Detecting Invisible Relevant Persons in a Homogeneous Social Network

Yoshiharu Maeno[1], Kiichi Ito[2], and Yukio Ohsawa[3]

[1] Graduate School of Systems Management, Tsukuba University,
3-29-11 Otsuka, Bunkyo-ku, Tokyo, 112-0012 Japan
maeno.yoshiharu@nifty.com
[2] Graduate School of Media and Governance, Keio University,
5322 Endo, Fujisawa-shi, Kanagawa 252-8520 Japan
[3] School of Engineering, the University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8563 Japan

Abstract. An algorithm to detect invisible relevant persons in a homogeneous social network is studied with computer simulation. The network is effective as a model for contemporary inter-working terrorists where large hub persons do not exist. Absense of large hub persons results in that the observed communication flow is also homogeneous. Clues regarding invisible relevant persons are hardly found in communication records. This task is, therefore, generally difficult. We demonstrate that our algorithm identifies the portion of the market baskets representing communication records, where invisible relevant persons are likely to be hidden, with good precision.

1 Problem

The activity of an organization is often under influence of invisible, but relevant persons. The influence is not directly observed. For example, a coordinator, who provides a number of activists with clever plans, communication means, attacking skill, money, etc., is hidden behind a terrorist network. The coordinator plays a role to synchronize the whole network toward a target. Understanding such invisible relevant persons from observed data provides important clues to invent hypothetical scenarios on opportunity and threat in business and social problems.

Network models such as a scale-free network (governed by a power law) [Barabási 1999] or a small-world (governed by an exponential law) [Watts 1998], have been studied as a generic abstraction of society, an economic system, and nature. It should, however, be noticed that individual systems have various concrete properties. (1) A model for contemporary inter-working terrorists, (2) a model for a self-organizing community, and (3) a model for a purposefully organized business team have quite different structures. Center-and-periphery structure, characterized by the existence of big hub persons, is a common property of the models (2) and (3). We call such a model an inhomogeneous social network. Invisible relevant persons are usually the hub persons. They can be detected by

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 74–81, 2007.

investigating communication from the hub persons toward peripheral persons. On the other hand, the model (1) is different in that it does not possess representative central persons. It is compared to networks of networks. We call this model a homogeneous social network.

In this paper, we study an algorithm to detect invisible relevant persons in a homogeneous social network. The algorithm aims at identifying precisely the portion of the observed communication, which includes traces and clues to the invisible relevant persons. We generate a test data set of market basket form representing communication among visible and invisible persons from a homogeneous social network with computer simulation. We demonstrate the precision characteristics of the algorithm with the test data set.

2 Social Network Model

We present a basic structure of a homogeneous social network, compared with inhomogeneous networks. Three models mentioned in 1 are described.

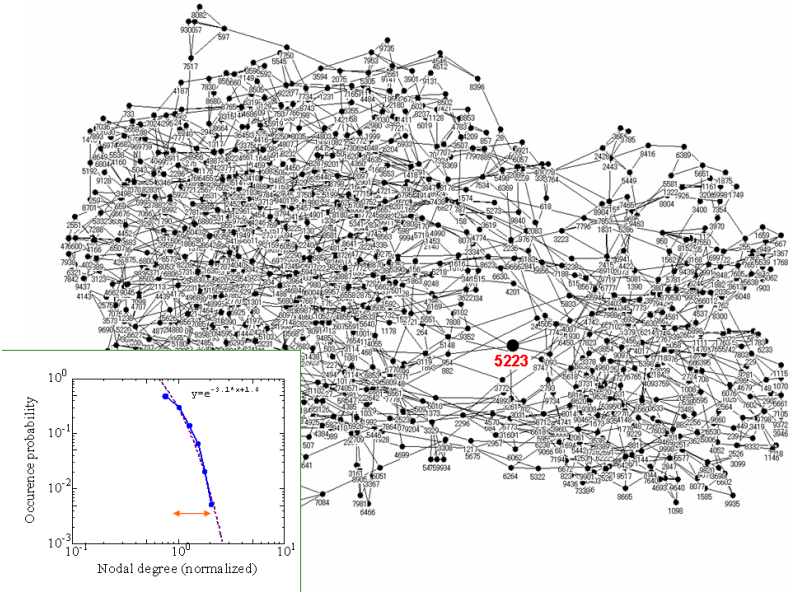

Fig. 1. Model for contemporary inter-working terrorists: a homogeneous network consisting of 995 nodes. The inset shows distribution of nodal degree. The persion ID=5223 is used in the evaluation.

