# Structure Pattern Analysis and Cascade Prediction in Social Networks

Bolei Zhang, Zhuzhong Qian(✉), and Sanglu Lu

State Key Laboratory for Novel Software Technology,
Nanjing University, Nanjing, China
zhangbolei@dislab.nju.edu.cn, {qzz,sanglu}@nju.edu.cn

**Abstract.** As information spreads across social links, it may reach different people and become cascades in social networks. However, the elusive micro-foundations of social behaviors and the complex underlying social networks make it very difficult to model and predict the information diffusion process precisely. From a different perspective, we can often observe the interplay between information diffusion and the cascade structures. On one hand, information driven by different mechanics may evolve into diverse structures; On the other hand, different cascade structures will reach different groups people and thus affect the diffusion process.

In this paper, we explore the relationships between information diffusion and the cascade structures in social networks. By embedding the cascades in a lower dimensional space and employing spectral clustering algorithm, we find that the cascades generally evolve into five typical structure patterns with distinguishable characteristics. In addition, these patterns can be identified by observing the initial footprints of the cascades. Based on this observation, we propose to predict cascade growth with the structure patterns. The experiment results show that the accuracy of predicting both the structure and virality of cascades can be improved significantly.

## 1 Introduction

How does information spread in social networks? When users observe information from their neighbors in the social network, they may make decisions and share the information to their friends. Starting from some root node, the information could then spread out and become cascades with tree structures. This phenomenon of information cascades has been observed ubiquitously in social networks across various domains. With the proliferation and emergence of online social networks, understanding and predicting how will the cascades evolve have attracted enormous attentions in opinion monitoring, advertising prediction and rumor control.

In a micro scope, we can model each user's behavior to predict the process of information diffusion [2,11,13]. The diffusion can often be described as a stochastic process where the nodes spread the information according to some


P. Frasconi et al. (Eds.): ECML PKDD 2016, Part I, LNAI 9851, pp. 524–539, 2016.
DOI: 10.1007/978-3-319-46128-1_33

predefined probability. According to the social sciences, the information diffusion probability is mainly dependent on two factors: homophily and social influence [3]. Homophily is the phenomenon that people tend to build social relationship and spread information with users in similar interests or background [18], while social influence occurs when users emotions or opinions are influenced by others. Despite extensive researches in studying the process of information diffusion, the complex micro-foundations of social diffusion and unpredictability of user decisions make it very difficult to extract the diffusion models precisely.

Instead, we can often observe the interplay between information diffusion and the cascade structures. On one hand, different mechanics of information diffusion can cause different cascade structures [6,8]. For example, influence-driven cascades usually evolve as rapid, complex structures, whereas homophily-driven cascades may become simple and star-like structures [3]. On the other hand, according to the widely used information diffusion models [13], different structures of the cascades may reach different people or communities, and thus affect the spread of information. Such phenomenon motivates us to study the relationships between information diffusion and cascade structures, despite that the intricate and diverse structures of the information cascades are often very difficult to analyze [15,20].

In this paper, we dive into the structures of information cascade in social networks and explore the relationships with information diffusion empirically. We propose that the structure patterns can be predictive of the cascade growth. In our first experiment, by dimension reduction and defining similarity measure between the cascades, we find that the information cascades in social networks generally evolve into five typical structure patterns with distinguishable statistics. In addition, these patterns can be detected from the early footprints of the cascades. Based on this observation, we predict the growth of the cascades by incorporating the structure patterns. The results show that the accuracy of predicting both the structure and virality of the cascades can be improved significantly when considering the structure patterns.

**Contributions.** The main contributions of this paper are:

- We propose a novel method for embedding the cascades in a lower dimensional space by incorporating social influence and homophily at the same time.
- By dimension reduction and spectral clustering, we find that the information cascades generally evolve into five typical structure patterns with distinguishable characteristics.
- We propose that the structure patterns can be predictive of the growth of information cascades. The experiment results show that the accuracy of predicting the growth of the cascades can be improved significantly by using the structure patterns as new features.

**Organization.** The rest of this paper is organized as follows: In Sect. 2, we review the research works related to this paper. Then we will introduce some preliminaries including the data set, the theories of information diffusion and cascade structures in Sect. 3. Section 4 formally presents the method for finding

the structure patterns of the cascades. In Sect. 5, we present the experiments of predicting the cascade growth. Finally, the paper is discussed and concluded in Sect. 6.

## 2 Related Works

There are three threads of researches related to our work: the mechanics of information diffusion, the prediction of cascade virality and analysis of cascade structures. We now introduce each of them respectively.

