out-of-town regions.

To adapt to *spatial dynamic of user interests*, Geo-SAGE leverages two types of wisdom of the crowds. First, we recognize two roles of an individual user in a spatial region: local or tourist. Given a region r, check-in records from local users are mined to learn *native preference* $\boldsymbol{\theta}_r^{native}$ as a vector of latent topics. Similarly, check-in records from tourists are used to learn *tourist preference* $\boldsymbol{\theta}_r^{tourist}$. Users with the same role at a region are more likely to have similar preferences and behavior patterns. Second, we exploit and integrate the social influence, as recent research (Cho et al. 2011) showed that there is a strong correlation between users' friendship and their mobility, especially when users travel out of town. Given a target user u and a target region r, we first find u's friends who have enough check-in records at region r. These friends' interests, represented as a vector of latent topics $\boldsymbol{\theta}_{u,r}^{social}$, provide important clues and references for recommendation. Thus, to recommend spatial items to a target user u at region r, we consider not only u's personal interests but also the crowd preferences (native preferences or tourist preferences) and the friends' interests.

Temporal influence plays an important role in analyzing users' daily activities in geo-social networks. For example, a user is more likely to go to a restaurant rather than a bar at noon. Therefore, the recommendation should be time-aware. To model the temporal influence, an intuitive solution is to split time into time slices at the predefined granularity (e.g., hourly or seasonal) and then model the temporal preference of a user based on the spatial items visited by the user in a time slice. However, splitting a user's activity data into multiple slices will make the data much sparser in a specific time slice. To effectively exploit the temporal influence and support time-aware recommendation, we extend Geo-SAGE by integrating the collective temporal preferences of the crowd $\left(\boldsymbol{\theta}_{r,t}^{native} \text{ or } \boldsymbol{\theta}_{r,t}^{tourist}\right)$ with the same role instead of personal temporal preferences.

Specifically, given a user, a specific time and his/her current location, we first find the crowd of users sharing same role (e.g., tourists), and then produce time-aware recommendations based on the crowd's temporal preferences, the user's personal interests, and his/her friends' preferences for the target region.

The generative process of Geo-SAGE for an activity record (u, v, t, s, l_v, V_v) in the user profile V_u is listed as follows, and the graphical representation of Geo-SAGE is shown in Fig. 4.

- Draw a region index r

$$r \sim P\left(r \mid \vartheta_u^{user}, \vartheta^0\right)$$

- Draw a topic index z

$$z \sim P\left(z \mid s, \boldsymbol{\theta}_u^{user}, \boldsymbol{\theta}_{r,t}^{native}, \boldsymbol{\theta}_{r,t}^{tourist}, \boldsymbol{\theta}_{u,r}^{social}\right)$$

- For each word w in Wv, draw

$$w \sim P\left(w \mid \phi_z^0, \phi_z^{topic}\right)$$

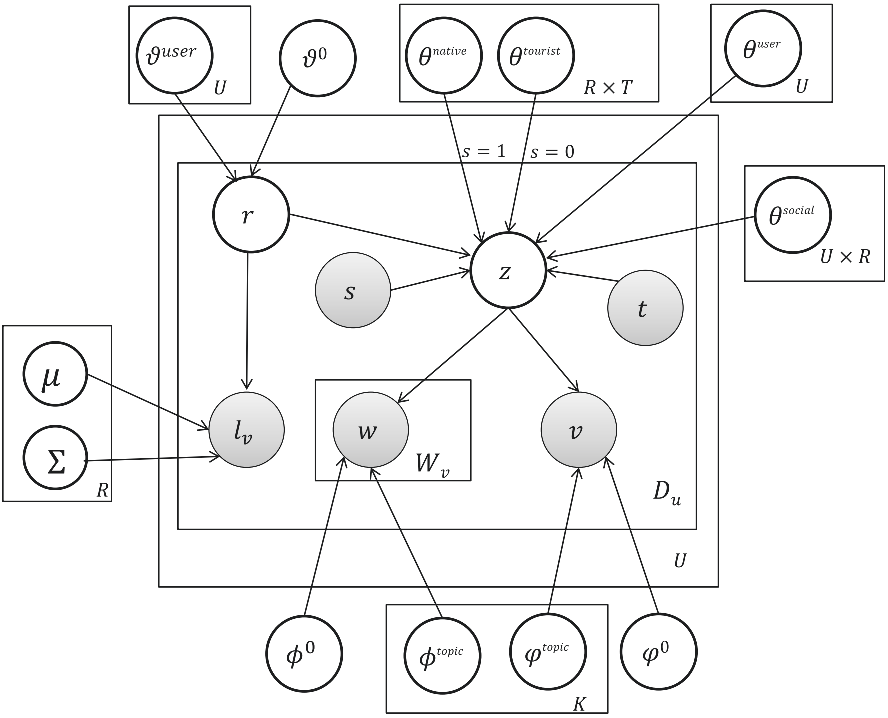
- Draw the ID of spatial item v

$$v \sim P\left(v \mid \varphi_z^0, \varphi_z^{topic}\right)$$

- Draw the geographical location of v

$$l_v \sim \mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$$

where v is the index of spatial item; s indicates the role of user u when she/he checked-in v: $s = 1$ indicates the user is a local while $s = 0$ indicates the user is a tourist. t is the check-in time, l_v is the geographical location of v, and W_v is a set of words describing v.



Spatiotemporal Recommendation in Geo-Social Networks, Fig. 4 Graphical representation of Geo-SAGE model

For each check-in record, Geo-SAGE first chooses the region where the check-in occurs. To generate the region index r, we utilize a multinomial model as follows.

$$P\left(r|\,\vartheta^0, \vartheta_u^{user}\right) = P\left(r|\,\vartheta^0 + \vartheta_u^{user}\right) = \frac{exp\left(\vartheta_r^0 + \vartheta_{u,r}^{user}\right)}{\sum_{r'} exp\left(\vartheta_{r'}^0 + \vartheta_{u,r'}^{user}\right)} \tag{2}$$

where ϑ^0 is a background model, i.e., a global distribution over regions. To generate the topic index z that the check-in is about, we utilize a multinomial model as follows.

$$\begin{aligned} &P\left(z|\,s, \boldsymbol{\theta}_u^{user}, \boldsymbol{\theta}_{r,t}^{native}, \boldsymbol{\theta}_{r,t}^{tourist}, \boldsymbol{\theta}_{u,r}^{social}\right) \\ &= P\left(z|\,\boldsymbol{\theta}_u^{user} + s \times \boldsymbol{\theta}_{r,t}^{native} + (1-s)\boldsymbol{\theta}_{r,t}^{tourist} + \boldsymbol{\theta}_{u,r}^{social}\right) \\ &= \frac{exp\left(\theta_{u,z}^{user} + s \times \theta_{r,t,z}^{native} + (1-s) \times \theta_{r,t,z}^{tourist} + \theta_{u,r,z}^{social}\right)}{\sum_{z'} exp\left(\theta_{u,z'}^{user} + s \times \theta_{r,t,z'}^{native} + (1-s) \times \theta_{r,t,z'}^{tourist} + \theta_{u,r,z'}^{social}\right)} \end{aligned} \tag{3}$$

where $\theta_{u,r,z}^{social} = \frac{\sum_{u' \in \mathcal{S}_{u,r}} \theta_{u',z}^{user}}{|\mathcal{S}_{u,r}|}$; $\mathcal{S}_{u,r}$ is a set of u's friends who have enough check-in records at region r.

Once the topic z is generated, the spatial item v and its associated content words W_v are generated as expressed in Eqs. 4 and 5, respectively.

$$P\left(v|\,\varphi_z^0, \varphi_z^{topic}\right) = P\left(v|\,\varphi_z^0 + \varphi_z^{topic}\right) = \frac{exp\left(\varphi_{z,v}^0 + \varphi_{z,v}^{topic}\right)}{\sum_{v'} exp\left(\varphi_{z,v'}^0 + \varphi_{z,v'}^{topic}\right)} \tag{4}$$

$$P\left(w|\,\phi_z^0, \phi_z^{topic}\right) = P\left(w|\,\phi_z^0 + \phi_z^{topic}\right) = \frac{exp\left(\phi_{z,w}^0 + \phi_{z,w}^{topic}\right)}{\sum_{w'} exp\left(\phi_{z,w'}^0 + \phi_{z,w'}^{topic}\right)} \tag{5}$$

To take full advantage of the strengths of both content-based and collaborative filtering-based recommendation methods, a topic z in our Geo-SAGE model is not only associated with a word distribution ϕ_z^{topic} but also with a distribution over spatial items φ_z^{topic}. This design enables ϕ_z^{topic} and φ_z^{topic} to be mutually influenced and enhanced during the topic discovery process by associating them. Thus, the discovered topic z, on the one hand, can cluster the content-similar items together. On the other hand, it can also capture the item co-occurrence patterns to link relevant items together, similar to item-based collaborative filtering methods. We also introduce two background models for words and items, respectively: ϕ^0 and φ^0. The purpose of using background models is to make the topics learned from the dataset more discriminative, since ϕ^0 and φ^0 assign high probabilities to nondiscriminative and noninformative words and items.

We employ a mixture of EM and a Monte Carlo sampler, called the Gibbs EM algorithm, to infer all the model parameters $\Theta = \{\vartheta^0, \vartheta^{0user}, \theta^{user}, \theta^{social}, \theta^{native}, \theta^{tourist}, \phi^{topic}, \phi^0, \varphi^{topic}, \varphi^0, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$. Once the model parameters Θ are learnt from the training dataset of user check-in records, given a querying user u_q at time t_q and location l_q (i.e., $q = (u_q, t_q, l_q)$), we first compute her/his current role $s_{q,}$, and then the probability of u_q visiting spatial item v is computed as in Eq. 6. We choose the top-k spatial items with highest probabilities as recommendations.

$$\begin{aligned} &P\left(v, l_v, \mathcal{W}_v|\,u_q, l_q, t_q, s_q, \Theta\right) = \sum_r P\left(r|\,l_q\right) \\ &\times P(l_v|\,\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r) \times \sum_z P\left(z|\,\boldsymbol{\theta}_u^{user} + s \times \boldsymbol{\theta}_{r,t_q}^{native}\right. \\ &\left.+(1-s_q) \times \boldsymbol{\theta}_{r,t_q}^{tourist} + \boldsymbol{\theta}_{u_q,r}^{social}\right) \\ &\times P\left(v|\,\varphi_z^0 + \varphi_z^{topic}\right) P\left(\mathcal{W}_v|\,\phi_z^0 + \phi_z^{topic}\right) \end{aligned} \tag{6}$$

where the probability of region r given location l_q, i.e., $P(r|l_q)$, is computed as follows:

$$P\left(r|\,l_q\right) = \frac{P(r)P\left(l_q|\,\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r\right)}{\sum_{r'} P(r')P\left(l_q|\,\boldsymbol{\mu}_{r'}, \boldsymbol{\Sigma}_{r'}\right)} \propto P(r) \times P\left(l_q|\,\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r\right) \tag{7}$$

$$P(r) = \sum_u P(r|\,u)P(u) = \sum_u \frac{N_u + \kappa}{\sum_{u'}(N_{u'} + \kappa)} \vartheta_{u,r} \tag{8}$$

where N_u denotes the number of check-ins generated by user u. In order to avoid overfitting, we introduce the Dirichlet prior parameter κ to play the role of pseudocount. Note that to support dynamic real-time recommendation, we compute the probability of u_q choosing region r according to his/her real-time location l_q instead of the spatial patterns (i.e., $\mathrm{q}_{u,\ r}$) learnt from his/her historical check-in records, which distinguishes this work from the static recommendation scheme adopted by most spatial item recommendation work (Ye et al. 2011c; Lian et al. 2014; Gao et al. 2013, 2015; Hu and Ester 2013; Zhao et al. 2015). The probability $P\left(\mathcal{W}_v \middle| \phi_z^0 + \phi_z^{topic}\right)$ is computed as follows:

$$P\left(\mathcal{W}_v \middle| \phi_z^0 + \phi_z^{topic}\right) = \left(\Pi_{w \in \mathcal{W}_v} P\left(w \middle| \phi_z^0 + \phi_z^{topic}\right)\right)^{\frac{1}{|\mathcal{W}_v|}} \tag{9}$$

where we adopt geometric mean for the probability of topic z generating word set W_v, considering that the number of words associated with different spatial items may be different.

Next Spatial Item Recommendation

Next spatial item recommendation (Cheng et al. 2013a, b; Feng et al. 2015) or successive spatial item recommendation, as a natural extension of general spatial item recommendation, is recently proposed and has attracted great research interest. Different from general spatial item recommendation that focuses only on estimating users' preferences on spatial items, next spatial item recommendation provides satisfied recommendations promptly based on users' most recent checked-in spatial items, which requires not only user preference modeling but also the deep exploration of sequential patterns of spatial items (i.e., the correlation between spatial items). Nevertheless, leveraging sequential information for spatial item recommendation is highly challenging, mainly due to the following problems:

Low-sampling rate There are a number of studies that predict next locations on GPS trajectories (Zheng et al. 2009; Song et al. 2004). At first glance, these approaches can be directly applied to geo-social data, since both the GPS and geo-social data contain location and time information. However, the analysis (Cheng et al. 2013b) of the check-in records collected from Gowalla, a popular geo-social network, shows that geo-social data has a low sampling rate in both space and time, compared to GPS trajectories, as the check-in process on geo-social social networks is user-driven (Noulas et al. 2011), i.e., the user decides whether to check in at a specific place or not due to certain privacy concerns. According to the analysis, only 10% of users have more than 58 check-in records over a 12-month period, representing a low check-in frequency over time. In addition, 40% of all consecutive check-ins have a spatial distance larger than 1 kilometer, much longer than the gap in GPS trajectories which is typically 5–10 meters (Zheng et al. 2009). Thus, it is difficult to model the dependency between two check-in locations in geo-social networks using the location prediction techniques on GPS trajectories.

Huge prediction space Sequential recommendation methods have been proposed in the literature (Cheng et al. 2013a; Zheng et al. 2012, 2014), most of which are based on Markov chains. Suppose there are a collection of V spatial items and the next item depends on the previous n items. The sequential recommendation methods then need to estimate V^{n+1} free parameters in the nth order Markov chain model, which is extremely computational-expensive. To reduce the size of the prediction space, most related studies (Cheng et al. 2013a; Zheng et al. 2012) exploit sequential influence using a first-order Markov chain, which considers only the last one in a sequence of locations visited by a user to recommend a new location for her. Although the parameter space can be decreased to V^2, it may still be huge considering that V is usually a large number in geo-social networks. However, in reality, the next location may not only rely on the latest location but also earlier ones visited by the user (Zhang et al. 2014). Hence, we aim to develop a new method to incorporate the influence from all recently visited locations, rather than just the last one, to make location recommendations within a small parameter space.

Time-Awareness Next spatial item recommendation is a time-subtle recommendation task since at different time, users would prefer different next spatial items. It is easy to imagine that a user may go to a restaurant after leaving from office at noon, while he/she may be more likely to go to a gym when he/she leaves office at night. However, most of next spatial item recommendation methods only highlight the modeling of correlations between spatial items within users' check-in sequences but neglect to model the temporal influence.

Unifying Personalization, Temporal Influence, and Sequential Effect On the one hand, most of existing spatial item recommendation methods focusing on personalization (Bao et al. 2012; Ference et al. 2013; Cho et al. 2011) make recommendations according to users' personal interests, but neglect the sequential orders between spatial items. On the other hand, existing sequential recommendation methods, such as Markov chain based approaches, capture sequential patterns by assuming equivalent transition probabilities between items for all users, and ignore personalization. Besides, some other work (Yuan et al. 2013; Gao et al. 2013) exploited the temporal cyclic effect to support time-aware recommendation. A recommendation system that only focuses on one of the three aspects cannot produce ideal results. Therefore, we aim to develop a recommendation method which combines personalization, temporal, and sequential influence between items, in a unified and principled manner.

In light of the aforementioned challenges, we extended the Geo-SAGE model to a **S**equential **P**ers**O**nalized spatial items **RE**commender system (SPORE) in (Wang et al. 2016), which seamlessly fuses the sequential influence of visited spatial items, the temporal influence, and the personal interests of individual users in a principled way. Technically, SPORE is a latent class probabilistic generative model designed to mimic users' decision-making process for choosing spatial items. We model personal interests, temporal influence, and sequential influence based on the latent variable *topic* in SPORE. A topic z corresponds to a category or a theme (i.e., a soft cluster of words describing a category or a theme of spatial items) and a geographical region (i.e., a soft cluster of locations of spatial items) at the same time. The generative process of users' check-in behaviors in SPORE is briefly illustrated as follows. Given a target user u at time t, SPORE first chooses a topic z for u based on her personal interests, the time t and her visited items before t. The selected topic z in turn generates a spatial item v following z's semantic and geographical distributions.

By introducing the latent factor topic, SPORE effectively overcomes the challenge posed by *low-sampling rate*. Specifically, SPORE addresses the sparsity of geo-social network data by considering the hidden variable topic, which groups spatial items with similar semantic meanings and geographic locations, rather than focusing on the fine granularity of data such as consecutive points in GPS trajectories.

Our proposed SPORE is able to reduce *the prediction space* effectively. In particular, for each spatial item v, we learn a vector $\boldsymbol{\theta}_v^{seq}$ of latent topics where each component $\theta_{v,z}^{seq}$ represents the likelihood of visiting the topic z after visiting v. An obvious advantage of predicting the topic of a user's activity at the next step is a significantly reduced prediction space, because the number of topics is much smaller than the number of spatial items. Additionally, to capture the influence from high order items, SPORE adds the influence of the previously visited items in the exponential space to avoid the inference of mixture weights for each visited item, inspired by the Sparse Additive Generative model (SAGE) (Eisenstein et al. 2011). In this way, SPORE accurately captures the influence from more items previously visited by the target user, and at the same time reduces the exponential complexity V^{n+1} of the classic nth order Markov Chain into linear complexity $V \times K$ (K is the number of topics).

The generative process of the SPORE model for a check-in activity (u, t, v, W_v) in the user profile D_u is as follows, and the graphical representation of SPORE model is presented in Fig. 5.

- Draw a topic index z

$$z \sim P\left(z \mid \boldsymbol{\theta}_t^{time} + \boldsymbol{\theta}_u^{user} + \boldsymbol{\theta}_{u,t}^{seq}\right)$$

- For each content word w in W_v, draw

$$w \sim P\left(w \mid \phi^0 + \phi_z^{topic}\right)$$

- Draw a spatial item v

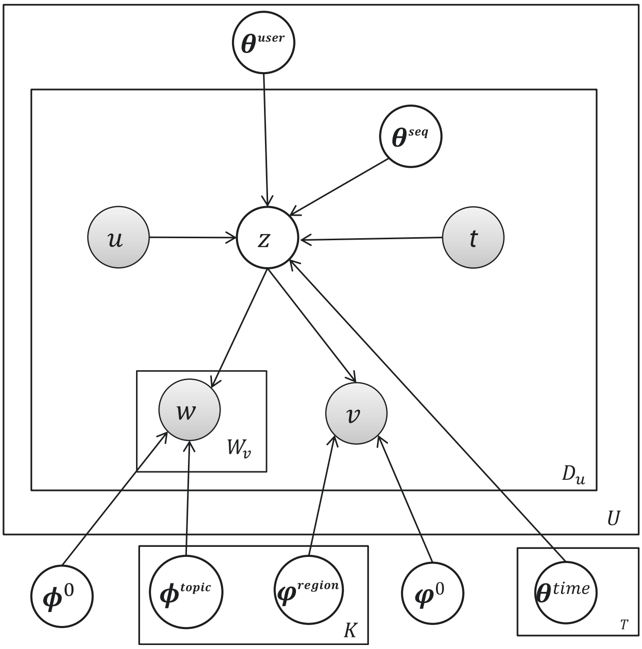
$$v \sim P\left(v \mid \varphi^0 + \varphi_z^{topic}\right)$$

where $\boldsymbol{\theta}_{u,t}^{seq} = \sum_{v' \in \mathcal{S}_{u,t}} \boldsymbol{\theta}_{v'}^{seq}$; $\mathcal{S}_{u,t}$ denotes the collection of spatial items recently visited by u before t, and $\boldsymbol{\theta}_t^{time}$ denotes the temporal preferences of the general public at time t. $P\left(z \mid \boldsymbol{\theta}_t^{time} + \boldsymbol{\theta}_u^{user} + \boldsymbol{\theta}_{u,t}^{seq}\right)$ is computed as in Eq. (10); $P\left(v \mid \varphi^0 + \varphi_z^{topic}\right)$ and $P\left(w \mid \phi^0 + \phi_z^{topic}\right)$ are computed as in the Geo-SAGE model, referring to Eqs. 4 and 5

$$P\left(z \mid \boldsymbol{\theta}_t^{time} + \boldsymbol{\theta}_u^{user} + \boldsymbol{\theta}_{u,t}^{seq}\right) = \frac{exp\left(\theta_{u,z}^{user} + \theta_{t,z}^{time} + \theta_{u,t,z}^{seq}\right)}{\sum_{z'} exp\left(\theta_{u,z'}^{user} + \theta_{t,z'}^{time} + \theta_{u,t,z'}^{seq}\right)} \tag{10}$$

We employ a mixture of EM and a Monte Carlo sampler, called the Gibbs EM algorithm, to infer all the model parameters Θ. Once the model parameters Θ are learnt from the training dataset of user check-in records, given a querying user u_q at time t_q, the probability of u_q visiting spatial item v in the next is computed as follows:

$$\begin{aligned} P&\left(v, l_v, \mathcal{W}_v \mid u_q, t_q, \mathcal{S}_{u_q,t_q}, \Theta\right) \\ &= \sum_z P\left(z \mid \boldsymbol{\theta}_{u_q}^{user} + \boldsymbol{\theta}_{t_q}^{time} + \boldsymbol{\theta}_{u_q,t_q}^{seq}\right) P\left(v \mid \varphi_z^0 + \varphi_z^{topic}\right) \\ &\times P\left(\mathcal{W}_v \mid \phi_z^0 + \phi_z^{topic}\right) \end{aligned} \tag{11}$$



Spatiotemporal Recommendation in Geo-Social Networks, Fig. 5 Graphical representation of SPORE model

Real-Time Recommendation

As the time goes on, users' interests may change and need different spatial items. This requires producing recommendation results in a real-time manner. However, our proposed Geo-SAGE model (actually almost all existing spatiotemporal recommender models) is incapable of supporting real-time recommendation due to the following reasons. First, although it exploits the temporal cyclic effect of the users' check-in behaviors to improve the topic discovery, it assumes that the individuals' interests are stable and ignores their dynamic in the long term. In reality, they are changing over time, as analyzed in (Yin et al. 2015b). For instance, users will naturally be interested in visiting parenting-related spatial items (e.g., the playground and amusement park) after they have a baby, and probably ignore their other interests. For another example, when people move from a city to another city where there are different urban compositions and cultures, their interests will be most likely to change. Accurately capturing this change in a real-time manner has been proved to be very commercially valuable since it indicates visiting and purchasing intents. Second, it is difficult to apply the current Geo-SAGE model for large-scale check-in data which arrives in a stream, since the batch learning algorithm developed for the Geo-SAGE in (Wang et al. 2015) needs to run through all check-in data for many times, and it is very time-consuming and infeasible. The batch learning algorithms usually suffer from the following two drawbacks when dealing with the situation mentioned above: (1) delay on model updates caused by the expensive time cost of rerunning the batch model; and (2) disability to track changing user interests and spatial mobility patterns due to the fact that latest check-in records used for updating recommendation models are often overwhelmed by the large data of the past. Therefore, to support real-time spatial item recommendation, we need to extend the batch Geo-SAGE model (Wang et al. 2015) to an online learning model Geo-SAGE-Online, which can efficiently process the check-in stream and track changing user interests. Actually, we can borrow the online learning techniques developed in (Yin et al. 2016a) to extend the Geo-SAGE model.

The real-time recommendation requires that the recommendation results should be time-aware (Yuan et al. 2013), location-based (Ference et al. 2013), and personalized, i.e., to recommend different ranked lists of spatial items for the same target user at different time and locations. Given a querying user u_q with the current location l_q and time t_q (i.e., $q = (u_q, l_q, t_q)$), the naive approach to produce online top-k recommendations is to first compute a ranking score for each spatial item and then select k ones with highest ranking scores. However, when the number of available spatial items becomes large, to produce a top-k ranked list using this brute-force method is very time-consuming and slow. To support real-time spatial recommendation, we developed efficient retrieval algorithms to speed up the process of online recommendation. Specifically, we proposed threshold-based algorithm TA in (Yin et al. 2013, 2014b), locality-sensitive hashing technique (LSH) in (Wang et al. 2016), and a clustering-based branch and bound algorithm (CBB) in (Yin et al. 2016a) to prune the item search space and facilitate fast retrieval of top-k recommendation.

Key Applications

The 2014 study by Econsultancy shows that 94% of companies agree that "personalized recommendation is critical to current and future success." For example, Amazon and Netflix recommendations are cited by many company observers as an outstanding means to improve revenues and profits. Especially in the era of mobile Internet, mobile recommender system usage will continue to grow faster than ever due to the exponential growth of personal activity and behavior data aptly called "THE BIG DATA." The resulting techniques presented in this entry can considerably improve the ability of business to better exploit the power of geo-social network data and provide higher quality mobile recommendation services, fostering the development of online to offline business. This exciting opportunity has emerged due to the prevalence of location-aware mobile devices and explosive growth of geo-social networks, and mobile recommendation will play a dominant role in the emerging O2O commerce. Smarter utilization of this data has far-reaching implications for other numerous business application areas such as personalized trip planning and tourism service, restaurant recommendation, mobile advertising, and marketing strategies. In addition, this entry has the potential to enable researchers from sociologists, psychologists, and anthropologists to study and analyze human mobility, interests, social structure, and behavior at a larger scale.

Future Directions

Spatial item recommendation is currently an active research area in spatial database, data mining, information retrieval and machine learning. Although a lot of progress has been made in the past, challenges remain.

- *Multiview User Modeling*. All the geo-social networks share the same geographical layers, and a user may have multiple accounts in different platforms. However, all existing spatiotemporal recommender models are designed for a single geo-social network, and the user's generated data in one single geo-social network only partial view of this user. To overcome the issue of data sparsity in a single social network and capture a comprehensive view of users, it will be a promising research direction to develop a multiview user model to learn a unified user representation from multiple geo-social social networks. For example, a user has checked-in at many infant-related point-of-interests at location-based social networks (e.g., foursquare.com) and attended many parenting-related lectures on event-based social networks (e.g., meetup.com), which suggests that this user may be a mom for a new-born baby. With these abundant available user online data in various geo-social networks, spatial item recommendation can be generated more efficiently and effectively.
- *Item-Partner Recommendation*. Different from traditional items, the spatial items are generally associated with social activities, e.g., a restaurant is associated with "having dinner" and a cinema can be mapped to "watching movies." Traditional spatiotemporal recommenders estimate the preference degree of a target user to candidate spatial items and focuses on finding top-*k* items to the user. However, these systems ignore an important characteristic of spatial items: people like to visit a POI or participate in a social event with their folks. This means that if a system recommends a spatial item alone, the user may reject the item, if she cannot think of a partner to attend the activity together. Therefore, another interesting topic is to recommend spatial items and partners together in a unified process.
- *Learning Representation for Spatial Items*. Most of the current research efforts on spatial item recommendation are devoted to the high-level recommender model development, while representation learning, one of the fundamental problems in machine learning, is largely ignored. The geo-social network data centered around spatial items exhibit diverse, heterogeneous, and collective characteristics. It is very common that multiple forms of data, e.g., text, image, geographical location, and time exist simultaneously on the same geo-social network platforms. One important promising research line is to learn the representation of spatial items from the geo-social network data based on the popular deep architecture. Based on the learnt unified feature representation of spatial items, spatiotemporal recommendation could be achieved with the off-the-shelf recommendation algorithms such as item-based collaborative filtering. It has been proved that the performance of machine learning algorithms heavily depends on the choice of data representation on which they are applied.

Cross-References

- ▶ Data Mining
- ▶ Location-Based Social Networks
- ▶ Machine Learning
- ▶ Mobile- and Context-Aware Applications of Social Networks
- ▶ Modeling and Analysis of Spatiotemporal Social Networks
- ▶ Probabilistic Graphical Models
- ▶ Recommender Systems: Models and Techniques
- ▶ Recommender Systems Using Social Network Analysis: Challenges and Future Trends
- ▶ Spatiotemporal Personalized Recommendation of Social Media Content
- ▶ Topic Model

References

Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng 17(6):734–749

Bao J, Zheng Y, Mokbel MF (2012) Location-based and preference-aware recommendation using sparse geo-social networking data. In: SIGSPATIAL, Redondo Beach, pp 199–208

Cheng C, Yang H, King I, Lyu MR (2012) Fused matrix factorization with geographical and social influence in location-based social networks. In: AAAI, Toronto, pp 17–23

Cheng C, Yang H, Lyu MR, King I (2013a) Where you like to go next: successive point-of-interest recommendation. In: IJCAI, Beijing, pp 2605–2611

Cheng H, Ye J, Zhu Z (2013b) What's your next move: user activity prediction in location-based social networks. In: SDM, SIAM, Austin, pp 171–179

Cho E, Myers SA, Leskovec J (2011) Friendship and mobility: user movement in location-based social networks. In: KDD, San Diego, pp 1082–1090

Dhar S, Varshney U (2011) Challenges and business models for mobile location-based services and advertising. Commun ACM 54(5):121–128

Eisenstein J, Ahmed A, Xing EP (2011) Sparse additive generative models of text. In: ICML, Washington, pp 1041–1048

Feng S, Li X, Zeng Y, Cong G, Chee YM, Yuan Q (2015) Personalized ranking metric embedding for next new poi recommendation. In: IJCAI, Buenos Aires, pp 2069–2075

Ference G, Ye M, Lee WC (2013) Location recommendation for out-of-town users in location-based social networks. In: CIKM, San Francisco, pp 721–726

Gao H, Liu H (2014) Data analysis on location-based social networks. In: Mobile social networking. Springer, New York, pp 165–194

Gao H, Tang J, Hu X, Liu H (2013) Exploring temporal effects for location recommendation on location-based social networks. In: RecSys, Hong Kong, pp 93–100

Gao H, Tang J, Hu X, Liu H (2015) Content-aware point of interest recommendation on location-based social networks. In: AAAI, Austin, pp 1721–1727

Hong L, Ahmed A, Gurumurthy S, Smola AJ, Tsioutsiouliklis K (2012) Discovering geographical topics in the twitter stream. In: WWW, Lyon, pp 769–778

Hu B, Ester M (2013) Spatial topic modeling in online social media for location recommendation. In: RecSys, Hong Kong, pp 25–32

Levandoski JJ, Sarwat M, Eldawy A, Mokbel MF (2012) Lars: a location-aware recommender system. In: ICDE, Washington, DC, pp 450–461

Li X, Cong G, Li XL, Pham TAN, Krishnaswamy S (2015) Rank-geofm: a ranking based geographical factorization method for point of interest recommendation. In: SIGIR, Santiago, pp 433–442

Lian D, Zhao C, Xie X, Sun G, Chen E, Rui Y (2014) Geomf: joint geographical modeling and matrix factorization for point-of-interest recommendation. In: KDD, New York, pp 831–840

Lichman M, Smyth P (2014) Modeling human location data with mixtures of kernel densities. In: KDD, New York, pp 35–44

Linden G, Smith B, York J (2003) Amazon.Com recommendations: item-to-item collaborative filtering. IEEE Internet Comput 7(1):76–80

Liu B, Xiong H (2013) Point-of-interest recommendation in location based social networks with topic and location awareness. In: SDM, Austin, pp 396–404

Liu J, Dolan P, Pedersen ER (2010) Personalized news recommendation based on click behavior. In: IUI, Hong Kong, pp 31–40

Liu X, He Q, Tian Y, Lee WC, McPherson J, Han J (2012) Event-based social networks: linking the online and offline social worlds. In: KDD, Beijing, pp 1032–1040

Liu B, Fu Y, Yao Z, Xiong H (2013) Learning geographical preferences for point-of-interest recommendation. In: KDD, Chicago, pp 1043–1051

Noulas A, Scellato S, Mascolo C, Pontil M (2011) An empirical study of geographic user activity patterns in foursquare. In: ICWSM, Barcelona, pp 70–573

Sarwar B, Karypis G, Konstan J, Riedl J (2001) Item-based collaborative filtering recommendation algorithms. In: WWW, Hong Kong, pp 285–295

Scellato S, Noulas A, Lambiotte R, Mascolo C (2011) Socio-spatial properties of online location-based social networks. In: ICWSM, Barcelona, 329–336

Song L, Kotz D, Jain R, He Xiaoing (2004) Evaluating location predictors with extensive wi-fi mobility data. In: INFOCOM, Hong Kong, pp 1414–1424

Stoyanovich J, Amer Yahia S, Marlow C, Yu C (2008) Leveraging tagging to model user interests in del.icio.us. In: AAAI, Chicago, pp 104–109

Wang W, Yin H, Chen L, Sun Y, Sadiq S, Zhou X (2015) Geo-sage: a geographical sparse additive generative model for spatial item recommendation. In: KDD, Sydney, pp 1255–1264

Wang W, Yin H, Sadiq S, Chen L, Xie M, Zhou X (2016) Spore: a sequential personalized spatial item recommender system. In: ICDE, Helsinki, pp 954–965

Xu Z, Zhang Y, Wu Y, Yang Q (2012) Modeling user posting behavior on social media. In: SIGIR, Portland, pp 545–554

Ye M, Yin P, Lee WC (2010) Location recommendation for location-based social networks. In: GIS, San Jose, pp 458–461

Ye M, Janowicz K, Mülligann C, Lee WC (2011a) What you are is when you are: the temporal dimension of feature types in location-based social networks. In: SIGSPATIAL GIS, Chicago, pp 102–111

Ye M, Shou D, Lee WC, Yin P, Janowicz K (2011b) On the semantic annotation of places in location-based social networks. In: KDD, San Diego, pp 520–528

Ye M, Yin P, Lee WC, Lee DL (2011c) Exploiting geographical influence for collaborative point-of-interest recommendation. In: SIGIR, Beijing, pp 325–334

Yin H, Sun Y, Cui B, Hu Z, Chen L (2013) Lcars: a location-content-aware recommender system. In: KDD, Chicago, pp 221–229

Yin H, Cui B, Chen L, Hu Z, Huang Z (2014a) A temporal context-aware model for user behavior modeling in social media systems. In: SIGMOD, Snowbird, pp 1543–1554

Yin H, Cui B, Sun Y, Hu Z, Chen L (2014b) Lcars: a spatial item recommender system. ACM Trans Inf Syst 32 (3):11:1–11:37

Yin H, Cui B, Chen L, Hu Z, Zhang C (2015a) Modeling location-based user rating profiles for personalized recommendation. ACM Trans Knowl Discov Data 9 (3):19:1–19:41

Yin H, Cui B, Chen L, Hu Z, Zhou X (2015b) Dynamic user modeling in social media systems. ACM Trans Inf Syst 33(3):10:1–10:44

Yin H, Cui B, Huang Z, Wang W, Wu X, Zhou X (2015c) Joint modeling of users' interests and mobility patterns for point-of-interest recommendation. In: ACM Multimedia, Brisbane, pp 819–822

Yin H, Zhou X, Shao Y, Wang H, Sadiq S (2015d) Joint modeling of user check-in behaviors for point-of-interest recommendation. In: CIKM, Melbourne, pp 1631–1640

Yin H, Cui B, Zhou X, Wang W, Huang Z, Sadiq S (2016a) Joint modeling of user check-in behaviors for real-time point-of-interest recommendation. ACM Trans Inf Syst 35(2):1–44

Yin H, Zhou X, Cui B, Wang H, Zheng K, Nguyen QVH (2016b) Adapting to user interest drift for poi recommendation. IEEE Trans Knowl Data Eng 28 (10):2566–2581

Yuan Q, Cong G, Ma Z, Sun A, Thalmann NM (2013) Time-aware point-of-interest recommendation. In: SIGIR, Dublin, pp 363–372

Zhang JD, Chow CY (2015) Spatiotemporal sequential influence modeling for location recommendations: a gravity-based approach. ACM Trans Intell Syst Technol 7(1):11:1–11:25

Zhang JD, Chow CY, Li Y (2014) Lore: Exploiting sequential influence for location recommendations. In: SIGSPATIAL, Dallas, pp 103–112

Zhao K, Cong G, Yuan Q, Zhu KQ (2015) Sar: a sentiment-aspect-region model for user preference analysis in geo-tagged reviews. In: ICDE, Seoul, pp 675–686

Zheng Y, Xie X (2011) Learning travel recommendations from user-generated gps traces. ACM Trans Intell Syst Technol 2(1):2:1–2:29

Zheng Y, Zhang L, Xie X, Ma WY (2009) Mining interesting locations and travel sequences from gps trajectories. In: WWW, Madrid, pp 791–800

Zheng VW, Zheng Y, Xie X, Yang Q (2010) Collaborative location and activity recommendations with gps history data. In: WWW, Raleigh, pp 1029–1038

Zheng YT, Zha ZJ, Chua TS (2012) Mining travel patterns from geotagged photos. ACM Trans Intell Syst Technol 3(3):56:1–56:18

Recommended Reading

Yin H, Cui B (2016) Spatio-temporal recommendation in social media. Springer, ISBN 978–981–10-0747-7, pp. 1–114

Spatiotemporal Topic Detection from Social Media

Kaiqi Zhao[1], Quan Yuan[2] and Gao Cong[1]
[1]School of Computer Engineering, Nanyang Technological University, Singapore, Singapore
[2]Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, IL, USA

Synonyms

Generative model; Social networks; Spatiotemporal data; Topic model

Glossary

Geotagged post	User-generated documents associated with geographical coordinates or points of interest. Typical geotagged posts include Foursquare checkins, geotagged tweets, and geotagged reviews
Check-in	The behavior that a user posted a geotagged post at some place at some time
Geographical coordinates	A pair of latitude and longitude recorded by GPS devices that indicates a geolocation
Point of interest (POI)	Functional venue in digital maps, such as restaurants, hotels, shops, etc. Each POI has geographical coordinates, categories, rating, etc
Topic	A latent variable that modeled by distributions over words, locations, or time
Latent region	A geographical cluster of locations. It is often modeled as a two-dimensional Gaussian distribution over coordinates
Latent time component	A cluster of timestamp. It is often modeled as a one-dimensional Gaussian distribution over time

Definition

With the widely adoption of GPS devices among social media users in recent years, a large volume of geotagged social media posts have been generated, each of which is associated with a location (e.g., a pair of geographical coordinates or a point of interest). The contents of these geotagged posts could be related to the creation time and the location where the post was curated. Spatiotemporal topic detection, which aims to find out relations among topics, latent regions, and time components, has attracted much attention recently. Because it can not only discover the spatiotemporal dynamics of the topics, but also facilitate many applications such as point-of-interest recommendation, travelogue recommendation, event detection, geotagging, etc.

Detecting spatiotemporal topics is difficult because neither regions nor time components are known beforehand. The short and low-quality text contents of social media posts make it even harder. This chapter reviews methods that tackle problems in mining spatiotemporal topics.

Introduction

Popular social media platforms, e.g., Foursquare, Facebook, and Twitter, support users to share contents at anytime and anywhere. Many of these social media posts contain the creation time and location information collected by GPS-equipped mobile devices. Spatiotemporal topic detection becomes a popular research topic in social media analysis and it has been proven to be beneficial to many social media applications, such as point-of-interest recommendation, event detection, geotagging, travelogue recommendation, etc.

Considering that time and locations are continuous variables, there exist two ways to detect spatiotemporal topics. One way is to divide the spatiotemporal space into discrete cells and then train a topic model for each cell. However, this method cannot capture the relations between adjacent cells, and it is usually difficult to determine the granularity of the cells. Large cells may lose local spatiotemporal dynamics in the cells, while small cells may suffer from the data sparsity problem. As a result, most of the existing work on spatiotemporal topic detection takes the other way, i.e., they assume time and locations are generated from some high-level latent spatiotemporal variables, namely, latent regions and time components. Because the latent regions and time components are learned automatically from the data, they can capture the correlations between adjacent timestamps and locations, and will not suffer from the granularity problem.