A model for contemporary inter-working terrorists is illustrated in Fig.1. The network is derived from a number of empirical studies, simulation analysis, and jounalistic articles on terrorism [Popp 2006], [Singh 2004]. The network consists

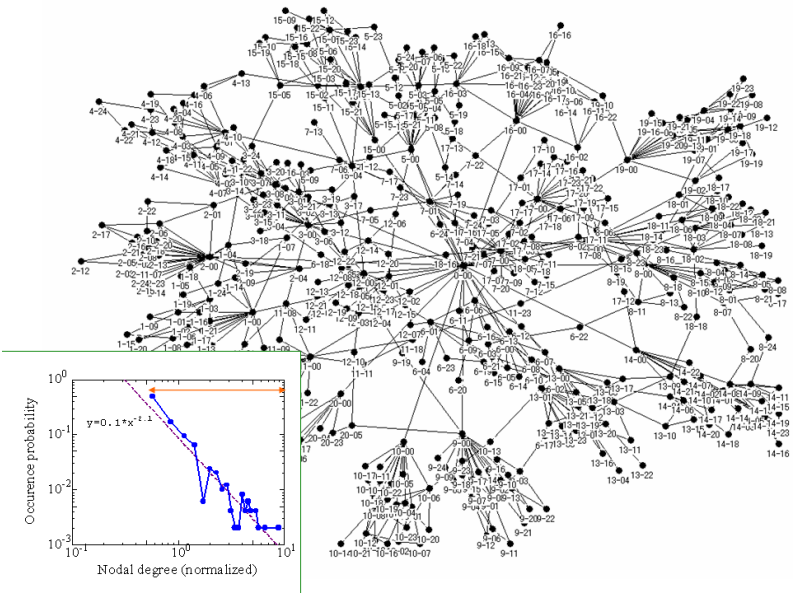

Fig. 2. Model for a self-organizing community: an inhomogeneous network consisting of 490 nodes. It is a scale-free network governed by a power low (Barabási-Albert model). The inset shows distribution of nodal degree.

of 995 nodes. As a whole, (1) the network seems to have two large groups: larger one on the left and smaller one on the right, (2) smaller groups seem to exit inside the two groups, (3) the boundary between the groups is not clear, and (4) the network does not possess big hub persons providing with a center facility among persons. The inset shows the occurrence distribution of nodal degree. The horizontal axis is normalized degree; degree devided by the average degree. It is governed by an exponential law; $y \propto e^{-3.1x}$. The degree ranges from 3 to 8. The average degree is 3.9. The deviation in the degree is small. It is the characteristics of the homogeneous network. Most persons are equivalent in that they have similar degree or centrality measure. Absense of large hub persons results in the fact that communication flow is also homogeneous. This model is used in the simulation study in 4.

A model for a self-organizing community is illustrated in Fig.2. It is a scale-free network derived from Barabási-Albert model [Barabási 1999]. The scale-free network is used to describe World Wide Web strucutre, scientist's collaboration network, etc.. The network consists of 490 nodes. The inset shows the occurrence distribution of nodal degree. It is governed by a power low; $y \propto x^{-2.1}$. The average degree is 3.6. The deviation in the degree is large. About 10 big hub persons are easily identified. The hub persons influence the way the network operates. Detecting invisible persons in such a network was studied in [Maeno 2006a].

A model for a purposefully organized business team is illustrated in Fig.3. The network is derived from empirical studies and analysis on communication via

email exchange within Enron [Keila 2006]. Enron was an energy company which ended in bankruptcy in 2001 because of the institutionalized accounting fraud. The network consists of 184 nodes. The inset shows the occurrence distribution of nodal degree. The average degree is 22.8. The deviation in the degree is large. The distribution is, however different from that for Fig.2. There are not small peripheral persons. Most persons contribute to functioning of the network. This may be a characteristic of the business team.