**Modeling information diffusion.** Modeling information diffusion is a central problem in studying social networks. Earlier research works proposed that information spreads like epidemics. The epidemic model generally assumes homogeneous networks [2,4] where the network is full clique and each person has the same probability for spreading the information. Typical diffusion models include the SIS (Susceptible, Infected), SIR (Susceptible, Infected, Recovered) etc. In [17], the authors proposed to fit the temporal curves of the cascade spikes using SI-like model. Compared to the biological viruses, extensive researches have been proposed to model the rumors, ideas, memes using social influence models, such as the independent cascade model [13], threshold models [11,23] and coverage models [21] etc. Following works have studied how to select the most influential nodes to maximize the spread of information diffusion [7,14]. Despite the algorithmic progress on selecting the most influential nodes, how to infer the diffusion models accurately remains a challenge.

**Cascade prediction.** From a macro scope, we can omit the diffusion process and predict the statistics of a cascade from its early footprints directly. Previous works usually considered the task as a regression problem [5,22] or a binary classification problem [12,24]. The growth of cascades may originate from multiple factors. In [20,22], the authors proposed to dive into the content of the information diffusion and analyze the spread of the information. The temporal features are also often used to predict the evolution of cascades [17,25]. The temporal dynamics of online usually falls into six different patterns [25]. In [17], the authors fit the model with one unified model. Recently, the structures of the social network are also taken into account to predict the evolution of cascades [1,6,19,24]. Generally, cascades that spread across multiple communities are considered to be more viral than those trapped in a single community. In comparison, we try to analyze the structure pattern of the cascade to observe how it evolves from time to time.

**Cascade structures.** When information starts to spread in the social network, it usually generates a tree structure. The properties of the structures have been studied over years. In [15], the authors find that the cascades in email network are usually very narrow and continually reaching people several hundred levels away. There are also implications that cascades with different topics spread in different structures [20]. For example, the political topics are usually more persistent than the conversational idioms. To quantitatively differ the structures of the cascades,

the authors in [9] proposed to use wiener index to characterize the structure of information cascades. The wiener index is the average distance between each pair of nodes in the cascade, which is often used to describe the complexity of the structure of a graph. Due to the intricate and diverse structures of social cascades, it is necessary to fully understand the cascade structures and explore the relationships with information diffusion.

## 3 Data Set, Information Diffusion, and Structure Patterns

In this section, we first present the collected data set for analysis. Then we will briefly introduce the mechanics of information diffusion in social networks. In the last part, we show how the information could shape the structure of the cascades.

### 3.1 Data Collection

We choose Weibo (http://weibo.com) as the basis social network platform for analysis. Weibo is a Twitter-like micro-blog platform in China. The network can be modeled as a directed graph $G = (V, E)$ where the nodes set $V$ represents users and the edges $E$ represent following/follower relationships between users. For a user $i$, we denote the followers of $i$ as $N(i)$. In Weibo, each user can post a message with at most 140 (Chinese) characters publicly. Once the message is posted, the neighbor users may observe the occurrence and share the message.

In Weibo, we can trace the information diffusion path by analyzing the spreading contents. For a message with content "A: xxx//@B: yyy//@C: zzz", it means that, user B first shares the origin message with content "yyy" from user C; then, user A shares B's message with content "xxx". A diffusion path as $C \rightarrow B \rightarrow A$ can be constructed from the above message. To get the full trace of each cascade, we start from the root message and get all the shared messages. Each cascade can then be modeled as a tree structure with the origin message from the root node. Figure 1 presents an example of the tree structure of a cascade.

In total, we crawled 16, 439, 997 messages from Weibo during April 1st 2015 to April 30th 2015, and extracted 33, 214 cascades with at least 10 shares. The social network has 6,738, 199 users and 11,271,789 following relationships. By adopting text clipping on the text and training a multinomial Naive Bayes classifier with a labeled data set, the messages are classified into 8 topics, including: Economy, Education, Technology, Culture, Sports, Health, Politics and Travel.

### 3.2 The Mechanics of Information Diffusion

Information diffusion is ubiquitous in online social networks. There are complex factors that drive the diffusion of information between people. Among them,

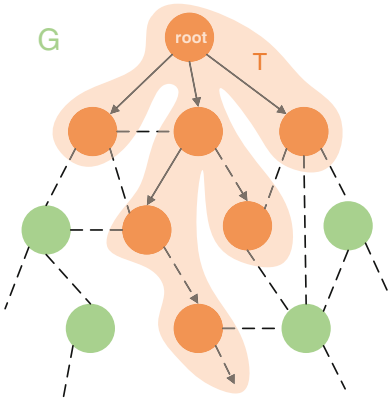



**Fig. 1.** Cascade example. A cascade example in the social network $G$, where the lines are social links between users. The cascade starts from the root node to spread. The cascade include the nodes with orange color, denoted as $T$. The green nodes are those in the social network but not engaged in the cascade. The information spreads only on the solid line. (Color figure online)

homophily and social influence are usually considered to be the most important ones [3]. *Homophily* refers to the tendency of individuals to associate with similar peers, such as age, gender, religion etc. In addition, researches show that homophily also contributes a lot to the information diffusion [26], since the behaviors of similar users are often correlated. Accordingly, information driven by homophily is more likely to spread to the close neighbors. In comparison, *Social influence* occurs when one's opinions are affected by others. When the information is driven by the social influence, it is more likely to spread across communities and in long distances. Thus, the information diffusion is more unpredictable in this case.