Since topics, regions, and time components are often latent, mining spatiotemporal topics faces two main challenges. First, the dependency relationships among topic, region, and time component are not clear. For example, how do topics distribute in different regions along time? Is there any underlying correlations among them? Second, the variables in a spatiotemporal topic model could be in different forms. For example, topics are often modeled as discrete distributions while regions are often Gaussians. How to combine these different types of variables in a unified model is a difficult problem.

To tackle the first challenge, latent regions and time components are modeled as Gaussian distributions in some research (Yuan et al. 2015b; Guo and Gong 2016). They use Bayesian models to connect all the latent variables based on intuitions observed from the data. For the second challenge, Kurashima et al. (2013) model the probability of a user visiting a POI as the multiplication of the POI distributions of both topics and regions. Yuan et al. (2013) propose to use linear combination instead of multiplication to avoid the POI distribution of region overwhelming that of topic. They also propose to discretize the Gaussian distribution to make the two POI distributions (of topics and regions) comparable. This chapter reviews recent spatiotemporal topic models to illustrate the key ideas of spatiotemporal topic modeling.

Key Points

The following two key points usually make differences among spatiotemporal models:

- How to model the spatial and temporal information?
- How do location, time, and topic interact with each other?

In this chapter, we first review different factors (location, time, topic, etc.) considered in the literatures. To understand how spatiotemporal topic models work, we introduce methods to model spatial and temporal patterns from the social media posts, and then illustrate how different factors interact with each other using several example models. We also present several applications of spatiotemporal topic modeling.

Historical Background

Spatial topic modeling has been a hot research topic recently. Hao et al. (2010) assume the topics of a travelogue come either from the topic distribution of point of interest or from some global topic distribution. Kurashima et al. (2010) model the topics on clusters of landmarks grouped by spatial proximity rather than individual POIs. In the consequent work (Kurashima et al. 2013), they propose a model for point-of-interest (POI) recommendation, which exploits user's historical check-in location in a topic model. Sizov (2010, 2012) proposes the *GeoFolk* model, which introduces latent regions in spatial topic modeling. In GeoFolk, each latent region is defined by both spatial and topical aspects. More precisely, it consists of two Gaussian distributions on latitude and longitude, respectively, and a multinomial distribution over words. Yin et al. (2011) split the latent region and topic as two latent variables and assume each region has a topic distribution. This model improves the GeoFolk model in two ways. First, it captures the topic distributions of regions while GeoFolk considers each region as one topic. Second, the topic distribution over geographical space is not restricted to Gaussian in this model, while GeoFolk assumes Gaussian distribution for each topic. However, in Yin's model, the number of regions needs to be specified. Ahmed et al. (2013) propose a nonparametric spatial topic model which can automatically fit the number of regions and other parameters to maximize the likelihood of the training data.

Early studies that consider the time factor in spatial topic modeling divide the time dimension into discrete intervals. Hu et al. (2013) introduce a topic distribution for each time interval to capture the temporal patterns of topics, while Vosecky et al. (2013) model a time distribution for each topic. Recently, spatiotemporal topic models are designed to be able to work on continuous time dimension. Similar to the modeling of spatial information, a mixture of Gaussian distributions is assumed on the time dimension to capture users' temporal check-in patterns. Jiang and Ng (2013) consider three temporal patterns of each topic, i.e., global, bursty, and periodic patterns, and use uniform, single Gaussian distribution and a mixture of Gaussian distributions to model these three patterns, respectively. Yuan et al. (2013) propose a W4 model (Who, Where, When, and What) to mine user behaviors from user check-in data. W4 models the temporal mobility patterns of individual users in latent regions and the content of a geotagged post is considered to be generated based on both latent regions and topics. In the follow-up work (Yuan et al. 2015b), they propose an enhanced W4 model, which uses a number of Gaussian distributions (time components) to represent the time dimension and each time component has a distribution over regions. Liu et al. (2015b) propose a downstream spatiotemporal topic model (DSTTM) and an upstream spatiotemporal topic model (USTTM). In DSTTM, a time distribution is associated with each region. In contrast, a region distribution is associated with each latent time component in USTTM. Guo and Gong (2016) propose a spatiotemporal topic model to discover events. Each event is modeled as a topic that consists of a word distribution, a region distribution, and a time component distribution.

Apart from location and time, some spatiotemporal topic models take other information such as user, social network, and sentiment into account to support more applications. For instance, Yuan et al. (2013) model a region distribution for each user to support user behavior modeling. Hu and Ester (2014) incorporate the social influence in spatial topic model to make POI recommendations. Zhao et al. (2015a) combine sentiment analysis and spatial topic modeling to analyze user's preferences on regions and the aspects of a POI.

Spatiotemporal Topic Model

Before introducing the key ideas of spatiotemporal topic models, we first formulate geotagged social media posts. Suppose we have a collection of geotagged social media posts D. Each post in D is a quadruple $d = \langle u_{d,}\ \mathbf{l}_{d,} t_{d,}\ \mathbf{w}_d \rangle$, indicating the user u_d posted the content $\mathbf{w}_d$ at the location $\mathbf{l}_d$ and timestamp $t_{d.}$ The location in a post could be geographical coordinates or a point of interest with or without coordinates.

A general spatiotemporal topic detection problem can be formulated as finding Z topics (represented by word distributions $p(w|z)$) in the data collection D, the topic distributions on latent regions $p(z|r)$, $\forall z \in [1, \boldsymbol{Z}]$, $\forall r \in [1, \boldsymbol{R}]$, and the topic distributions on time components $p(z|c)$, $\forall z \in [1, \boldsymbol{Z}]$, $\forall c \in [1, \boldsymbol{C}]$. Note that the latent regions can either be clusters of POIs (Kling et al. 2014) or probabilistic distributions over POIs (Sizov 2010). Time components can either be explicitly discretized to time intervals (Hu et al. 2013) or probabilistic distribution over timestamps (Yuan et al. 2015b).

Several modeling methods have been used to detect spatiotemporal topics, such as matrix factorization (Liu and Xiong 2013; Liu et al. 2013, 2015a) and generative models (Yuan et al. 2013; Yin et al. 2013). In this chapter, we use generative models as examples to illustrate the key ideas of spatiotemporal topic detection.

Modeling Regions

Region and topic: The simplest way to add spatial information to a topic model is to attach a bivariate Gaussian distribution to each topic/region as in GeoFolk (Sizov 2010). In this method, the detected topics are spatially isolated to each other and each region only represents one topic. However, it is often useful to reveal the relationship between region and topic. Recent models separate the two variables to support modeling the relationship between them. According to different applications, the dependencies between the two variables can be different. For example, a topic is often assumed to be picked from the topic distribution of a region in models for POI recommendations (Yuan et al. 2013), while a region is generated according to the region distribution of a topic in spatiotemporal event detection models (Guo and Gong 2016).

Gaussian vs. non-gaussian: In most work on spatiotemporal topic modeling, a region is represented by a Gaussian distribution over POIs. POIs are then generated from a Gaussian mixture model. There also exist some studies using non-gaussian methods to model a region.

Lichman and Smyth (2014) prove that kernel density estimation methods perform better than Gaussian mixture models in capturing irregular shaped, long-tailed spatial patterns through extensive experiments. Kling et al. (2014) propose to first perform spatial clustering on the social media posts and learn topics on each spatial cluster (region). However, compared to Gaussian distribution, non-gaussian methods are hard to be applied in a full Bayesian inference framework. Because kernel density functions and clusters cannot be mathematically integrable (or "collapsed") while we often need to collapse some latent variables to estimate others in full Bayesian inference of a complex model. Consider that we often have complex Bayesian networks among topic, region, time component, and even other variables (e.g., users, categories, social friends) in the spatiotemporal topic detection scenario, modeling regions with Gaussian distribution is often more flexible than non-gaussian methods. Moreover, most non-gaussian methods detect topics and regions in different steps, making them incapable of capturing topical regions such as sightseeing areas, shopping street, etc. In the remaining of this chapter, we mainly focus on Gaussian-based methods.

Semantic regions: In addition to spatial information, Yuan et al. (2013, 2015b) also explore the semantic information in a latent region. They assume the regions affect the content of the social media posts. For example, if a post was created in New York City, it may contain some region-specific words like "new," "york," "US," etc. In these models, each region has a region-specific word distribution. Incorporating the word distribution will result in more condensed regions because the spatial (Gaussian) and word distributions make the POIs in a region to be both spatially and semantically close.

Modeling Time Components

In some studies of spatiotemporal topic modeling, time is discretized to a number of intervals (Hu et al. 2013; Liu et al. 2014; Zheng et al. 2016). Each time interval has a topic distribution. This treatment simplifies the model training but does not capture the relations between time intervals. According to Yuan's studies (2013, 2015b), social media users act with some temporal patterns. For example, a user may be active in the working region during the daytime but visit some entertainment region at night. The consecutive time intervals may form a single activity pattern. Instead of modeling with discrete time intervals, modeling with latent time components (each represents a temporal activity pattern) can support topic analysis on different temporal mobility patterns. Recent studies (Yuan et al. 2013, 2015b; Liu et al. 2015b) model the latent time components with Gaussian distributions, which is similar to the modeling of regions. The Gaussian parameters of each latent time component are learnt from the geotagged posts automatically.

Spatiotemporal Joint Modeling

Generally, there are three latent variables in spatiotemporal joint models: topic, region, and time component. The dependency relationship among these three variables significantly affects the applications of the models, and thus varies from models to models. We present several example spatiotemporal joint models to illustrate different dependency relationships.

W4 and EW4. Yuan et al. (2013) propose a generative model for spatiotemporal user behavior modeling to capture user (who), location (where), time (when), and topic (what) in a joint model (W4). The W4 model is the first attempt to jointly model the four factors of social media activities. In W4, the following assumptions are made to simulate a user's behavior.

- A user is assumed to have different interests on regions at different time in weekdays and weekends.
- In each region, a user has preferences on different topics.
- From the preferences on topics and regions, a user selects a region (e.g., downtown) and a topic (e.g., restaurants) in mind, and then visits a POI that is in the selected topic and close to the selected region.
- After a user visited the POI, he/she writes a social media post with words that come from both the current region and topic.

According to these assumptions, the relation among variables can be determined. According to the first assumption, we can model a user's regional preferences $p(r|u,\ s,\ t)$ as the probability of user u visits region r at time interval t on weekday ($s = 1$) or weekend ($s = 0$). Similarly, in each region r, the user u has preferences on topics $p(z|r,\ u)$. The selection of a POI is determined by both spatial and topical preferences as follows:

$$p(l|\, z, r) = \kappa p(l|\, r) + (1 - \kappa)p(l|\, z),$$

where $p(l|r)$ follows the Gaussian distribution of region r, and $p(l|z)$ is the probability of POI l in topic z. The weight $\boldsymbol{\kappa} \in [0,\ 1]$ is used to balance the importance of the two factors. It can also be seen as the probability of selecting one of the two factors from the view of a mixture model. Note that $p(l|r)$ is a continuous distribution. To make the linear combination computable, the authors suggest to discretize the continuous Gaussian distribution to a categorical distribution according to different contour values.

Similarly, the selection of words to write in the social media posts is also modeled as a linear combination of the word distributions of region $p(w|r)$ and topic $p(w|z)$ with weight $\lambda \in [0,\ 1]$:

$$p(w|\, z, r) = \lambda p(w|\, r) + (1 - \lambda)p(w|\, z).$$

With these relations, we can apply the Bayesian chain rule to compute the likelihood of the dataset and estimate the parameters with an Expectation-Maximization (EM) algorithm. After training the model, the topics $p(w|z)$, topic distribution in region $p(z|r)$ can be directly obtained, and the topic distribution at time t can be also computed using the Bayesian chain rule $p(z|s,t) = \sum_u \sum_r p(z|r)p(r|u,\ s,\ t)p(u)$.

The enhanced W4 model (EW4) (Yuan et al. 2015b) introduces latent time components to W4. A user is assumed to have different region preferences in weekdays and weekends, and each region is assumed to have different time patterns (latent time components) in weekdays and weekends. For example, a user may visit the working region during the daytime in weekdays while stay in the home region the whole day in weekends. Moreover, EW4 is a nonparametric model which automatically learns the numbers of regions and topics from data instead of requiring experts to provide the numbers as in W4.

Both W4 and EW4 are successful in mining user's topical interests in regions, and user's regional interests at different time intervals to provide good prediction capability for many applications. For example, modeling user's regional interests benefits the prediction of user's location at some timestamp because the model understands the user's spatial activity patterns (in regions) at different time. For another example, modeling user's topic interests in regions help predict the user's activity in some given region.

DSTTM and USTTM. Liu et al. (2015b) propose two joint models, namely downstream spatiotemporal topic model (DSTTM) and upstream spatiotemporal topic model (USTTM), for spatiotemporal data. The main difference between the two models lies in the modeling of time. In DSTTM, time is supposed to be determined by both topic and region, because it is assumed that different combinations of topics and regions could have different time patterns. As an example, the same topic "food" in downtown would have different check-in time in a nearby place. As another example, the topic "lunch" and "beer" in the same region would have different time patterns.

In contrast, USTTM assumes that each user may have different topic and region interests, and thus the time variable is used to determine topics and regions. A Gaussian mixture model is learnt to generalize the timestamp to latent time components. Each user in each time component has a topic distribution and a region distribution. A user selects a POI to visit according to the region and topic chosen based on those distributions.

Different from W4 and EW4, DSTTM and USTTM treat topic and region independently. In other words, there is no topic distribution for latent regions. This is because DSTTM and USTTM are not designed for exploring the user's topical preferences in different regions as in W4 and EW4. DSTTM and USTTM are built in order to provide a good POI recommender and a good location predictor, while W4 and EW4 aim to support more applications. In other words, these spatiotemporal topic models consider different relationships among variables to cater the requirements of different applications.

Key Applications

Spatiotemporal topic models are widely used in many social media applications.

POI recommendation: POI recommendation is a key application of many spatiotemporal topic models. Some studies only consider spatial information, and recommend POIs to a user by exploring the user's topical preferences under different spatial conditions. Some other studies (Liu et al. 2013; Liu and Xiong 2013) apply LDA to extract topics for both users and POIs from textual information (e.g., tags, categories of POIs), and incorporate the topics and spatial information in matrix factorization model to learn latent spatial topic vectors for both users and POIs. They sort POIs based on the dot products of the two vectors and return the top-ranked ones to users. The authors further develop a more general matrix factorization framework for incorporating other information in their consequent work (Liu et al. 2015a). Hu and Ester (2013) recommend a POI to a user when the POI both spatially and topically satisfies the user's preferences. Zheng et al. (2016) define a spatiotemporal distance–based weighting scheme in matrix factorization to incorporate the spatiotemporal influence. Yin et al. (2013, 2014) assume a user may visit a POI based on either the user's topic preferences or the local topic interests. Local interests are modeled as the overall topic distribution over some fine-grained predefined regions, e.g., cities.

Considering that time is also an important factor in POI recommendations, recent studies explore both spatial and temporal information. Hu et al. (2013) mine user's topical preferences over POIs for different time intervals. Yuan et al. (2013, 2015b) discover both user's preferences over latent regions in different time and the user's topical preferences in different regions. The ranking score of a POI to a user is proportional to the agreement between the POI and user on topical and regional preferences. Liu et al. (2015b) assume each user may have different topic and region interests at different time, and the recommendation is made based on the user's topic and region interests at current time.

Apart from spatiotemporal information, other factors are also considered for recommendation. Zhao et al. (2015a) learn the user's preferences on latent region, categories, and aspects (e.g., environment) of a POI, and combine spatial information with user sentiments to make POI recommendations. Hu and Ester (2014) consider the social relation between users. POIs of topics that are commonly preferred by the user's friends are recommended.

Event detection: Zhao et al. (2015b) divide the map into grids and learn local topics for each grid in each time interval. Based on the topics and spatiotemporal burstiness, they classify the tweet sequences in the current time interval to be event-related or not. Guo and Gong (2016) develop a nonparametric method for local event discovery on a static dataset. A local event is modeled as a topic that contains a sparse distribution on both latent regions and time components.

Travel planning: Spatiotemporal topic models are also applied to help tourists in travel planning (Hao et al. 2010; Kurashima et al. 2010). Hao et al. (2010) distinguish local topics (for each POI) and global topics in a spatial topic model trained on travelogues to support tourists by providing POI recommendation, highlighting informative travelogues, etc. Kurashima et al. (2010) incorporate user's topical preferences on locations to hidden markov model for generating travel routes. Liu et al. (2011, 2014) propose to recommending travel packages (each contains a set of landmarks in one or more regions) by considering topics in different regions and seasons.

Geotagging: Hong et al. (2012) consider user's preferences on latent regions and topics to predict the coordinates of tweets that have no geotags. Zhao et al. (2016) design a supervised model to annotate POIs with relevant tweets by combining the spatial topic model and a regression model.

There also exist other applications. Yuan et al. (2012, 2015a) discover functional regions each of which has different topics. Spatiotemporal topics of queries are mined to support location-time-aware web content search (Jiang and Ng 2013; Jiang et al. 2015). Kim et al. (2015) learn topical trajectory patterns for geotagged data. Yu et al. (2016) segment the map to topical regions and predict the count of different types of spatial events (e.g., open an app, or click on a news article) in the regions.

Future Directions

Spatiotemporal topic detection is a hot research topic that benefits many social media applications. Many models are designed under different assumptions on data and applications. Although many contributions are made, challenges still exist.

- **Big data and scalability:** As the number of latent variables in a spatiotemporal topic model increases, the computation complexity of the training algorithm increases dramatically. For example, the complexity of training the W4 model (Yuan et al. 2013) is proportional to the multiplication of (1) the number of topics, (2) the number of regions, and (3) the number of word tokens in the social media posts. It could be more time consuming when the topic distribution of time components is included as it is in (Guo and Gong 2016). As the scale of social media becomes large, training these models is extremely expensive. The scalability problem on big data raises the following two future directions: (1) Is there any faster method to reduce the training complexity? (2) How to train spatiotemporal topic models in parallel or in distributed environments?
- **Streaming data:** Most of the spatiotemporal topic models are trained on static data and make predictions on a test dataset. Whenever new social media posts are submitted, the whole model needs to be retrained to update the topics. Can we develop efficient incremental training models for spatiotemporal topic detection?
- **Modeling topic dynamics:** Topics are often static in most of the spatiotemporal topic models. Is it possible to learn how topics evolve over time and how topics spread over spatial areas along time? Modeling the spatiotemporal dynamics can benefit event monitoring and prediction, e.g., predicting how some disease will spread out in the next several hours.

Cross-References

▶ Location-Based Social Networks
▶ Modeling and Analysis of Spatiotemporal Social Networks
▶ Probabilistic Graphical Models
▶ Spatiotemporal Recommendation in Geo-Social Networks

References

Ahmed A, Hong L, Smola AJ (2013) Hierarchical geographical modeling of user locations from social media posts. In: WWW, pp 25–36

Guo J, Gong Z (2016) A nonparametric model for event discovery in the geospatialtemporal space. In: CIKM, pp 499–508

Hao Q, Cai R, Wang C, Xiao R, Yang JM, Pang Y, Zhang L (2010) Equip tourists with knowledge mined from travelogues. In: WWW, pp 1–10

Hong L, Ahmed A, Gurumurthy S, Smola AJ, Tsioutsiouliklis K (2012) Discovering geographical topics in the twitter stream. In: WWW, p 769

Hu B, Ester M (2013) Spatial topic modeling in online social media for location recommendation. In: RecSys, pp 25–32

Hu B, Ester M (2014) Social topic modeling for point-of-interest recommendation in location-based social networks. In: ICDM, pp 845–850

Hu B, Jamali M, Ester M (2013) Spatio-temporal topic modeling in mobile social media for location recommendation. In: ICDM, IEEE, pp 1073–1078

Jiang D, Ng W (2013) Mining web search topics with diverse spatiotemporal patterns. In: SIGIR, pp 881–884

Jiang D, Vosecky J, Leung KW, Yang L, Ng W (2015) SG-WSTD: a framework for scalable geographic web search topic discovery. Knowl-Based Syst 84:18–33

Kim Y, Han J, Yuan C (2015) Toptrac: topical trajectory pattern mining. In: KDD, ACM, pp 587–596

Kling CC, Kunegis J, Sizov S, Staab S (2014) Detecting non-gaussian geographical topics in tagged photo collections. In: WSDM, pp 603–612

Kurashima T, Iwata T, Irie G, Fujimura K (2010) Travel route recommendation using geotags in photo sharing sites. In: CIKM, pp 579–588

Kurashima T, Iwata T, Hoshide T, Takaya N, Fujimura K (2013) Geo topic model: joint modeling of user's activity area and interests for location recommendation. In: WSDM, ACM, pp 375–384

Lichman M, Smyth P (2014) Modeling human location data with mixtures of kernel densities. In: KDD, pp 35–44

Liu B, Xiong H (2013) Point-of-interest recommendation in location based social networks with topic and location awareness. In: SDM, SIAM, pp 396–404

Liu Q, Ge Y, Li Z, Chen E, Xiong H (2011) Personalized travel package recommendation. In: ICDM, pp 407–416

Liu B, Fu Y, Yao Z, Xiong H (2013) Learning geographical preferences for point of-interest recommendation. In: KDD, pp 1043–1051

Liu Q, Chen E, Xiong H, Ge Y, Li Z, Wu X (2014) A cocktail approach for travel package recommendation. TKDE 26(2):278–293

Liu B, Xiong H, Papadimitriou S, Fu Y, Yao Z (2015a) A general geographical probabilistic factor model for point of interest recommendation. TKDE 27(5):1167–1179

Liu Y, Ester M, Hu B, Cheung DW (2015b) Spatio-temporal topic models for checkin data. In: ICDM, pp 889–894

Sizov S (2010) Geofolk: Latent spatial semantics in web 2.0 social media. In: WSDM, pp 281–290

Sizov S (2012) Latent geospatial semantics of social media. TIST 3(4):1–20

Vosecky J, Jiang D, Leung KWT, Ng W (2013) Dynamic multi-faceted topic discovery in twitter. In: CIKM, pp 879–884

Yin Z, Cao L, Han J, Zhai C, Huang T (2011) Geographical topic discovery and comparison. In: WWW, pp 247–256

Yin H, Sun Y, Cui B, Hu Z, Chen L (2013) Lcars: a location-content-aware recommender system. In: KDD, pp 221–229

Yin H, Cui B, Sun Y, Hu Z, Chen L (2014) LCARS: a spatial item recommender system. TOIS 32(3):11:1–11:37

Yu R, Gelfand A, Rajan S, Shahabi C, Liu Y (2016) Geographic segmentation via latent poisson factor model. In: WSDM, pp 357–366

Yuan J, Zheng Y, Xie X (2012) Discovering regions of different functions in a city using human mobility and pois. In: KDD, pp 186–194

Yuan Q, Cong G, Ma Z, Sun A, Magnenat-Thalmann N (2013) Who, where, when and what: discover spatio-temporal topics for twitter users. In: The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, 11–14 Aug 2013, pp 605–613

Yuan NJ, Zheng Y, Xie X, Wang Y, Zheng K, Xiong H (2015a) Discovering urban functional zones using latent activity trajectories. TKDE 27(3):712–725

Yuan Q, Cong G, Zhao K, Ma Z, Sun A (2015b) Who, where, when, and what: A nonparametric bayesian approach to context-aware recommendation and search for twitter users. ACM Trans Inf Syst 33(1):2:1–2:33

Zhao K, Cong G, Yuan Q, Zhu KQ (2015a) Sar: a sentiment-aspect-region model for user preference analysis in geo-tagged reviews. In: ICDE, pp 675–686

Zhao L, Chen F, Lu CT, Ramakrishnan N (2015b) Spatio-temporal event forecasting in social media. In: SDM, SIAM, vol 15, pp 963–971

Zhao K, Cong G, Sun A (2016) Annotating points of interest with geo-tagged tweets. In: CIKM, pp 417–426

Zheng C, E Haihong, Song M, Song J (2016) TGTM: temporal-geographical topic model for point-of-interest recommendation. In: DASFAA, pp 348–363

Spatiotemporal Web

► Spatiotemporal Information for the Web

Spatio-temporal-Thematic Analysis

► Twitris: A System for Collective Social Intelligence

Spectral Analysis

Xiao-Dong Zhang
School of Mathematical Science, Ministry of Education Key Laboratory of Scientific and Engineering Computing, Shanghai Jiao Tong University, Shanghai, China

Synonyms

Spectral Graph Analysis; Spectral Network Analysis; Spectral Technique

Glossary

Network (graph) A network G is a triple consisting of a node set $V(G)$, a link set $E(G)$, and a relation that associates each link with two nodes

Adjacency matrix Let $G = (V(G), E(G))$ be a network with $V(G) = \{v_1, \cdots, v_n\}$. The adjacency matrix $A(G) = (a_{ij})$ of G is $n \times n$ matrix with $a_{ij} = 1$ if v_i is adjacent to v_j, and 0 otherwise

Eigenvalues of a graph All eigenvalues of the adjacency matrix $A(G)$ of a graph G are called eigenvalues of G and denoted by $\lambda_1 \geq \lambda_2 \geq \ \ldots \ \geq \lambda_n$

Degree diagonal matrix The degree diagonal matrix $D(G)$ of a network G is the diagonal matrix whose diagonal entries are degrees of the corresponding nodes

Laplacian matrix The Laplacian matrix $L(G)$ is defined be $L(G) = D(G) - A(G)$, where $D(G)$ is the degree diagonal matrix and $A(G)$ is the adjacency matrix

Laplacian eigenvalues of a graph All eigenvalues of the Laplacian matrix $L(G)$ of a graph G are called the Laplacian eigenvalues of G and denoted by $\mu_1 \geq \mu_2 \geq \ \ldots \ \geq \mu_n = 0$

Normal matrix The normal matrix $N(G)$ is defined to as the product of the inverse of degree diagonal matrix and the adjacency matrix

Normal eigenvalues of a graph All eigenvalues of the normal matrix $N(G)$ of a graph are called the normal eigenvalues of G and denoted by $1 = \nu_1 \geq \nu_2 \geq \ \ldots \ \geq \nu_n$

Adjacency (Laplacian, normal) spectrum The set of all eigenvalues of the adjacency (Laplacian, normal) matrix

Walk A walk is list $v_0, e_1, \ldots, e_k, v_k$ of nodes and links such that, for $1 \leq i \leq k$, the link e_i has endpoints v_{i-1} and v_i. The length of a walk is its number of links

Bipartite network A network $G = (V, E)$ is called bipartite if V is decomposed into two disjoint sets such that each link has its ends in different sets

Definition

In this entry, we describe spectral analysis of networks. Generally speaking, the eigenvalues and eigenvectors of the different matrices associated with networks are intimately connected to important topological features, such as diameter, community structure, node centrality, etc.

Introduction

In the past decade, networks have attracted considerable attention in many disciplines such as statistical physics, social science, and applied mathematics. Generally speaking, the study of network representations of physical, biological, and social phenomena leading to predictive models of these phenomena is now called network science. Newman (Newman 2003) reviews some features of real-world networks and properties of several network models, including random graphs, the small-world model, models of network growth, and epidemiological processes. Boccaletti et al. (Boccaletti et al. 2006) survey the important concepts and results in the network science, in particular the topological structure and synchronization and collective dynamics of complex networks. One important class of complex networks is the class of social networks, which are social structures consisting of individuals (or groups) called nodes and their relationships, such as friendship, common interest, and financial exchange, called links. The analysis of social networks is used in epidemiology, mass surveillance, diffusion of innovations, etc. There are many measures (metrics) in social network analysis, such as betweenness, centrality, clustering coefficient, etc. The community structure, or clusters, is one of the most important features in sociology. Recently, Fortunato (Fortunato 2010) gave a thorough exposition of community deletion, from the several main definitions of the community problem to the presentation of most methods developed.

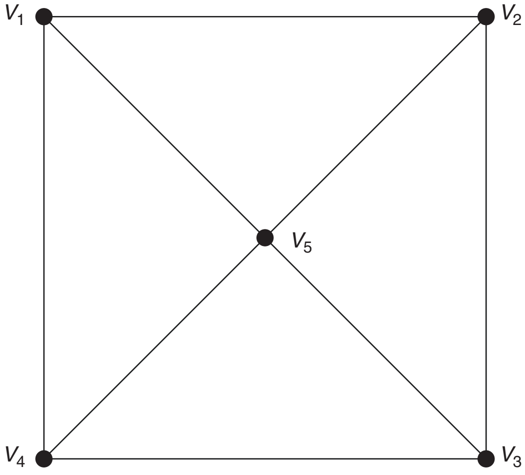
In mathematics, spectral graph theory (or analysis) is the study of properties of a graph (network) in relationship to the characteristic polynomial, eigenvalues, and eigenvectors of matrices associated to the graph, such as its adjacency matrix or Laplacian matrix. There are two excellent books, i.e., *Spectral Graph Theory* (Chung 1997) and *Spectra of Graphs: Theory and Applications* (D. Cvetkovi'c et al. 1980), which focus on deducing the properties and structure of a graph from its graph spectrum and reveal increasingly rich connections with many areas of mathematics and other disciplines, such as quantum chemistry, statistical physics, and computer science. The canonical example is the use of eigenvalue techniques to prove that certain extremal graphs cannot exist. The eigenvalues of a network are intimately connected to important topological features such as diameter (average distance), clustering coefficient, connectivity, and how random the network is. The associated eigenvectors can be used to detect community structure, or clustering. What is more, some important results purely on networks cannot be proved without resorting to algebraic methods, involving a consideration of eigenvalues of adjacency (Laplacian) matrices of graphs. For example, Gkantsidis et al. (Gkantsidis et al. 2003) use the weights of the eigenvector corresponding to the largest eigenvalue of the adjacency matrix to obtain an alternative hierarchical ranking of the autonomous system. Spectral analysis has been successfully applied to the detection of community structure of networks, being based on the adjacency matrix, the Laplacian matrix, the normalized Laplacian matrix, etc. Moreover, many real networks may be visualized by spectral methods (see (Seary and Richards 2005)). Social network analysis can be dated back in the early 1920s and has now become one of the most important methods in investigating the features and structures of social systems (see (Scott 2000; Wu et al. 2011; Wasserman and Faust 1994).

In the article we introduce results about spectral analysis of social networks and explain how to find the community structure and centrality of social networks by the means of spectral network theory.

Adjacency Spectrum

There are several matrices associated with a network. For a network $G = (V, E)$ with $V = \{v_1, \cdots, v_n\}$, the most commonly used matrix may be the adjacency matrix $A(G) = (a_{ij})$ of order n. Clearly the adjacency matrix of a network is symmetric and the entries of the main diagonal are zeros. In this way, there is one-to-one correspondence between networks and (0, 1)-symmetric matrices with zeros on the main diagonal. So all information of networks can be presented and obtained by the properties of matrices. By the way, the adjacency matrix can also be generalized to represent weighted networks.

Example 0.1 *For the graph G in Fig. 1: G: five vertices and eight edges.*



Spectral Analysis, Fig. 1G

Then the adjacency matrix is

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}.$$

One of the important sets associated with a network is the set of all walks between any pair of nodes v_i and v_j.

Proposition 0.2 *Let $G = (V, E)$ be a network associated with the adjacency matrix A. Then the number of walks of length k starting at node v_i and ending at node v_j is the (i, j) entry of A^k.*

From this proposition, if the entry (i, j) of A^k is positive, then there exist at least one path from node v_i to node v_j. Hence we can conclude that the diameter of a network is at most d if there exists an integer d such that all entries of A^d are positive.

Another property of networks is revealed by the entries of A.

Proposition 0.3 *Let $G = (V, E)$ be a network associated with the $n \times n$ adjacency matrix A. Then G is bipartite if and only if for some odd integer r, the diagonal entries of A^r are all zero.*

Since the adjacency matrix of G is symmetric, all eigenvalues of A are real and can be denoted by

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n.$$

Moreover, A is nonnegative matrix. By the Perron-Frobenius theorem, λ_1 must be positive and be greater than or equal to the absolute values of the other eigenvalue, i.e., $\lambda_1 \geq |\lambda_i|$ for $i = 2,\ldots, n$. Further there exists, up to constant multiplication, only one eigenvector with all nonnegative entries, corresponding to the eigenvalue λ_1. For example, the eigenvalues of the star network of order n are

$$\sqrt{n-1}, \ -\sqrt{n-1}, 0, \ldots 0.$$

The eigenvalues of the complete network of order n are

$$n-1, \ -1, \quad \ldots, \ -1.$$

The eigenvalues of a path of order n are

$$2\cos\frac{\pi}{2(n+1)}, 2\cos\frac{2\pi}{2(n+1)}, \ldots, \\ 2\cos\frac{k\pi}{2(n+1)}, \ldots, 2\cos\frac{n\pi}{2(n+1)}.$$

The eigenvalues of cycle are

$$2, 2\cos\frac{2\pi}{n}, 2\cos\frac{4\pi}{n}, 2\cos\frac{2k\pi}{n}, \ldots, \\ 2\cos\frac{2(n-1)\pi}{n}.$$

The eigenvalues of G in Fig. 1 are

$$3.2361, 0, 0, \ -1.2361, \ -2,$$

and corresponding to eigenvectors,

$$(0.4253, 0.4253, 0.4253, 0.4253, 0.5257)^T, \\ (-0.6932, \ -0.1398, 0.6932, 0.1398, 0.0000)^T, \\ (-0.1398, 0.6932, 0.1398, \ -0.6932, \ -0.0000)^T, \\ (-0.2629, \ -0.2629, \ -0.2629, \ -0.2629, 0.8507)^T, \\ (0.5000, \ -0.5000, 0.5000, \ -0.5000, 0.0000)^T.$$

The following proposition reveals the relationship between the structure of the network and its spectrum.

Proposition 0.4 *Let G be a network with adjacency matrix $A(G)$. Then*

(i). *G is bipartite if and only if $\lambda_1 = -\lambda_n$.*
(ii). *$\sum_{i=1}^{n} \lambda_i^k = tr\,A^k$, where trM is the trace of a matrix M is equal to the sum of all entries of the main diagonal of M.*
(iii). *If H is subnetwork of a G, then*

$$\lambda_{min}(G) \leq \lambda_{min}(H) \leq \lambda_{max}(H) \leq \lambda_{max}(G).$$

In fact, (ii) of Proposition 0.4 are trues for the Laplacian and Normal matrices, while (iii) of Proposition 0.4 is not true for the normal matrix.

For example, let G be a complete network of order 3 and H be a subnetwork of G by deleting an edge. Then

$$\lambda_1(G) = 2, \lambda_2(G) = \lambda_3(G) = -1, \lambda_1(H) \\ = \sqrt{2}, \lambda_2(H) = 0, \lambda_3(H) - \sqrt{2},$$

$$\mu_1(G) = 3 = \mu_2(G) = 3, \mu_3(G) = 0, \mu_1(H) \\ = 3, \mu_2 = 1, \mu(H) = 0,$$

$$v_1(G) = 1, v_2(G) = v_3 = -\frac{1}{2}, v_1(H) = 1, v_2(H) \\ = 0, v_3(H) = -1.$$

Laplacian Spectrum

The Laplacian matrix of a network dates back to the Kirchhoff matrix-tree theorem. The discrete graph Laplacian shares many important properties with the well-known continuous Laplacian operator of mathematical physics. It is easy to see that the Laplacian matrix $L(G)$ of a network is symmetric positive semi-definite, that zero is its smallest eigenvalue, and that this eigenvalue corresponds to the eigenvector $x = (1,..., 1)^T$. One of the most important results is the following:

Proposition 0.5 *Let G be a graph of order n with the Laplacian matrix L. Then the number of nonidentical spanning trees of G is equal to any cofactor of L. Moreover, the number of nonidentical spanning trees of G is equal to* $\frac{1}{n}\mu_1 \cdots \mu_{n-1}$˙

From this proposition, we can see that the second smallest eigenvalue is positive if and only if G has a spanning tree. In other words, G is connected if and only if G has only one zero Laplacian eigenvalue. Further, there is relationship between the number of zero eigenvalues and the number of connected components of a graph.

Proposition 0.6 *Let G be a simple network with the Laplacian matrix L. Then the number of zero eigenvalues of L is equal to the number of connected components of G.*

This proposition asserts that G is connected if and only of $\mu_{n-1} > 0$. Hence Fiedler (Fiedler 1973) called μ_{n-1} the algebraic connectivity of G and denoted α or $\alpha(G)$. He also proved the following:

Proposition 0.7 *Let G be a simple graph other than the complete graph. Then the algebraic connectivity is no more than the vertex connectivity. In other words, the algebraic connectivity of G is bounded above by the vertex connectivity of G.*

This proposition suggests that the algebraic connectivity is a suitable measure for connectivity of a network.

Normal Spectrum

It is easy to see that the normal matrix of G $N = D(G)^{-1}A(G)$ is a stochastic matrix and serves as the probability transition matrix of a random walk, where $D(G)$ is degree diagonal matrix and $A(G)$ is the adjacency matrix of G. Consider a random walk on a network G, starting at a node v_i, at each step to each neighbor with probability $1/d(v_i)$, where $d(v_i)$ is the degree of vertex v_i. Random walks arise in many models in mathematics and physics and have important algorithmic applications. They can be used to reach "obscure" parts of large sets and also generate random elements in large and complicated sets. We observe that the random walk on a connected network is a Markov chain, and the probability distribution is proportional to the degree distribution. But a major problem is how to determine the number of steps k required for the distributions of a random walk to become close to its stationary distribution, given an arbitrary initial distribution. Here the second modular eigenvalue v of N plays an important role in the analysis of rapidly mixing Markov chains: if v is far from to 1, i.e., there is a large eigenvalue gap, then the walk quickly forgets where it started. If v closes to one, then there must be parts of the network that are not easy to reach in a random walk, implying long paths or a nearly disconnected network. An important measure of the speed of convergence, called the relative pointwise distance, is given by

$$\Delta(k) = \max_{x,y} \frac{N^k(y,x) - \pi(x)}{\pi(x)},$$

where N^k (y, x) is value of the (y, x) entry of N^k and π is the stationary distribution of the random walk. Chung (Chung 1997) showed that

$$\Delta(t) \le e^{-t\mu} \frac{vol(G)}{\min_x d_x},$$

where Vol(G) is the sum of all degrees in G.

Weinan et al. (Weinan et al. 2008; Li et al. 2009) proposed several effective algorithms for network partition based on the framework of optimal predictions and probabilistic framework which is related to a discrete-time Markov chain with the normal matrix N of a network.

Finding Community Structure

A common feature of many networks in biological and social system is "community structure," which means that network nodes can be divided into groups, with dense connections within groups and sparse connections between them. For example, in the friendship school studied by Moody (Moody 2001), one of the principal divisions in the network is by individuals' race. The analysis of the community structure in large collaboration networks can be used to reveal the informal organization and the nature of information flows through the whole system. It will be of interest and practical importance if we are able to find community structure from the networks. Several methods to detect community structure have already been proposed. Newman and Girvan (Girvan and Newman 2002) proposed a fast and effective algorithm, based on the link betweenness, which measures the fraction of all shortest paths passing through a given link. But it does not give an indication of the resolution of the clustering. An alternative way to deal with the communities is by spectral analysis. Further, Newman (Newman 2006a, b) proposed a number of possible algorithms for detecting community structure by means of the Laplacian eigenvectors. Recently, Bickela and Chen (Bickel and Chen 2009) proposed the random graph models which are aimed at unifying points of view and analyses of networks from social sciences. If a network has k clearly distinct communities, then the largest $k - 1$ eigenvalues of the normal matrix close to 1, the other eigenvalues being far from 1. Hence there exists an eigenvector among $k - 1$ eigenvectors corresponding to the largest $k - 1$ eigenvalues, whose components are approximatively constant values on nodes belonging to the same community. Servedio et al. (Servedio et al. 2004) proposed an optimization problem based on the matrix of a network. The objective function is

$$z(x) = \frac{1}{2} \sum_{(i,j) \in E(G)} \left(x_i - x_j\right)^2 w_{ij},$$

where the sum is taken over all edges $(i, j) \in E(G)$ and x_i are values assigned to the nodes, with the constraint function

$$\sum_{i,j=1}^{n} x_i x_j m_{ij} = 1,$$

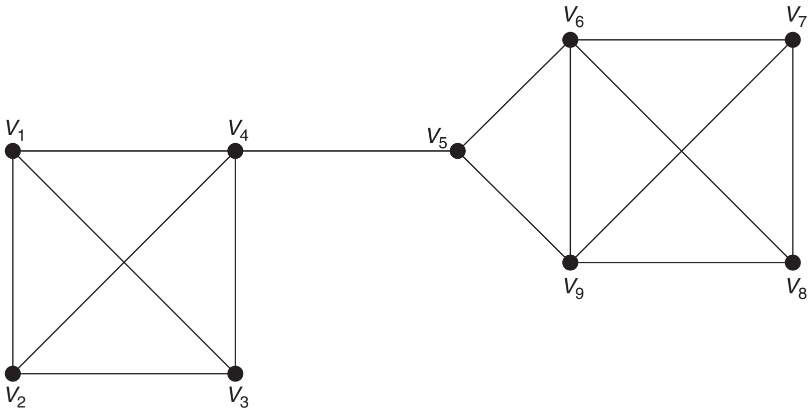
where m_{ij} are elements of a given symmetric matrix M. Then the stationary points of z are the solutions of

$$(D - A)x = \lambda M x,$$

where D is the diagonal matrix and λ is a Lagrange multiplier. If we choose $M = D$, then the solution becomes the eigenvalue problem of $D^{-1}Wx = (1 - 2\lambda)x$. If we choose $M = I$, then the solution becomes the Laplacian eigenvalue problem $(D - W)x = \lambda x$.

