

Discovering Information Propagation Patterns in Microblogging Services

ZHIWEN YU, ZHU WANG, and HUILEI HE, Northwestern Polytechnical University

JILEI TIAN, BMW Technology

XINJIANG LU and BIN GUO, Northwestern Polytechnical University

During the last decade, microblog has become an important social networking service with billions of users all over the world, acting as a novel and efficient platform for the creation and dissemination of real-time information. Modeling and revealing the information propagation patterns in microblogging services cannot only lead to more accurate understanding of user behaviors and provide insights into the underlying sociology, but also enable useful applications such as **trending prediction, recommendation and filtering, spam detection and viral marketing**. In this article, we aim to reveal the information propagation patterns in Sina Weibo, the biggest microblogging service in China. **First, the cascade of each message is represented as a tree based on its retweeting process. Afterwards, we divide the information propagation pattern into two levels, that is, the macro level and the micro level.** On one hand, the macro propagation patterns refer to general propagation modes that are extracted by grouping propagation trees based on hierarchical clustering. On the other hand, the **micro propagation patterns are frequent information flow patterns that are discovered using tree-based mining techniques**. Experimental results show that several interesting patterns are extracted, such as popular message propagation, artificial propagation, and typical information flows between different types of users.

Categories and Subject Descriptors: H.1.2 [User/Machine Systems]: Human Factors; H.2.8 [Database Applications]: Data Mining

General Terms: Algorithms, Human Factors

Additional Key Words and Phrases: Information propagation pattern, message cascade, propagation tree, microblogging services.

ACM Reference Format:

Zhiwen Yu, Zhu Wang, Huilei He, Jilei Tian, Xinjiang Lu, and Bin Guo. 2015. Discovering information propagation patterns in microblogging services. *ACM Trans. Knowl. Discov. Data* 10, 1, Article 7 (July 2015), 22 pages.

DOI: <http://dx.doi.org/10.1145/2742801>

This work was partially supported by the National Basic Research Program of China (No. 2012CB316400), the National Natural Science Foundation of China (No. 61222209, 61373119, 61332005, 61402369), the Program for New Century Excellent Talents in University (No. NCET-12-0466), and the Specialized Research Fund for the Doctoral Program of Higher Education (No. 20126102110043).

Authors' addresses: Z. Yu, Z. Wang, H. He, X. Lu, and B. Guo, School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi, China, 710072; emails: {zhiwenyu, wangzhu}@nwpu.edu.cn, huilei@mail.nwpu.edu.cn, ramber1836@gmail.com, guob@nwpu.edu.cn; J. Tian, BMW Technology, Chicago, USA; email: jilei_tian@hotmail.com.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

© 2015 ACM 1556-4681/2015/07-ART7 \$15.00

DOI: <http://dx.doi.org/10.1145/2742801>

1. INTRODUCTION

With the breakthrough of web 2.0 technologies, there has been an unprecedented booming of social media services. As a representative category of such services, the microblogging service, such as Twitter¹ and Sina Weibo², emerges and quickly becomes extremely popular. Millions of people are using these services on a daily basis to create, communicate, and share contents on an enormous scale. With its huge number of users and contents, **microblog is now recognized as one of the most important types of media**, which has profoundly changed the way people acquire knowledge, share information and interact with each other on a societal scale.

As a high-quality information propagation platform, the microblog acts as a central domain for the creation and dissemination of real-time information. **Users may adopt the microblogging service for many purposes, such as sharing news, promoting political views, showing off, marketing, and tracking real-time events [Boyd et al. 2010; Java et al. 2007; Zhao and Rosson 2009].** Specifically, users post short messages of up to 140 characters containing various contents, ranging from daily activity updates, discussions, photos to interesting URLs and spontaneous thoughts. Meanwhile, people may *follow* one another to receive up-to-date messages published by interested users and get *followed* by other users to spread out their messages. Those who follow a user are called followers and those whom a user follows are called followees. Since following someone does not necessarily mean that she/he will follow you back, **the followees-followers network is a directed graph [Zhou et al. 2010].** A user can retweet messages posted by other people and also use @ to mention others and address messages to them directly.

The intuitive usage of tweets has made it possible for the **swift** propagation of news and messages in the microblog network [Java et al. 2007]. Specifically, the *followees-followers* social graph forms a unique and dynamic information propagation platform that might be leveraged to facilitate various social needs, for example, **the propagation of breaking news, the broadcasting of emergencies, the marketing of products and services, et al.** Modeling and understanding the information flows in microblogging services can potentially lead to more effective use of this new social media and provide insights into the underlying sociology.

Microblogging services have been appreciated in people's daily life in China. However, unlike Twitter which has been well studied, fewer studies have been done on Sina Weibo. Even though Sina Weibo is much younger than Twitter, its number of users and messages are tremendous. Meanwhile, the unique cultural and social environments in China also suggest that people's online behaviors might be different from that of western counterparts. **For instance, Yu et al. [2012a] found that the effect of retweets is much larger in Sina Weibo where users are more likely to learn about a particular topic through retweets.** To this end, we mainly focus on exploring the retweeting patterns of Sina Weibo in this article. To simplify the discussion, we will follow the terminology used in Twitter.