Due to complex micro-foundations of user decisions, it is often difficult to distinguish the two factors from information diffusion processes. Moreover, in most cases, both the factors may play a role in driving the spread of information. From a different perspective, we can often observe the interplay between information diffusion and the cascade structures. Thus, if we could identify the different structure patterns of information cascade, it may help us understand the mechanics of information diffusion and accordingly predict the cascade growth.

### 3.3 Cascade Structures

We propose that the cascade structures can often reflect the mechanics of the information diffusion. For instance, the deep and complex structure of a cascade tree means that the information is spreading in different communities and long distances, which may be the result of the social influence. Such observations motivate us to explore the structure patterns of information cascades.

In Fig. 2 we show two illustrative examples of the evolution of cascade structures in social networks. The topic of the first message belongs to Culture, which is possibly driven by homophily. As shown in the figure, it has a star-like structure, since the information mainly spreads to users not too far away. While in the second example, the message of the information is about Technology, which is more likely to be driven by social influence. There may be professional discussion and complex diffusion in this topic. But the information may not be interested to a wide range of the near neighbors, since the social links are probably built on homophily.

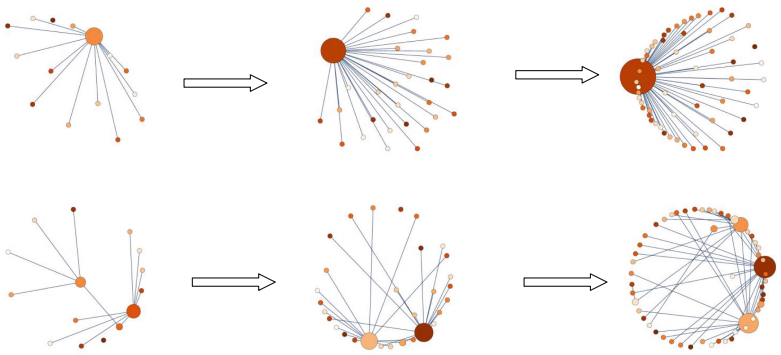


**Fig. 2.** Evolution of the cascades. Two illustrative examples of the cascade structures. The size of the nodes reflects the degree of the nodes in the tree.

## 4 Finding the Structural Patterns

In this section, we formally present the method for finding the structural patterns. We first embed the cascades into a lower dimensional space, and then apply spectral clustering to find structure patterns. Finally we will show the properties of the patterns.

### 4.1 Cascade Embedding

Due to the intricate structures of the information cascades, we first try to embed the cascades in a lower dimensional space to find the structure patterns. To preserve the structure characteristics, the following prerequisites should be considered: 1. The embedded cascades can distinguish cascades with different structures, such as star-like trees and the trees with deep complex structure; 2. It can reflect the virality of the cascades; 3. In the embedded cascade, users with close proximity are less important since information is more likely to spread between these users as a result of homophily.

Considering the above aspects, we propose to extract a centrality value for each of the nodes in the cascades. The basic idea is that the centrality increases with the influence of a user, but is offset by the proximity to the root of the cascade. First, we assume that the information diffusion on each edge depends on the difference of the posting time: $\Delta_{ij} = |t_i - t_j|$, which has been extensively studied in papers such as [10,16]. The weight of influence strength from $i$ to $j$ can be formulated as:

$$w_{ij} = \alpha_{ij} e^{-\alpha_{ij}\Delta_{ij}} = \alpha_{ij} e^{-\alpha_{ij}(t_j - t_i)}$$

Thus, an edge with larger weight indicates the information is more "viral" on the edge. The influence of the user can then be measured as the sum of the influence to all users, i.e. $\sum_{j \in N(i)} w_{ij}$. We also consider the effect of homophily: the users with close proximity to the root is less central in the cascade since the homophily may play a more important role. Thus, we propose an amplified factor of $d_i + \gamma$ parameterized by $\gamma$, where $d_i$ is the distance between user $i$ and the root user. The centrality of a user $i$ can be formally represented as:

$$w_i = (d_i + \gamma) \sum_{j \in N(i)} w_{ij}$$

In practice, we empirically choose uniform value for $\alpha_{ij}$ as 1.0 and $\gamma$ as 5.0. Finally, we extract the skeleton of a cascade by sorting the nodes according to their centralities in decreasing order as a vector $\mathbf{w}$.