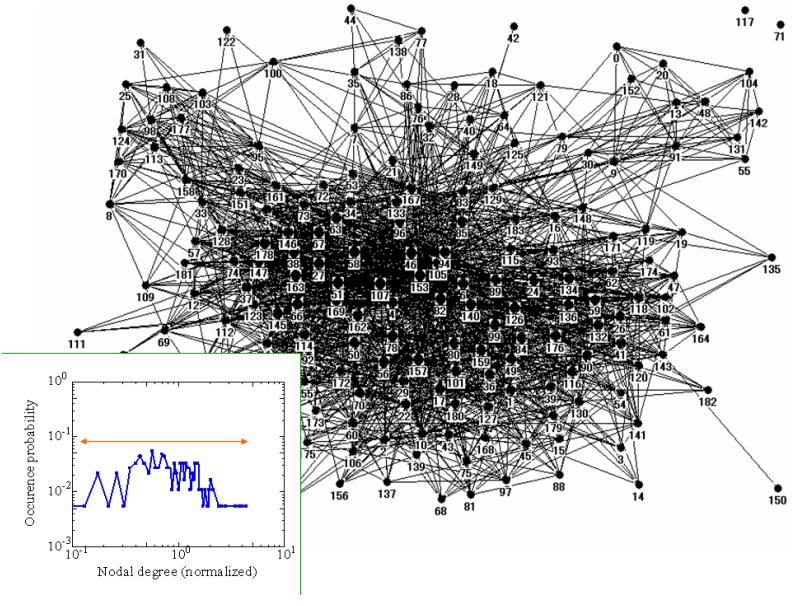

Fig. 3. Model for a purposefully organized business team: an inhomogeneous network consisting of 184 nodes. The inset shows distribution of nodal degree.

3 Algorithm

The algorithm to detect invisible persons in a homogeneous social network is presented. The algorthm is based on the crystallization algorithm [Ohsawa 2005] used in the human-interactive annealing process [Maeno 2007], [Maeno 2006b]. The input to the algorithm is observation data having market basket form; $b_i = \{e_j\}$. An individual event is denoted by e_j. A market basket is a set of events occuring simultaneously, spatially, or related strongly under a given subject. The output is ranking of b_i which indicates relative likeliness that the market basket provides clues to invisible events. The invisible event, which should have been in the market basket originally, is missing in the market baskets ranked highly.

The events are clustered into groups based on co-occurrence as a measure of similarity between events. Jaccard coefficient for the all pairs of events is

calculated from eq.(1). Jaccard coefficient is often used as a measure of co-occurrence in web mining and text analysis applications [Ohsawa 2006]. We can utilize various expertise on clustering. For example, k-medoids [Hastie 2001] or hierarchical clustering are simple, but efficient techniques. The k-medoids clustering is an EM algorithm similar to k-means algorithm for numerical data. A medoid event $e_{\mathrm{med}(j)}$ is an event locating most centrally within a cluster c_j. They are initially selected at random. Other $|E| - |C|$ events are classified into the clusters based on the Jaccard coefficint to the medoids. A new medoid is selected within the individual cluster so that the sum of Jaccard coefficints from events within the cluster to the modoid can be maximal. This is repeated until the medoid events converge. Advanced algorithms for unsupervised learning like self-oraganization mapping can also be employed.

$$J(i,j) \equiv \frac{\mathrm{Freq}(e_i \cap e_j)}{\mathrm{Freq}(e_i \cup e_j)}. \tag{1}$$

The ranking of b_i is evaluated as follows. A dummy event DE_i, representing invisible events, is inserted into b_i, that results in $b_i \rightarrow \{e_j\} \cup \mathrm{DE}_i$. The index i can be used to identify the market basket where the corresponding dummy event was inserted. The mixture of clusters resulting from the invisible events in the market basket is calculated with the Jaccard coefficient between the dummy event and the events in the clusters. It is evaluated according to eq.(2) using eq.(3). The market baskets which are ranked highly according to the largeness of eq.(2) are retrieved as candidate market baskets where invisible events should have been hidden.