For example, the graph in Fig. 2 below H: 9 vertices and 15 edges.

Spectral Analysis, Fig. 2H



Then the Laplacian matrix of H is

$$L(H) = \begin{pmatrix} 3 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 3 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & -1 & 4 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 4 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & 3 & -1 \\ 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & 3 \end{pmatrix},$$

where vertex v_i is corresponding to the $i-$th rows in H. The eigenvector of H corresponding to the second smallest eigenvalue 0.2783 is 0.3911, 0.3911, 0.3911, 0.2822, -0.1228, *and* -0.3082.

Clearly, the positive components of the eigenvector correspond with the nodes v_1, v_2, v_3, v_4, and the negative components of the eigenvector correspond with nodes v_5, v_6, v_7, v_8, v_9. Obviously the nodes of the network can be divided into two groups, with one group v_1, v_2, v_3, v_4 and the other group v_5, v_6, v_7, v_8, v_9. In addition, Chauhan and Girvan (Chauhan et al. 2009) investigate the properties of spectra of networks with community structure. Nascimento and Carvalho (Nascimento and Carvalho 2011) presented a survey of graph clustering algorithms and different graph clustering formulations in literature, while Newman (Newman 2012) discussed the relations among communities, modules, and large-scale structure in networks.

Eigenvector Centrality Structure

The concept of centrality in a network plays an important role in the analysis of its structure. However, there are many different features which have been used to create measures of a centrality. Ruhnau (Ruhnau 2000) gave the following definition:

Definition 0.8 *Let $G = (V, E)$ be a connected network with $|V| = n$ and nc be a function which assigns a real value to every node of G. $nc(v_i)$ is called a node centrality of node v_i if.*

(i). $nc(v_i) \in [0, 1]$ for every $v_i \in V$.

(ii) $nc(v_i) = 1$ if and only if G is the star $S_{1,n-1}$ and $i = 1$.

Bonacich (Bonacich 1972) defines the centrality $c(v_i)$ of a node v_i to be a positive multiple of the sum of adjacent centralities, i.e.,

$$\lambda c(v_i) = \sum_{i,j=1}^{n} a_{ij} c(v_j), \quad \forall i,$$

where $A = (a_{ij})$ is the adjacency matrix of a network. The equations are equivalent to the eigenvalue-eigenvector problem of A. By the Perron-Frobenius theorem, there exists, for connected graphs, an eigenvector corresponding to the largest eigenvalue, with all positive entries. The entry $c(v_i)$ is called the eigenvector centrality of node. Then the function

$$nc_e(v_i) \equiv \frac{\sqrt{2}c(v_i)}{\sqrt{\sum_{i=1}^{n} c(v_i)^2}}$$

is node centrality. Ruhnau analyzes the structure of networks by using several centrality concepts, including degree centrality, closeness, betweenness, and eigenvector centrality. On the other hand, network, Mieghem et al. (Van Mieghem et al. 2010) established some relationships among the spectrum, the maximum modularity, and assortativity. In particular, the maximum modularity increases when the number of clusters decreases, and the average hop count and the effective graph resistance increase with increasing assortativity.

Conclusions

In this entry, we have described some properties of three kinds of matrices associated with a network, which are used to analyze the topological structure, community structure, and centrality.

Cross-References

- ► Clustering Algorithms
- ► Eigenvalues: Singular Value Decomposition
- ► Iterative Methods for Eigenvalues/Eigenvectors
- ► Matrix Algebra, Basics of
- ► Matrix Decomposition
- ► Probability Matrices
- ► Ranking Methods for Networks
- ► Semirings and Matrix Analysis of Networks
- ► Spectral Evolution of Social Networks

Acknowledgments This work is supported by the National Natural Science Foundation of China (Nos. 11531001 and 11271256), the Joint NSFC-ISF Research Program (jointly funded by the National Natural Science Foundation of China and the Israel Science Foundation (No. 11561141001), Innovation Program of Shanghai Municipal Education Commission (No. 14ZZ016), and Specialized Research Fund for the Doctoral Program of Higher Education (No. 20130073110075).

References

Bickel PJ, Chen A (2009) A nonparametric view of network models and Newman-Girvan and other modularities. Proc Natl Acad Sci U S A 106:21068–21073

Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U (2006) Complex networks: structure and dynamics. Phys Rep 424:275–308

Bonacich P (1972) Factoring and weighting approaches to status scores and clique identification. J Math Sociol 2:113–120

Chauhan S, Girvan M, Ott E (2009) Spectral properties of networks with community structure. Phys Rev E 80:0561104

Chung FRK (1997) Spectral graph theory. AMS Publications, Providence

Cvetkovi’c D, Doob M, Sachs H (1980) Spectra of graphs-theory and applications. Academic Press, New Work. Third edition, 1995

Fiedler M (1973) Algebra connectivity of graphs. Czechoslovake Mathematical Journal 23(98):298–305

Fortunato S (2010) Community detection in graphs. Phys Rep 48:75–174

Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci U S A 99:7821

Gkantsidis C, Mihail M, Zegura E (2003) Spectral analysis of internet topologies. In: IEEE INFOCOM. San Francisco, CA, USA

Li T, Liu J, Weinan E (2009) Probabilistic framework for network partition. Phys Rev E 80:026106

Moody J (2001) Race, school integration, and friendship segregation in America. Amer J Sociol 107:679–716

Nascimento MCV, Carvalho ACPF d (2011) Spectral methods for graph clustering-a survey. European J Oper Res 211:221–231

Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45(2):167–245

Newman MEJ (2006a) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74:036104

Newman MEJ (2006b) Modularity and community structure in networks. Proc Natl Acad Sci U S A 103:8577–8582

Newman MEJ (2012) Communities modules and large-scale structure in networks. Nat Phys 8:25–31

Ruhnau B (2000) Eigenvector-centrality – a node-centrality? Soc Networks 22:357–365

Scott J (2000) Social network analysis: a handbook. Sage Publications, London

Seary AJ, Richards WD (2005) Spectral methods for analyzing and visualizing networks: an introduction. In:

Breiger R, Carley KM, Pattison P (eds) Dynamic social network Modeling and analysis. National Academies Press, Washington, DC, pp 209–228

Servedio VDP, Colaiori F, Capocci A, Caldarelli G (2004) Community structure from spectral properties in complex network. In: Mendes JFF, Dorogovtsev SN, Abreu FV, Oliveira JG (eds) Science of complex networks: from biology to the internet and WWW; CNRT, pp 277–286

Van Mieghem P, Ge X, Schumm P, Trajanovski S, Wang H (2010) Spectral graph analysis of modularity and assortativity. Phys Rev E 82:056113

Wasserman S, Faust K (1994) Social network analysis. Cambridge University Press, Cambridge

Weinan E, Li T, Vanden-Eijnden E (2008) Optimal partition and effective dynamics of complex networks. Proc Natl Acad Sci U S A 105:7907–7912

Wu L, Ying X, Wu X, Zhou Z.-H (2011) Line orthogonality in adjacency eigenspace with application to community partition. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI11), Barcelona, July 16–22

Spectral Evolution Model

► Spectral Evolution of Social Networks

Spectral Evolution of Social Networks

Jérôme Kunegis
Institute for Web Science and Technologies,
University of Koblenz-Landau, Koblenz,
Germany

Synonyms

Spectral evolution model

Glossary

Adjacency Matrix	A characteristic matrix of a social network, typically denoted **A**. If the social network contains **n** persons, the adjacency matrix is a 0/1 $n \times n$ that contains 1 in the entries **A***ij* that correspond to an edge $\{i, j\}$ and 0 otherwise
Eigenvalue Decomposition	A decomposition of a square matrix giving $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}$, in which **U** contains the eigenvectors of **A** and $\mathbf{\Lambda}$ contains the eigenvalues
Singular Value Decomposition	A decomposition of any matrix giving $\mathbf{A} = \mathbf{U}\sum\mathbf{V}^{\mathrm{T}}$, in which $\sum$ contains the singular values of **A**
Spectral Evolution Model	The model that states that over time, eigenvectors stay constant and eigenvalues change
Spectrum	The set of eigenvalues or singular values of a matrix

Definition

The term *spectral evolution* describes a model of the evolution of network based on matrix decompositions. When applied to social networks, this model can be used to predict friendships, recommend friends, and implement other learning problems.

Introduction

The analysis of the evolution of social networks is an important field of study in the areas of information retrieval, data mining, recommender systems, and network science. As an example, models of the evolution of social networks can be used to solve the problem of link prediction, i.e., to predict which edges will appear in a network in the future (Liben-Nowell and Kleinberg 2003). Another common problem associated to models of network evolution is the friend recommendation problem, in which users of social networking sites are recommended to other users.

The spectral evolution model describes the evolution of network using matrix decompositions, in particular the eigenvalue and singular value decompositions of matrices associated with a network, such as the adjacency matrix and the Laplacian matrix. In its most generic version, the spectral evolution is based on the eigenvalue decomposition of the symmetric adjacency matrix of an undirected social network and can be stated in terms of eigenvectors and eigenvalues. The spectral evolution model then asserts that over time, the eigenvectors stay constant and the eigenvalues grow. Other similar formulations exist for other matrix decompositions, other characteristic graph matrices, and other types of social networks, such as directed networks.

Historical Background

In order to analyze graphs, algebraic graph theory is a common approach. In algebraic graph theory, a graph with n vertices is represented by an $n \times n$ matrix called the adjacency matrix, from which other matrices can be derived.

The edge set of an undirected graph $G = (V, E)$ can be represented by a matrix whose characteristics follow those of the graph. An unweighted undirected graph on n vertices can be represented by an $n \times n$ 0/1 matrix $\mathbf{A}$ defined by

$$\mathbf{A}_{ij} = \begin{cases} 1 & \text{if}\{i,j\} \in E \\ 0 & \text{otherwise} \end{cases}.$$

The matrix $\mathbf{A}$ is called the adjacency matrix of G.

Spectral graph theory is a branch of algebraic graph theory that applies matrix decompositions to characteristic graph matrices in order to study a graph's properties (Chung 1997; Cvetković et al. 1997). The word *spectral* refers to the spectrum of networks, which is given by the eigenvalue decomposition of a graph's adjacency or Laplacian matrix. Spectral graph theory can be used to study graph properties such as connectivity, centrality, balance, and clustering.

A square symmetric matrix $\mathbf{A}$ can be written in the following way:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}} \tag{1}$$

where $\mathbf{U}$ is an $n \times n$ orthogonal matrix and Λ is an $n \times n$ diagonal matrix. A matrix $\mathbf{U}$ is orthogonal when $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}$ or equivalently when $\mathbf{U}^{\mathrm{T}}\,\mathbf{U} = \mathbf{I}$. Another characterization of an orthogonal matrix $\mathbf{U}$ is that its columns are pairwise orthogonal vectors and each has unit norm. The values Λ_{kk} are the eigenvalues of $\mathbf{A}$, and the columns of $\mathbf{U}$ are its eigenvectors. We will designate the eigenvalues by $\lambda_k = \Lambda_{kk}$ and the eigenvectors by $\mathbf{u}_k = \mathbf{U}_k$ for $1 \leq k \leq n$.

A certain number of interesting graph properties can be described spectrally, such as connectivity (Mohar 1991), centrality (Brin and Page 1998), conflict and balance (Kunegis et al. 2010c), and clustering (Luxburg 2007). Spectral transformations were considered in 2009 in Kunegis and Lommatzsch. The spectral evolution model itself was introduced in 2010 (Kunegis et al. 2010b) and in detail in 2011 (Kunegis 2011).

The Spectral Evolution Model

We first describe the spectral evolution model for unweighted, undirected social networks based on the eigenvalue decomposition of the adjacency matrix, and will then review extensions of it to other types of networks and other characteristic graph matrices and decompositions.

Let $G_t = (V, E_t)$ be a social network that evolves over time, at time t. We assume that the set of vertices V is constant and will only consider evolving sets of edges E_t. Let $\mathbf{A}_t$ be the adjacency matrix of the social network as time t. We can now consider the eigenvalue decomposition

$$\mathbf{A}_t = \mathbf{U}_t\boldsymbol{\Lambda}\mathbf{U}_t^{\mathrm{T}}.$$

A priori, this eigenvalue decomposition will change from timepoint to timepoint. The spectral evolution model can now be stated as:

Definition 1 (Spectral Evolution Model) A network that changes over time is said to follow the spectral evolution model when its eigenvalues Λ_t evolves while its eigenvectors $\mathbf{U}_t$ stay approximately constant.

The spectral evolution model is a quantitative statement: The eigenvectors do not need to be exactly constant. In the general case, the spectral evolution model can be stated to hold when the eigenvectors change less than predicted by a random graph model and the eigenvalues change more than predicted by a random graph model.

Relationship to Link Prediction

The spectral evolution model can be compared to a number of link prediction models that are special cases of it. A link prediction function is a function used to implement the link prediction problem in social networks (Liben-Nowell and Kleinberg 2003).

Let $\mathbf{A}_1$ be the current adjacency matrix of the social network. A link prediction function is a function

$$f : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$$

that maps the current adjacency matrix $\mathbf{A}_1$ to its predicted value in the future $\mathbf{A}_2$. We will call f a spectral transformation when it can be expressed using the eigenvalue decomposition $\mathbf{A}_1 = \mathbf{U}\Lambda\mathbf{U}^{\mathrm{T}}$ as

$$f(\mathbf{A}_1) = \mathbf{U}f(\mathbf{\Lambda})\mathbf{U}^{\mathrm{T}}.$$

The spectral evolution model can then be stated as:

Definition 2 (Spectral Evolution Model, Alternative Definition) A network follows the spectral evolution model when a future value of its adjacency matrix can be predicted by application of a spectral evolution function.

Friend of a Friend Count

For instance, the friend of a friend count (or **common neighbor count**) is one such model: Given two users $i, j \in V$, the number of common friends of i and j can be used as a link prediction function. The higher the number of common neighbors, the more likely it is that an edge will appear between them in the social network. An example of that method is used on the social network Facebook (http://www.facebook.com) for recommending new friends. Mathematically, the common neighbor count can be expressed using the social network's adjacency matrix as the square $\mathbf{A}^2$. In fact, the entry $(\mathbf{A}^2)_{ij}$ equals the number of common friends of users i and j. Assuming that the probability that an edge will appear between i and j is proportional to $(\mathbf{A}^2)_{ij}$, there is a constant α such that the adjacency matrix in the future can be expressed as a spectral transformation of the original adjacency matrix:

$$\begin{aligned}\mathbf{A}_2 &= \mathbf{A} + \alpha\mathbf{A}^2 \\ &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}} + \alpha\left(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}\right)^2 \\ &= \mathbf{U}\left(\mathbf{\Lambda} + \alpha\mathbf{\Lambda}^2\right)\mathbf{U}^{\mathrm{T}}.\end{aligned}$$

Thus, the friend of a friend model predicts a spectral transformation of A and thus justifies the spectral evolution model of social networks. This argument can be extended to *the friend-of-a-friend-of-a-friend* model and models based paths of any length.

Graph Kernels

A related class of link prediction functions are given by graph kernels. The exponential graph kernel is defined as the exponential function of the adjacency matrix (Kondor and Lafferty 2002):

$$e^{\alpha\mathbf{A}} = \mathbf{I} + \alpha\mathbf{A} + \frac{\alpha^2}{2}\mathbf{A}^2 + \frac{\alpha^3}{6}\mathbf{A}^3 + \ldots.$$

The Neumann graph kernel is defined using matrix inversion (Kandola et al. 2002):

$$(\mathbf{I} - \alpha\mathbf{A})^{-1} = \mathbf{I} + \alpha\mathbf{A} + \alpha^2\mathbf{A}^2 + \alpha^3\frac{1}{6}\mathbf{A}^3 + \ldots.$$

Both graph kernels can be expressed as a spectral transformation:

$$\begin{aligned}e^{\alpha\mathbf{A}} &= \mathbf{U}e^{\alpha\mathbf{\Lambda}}\mathbf{U}^{\mathrm{T}} \\ (\mathbf{I} - \alpha\mathbf{A})^{-1} &= \mathbf{U}(\mathbf{I} - \alpha\mathbf{\Lambda})^{-1}\mathbf{U}^{\mathrm{T}}\end{aligned}$$

Thus, the two graph kernels justify the spectral evolution model in the sense that if they produce accurate link predictions, the social network will grow according to the spectral evolution model.

Preferential Attachment

Preferential attachment is a simple link prediction model based on the idea that the probability of a new link being formed is proportional to the degrees of the nodes it connects. This idea can be extended to the decomposition of a graph's adjacency matrix, resulting in the latent preferential attachment model, first described in Kunegis (2011).

The eigenvalue decomposition of **A** can be written as a sum of rank-one matrices:

$$\mathbf{A} \approx \sum_{k=1}^{r} \mathbf{\Lambda}_k \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}}$$

where r is the rank of the decomposition. The usual interpretation of a matrix factorization is that each latent dimension k represents a *topic* in the network. Then $\mathbf{U}_{ik}$ represents the importance of vertex i in topic k, and λ_k represents the overall importance of topic k. Each rank-one matrix $\mathbf{A}^{(k)} = \mathbf{\Lambda}_k \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}}$ can be interpreted as the adjacency matrix of a weighted graph. Now, assume that preferential attachment is happening in the network, but restricted to the subgraph G_k. Then the probability of the edge $\{i, j\}$ appearing will be proportional to $d_k\,(i)\,d_k\,(j)$, where $d_k\,(i)$ is the degree of node i in the graph G_k. This degree can be written as the sum over edge weights in G_k:

$$\begin{aligned} d_k(i) &= \sum_l \mathbf{A}_{il}^{(k)} = \sum_l \mathbf{\Lambda}_k \mathbf{U}_{ik} \mathbf{U}_{lk} \\ &= \mathbf{U}_{ik} \mathbf{\Lambda}_k \sum_l \mathbf{U}_{lk} \sim \mathbf{U}_{ik}. \end{aligned}$$

In other words, d_k (i) is proportional to **U***ik*. Therefore, the preferential attachment value is proportional to the corresponding entry in **A**(^{k}):

$$d_k(i) d_k(j) \sim \mathbf{U}_{ik} \mathbf{U}_{jk}.$$

These values can be aggregated into a matrix **P**(k) giving the preferential attachment values for all pairs (i, j):

$$\mathbf{P}^{(k)} \sim \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}}.$$

Assuming that a preferential attachment process is happening for each subgraph G_k separately, with a weight ε_k depending on the topic k, then the overall preferential attachment prediction can be written as $\mathbf{P} = \sum_k \varepsilon_k \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}}$. Here, we replace proportionality by equality since the proportionally constants are absorbed by the constants ε_k. The matrix **P** can then be written in the following form, giving its eigenvalue decomposition $\mathbf{P} = \mathbf{U}\mathbf{E}\mathbf{U}^{\mathrm{T}}$, where **E** is the diagonal matrix containing the individual topic weights $\mathbf{E}_{kk} = \varepsilon_k$. This prediction matrix is a spectral transformation of the adjacency matrix **A**. Under this model, network growth can be interpreted as the replacement of the eigenvalues Λ by $\Lambda + \mathbf{E}$:

$$f(\mathbf{A}_1) = \mathbf{A} + \mathbf{P} = \mathbf{U}(\mathbf{\Lambda} + \mathbf{E})\mathbf{U}^{\mathrm{T}}.$$

Since the values **E** are not modeled by the latent preferential attachment model, every spectral transformation can be interpreted as latent preferential attachment, and thus, the latent preferential attachment model is equivalent to the spectral evolution model.

Learning Spectral Transformations

Under the assumption that a social network evolves according to the spectral evolution model, the best possible link prediction function can be learned using curve fitting (Kunegis and Lommatzsch 2009).

Given the current adjacency matrix $\mathbf{A}_1$ and the future adjacency matrix $\mathbf{A}_2$, the best possible link prediction function f that maps $\mathbf{A}_1$ to $\mathbf{A}_2$ is given by the following minimization problem:

$$\min_f \|f(\mathbf{A}_1) - \mathbf{A}_2\|_{\mathrm{F}}.$$

Using the eigenvalue decomposition $\mathbf{A}_1 = \mathbf{U}\Lambda$ $\mathbf{U}^{\mathrm{T}}$ of rank r, this problem is equivalent to

$$\min_f \left\| f(\mathbf{\Lambda}) - \mathbf{U}^{\mathrm{T}} \mathbf{A}_2 \mathbf{U} \right\|_{\mathrm{F}}.$$

Since Λ is diagonal and $f(\Lambda)$ is diagonal too, only the diagonal elements of $\mathbf{U}^{\mathrm{T}}$ $\mathbf{A}_2$ $\mathbf{U}$ influence the minimization problem. Thus, the minimization problem is equivalent to

$$\min_f \sum_{i=1}^{r} \left(f(\mathbf{\Lambda}_{ii}) - \left(\mathbf{U}^{\mathrm{T}}\mathbf{A}_2\mathbf{U}\right)_{ii} \right)^2. \qquad (2)$$

This is a one-dimensional curve-fitting problem with r parameters and can be solved efficiently. For each spectral link prediction function, the corresponding spectral transformation function can be fitted to solve the optimization problem in Eq. 2, learning its parameters in the process, for instance, the parameter α for the exponential and Neumann graph kernels.

An alternative way of learning spectral transformations is based on the extrapolation of the eigenvalues into the future (Kunegis et al. 2010b). This gives new values for the eigenvalues, which can be combined with the unchanging eigenvectors to give the predicted value of the adjacency matrix.

Tests of the Spectral Evolution Model

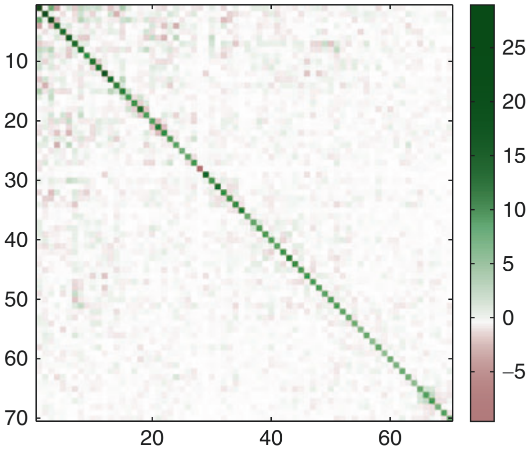
In addition to the fact that the spectral evolution model has known link prediction functions as special cases, it can be verified experimentally by measuring the change in the eigenvectors and eigenvalues of actual social networks, of which the temporal evolution is known. These observations can then be combined with the changes predicted by a random graph growth model in which edges are added randomly to a network.

When adding a small random perturbation $\mathbf{E}$ of size $\|\mathbf{E}\|_{\mathrm{F}} = \varepsilon$ to the adjacency matrix $\mathbf{A}$ to give $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$, the expected change in the new eigenvalues $\widetilde{\mathbf{\Lambda}}$ and the new eigenvectors $\tilde{\mathbf{U}}$ is given by

$$\begin{aligned} \left\| \mathbf{\Lambda} - \widetilde{\mathbf{\Lambda}} \right\|_{\mathrm{F}} &= O(\varepsilon^2) \\ \left| \mathbf{U}_k \cdot \tilde{\mathbf{U}}_k \right| &= O(\varepsilon). \end{aligned}$$

These results can be shown by a perturbation argument (Stewart 1990) and ultimately can be derived from theorems by Weyl (1912) and Wedin (1972). As a result, eigenvectors are expected to change faster than eigenvalues for random additions to the adjacency matrix, justifying the spectral evolution model for social networks.

When the growth of actual social network can be observed over time, the spectral evolution model can be verified directly. As an example for a method of achieving this, we describe the spectral diagonality test. The spectral diagonality test can be computed from the snapshot of a network at two different times 1 and 2, using the adjacency matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ (Kunegis et al. 2010b). Using the eigenvalue decomposition $\mathbf{A}_1 = \mathbf{U}\Lambda\ \mathbf{U}^{\ \mathrm{T}}$, the spectral diagonality test consists in verifying the diagonality of the matrix $\Delta = \mathbf{U}^{\mathrm{T}}\ \mathbf{A}_2\ \mathbf{U}$. If the matrix Δ is diagonal, the evolution of the social network is perfectly spectral. In practice, Δ is not perfectly spectral, but almost so. An example of the matrix Δ is given Fig. 1 for a subset of the Facebook social network (Viswanath et al. 2009). In this instance of the spectral diagonality test, Δ is indeed almost diagonal, and the evolution of that network can be concluded to follow the spectral evolution model.



Spectral Evolution of Social Networks, Fig. 1 The spectral diagonality test matrix Δ for a subset of the Facebook social network (Viswanath et al. 2009). Since the matrix is almost diagonal, the test shows that the evolution of that subset of Facebook follows the spectral evolution model

Normalized Adjacency Matrix

The spectral evolution model can be extended to the normalized adjacency matrix, defined as $\mathbf{N} = \mathbf{D}^{-1/2}\ \mathbf{AD}^{-1/2}$, in which $\mathbf{D}$ denotes the diagonal degree matrix with $\mathbf{D}_{ii}$ being the degree of node i. In this definition, we assume that the social network does not contain any isolated nodes, i.e., users without friends.

The theory of spectral network evolution can be extended to using the matrix **N** instead of the matrix **A** without much change. A key difference to the unnormalized case is in the evolution of the eigenvalues over time: While the eigenvalues of **A** grow in the general case, the eigenvalues of **N** cannot grow without bounds, as by construction, they lie in the interval [−1,+1]. In fact, the eigenvalues of **N** will typically shrink over time (Kunegis et al. 2012).

Another difference in using **N** over **A** lies in the interpretation of corresponding link prediction functions. For instance, the exponential and Neumann graph kernels give the following link prediction functions in the normalized case:

$$\begin{aligned} e^{\alpha\mathbf{N}} &= \mathbf{I} + \alpha\mathbf{N} + \frac{\alpha^2}{2}\mathbf{N}^2 + \frac{\alpha^3}{6}\mathbf{N}^3 + \ldots \\ (\mathbf{I} - \alpha\mathbf{N})^{-1} &= \mathbf{I} + \alpha\mathbf{N} + \alpha^2\mathbf{N}^2 + \alpha^3\frac{1}{6}\mathbf{N}^3 + \ldots . \end{aligned}$$

Laplacian Matrix

The Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{A}$, too, can be used as the basis for studying the spectral evolution of networks. This leads to a more complicated situation. Although the eigenvectors do stay constant in the general case, the eigenvalues will not change continuously, but grow in steps, which makes the diagonality test impracticable. However, link prediction function can still be used in that case. These include the regularized commute-time kernel $(\mathbf{I} + \alpha\ \mathbf{L})^{-1}$ and the heat diffusion kernel $e^{-\alpha\mathrm{L}}$.

Bipartite Networks

Bipartite networks are networks in which the set of nodes V can be partitioned into two sets $V = V_1 \cup V_2$ such that all edges connect a node in V_1 with a node in V_2. Social networks are not bipartite in the general case, since they contain triangles. Still, many bipartite networks can be found in social media, for instance, user–group inclusion networks or user–item rating networks. In such networks, the spectral evolution model can be applied as is with good results. However, a simplification of the expression is possible, due to the special structure of the networks (Kunegis et al. 2010a).

The adjacency matrix **A** of a bipartite network can always be written as

$$\mathbf{A} = \begin{bmatrix} 0 & \mathbf{B} \\ \mathbf{B}^{\mathrm{T}} & 0 \end{bmatrix}$$

for a matrix **B** of size $|V_1| \times |V_2|$. The matrix **B** is then called the biadjacency matrix of the network. This can be exploited to reduce the eigenvalue decomposition of **A** to the singular value decomposition of **B**. Given the singular value decomposition $\mathbf{B} = \mathbf{U}\sum V^{\mathrm{T}}$, the eigenvalue decomposition of **A** is given by

$$\mathbf{A} = \begin{bmatrix} \overline{\mathbf{U}} & \overline{\mathbf{U}} \\ \overline{\mathbf{V}} & -\overline{\mathbf{V}} \end{bmatrix} \begin{bmatrix} +\boldsymbol{\Sigma} & 0 \\ 0 & -\boldsymbol{\Sigma} \end{bmatrix} \begin{bmatrix} \overline{\mathbf{U}} & \overline{\mathbf{U}} \\ \overline{\mathbf{V}} & -\overline{\mathbf{V}} \end{bmatrix}^{\mathrm{T}} \quad (3)$$

with $\overline{\mathbf{U}} = \mathbf{U}/\sqrt{2}$ and $\overline{\mathbf{V}} = \mathbf{V}/\sqrt{2}$. In this decomposition, each singular value a corresponds to the eigenvalue pair $\{\pm\sigma\}$. Odd powers of **A** then have the form

$$\mathbf{A}^{2k+1} = \begin{bmatrix} 0 & \left(\mathbf{BB}^{\mathrm{T}}\right)^k\mathbf{B} \\ \left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)^k\mathbf{B}^{\mathrm{T}} & 0 \end{bmatrix},$$

where the alternating power $(\mathbf{BB}^{\mathrm{T}})^{\mathrm{K}}\ \mathbf{B}$ can be explained by the fact that in the bipartite network, a path will follow edges from one vertex set to the other in alternating directions, corresponding to the alternating transpositions of **B**.

Thus, it is sufficient, in a bipartite network, to consider only odd functions of the biadjacency matrix **B**. Here, an odd function is to be understood as a function/for which it holds that $f(-\mathbf{A}) = -(\mathbf{A})$. Examples of resulting odd link prediction functions are the matrix hyperbolic sine sinh (α **A**) and the Neumann pseudokernel

$\alpha\,\mathbf{A}(\mathbf{I}-\alpha^2\,\mathbf{A})^{-1}$. These functions are pseudokernels and not kernels, as they are not positive definite.

Directed Networks

The case of directed networks is more complicated than the other cases, since the eigenvectors of the adjacency matrix are not orthogonal anymore in that general case. Four methods can be used for directed networks:

- **Ignoring Edge Directions** By ignoring edge directions, the problem is reduced to the undirected case. This is sensible in social networks that tend to be symmetric, such as communication networks, but does not give good results in networks that are inherently directed, such as trust networks.
- **Working on the Bipartite Double Cover** By considering the bipartite double cover of a directed network, the problem reduces to the bipartite case. The bipartite cover of a directed graph is constructed by replacing each node by two nodes, one that keeps all in-edges and one that keeps all out-edges. The resulting link prediction methods work well when the primary mechanism of graph growth follows paths of alternating signs. An example of such networks are citation networks, in which co-citation can be interpreted using paths of alternating directions.
- **Non-orthogonal Decomposition** The non-orthogonal eigenvalue decomposition of a directed network can be used with difficulty. Since the matrix **A** is asymmetric, the eigenvalue decomposition must be written as $\mathbf{A}=\mathbf{U}\Lambda\,\mathbf{U}^{-1}$ and will contain complex eigenvalues. The link prediction methods described in the previous sections do not perform well in that case. In the extreme case, if a directed network is acyclic, for instance, a scientific citation network, then all eigenvalues are zero, and all graph kernels and other link prediction methods return only the value zero.
- **DEDICOM** The last variant uses matrix decompositions of the form $\mathbf{A}=\mathbf{U}\mathbf{X}\mathbf{U}^{\mathrm{T}}$ in which **U** is orthogonal and **X** is not diagonal. Such decompositions are called DEDICOM (decomposition into directed components) (Harshman 1978). This decomposition is not unique, and thus, there are multiple variants of DEDICOMs. In general, the choice of a variant will involve the tradeoff between a fast computation and an accurate decomposition. This method is best suited to networks in which directed triangle closing is the main mechanism by which new edges are formed, for instance, in trust networks.

Key Applications

The spectral evolution model can be used to implement link prediction functions which themselves can be used to solve several different kinds of problems in social networks:

- Applying the link prediction problem to an ordinary social network leads to the recommendation of new friends. In this case, edges are unweighted, and the links to be predicted describe the similarity between nodes.
- Trust prediction in a social network consists of predicting trust edges in a directed social network consisting of trust edges. In some cases, distrust edges are additionally known.
- Rating prediction is a special case of link prediction, where edges are weighted. An important application of rating prediction is collaborative filtering, in which the network is either unipartite when users are rated as in dating sites or items are rated as in movie rating sites.
- In a signed network, the prediction of an edge's sign, knowing that the edge is part of the network, is known as the link sign prediction problem.
- To predict future interactions in social networks, for instance, emails or scientific coau-thorship, link prediction can be performed in a network with multiple edges.

The spectral evolution model applies equally to all these variants of the link prediction problem, with appropriate choice of matrix and decomposition type.

Future Directions

As of 2012, the link prediction problem in all its variants is not fully covered by research, and new applications are still being published. In particular, the application of social network analysis methods such as the spectral evolution model is increasingly applied to other kinds of networks, such as content networks or hyperlink networks. Another area of research lies in the exploration of more complex matrix decompositions, such as nonnegative decompositions and tensor decompositions.

Cross-References

- ▶ Community Evolution
- ▶ Data Mining
- ▶ Eigenvalues: Singular Value Decomposition
- ▶ Link Prediction: A Primer
- ▶ Matrix Decomposition
- ▶ Network Models
- ▶ Recommender Systems: Models and Techniques
- ▶ Spectral Analysis
- ▶ Temporal Networks

Acknowledgments We thank our collaborators on previous work: Christian Bauckhage, Damien Fay, and Andreas Lommatzsch. The author of this work has received funding from the European Community's Seventh Frame Programme under grant agreement no 257859, ROBUST.

References

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30 (1–7):107–117

Chung F (1997) Spectral graph theory. American Mathematical Society, Providence

Cvetković D, Rowlinson P, Simic S (1997) Eigenspaces of graphs. Cambridge University Press, Cambridge

Harshman RA (1978) Models for analysis of asymmetrical relationships among n objects or stimuli. In: Proceedings of the first meeting of the psychometric society and the society for mathematical psychology, Hamilton

Kandola J, Shawe-Taylor J, Cristianini N (2002) Learning semantic similarity. In: Advances in neural information processing systems, Whistler, pp 657–664

Kondor R, Lafferty J (2002) Diffusion kernels on graphs and other discrete structures. In: Proceedings of the international conference on machine learning, Sydney, pp 315–322

Kunegis J (2011) On the spectral evolution of large networks. PhD thesis, University of Koblenz–Landau. http://userpages.uni-koblenz.de/~kunegis/paper/kunegis-phd-thesis-on-the-spectral-evolution-of-large-networks.pdf

Kunegis J, Lommatzsch A (2009) Learning spectral graph transformations for link prediction. In: Proceedings of the international conference on machine learning, Montréal, pp 561–568. http://uni-koblenz.de/~kunegis/paper/kunegis-spectral-transformation.pdf

Kunegis J, De Luca EW, Albayrak S (2010a) The link prediction problem in bipartite networks. In: Proceeding of the international conference in information processing and management of uncertainty in knowledge-based systems, Dortmund, pp 380–389. http://uni-koblenz.de/~kunegis/paper/kunegis-hyperbolic-sine.pdf

Kunegis J, Fay D, Bauckhage C (2010b) Network growth and the spectral evolution model. In: Proceeding of the international conference on information and knowledge management, Toronto, pp 739–748. http://uni-koblenz.de/~kunegis/paper/kunegis-spectral-network-evolution.pdf

Kunegis J, Schmidt S, Lommatzsch A, Lerner J (2010c) Spectral analysis of signed graphs for clustering, prediction and visualization. In: Proceedings SIAM international conferences on data mining, Columbus, pp 559–570. http://uni-koblenz.de/~kunegis/paper/kunegis-spectral-analysis-of-signed-graphs.pdf

Kunegis J, Sizov S, Schwagereit F, Fay D (2012) Diversity dynamics in online networks. In: Proceedings of the conference on hypertext and social media, Milwaukee, pp 255–264. http://userpages.uni-koblenz.de/~kunegis/paper/kunegis-diversity-dynamics-in-online-networks.pdf

Liben-Nowell D, Kleinberg J (2003) The link prediction problem for social networks. In: Proceedings international conference on information and knowledge management, New Orleans, pp 556–559

Luxburg UV (2007) A tutorial on spectral clustering. Stat Comput 17(4):395–416

Mohar B (1991) The Laplacian spectrum of graphs. Graph Theory Combin Appl 2:871–898

Stewart GW (1990) Perturbation theory for the singular value decomposition. Technical report, University of Maryland, College Park

Viswanath B, Mislove A, Cha M, Gummadi KP (2009) On the evolution of user interaction in Facebook. In: Proceeding of the workshop on online social networks, Barcelona, pp 37–42

Wedin PÅ (1972) Perturbation bounds in connection with singular value decomposition. BIT Numer Math 12(1):99–111

Weyl H (1912) Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differenzialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). Math Ann 71(4):441–479

Spectral Graph Analysis

▶ Spectral Analysis

Spectral Network Analysis

▶ Spectral Analysis

Spectral Technique

▶ Spectral Analysis

Stability and Evolution of Scientific Networks

Eugenia Galeota[1], Susanna Liberti[2], Frédéric Amblard[3] and Walter Quattrociocchi[4,5]
[1]Computational Epigenetics Group, Center for Genomic Science IIT@SEMM, Istituto Italiano di Tecnologia, Milano, Italy
[2]Department of Biology, University of Rome Tor Vergata, Rome, Italy
[3]IRIT – Université Toulouse 1 Capitole, Toulouse, France
[4]London Institute of Mathematical Sciences, London, Mayfair, UK
[5]Natwork Department, IMT Alti Studi Lucca, Lucca, Italy

Synonyms

Emergence; Evolving structures; Scientific communities; Social networks analysis; Social selection; Temporal metrics; Time-varying graphs

Glossary

APS	American Physical Society
DOI	Digital Object unique Identifiers associated to papers
Egalitarian growth	the growth benefiting on average equally to each node
Evolution of social networks	The change in time of the structure of a social network due to changing interactions between the components
TVG	Time Varying Graphs

Introduction

Nowadays one of the most pressing as well as interesting scientific challenges deals with the analysis and the understanding of social systems' dynamics and how these evolve according to the interactions among their components. The efforts in this area strive to understand what are the driving forces behind the evolution of social networks and how they are articulated together with social dynamics – e.g., opinion dynamics, the epidemic or innovation diffusion, the teams formation, and so forth (Deffuant et al. 2001; Moore and Newman 2000; Lelarge 2008; Carter 2007; Powell et al. 2005; Guimera et al. 2005; Quattrociocchi et al. 2009, 2010). In this paper we approach the challenge of depicting the evolution of social systems from a network science's perspective. As an example, we chose the case of scientific communities by analyzing a portion of the American Physical Society dataset (APS). The analysis addresses the coexistence of coauthorships' and citations' behaviors of scientists. On one hand, the studies on scientific network dynamics deal with the understanding of the factors that play a significant role in their evolution, not all of them being neither objective nor rational, e.g., the existence of a star system (Wagner and Leydesdorff 2005; Newman 2001a, 2004a; Barabasi et al. 2002), the blind imitation concerning the citations (MacRoberts and MacRoberts 1996), the reputation, and community affiliation bias (Gilbert 1977). On the other hand, having some elements to understand such dynamics could enable for a better detection of the hot topics and of the vivid subfields and how the scientific production is advanced with respect to selection process inside the community itself. Among the available data to analyze such a system, a subset of the publications in a given field is the most frequently used such as in (De Solla Price 1965; Newman 2001b; 2004b; Quattrociocchi et al. 2012; Amblard et al. 2011; Santoro et al. 2011; Radicchi et al. 2009). The scientific publications correspond to the production of such a system and clearly identify who are the producers (the authors), which institution they belong to (the affiliation), which funded project they are working on (the acknowledgement), and what are the related publications (the citations), having most of the time a public access to these data explain also a part of its frequent use in the analyses of the scientific field. Classical analyses concern either the coauthorships network (Newman 2001a; Barabasi et al. 2002) or the citation network (Hummon and Dereian 1989; Redner 2005), more rarely the institutional network (Powell et al. 2005). Moreover, such networks are often considered as static and their structure is rarely analyzed overtime (an exception is the one performed by (Radicchi et al. 2009)). The illustrative analysis presented in the paper passes through different data transformations aimed at providing different perspectives on the APS network and its evolution. In (Newman 2001a) the network of scientific collaborations, explored upon several databases, shows a clustered and small-world structure. Moreover, several differences between the collaborations' patterns of the different fields studied are captured. Such differences have been deepened in (Newman 2004a) with respect to the number of papers produced by a given group of authors, the number of collaborations, and the topological distances between scientists. Peltomaki and Alava in (Peltomaki and Alava 2006) propose a new emulative model aimed at approximating the growth of scientific networks by incorporating bipartition and sublinear preferential attachment. A model for the self-assembly of creative teams based on three parameters (e.g., team size, the rate of newcomers in the scientific production, and the tendency of authors to collaborate with the same group) has been outlined in (Guimera et al. 2005). Connectivity patterns in a citations network have been studied with respect to the development of the DNA theory (Hummon and Dereian 1989). The work of Klemm and Eguiluz (2002) observed that real networks' (e.g., movie actors, coauthorship in science, and word synonyms) growing patterns are characterized by a clustering trend that reaches an asymptotic value larger than regular lattices of the same average connectivity. In the field of social network analysis several works have approached the problem of temporal metrics (Holme 2005; Kostakos 2009; Kossinets et al. 2008). The focus is on the definition of instruments able to capture the intrinsic properties of complex systems' evolution, that is, characterizing the interdependencies and the coexistence between local behaviors (interactions) and their global effects (emergence) (Deffuant et al. 2001; Quattrociocchi et al. 2010; Davidsen et al. 2002; Mataric 1992; Woolley 1994). The work of (Wang et al. 2016) provides an analysis of the temporal dynamics of collaboration within scientific consortia, showing a higher modularity when the network was built from authors participating in the consortia with respect to a collaboration network of authors that weren't involved in any consortia. The research approach to characterize the evolution patterns of social networks, at the very beginning, was mainly based upon simulations, while in the past few years, due to the large availability of real datasets, either the methodology of analysis and the object of research have changed (Kossinets et al. 2008; Taramasco et al. 2010; Leskovec et al. 2007). Actually the field of social network analysis has been proved to have many practical applications. The analysis of scientific collaboration networks has been used to support the development of efficient research plans in health, to map priority thematic areas, and investigate the relationship of these networks with the technological progress (Fonseca et al. 2016), the globalization, and historical/political events (Pan et al. 2012).