### 4.2 Spectral Clustering

After embedding the cascade in a vector space, the distance between two embedded trees $\mathbf{w}^i$ and $\mathbf{w}^j$ can be computed with the Euclidean distance, which is defined as:

$$d(\mathbf{w}^{(i)}, \mathbf{w}^{(j)}) = \sum_t ||\mathbf{w}_t^{(i)} - \mathbf{w}_t^{(j)}||$$

where $||\cdot||$ is the $l_2$ norm. The Euclidean distance is then converted to similarity as:

$$S_{ij} = e^{-\frac{d(\mathbf{w}^{(i)}, \mathbf{w}^{(j)})}{\sigma}}$$

where $\sigma$ is the standard variance of the derived distance matrix. Given the similarity measure, we use spectral clustering to find the structure patterns of the cascades, so that trees with similar structures are clustered into the same group. The spectral clustering techniques make use of the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction. The first step of the spectral clustering algorithm is to get the graph Laplacian matrix $L$.

$$L = D - S$$

where D is the diagonal matrix $D_{ii} = \sum_j S_{ij}$. Then, we get the first $k$ eigenvectors of $L$, denote as $u_1, u_2, ..., u_k$. Let $y_i$ be a row vector of the matrix $U = [u_1, u_2, ..., u_k]$, i.e., $y_i = [u_{1i}, u_{2i}, ..., u_{ki}]$. The results can be derived by applying k-means clustering on $\mathbf{y}$ to get the cluster results.

### 4.3 Structure Patterns of Cascades

In this experiment, we will apply spectral clustering algorithm on the diffusion trees to find the structure patterns. Specifically, we explore the relationship between the early structure patterns and the eventual growth of the cascades. The clustering algorithm is employed on the first $k = 30$ nodes of each cascade. The value of $k$ is chosen empirically and the reason will be explained in Sect. 5.

To apply the clustering algorithm, the first step is to choose the number of clusters. Here we use the average silhouette score to evaluate the performance of the results of clustering algorithm. The higher silhouette score indicates better cluster results. We find that the scores are almost the same ranging from 0.1 to 0.2 with cluster number from 3 to 8. We choose 5 as the number of the clusters when the silhouette score is 0.14. Despite that the silhouette score is higher with fewer clusters, it reveals less information about the structures.

After clustering the cascades, we observe the statistics of the cascades in each cluster in Table 1. We show the average size, average depth, largest degree and average wiener index of the cascades in each cluster. Generally, higher wiener index indicates more complex structures. According to the statistics, even though the clustering algorithm is employed on the initial structures (first 30 nodes) of the cascades, the eventual statistics in the clusters are quite distinguishable from each other. For example, The cascades in C2 have almost the same size (243.14) as the cascades in C3 (292.34). However, the cascades depth (4.07) is much higher than that in C3 (2.62). In C4, the cascades have more nodes (308.30) than both C2 and C3, but the average depth (3.56) is between them. In C1 and C5, the cascades have significantly more nodes (1265.57 and 2448.77) than other clusters. And the depth in C1 (4.70) is almost the same as that in C5 (4.93). The cascades in C1 and C5 usually have a dominated node since the largest degree is very close to the size of the cascades. Moreover, the cascades in C2 have more complex structures, since the wiener index (2.28) is higher than other cascades.

**Table 1.** Table of statistics.

| | C1 | C2 | C3 | C4 | C5 |
|---|---|---|---|---|---|
| Cascade number | 4537 | 12482 | 9409 | 3762 | 3024 |
| Average size | 1265.57 | 243.14 | 292.34 | 308.30 | 2448.77 |
| Average depth | 4.70 | 4.07 | 2.62 | 3.56 | 4.93 |
| Largest degree | 999.25 | 128.71 | 56.69 | 218.27 | 2054.28 |
| Average wiener | 2.14 | 2.28 | 2.01 | 2.09 | 2.13 |

In addition to the statistics, we also plot the cumulative distribution of the cascades size and depth in each cluster in Fig. 3. It can be observed that cascades in the same clusters tend to have the similar statistics, since the cumulative distribution curves increase steeply around the average size or depth. This also strongly implies that the early structure of a cascade may be predictive of the cascade growth.