$$I_{\mathrm{nu}}(i) \equiv \sum_{j=0}^{|c|-1} u(\max_{e_k \in c_j, b_i} J(\mathrm{DE}_i, e_k)). \tag{2}$$

$$u(x) = \begin{cases} 1 \text{ for } x > 0 \\ 0 \text{ otherwise} \end{cases}. \tag{3}$$

4 Evaluation

The objective of evaluation is to study how precise information regarding invisible persons the algorithm can retrieve from the test data with computer simulation. We use precision as a measure. In information retrieval, precision has been used as evaluation criteria. Precision is the fraction of relevant data among the all data returned by search.

4.1 Test Data

We describe the detail of the test data used for evaluation. It is generated from a homogeneous social network in Fig.3, as communication records among persons in the two steps below.

1. Collecting communication records into market basket: Market baskets representing communication among neighbor persons is generated. Persons within a specific distance (hop count) from a communication initiator are grouped into a market basket. This corresponds to a single conversation taking place regarding a specific subject the communication initiator concerns. An example market basket is $b_1 = \{954, 1930, 3261, 5093, 5223, 7743, 7808, ...\}$, representing communication initiated by the person ID=954.
2. Configuring invisible relevant persons: A latent structure regarding persons of interest is configured to the market basket by deleting the person from the data. Deletion made the structure invisible. As a result, the deleted persons and the links inter-connecting them become a latent structure hidden behind the market basket. The example market basket becomes $b_1 = \{954, 1930, 3261, 5093, 7743, 7808, ...\}$ if the person ID=5223 is an invisible person focused in evaluation. The resulting market baskets are like a bundle of email exchange records which lacks in oral communication from the invisible persons. The market baskets are the input to the algorithm. The algorithm attempts to identify b_1 as a candidate market basket where invisible persons should have been hidden.

For evaluation in 4.2, market baskets are configured from persons within five hops from individual persons in the first step 1. One hop is as long as one edge on the network shown in Fig.1. The number of persons within five hops is about 20% of the whole persons on the average. This is a relatively long distance communication. The latent structure of interest includes fifteen persons within two hops from the person ID=5223. These persons are remarkable in that they are equally close to every person in the network. They are not like a CEO governing a whole company in a hierarchical construct. We focus on them regarding them as a coordinating strategist group, who provides a number of activists inter-working in the terrorist network with clever plans, communication means, attacking skill, money, etc.. Persons in either the large cluster on the left or the small cluster on the right do not occupy such unbiased position. These persons are deleted from the market basket in the second step, so that they can be made invisible within the market basket data input to the algorithm.

4.2 Precision

Here, precision is evaluated by calculating the ratio of correct market baskets within the market baskets retrieved by the algorithm. The correct market baskets are those where persons had been deleted in the second step of 4.1. A single simulation condition is demonstrated in this paper. Systematic study on various conditions is planned.

Fig.4 shows the calculated precision. The horizontal axis is the number of retrieved basket data according to the ranking the algorithm outputs. The box in the figure lists 11 of highly ranked market baskets. The top 10 baskets retrieved as candidates are correct. Precision is good. This indicates the algorithm provides relevant suggestions regarding invisible persons. Clues to trace the invisible persons themselves will be found by monitoring the communication of the

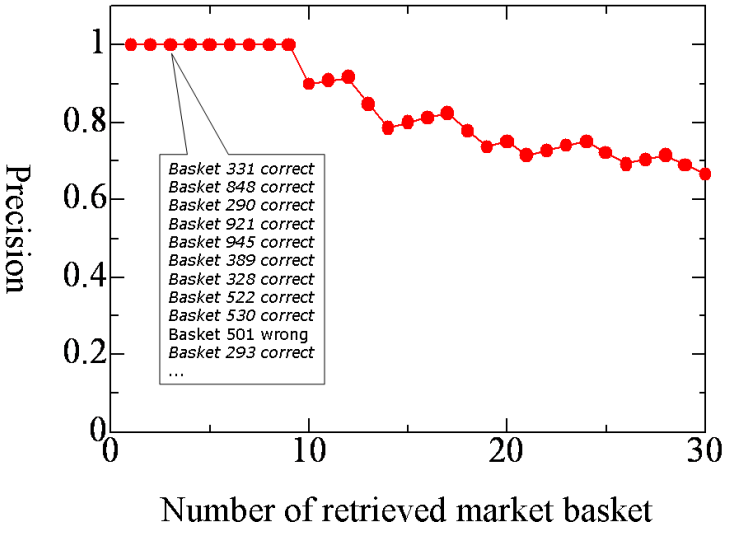

Fig. 4. Precision to identify the market baskets, where persons were made invisible, as a function of the number of retrieved market baskets