Analysis of Scientific Networks Dynamics

In this work we present a very basic analysis aiming at understanding the social aspects of the scientific systems by coupling the collaborations between scientists and their effect on the scientific community itself through the citation network. The data to build up the networks analyzed in this work has been extracted from the APS (American Physical Society) dataset, made available upon request by the APS for research purposes. The database contains information about 463,343 articles published on 11 journals of the APS in a time span ranging from 1892 to 2009. For the citations network, we used a list of 2,944,144 DOI pairs in which the first DOI identifies an article containing a reference to the article identified by the second DOI. A date flag corresponding to the issue date of the citing article has been associated to each couple of DOIs in the list to represent the citation date. Such information has been obtained from the "Article metadata" part of the database which is divided by journal and provides for each paper the following fields: DOI, journal, volume, issue, first page and last page or article id and number of pages, title, authors, affiliations, publication history, PACS codes, table of contents, heading, article type, and copyright information. The list of authors provided for each DOI has been used to generate the collaboration network where authors of the same paper form small coauthorship cliques. Starting from the metadata a list of 17,069,841 total coauthorships has been generated for 119,172 unique authors' surnames. In order to assign a date to the collaboration the submission date of the coauthored article has been associated to each couple of authors. The data transformation is performed through the *time-varying graphs* (TVG) formalism. The *time-varying graph* (TVG) formalism, introduced in (Casteigts et al. 2010), is a graph formalism based on an *interaction-centric* point of view and offers concise and elegant formulation of temporal concepts and properties (Santoro et al. 2010). Let us consider a set of entities V (or *nodes*), a set of relations E among entities (*edges*), and an alphabet L labeling any property of a relation (*label*), that is, $E \subseteq V \times V \times L$. The set E enables multiple relations between any given pair of entities, as long as these relations have different properties, that is, for any $e_1 = (x_1, y_1, \lambda_1) \in E$, $e_2 = (x_2, y_2, \lambda_2) \in E, (x_1 = x_2 \wedge y_1 = y_2 \wedge \lambda_1 = \lambda_2) \Rightarrow e_1 = e_2$. Relationships between entities are assumed to occur over a time span $\mathbb{T} \subseteq \mathbb{T}$, namely the *lifetime* of the system. The temporal domain $\mathbb{T}$ is assumed to be $\mathbb{N}$ for discrete-time systems or $\mathbb{R}$ for continuous-time systems. The time-varying graph structure is denoted by the set $\mathcal{G} = (V, E, \mathbb{T}, \rho, \zeta)$, where $\rho : E \times \mathbb{T} \rightarrow \{0, 1\}$, called *presence function*, indicates whether a given edge is present at a given time, and $\zeta : E \times \mathbb{T} \rightarrow \mathbb{T}$, called *latency function,* indicates the time it takes to cross a given edge if starting at a given date. As in this paper the focus is on the temporal and structural analysis of a social network, we will deliberately omit the latency function and consider TVGs described as $\mathcal{G} = (V, E, \mathbb{T}, \rho)$. Given a TVG $\mathcal{G} = (V, E, \mathbb{T}, \rho)$, one can define the *footprint* of this graph from t_1 to t_2 as the static graph $G^{[t_1, t_2)} = (VE^{[t_1, t_2)})$ such that $\forall e \in E, e \in E^{[t_1, t_2)} \Leftrightarrow \exists t \in [t_1, t_2), \rho(e, t) = 1$. In other words, the footprint aggregates interactions over a given time window into static graphs. Let the lifetime $\mathbb{T}$ of the time-varying graph be partitioned in consecutive subintervals $\tau = [t_0, t_1), [t_1, t_2) \ldots [t_i, t_{i+1}), \ldots$, where each $[t_k, t_{k+1})$ can be noted τ_k. We call *sequence of footprints* of $\mathcal{G}$ according to τ the sequence $SF(\tau) = G^{\tau 0}, G^{\tau 1}, \ldots$.

Hence, we derive two time-varying graphs: the *temporal coauthorships network*, with undirected edges and authors as nodes where a link stands for the relations of coauthoring a paper, and the *temporal citations network* having papers as nodes and the links (directed) representing the citations from a paper to another one. The temporal dimension of both networks is derived by the paper's submission date. The temporal coauthorship network has edges labeled with the date of submission, while the temporal citations network has the nodes labeled with the publication date of papers citing other papers.

More formally, we can define this system as two networks:

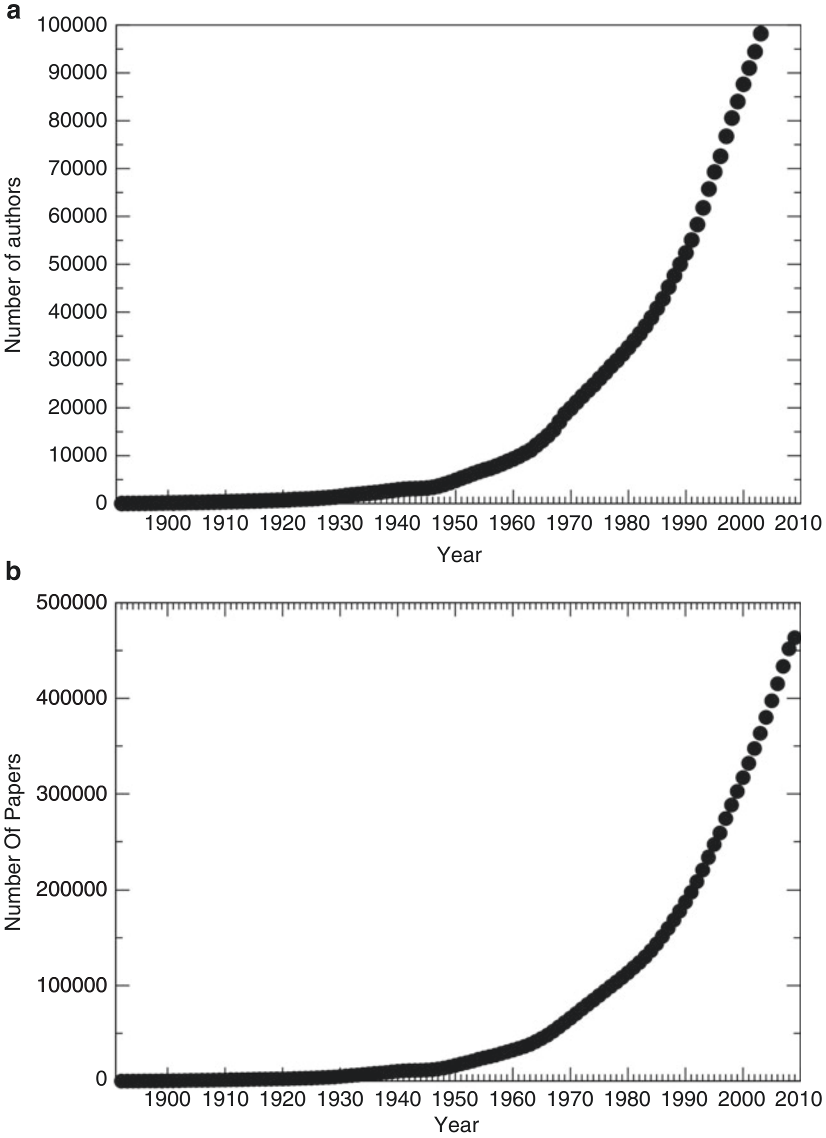
- The **temporal coauthorships network** as a quadruplet $G_a^t : (V, E, \mathbb{T}, \rho)$ where the nodes in $v \in V$ are the authors and links $e \in E$ connect a couple of scientists coauthoring a paper. The temporal domain $\mathbb{T} = [t_a, t_b)$ of the function ρ is the *lifetime* of each node v that in this context is assumed as t_a to be the submission date of the paper and $t_b = \infty$.
- The **temporal citations network** as a quadruplet $G_c^t : (V, E, \mathbb{T}, \rho)$ where the nodes in the set V are the papers and each edge $e \in E$ corresponds to a citation to another paper. As for the coauthorships network, the temporal dimension $\mathbb{T} = [t_a, t_b)$ of the presence function p of GC is defined within the submission date of papers and ∞.

Networks Evolution

In Fig. 1 we show the number of authors and the number of papers for each year. One can observe from such figures an exponential growth of both the number of authors and of papers along time. Such results are not surprising and have been highlighted by several former works (for instance in (Radicchi et al. 2009)). The exponential growth in the number of publications is more or less directly attributed to a change in the behavior of scientists induced by the pressure to publish all along their career (it has been popularized through the proverb publish or perish). The exponential growth of the number authors is more surprising at a first attempt, as it does not translate an exponential increase of the positions in research that does not exist. It is much more seriously explained by an indirect effect of the exponential growth of publications. We have to remind that this dataset concerns the APS publications and such publications do not render the effective number of physicists. As the popularity of the APS journals increased, they probably attract more and more physicists worldwide and we can expect a stabilization of such tendency once as the APS will tend to reference nearly the whole population of physicists worldwide.

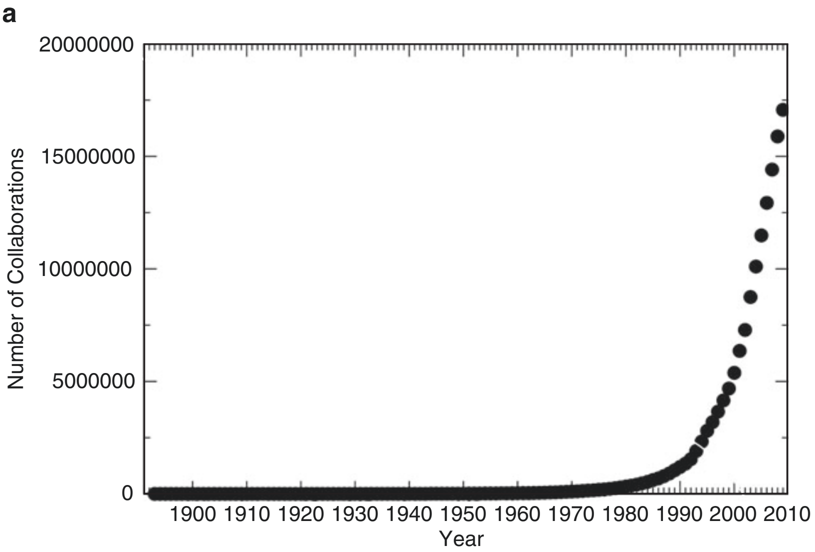
In Fig. 2 we show the number of collaborations within authors and the number of citations within papers. Those two measures correspond basically to the number of edges in each of the two networks. The first important element concerns the increase of the number of collaborations that scales as a power law rather than an exponential. This feature results clearly of a double effect over the past few years. The first one is directly linked to the increase in the number of papers that increases the potential of collaboration among authors. The other effect comes from the progressive increase of the number of authors per paper. Translated into network terms, it means that each paper coauthored by N scientist creates N*(N − 1)/2 links in the collaboration network, therefore if you follow the current tendency to increase the number of authors, your increase with a power coefficient, the number of links among them. Concerning the other figure, the exponential growth is probably less essential but again it results from two combined effects. On the one hand, the increase in the number of papers published increase in the same way the total number of citations. On the other hand, the slight tendency to progressively increase the number of papers cited in each paper straightened again the slope. Considering the two graphs mentioned in Fig. 1, the basic feature that we can observe is a global tendency of the increase of the number of nodes in the corresponding networks. The point that the number of links on each graph increases more rapidly than the number of nodes lead to the conclusion that the coauthorship and the citation graphs tend to grow and to densify as well. However, we don’t have any clue concerning the properties of such a density growth; mainly is it an egalitarian growth or it is an elitist system with some few nodes benefiting from this increase in density and the majority of nodes being left behind without many links. The measure of the evolution of the clustering coefficient on such networks can bring arguments for this distinction.

In Fig. 3 we show the clustering coefficient, i.e., the transitivity among nodes, for the collaboration and citation network. Qualitatively, the curves are totally different on the two networks. Whereas on the coauthorship network, the evolution follows



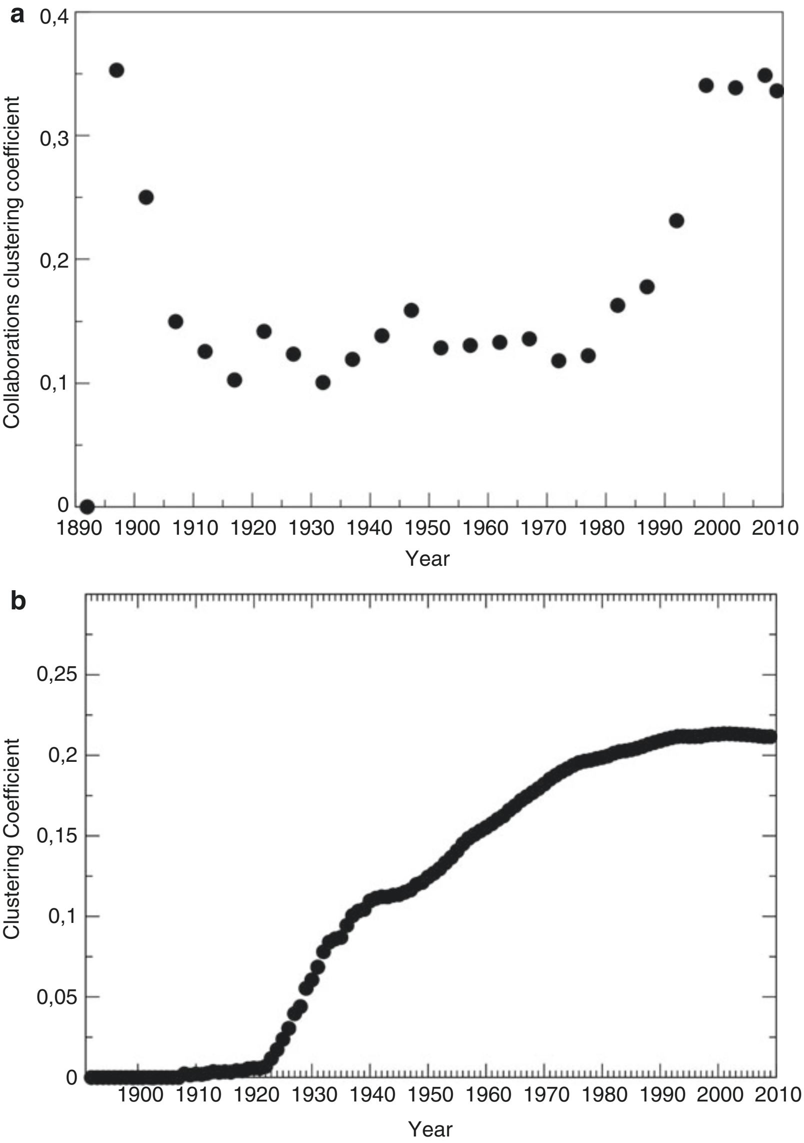
Stability and Evolution of Scientific Networks, Fig. 1 Number of authors and number of papers

first an important decrease then stabilized before than to increase again; on the citation network, we can observe an increase that tends to stabilize in the last 20 years. The elements of interpretation behind those two figures are the following. From the coauthorship network, the first global decrease can be explained mainly because it starts from an important number of nonconnected components in the network. Therefore, the creation of new links among those components (or communities) that corresponds to a porosity of the different communities in physics results in a global decrease of the clustering coefficient as it tends to dissolve locally the density of each component. Once a global giant component is created (corresponding to the observed plate on the figure), then there is a stabilization of the clustering coefficient. The final increase is maybe the most interesting feature of this analysis as it corresponds to the case where, in a global single component, the clustering is



Stability and Evolution of Scientific Networks, Fig. 2 Number of collaborations and number of citations

increasing. This is the case where communities tend to emerge from a global network. Therefore this last increase could be interpreted as the formation and the radicalization of scientific communities on the global network. These communities correspond to the effective work in the scientific communities, i.e., coauthorship. Concerning the evolution of the clustering coefficient in the citation network, the first observation we have to make is that the global big component appears very soon on this network (this is much more probable to cite works from outside the field than to collaborate with people from outside the field). Therefore this global and progressive increase of the clustering coefficient corresponds solely to the progressive formation of scientific communities on the network. However, the final stabilization of the index results from a consolidation of the communities that have reached a relative equilibrium. We have to notice that in the case of the creation of new communities or emerging fields, and then we could see the global clustering coefficient increase again. Such



Stability and Evolution of Scientific Networks, Fig. 3 The evolution of the clustering coefficient

an observation can be made on the figure where around 1940, we can observe a global stabilization of the index and therefore of the corresponding communities before than to increase again, such a new burst being the result of the inclusion of new communities in the network. However, in order to relativize such an effect, we have to remind that the dataset we analyze corresponds to the publications of the APS and such an inclusion of new communities can result simply from an editorial choice corresponding to the launch of new journals on new thematics for the APS, but not necessarily for the scientific domain of physics.

Conclusions

Scientific networks represent a wealth of knowledge for the study of the relationships that occur between science-related entities like single scientists, groups, journals, and related institutions. Network analysis and modeling has been widely used for observing collaboration and citation patterns of scientists and papers, discovering scientific communities, how these evolve in time and space, and how they impact scientific productivity and profitability. Collaboration networks can give insights for the improvement of the quality of the research topics, show how coauthorship patterns shape field developments and which are the leading scientists and institutions eligible for fundings.

In this paper we characterize the evolution of a scientific community extracted by the APS dataset. The temporal dimension and the metrics used for the analysis were formalized using Time-Varying Graphs (TVG), a mathematical framework designed to represent the interactions and their evolution in dynamically changing environments.

Since we are interested in the relationships between collaborations and citations behaviors of scientists, we focus on the network of most cited authors and on its structural evolution where several interesting aspects emerge. Through our approach, we capture the role played by famous authors on coauthorship behaviors. They act as attractors on the community. The driving force is a sort of preferential attachment driven by the number of citations received by a given group, which in terms of the goal of any scientific community indicates a strategy oriented to the community belonging.

Furthermore, the evolution of the network from a sparse and modular structure to a denser and homogeneous one can be interpreted as a threefold process reflecting the natural selection. The first phase is the exploration of ideas by means of separated works; once some ideas start to be cited (selected) more than others, then authors tend to join groups that have produced highly cited works. The selection is performed by individuals in a goal-oriented environment and such a (social) selection produces self-organization, because it is played by a group of individuals which act, compete, and collaborate in order to advance Science. In fact, the driving force is an emergent effect of the interdependencies between citations and the goal of the scientific production since the social selection determines the emergence of a topic and of the scientists working on it by determining the so-called preferential attachment toward groups and topics having high potential of citations.

Cross-References

- ▶ Analysis and Visualization of Dynamic Networks
- ▶ Community Evolution
- ▶ Dynamic Community Detection
- ▶ Modeling and Analysis of Spatiotemporal Social Networks
- ▶ Spectral Evolution of Social Networks

Acknowledgements Thanks to the American Physical Society for allowing us to use the APS dataset.

References

Amblard F, Casteigts A, Flocchini P, Quattrociocchi W, Santoro N (2011) On the temporal analysis of scientific network evolution. In: Computational aspects of social networks (CASoN), 2011 international conference on. pp 169–174. IEEE, Salamanca, Spain

Barabasi AL, Jeong H, Neda Z, Ravasz E, Schubert A, Vicsek T (2002) Evolution of the social network of scientific collaborations. Physica A 311:590–614

Carter TB (2007) Structural change and homeostasis in organizations: a decision-theoretic approach. J Math Sociol 31(4):295–321

Casteigts A, Flocchini P, Quattrociocchi W, Santoro N (2010) Time-varying graphs and dynamic networks. Technical Report University of Carleton, Ottawa

Davidsen J, Ebel H, Bornholdt S (2002) Emergence of a small world from local interactions: modeling acquaintance networks. Phys Rev Lett 88(12):128701

De Solla Price DJ (1965) Networks of scientific papers. Science 149(3683):510–515

Deffuant G, Neau D, Amblard F, Weisbuch G (2001) Mixing beliefs among interacting agents. Adv Complex Syst 3:87–98

de e Fonseca BPF, Sampaio RB, de Fonseca MVA, Zicker F (2016) Co-authorship network analysis in health research: method and potential use. Health Res Policy Syst 14(1):34

Gilbert N (1977) Referencing as persuasion. Soc Stud Sci 7:113–122

Guimera R, Uzzi B, Spiro J, Amaral LA (2005) Team assembly mechanisms determine collaboration network structure and team performance. Science 308 (5722):697–702

Holme P (2005) Network reachability of real-world contact sequences. Phys Rev E 71(4):46119

Hummon NP, Dereian P (1989) Connectivity in a citation network: the development of DNA theory. Soc Networks 11(1):39–63

Klemm K, Eguíluz VM (2002) Highly clustered scale-free networks. Phys Rev E 65(3):036123

Kossinets G, Kleinberg J, Watts D (2008) The structure of information pathways in a social communication network. In Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2008), pp 435–443. Paris, France

Kostakos V (2009) Temporal graphs. Physica A 388 (6):1007–1023

Lelarge M (2008) Diffusion of innovations on random networks: understanding the chasm. Intern Netw Econ:178–185

Leskovec J, Kleinberg JM, Faloutsos C (2007) Graph evolution: densification and shrinking diameters. TKDD 1(1):2–1

MacRoberts MH, MacRoberts BR (1996) Problems of citation analysis. Scientometrics 36(3):435–444

Mataric M (1992) Designing emergent behaviors: from local interactions to collective intelligence. In: Proceedings of the international conference on simulation of adaptive behavior: from animals to animats, vol 2. pp 432–441. Honolulu, HI

Moore C, Newman MEJ (2000) Epidemics and percolation in small-world networks. Phys Rev E 61: 5678–5682

Newman MEJ (2001a) The structure of scientific collaboration networks. Proc Natl Acad Sci 98(2):404–409

Newman MEJ (2001b) Clustering and preferential attachment in growing networks. Phys Rev E 64:025102

Newman MEJ (2004a) Coauthorship networks and patterns of scientific collaboration. Proc Natl Acad Sci 101:5200–5205

Newman MEJ (2004b) Who is the best connected scientist? A study of scientific coauthorship networks. Complex Netw 650:337

Pan RK, Kaski K, Fortunato S (2012) World citation and collaboration networks: uncovering the role of geography in science. arXiv preprint arXiv 1209:0781

Peltomaki M, Alava M (2006) Correlations in bipartite collaboration networks. J Stat Mech 2006:P01010

Powell WW, White DR, Koput KW (2005) Network dynamics and field evolution: the growth of interorganizational collaboration in the life sciences. Am J Sociol 110(4):1132–1205

Quattrociocchi W, Paolucci M, Conte R (2009) On the effects of informational cheating on social evaluations: image and reputation through gossip. Int J Knowl Learn 5(5/6):457–471

Quattrociocchi W, Conte R, Lodi E (2010) Simulating opinion dynamics in heterogeneous communication systems. In: ECCS 2010 – Lisbon Portugal

Quattrociocchi W, Amblard F, Galeota E (2012) Selection in scientific networks. Soc Netw Anal Min 2(3): 229–237

Radicchi F, Fortunato S, Markiness B, Vespignani A (2009) Diffusion of scientific credits and the ranking of scientists. Phys Rev E 80

Redner S (2005) Citation statistics from 110 years of physical review. Phys Rev Phys Today 58:49–54

Santoro N, Quattrociocchi W, Flocchini P, Casteigts A, Amblard F (2010) Time varying graphs and social network analysis: temporal indicators and metrics. Technical Report University of Carleton, Ottawa

Santoro N, Quattrociocchi W, Flocchini P, Casteigts A, Amblard F (2011) Time-varying graphs and social network analysis: temporal indicators and metrics. In: 3rd AISB social networks and multiagent systems symposium (SNA-MAS). pp 32–38. University of York, UK

Taramasco C, Cointet J-P, Roth C (2010) Academic team formation as evolving hypergraphs. Scientometrics

Wagner CS, Leydesdorff K (2005) Network structure, self-organization, and the growth of international collaboration in science. Res Policy 34(10):1608–1618

Wang D, Yan K-K, Rozowsky J, Pan E, Gerstein M (2016) Temporal dynamics of collaborative networks in large scientific consortia. Trends Genet 32(5):251–253

Woolley DR (1994) Plato: the emergence of online community. Comput-Med Commun Mag 1(3):5

Stance Detection

▶ Social Media Analysis for Monitoring Political Sentiment

Statistical Analysis

▶ Assessing Individual and Group Behavior from Mobility Data: Technological Advances and Emerging Applications

Statistical Inference

▶ Theory of Statistics: Basics and Fundamentals

Statistical Modeling

▶ Siena: Statistical Modeling of Longitudinal Network Data

Statistical Models

▶ Theory of Statistics: Basics and Fundamentals

Statistical Relational Learning

▶ Probabilistic Logic and Relational Models
▶ Relational Models

Statistical Relational Models

▶ Relational Models

Statistical Research in Networks: Looking Forward

Eric D. Kolaczyk
Department of Mathematics and Statistics, Boston University, Boston, MA, USA

Synonyms

Propagation of uncertainty; Research challenges

Glossary

Network summary statistic	A statistic summarizing a network graph
Propagation of uncertainty	Understanding the effect of uncertainty in an initial set of measurements on functions thereof

Introduction

The emerging field of network analysis, through its roots in social network analysis, has had a nontrivial statistical component from the start. In the ensuing years, problems in network analysis have motivated – and continue to motivate – new research in the field of statistics. Conversely, new developments in statistics are routinely integrated into network research. It is therefore rather surprising that despite the many interesting and important statistical challenges in network analysis to which researchers have already been able to respond, there, nevertheless, are a number of challenges of an entirely fundamental nature that remain almost untouched!

We will support this central claim through two examples. Additional examples will be mentioned in passing at the end. All of these examples relate to the basic problem of *propagation of uncertainty* – understanding the effect of uncertainty in an initial set of measurements on functions thereof.

Network construction is often a sophisticated (and frequently complicated) process. Importantly, the network graph $G = (V, E)$ that is considered to be "observed" typically is the result of taking some sort of basic measurements and then going through a series of steps (whether formal or informal) before arriving at a set of vertices V and edges E. Accordingly, to the extent that there is uncertainty in the basic underlying measurements, there will be uncertainty in the corresponding edges (and perhaps vertices as well, depending on context). In turn, therefore, this uncertainty will impact any further processing of G.

The examples of such "processing" that we will consider are (i) network summary statistics and (ii) network modeling. Before doing so, however, the following motivating illustration from the context of Internet traffic data analysis will be useful.

S

Illustration: Network Traffic Graphs

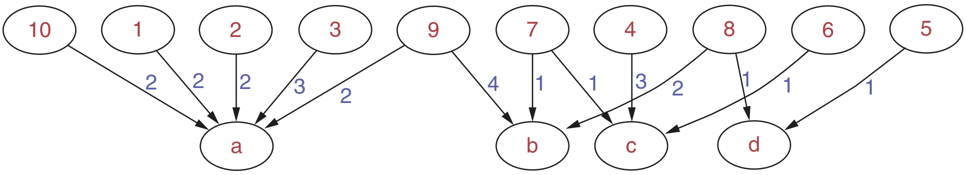
Consider the recent trend toward using social network principles and analysis in the study of Internet traffic flow data. Each time an actor, working from an Internet-capable device (e.g., a smartphone), uses an Internet application (e.g., an iPhone or Android app), traffic is generated in the form of packets of information that are exchanged between the device and the relevant Internet destination(s). The collection of such packets relevant to a given basic task (e.g., opening a Web page in a browser or downloading a new song) is termed a *flow*. It is possible, using measurement technology deployed within the physical layer of the Internet (typically at Internet routing devices), to capture information on such flows. Researchers then use bipartite graphs to represent these flows, with vertices corresponding to origins (e.g., the IP address of an actor's iPhone) and destinations (e.g., the IP address of a server hosting a Web page) and edges indicating that a flow was declared to have passed between the two. A small graph of this sort is shown in Fig. 1.

Social network tools are then used to study these typically massive networks, or variations thereof. For example, beginning with the bipartite representation just described, Ding et al. (2012) construct networks $G = (V; E)$ of actors $v \in V$ (formally, IP addresses understood to correspond to users) that have an edge $e = (u, v) \in E$ between them when there is least one Web server with which both u and v exchanged flows, indicating some level of common behavior. (Formally, these authors construct a one-mode projection of the bipartite network traffic graph). An example may be found in Fig. 2. Exploiting data on IP addresses known to have exhibited malicious behavior during the same time period, they find that the corresponding vertices, say $v^* \in V$, in the graph G tend to be overrepresented in regions of G falling *between* natural communities (e.g., as bridge nodes between communities). Thus, these nodes demonstrate a curious (anti) social behavior.

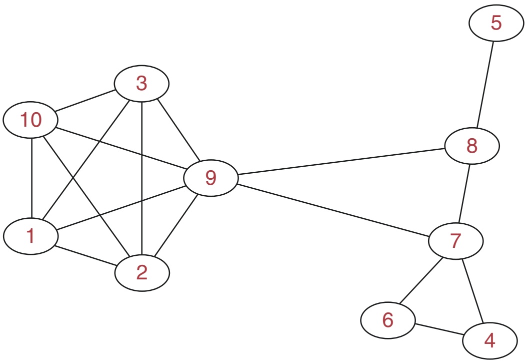
Ding et al. (2012) use this observation to create an anomaly detection strategy for finding IP addresses participating in malicious behavior. Similarly, other authors have used networks of this type (termed "traffic activity graphs" or "traffic dispersion graphs") for a variety of purposes, ranging from characterization to detection. See, for example, Jiang et al. (2010), Jin et al. (2009), Iliofotou et al. (2009), and Iliofotou et al. (2007).

Notwithstanding such successes, it is important to note that there are various sources of uncertainty in the basic measurements underlying these networks, and hence in the networks themselves. Most fundamentally, Internet routers typically do not keep record of every packet to pass through them. Rather, they record only some fraction of those packets (e.g., 1 in 1000), through either random or deterministic sampling. Hence, some flows will not be observed at all, corresponding to missing edges in the initial bipartite graph representation. In addition, again due to sampling, a single flow initially may be recorded as multiple flows. As a result, what should be a single edge in the bipartite graph ends up being represented as multiple edges. Postprocessing typically is done to ameliorate this latter effect of sampling, but cannot be expected to succeed completely.

The extent to which such low-level sources of uncertainty in the initial underlying flow measurements affect high-level Internet traffic analysis tasks, like characterization and anomaly detection, appears not to have been studied systematically to date. However, there has been promising work understanding and statistically correcting for sampling artifacts at the level of flow summary statistics (e.g., distributions of flow counts and lengths) and queries thereof. See, for example, Duffield et al. (2005a, b) and Cohen et al. (2008). For a general overview of Internet traffic packet sampling and related issues, see Duffield (2004).



Statistical Research in Networks: Looking Forward, Fig. 1 A bipartite representation of Internet traffic flow measurements, from ten sources (i.e., *1* through *10*) to four destinations (i.e., *a* through *d*).



Statistical Research in Networks: Looking Forward, Fig. 2 One-mode projection of the bipartite network in Fig. 1

Propagation of Uncertainty in Network Analysis

As noted earlier, the issues just described pertain to the problem of propagation of uncertainty – from the initial measurements to the network graph G to any further processing of G. Of course, the statistical analysis of network data – and, in particular, attempting to account for uncertainty inherent to the data – is by no means new. See Kolaczyk (2009), for example, for a recent overview of statistical methods and models in network analysis. However, there remains much to be done and, surprisingly, some of it of a particularly fundamental nature, that is, "fundamental" in the sense that any student of a one-semester elementary statistics course can be expected to have tools for doing analogous tasks with classical data (i.e., independent and identically distributed observations). Yet we lack these same tools in the context of network analysis. We describe two examples in detail below.

Uncertainty in Network Summary Statistics

A standard paradigm in network analysis goes as follows. Data are obtained from a complex system of interest, a network graph G is constructed, and summary statistics of G, say $\eta(G)$, are reported (e.g., degree distribution, clustering coefficient, and various measures of centrality). When G itself is the primary object of interest, this is a sensible paradigm. However, when in reality we have only a "noisy" version of G, say G^*, then the statistic we calculate, that is, $\eta(G^*)$, is only a noisy version of $\eta(G)$. In that case, interpreting $\eta(G^*)$ as a point estimate of $\eta(G)$, it is natural to wish to equip this estimate with some quantification of its inherent uncertainty, such as a standard error or, more ambitiously, confidence intervals.

To date, there is little in the way of general statistical methodology for this problem. Some progress has been had for (a) certain sources of uncertainty and (b) certain summary statistics $\eta(\cdot)$. Frank and colleagues have, for example, established standard errors for statistics in the form of dyad and triad sums, when G^* is obtained from G through specific sampling plans, such as induced subgraph sampling. See Frank (2004) for an overview of such results.

More generally, Viles (2013) has recently established the limiting distribution of statistics $\eta(\cdot)$ that take the form of sums of configurations of G^*, such as dyads and triads, under a general measurement error model, in the limit as the number of vertices tends to infinity. Intriguingly, rather than the seemingly ubiquitous standard normal distribution, this limiting distribution is a so-called Skellam distribution – the difference of two independent Poisson random variables.

To see intuitively why this distribution might arise, consider the case where $\eta(\cdot)$ simply counts the number of edges in its argument. The error in $\eta(G^*)$ in estimating the true $\eta(G)$ will be (proportional to) the difference of (i) the total number of edges in G^* that are false (i.e., not in G) and (ii) the total number of nonedges in G^* that are false (i.e., in G). These totals are each sum of binary random variables and hence might be expected to possess characteristics of a Poisson distribution, under appropriate conditions. That this is the case, however, is nontrivial to demonstrate, since the binary variables are dependent. The arguments in Viles (2013) make use of Stein's method, typically part of the toolset learned in advanced graduate statistics courses.

These preliminary results suggest that the problem of propagating uncertainty to network summary statistics, although fundamental, has a complexity associated with it that goes beyond that of, say, a simple sample mean as encountered in "Statistics 101." More general results will likely depend on the smoothness of $\eta(\cdot)$, with respect to changes in G^* and the measurement error involved in obtaining G^*, as well as characteristics of the true underlying G itself.

Formal statistical arguments establishing such results will require techniques beyond those employed in the classical setting and quite possibly the development of new techniques altogether.

Uncertainty in Network Modeling

While the above paradigm, in which we conceptualize ourselves as having observed a "noisy" version G^* of a "true" graph G, is appropriate for many contexts, another useful perspective is that in which we think of G^* as having derived from some network distribution, $\mathbb{P}(G)$. This perspective underlies, for example, the large body of work in social network analysis using exponential random graph models (ERGMs).

With a history going back roughly 30 years, ERGMs have become a mainstay of social network analysis. See the edited volume by Luscher et al. (2012), for example, for a recent overview. This class of models specifies that the distribution of the adjacency matrix, say Y, for a random graph G follows an exponential family form, that is, $p^{\theta}(Y = y) \propto \exp(\theta^T g(y))$, for vectors θ of parameters and $g(\cdot)$ of sufficient statistics. However, despite this seemingly appealing feature, work in the last 5 years has shown that exponential random graph models must be handled with some care, as both their theoretical properties and computational tractability can be rather sensitive to model specification. See Robins et al. (2007), for example, and Chatterjee and Diaconis (2011) for a more theoretical treatment. Benefiting from these findings, software is now available for dependably fitting well-posed ERGM models and typically estimates of model parameters θ are accompanied by standard errors, where the latter are based on standard arguments for exponential families.

Unfortunately, while such standard errors are perhaps useful in summarizing relative levels of uncertainty associated with estimates of the parameters in θ, there is to date no general theory supporting their use in creating confidence intervals or performing tests. Even more unfortunate is that the importance of this fact does not appear to be universally appreciated, since it is not unheard of to find applied papers in social network analysis in which ERGM parameter estimates are cited with what are purported to be confidence intervals or results of significance tests! Contrast this situation with that of, say, linear regression analysis, for which a "Statistics 101" student would, by the end of a single semester, typically have been exposed to methods for confidence intervals and tests of all sorts in the classical setting.

In recent work, Kolaczyk and Krivitsky (2013) have demonstrated that the asymptotic analysis necessary to establish a parallel theory for ERGMs likely will be rather more subtle. These authors concentrate on the simplest of ERGM models, in which dyads y^{ij} , y_{ji} are independent, and focus on a comparison of the cases of sparse versus nonsparse networks. (We consider a network graph G to be sparse if the number of edges is of the order of the number of vertices, that is, $N_e = O(N_v)$, rather than the square of that number, that is, $N_e = O\left(N_v^2\right)$. In that setting, they demonstrate that the very order of the asymptotics will depend critically on the sparseness of G. More specifically, they show that the maximum likelihood estimates of the ERGM parameters for attraction and mutuality will converge asymptotically to a bivariate normal distribution in both sparse and nonsparse cases but at rates $N_v^{1/2}$ and N_v, respectively.