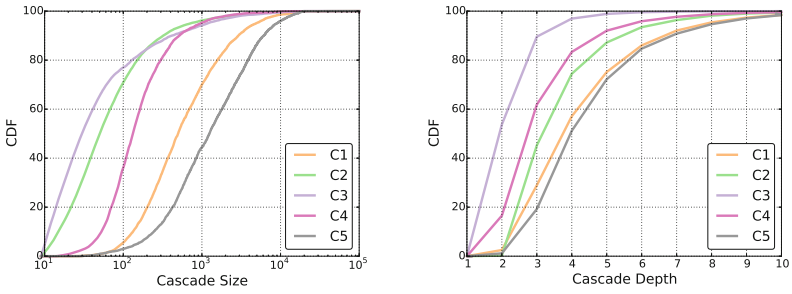



**Fig. 3.** CDF of cascade depth and size. In the first figure, the x-axis is the log of the cascade size. The cascades in different clusters are distinguished in different colors. (Color figure online)

Now we plot the representatives of from C1 to C5 to observe their structures in Fig. 4. The representatives are selected as the cascade which has the closest distance to all other cascades in the same cluster. As observed from the figure, the structures of cascades from different clusters vary significantly from each other. The representatives of C3 and C4 have fewer nodes and star-like structure; and the representatives of C1 and C5 have more nodes and complex structures. In particular, in C2, there are relevant as many nodes as that in C3 and C4, while the structure of C2 seems to be much more complex.

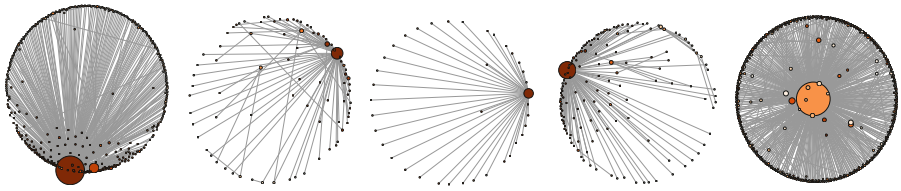


**Fig. 4.** Representative cascade in each cluster.

Finally, we explore the topic distributions of different clusters of the cascades. Figure 5 shows the proportion of the cascades of each cluster in the 8 topics. As

observed, the cascades with topics such as Culture, Health, Travel are more likely to have wide and simple structures in C3, C4 and C5. These topics may be related to user interests or public opinions that may be driven by homophily. Meanwhile, the cascades with topics such as Economics, Education, Politics, Sports, Technology are more likely to grow as complex structures in C1 and C2. The topics are more likely to be controversial or professional that may cause social influence between users. This result is consistent with the "persistence" of topics as introduced in [20].

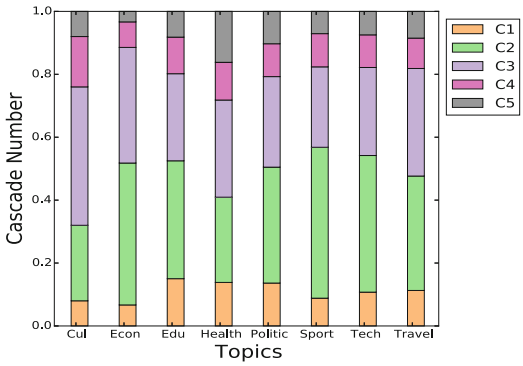



**Fig. 5.** The proportion of cascades in each topics.

## 5 Predicting Cascade Growth

By spectral clustering, we identify five patterns of cascades with distinguishable characteristics. The observation implies that the early structures can be predictive of the future growth of the cascades. In this section, we use the patterns to predict the structure and virality of the cascades growth respectively.

### 5.1 Experiment Setup

In predicting the cascade growth, our general method is to use the machine learning techniques to predict the labels of the target data with a set of features. Here, we use the method of *logistic regression* algorithm. Other classification methods such as random forest, SVM etc. were also tried. The results show that the performances of different classification methods do not vary a lot from each other. We use 5-fold cross validation for training and testing.

**Features.** We select a set of features that might be correlated with the growth of the cascade. The features are extracted from the first $k$ nodes of the cascades on 5 dimensions: content, temporal, structural, root and structure pattern features. In total, we extract 50 features for each of the cascade.

- Content Features. The content features include some origin content related statistics: whether the content has a *link*, a *hashtag* or a *mention*; and whether the content belongs to one of the 8 topics.
- Temporal Features. In the temporal features, we first use the *average time* between the first and the last $k/2$ shares; and the temporal features also include the time elapsed between the original and the first 10 shares.
- Structural Features. In the structural features, we have the total number of *friends* of the root node in the first $k$ shares, the total number of *uninfected friends* of the first $k$ sharers, the *average depth* of the first $k$ users, and the out-degree of the first 10 shares.
- Root Features. The root features include the *number of followers* of the root user, the *gender*, the *verification status* and the *number of messages* that has been posted by the root user.
- Structure Pattern Features. Finally, we employ the cascade embedding on each of the cascade, and extract the first 10 nodes with highest centralities.