persons included in the market baskets. The algorithm is successful in a difficult task; detecting invisible persons in a homogeneous social network.

5 Discussion

An algorithm to detect invisible persons in a homogeneous social network was studied. The portion of the market baskets, where invisible persons are likely to exist, was identified with good precision. There are still remaining issues on the algorithm. We need to investigate two aspects more closely; (1) variability of social network structure and (2) communication pattern in terms of time sequence. Although we showed three models of social network, a number of alternative models exist for various domains of society, an economic system, and nature. Is a single algorithm applicable to such a wide range of models ? Communication sequence provides useful information to infer the influence flowing in a social network. Can the algorithm be modified with a time sequence analysis ? The topology of the network itself evolves as the communication pattern and consequently emerging inter-dependency between persons change. Understanding such dynamical nature is important to become aware of potential opportunity and threat arising from the latent structure hidden behind the surface behaviors.

We need to take a next step to understand the various impacts resulted from the social network, latent structures, and environment. For example, how do you avoid inconvenience caused by a specific action of a specific person in the social network influenced by an invisible relevant person's specific desicion making in a specific circumstance ? This is a hypothetical but concrete scenario invented to

obtain opportunity and to eliminate threat. For this purpose, we believe that it is essential to visualize various concepts and relationships in a map to recognize prior understanding and cognition to observation [Ohsawa 2006]. The human-interactive annealing process [Maeno 2007], [Maeno 2006b] is an effort toward such an objective.

References

[Barabási 1999] A. L. Barabási, R. Albert, and H. Jeong: Mean-field theory for scale-free random networks, Physica A **272**, 173-187 (1999).

[Hastie 2001] T. Hastie, R. Tibshirani, and J. Friedman: The elements of statistical learning: Data mining, inference, and prediction (Springer series in statistics). Springer-Verlag (2001).

[Keila 2006] P. S. Keila, and D. B. Skillicorn: Structure in the Enron email dataset, J. Computational & Mathematical Organization Theory **11**, 183-199 (2006).

[Maeno 2007] Y. Maeno, and Y. Ohsawa: Human-computer interactive annealing for discovering invisible dark events, to appear, IEEE Trans. Industrial Electronics (2007).

[Maeno 2006a] Y. Maeno, and Y. Ohsawa: Stable deterministic crystallization for discovering hidden hubs, Proc. IEEE Int'l. Conf. Systems, Man & Cybernetics, Taipei (2006).

[Maeno 2006b] Y. Maeno, K. Ito, K. Horie, and Y. Ohsawa: Human-interactive annealing for turning threat to opportunity in technology development, Proc. IEEE Int'l. Conf. Data Mining, Workshops, Hong Kong, 714-717 (2006).

[Ohsawa 2006] Y. Ohsawa eds.: Chance discovery in real world decision making. Springer-Verlag (2006).

[Ohsawa 2005] Y. Ohsawa: Data crystallization: chance discovery extended for dealing with unobservable events, New Mathematics and Natural Computation **1**, 373-392 (2005).

[Popp 2006] R. L. Popp, and J. Yen: Emergent information technologies and enabling policies for counter-terrorism, IEEE Press (2006).

[Singh 2004] S. Singh, J. Allanach, T. Haiying, K. Pattipati, and P. Willett: Stochastic modeling of a terrorist event via the ASAM system, Proc. IEEE Int'l. Conf. Systems, Man & Cybernetics, Hague, 6/5673-6/5678 (2004).

[Watts 1998] D. J. Watts, and S. H. Strogatz: Collective dynamics of small-world networks, Nature **398**, 440-442 (1998).