At an intuitive level, these results say that the nature of the dependency in the relational measurements y_{ij} leads to variations in the effective sample size. For nonsparse networks, we have effectively $O\left(N_v^2\right)$ measurements – in fact, the same number as entries in the adjacency matrix Y. But for sparse networks, we have effectively only N_v measurements! Since the effective sample size drives the relative magnitude of the standard error, as a function of network order, it is a critical factor in establishing asymptotic results justifying confidence intervals and tests based on the latter.

While the results of Kolaczyk and Krivitsky (2013) use tools from, say, the latter part of a first-year course in theoretical statistics (i.e., stochastic convergence of estimating equations coupled with a double-array central limit theorem), it is likely that additional traction on this problem for ERGM models with more complex forms of dependency (e.g., stars, triadic structure) will require the development of new tools.

Future Directions

In looking forward at the challenges facing us for statistical research in relation to social network analysis, there is a curious feeling of looking back as well, that is, "looking back" in the sense that we realize there is much to be done in this context when we note what has already been done previously in more established contexts. Much that is now considered foundational, in that it is part of a now-standard toolset.

Certainly, the classical case of independent and identically distributed observations forms the gold standard. As noted earlier, much of the foundational material in the classical setting – such as confidence intervals for summary statistics and regression model parameters – forms part of the core of what is presented to students at the very earliest stage of statistics education. Social network analysis currently lacks a number of such foundational aspects. Yet it can arguably take hope from the example of time series analysis and spatial data analysis. In both cases, the data deviate from the classical case in that they are dependent. And, moreover, in both cases, over time, the analogous foundations were laid. See, for example, the books Brockwell and Davis (2009) and Cressie (1993) for time series and spatial analysis, respectively.

It can be expected that new tools and techniques will be required from statisticians to fill the gaps in the foundations for network analysis. Networks share dependency with time series and spatial data but lack the temporal and geometric aspects of the latter (More formally, they lack the properties of Euclidean space that can be exploited in the context of time series and spatial data.). These aspects were critical for successfully extending classical results to time series and spatial data, due to the fact that they facilitate a notion of "local" dependence (e.g., local in time or in space) that emerges naturally as a "loosening" of the stricter assumption of independent and identically distributed. Developing and working with analogous notions of "localness" in the network context is a key hurdle to be faced.

In terms of specific areas requiring work, from the descriptions above, it should be clear that progress is only just starting to be made on the two examples cited. For confidence intervals for summary statistics, there is much to be explored in understanding the interaction between (a) characteristics of the sources of uncertainty (e.g., based on sampling, measurement, missingness), on the one hand, and (b) the nature of the summary statistics to be computed (e.g., smooth or unsmooth, in an appropriate sense), on the other hand. Furthermore, establishing limiting parametric distributions (such as the Skellam distribution in Viles (2013)) is one key way to facilitate the construction of confidence intervals; it would be useful to see a version of bootstrapping or related resampling approaches justified in the context of networks.

Similarly, asymptotic theory supporting methods for the construction of confidence intervals for network parameters is only beginning to emerge. The most traction appears to have been gained in the context of stochastic block models (e.g., Bickel and Chen 2009; Choi et al. 2010; Celisse et al. 2011; Rohe et al. 2011), although progress is beginning to be had with exponential random graph models as well (e.g., Chatterjee et al. 2011; Chatterjee and Diaconis 2011; Rinaldo et al. 2013). Most of these works present consistency results for maximum likelihood and related estimators, with the exception of Bickel and Chen (2009), which also includes results on asymptotic normality of estimators. See Haberman (1981) for another contribution in this direction, proposed as part of the discussion of the original paper of Holland and Leinhardt (1981). Finally, for some initial (nonasymptotic) results in the context of more mathematical models (e.g., preferential attachment, copying), see Wiuf et al. (2006).

While there are various other similarly important statistical topics that remain to be explored in network analysis, arguably one of the most pressing of those is that of missing data. It is known, again in the classical setting first and foremost, that depending on the mechanism of missingness, the impact of missing data on statistical inference can range anywhere from mild to devastating. See Kolaczyk (2009, Chap. 3) for some general discussion, including comparisons to the importance of missingness and related notions in the context of spatial data analysis. A general framework for thinking about the impact of missingness on network modeling has recently been initiated in Handcock and Gile (2010) and Jiang and Kolaczyk (2012) have recently demonstrated that accounting for observation errors that include missingness (using a hierarchical modeling formulation) can lead to marked improvement in accuracy of link prediction. But, as with the other areas cited above, much remains to be done to explore and develop the necessary statistical infrastructure for understanding and dealing with missingness generally in network contexts.

Cross-References

- ▶ Exponential Random Graph Models
- ▶ Learning Networks
- ▶ Network Models
- ▶ Network Representations of Complex Data
- ▶ Probabilistic Analysis
- ▶ Sampling Effects in Social Network Analysis
- ▶ Sources of Network Data
- ▶ Theory of Statistics: Basics and Fundamentals

References

Bickel PJ, Chen A (2009) A nonparametric view of network models and Newman–Girvan and other modularities. Proc Natl Acad Sci 106(50):21068

Brockwell PJ, Davis RA (2009) Time series: theory and methods. Springer, New Delhi

Celisse A, Daudin JJ, Pierre L (2011) Consistency of maximum-likelihood and variational estimators in the stochastic block model. Arxiv preprint arXiv: 1105.3288

Chatterjee S, Diaconis P (2011) Estimating and understanding exponential random graph models. Arxiv preprint arXiv: 1102.2650

Chatterjee S, Diaconis P, Sly A (2011) Random graphs with a given degree sequence. Ann Appl Probab 21 (4):1400–1435

Choi DS, Wolfe PJ, Airoldi EM (2010) Stochastic blockmodels with growing number of classes. Arxiv preprint arXiv: 1011.4644

Cohen E, Duffield N, Lund C, Thorup M (2008) Confident estimation for multistage measurement sampling and aggregation. ACM SIGMETRICS Per Eval Rev 36:109–120

Cressie NAC (1993) Statistics for spatial data, revised edn. Wiley, New York

Ding Q, Katenka N, Barford P, Kolaczyk ED, Crovella M (2012) Intrusion as (anti)social communication: Characterization and detection. In: Proceedings of the 18th ACM SIGKDD conference on knowledge discovery and data mining. ACM, Beijing, pp 886–894

Duffield N (2004) Sampling for passive internet measurement: a review. Stat Sci 19(3):472–498

Duffield N, Lund C, Thorup M (2005a) Estimating flow distributions from sampled flow statistics. IEEE/ACM Trans Netw 13(5):933–946

Duffield N, Lund C, Thorup M (2005b) Optimal combination of sampled network measurements. In: Proceedings of the 5th ACM SIGCOMM conference on internet measurement. USENIX Association, Berkeleye, pp 8–8

Frank O (2004) Network sampling and model fitting. In: Carrington PJ, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, New York

Haberman SJ (1981) An exponential family of probability distributions for directed graphs: comment. J Am Stat Assoc 76(373):60–61

Handcock MS, Gile KJ (2010) Modeling social networks from sampled data. Ann Appl Stat 4(1):5–25

Holland PW, Leinhardt S (1981) An exponential family of probability distributions for directed graphs. J Am Stat Assoc 76:33–50

Iliofotou M, Pappu P, Faloutsos M, Mitzenmacher M, Singh S, Varghese G (2007) Network monitoring using traffic dispersion graphs (tdgs). In: Proceedings of the 7th ACM SIGCOMM conference on internet measurement. ACM, San Diego, pp 315–320

Iliofotou M, Faloutsos M, Mitzenmacher M (2009) Exploiting dynamicity in graph-based traffic analysis: techniques and applications. In: Proceedings of the 5th international conference on emerging networking experiments and technologies. ACM, Rome, pp 241–252

Jiang X, Kolaczyk ED (2012) A latent eigenprobit model with link uncertainty for prediction of protein–protein interactions. Stat Biosci 4(1):84–104

Jiang N, Cao J, Jin Y, Li LE, Zhang ZL (2010) Identifying suspicious activities through DNS failure graph analysis. In: 18th IEEE international conference on network protocols (ICNP) 2010. IEEE, Kyoto, pp 144–153

Jin Y, Sharafuddin E, Zhang ZL (2009) Unveiling core network-wide communication patterns through application traffic activity graph decomposition. In: Proceedings of the 11th international joint conference on measurement and modeling of computer systems. ACM, Seattle, pp 49–60

Kolaczyk ED (2009) Statistical analysis of network data: methods and models. Springer, New York/London

Kolaczyk ED, Krivitsky PN (2013) On the question of effective sample size in network modeling. Stat Sci (under invited revision)

Luscher D, Koskinens J, Robins G (2012) Exponential random graph models for social networks: Theory, methods, and applications. Cambridge University Press, Cambridge

Rinaldo A, Petrovic S, Fienberg SE (2013) Maximum likelihood estimation in the beta model. Ann Stat (to appear)

Robins G, Snijders T, Wang P, Handcock M, Pattison P (2007) Recent developments in exponential random graph (p*) models for social networks. Soc Networks 29(2):192–215

Rohe K, Chatterjee S, Yu B (2011) Spectral clustering and the high-dimensional stochastic blockmodel. Ann Stat 39(4):1878–1915

Viles WE (2013) Uncertainty propagation from network inference to network characterization. Ph.D. thesis, Department of Mathematics & Statistics, Boston University

Wiuf C, Brameier M, Hagberg O, Stumpf MPH (2006) A likelihood approach to analysis of network data. Proc Natl Acad Sci 103(20):7566–7570

Statistical Simulation

▶ Simulated Datasets

Statnet

▶ R Packages for Social Network Analysis

Status

▶ Time- and Event-Driven Modeling of Blogger Influence

Status Update

▶ Microtext Processing

Stochastic Actor-Based Models

▶ Actor-Based Models for Longitudinal Networks

Stochastic Diffusion Model for Propagation of Influence in a Social Network

▶ Mathematical Model for Propagation of Influence in a Social Network

Stochastic Matrix

▶ Probability Matrices

Stochastic Models

▶ Probabilistic Analysis

Storage

▶ Network Management and Governance

Strain Model

▶ Visualization of Large Networks

Strategic Allocation of Resources

► Network Management and Governance

Strategic Decision-Making

► Top Management Team Networks

Stream Querying and Reasoning on Social Data

Jayanta Mondal[1] and Amol Deshpande[2]
[1]Microsoft Corporation, Redmond, WA, USA
[2]Department of Computer Science, University of Maryland, College Park, MD, USA

Synonyms

Continuous query processing; Dynamic social networks; Incremental computation; Temporal analytics

Glossary

CEP	Complex event processing
CQP	Continuous query processing
SNA	Social network analysis
Social data stream	A time-stamped sequence of updates to a social network

Definition

We define social data to be comprised of a *network component* that captures the relationships among its entities, as well as the constant *stream of information* generated by the entities. In turn, we define stream querying and reasoning on social data to be tasks that need to process the data in a continuous fashion to produce answers and insights.

Introduction

Since the inception of online social networks, the amount of social data that is being published on a daily basis has been increasing at an unprecedented rate. Smart, GPS-enabled, always-connected personal devices have taken the data generation to a new level by making it tremendously easy to generate and share social content like *check-in* information, *likes*, *microblogs* (e.g., Twitter), multimedia data, and so on. There is an enormous value in reasoning about such streaming data and deriving meaningful insights from it in real time. Examples of potential applications include advertising, sentiment analysis, detecting natural disasters, social recommendations, personalized trends, and spam detection, to name a few. There is thus an increasing need to build scalable systems to support such applications. Complex nature of social networks and their rapid evolution, coupled with the huge volume of streaming social data and the need for real-time processing, raise many computational challenges that have not been addressed in prior work.

Social network data comprises two major components. First, there is a network (*linkage*) component that captures the underlying interconnection structure among the entities in the social network. Second, there is *content data* that is typically associated with the nodes and the edges in the social network. The social network data *stream* contains updates to both these components. The structure of the network may itself change rapidly in many cases, especially when things like webpages and user tags (e.g., Twitter *hashtags*) are treated as nodes of the network. However, most of the social network data stream consists of updates to the data associated with the nodes and the edges, for example, status updates and other content uploaded by the users, communication among the users, and so on. There is interest in performing a wide variety of queries and analytics over such data streams in real time.

Example 1 A natural structural pattern in a graph is a *star* that captures a node along with its connections. However, when we incorporate the activity behaviors (i.e., the content stream) of the participating nodes in a star, the star patterns often reveal different event semantics. For example, in a communication network, if the central node/device always sends one-way messages to the peripheral nodes/device, it might be an example of command-and-control behavior shown by a botnet master. Considering an example from social media, a Facebook page is connected with its followers in a star pattern, and frequent two-way communications between the page administrator and the followers indicate that the Facebook page is well-maintained.

Example 2 In social networks, users are often recommended items based on what their friends have liked or posted. However, quality of such recommendations would increase if the users could be categorized based on topic affinity (i.e., whether they like/post about sports, politics, technology), and they are recommended using items from friends who show similar topic affinities.

Example 3 Synchronized behavior (i.e., a set of nodes repeatedly performing the same activities over a period of time) in many application domains indicates unusual behavior. For example, such behavior in a user recommendation system like Yelp might indicate a coordinated effort to influence the ratings of some businesses (Jiang 2014). Some researchers have found that, sometimes unethical Yelp-users review businesses for money. The same/similar set of users was found to be reviewing businesses together (i.e., within a short period of time) multiple times.

Example 4 A clique of users in a social network indicates that they all know each other. However, an active clique, that is, a clique where all the users are active at the same time, can have a variety of interpretations and utility. An active clique where the nodes rarely communicate with the nodes outside the clique would indicate suspicious behavior. Moreover, by analyzing the number of activities by the nodes within a clique, it is possible to identify the group dynamics (i.e., leader vs. subordinate).

In terms of broad categories, these queries can range from simple publish-subscribe queries, where a user is interested in being notified when something happens in his or her friend circle, to complex anomaly detection queries, where the goal is to identify anomalous behavior as early as possible.

In this entry, we present an introduction to this new research area of *stream querying and reasoning* over social data. This area combines aspects from several well-studied research areas, chief among them, social network analysis, graph databases, and data streams. We provide a formal definition of the problem, survey the related prior work, and discuss some of the key research challenges that need to be addressed (and some of the solutions that have been proposed). We note that we use the term *stream reasoning* in this entry to encompass a broad range of tasks including various types of analytics, probabilistic reasoning, statistical inference, and logical reasoning. We contrast our use of this term with the recent work by Valle et al. (2008, 2009) who define this term more specifically to refer to integration of logical reasoning systems with data streams in the context of the Semantic Web. Given the vast amount of work on this and related topics, it is not our intention to be comprehensive in this brief overview. Rather we aim to cover some of the key ideas and representative work.

Key Points

Social data encodes rich metadata about its entities, complex interconnections among them, as well as the individual and community behavior of the entities in presence of social stimuli. While there has been much work that analyzes social data in static settings, performing such analysis in highly dynamic settings is still relatively unexplored. The dynamic setting not only gives rise to different type of reasoning tasks, it also calls for different types of problem abstractions, as well as different types of system designs where such tasks can be defined effortlessly, and can be executed in an efficient manner. In this entry, we attempt to define a set of abstractions for such tasks, survey related works, and lay the ground rules for executing such tasks efficiently in a general-purpose large-scale system.

Historical Background

Stream querying and reasoning over social networks combines aspects from several different research areas that have been very well studied over the last few decades. Here we will provide very brief background on three of the most closely related research areas: social network analysis, data streams, and graph databases. A more detailed background, including references to related work, can be found in an extended version of this article (Mondal and Deshpande 2013).

Social Network Analysis (SNA) Social network analysis, sometimes called *network science*, has been a very active area of research over the last decade, with much work on network evolution and information diffusion models, community detection, centrality computation, and so on. We refer the reader to well-known surveys and textbooks on that topic (see, e.g., Newman 2003; Scott 2012; Boccaletti et al. 2006). There has been an increasing interest in dynamic or temporal network analysis in recent years, fueled by the increasing availability of large volumes of temporally annotated network data and the real-time requirements of various popular online services. Such analysis has the potential to lend much better insights into various phenomena, especially those relating to the temporal or evolutionary aspects of the network. Many works have focused on designing analytical models that capture how a network evolves, with a primary focus on social networks and the Web. There is also much work on understanding how communities evolve, identifying key individuals, locating hidden groups, identifying changes, and visualizing the temporal evolution in dynamic networks. Most of that prior work, however, focuses on off-line analysis of static datasets.

Data Streams Data stream management is another research area that has seen tremendous amount of work over the last decade (see Aggarwal 2007; Muthukrishnan 2005; Garofalakis et al. 2011 for comprehensive surveys), resulting in several data management systems being built. Several SQL extensions have also been proposed to express continuous queries over data streams. Similarly, languages have also been designed for specifying event patterns to be matched against data streams. Continuous query processing (CQP) also bears strong resemblance to materialized view maintenance, an area that has also seen much work (Gupta and Mumick 1999). The key difference between the two research areas has been that CQP systems are designed to simultaneously support large numbers of relatively simple queries over highly dynamic data, whereas view maintenance techniques usually focus on a small number (usually just one) of more complex queries. The former also tend to build intermediate data structures like *predicate indexes* to efficiently identify the queries whose results are affected by new updates. Another line of work has focused on development of *one-pass* algorithms that can incrementally compute some quantities of interest over very large volumes of data (e.g., statistics or aggregates) while using very small amounts of memory (see, e.g., Muthukrishnan 2005).

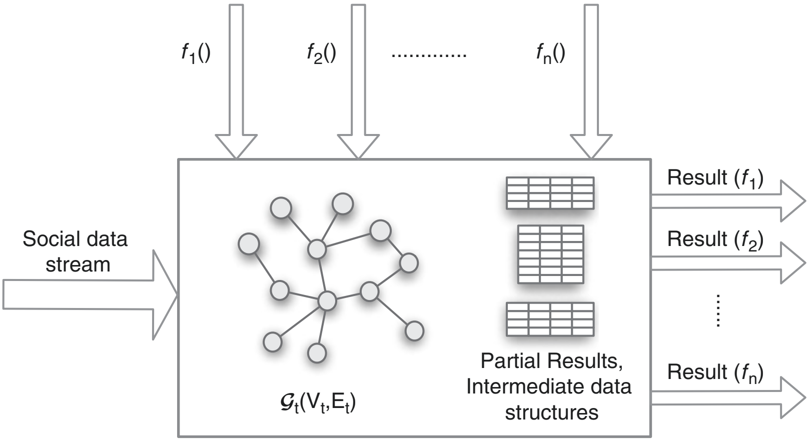
Graph Databases Since social networks are naturally represented as graphs, specialized graph data management systems are a natural option to store social network data. There has been much work on single-site graph databases and, in recent years, on distributed graph databases and programing frameworks for specifying batch analysis tasks over graphs. There is also much work on executing specific types of queries efficiently over graphs (both in centralized or distributed settings) through strategic traversal of the underlying graph, for example, reachability, keyword search queries, subgraph pattern matching, and shortest path queries. However, distributed management of dynamic graph data is not as well studied, especially in the data management research community.

Problem Definition

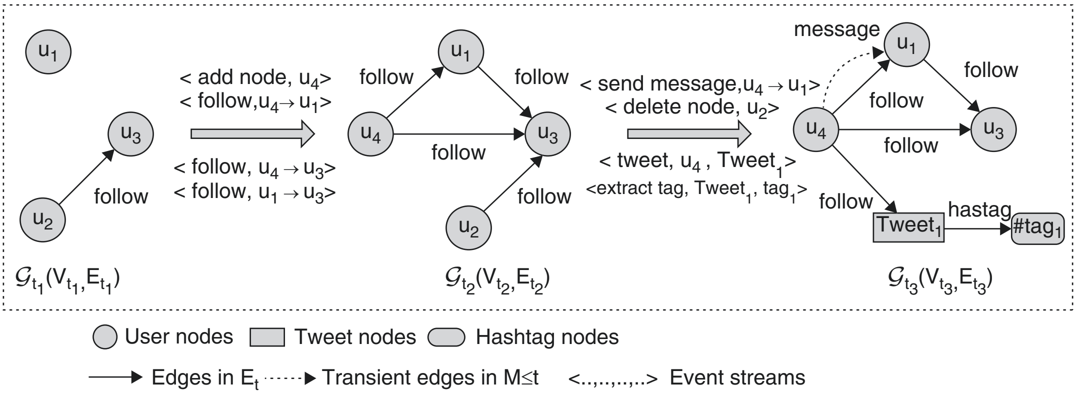
An online social network is defined to be a community of people (called *users*) connected via a variety of social relations that use online technologies to communicate with each other and share information. Social data is defined to be the data arising in the context of a social network that includes both the embedded structural information as well as the data generated by the users. Online social networks continuously generate a huge volume of such social data that includes both structural changes to the network and updates that are associated with the nodes or the edges of the network. The task of "stream querying and reasoning" refers to ingesting and managing such continuously generated data and querying and/or reasoning over it in real time as the data arrives (Fig. 1).

To make the discussion more concrete and formal, let (($\mathcal{G}$) *t*. *Vt*, *Et*) denote the underlying social graph at time *t*, with *Vt* and *Et* denoting the sets of nodes and edges at time *t*, respectively. In general, ($\mathcal{G}$) *t* is a heterogeneous, multi-relational graph that may contain many different types of nodes and may contain both directed and undirected edges (Fig. 2 shows an example graph). Along with nodes representing the users of the network, *Vt* may include other types of nodes, for example, nodes representing communities or groups, user tags, webpages, and so on. Similarly, *Et* includes not only symmetric *friendship* (or analogous) edges but may include asymmetric *follows* edges, *membership* edges, and other types of semipermanent edges that are usually in existence from the time they are formed till the time they are deleted (or till the current time). We distinguish such edges from *transient* edges that can be used to capture specific interaction between two nodes in *Vt* (e.g., a message being sent from one node to another). A transient edge is typically time-stamped and is only valid for the specific time instance. To allow us to clearly distinguish between these two types of edges, we do not include such transient edges in *Et*; instead, we use $M \leq t$ to denote all such transient edges that were generated from the beginning (i.e., from time 0) till time *t*. This distinction is not necessary but affords clearer distinctions between different types of stream reasoning tasks in many cases.

The information associated with the nodes and edges can be captured through a set of *key-value* pairs (also called *attribute-value* pairs) associated with them. We once again can make a distinction between semipermanent information associated with the nodes or the edges (e.g., user *names*, *interests*, or *locations*) and transient information associated with them (e.g., *status updates*). The former type of information can be seen as being valid for a given time period, whereas the latter is typically associated with a single time instance. Given this, we define a stream reasoning or querying task to be a declaratively specified query or an analysis or reasoning task that is posed (submitted) once by the user, but is executed continuously (or periodically with a user-specified



Stream Querying and Reasoning on Social Data, Fig. 1 High-level overview of a stream querying system


u_1
u_3
u_2
follow
$\mathcal{G}_{t_1}(V_{t_1},E_{t_1})$
< add node, u_4>
< follow,$u_4 \rightarrow u_1$>
< follow, $u_4 \rightarrow u_3$>
< follow, $u_1 \rightarrow u_3$>
follow
u_1
follow
u_4
follow
u_3
follow
u_2
$\mathcal{G}_{t_2}(V_{t_2},E_{t_2})$
< send message,$u_4 \rightarrow u_1$>
< delete node, u_2>
< tweet, u_4 , $Tweet_1$>
<extract tag, $Tweet_1$, tag_1>
message
u_1
follow
follow
u_4
follow
u_3
follow
$Tweet_1$
hastag
$\#tag_1$
$\mathcal{G}_{t_3}(V_{t_3},E_{t_3})$
User nodes
Tweet nodes
Hashtag nodes
Edges in E_t
Transient edges in $M \leq t$
<..,..,..,..> Event streams

Stream Querying and Reasoning on Social Data, Fig. 2 Example of a multi-relational, heterogeneous dynamic graph

frequency) as updates arrive into the system (Fig. 1). Along with a task, denoted *f*(), the user must specify what forms the *input* to the task, when to compute the *output*, and when to *return* the output to the user.

In many cases, the input is the *current* graph, that is, the input is $((\mathcal{G})\ t.\ Vt,\ Et)$ (that is continuously changing). An example of such a task is dense *subgraph maintenance* (Angel et al. 2012) where the goal is to compute and maintain the dense subgraphs in a dynamically changing graph. In other cases, the input to *f*() may be defined using a *sliding window*, that is, it may be defined as the set of all updates that arrived in recent past. An example of such a task is continuously identifying dense subgraphs in the graph formed by all message edges over say the last 24 h (i.e., the input to the task is $M \leq t - M - t{-}24\ h$). As time progresses, the window slides and new message edges will be added to the graph, and old message edges (that fall out of the window) will be deleted (Table 1).

The second key issue is when to compute the output and when to return it to the user. In some cases, the user may desire continuous execution of the query, that is, for every relevant change in the input, *f*() needs to be recomputed (from either scratch or incrementally). Anomaly detection queries typically need to be executed in this fashion since anomalies must be detected as soon as they are formed. But in other cases, the user may specify a frequency with which to execute the query or the task (e.g., every hour or every day). Finally, for simplicity, we will assume that the user should be notified anytime the output of *f*() is computed and is different from the prior output. However, in many cases, the output may need to be returned to the user only when he asks for it. In those cases, partial precomputation of the query results (with the rest of the processing performed at query time) becomes a possibility.

Stream Querying and Reasoning on Social Data, Table 1 Notation

Notation	Description
$((\mathcal{G})\ t,\ Vt,\ Et)$	Current state of the network
$M \leq t$	Transient edges generated till time *t*
$f_1(), f_2(), \ldots$	Stream querying or reasoning tasks
$N\,k\,(v)$	*k*-hop ego network of node *v*

Proposed Solution and Methodology

The area of stream querying and reasoning over social networks is still in its infancy, and as a result, the research in this area is somewhat fragmented with several ongoing attempts at unifying the different research themes. Here we begin with a broad classification of the different types of stream querying and reasoning tasks and give examples of different types of tasks that have been studied in prior literature. We then discuss some of the key research challenges in effective stream querying and reasoning that need to be addressed.

Classifying Tasks by Scope

Here we attempt to classify stream reasoning and querying tasks by their input scope, that is, what data forms the input to the task at any time. Broadly speaking, there are two crucial dimensions along which the tasks may differ.

Temporal Scope

The first key dimension captures the temporal scope of the task and has a direct impact on the amount of state that must be stored and reasoned about.

Entire Stream At one extreme, the temporal scope of a stream reasoning task may stretch from the beginning of the stream to the current time. Note that not all the data generated so far may be of interest – for example, the task may only see a subset of the data by choosing to focus only on certain attributes of the nodes or edges. However, the data of interest may have arrived into the system at any point in the past. For example, in a social network with location data, a stream reasoning task may wish to process all the location updates ever produced by a user for predicting future user movements. We expect such types of stream reasoning tasks to be somewhat uncommon given the large volumes of data generated in most online social networks.

Current State of the Network Many stream reasoning tasks will take the current state of the network (i.e., $Gt\,(Vt, Et)$) as the input. An example of this task is online dense subgraph maintenance (Angel et al. 2012) where the goal is to maintain the dense subgraphs of the current social network at all times.

Sliding Window The third alternative that falls in between the two extremes above is that the reasoning task defines a sliding window on the data stream and the input consists of all updates that arrive during that window. For instance, one may be interested in analyzing all messages that were exchanged during the last 24 h among the users of a network to identify anomalous behavior in real time. Another example of such a task is detection of personalized trends where the goal is to find the most commonly seen words or phrases in the recent status updates or blog posts by the friends of a user.

Network Traversal Scope

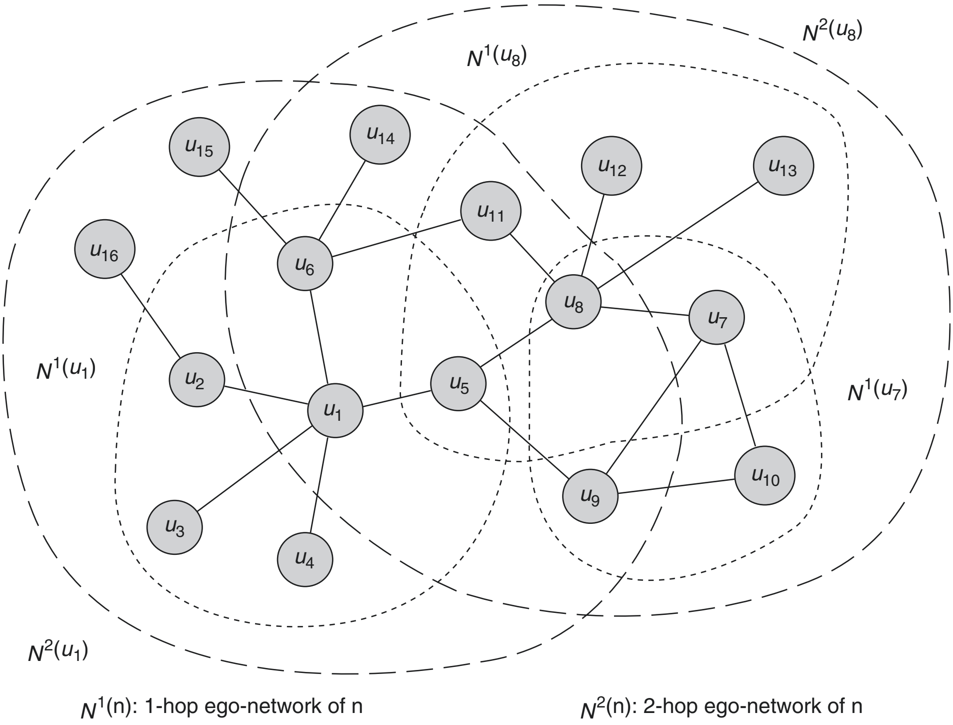
The second key dimension is what we call *network traversal scope* of a query, which refers to the portion of the network that provides the input to a stream reasoning query or task.

Global Scope Many stream reasoning tasks require reasoning over the entire network. An example of such a task is computation of *PageRank* (or other centrality measures like *betweenness centrality*, *eigenvector centrality*, etc.). Dense subgraph maintenance task discussed above is also an example of a task with global scope.

Egocentric Scope On the other hand, in many cases, a reasoning task or a query may only focus on a local neighborhood in the network, often termed *ego networks*. For example, if the goal is to identify *social circles* for a user (Mcauley and Leskovec 2012), then only a 1- or 2-hop neighborhood around the user may be of interest (Fig. 3). Personalized trend detection task, discussed above, is another example of such a task. Note that, in many cases, we may want to execute the same task for every node in the network (e.g., we may wish to do continuous trend detection for every user of the social network), and in total, updates in the entire network may need to examined. However, those should be treated as separate tasks, each of which is egocentric in scope. The most common example of an ego network is the network over the immediate set of neighbors of a node. However, in general, an ego network of node could be defined as *k-hop neighborhood* containing all nodes reachable within k hops from the node (and all the incident edges among those).

Key Applications

Next we attempt to provide a categorization of different stream reasoning and querying tasks by type. Given the wide variety in the stream



Stream Querying and Reasoning on Social Data, Fig. 3 Stream queries often have ego-centric scope: figure shows 1-hop ego-networks of u_1; u_7 and u_8 and 2-hop ego-networks of u_1 and u_8

reasoning tasks of interest, unlike the categorization by scope, the categorization that follows is less precise and not fully disjoint. Our intention here is not to be comprehensive, but rather to discuss some representative stream reasoning tasks.

Publish-Subscribe Queries

Perhaps the simplest kind of queries over streaming data is what are commonly referred to as *publish-subscribe* queries. These queries form a subclass of the more general class of *event monitoring* queries, where the users specify events or updates of interest and they should be notified as soon as a matching event is detected in the data stream. We make a loose distinction between simple event monitoring queries (what we call publish-subscribe queries) and more complex event monitoring or anomaly detection queries (discussed subsequently). For publish-subscribe queries, the events are typically defined over one or a few data stream updates (i.e., they have very limited temporal and traversal scopes). For example, a user may be interested in tweets that contain a particular key word, or a user may want to know as soon as a friend is online. In a location-enabled social network, a user may be interested in getting notified when one of his friends checks-in in a nearby restaurant or cafe. The key challenge with executing simple event monitoring queries is not so much the complexity of detecting the events, but rather dealing with the very large update rates as well as a very large number of queries.

Complex Event Processing (CEP)

On the other hand, in complex event processing, the events (often called *patterns*) to be detected often have larger temporal or network traversal scopes or both. Hence, unlike simpler publish-subscribe queries, efficiently detecting the events can be a major challenge in CEP. An example of such a query is a continuous subgraph pattern matching query, where the goal is to detect matches to a given query graph in real time. Wang and Chen (2012) use such queries for continuous detection of accidents from incoming traffic information. CEP systems often support specification of the events using a high-level declarative language. For example, in recent work, Anicic et al. (2011) proposed a language called EP-SPARQL that extends the SPARQL query language with support for specifying complex event processing queries over RDF data streams. Similarly, Mozafari et al. (2012) present a language for detecting hierarchical patterns over hierarchical data (e.g., XML data) that may be generalizable to graph-structured data as well.

Anomaly Detection

Anomaly detection queries can be seen as a form of complex event processing; however, due to their importance, we discuss them separately. The goal with real-time anomaly detection is to identify anomalous behavior in a dynamic network as quickly as possible. Two issues need to be addressed: (1) how to define what constitutes an "anomaly?" and (2) how to efficiently detect anomalies in presence of very high data rates? Generally speaking, anomalous behavior can be defined as behavior that deviates significantly from normal behavior. However, in highly dynamic and rapidly changing environments like an online social network, there is often no clear definition of normal behavior, making it a challenge to identify anomalous behavior. There have been many proposals for defining anomalous behavior in social networks over the years. For example, Akoglu et al. (2010) present an approach called Oddball that is based on analyzing the ego-networks of the nodes in the network.

Aggarwal et al. (2011) propose a probabilistic algorithm that maintains summary structure models about graph streams to detect outliers. We refer the reader to the tutorial by Akoglu and Faloutsos (2013) for a more comprehensive discussion of different anomaly detection algorithms.

Perhaps because of a lack of a clear definition of an anomaly, there is much less work on efficient techniques for real-time anomaly detection. From the efficiency perspective, an important issue is the scope (both temporal and network traversal) of an anomaly detection task. For example, if the goal is to identify users with anomalous behavior, then the network traversal scope could be limited to ego networks of the users. However, in many cases, detecting anomalous behavior may require global reasoning over the entire network.

Continuous Aggregates/Statistics Computation

In these types of queries, the goal is to incrementally maintain or compute an aggregate or a statistic over the network (Mondal and Deshpande 2013; Mondal 2014). An example of such a task is maintaining the top- *k* trending *hashtags* in Twitter, that is, hashtags with the highest activity over a recent window in past. Another well-studied task is the computation of *global clustering coefficient* in presence of streaming updates to the network structure (Jowhari and Ghodsi 2005; Becchetti et al. 2008). A simpler aggregate query might be to continuously maintain, for all users, their friends that are (physically) closest to them (the aggregate function here is MIN). There are two key properties of aggregate functions that have significant impact on the computational complexity of the computation task: *duplicate sensitivity* and *decomposability*. A duplicate-insensitive aggregate function will return the same value even if some of its inputs are repeated. Examples include MAX, MIN, and UNIQUE. Duplicate-insensitive aggregates are amenable to additional optimizations during computation (Madden et al. 2002a). On the other hand, whether the aggregate function is *holistic* or decomposable has a significant impact on the optimizations that we can perform (Madden et al. 2002a). A holistic aggregate function (e.g., MEDIAN) requires all the input values to compute the final result, whereas decomposable aggregate functions are amenable to optimizations centered around partial aggregate computation and can be computed with much less memory. Clustering coefficient is an example of the latter type of aggregate function since the number of triangles can be counted (mostly) independently for each node.

Maintenance of Views or Other Derived Information

In this type of a task, the goal is to incrementally maintain the result of running an algorithm or performing a computation on the social network in presence of updates. Such tasks can be seen as a generalization of *materialized view maintenance* in traditional relational databases. In traditional view maintenance, the goal is to incrementally maintain the result of a declaratively specified query; however, in social networks, the focus is often on more complex reasoning tasks. Examples of such tasks include incremental maintenance of *PageRank*, *dense subgraphs*, *spanning trees*, *shortest paths*, and *communities*. In general, for any graph algorithm that is of interest in SNA, the question of incremental maintenance of the result in a dynamic setting may need to be addressed. For example, Bahmani et al. (2010) address the problem of incrementally maintaining PageRank over a social network. Several works have considered the problem of incremental maintenance of dense subgraphs (e.g., Angel et al. 2012). The key challenge here is to avoid recomputation from scratch, and so far, most of the proposed techniques are heavily focused on a specific task.

Research Challenges

In this section, we look at some of the key research challenges in supporting stream reasoning and querying tasks over social networks and briefly review the prior work on addressing those challenges. We stress that the area of stream reasoning over social network is still in its infancy, and the solutions discussed here should be considered as the starting point for future research on this topic.

Query Language

One of the major challenges in building general-purpose data management techniques or systems for stream reasoning over social networks is the lack of a high-level declarative query language for specifying the tasks. This issue arises in the context of graph data management in static settings as well. Well-established relational or XML query languages are not appropriate for graph-structured data because they lack support for specifying graph traversals. Although there have been proposals for graph query languages, none has gained wide acceptance; perhaps the only exception is the SPARQL query language, but the use of that query language has been largely limited to RDF datasets. This lack of a declarative language has led to a significant repetition of work by researchers that are developing tools for stream reasoning and querying over social networks. Clearly it is impossible to specify all of the wide range of tasks that we discussed in the previous section using a high-level, declarative language. However, we believe that it is possible to develop a declarative query language that will serve the needs of many stream reasoning and querying tasks; further, those tasks that cannot be fully expressed in the language can use the language to do part of the computation, with the remaining part done using a program written in a procedural language that ingests the result (analogous to how *user-defined functions* (*UDFs*) are often used in conjunction with SQL in relational databases).

There are several starting points for designing such a query language. Several languages have been proposed in recent years that build upon SPARQL, for example, streaming SPARQL (Bolles et al. 2008), continuous SPARQL (C-SPARQL) (Barbieri et al. 2009), and EP-SPARQL (Anicic et al. 2011). Although these languages focus on RDF data streams, they could be adapted to use in social networks by treating social network data as RDF data. Figure 4 shows a C-SPARQL query that, given a stream of tweets along with the identified *hashtags* in it, returns all the hashtags with their cumulative frequencies within the last hour. Some of the key extensions to SPARQL include the use of "REGISTER QUERY" keyword to specify a continuous query that should be evaluated continuously and a way to specify a window over the stream (using keyword "RANGE").

Another option is to generalize XPath. For example, Mozafari et al. (2012) propose XSeq, an extension to XPath to express both sequential and Kleene-closure expressions for XML streams.

Example 1: C-SPARQL Example (Barbieri et al. 2009). Given the static user information and a stream of tweets, compute the total number of tweets per hashtag in last hour.

```
REGISTER QUERY
   NumberOfTweetsPerHashTag COMPUTE
   EVERY 10m AS
PREFIX ex: <http://example/>
SELECT DISTINCT ?hashtag ?total
FROM STREAM <http://twitter.com/alltweets>
   [RANGE 1h STEP 10m]
WHERE
?user ex:from ?country .
?user ex:tweets ?tweet .
?tweet ex:has ?hashtag FILTER
   (?country="USA")
AGGREGATE { (?total. COUNT(?tweet).
  ?hashtag }
```

Stream Querying and Reasoning on Social Data, Fig. 4 C-SPARQL example (Barbieri et al. 2009). Given the static user information and a stream of tweets, compute the total number of tweets per hashtag in last hour

Example 2: XSeq Example: In a stream of tweets, report users who have been active over a month. A user is active if the posts at least a tweet every 2 days.

```
return first(T)@userid
from/twitter/Z* (ST)*
where tag(Z) = 'tweet' and tag(T) ='tweet'
and T@date-prev(T)@date < 2
and last(T)@date-first(T)@date > 30
partition by /twitter/tweet@userid
```

Stream Querying and Reasoning on Social Data, Fig. 5 XSeq example: In a stream of tweets, report users who have been active over a month. A user is active if he posts at least a tweet every 2 days

Figure 5 shows an XSeq query that reports Twitter users who have been active for over a month. A key challenge here is that XPath is designed to operate on tree-structured data, not graph-structured data. However, recent theoretical work suggests that it may be possible to use XPath for specifying graph queries (Libkin et al. 2013).