**Comparison Methods.** Denote our method as SP-based (Structure Pattern based). We compare our method with the following algorithms:

- SP-blind: In SP-blind method, we exclude the structure pattern features and employ the logistic regression method for prediction.
- PCA-based: Instead of using the structure pattern features, in PCA method, we use PCA (principle component analysis) for dimension reduction of the cascade matrix $T$ and get the eigenvector of the covariance matrix with largest eigenvalue as features.
- Wiener-based: Similar to PCA-based method, we compute the wiener index of the first $k$ nodes as a feature for prediction.
- Random: We randomly guess the result to be positive or negative.

### 5.2 Predicting Cascade Structure

We begin by predicting the structures of the cascades. The intuition is that: a cascade that is initially wide is more likely to evolve as star-like structure, while a cascade with complex structure initially would also grow to be complex in the future.

First, we observe the average Pearson correlation between the node centralities and the wiener index of the eventual structure of the cascades. The higher absolute value of the correlation implies higher importance of the node centrality in the feature space. Figure 6 shows the changes of the feature importance of the first 6 nodes with respect to different values of $k$. As expected, the node centralities are positive correlated with the wiener index (or complexity) of the cascade. This is because a node with high centrality has high influence and low homophily, which is more likely to cause rapid and complex structures. The correlation grows with the number of observed nodes $k$ in the early cascade. But when $k$ reaches 30, the correlation has diminishing returns and stablizes around a certain value.

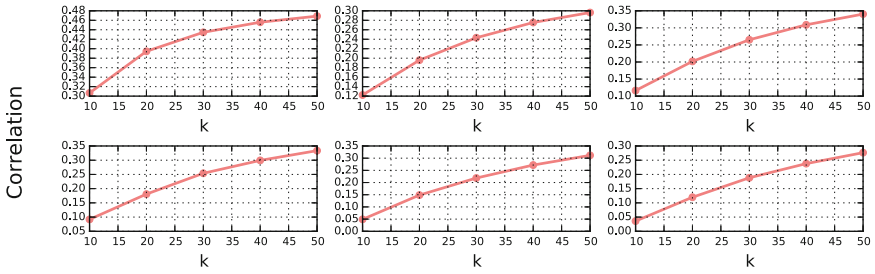



**Fig. 6.** Average Pearson correlation between node centralities and the cascade wiener index.

According to the above observation, we now predict whether a cascade will have a wide structure or a deep structure by observing the first $k$ nodes of the cascade. We use wiener index to measure the structure of a cascade. Generally, a tree with wiener index approximately 2.0 is wide shallow, while a tree with higher index indicates it has complex structure. In Table 2, we show our results for predicting whether the wiener index of the cascade will evolve to be below or above a chosen value of wiener index. The results include the precision, recall, F1-score and accuracy. To avoid the imbalance of data set, we set the value of wiener index for prediction as 2.05. As presented in Table 2, our SP-based method can reach the best result in almost all cases, showing the effectiveness of the structure pattern features. In comparison to our method, the SP-blind, PCA-based and Wiener-based based method have worse results, since they did not consider the mechanics of information diffusion. According to the selection of wiener index, the random method reaches almost 50 % in every result.

**Table 2.** Predicting cascade structures.

| Algorithm | Prec. | Rec. | F1 | Accu. |
|---|---|---|---|---|
| SP-based | **0.775** | 0.641 | **0.697** | **0.722** |
| SP-blind | 0.591 | **0.643** | 0.613 | 0.596 |
| PCA-based | 0.597 | 0.628 | 0.607 | 0.598 |
| Wiener-based | 0.596 | 0.642 | 0.615 | 0.601 |
| Random | 0.504 | 0.505 | 0.500 | 0.499 |

### 5.3 Predicting Cascade Virality

Another important application of the structure pattern is to predict the virality of a cascade by observing its early footprints. As shown in Fig. 3, the cascades in the same cluster tend to have the same size. Based on this observation, in

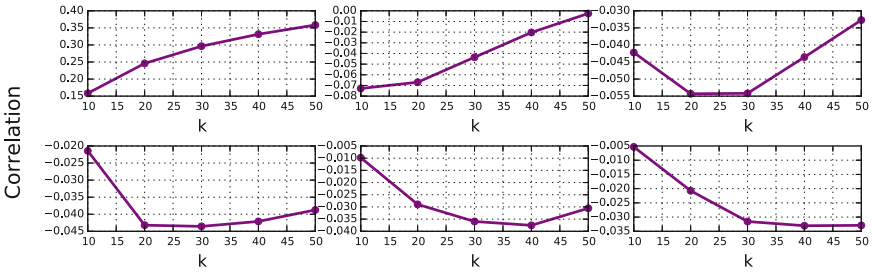



**Fig. 7.** Average Pearson correlation between the cascade size and node centralities.

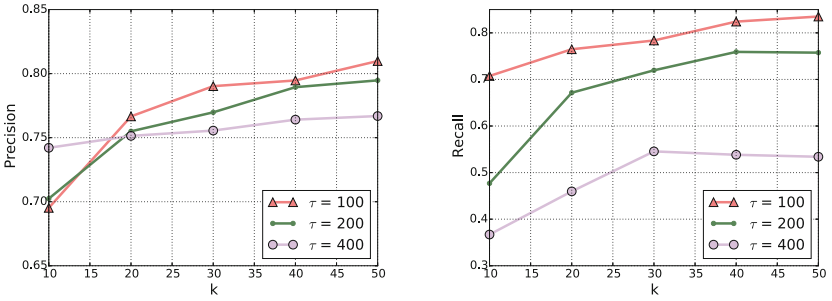
this section, we aim to predict the virality of the cascades using the structural patterns.

We start by observing the Pearson correlation between the node centralities and the size of the eventual cascades in Fig. 7. Surprisingly, we find that from the second node, the centralities are negative correlated with the size of the cascades. This implies that the cascades are more likely to be viral if the structures of cascade are simple and wide. The centrality of the second node becomes less important when observing more nodes. This may be the reason that a node with high centrality is also common in homophily-driven cascades. And in most other cases, the absolute value of the correlation would increase with the size initial cascade.

**Table 3.** Predicting cascade virality with respect to different thresholds.

| Threshold | Algorithm | Prec. | Rec. | F1 | Accu. |
|---|---|---|---|---|---|
| 100 | SP-based | **0.790** | **0.783** | **0.775** | **0.809** |
| | SP-blind | 0.742 | 0.732 | 0.727 | 0.753 |
| | PCA-based | 0.768 | 0.776 | 0.755 | 0.788 |
| | Wiener-based | 0.742 | 0.737 | 0.730 | 0.755 |
| | Random | 0.463 | 0.498 | 0.477 | 0.499 |
| 200 | SP-based | **0.769** | **0.719** | **0.727** | **0.836** |
| | SP-blind | 0.690 | 0.493 | 0.534 | 0.738 |
| | PCA-based | 0.698 | 0.465 | 0.517 | 0.738 |
| | Wiener-based | 0.674 | 0.485 | 0.529 | 0.742 |
| | Random | 0.327 | 0.495 | 0.395 | 0.499 |
| 400 | SP-based | 0.755 | **0.545** | **0.589** | **0.847** |
| | SP-blind | 0.731 | 0.368 | 0.426 | 0.817 |
| | PCA-based | **0.771** | 0.455 | 0.516 | 0.835 |
| | Wiener-based | 0.733 | 0.368 | 0.425 | 0.813 |
| | Random | 0.229 | 0.493 | 0.310 | 0.499 |

Next, we predict whether the cascade will reach a certain number of nodes. We try different values of the threshold as 100, 200 and 400. Table 3 shows the results of the predictions. Obviously, our SP-based method performs the best in almost all cases. And according to the setting of the experiment, the accuracy of the Random method is almost around 50 % when the threshold is 100. When the threshold increases, generally, the precision and the recall will decrease as the cascades become more and more unpredictable. However, our SP-based algorithm could still reach a high F1-measure. The significant improvement in the results validates the effectiveness of the importance of the structure pattern features. In predicting the virality, we should try to identify as many viral cascades as possible. Thus, recall is often a critical measure. And in all cases, our SP-based method can has the highest recall of all.




(a) (b)

**Fig. 8.** Predicting virality with respect to different values of $k$.

Finally, we examine the effect of the prediction with respect to different values of $k$ from 10 to 50. The results are shown in Fig. 8. As presented, the precision and recall are almost monotone with the increase of $k$. The values increases quickly at early steps. But when $k$ exceeds 30, the values are not likely to grow too much. This also demonstrates that $k = 30$ is reasonable for predicting the growth of the cascades.

## 6 Conclusions

In this paper, we studied the structures of cascades in online social networks and explore the relationships with information diffusion. By embedding the cascades in a lower dimensional space and employing the spectral clustering algorithm, we can identify five typical patterns of the cascade structures with distinguishable characteristics. In addition, since the structure patterns of the cascades can be identified based on the early footprints, we can incorporate the structure patterns to predict the growth of information cascades.

The analysis and experiments of the results are based on the Weibo platform. We believe that the Weibo data set is comprehensive for the information cascades

since it has large scale of data and includes topics across different disciplines. For the future work, first, we would empirically analyze the effect of our algorithm on other data sets; On the other hand, we would use the structure patterns to guide the micro analysis of user behaviors in social networks, so that we can predict even the individual behaviors more accurately.