Example 3: Datalog Example (Moustafa et al. 2011): Compute the clustering coefficient of each node.

```
NeighborCluster(X, COUNT <Y, Z>) :=
                                Edge(X,Y),
Edge(X,Z), Edge(Y,Z)
Degree(X, COUNT<Y>) := Edge(X, Y)
ClusteringCoeff(X, C) :=
                   NeighborCluster(X,N), Degree(X,D),
C=2*N/D*(D-1)
```

Stream Querying and Reasoning on Social Data, Fig. 6 Datalog example (Moustafa et al. 2011): Compute the clustering coefficient of each node

Finally, the option that we have taken in our work (Moustafa et al. 2011; Mondal and Deshpande 2013) is to extend Datalog (Ramakrishnan and Ullman 1995) for this purpose. In recent years, Datalog has been shown to be an effective centerpiece in enabling declarative specification in a range of domains including networking, data cleaning, machine learning, and SNA. Compared to the above two languages, Datalog seems more amenable to be extended to support a large class of complex aggregate queries (e.g., global queries like *PageRank* computation and *shortest paths* can be specified using *recursion*).

Datalog snippet in Fig. 6 specifies computation of *local clustering coefficient*, a measure of connectedness of a node's neighborhood. With some extensions, Datalog can also be used to specify social network transformation tasks as we showed in our prior work (Moustafa et al. 2011, 2013). Such flexibility may make a Datalog-based language, a superior option in the end to specify a wide variety of stream reasoning tasks over social data.

Efficient Execution Strategies

Irrespective of how the stream reasoning tasks are specified, we must devise efficient execution strategies that can handle the very high update rates expected in online social networks. Below we briefly survey the key ideas that have been used successfully in past research on data streams for low-latency execution.

Incremental Computation The naive option of re-executing a query or a reasoning task when a new update arrives is likely to be infeasible except for very low-rate data streams. Instead the goal of incremental computation is to maintain sufficient intermediate state in memory so that the new answer can be computed in an incremental fashion with minimal work. Such incremental techniques are unfortunately often specific to the task at hand. Eppstein et al. (1999) did an early survey on the related topic of dynamic graph algorithms. In a recent work, Angel et al. (2012) and Agarwal et al. (2012) devise techniques for maintaining dense subgraphs; Bahmani et al. (2010) present an approach to incremental computation of PageRank; Kutzkov and Pagh (2013) present an incremental algorithm for computing clustering coefficient; and so on. A key research challenge here is to identify incremental techniques that are applicable to a wide variety of tasks (one way to do that is to focus on a high-level query language as we discussed in the previous section, e.g., C-SPARQL (Barbieri et al. 2010)). There is also often a natural trade-off between the amount of intermediate state that is maintained and the amount of work that needs to be done when a new update arrives. In one of our recent works (Mondal 2016) we explore such trade-off among incremental computing, on-demand computing, and model-based computing. Better understanding of this trade-off also presents a rich area for future work.

Sharing Across Multiple Queries Unlike traditional data management systems, in stream query processing systems, we may have thousands to millions of continuous queries running simultaneously. For instance, a personalized trend detection query where the goal is to monitor trends in every user's ego network can be seen as a collection of a large number of independent queries, one for each user. Sharing of computation across these queries is crucial in order to limit the computational cost. Such sharing has been shown to be an effective way to deal with high-rate data streams in past work on data streaming systems (Madden et al. 2002b; Diao et al. 2002). However, these types of techniques have not been well studied in social network setting. In a recent work, we designed novel techniques based on graph compression to exploit such sharing for continuous aggregate computation in social networks (Mondal and Deshpande 2013; Mondal 2014).

Approximate Computation One way to mitigate the execution complexity is to consider computing approximate answers instead of exact answers. This is especially attractive in scenarios where exact computation can be shown to be prohibitively expensive. For example, Becchetti et al. (2008) show how one can incrementally compute local clustering coefficient with small error bounds where the exact algorithm (Alon et al. 1997) can require $O(n^{2.3727})$ time. Although there is much work on this topic in the data streams community, only recently have researchers started investigating similar problems for network algorithms. Zhao et al. (2011) present a graph-sketching technique, called gSketch, and show it can be used to answer several primitive frequency estimation techniques. Similarly, Ahn et al. (2012) present graph-sketching techniques for approximating *cut* values and for approximating the number of matches to a subgraph pattern query.

Sampling Another general technique to deal with the high update rates is use random sampling to reduce the size of the data that needs to be processed. We may sample at two levels in a social network: first, we can try to sample from the network structure itself to reduce the size of the graph that needs to be processed; second, we can sample from the updates to the content. The latter is generally well understood, and the theory developed in the data streams literature could be extended for some types of queries. However, sampling the network structure is tricky since a naive random sample is likely to yield a network with very different properties than the original network. We refer the reader to Ahmed et al. (2014) for a detailed discussion of network sampling, both in static and streaming settings.

Parallel Computation The increasing scale of most online social networks necessitates use of parallel and distributed solutions. Unfortunately, computations on social networks are not easily distributable because of their highly interconnected nature. In fact, partitioning a social graph, which is key to distributed graph processing, is a hard problem to tackle because of overlapping community structure and existence of highly connected dense components (cores) in most social networks. One of the simplest examples of a stream query on social data is a publish-subscribe query that asks to *fetch all updates from all friends* (this is also called *feed following*). Answering such queries with very low latencies is challenging if the data is distributed across a set of machines – for most users, their friends' data is likely to be located across multiple machines necessitating expensive distributed traversals. One extreme option is to replicate the data sufficiently so that, for each user, the required data (i.e., status updates of all their friends) is located on some machine (Pujol et al. 2010). However, both the memory overhead and the replica maintenance overhead can be very high for that solution (Mondal and Deshpande 2012). More intelligent and sophisticated techniques for partitioning and replica maintenance must be developed to address these issues for more general stream reasoning and querying tasks. Another key challenge is designing appropriate distributed programing frameworks to support specifying general-purpose stream querying and reasoning tasks. Although there has been some progress on addressing this challenge in recent years (e.g., Kineograph (Cheng et al. 2012), GraphInc (Cai et al. 2012)), much more needs to be done to scalably support a variety of complex stream querying and reasoning tasks.

Future Directions

Stream querying and reasoning over social data is an emerging research area that combines aspects from social network analysis, graph databases, and data streams and is motivated by an increasing need for real-time processing of continuously generated social data. In this entry, we presented a brief overview of this field and discussed some of the key research challenges therein. There has been much work on specific problems in this field over the last few years (e.g., detecting specific types of events or anomalies, incremental maintenance of derived structures like dense subgraphs, approximating different types of summary statistics). However, designing general-purpose data management systems that enable declarative specification of stream querying and reasoning tasks and that can efficiently execute such tasks over high-rate data streams remains a fruitful direction for future research. Moreover, a general-purpose system needs to consider the usability of the system by considering the analysists' point of view. A stream reasoning task is often complete until the output of such a task is considered in an appropriate context. Context of an event could be seen as additional qualifying information that puts an event in perspective, thereby explaining the event and making it consumable and actionable. At an abstract level, such context-driven analysis requires us to execute an explanation algorithm on some accessible form of context information. Both the explanation task and the context information, however, depending on the application, could come in various flavors. The simplest and most general form of context information, perhaps, could be the stored history of the nodes that were a part of a detected event. On the other hand, the explanation algorithm might be any arbitrary analysis task on such history.

Example 5 In a social network, influential users are often defined to be the ones that induce a lot of activities among their followers. A single instance of such a behavior could be extracted by identifying a star pattern where all the nodes are active within a small period of time, especially when the peripheral nodes are not well connected. This, however, can't guarantee that the central node initiated the activities of the peripheral node. But, if the same active star repeats over time (i.e., the central node remains the same, and the peripheral nodes are very similar), then we can say with higher confidence that the central node is responsible for the activities. In connection with the ongoing discussion about context-aware stream processing, in this example, the context information could be a list of all the active stars that have been detected by the system. While the explanation algorithm figures out, for a given center of an active star, whether the central node has participated in many active stars over time, thereby deeming the node as an influential user.

Example 6 Synchronized behavior: Synchronized behavior by nodes in a network has been shown to be indicative of anomalous behavior. In a recent work Jiang (2014) has shown that new Yelp businesses often buy reviews from users, and the unethical users typically respond within a time period (i.e., they are active within a window) by writing reviews for the concerned business. Moreover, the same set of users tends to write reviews for multiple businesses over time. This again could be seen as an example of active star, where only the set of peripheral nodes (the unethical users) repeats over time in groups and the central node changes (i.e., the new businesses). In this example, the explanation task needs to look for historical evidence for repetitive grouped review writing by a set of users.

So far, we have discussed some examples of explanation algorithms, however, with a rather simple form of context information, that is, history-based context which is static in nature. A more complex situation arises when we allow such contexts to be dynamic in nature. For example, detection of the interesting events could be relative to other events detected in the system. An event with low selectivity (i.e., many such events are simultaneously detected) might be of reduced interest, and an application might require them to be filtered out. There have been some works on combining stream processing in the context of historical data (Chandrasekaran 2004; Reiss 2007; Chandramouli 2012). However, none of that work supports graph-structured data and graph-based event detection. There have been several recent efforts, generally termed as Lambda architecture that also try to combine stream processing and batch processing, and produces unified results. Even though, at a high level, Lambda architecture is similar to context-aware stream processing, they are fundamentally different, as Lambda architecture is designed to perform both streaming and batch processing on the same data, so that final results could be produced by combining output from both the systems. Whereas, in our case, we need a context server that can act as a knowledge base whom the stream-reasoning tasks can consult to evaluate, curate, and demystify their output.

Cross-References

- ▶ Analysis and Mining of Tags, (Micro)Blogs, and Virtual Communities
- ▶ Analysis and Visualization of Dynamic Networks
- ▶ Anomaly Detection
- ▶ Behavior Analysis in Social Networks
- ▶ Behavior Modeling in Social Networks
- ▶ Modeling and Analysis of Spatiotemporal Social Networks
- ▶ Querying Volatile and Dynamic Networks
- ▶ Real-Time Social Media Analysis
- ▶ SPARQL
- ▶ Temporal Networks

References

Agarwal MK, Ramamritham K, Bhide M (2012) Real time discovery of dense clusters in highly dynamic graphs: identifying real world events in highly dynamic environments. Proc VLDB Endow 5(10):980–991

Aggarwal C (ed) (2007) Data streams: models and algorithms. Springer, New York

Aggarwal C, Zhao Y, Yu P (2011) Outlier detection in graph streams. In: 27th international conference on data engineering (ICDE), Hannover, pp 399–409

Ahmed NK, Neville J, Kompella R (2014) Network sampling: from static to streaming graphs. ACM Transactions on Knowledge Discovery from Data (TKDD), 2014;8(2):7

Ahn KJ, Guha S, McGregor A (2012) Graph sketches: sparsification, spanners, and subgraphs. In: Proceedings of the 31st ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems, PODS 2012, Scottsdale, 20–24 May 2012, pp 5–14

Akoglu L, Faloutsos C (2013) Anomaly, event, and fraud detection in large network datasets. In: WSDM'13 Proceedings of the sixth ACM international conference on web search and data mining, Rome, 4–8 Feb 2013, pp 773–774

Akoglu L, McGlohon M, Faloutsos C (2010) Oddball: spotting anomalies in weighted graphs. In: Proceedings of the 14th Pacific–Asia conference on advances in knowledge discovery and data mining (PAKDD), Hyderabad, 21–24 June 2010, pp 410–421

Alon N, Yuster R, Zwick U (1997) Finding and counting given length cycles. Algorithmica 17:209–223

Angel A, Sarkas N, Koudas N, Srivastava D (2012) Dense subgraph maintenance under streaming edge weight updates for real-time story identification. VLDB 5:574–585

Anicic D, Fodor P, Rudolph S, Stojanovic N (2011) EP-SPARQL: a unified language for event processing and stream reasoning. In: WWW 2011, Hyderabad, pp 635–644

Bahmani B, Chowdhury A, Goel A (2010) Fast incremental and personalized pagerank. Proc VLDB Endow 4:173–184

Barbieri DF, Braga D, Ceri S, Della Valle E, Grossniklaus M (2009) C-SPARQL: SPARQL for continuous querying. In: Proceedings of the 18th international conference on World wide web, Madrid, pp 1061–1062

Barbieri DF, Braga D, Ceri S, Grossniklaus M (2010) An execution environment for C-SPARQL queries. In: Proceedings of the 13th international conference on extending database technology, EDBT'10, Lausanne, pp 441–452

Becchetti L, Boldi P, Castillo C, Gionis A (2008) Efficient semi-streaming algorithms for local triangle counting in massive graphs. In: Proceedings of ACM KDD, Las Vegas, Aug 2008, pp 16–24

Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U (2006) Complex networks: structure and dynamics. Phys Rep 424(4):175–308

Bolles A, Grawunder M, Jacobi J (2008) Streaming SPARQL: extending SPARQL to process data streams. In: The semantic web: research and applications. Springer, New York, pp 448–462

Cai Z, Logothetis D, Siganos G (2012) Facilitating real-time graph mining. In: Proceedings of the fourth international workshop on cloud data management, CloudDB'12, Sheraton, Maui, 29 Oct 2012, pp 1–8

Chandramouli BJ (2012). Temporal analytics on big data for web advertising. In: IEEE 28th international conference on data engineering (ICDE), Apr 2012, pp 90–101

Chandrasekaran S (2004) Remembrance of streams past: overload-sensitive management of archived streams. In: VLDB'04 proceedings of the 30th international conference on very large data bases – volume 30, Toronto, 31 Aug–3 Sept 2004, pp 348–359

Cheng R, Hong J, Kyrola A, Miao Y, Weng X, Wu M, Yang F, Zhou L, Zhao F, Chen E (2012) Kineograph: taking the pulse of a fast-changing and connected world. In: Proceedings of the 7th ACM European conference on computer systems, EuroSys'12, Bern, pp 85–98

Diao Y, Fischer P, Franklin MJ, To R (2002) YFilter: efficient and scalable filtering of XML documents. In: Proceedings of the 18th international conference on data engineering, IEEE, San Jose, pp 341–342

Eppstein D, Galil Z, Italiano GF (1999) Dynamic graph algorithms, chapter 8. In: Atallah MJ (ed) Algorithms and theory of computation handbook. CRC, Boca Raton

Garofalakis M, Gehrke J, Rastogi R (eds) (2011) Data stream management – processing high-speed data streams, Data-centric systems and applications series. Springer, New York

Gupta A, Mumick IS (1999) Materialized views: techniques, implementations, and applications. MIT, Cambridge

Jiang MB (2014) Catchsync: catching synchronized behavior in large directed graphs. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, New York, 24–27 Aug 2014, pp 941–950

Jowhari H, Ghodsi M (2005) New streaming algorithms for counting triangles in graphs. In: Wang L (ed) Computing and combinatorics, Lecture notes in computer science, vol 3595. Springer, Berlin/Heidelberg, pp 710–716

Kutzkov K, Pagh R (2013) On the streaming complexity of computing local clustering coefficients. In: Proceedings of the 6th ACM international conference on web search and data mining, WSDM, Rome, 4–8 Feb 2013, pp 677–686

Libkin L, Martens W, Vrgoc D (2013) Querying graph databases with XPath. In: Proceedings of the 16th international conference on database theory, ICDT, Genoa, 18–22 Mar 2013, pp 129–140

Madden S, Franklin MJ, Hellerstein JM, Hong W (2002a) TAG: a tiny aggregation service for ad-hoc sensor networks. In: Proceedings of the 5th symposium on operating systems design and implementation, OSDI, Boston, pp 131–146

Madden S, Shah MA, Hellerstein JM, Raman V (2002b) Continuously adaptive continuous queries over streams. In: Proceedings of the 2002 ACM SIGMOD international conference on management of data, Madison, 3–6 June 2002, pp 49–60

Mcauley J, Leskovec J (2014) Discovering social circles in ego networks. ACM Transactions on Knowledge Discovery from Data (TKDD), 2014;8(1):4

Mondal J (2014) EAGr: supporting continuous ego-centric aggregate queries over large dynamic graphs. In: Proceedings of the 2014 ACM SIGMOD international conference on management of data, Snowbird, 22–27 June 2014, pp 1335–1346

Mondal J (2016) CASQD: continuous detection of activity-based subgraph pattern queries on dynamic graphs. In: Proceedings of the 10th ACM international conference on distributed and event-based systems, Irvine, 20–24 June 2016, pp 226–237

Mondal J, Deshpande A (2012) Managing large dynamic graphs efficiently. In: Proceedings of the 2012 ACM SIGMOD international conference on management of data, Scottsdale, 20–24 May 2014, pp 145–156

Mondal J, Deshpande A (2013) Stream querying and reasoning on social data. http://www.cs.umd.edu/~jayanta/papers/SRQ-ESNAM.pdf. Accessed 18 Apr 2017

Moustafa WE, Namata G, Deshpande A, Getoor L (2011) Declarative analysis of noisy information networks. In: ICDE GDM workshop, Hannover, 11–16 Apr 2011

Moustafa WE, Miao H, Deshpande A, Getoor L (2013) GrDB: a system for declarative and interactive analysis of noisy information networks: demo. SIGMOD, New York

Mozafari B, Zeng K, Zaniolo C (2012) High-performance complex event processing over XML streams. In: Proceedings of the 2012 ACM SIGMOD international conference on management of data, Scottsdale, 20–24 May 2012, pp 253–264

Muthukrishnan S (2005) Data streams: algorithms and applications. Now Publishers, Boston/Hanover

Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256

Pujol J, Erramilli V, Siganos G, Yang X, Laoutaris N, Chhabra P, Rodriguez P (2010) The little engine (s) that could: scaling online social networks. In: Proceedings of the ACM SIGCOMM 2010 conference, New Delhi, pp 375–386

Ramakrishnan R, Ullman JD (1995) A survey of deductive database systems. J Log Program 23(2):125–149

Reiss FS (2007). Enabling real-time querying of live and historical stream data. In: 19th international conference on scientific and statistical database management, SSDBM 2007, Banff, 9–11 July 2007, p 28

Scott J (2012) Social network analysis. Sage, London

Valle ED, Ceri S, Barbieri DF, Braga D, Campi A (2008) A first step towards stream reasoning. In: FIS, Vienna, 28–30 Sept 2008, pp 72–81

Valle ED, Ceri S, van Harmelen F, Fensel D (2009) It's a streaming world! Reasoning upon rapidly changing information. IEEE Intell Syst 24(6):83–89

Wang C, Chen L (2009) Continuous subgraph pattern search over graph streams. IEEE 25th International Conference on Data Engineering (ICDE), IEEE 2009; 29:393–404

Zhao P, Aggarwal CC, Wang M (2011) gSketch: on query estimation in graph streams. Proc VLDB Endow 5:193–204

Stress Model

► Visualization of Large Networks

Structural and Locational Properties

► Path-Based and Whole-Network Measures

Structural Attribute

► Collective Classification: Structural Features

Structural Autonomy

► Structural Holes

Structural Holes

Alona Labun[1] and Rafael Wittek[2]
[1]Jeugdhulp Friesland, Leeuwarden, The Netherlands
[2]Theoretical Sociology – Department of Sociology, University of Groningen, Groningen, The Netherlands

Synonyms

Brokerage; Middlemen; Network entrepreneurs; Social capital; Structural autonomy

Glossary

Aggregate constraint	The sum of dyadic constraints imposed on a focal actor by all his contacts
Dyadic constraint	Degree to which a focal actor's primary contact can constrain exchange opportunities with third parties
Effective size	The number of non-redundant contacts in a focal actor's personal network
Redundant tie	A tie to a cluster of contacts to which a focal actor already has ties with other actors
Secondary hole	Gaps in the networks of a focal actor's primary contacts

Definition

A structural hole refers to an "empty space" between contacts in a person's network. It means that these contacts do not interact closely (though they may be aware of one another). Actors on either side of the structural hole have access to different flows of information (see Fig. 1). Structural holes therefore reflect "an opportunity to *broker* the flow of *information* between people, and *control* the projects that bring together people from opposite sides of the hole" (Burt 2000). Several measures are used to capture structural-hole networks.

Effective network size is an elementary building block in all structural-hole measures. It is composed of three elements: First, the *proportion* of an actor i's time and energy invested in a relation with q:

$$p_{iq} = \left(z_{iq} + z_{qi}\right) / \left[\sum_j \left(z_{ij} + z_{ji}\right)\right], \tag{1}$$

Z_{iq} , Z_{qi} , Z_{ij} , Z_{ji} represent time or energy actor i invests in q, i in j, and j in i, respectively.

Second, the marginal strength of j's relation with q:

$$m_{jq} = \left(z_{jq} + z_{qj}\right) / \max\left(z_{jk} + z_{kj}\right) \quad j \neq k. \tag{2}$$

m_{jq} is the marginal strength of contact j's relation with actor q. Z_{jq} is the network variable measuring the strength of the relation from j to q and max(Z_{jq} is the largest of j's relations with anyone (Burt 1992: 51).

Third, the *redundant portion* (RP) of i's network. The portion of i's relation with j that is redundant to i's relations with other primary contacts is defined as the following:

$$\mathrm{RP} = \sum_q p_{iq} m_{jq}. \tag{3}$$

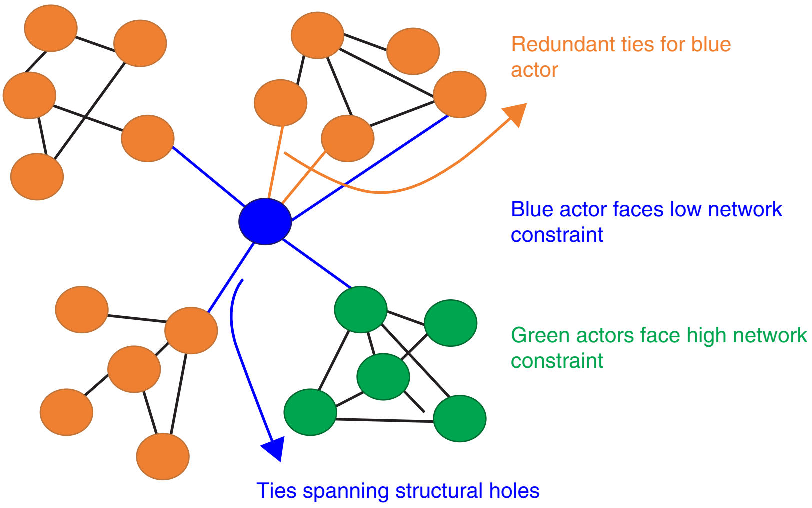
Effective size (ES) is obtained by aggregating across all of i's primary contacts j:

$$\mathrm{ES} = \sum_i \left[1 - \sum_q p_{iq} m_{jq}\right]. \tag{4}$$

The effective size of i's network ranges from 1 (network provides one single contact) to N (all contacts are non-redundant), with N being the number of all contacts in i's network. The *efficiency* of an actor's network is computed as the effective size divided by the number of actors in the network.

Dyadic constraint C_{ij} measures the degree to which an actor j imposes structural constraint on the focal actor i. Dyadic constraint is highest in a situation where the focal actor's network is inefficient (i.e., he or she invests time and energy in the relation to someone whose network lacks structural holes and is also tied to other contacts in the focal person's network). A low dyadic constraint originates from actors who do not have

Structural Holes, Fig. 1 Schematic representation of structural holes in social network


Redundant ties for blue actor
Blue actor faces low network constraint
Green actors face high network constraint
Ties spanning structural holes

many ties to a focal person's contacts. Dyadic constraint is a function of effective size:

$$C_{ij} = \left(p_{ij} + \sum_{q} p_{iq}p_{qi}\right)^{2}. \qquad (5)$$

C_{ij} + level of constraint that contact *j* poses on focal actor *i*; p_{ij}, p_{iq}, p_{qj} see Eq. 1.

Aggregate constraint indicates the extent to which an actor is constrained by the structure of the network involving other members of his or her group. High constraint values indicate low autonomy: the actor has few structural holes, i.e., little entrepreneurial opportunities. Technically, aggregate constraint is the sum of all contact-specific dyadic constraints in an actor's network. This indicator is also the most frequently used one in structural-hole research.

Hierarchy (*H*) indicates the extent to which aggregate constraint on ego is concentrated in a single alter. If the total constraint on the person is concentrated in a single other actor, the hierarchy measure will have a higher value. If the constraint results more equally from multiple actors in a person's network, hierarchy will be less. The hierarchy measure, in itself, does not assess the degree of constraint. Independently of the constraint on a focal actor, it measures inequality in the distribution of constraints on a focal person across the other actors in its neighborhood.

$$H = \left(\frac{C_{ij}}{C/N}\right). \qquad (6)$$

C_{ij} + level of constraint that *j* poses on *i*; *C* + sum of constraint (from an actor's network) across all *N* relationships of an actor; *N* + number of contacts in the actor's network; *C*+*N* + mean level of constraint per contact; and the ratio is 1 for contact *j* posing an average level of constraint.

Oligopoly Primary structural holes were defined as the aggregate of all dyadic constraint on a focal actor. Contact *j*'s constraint on a focal actor *i* was defined as the product of two terms (Burt 1992: 62): (1) the network time and energy *i* invested to reach *j* multiplied by (2) the lack of structural holes around *j*. The second term, in turn, is the product of two conditions: (a) the lack of primary structural holes between the contact *j* and others in the player's network and (b) the lack of secondary structural holes between the contact and others outside the network who could replace the contact. Burt refers to this second term as the *oligopoly*: "a measure of the organization of players within the cluster around contact j such that it would be difficult to replace j, or threaten him with being replaced, by some other player in the cluster" (Burt 1992: 62).

Hole signatures of a focal actor's network describe "the distribution of opportunity and constraint across the individual relationships in a player's network" (Burt 1992: 62). Hole signatures can be graphically represented, with the time and energy devoted by a focal actor *i* to a specific alter *j* (p_{ij}) delimiting the upper boundary and the dyadic constraint (c_{ij}) defining the lower boundary. Hole signatures allow to assess structural features of a focal actor's network (clique, center-periphery, leader hierarchy, and leaderless hierarchy).

Hole depth The depth of a structural hole reflects "the ease with which it can be developed for control and information benefits" (Burt 1992: 42–44). The depth of a hole between two actors is a function of both the degree of cohesion between two players and the degree of structural equivalence of their ties to others: in the ideal-typical structural hole, both actors are neither connected nor do they have equivalent relations to others. A deep structural hole characterizes two unrelated actors with equivalent ties to third parties: they are "competitors in the same market." In a shallow structural hole, two actors have a tie, but do not share equivalent relations to third parties.

Historical Background

Structural-hole theory and the related measures can be seen as the confluence of three streams of work. First, during the late 1960s and early 1970s, Harrison White and his group (now often referred to as the Harvard School), formalized ideas focusing on the absence of ties between individuals ("gaps"). This resulted in the development of *blockmodeling* algorithms, which grouped structurally equivalent nodes into blocks, and identified "zero blocks" – nodes that did not share similar relations with third parties. These "zero blocks" have qualities similar to structural holes.

Second, the article "The Strength of *Weak Ties*" by one of White's graduate students (Granovetter 1973) produced the counterintuitive empirical finding that in some situations – like job search, the topic of Granovetter's study – individuals benefit more from weak ties (like acquaintances) rather than strong ties (like friends or relatives), because one's weak ties can provide access to circles of information we usually are not familiar with. The "strength" of an interpersonal tie is a linear combination of the amount of time, the emotional intensity, the intimacy (or mutual confiding), and the reciprocal services which characterize each tie. Strong ties represent closer friendship and greater frequency of interaction, whereas weak ties correspond to acquaintances (Granovetter 1973). Members of closely knit groups connected through strong ties tend to be exposed to similar sources of information. Truly novel, valuable information is often likely to come from more distant acquaintances who may serve as a conduit to hard-to-reach parts of the network. A key proposition in Granovetter's argument is that "all bridges are weak ties," which rules out that strong ties can be bridges (also known as the "forbidden triad" assumption). However, Burt (1992: 27) argues that the main source of benefits in a network is not the weakness of the tie, but the hole it spans. From this perspective, the focus on the weakness or strength of a tie even obscures the importance of control benefits. "Bridge strength is an aside in the structural hole argument, since information benefits are expected to travel over all bridges. Benefits vary between redundant and non-redundant ties" (Burt 1992: 30).

But Granovetter's article by now is among the most frequently cited papers in the social sciences. In addition to stimulating much substantive research, e.g., on job search, it also sparked the interest for social network indicators reflecting an individual's *centrality* in the network (Freeman 1979). *Degree* centrality captures communication *activity* and has been defined by the number of ties an actor has with others in the network or the number of others who choose a focal actor. *Betweenness* centrality reflects the potential for *control* of communication and has been defined as the extent to which an actor has control over other actors' access to various regions of the network. Closeness centrality captures either *independence* or *efficiency* and has been conceptualized as an actor's ability to access independently all other members of the network. *Eigenvector centrality* (Bonacich 1987: 1172) measures centrality as the summed connection to others, weighted by their centralities. This measure allows to distinguish situations in which being connected to others with many contacts (powerful others) is advantageous for a focal actor (as is the case in communication networks), from situations in which being connected to powerful others is a liability (as is the case in bargaining situations). These centrality measures only partly capture the essence of structural holes, mainly because they are less sensitive to the gaps in the networks of a focal actor's primary contacts.

Third, Burt was among the first who did a serious effort to ground structuralist reasoning on a behavioral micro-foundation. Many of the ideas presented in his 1992 book on structural holes – including the core argument on structural autonomy – had actually been elaborated in detail about a decade earlier in his *Toward a Structural Theory of Action. Network Models of Social Structure, Perception and Action* (Burt 1982). Here, he exposes the rational choice framework underlying structural-hole theory. A key assumption is that individuals are purposeful actors, who strive for improving their well-being by evaluating the costs and benefits of different action alternatives, taking into consideration structural constraints. Individuals in similar network positions face similar constraints. As a result, the network is simultaneously an indicator of entrepreneurial opportunity and of motivation (Burt 1992: 35).

By combining an innovative structural approach with a theory of action, Burt's structural-hole framework significantly advanced previous network research, which clearly lacked a behavioral micro-foundation.

Structural-Hole Theory

In social networks, access to advantageous structural positions is not equally distributed across all actors: some group members may be positioned at the interface between multiple groups with access to boundary-spanning links, while others are positioned in the middle of a single tightly knit group. Structural holes offer two main benefits.

Information benefits come in three forms: access, timing, and referrals. A network rich in structural holes provides one with *access* to nonredundant sources of information originating in multiple, noninteracting parts of the network. It also increases the likelihood of receiving information earlier than individuals in less advantageous network positions (*timing*) and that others talk positively about the focal actor in their own networks (*referrals*).

Control benefits of structural holes result from the opportunity to either play two unrelated parties out against each other (*tertius gaudens*) or to bring them together (*tertius iungens*). In both cases, the third party can reap benefits.

Structural-hole theory further assumes actors to strategically and proactively creating and manufacturing their social network. This means that actors will actively develop the information and control benefits of existing structural holes and manage the constraint of absent structural holes (Burt 1992: 230). They have three strategies to achieve this: they can *withdraw* from a contact, they can *expand* their network by adding a contact's competitor to their network, or they can "leave the constraint-generating network in place but to manage the offending constraint by *embedding* it in a second relationship over which you have more control" (Burt 1992: 233).

Key Applications

Structural-hole theory has stimulated considerable empirical research on networks, mostly in and between organizations, as well as on entrepreneurship. It was used to explain a wide range of outcomes at the level of individuals and organizations.

Performance With information being a critical resource in organizational settings (McCall 1979; Mechanic 1962; Pettigrew 1972; Pfeffer 1981), individuals rich in structural holes have a better opportunity to manipulate information for their purpose. According to a meta-analysis (Balkundi et al. 2009), and a recent review (Brass 2011), spanning structural holes increased performance or innovation for the focal actor (Ahuja 2000; Burt 1992, 2004; Mehra et al. 2001; Seibert et al. 2001). Disconnected networks help brokers realize value by offering them the opportunity to transfer ideas from one isolated group to another, a process that involves recognizing when solutions current in one part of the network are likely to have applications elsewhere in the network (Hargadon and Sutton 1997).

Promotions Knowing whom to consult for information and aid becomes of crucial importance at times of competition for career opportunities within organizations. In his work *"Structural Holes"* (1992), Burt has systematically explored the network effects on career advancement within the firm. According to his analysis, a configuration of network ties that creates opportunities for brokering and entrepreneurialism (i.e., a network full of structural holes) enhances career opportunities for actors competing for promotions within organizations (Burt 1992, 2005). The findings of another study on social networks and mobility at the workplace further substantiated Burt's claims that the network structures most conducive to maximizing access to information, resources, and brokerage opportunities (i.e., large, sparse networks) are a meaningful determinant of intraor-ganizational advancement (Podolny and Baron 1997).

Creativity A network "rich in structural holes" has also been found to facilitate the development of novel valuable ideas by increasing the actor's ability to merge the distinct sources of information in new ways, thus boosting individual creativity. The empirical findings suggest that between-group brokers are more likely to have a vision advantage, express ideas evaluated as valuable, and are less likely to have ideas dismissed (Burt 2004). Moreover, brokerage appears to provide the opportunity for social "gatekeeping" –

regulating the access of others to the tightly knit group one belongs to, while at the same time controlling the ways in which one's own group members learn about information coming from other groups (Burt 2004).

Power Occupying a strong or weak structural position in the network has recently been found to affect the inferences organizational actors draw about one another (Labun 2012). In particular, the empirical evidence suggests that the more an individual is constrained by the structure of his network, the more likely he is to attribute power to others. Embeddedness in networks "poor in structural holes" implies a condition of dependence and limited autonomy (Burt 1992), potentially triggering feelings of helplessness and apprehension, and thereby contributing to increased number of power attributions to other group members (Labun 2012).

Trust and gossip Trustworthy and confidential collegial environment may be advantageous when establishing informal cooperation and forming alliances against powerful third parties. According to Burt's study on trust and gossip in social networks (2001), gossip can act as a strategic tool in this process, allowing the group members to control their fellow members' actions and to weaken the reputation of competitors. The manipulation of information flow to one's own advantage becomes easier when employees occupy brokerage positions in the organizational network – connect to colleagues who are not connected with one another. The more trust exists in an employee network, the further negative gossip echoes, so that single incidents of negative gossip can have far-reaching impacts (Burt 2001). Thus, people may ensure norms of cooperation and punish the uncooperative actors (i.e., the untrustworthy group members) through gossiping – by spreading reputation-harming information about them in the broader informal network (Burt 2005).

The gender contingency effects The synthesis of the informal social network theories with research related to career advancement of women has generated interesting insights. Burt (1998) argued that women often lack sufficient legitimacy in their organizations and therefore need to "borrow" social capital (i.e., structural holes) from a strategic partner (sponsor) in order to get promoted. Whereas senior male managers indeed benefit more from a personal network rich in structural holes, women (as well as junior and non-White managers) fare better with a hierarchical network, in which a tie to an influential "sponsor" provides access to this person's entrepreneurial network (Burt 1998).

The hierarchy contingency effects Actor's position in an organizational hierarchy may serve as one of the conditions under which either structural-hole networks or cohesive networks are likely to provide the focal actor with advantages. Burt (1997) showed that the benefits of structural holes flow mainly to members of senior management. Other research has shown that the benefits of cohesion flow mainly to people occupying lower hierarchical levels in organizations, for whom issues of organizational identity and belonging remain salient for potential career advancement (Podolny and Baron 1997).

The cultural contingency effects Another contingent factor that has been found to moderate the effect of structural holes includes the specific cultural and organizational context in which the mechanisms of social capital operate. In stark contrast to the results of studies using Western samples, the empirical findings of Xiao and Tsui (2007) show that in a collectivistic Chinese culture, structural holes in an employee's career network tend to be detrimental to the employee's career development. Moreover, it has been suggested that the network consequences of social capital may differ across organizations: whereas in a market-like, low commitment organizational culture, structural holes bring positive returns to individual actors, it is network closure that appears to bring advantages to the actors by facilitating trust, reciprocity, and reputation in a clan-like, cohesive, high-commitment organization with a strong cooperative culture (Lazega 2001; Xiao and Tsui 2007).

Future Directions

The existing work utilizing the insights from Burt's structural-hole theory has recently been extended in a number of interesting directions, namely, explicit inclusion of actor characteristics, agency, and cognitions, as well as increasing use of longitudinal (dynamic) research designs. Drawing inspiration from the leading ideas of social network research, new theory and innovative hypotheses are being proposed, providing additional valuable insights.

Actor characteristics Researchers have increasingly started to incorporate personality variables in their study designs (e.g., self-monitoring) as potential predictors of variance in network outcomes (Kilduff and Krackhardt 2008; Mehra et al. 2001). People with different self-monitoring orientations have been suggested to occupy different structural positions. High self-monitors, relative to low self-monitors, tend to ingratiate themselves into distinctly different social circles of acquaintances with few links between these clusters and thereby occupy structural holes. Burt's (2005: 34) "structural entrepreneur personality index" quantifies the individual inclination to exploit social resources. Structural entrepreneurs recognize the opportunities offered by structurally advantageous positions and place themselves in the "hole" by initiating ties with actors from opposite sides of the hole who can subsequently be played off against each other. This recent work challenges the ideological refusal of the traditional social network research to acknowledge ways in which individual actors differ in their attributes and actively explore the possibility of complementary synergies between actors and network structure (Kilduff and Brass 2010). Future research on personality and social networks is likely to be generative of compelling insights on the link between individual attributes and structural outcomes.

Agency Social network research also moves forward by explicitly assuming that actors differ in their abilities, skills, and motivation to take advantage of advantageous network positions. The earlier research has shown that some individuals can choose not to reap the profits derived from their network (Burt 1992). Drawing on these earlier findings, the more recent studies suggest that the more strategically skilled group members enjoy greater access to network resources and appear to be more competent at utilizing and leveraging these resources to advance their career and performance (Ferris et al. 2007; Labun 2012; Wei et al. 2012). This work uncovers the comprehensive role that individual strategic skills may play in the process of network resource building. Following this line of analysis, the incorporation of additional types of personal or social influence skills that may affect network resource development would be an interesting and fruitful avenue for future research. Moreover, future work might consider more closely the question of how much control actors have over the networks that constrain and enable their behaviors (Kilduff and Brass 2010).

Cognition Another research area drawing from the core concepts of social network program puts a special emphasis on subjective meanings (i.e., cognitions) inherent in networks rather than on "concrete" relations such as exchanges between actors (Kilduff and Brass 2010). The cognitive social network research line has led to the conceptualization of networks as "prisms" through which others' reputations and potentials are perceived, as well as "pipes" through which resources flow (Podolny 2001). Perceived status of one's exchange partners may indeed act as a distorting prism filtering attributions concerning the focal individual (Labun 2012): having a trust relationship with a superior had a significant positive effect on other's perceptions of one's power. The role of cognitions inherent in networks was further accentuated in a study demonstrating that individuals tend to bias perceptions to highlight small world features of clustering and connectivity (Kilduff et al. 2008): across four different organizational friendship networks, people have been found to perceive more "small worldedness" than was actually the case, including the perception of more network clustering than actually existed and the attribution of more popularity and brokerage to the perceived popular than to the actually popular.

Network dynamics Finally, longitudinal research designs that allow considering and effectively addressing the dynamic nature of networks is likely to drive the social network research program forward. The very recent analytical developments (Snijders et al. 2010) allow unraveling and tackling the intriguing novel phenomena concerning interpersonal network change, coevolution of networks, and individual behavior (e.g., friendship, music preferences, and alcohol consumption (Steglich et al. 2006); friendship and smoking behavior (Mercken et al. 2010)), as well as different types of networks (e.g., friendship and gossip (Ellwardt et al. 2012); friendship and power (Labun 2012)). For example, the friendship and power study showed that power perceptions breed friendship (Labun 2012). Through a power attribution to a colleague, an individual may signal his or her trust in the colleague's competence, thereby triggering a friendship nomination from/facilitating friendship with him or her. The multiplex effect showed up also when analyzing the conditions that influence the formation of social ties (i.e., friendship) to the high-power organizational actors. However, in this case, the relationship between two networks appeared to depend on individual's strategic orientations (Labun 2012). This emergent research contributes to a better understanding of the coevolution of multiplex networks as well as networks and individual behavior, thereby allowing us to fully grasp the antecedents, dynamics, and consequences of the "informal organization."

Using a game theoretic model of network formation, Buskens and Van de Rijt (2008) confirm Burt's own speculation that when the monopoly on structural entrepreneurship is lifted, structural advantages most likely disappear (Burt 2005): when everyone strives for structural holes, no one will be able to maintain a structural advantage in the long run (Buskens and Van de Rijt 2008).

It would be interesting to perform further empirical studies in different types of organizational settings to help elucidate the dynamics of structural holes. The ongoing methodological advancements and the theoretical insights gained from the above-mentioned recent work are certainly beneficial for the future development and possible extension of existing structural-hole research.

Cross-References

- ▶ Actor-Based Models for Longitudinal Networks
- ▶ Centrality Measures
- ▶ Exchange Networks
- ▶ Exponential Random Graph Models
- ▶ Futures of Social Networks: Where Are Trends Heading?
- ▶ Game Theory and Social Networks
- ▶ Personal Networks: Ties, Space, and the Internet
- ▶ Social Capital
- ▶ Social Influence Analysis
- ▶ Trust in Social Networks

References

Ahuja G (2000) Collaboration networks, structural holes, and innovation: a longitudinal study. Adm Sci Q 45:425–455

Balkundi P, Wang L, Harrison DA (2009) Bridging the gap: consequences of structural hole spanning at multiple levels. Working paper, SUNY, Buffalo

Bonacich P (1987) Power and centrality: a family of measures. Am J Sociol 92:1170–1182

Brass DJ (2011) A social network perspective on organizational psychology. In: Kozlowski SWJ (ed) The Oxford handbook of organizational psychology. Oxford University Press, New York

Burt RS (1982) Toward a structural theory of action. Academic, New York

Burt RS (1992) Structural holes: the social structure of competition. Harvard University Press, Cambridge

Burt RS (1997) The contingent value of social capital. Adm Sci Q 42:339–365

Burt RS (1998) The gender of social capital. Ration Soc 10:5–46

Burt RS (2000) The network structure of social capital. Res Organ Behav 22:345–423

Burt RS (2001) Bandwidth and echo: trust, information, and gossip in social networks. In: Casella A, Rauch JE (eds) Networks and markets: contributions from economics and sociology. Russell Sage, New York, pp 30–74

Burt RS (2004) Structural holes and good ideas. Am J Sociol 110:349–399

Burt RS (2005) Brokerage and closure: an introduction to social capital. Oxford University Press, Oxford

Ferris GR, Treadway DC, Perrewe PL, Brouer RL, Douglas C, Lux S (2007) Political skill in organizations. J Manag 33:290–320

Freeman LC (1979) Centrality in social networks: conceptual clarification. Soc Netw 1:215–239

Granovetter M (1973) The strength of weak ties. Am J Sociol 6:1360–1380

Hargadon AB, Sutton RI (1997) Technology brokering and innovation in a product development firm. Adm Sci Q 42:716–749

Kilduff M, Brass DJ (2010) Organizational social network research: core ideas and key debates. Acad Manag Ann 4:317–357

Kilduff M, Krackhardt D (2008) Interpersonal networks in organizations. Cambridge University Press, Cambridge

Labun A (2012) Social networks and informal power in organizations. ICS Dissertation series, Groningen, p 194

Lazega E (2001) The collegial phenomenon: the social mechanisms of cooperation among peers in a corporate law partnership. Oxford University Press, Oxford

McCall MW (1979) Power, authority, and influence. In: Kerr S (ed) Organizational behavior. Grid, Columbus, pp 185–206

Mechanic D (1962) Sources of power of lower participants in complex organizations. Adm Sci Q 7:349–364

Mehra A, Kilduff M, Brass DJ (2001) The social networks of high and low self-monitors: implications for workplace performance. Adm Sci Q 46:121–146

Pettigrew AM (1972) Informational control as a power resource. Sociology 6:187–204

Pfeffer J (1981) Power in organizations. Pitman, Marshfield

Podolny JM (2001) Networks as the pipes and prisms of the market. Am J Sociol 107:33–60

Podolny JM, Baron JN (1997) Resources and relationships: social networks and mobility in the workplace. Am Sociol Rev 62:673–693

Seibert SE, Kraimer ML, Liden RC (2001) A social capital theory of career success. Acad Manag J 44:219–237

Wei L, Chiang FFT, Wu L (2012) Developing and utilizing network resources: roles of political skill. J Manag Stud 49:381–402

Xiao Z, Tsui AS (2007) When brokers may not work: the cultural contingency of social capital in Chinese high-tech firms. Adm Sci Q 52:1–31

Recommended Reading

Buskens V, Van de Rijt A (2008) Dynamics of networks if everyone strives for structural holes. Am J Sociol 114:371–407

Ellwardt L, Steglich CEG, Wittek R (2012) The co-evolution of friendship and gossip in workplace social networks. Soc Netw 34:623–633

Kilduff M, Crossland C, Tsai W, Krackhardt D (2008) Network perceptions versus reality: a small world after all? Organ Behav Hum Decis Process 107:15–28

Mercken L, Snijders TAB, Steglich CEG, Vartiainen E, de Vries H (2010) Dynamics of adolescent friendship networks and smoking behavior. Soc Netw 32:72–81

Snijders TAB, Van de Bunt GG, Steglich CEG (2010) Introduction to stochastic actor-based models for network dynamics. Soc Netw 32:44–60

Steglich CEG, Snijders TAB, West P (2006) Applying SIENA: an illustrative analysis of the co-evolution of adolescent's friendship networks, taste in music, and alcohol consumption. Methodology 2:48–56

Structural Measure

▶ Role Discovery

Structural Roles

▶ Querying Volatile and Dynamic Networks

Structuralism

▶ Network Analysis in French Sociology and Anthropology

Subgraph Count

▶ Motif Analysis

Subgraph Discovery

▶ Subgraph Extraction for Trust Inference in Social Networks

Subgraph Evolution

▶ Motif Analysis

Subgraph Extraction for Trust Inference in Social Networks

Yuan Yao[1], Hanghang Tong[2], Feng Xu[1] and Jian Lu[1]
[1]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, Jiangsu, China
[2]CUNY City College, New York, NY, USA

Synonyms

Interaction network; Subgraph discovery; Trust evaluation; Trust network; Trust prediction

Glossary

Social network	A graph in which the nodes represent the participants in the network and the edges represent relationships
Trust-based social network	A directed weighted graph in which the nodes represent the participants in the network, the edges represent trust relationships and the weight on each edge indicates the local trust value derived from the historical interactions
Trust inference	A mechanism to build new trust relationships based on existing ones
Subgraph	A subgraph of graph *G* is a graph whose node set is a subset of that of *G*, and whose edge set is a subset of that of *G* restricted to the node subset
Subgraph extraction	Discovery of a subgraph from a whole graph

Definition

Trust-based social networks might contain a large amount of redundant information, making existing trust inference suffer from the scalability and usability issues. Therefore, it is natural to apply subgraph extraction as an intermediate step to speed up as well as to interpret the trust inference process.

Introduction

Trust inference, which aims to infer a trustworthiness score from the trustor to the trustee in the underlying social network, is an essential task in many real-world applications including e-commerce (Xiong and Liu 2004), peer-to-peer networks (Kamvar et al. 2003), and mobile ad hoc networks (Buchegger and Le Boudec 2004).

To date, many trust inference algorithms have been proposed, which can be categorized into two main classes (see the next section for a review): (a) path-based inference (Mui et al. 2002; Wang and Singh 2006; Hang et al. 2009; Wang and Wu 2011) and (b) component-based inference (Guha et al. 2004; Massa and Avesani 2005; Ziegler and Lausen 2005; Zhou and Hwang 2007).

Despite their own success, most of the existing inference algorithms have two limitations. The first challenge lies in *scalability* – many existing algorithms become very time-consuming or even computationally infeasible for the graphs with more than thousands of nodes. Additionally, some algorithms assume the existence of a subgraph while how to construct such a subgraph remains an open issue (Wang and Wu 2011). The second challenge is the *usability* of the inference results. Most, if not all, of the existing inference algorithms output an abstract numerical trustworthiness score. This gives a quantitative measure of *to what extent* the trustor should trust the trustee but gives few cues on *how* the trustworthiness score is inferred. This usability/interpretation issue becomes more evident when the size of the underlying graph increases, since we cannot even display the entire graph to the end users (see Fig. 9 for an example).

In this article, we propose subgraph extraction to address these challenges. The core of our subgraph extraction consists of two stages: *path selection* and *component induction*. In the first (path selection) stage, we extract a few, important paths from the trustor to the trustee. In the second (component induction) stage, we propose a novel evolutionary algorithm to generate a small subgraph based on the extracted paths. The outputs of these two stages are then used as an intermediate step to speed up the path-based inference and component-based inference algorithms, respectively. Our experimental evaluations on real graphs show that the proposed method can significantly accelerate existing trust inference algorithms (up to 2,400× speedup) while maintaining high accuracy (P-error is less than 0.04). In addition, the extracted subgraph provides an intuitive way to interpret the resulting trustworthiness score by presenting a concise summarization on the relationship from the trustor to the trustee. To the best of our knowledge, we are the first to propose subgraph extraction for trust inference. We believe that our work can improve most of the existing trust inference algorithms by (1) scaling up as well as (2) delivering more usable (i.e., interpretation-friendly) inference results to the end users.

Historical Background

We review the historical background in this section, which can be categorized into two parts: trust inference algorithms and subgraph extraction.

Trust Inference

We categorize existing trust inference algorithms into two main classes: path-based trust inference and component-based trust inference.

In the first class of path-based inference, trust is propagated along a path from the trustor to the trustee, and the propagated trust from multiple paths can be combined to form a final trustworthiness score. For example, Wang and Singh (2006, 2007) as well as Hang et al. (2009) propose operators to concatenate trust along a path and aggregate trust from multiple paths. Liu et al. (2010) argue that not only trust values but social relationships and recommendation role are important for trust inference. However, these algorithms are only suitable for small networks due to their complexity. Some other path-based trust inference algorithms, such as Mui et al. (2002) and Wang and Wu (2011), assume the existence of an extracted subgraph while how to construct such a subgraph remains an open issue (Wang and Wu 2011).

In the second class of component-based inference, EigenTrust Kamvar et al. (2003) tries to compute an objective trustworthiness score for each node in the graph. In contrast to EigenTrust, our main focus is to provide support for subjective trust metrics where different trustors can form different opinions on the same trustee. In contrast to path-based trust inference algorithms, there is no explicit concept of paths in component-based trust inference. Instead, existing subjective trust algorithms, including Guha et al. (2004), Massa and Avesani (2005), Ziegler and Lausen (2005), and Nordheimer et al. (2010), take the initial graph as input and treat trust as random walks on a Markov chain or on a graph (Richardson et al. 2003). For example, in MoleTrust (Massa and Avesani 2005) and Appleseed (Ziegler and Lausen 2005), trust propagates along the edges according to the trust values on the edges. Our subgraph extraction method not only can speed up many of these algorithms but also can provide interpretive result which is not considered by the existing algorithms.

Overall, our subgraph extraction is motivated to address the two common challenges (i.e., scalability and usability) shared by most of these existing trust inference algorithms.

Subgraph Extraction

Several end-to-end subgraph extraction algorithms are developed to solve different problems.

In the field of graph mining, Faloutsos et al. (2004) refer to the idea of electrical current where trust relationships are modeled as resistors and try to find a connection subgraph that maximizes the current flowing from source to target. Later, Tong et al. (2007) generalize the connection subgraph to directed graphs and use the subgraph to compute proximities between nodes. Similar to Tong et al., Koren et al. (2006) also try to induce a subgraph for proximity computation. In addition, Koren et al. search the k-shortest paths to provide a basis for measuring the proximity.

Recently, several algorithms are proposed for reliable subgraph extraction. Among them, Monte Carlo pruning (Hintsanen and Toivonen 2008) measures the relevance of each edge by Monte Carlo simulations and tends to remove the edge of lowest relevance one by one. The most related work is perhaps the randomized Path Covering algorithm (Hintsanen et al. 2010) which also consists of two stages of path sampling and subgraph construction. However, both Monte Carlo pruning and Path Covering tend to find a subgraph with highest probability to be connected, while we aim to find a subgraph to address the scalability and usability issues in trust inference.

The Proposed Subgraph Extraction Method

In this section, we first formalize the subgraph extraction problem for trust inference in social networks and then introduce our proposed solution which consists of two stages: path selection and component induction.

Problem Definition

Following the standard notations in the existing trust inference algorithms, we model the trust relationships in social networks as a weighted directed graph (Barbian 2011; Yao et al. 2011). The nodes of the graph represent the participants in the network, and the weight on each edge indicates the local trust value derived from the historical interactions.

We then categorize the existing trust inference algorithms into two major classes: *path-based trust inference* and *component-based trust inference*.

Definition 1 Path-Based Trust Inference Path-based trust inference includes the approaches, which are started by the trustor, to evaluating the trustworthiness of the trustee, through a set of paths from the trustor to the trustee in the network.

Definition 2 Component-Based Trust Inference Component-based trust inference includes the approaches, which are started by the trustor, to evaluating the trustworthiness of the trustee, through a connected component from the trustor to the trustee in the network.

Both classes belong to the subjective trust metrics (Ziegler and Lausen 2005), where different trustors can form different opinions on the same trustee. Accordingly, path-based trust inference such as Mui et al. (2002), Wang and Singh (2006), Liu et al. (2010), Hang et al. (2009), and Wang and Wu (2011) and component-based inference such as (Guha et al. 2004), Massa and Avesani (2005), Ziegler and Lausen (2005), and Zhou and Hwang (2007) all belong to trust inference algorithms. Although the main focus of this article is on the subjective metrics, our proposed subgraph extraction can also be applied to the objective trust metrics.

Despite the success of most existing inference algorithms, they share the scalability and usability limitations. To address these issues, we propose subgraph extraction for trust inference. The core of our subgraph extraction consists of two stages. The first stage, which serves for path-based trust inference, selects a set of paths from the trustor to the trustee. The second stage aims to produce a connected component between the trustor and the trustee for component-based trust inference. In addition, the second stage of our subgraph extraction produces a relatively small subgraph which can be clearly displayed and help the end user better understand the inference result.

We now formally define the subgraph extraction problem for trust inference. In accordance to the corresponding two stages, the problem is divided into two subproblems: *path selection problem* and *component induction problem*.

Definition 3 Path Selection Problem Given: a weighted directed graph $G(V, E)$; two nodes s; $t \vee V$; and an integer K

Find: a set C with K paths from s to t that minimizes the error function $f(C)$

Definition 4 Component Induction Problem Given: a set C of paths from s to t and an integer N

Find: an induced component $H.(V', E')$ with at most N edges that minimizes the error function g

(H), where $V' \subseteq \{v|\ (u,\ v)$ or . $(v,\ u) \in P,\ P \in C\}$ and $E' \subseteq \{e|e \in P\ ;\ P \in$ C)

We next discuss the error function in the definitions. The error function $f(C)$ in Definition 3 indicates the goodness of the extracted paths, and $f(C)$ reaches its minimum value when C contains all the possible paths from s to t. Similarly, the error function $g(H)$ in Definition 4 reaches its minimum value if $H = G$. In this article, we use P-error, which is defined as follows, as the error function for both subproblems, i. e., $f = g =$ P-error.

Definition 5 P-error For a given trustor-trustee pair, the error function P-error is defined as

$$P - \text{error} = |\ p_{\text{sub}} - p_{\text{whole}}\ |\ ,$$

where p_{sub} is the trustworthiness score inferred from the subgraph and p_{whole}, which serves as a ground truth, is the trustworthiness score inferred from the whole graph.

Path Selection

In the path selection stage, we aim to extract a few paths from the trustor to the trustee as an intermediate step to speed up path-based trust inference algorithms. These extracted paths will also serve as the input for the component induction stage.

There are two preprocessing steps in our extraction method. First of all, trust is interpreted as the probability by which the trustor expects that the trustee will perform a given action. This interpretation of trust is adopted by many existing trust inference algorithms, and it allows trust to be multiplicatively propagated along a path (Liu et al. 2010). Second, we transform probability into weight by negative logarithm. Namely, the local trust value on the edge e is interpreted as probability $p(e)$, and the probability $p(e)$ is transformed to weight $w(e) = -\log\ (p(e))$. Based on these two steps, the weight of a path P can be presented as

$$\begin{aligned} w(P) &= \sum_{e \in P} -\log(p(e)) = -\log\left(\prod_{e \in P} p(e)\right) \\ &= -\log(Pr(P)). \end{aligned}$$

As a result, finding a path of high trustworthiness in the original network is equivalent to finding a short path in the transformed network. We will use this transformed weighted graph $G(V,\ E)$ as the input of our method.

Then, the path selection problem becomes to extract top-k short paths from the trustor to the trustee in the transformed graph $G(V,\ E)$. Many existing algorithms can be plugged into this stage, such as Yen's k-shortest loopless paths (KS) (Yen 1971), and path sampling (PS) (Hintsanen et al. 2010). In our experiments, we found that KS algorithm performs best even if the multiplicative property of the interpretation does not hold, and we therefore recommend KS in this stage. A brief skeleton of the KS algorithm is shown in Algorithm 1, and the detailed algorithms for KS and PS are presented in the appendix for completeness.

Algorithm 1 KS Algorithm (See the Appendix for the Details)

```
Input: Weighted directed graph $G(V,\ E)$, two nodes $s,\ t \in V$, and a parameter $K$ of path number
Output: Set $C$ with $K$ paths from $s$ to $t$
$C$ = k-shortest($G$, $s$, $t$, $K$)
return $C$
```

Algorithm Analysis

The worst-case time complexity of KS is $O(K|V|\ (|E| + |V|\ \log\ |V|))$, which is known as the best result to ensure that k-shortest loopless paths can be found in a directed graph (Hershberger et al. 2007). However, the actual wall-clock time of KS on many real graphs is often much better than such worst-case scenario (Martins and Pascoal 2003). In fact, based on our experiments, we find that it empirically scales near linearly wrt the graph size $|V|$ in the chosen datasets.

Component Induction

In the component induction, we take the output of path selection stage (i.e., a set of K paths) as input and output a small connected component from the trustor to the trustee. The output of the component induction stage not only acts as an intermediate step to speed up component-based trust inference algorithms but also helps to improve the usability of trust inference by interpreting the inference results for the end users. Notice that although our upcoming proposed algorithm EVO could also be applied on the whole graph, we do not recommend it in practice for the following two reasons: (1) most trustworthy paths have already been captured by the path selection stage (i.e., KS), and (2) applying EVO on the whole graph would cost more memory and time to achieve high accuracy. We will present more detailed experimental evaluations to validate this in the next section.

In general, our proposed EVO algorithm (shown in Algorithm 2) belongs to the so-called evolutionary methods (Bäck 1996). It aims to minimize P-error under the constraint of edge number. The input component $G^c.(V^c; E^c)$ is directly induced from the set C of paths from s to t, where $V^c = \{v| .(u, v) \in P or(v, u) \in P; P \in C\}$and $E^c = \{e| e \in P; P \in C\}$. There are two implicit parameters in the algorithm, i.e., the initial vector number m and iteration number iter.

Algorithm 2 EVO Algorithm

```
Input: Set $C$ of paths from $s$ to $t$ and the directly induced component $G^C(V^C, E^C)$, as well as a constraint $N$ of the edge number
Output: Induced component $\boldsymbol{H}(\boldsymbol{V}', \boldsymbol{E}')$ with at most $\boldsymbol{N}$ edges
define 0/1 vector $\boldsymbol{B}$ of size $|E^C|$ where each element in $\boldsymbol{B}$ stands for the existence of a corresponding edge in $\boldsymbol{G}^C$
initialize $m$ vectors $S \leftarrow \{B_1, B_2,\ldots, B_m\}$, with at most $N$ 1-bits for each vector
while $iter > 0$ do
  for each vector $\boldsymbol{B}_i$ in $\boldsymbol{S}$ do
    repeat
      mutate $\boldsymbol{B}_i$ to $\boldsymbol{B}_{i+m}$ with mutation probability $1/|E^C|$
    until the number of 1-bits in $B_{i+m} \leqslant N$
  end for
  compute P-error results for the $2m$ vectors $\{\boldsymbol{B}_1, \boldsymbol{B}_2,\ldots,\boldsymbol{B}_{2m}\}$
  $S \leftarrow$ the best $m$ vectors from the $2m$ ones
  $iter \leftarrow iter - 1$
end while
$\boldsymbol{B}_{final} \leftarrow$ the best vector in $\boldsymbol{S}$
return the corresponding component $\boldsymbol{H}(\boldsymbol{V}', \boldsymbol{E}')$ of $\boldsymbol{B}_{final}$
```

We now explain EVO in detail. The first step of EVO is to establish a one-to-one correspondence between the edges in G^c and the elements in vector B. Each element of B is a 0/1 bit where 1 indicates that the corresponding edge exists and 0 indicates otherwise. The vector has exactly $|E^c|$ bits where $|E^c|$ is the edge number of G^c. In the second step, the algorithm generates m vectors $B_1; B_2,\ldots, B_m$, and each of them has at most N 1-bits. In our implementation, we apply a constant-time search in C to find a subset of paths with minimized P-error. In the following steps, EVO adopts *mutation* on each of these vectors to separately generate m new vectors $B_{m+1}\ B_{m+2};\ldots; B_{2m}$. In the mutation from B_i to B_{i+m}, each bit of B_i is changed with probability $1/|E^c|$. If the resulting vector has more than N 1-bits, the mutation operation is redone. The error function, which is P-error in our case, is then computed on each of these $2m$ vectors, and the m vectors with smallest P-error are kept to the next iteration. For efficiency, the P-error computation on vector B herein means computing the P-error between $G^c(V^c; E^c)$ and the component corresponding to the vector B. Namely, we use the input component $G^c(V^c; E^c)$ as an approximation of the ground truth in this stage.

Algorithm Analysis

The time complexity of EVO is summarized in the following lemma, which basically says that the expected time complexity of EVO scales linearly wrt both initial vector number m and iteration number iter.

Lemma 1 *The average-case time complexity of EVO is* $O(\text{iter} \cdot m(| E^c| /N + \theta))$, *where* θ *is the time complexity of the error function computation.*

Proof In the mutation step of EVO, with mutation probability $1/|E^c|$, the expected number of bit changes is 1. This step is expected to be redone only when the number of 1-bits is N and the bit change is from 0 to 1. Under this condition, the probability of bit change from 0 to 1 is $(|E^c| - N)/|E^c|$. Therefore, the expected iteration number of the mutation step is $|E^c|/N$. Therefore, the whole expected time complexity of EVO is $O(\text{iter}(m\cdot|E^c|/N + m\theta)) = \text{O}(\text{iter} \cdot m(|E^c|/\text{N} + \theta))$, which completes the proof.

Experimental Evaluation

In this section, we first describe the experimental setup and then present the results.

Experimental Setup

We first describe the datasets and the representatives of path-based and component-based trust inference algorithms. All algorithms are implemented in Java and have been run on a T400 ThinkPad with 1,280 m jvm heap space. Few other activities are done during the experiments.

Datasets Description

We use the advogato (http://www.trustlet.org/wiki/Advogato_dataset) datasets in our experiments, because advogato is a trust-based social network and it contains multilevel trust assertions. There are four levels of trust assertions in the network, i.e., "Observer," "Apprentice," "Journeyer," and "Master." These assertions can be mapped into real numbers in [0,1]. In our experiments, we map "Observer," "Apprentice," "Journeyer," and "Master" to 0.1, 0.4, 0.7, and 0.9, respectively. The statistics of the datasets is shown in Table 1.

Trust Inference Representatives

To evaluate our subgraph extraction method, we need to apply trust inference algorithms on the whole graph and on our extracted subgraph to compare their effectiveness and efficiency. We chose *CertProp* (Hang et al. 2009) as the representative of path-based inference algorithms, and *Appleseed* (Ziegler and Lausen 2005) as the representative of component-based inference algorithms.

P-error computation in CertProp needs to first compute the ground truth p_{whole} by finding all paths from the trustor to the trustee in the whole graph. This computation, however, easily causes the overflow of the jvm heap space even on the advogato-1 graph. Following the suggestions in the original CertProp (Hang et al. 2009), we apply the fixed search strategy and search all paths whose length is not longer than seven as an approximation of the ground truth. For CertProp, we define *collapsed samples* as the trustor-trustee pairs of which the P-error computation either exceeds the range of Java.lang.Double or runs out of the jvm heap space. We randomly select 100 node pairs out of 122 samples, where the rest 22 of them are collapsed samples. Our experimental results are all based on the average of these 100 samples. Notice that, as discussed in the path selection section, the multiplicative property of the probability interpretation does not hold in CertProp. As to Appleseed, we apply linear normalization on the outputs, since the algorithm can produce arbitrary trustworthiness scores.

Subgraph Extraction for Trust Inference in Social Networks, Table 1 High-level statistics of advogato datasets

Graph	Nodes	Edges	Avg. degree	Avg. clustering	Avg. diameter	Date
Advogato-1	279	2,109	15.1	0.45	4.62	2000-02-05
Advogato-2	1,261	12,176	19.3	0.36	4.71	2000-07-18
Advogato-3	2,443	22,486	18.4	0.31	4.67	2001-03-06
Advogato-4	3,279	32,743	20.0	0.33	4.74	2002-01-14
Advogato-5	4,158	41,308	19.9	0.33	4.83	2003-03-04
Advogato-6	5,428	51,493	19.0	0.31	4.82	2011-06-23

Experimental Results

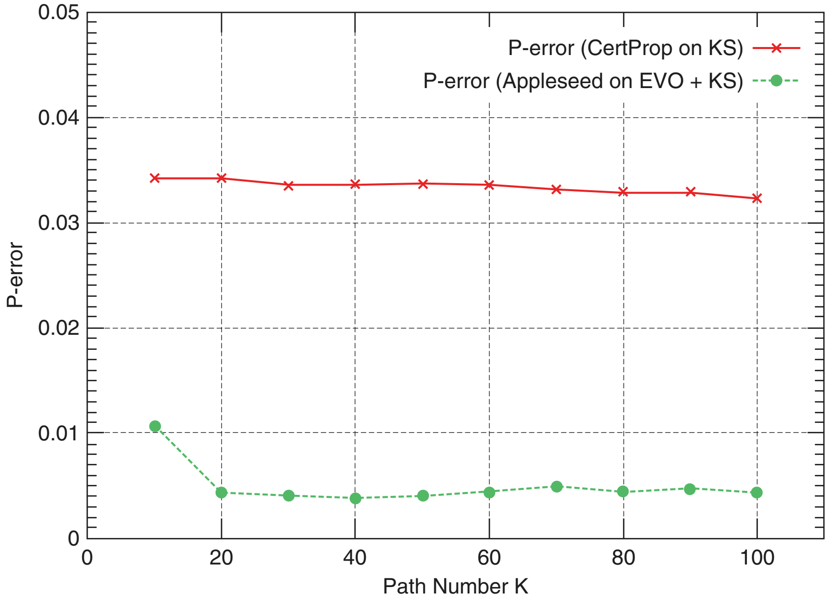
We now present the experimental results of our subgraph extraction method. In our experiments, the effectiveness, efficiency comparisons, and interpretation results are all based on the advogato-1 graph, as we found CertProp on the whole graph becomes computationally infeasible on all the other larger datasets. We evaluate the scalability of our method using all the datasets (i.e., advogato-1 to advogato-6). As for EVO, we set $m = 5$ and iter D 10 unless otherwise specified. The edge constraint N is set as $K/2$.

Effectiveness

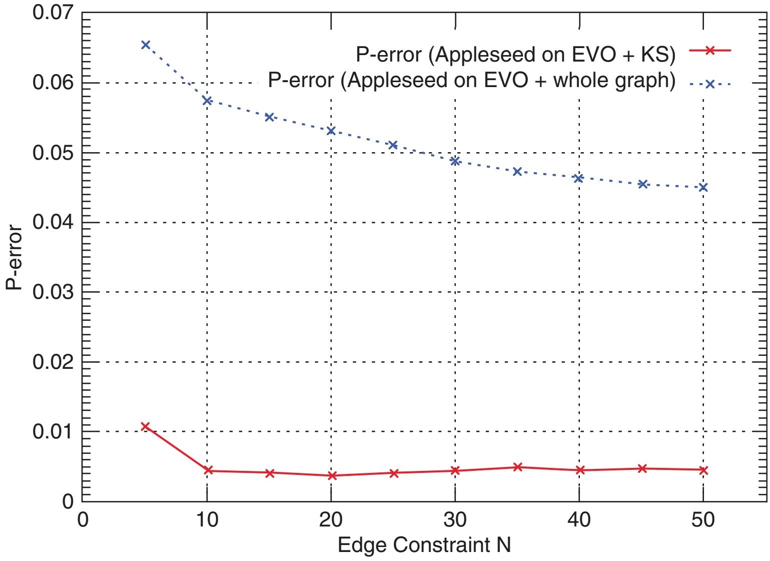
For effectiveness, we first study how CertProp and Appleseed perform on the KS subgraph (the output of path selection stage) and EVO subgraph (the output of component induction stage), respectively. The results are shown in Fig. 1. We can observe that all the P-error values of CertProp and Appleseed are less than 0.04, indicating that our extracted subgraphs, which are based on a small set of carefully selected paths and an evolutionary strategy, provide high accuracy for the trust inference algorithms.

Remember that the proposed EVO is always applied on the output of the path selection stage (referred to as “EVO + KS”). Here, for comparison purpose, we also apply EVO on the entire graph (referred to as “EVO + whole graph”). With the same parameter setting, the results are shown in Fig. 2. It can be seen that EVO on KS outperforms EVO on the whole graph. The reason is as follows. As an evolutionary algorithm, EVO (either on KS or on the entire graph) finds a local minima. By restricting the search space to those highly trustworthy paths (i.e., the output of KS), it converges to a better local minima in terms of P-error.

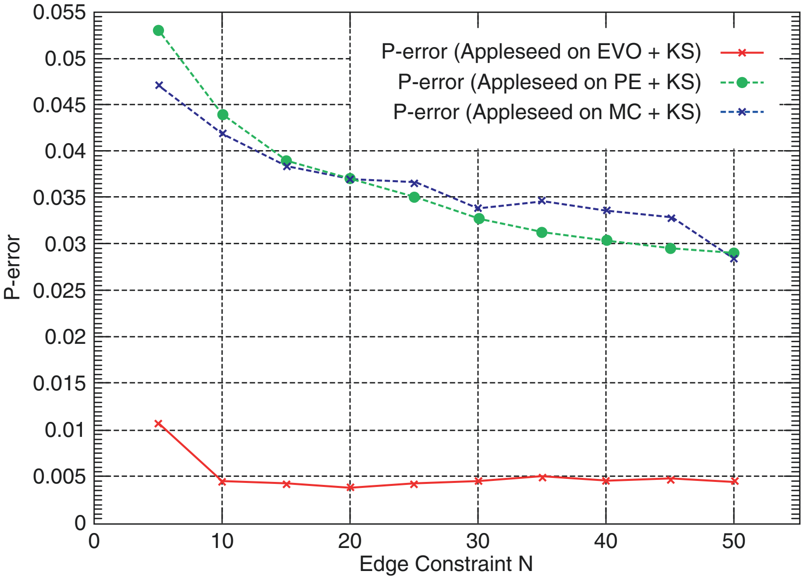
Finally, to compare EVO with existing component induction algorithms, we implement the *Monte Carlo pruning* (*MC*) method (Hintsanen and Toivonen 2008) and the *proximity extraction* (*PE*) method (Koren et al. 2006). As mentioned in the historical background, MC is proposed for the reliable subgraph extraction problem. The key idea of MC is to measure a relevance score for each edge by Monte Carlo simulations and then remove the edges of lowest relevance scores. On the other hand, PE is proposed for the proximity



Subgraph Extraction for Trust Inference in Social Networks, Fig. 1 Effectiveness of our subgraph extraction method with edge number constraint $N =$ D $K/2$. In all cases, the P-error is less than 0.04



Subgraph Extraction for Trust Inference in Social Networks, Fig. 2 Comparison of EVO on KS versus EVO on the whole graph with edge number constraint $N = K/2$. EVO on KS outperforms EVO on the whole graph

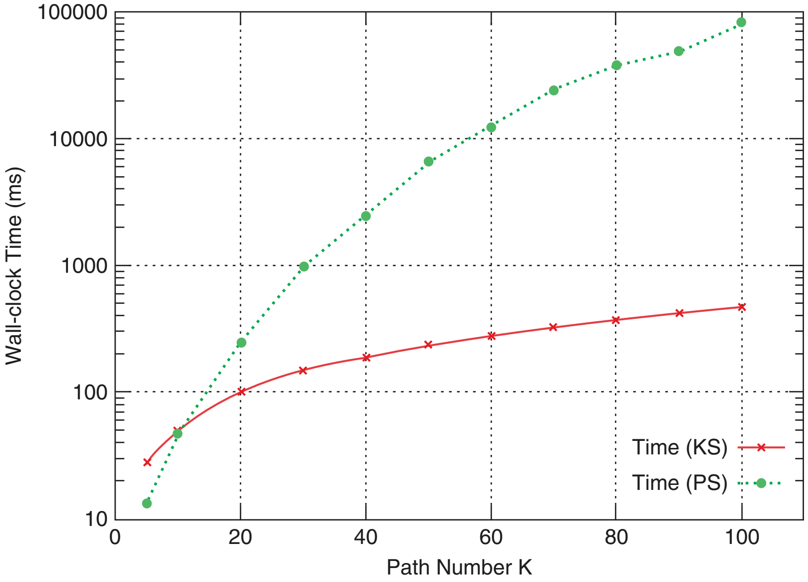


Subgraph Extraction for Trust Inference in Social Networks, Fig. 3 Comparison of different component induction algorithms with edge number constraint $N = K/2$. EVO outperforms the existing component induction algorithms

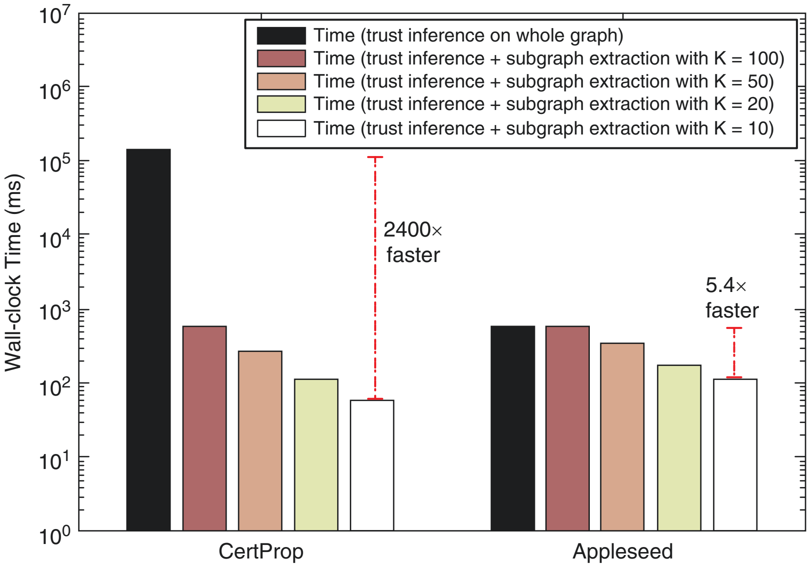
computation problem where a small set of paths are selected to maximize the proposed proximity objective function. We plot the comparison results in Fig. 3. Again, we can see that EVO outperforms both MC and PE wrt P-error. In fact, MC induces a component by successively deleting edges (edge-level component induction), while PE only selects a smaller set of paths (path-level component induction). Our EVO algorithm outperforms MC and PE because EVO combines these two levels of component induction by searching a smaller set of paths in the initial step and then evolving the resulting component on the edge level.

Efficiency

First, we compare the different algorithmic choices in the path selection stage. To this end, we compare the wall-clock time of KS with an



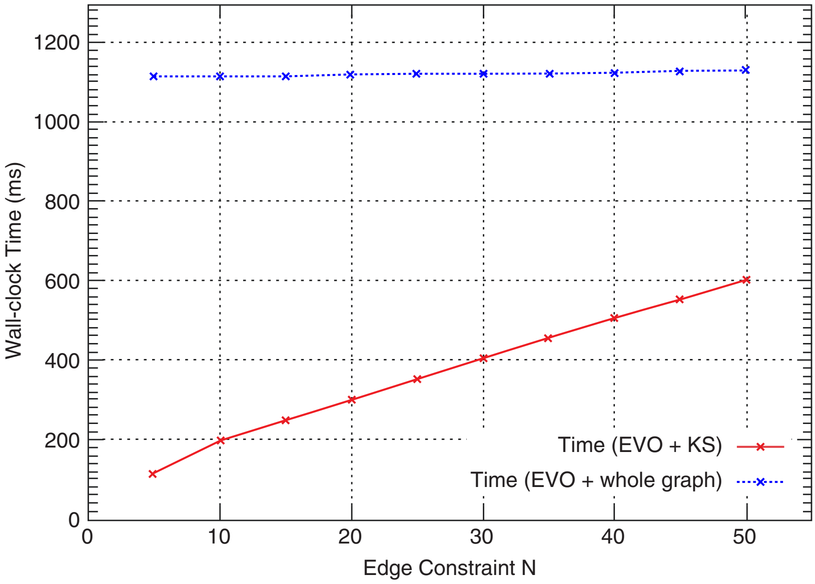
Subgraph Extraction for Trust Inference in Social Networks, Fig. 4 The average wall-clock time of KS and PS. The average wall-clock time of KS is much faster than that of PS when *K* is greater than 30



Subgraph Extraction for Trust Inference in Social Networks, Fig. 5 The average wall-clock time of CertProp on KS and Appleseed on KS + EVO. We achieve up to 2,400× speedup

alternative path selection algorithm *path sampling* (*PS*) (Hintsanen et al. 2010). The results are shown in Fig. 4. Note that the y-axis is of log scale. As we can see from the figure, although PS is slightly faster than KS when $K = 5$, the wall-clock time of PS is much longer than that of KS when K is greater than 30. For example, the wall-clock time of PS is more than 170× longer than that of KS when $K = 100$. Therefore, we recommend using KS for path selection.

Next, we study the computational savings by applying the proposed subgraph extraction as the intermediate steps for the existing trust inference algorithms. To this end, we report the wall-clock time of CertProp on the output of the path selection stage and Appleseed on the output of the component induction stage, respectively. The results are shown in Fig. 5 where the y-axis is of log scale. Notice that the reported time includes the wall-clock time of both subgraph extraction



Subgraph Extraction for Trust Inference in Social Networks, Fig. 6 The average wall-clock time of EVO on KS and EVO on the whole graph with edge number constraint $N = K/2$. EVO on KS is much faster

and trust inference. In the figure, we also plot the wall-clock time of CertProp and Appleseed on the entire graph for comparison. We can see that our subgraph extraction method saves the wall-clock time for both path-based trust inference and component-based trust inference, especially for the former one. For example, when $K = 10$, our subgraph extraction method achieves up to 2,400× and 5.4× speedup for Cert-Prop and Appleseed, respectively. Even when K grows to more than 60, our method can still achieve 200–400× speedup for CertProp.

Next, we compare the efficiency between applying EVO on KS and applying EVO on the whole graph. With $N = K/2$, the results are shown in Fig. 6. As we can see, the wall-clock time of EVO on KS (which includes the wall-clock time of both EVO and KS) is much faster than EVO on the whole graph. Together with the effectiveness results (Fig. 2), we recommend running EVO on the KS subgraph in practice.

Finally, we evaluate how the parameters m and iter in EVO affect the wall-clock time. In this experiment, we fix $K = 20$ and $N = 10$, and the results are shown in Fig. 7. We can observe that the wall-clock time of EVO scales linearly wrt iter for any fixed m, which is consistent with the time complexity analysis shown before.