**Acknowledgements.** This work was partially supported by the National Natural Science Foundation of China under Grant Nos. 61321491, 61472181, 91218302, the Natural Science Foundation of Jiangsu Province of China under Grant No. BK20151392, Jiangsu Key Technique Project (industry) under Grant No. BE2013116, EU FP7 IRSES MobileCloud Project under Grant No. 612212, the Program B for Outstanding PhD candidate of Nanjing University, and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

## References

1. Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J., Tiwari, M.: Global diffusion via cascading invitations: structure, growth, and homophily. In: Proceedings of the 24th International Conference on World Wide Web, pp. 66–76. International World Wide Web Conferences Steering Committee (2015)
2. Anderson, R.M., May, R.M., Anderson, B.: Infectious diseases of humans: dynamics and control, vol. 28. Wiley Online Library (1992)
3. Aral, S., Muchnik, L., Sundararajan, A.: Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proc. Natl. Acad. Sci. **106**(51), 21544–21549 (2009)
4. Bailey, N.T., et al.: The mathematical theory of infectious diseases and its applications. Charles Griffin & Company Ltd., 5a Crendon Street, High Wycombe, Bucks HP13 6LE (1975)
5. Bakshy, E., Karrer, B., Adamic, L.A.: Social influence and the diffusion of user-created content. In: Proceedings of the 10th ACM Conference on Electronic Commerce, pp. 325–334. ACM (2009)
6. Budak, C., Agrawal, D., El Abbadi, A.: Structural trend analysis for online social networks. Proc. VLDB Endowment **4**(10), 646–656 (2011)
7. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 199–208. ACM (2009)
8. Cheng, J., Adamic, L., Dow, P.A., Kleinberg, J.M., Leskovec, J.: Can cascades be predicted? In: Proceedings of the 23rd International Conference on World Wide Web, pp. 925–936. ACM (2014)
9. Goel, S., Watts, D.J., Goldstein, D.G.: The structure of online diffusion networks. In: Proceedings of the 13th ACM Conference on Electronic Commerce, pp. 623–638. ACM (2012)
10. Gomez Rodriguez, M., Leskovec, J., Krause, A.: Inferring networks of diffusion and influence. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1019–1028. ACM (2010)
11. Granovetter, M.: Threshold models of collective behavior. Am. J. Sociol. **83**, 1420–1443 (1978)
12. Hong, L., Dan, O., Davison, B.D.: Predicting popular messages in twitter. In: Proceedings of the 20th International Conference Companion on World Wide Web, pp. 57–58. ACM (2011)

13. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 137–146. ACM (2003)
14. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 420–429. ACM (2007)
15. Liben-Nowell, D., Kleinberg, J.: Tracing information flow on a global scale using internet chain-letter data. Proc. Natl. Acad. Sci. **105**(12), 4633–4638 (2008)
16. Malmgren, R.D., Stouffer, D.B., Motter, A.E., Amaral, L.A.: A poissonian explanation for heavy tails in e-mail communication. Proc. Natl. Acad. Sci. **105**(47), 18153–18158 (2008)
17. Matsubara, Y., Sakurai, Y., Prakash, B.A., Li, L., Faloutsos, C.: Rise and fall patterns of information diffusion: model and implications. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 6–14. ACM (2012)
18. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Ann. Rev. Sociol. **27**, 415–444 (2001)
19. Nematzadeh, A., Ferrara, E., Flammini, A., Ahn, Y.Y.: Optimal network modularity for information diffusion. Phys. Rev. Lett. **113**(8), 088701 (2014)
20. Romero, D.M., Meeder, B., Kleinberg, J.: Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In: Proceedings of the 20th International Conference on World Wide Web, pp. 695–704. ACM (2011)
21. Singer, Y.: How to win friends and influence people, truthfully: influence maximization mechanisms for social networks. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, pp. 733–742. ACM (2012)
22. Tsur, O., Rappoport, A.: What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, pp. 643–652. ACM (2012)
23. Ugander, J., Backstrom, L., Marlow, C., Kleinberg, J.: Structural diversity in social contagion. Proc. Natl. Acad. Sci. **109**(16), 5962–5966 (2012)
24. Weng, L., Menczer, F., Ahn, Y.Y.: Virality prediction and community structure in social networks. Sci. Rep. **3**, 2522 (2013)
25. Yang, J., Leskovec, J.: Patterns of temporal variation in online media. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 177–186. ACM (2011)
26. Yavaş, M., Yücel, G.: Impact of homophily on diffusion dynamics over social networks. Soc. Sci. Comput. Rev. **32**, 0894439313512464 (2014)