Scalability

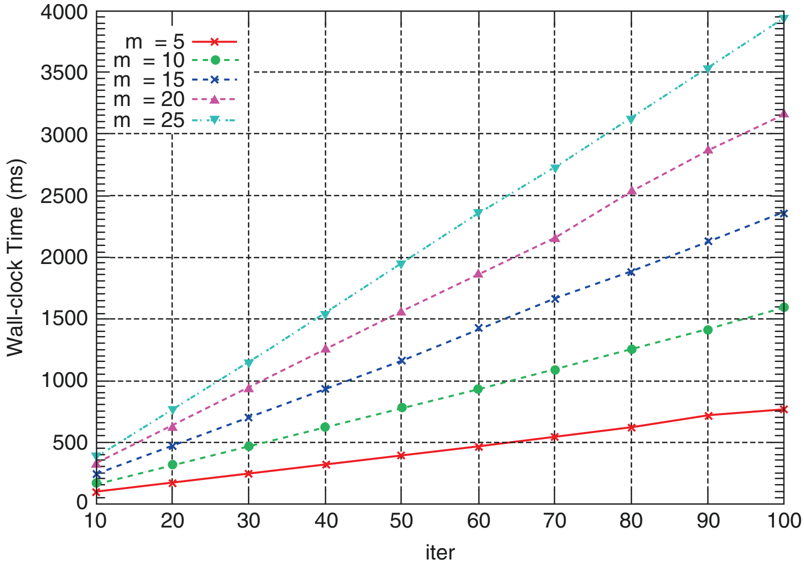
We now evaluate the scalability of our method on datasets advogato-1 to advogato-6. Figure 8 shows the results, where the y-axis is of log scale. In this experiment, we fix $K = 10$ and $N = 5$.

We can observe from the figure that even on the largest graph of 5,428 nodes and 51,293 edges, KS can help to infer the trustworthiness score within 25 s. In addition, KS scales near linearly wrt the underlying graph size. As to EVO, the wall-clock time stays stable in spite of the growth of the graph size. The reason is that $|E^c|$ scales near linearly to K due to many overlapping edges and N is set to $K/2$. Consequently, $|E^c|$/N is close to a constant, and the time complexity of EVO can be approximated to $O(\text{iter} \cdot m \cdot \delta)$.

Usability/Interpretation

Another important goal of the proposed EVO is to improve the usability in trust inference by interpreting the inferred trustworthiness score for end users. An illustrative example is shown in

Subgraph Extraction for Trust Inference in Social Networks, Fig. 7 The average wall-clock time of EVO with $K = 20$ and $N = 10$. EVO scales linearly wrt iter for the fixed m



Subgraph Extraction for Trust Inference in Social Networks, Fig. 8 The scalability of our subgraph extraction method. KS scales near linearly wrt the graph size, while the wall-clock time of EVO stays almost constant

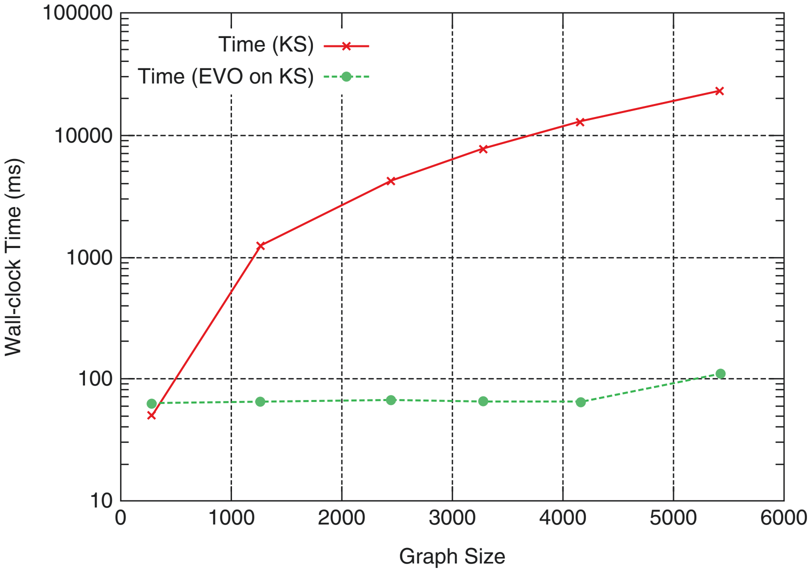
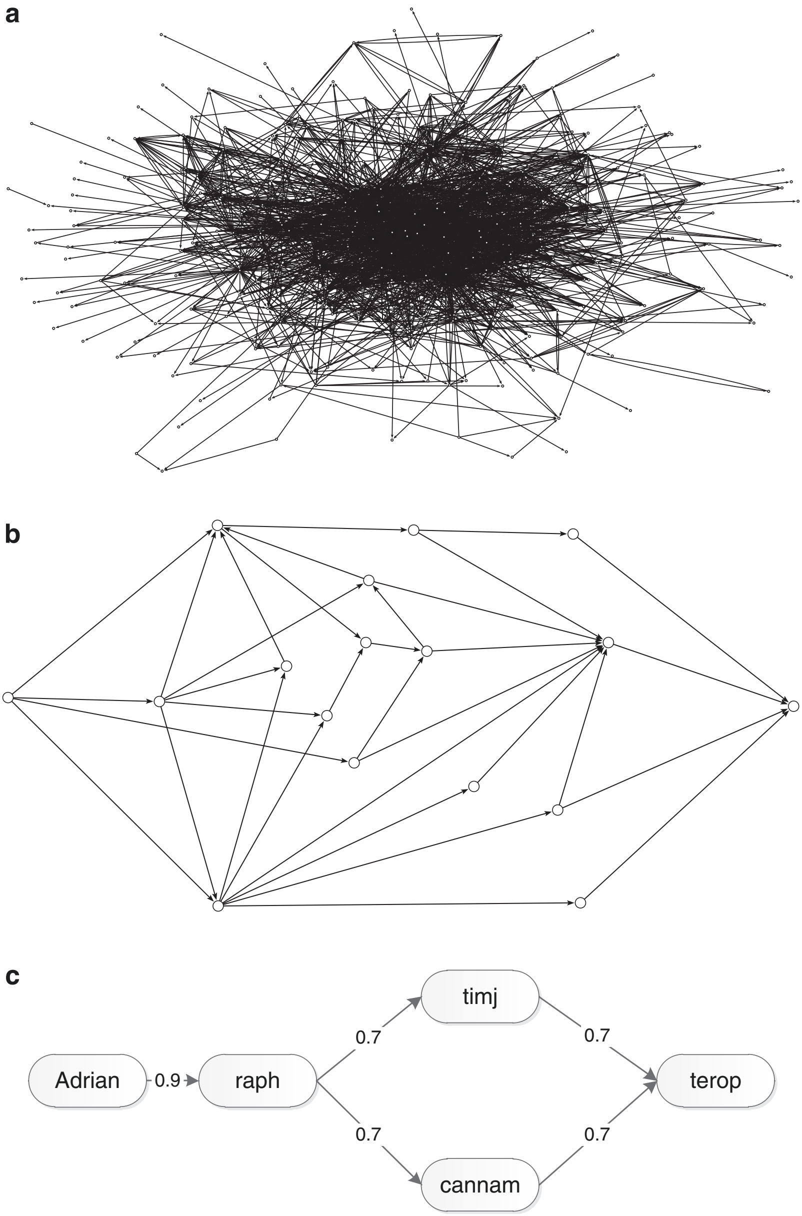


Fig. 9. The whole graph and the induced KS subgraph by the path selection stage are also plotted for comparison.

From the figures, we can see that the whole graph is hard for interpretation. As to the KS subgraph, although the number of edges has significantly decreased compared with the original whole graph, there are still some redundant edges which might diverge end users' attention. On the other hand, the EVO subgraph only presents the most important participants and their trust opinions, providing a much clearer explanation on how the trustworthiness score is inferred.

Future Directions

On one hand, much of the research in trust inference focuses on the inference accuracy, while inference efficiency is also important in



Subgraph Extraction for Trust Inference in Social Networks, Fig. 9 The interpretation example of the whole graph, KS-20, and EVO-10 on KS-20. (**a**) The original whole graph. (**b**) KS subgraph with $K = 20$. The paths are from "Adrian" (the leftmost node) to "terop" (the rightmost node). (**c**) EVO subgraph with $N = 10$ on KS-20. The component is from "Adrian" to "terop"

real-world trust inference applications, especially in those online applications. Future work should be able to find the best trade-offs between effectiveness and efficiency according to the specific applications.

On the other hand, we believe that usability is becoming a new requirement for trust inference. Users start to care about not only who they should trust but also why they should trust. It is also interesting to incorporate distrust in the subgraph extraction as users may also concern about why they should not trust someone.

Cross-References

▶ Computational Trust Models
▶ Trust in Social Networks

Acknowledgment This work is supported by the National 863 Program of China (No. 2012AA011205), and the National Natural Science Foundation of China (No. 91318301, 61021062, 61073030). The second author was partly sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053.

Appendix

To find K short paths from graph $G(V, E)$ in the path selection stage, many existing algorithms can be used. We consider two representative algorithms from the literature. Here, we present the detailed algorithm description for completeness.

The first algorithm is *Yen's k-shortest loopless paths* (*KS*) algorithm (Yen 1971), which is shown in Algorithm 3.

Algorithm 3 Detailed KS Algorithm

```
Input: Weighted directed graph $G(V, E)$, two nodes $s, t \in V$, and a parameter $K$ of path number
Output: Set $C$ with $K$ paths from $s$ to $t$
X ← shortest path from s to t
C ← shortest path from s to t
while $|C| < K$ and $X \neq \emptyset$ do
    P ← remove the shortest path in X
    d ← the deviation node of P
    for each node v between d (inclusive) and trustee t (exclusive) in P do
        pre ← subpath from trustor s to v in P
        post ← the deviated shortest path from v to t
        combine pre and post, and add it to X
    end for
    C ← C + the shortest path in X
end while
return C
```

In the algorithm, we use Dijkstra's algorithm for finding a shortest path. All the computed paths are loopless by temporarily removing visited nodes. The key idea of the KS algorithm is *deviation*. The *deviation node d* of path P is the node that makes P deviate from existing paths in the candidate set C. For each node v between d (inclusive) and trustee t (exclusive) in P, the *deviated shortest path* from node v to t is computed by temporarily removing the edge starting at v in P. The computed deviated shortest path *post* and the subpath *pre* (the path from s to v in P) are combined to form a possible path candidate. For the nodes before d, possible shortest paths are already computed and included in X. Based on deviation, KS finds the K-shortest paths from trustor s to trustee t one by one. Following Martins and Pascoal's implementation (Martins and Pascoal 2003), we compute the deviated shortest path from deviation node d to the trustee in a reverse order.

The other algorithm is the randomized algorithm *path sampling* (*PS*) (Hintsanen et al. 2010), which is proposed for the *most reliable subgraph problem* (Hintsanen and Toivonen 2008). While PS is proposed for undirected graphs, trust relationships in social networks should be directed as trust is asymmetric in nature (Golbeck and Hendler 2006). Therefore, we adapt PS (as shown in Algorithm 4) for a directed graph.

Algorithm 4 PS Algorithm

```
Input: Weighted directed graph $G(V, E)$. two nodes $s, t \in V$, and a parameter $K$ of path number
Output: Set $C$ with $K$ paths from $s$ to $t$
C ← shortest path from s to t
while $|C| < K$ do
    re-decide all the edges in E
    for each path P in C do
        if P is decided as true then
            F ← F + P
        end if
    end for
    while $F \neq \emptyset$ do
        re-decide the most overlapped edge in F as failed
        remove failed paths from F, if there are any
    end while
    P ← the shortest path among the non-failed edges from s to t
    if $\boldsymbol{P} \neq \emptyset$ then
        $\boldsymbol{C} \leftarrow \boldsymbol{C} + \boldsymbol{P}$
    end if
end while
return $\boldsymbol{C}$
```

PS considers the input graph as a Bernoulli random graph (Robins et al. 2007), and the algorithm is based on the *edge decision* of this random graph. An edge is randomly decided as true with probability $p(e)$, and a path is decided as true if all the edges on the path are decided as true. At the beginning of each iteration, all the edges of the graph are re-decided, and these *graph decisions* provide opportunities for distrust information to be contained. Like KS, PS first adds a shortest path into candidate set C. PS then tries to find a graph decision based on which none of the paths in C are true. To avoid the situation when this graph decision is hardly found, PS stores the true paths in C to a temporary set F and deliberately fails the most overlapping edges in F until none of the paths in F are true. Finally, based on the results of graph decision and edge failing, PS finds the shortest path P among the non-failed edges from trustor s to trustee t and adds it to C. The algorithm ends until K paths are found.

PS allows some distrust information to be incorporated into the extracted subgraph, which could in turn lower the P-error based on our experiments. However, the time complexity of PS is difficult to estimate, since the wall-clock time depends on the graph density. In addition, as shown in our experiments, the wall-clock time of PS is especially long when K becomes sufficiently large. We conjecture that PS can be used in dense graphs where numerous paths exist between node pairs.

References

Bäck T (1996) Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms. Oxford University Press, New York

Barbian G (2011) Assessing trust by disclosure in online social networks. In: Proceedings of the international conference on advances in social networks analysis and mining. In: ASONAM'11, Kaohsiung, pp 163–170

Buchegger S, Le Boudec JY (2004) A robust reputation system for mobile ad-hoc networks. Technical report, KTH Royal Institute of Technology, Theoretical Computer Science Group

Faloutsos C, McCurley KS, Tomkins A (2004) Fast discovery of connection subgraphs. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, ACM, KDD'04, Seattle, pp 118–127

Golbeck J, Hendler J (2006) Inferring binary trust relationships in web-based social networks. ACM Trans Internet Technol 6:497–529

Guha R, Kumar R, Raghavan P, Tomkins A (2004) Propagation of trust and distrust. In: Proceedings of the 13th international conference on world wide web, WWW'04, New York. ACM, pp 403–412

Hang CW, Wang Y, Singh MP (2009) Operators for propagating trust and their evaluation in social networks. In: Proceedings of the 8th international conference on autonomous agents and multiagent systems. In AAMAS'09, Budapest, vol 2. International Foundation for Autonomous Agents and Multiagent Systems, pp 1025–1032

Hershberger J, Maxel M, Suri S (2007) Finding the k shortest simple paths: a new algorithm and its implementation. ACM Trans Algorithms 3(4):45

Hintsanen P, Toivonen H (2008) Finding reliable subgraphs from large probabilistic graphs. Data Mini Knowl Disc 17(1):3–23

Hintsanen P, Toivonen H, Sevon P (2010) Fast discovery of reliable subnetworks. In: Proceedings of the international conference on advances in social networks analysis and mining, ASONAM'10, Odense, pp 104–111

Kamvar SD, Schlosser MT, Garcia-Molina H (2003) The eigentrust algorithm for reputation management in p2p networks. In: Proceedings of the 12th international conference on world wide web, WWW'03, Budapest. ACM, pp 640–651

Koren Y, North S, Volinsky C (2006) Measuring and extracting proximity in networks. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, KDD'06, Philadelphia. ACM, pp 245–255

Liu G, Wang Y, Orgun M (2010) Optimal social trust path selection in complex social networks. In: Proceedings of the twenty-fourth AAAI conference on artificial intelligence, AAAI'10, Atlanta, pp 1391–1398

Martins E, Pascoal M (2003) A new implementation of Yen's ranking loopless paths algorithm. 4OR Q J Oper Res 1(2):121–133

Massa P, Avesani P (2005) Controversial users demand local trust metrics: an experimental study on epinions.com community. In: Proceedings of the AAAI conference on artificial intelligence, AAAI'05, Pittsburgh, pp 121–126

Mui L, Mohtashemi M, Halberstadt A (2002) A computational model of trust and reputation. In:

Proceedings of the 35th annual Hawaii international conference on system sciences, HICSS'02, Big Island. IEEE, pp 2431–2439

Nordheimer K, Schulze T, Veit D (2010) Trustworthiness in networks: a simulation approach for approximating local trust and distrust values. In: Trust management IV, Morioka. IFIP advances in information and communication technology, vol 321. Springer, Boston, pp 157–171

Richardson M, Agrawal R, Domingos P (2003) Trust management for the semantic web. In: The semantic web, Sanibel Island. Lecture notes in computer science, vol 2870. Springer, Berlin/Heidelberg, pp 351–368

Robins G, Pattison P, Kalish Y, Lusher D (2007) An introduction to exponential random graph (p*) models for social networks. Soc Networks 29(2):173–191

Tong H, Faloutsos C, Koren Y (2007) Fast direction-aware proximity for graph mining. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining, KDD'07, San Jose. ACM, pp 747–756

Wang Y, Singh MP (2006) Trust representation and aggregation in a distributed agent system. In: Proceedings of the AAAI conference on artificial intelligence, AAAI'06, Boston, pp 1425–1430

Wang Y, Singh MP (2007) Formal trust model for multiagent systems. In: Proceedings of the 20th international joint conference on artificial intelligence, IJCAI'07, Hyderabad, pp 1551–1556

Wang G, Wu J (2011) Multi-dimensional evidence-based trust management with multi-trusted paths. Futur Gener Comput Syst 27(5):529–538

Xiong L, Liu L (2004) Peertrust: supporting reputation-based trust for peer-to-peer electronic communities. IEEE Trans Knowl Data Eng 16(7):843–857

Yao Y, Zhou J, Han L, Xu F, Lü J (2011) Comparing linkage graph and activity graph of online social networks. In: Social informatics, Singapore. Lecture notes in computer science, vol 6984. Springer, Berlin/Heidelberg, pp 84–97

Yen J (1971) Finding the k shortest loopless paths in a network. Manag Sci 17(11):712–716

Zhou R, Hwang K (2007) Powertrust: a robust and scalable reputation system for trusted peer-to-peer computing. IEEE Trans Parallel Distrib Syst 18(4):460–473

Ziegler C, Lausen G (2005) Propagation models for trust and distrust in social networks. Inf Syst Front 7 (4):337–358

Subgraph Identification

▶ Scaling Subgraph Matching Queries in Huge Networks

Subgraph Isomorphic Queries

▶ Scaling Subgraph Matching Queries in Huge Networks

Subgraph Matching

▶ Fraud Detection Using Social Network Analysis: A Case Study

Subgraph Mining

▶ Motif Analysis

Subgroup Discovery

▶ Community Detection and Analysis on Attributed Social Networks

Subjectivity Detection

▶ Sentiment Analysis, Basic Tasks of

Substitution Matrix

▶ Probability Matrices

Successive POI Recommendation

▶ Spatiotemporal Recommendation in Geo-Social Networks

Summary Statistics

▶ Univariate Descriptive Statistics

Supplier Networks

▶ Interorganizational Networks

Supply Chain Integration

▶ Supply Chain Networks

Supply Chain Management

▶ Supply Chain Networks

Supply Chain Networks

Dirk Pieter van Donk
Department Operations, Faculty of Economics
and Business, University of Groningen,
Groningen, The Netherlands

Synonyms

Buyer-supplier relationships; Supply chain integration; Supply chain management; Supply chain relations

Glossary

Supply chain network	A set of interconnected organizations that produce value
Supply chain integration	Cooperation between members of a supply chain network

Definition

A supply chain network can be defined as a set of interconnected organizations whose different processes and activities produce value (following Slack and Lewis 2011, p. 144), which is closely related to the definition of supply chain (management) by Christopher (2011) defined as the management of “a network of connected and interdependent organisations mutually and cooperatively working together to control, manage and improve the flow of goods and materials and information from suppliers to end users” (p. 19). In other words, it is the network of organizations that are involved, through upstream and downstream linkages, in the different processes and activities that produce value in the form of products and services in the hands of the ultimate consumer. Thus, for example, a shirt manufacturer is a part of a supply chain that extends upstream through the weavers of fabrics to the manufacturers of fibers and downstream through distributors and retailers to the final consumer.

Introduction

Over the last two decades, supply chain management has become one of the major fields of attention both in organizational practice and in academia. Consequently, managers have shifted from concentrating on internal flows and internal processes primarily to managing buyer-supplier relationships and even more managing relationships across the whole chain and across the network that supplies to and buys goods and services from an organization. Academics have followed this shift by extensively studying supply chains and networks and their management. According to Christopher (2011), some of the underlying reasons for the shift are that organizations compete globally, aim at being good at one specific task, and outsource all remaining activities. Additionally, consumers demand increasingly customized products at the same prices as normal products, while requirements and customer wishes change frequently. In order to be responsive and at the same time cost-effective, increased focus on the management of the supply chain or network is needed, often enabled by the use of novel ICT developments that link organizations to their suppliers and buyers. In this perspective, it is stated that "Competing is between supply chains instead of firms" (Christopher 2011). Below we will sketch what supply chain networks are and how to map and understand them. The concluding remarks will relate to some empirical findings and future research directions.

Key Points

As most companies do not produce their products and services alone, the relationships with suppliers and buyers, together forming the supply chain network, are pivotal. Consequently, making deliberate choices on with which of the members in the supply chain network to work intensively together is important. Empirical findings show that not only these choices influence the level of cooperation and supply chain integration, but also market and production characteristics might play a role in what is possible or required in terms of integration.

Historical Background

The importance of managing the entire chain from raw materials up to final customer has never been doubted. Still, in the past, parts of that overall chain have been relatively ignored. Driven by an increase in the number and variety of products as required by consumers and offered by manufacturers, paying closer attention to supply and distribution was forced in order to reduce costs levels, while more recently the attention for the overall supply chain network has been further increased driven by a variety of trends as increased outsourcing, global networks, better information systems, and the need to keep costs low.

Supply Chain Networks

Supply chain networks cover in principle all organizations that together produce services or products starting from basic raw materials until the final point of consumption. Such an end-to-end, integrated point of view is also reflected in statements like "from paddock to plate," "from mine to motorcar," or "from field to flower" that companies use to reflect their concern for the whole process. In order to better understand supply chain networks and to be able to map and manage them, Lambert and Cooper (2000) propose three key issues or questions to be answered by managers, e.g., to evaluate their current situation, while at the same time offering a framework of interest to researchers:

1. Who are the key supply chain members with whom to link processes?
2. What processes should be linked with each of these key supply chain members?
3. What level of integration and management should be applied for each process link? Below we will shortly address each of these.

Structure

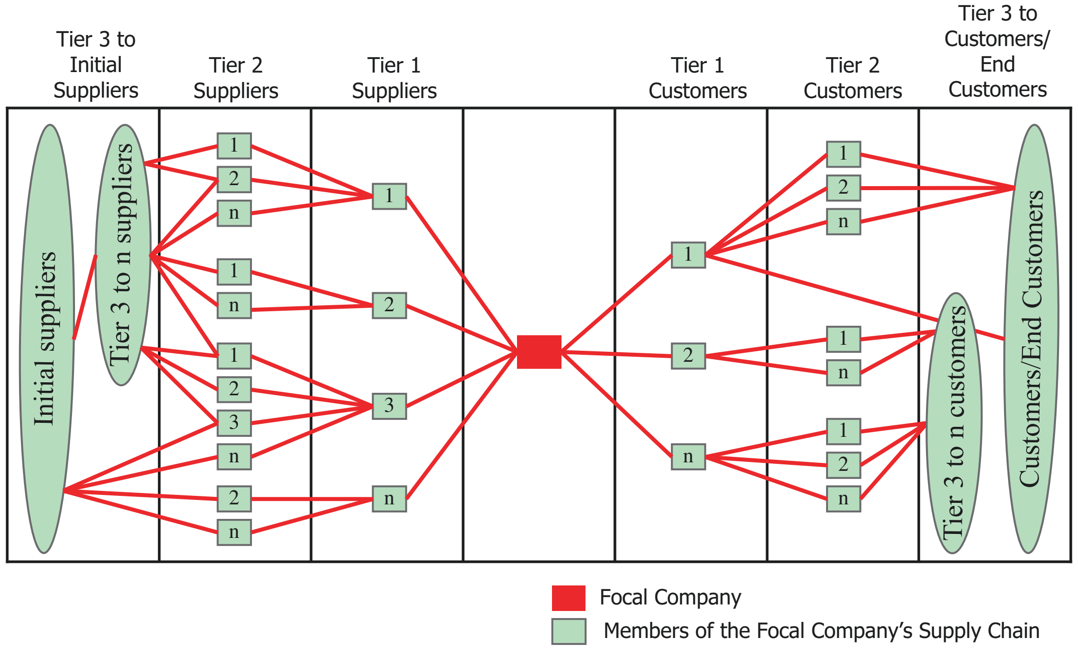
Companies are engaged in multiple networks such as the Chamber of Commerce and local industrial networks. From a supply chain network perspective, we are mainly interested in those companies that directly are involved in the value-adding activities for particular customers. Other organizations are addressed as supportive, such as banks and local authorities.

Networks can be distinguished along the number of partners in the chain (the number of tiers or horizontal structure) and the number of organizations (competitors, suppliers, or customers) at each level or tier in the network (vertical structure). Probably different from social networks, supply chain networks are always considered from a specific point of view, taking the perspective of one single company that is labeled as the focal company. Therefore, in drawing a network, there is always one central point, and it might be evident that the position of the focal firm in the network (being more located to the source of raw materials or more toward the final consumer), together with the vertical and horizontal position, is an important factor for the possibilities and also the wish to manage the – entire – network.

Examples and studies often focus on large and influential companies and their networks, e.g., Walmart, Ford, or Toyota, which are all powerful organizations that are able to direct and manage actively their networks. This might distort a general applicable approach. It might be evident that all organizations have their own network, albeit with rather different characteristics: vertical and horizontal structure and position in the network together indicate the complexity of the network. Figure 1 sketches a network structure with multiple suppliers/customers and also suppliers‘ suppliers and customers‘ customers, etc. These are, respectively, indicated as first tier, second tier, etc. In order to keep Fig. 1, relatively simple, competitors and relations of competitors with, e.g., first tier customers and/or suppliers, are left out. Networks can further become rather complex as suppliers might be both first and second supplier and even additionally have different roles in different supply networks as is, e.g., the case in the networks of Toyota and Nissan. Of course, both competitors and their relations with partners in the focal company’s network are in general relevant, dependent among upon others the size and power of the focal company.

Processes

The key process in any supply chain network is the provision of products or services to final consumers. As such the product flow is the point of departure of any network. The physical product flow involves processes such as transportation, warehousing, manufacturing, and distribution of finished goods. Mostly, depending on the nature of product and the network, each of these physical processes will be executed several times when products go from one stage to another stage in the network. In order to be able to execute those processes adequately, information processes have to be well organized as well, often enabled by ICT. Such processes relate to the physical flow directly such as ordering and purchasing processes, while others are more supportive and indirect such as customer service management and customer relationship management. The central tenet of supply chain management is that all such processes need to be well aligned or



Supply Chain Networks, Fig. 1 Supply chain network structure (Source: Lambert and Cooper 2000) integrated in order to be successful both within organizations (removing functional barriers) as well as along the whole network between organizations. While often formal alignment is stressed, based on formal ICT systems, there is evidence that personnel contacts between employees of different organizations in a network are important for proper functioning, as well (e.g., Ambrose et al. 2008).

Integration

A supply chain network might consist of numerous links and for each of the links different processes have to be dealt with. In order to be able to manage the network, three types of links can be distinguished: managed links, monitored links, and non-managed links. Managed links will most likely be the links with the main first tier suppliers and first tier customers but might also be the links with second tier suppliers or customers. For example, it is quite common that car manufacturers manage the relationship between a part manufacturer and a module manufacturer. Monitored links are links that are not actively managed, but some control is needed to be sure that such links are properly managed without directly interfering with the day-to-day management of the link. Finally, all other links are non-managed as they are of less interest for the focal company. Even in managed links, it seems likely that not all processes need to be firmly tuned, as will be shortly discussed below.

Reflection and Critical Issues

While theory explains and shows the benefits of supply chain management, there is little empirical evidence of management of whole supply chain networks. Part of that stems from the complexity explained above, and partly it might stem from the difference between the rhetoric and the practice of supply chain management. As Storey et al. (2006) explain, literature (as probably in more management areas) is not always clear in distinguishing between description and prescription. Part of the reality of supply chain management is that only a limited part of the chain is managed and mostly at the buyer-supplier relationship level and not along the whole chain or network. There is sufficient empirical evidence that shows the benefits of even such seemingly limited types of supply chain integration and management. Apart from the limited scope, there are serious barriers to supply chain integration such as misalignment of organizational and interorganizational performance measurement systems, along with misaligned – interorganizational – information systems and limited information transparency (e.g., Storey et al. 2006). In addition, different contexts (e.g., depending on product and/or market characteristics) might need different approaches, as Fisher (1997) argued. Also here, there is growing empirical evidence that shows the influence of such contextual factors (e.g., Van der Vaart and van Donk 2006; Giménez et al. 2012). Extending the benefits of buyer-supplier relationship management to the network level, while taking into account possible limitations and removing barriers, is one of the challenging issues in supply chain networks and their management. Social network theory is certainly one of the theoretical stances that can help to pave the way for further exploration and understanding of supply chain networks and their behavior.

Key Applications

Managing and governing the supply chain network is at the heart of every modern business as all companies have to work with the core members of their network in order to serve the final customers. Managing the network in the best way, given the changes in customer wishes and the trends that change the current business landscape, makes managing the supply chain network one of the core management areas.

Future Directions

There are several issues that recently have been mentioned as future directions for attention. The possibilities and options given by further developments in information and communication technology enable better cooperation, while of course the current rapid increase of e-commerce; online ordering also has changed the landscape of supply chain networks and supply chain management. Specifically, the final stage of distribution to the final customer has huge consequences. Another important development is the increased attention for sustainability and corporate social responsibility as an emerging field of attention in supply chain management. Specifically outsourcing to developing countries along with sourcing from such countries, together with the wish of customers to be sure that products have been produced not only safe but also in an environmental and social decent way, poses a burden on companies to control and safeguard members in the network. This is also an area of intense research on what practices and policies work and how. Another new direction is the application of insights form supply chain networks in production contexts to service supply chains both in the private and public domain.

Cross-References

▶ Business-to-Business Marketing
▶ Entrepreneurial Networks
▶ Innovator Networks
▶ Interorganizational Networks
▶ Location-Based Social Networks
▶ Network Models
▶ Relationships
▶ Visualization of Large Networks

References

Ambrose E, Marshall D, Fynes B, Lynch D (2008) Communication media selection in buyer-supplier relationships. Int J Oper Prod Manag 27(4):360–379

Christopher M (2011) Logistics and supply chain management, 4th edn. FT Press Pearson Education, Harlow

Fisher ML (1997) What is the right supply chain for your product? Harv Bus Rev 75(2):105–116

Giménez C, Van der Vaart T, van Donk DP (2012) Supply chain integration and performance: the moderating effect of supply complexity. Int J Oper Prod Manag 32(5):583–610

Lambert DM, Cooper MC (2000) Issues in supply chain management. Ind Mark Manag 29:65–83

Slack N, Lewis M (2011) Operations strategy, 3rd edn. Pearson Education, Harlow

Storey J, Emberson C, Godsell J, Harrison A (2006) Supply chain management: theory, practice and future challenges. Int J Oper Prod Manag 26(7):754–774

Van der Vaart T, van Donk DP (2006) Buyer-focused operations as a supply chain strategy: identifying the influence of business characteristics. Int J Oper Prod Manag 26(1):8–23

Recommended Reading

Handfield RB, Nichols EL (1999) Introduction to supply chain management. Prentice Hall, Upper Saddle River (basic text)

Hsuan J, Skjott-Larsen T, Kinra A, Kotzab H (2015) Managing the global supply chain, 4th edn. Copenhagen Business School Press, Copenhagen (advanced text)

Supply Chain Relations

▶ Supply Chain Networks

Surveillance

▶ Crime Prevention, Dataveillance, and the Regulation of Information Communication Technologies

Surveys

▶ Questionnaires for Measuring Social Network Contacts
▶ Sources of Network Data

Suspicious

▶ Social Phishing

Suspicious Behavior Detection

▶ Spam Detection on Social Networks

Symmetric and Skew-Symmetric Matrices

▶ Matrix Algebra, Basics of

Synthetic Data Sets

Sargur N. Srihari
Department of Computer Science and
Engineering, University at Buffalo, The State
University of New York, Buffalo, NY, USA

Synonyms

Anonymized Data; Augmented Data

Glossary

Machine Learning	Method to perform a task from examples
Probabilistic Graphical Model (PGM)	A graphical representation of joint probability distributions where nodes represent variables and edges represent influences
Regularization	Method used to generalize to previously unobserved evidence

Definition

Synthetic data, as used in computing, consists of symbols and values relevant to a given situation that are not directly observed but are artificially constructed.

Introduction

Synthetic data is created to design or improve performance of information processing systems. Principal uses of synthetic data are in designing machine learning systems to improve their performance and in the design of privacy-preserving algorithms that need to filter information to preserve confidentiality.

Key Points

The design of many information processing systems rely on representative samples of data. Such data may not be readily available due to the cost of acquiring the data or due to privacy concerns. This situation can be alleviated by constructing representative samples from available data sets. This can be done by applying appropriate transformations, e.g., injecting noise, that preserve semantic relationships, e.g., class labels.

Historical Background

Synthetic data has long been used to demonstrate design principles in pedagogy. In the early days of machine learning research in the 1950s statistically representative datasets were hard to obtain. Thus augmenting data sets by using appropriately transformed data was a work-around. Even today, state-of-the-art methods such as deep learning improve their performance when provided with synthetic data. In information systems that deal with human information, synthetic data is useful to anticipate unforeseen situations in which data may be compromised.

S

Synthetic Data Sets

A prerequisite for the construction of many algorithms with real-world applications is the availability of representative data sets. Artificial intelligence algorithms based on machine learning, e.g., to perform speech recognition or recognize objects in scenes, require data to determine necessary parameters. Supervised learning algorithms, such as neural networks, need examples of input and output, commonly referred to as *ground truth* to determine weights of connections. Even after models are constructed, sample data may be needed to perform inference with the models. *Probabilistic graphical models* (Koller and Friedman 2009; Srihari 2017), which are widely used for artificial intelligence applications, are notoriously intractable when it is necessary to respond to queries. The most popular approximate inference methods depend on generating samples (also referred to as *particles*) from the model.

Ground-truthed data sets that are large enough to make statistically significant conclusions are rarely available. This may be due to privacy concerns, such as with social network data. In such cases, automatic construction of synthetic data sets with characteristics similar to real-world data is necessitated. Many machine learning algorithms, including those based on *deep learning* improve their performance when synthetic data sets are used.

Data Set Augmentation for Machine Learning

The central problem of machine learning is to design algorithms that will perform well not just on training data but on new inputs as well. Regularization refers to any modification we make to a learning algorithm to reduce its generalization error but not its training error. The goal is to reduce the test error even at the expense of increasing training error.

Many forms of regularization available. Principal goals of regularization are to encode prior knowledge and to express preference for a simpler model. Some methods put extra constraints on the objective function that is optimized by the learning algorithm – they are equivalent to a soft constraint on parameter values, resulting in improved performance on a test set. Some strategies for regularization are: parameter norm penalties (L_2- and L_1- regularization), norm penalties as constrained optimization, data set augmentation, early stopping, sparse representations, bagging and other ensemble methods, dropout, and tangent methods. Data augmentation is one of the most effective methods of regularization, since a model generalizes better when it is trained on more data.

Data Sets to Overcome Privacy

Privacy issues are paramount when dealing with domains such as social networks. Examples of social network analytics are: attribute prediction given user social connections and phone usage log. It is of most interest to service providers and mobile equipment manufacturers to know particular user attributes and behavior patterns in order to provide better service and appropriate hardware technologies. This makes prediction of user attributes and behavior patterns one of the most actual research issues. Another is predicting future states of the social network. This would be useful for planning and marketing purposes.

The social network data set should suit most of the needs of the proposed analytic algorithms, e.g., it contains data for a large number of users for a long period of time (more than a year), has a unique combination of social interactions data and contextual attributes and basic user attributes (gender, age, marital status, etc.), and personal data in combination with location data and phone status information makes it possible to reveal novel patterns and relationships between users semantic location and its phone usage, etc.

Methods for Generating Synthetic Data

Data Augmentation

Data augmentation is easiest for classification, A classifier takes high-dimensional input $\mathbf{x}$ and summarizes it with a single category identity y. Since the main task of the classifier is to be invariant to a wide variety of transformations, we can generate new samples $(\mathbf{x}, y)$ just by transforming inputs. This approach is not easily generalized to other problems, e.g., for the density estimation problem it is not possible generate new data without solving density estimation.

Data set augmentation is very effective for the classification problem of object recognition. Images are high-dimensional and include a variety of variations, and may be easily simulated.

The transformations include translation, rotation, scaling, and compression. Translating the images a few pixels can greatly improve performance. This is observed even when a deep learning classifier has been designed to be invariant using convolution and pooling. Rotating and scaling are also effective.

A precaution in data augmentation for classification is that we should not apply transformations that would change the class. For instance, in the task of recognizing handwritten characters, e.g., *b* versus *d* and *6* versus *9*, horizontal flips and 180 degree rotations are not appropriate transformations. Some transformations are not easy to perform, e.g., out-of-plane rotation cannot be implemented as a simple geometric operation on pixels.

Injecting noise into the input of a neural network can be seen as data augmentation. Neural networks are not robust to noise. To improve robustness, train them with random noise applied to their inputs. Part of some unsupervised learning, such as denoising autoencoder. Noise can also be applied to hidden units. Dropout, a powerful regularization strategy, can be viewed as constructing new inputs by multiplying by noise.

Hand-designed data set augmentation can dramatically improve performance. When comparing algorithms *A* and *B*, the same data set augmentation technique should be used for both. If *A* performs poorly with no dataset augmentation and B performs well with synthetic transformations of the input, reason may be the data set rather than algorithm. Adding Gaussian noise is considered part of learning while cropping input images is not.

Generative Models and Sampling

Data can be synthesized from probabilistic generative models. Such models are typically in the form of directed or undirected graphs where nodes represent variables and edges represent influences. The model can be constructed from available data and any number of samples can be synthesized from the resulting model.

Synthesized data plays a role not only in constructing models but also in performing inference with models. A typical inference task is to determine the conditional probability of a variable when value assignments to evidence variables are given. A naive approach is to eliminate the irrelevant variables by summing over all possible values that they can take. But this is intractable since the number of terms increases exponentially with the number of variables.

In the *particle-based* approach to obtain an approximate answer, the data are generated appropriate to the query. Particle-based inference methods, also known as *Monte Carlo* methods, synthesize samples from a probability distribution. Sampling provides a flexible way to approximate many sums and integrals at reduced cost. Sometimes for speedup of a costly but tractable sum, e.g., subsample training cost with minibatches. In other cases, learning algorithms require us to approximate an intractable sum or integral, e.g., gradient of the log partition function of an undirected model.

If the model is in the form of a directed acyclic graph (also called a Bayesian network), samples can be obtained by proceeding through the variables in such a way that the parent nodes are sampled before the descendant nodes. This is known as ancestral or forward sampling. In the case where the query is conditional – when evidence variables are observed – it becomes necessary to modify forward sampling by using a method such as importance sampling.

When the model is an undirected graph, a *Markov chain Monte Carlo (MCMC)* method is used. In one such method known as *Gibbs sampling*, one of the variables is changed while holding the other variables constant. It is the conceptually simplest approach for drawing samples from an undirected graph. It works as follows. Suppose we have a graphical model over an *n*-dimensional vector of random variables $\mathbf{x}$. We iteratively visit each variable $\mathbf{x}_i$ and draw a sample conditioned on all the other variables, i.e., from $p(\mathbf{x}_i|\mathbf{x}_{-i})$. Due to the separation properties of the graphical model, we can equivalently condition on only the neighbors of $\mathbf{x}_i$.

Key Applications

Some uses of synthetic data are: (i) providing machine learning systems with more data than what is readily available so that they can improve their generalization performance, (ii) performing tractable inference from probabilistic models, and (iii) in designing filters for information systems so that confidentiality can be preserved.

Future Directions

An area of current research in deep learning is disentangling factors of variation in data (Goodfellow et al. 2016). Once such factors are discovered they can be used within generative models, which can then be sampled to synthesize realistic data sets.

Cross-References

- ▶ Gibbs Sampling
- ▶ Markov Chain Monte Carlo Model
- ▶ Probabilistic Graphical Models

References

Bishop C (2006) Pattern recognition and machine learning. Springer, New York, NY

Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge, MA

Koller D, Friedmana N (2009) Probabilistic graphical models: principles and techniques. MIT Press, Cambridge, MA

Srihari S (2017) Lecture slides and videos on machine learning and PGMs. http://www.cedar.buffalo.edu/~srihari/CSE674