Retweeting is one of the most important features in Sina Weibo, which relays a certain tweet that has been written by another user. When a user finds an interesting tweet published by someone and intends to share it, she/he can simply retweet the message. Retweeting can be considered as an efficient way of information propagation since the original tweet is propagated to a new set of users, that is, followers of the retweeter. **Currently, there are two different lines of research that are concentrated on the retweet feature of microblogging services.** The first line of research focuses on

¹<http://www.twitter.com>.

²<http://www.weibo.com>.

information propagation in the microblog sphere, which show that the message content and author influence act as two major factors impacting the dissemination of microblogging messages. For example, some studies revealed that the content and sentiment of messages play important roles in the propagation [Cheong and Lee 2010; Wu and Ren 2011; Zhang et al. 2010]. Some other studies devoted to quantifying the user influence by analyzing message cascades or user relationships [Fan et al. 2011; Suh et al. 2010]. The other line of research focuses on the retweeting behavior of microblog users. For example, some studies explored how the user's retweeting behavior or retweeting probability is influenced by different factors [Boyd et al. 2010; Galuba et al. 2010; Yang et al. 2010; Zhou et al. 2010], for example, content, authorship, and attributes. Some other studies examined retweeting behaviors from the perspective of information spreading [Kwak et al. 2010; Romero et al. 2011; Yang and Counts 2010], for example, how retweeting impacts the speed, scale, and range of information diffusion. However, none of these studies had revealed the information propagation patterns by exploiting the characteristics of message cascades (e.g., propagation range and timeliness) or the frequent information flow patterns from the user's perspective (e.g., how does the message propagate between users with different influence). Furthermore, Sina Weibo is different from Twitter in several aspects [Wang and Wu 2011; Wang et al. 2012; Chen et al. 2011], which creates new research opportunities. For example, Twitter had been text only for a long time, while Sina Weibo allows users to post images, videos, audios, emotions and even launching votes as attachments at the very beginning. Moreover, in Sina Weibo the most propagation happens due to retweets of media contents such as jokes, images, and videos, whereas the current global events and news cause most retweets in Twitter [Yu et al. 2011].

To distinguish different information propagation patterns and discover frequent information flow patterns on Sina Weibo, we first crawl the diffusion trace data for each selected message. Afterwards, we explore the information propagation pattern of Weibo messages from both macro and micro perspectives by leveraging a tree-based propagation model. Specifically, the contributions of this article are threefold:

- We propose a tree-based model to represent the information diffusion process of each message, in which each node corresponds to a user while the root node represents the originator of a cascaded message, each directed edge stands for a certain retweet action, and each node can have many child nodes.
- Hierarchical clustering is applied to group propagation trees and reveal propagation patterns from the macro level perspective. Particularly, 24 features are selected to characterize the message cascade, which mainly include three sets of properties: the number of participants, the topology of the propagation trace, and the promptness of the information dissimulation in terms of time. Eight distinct clusters of message cascade are extracted, which correspond to different propagation patterns. Afterwards, we investigate two significant patterns in more depth: the popular message propagation and the artificial inflation propagation.
- Tree-based frequent pattern mining is used to discover the frequent information flow patterns from the micro level perspective. In particular, we propose an incremental algorithm to discover micro propagation patterns by simultaneously constructing the set of frequent patterns and their occurrences level by level, based on which we obtain a number of interesting findings. For instance, there is a significant polarization phenomenon among users in the Weibo sphere. Although both the high-level and low-level users play important roles in information propagation, the middle-level users are less determinative for the popularity of Weibo messages.

The rest of the article is structured as follows. Section 2 discusses the related work, followed by the representation of information cascade in Section 3. Section 4 presents

the clustering of the propagation trees, and the mining of information cascade patterns is described in Section 5. Experimental results and potential applications are depicted in Sections 6 and 7, and then we conclude our work in Section 8.

2. RELATED WORK

Microblog has attracted much attention from the research community since it became an important social networking service in the last decade. As a typical and worldwide microblogging service, Twitter has been well studied. Most early studies aim to explore the basic properties of the Twitter network. For example, Kwak et al. [2010] conducted a large-scale study to analyze the topology characteristics of Twitter and its power as a new medium of information sharing. Java et al. [2007] provided initial analysis on the topological and geographical properties of the Twitter social graph along with observations on what type of contents people used to tweet. Huberman et al. [2009] performed a more detailed investigation on the social network, trying to better understand the nature of social interactions on Twitter. Cha et al. [2010] developed a framework to measure and model the individual's influence on Twitter, and found that a high follower count does not necessarily lead to a large number of retweets. Meanwhile, a number of studies have been focused on developing tools on top of Twitter. For example, Jansen et al. [2009] used Twitter to share consumer opinions about brands. Ramage et al. [2010] built tools to group tweets into topics to support fast browsing. Chen et al. [2010] examined the personal recommendation of URL items in the Twitter stream. In general, while the earlier-mentioned studies mainly aim at analyzing the basic characteristics of the Twitter network and designing tools on top of the Twitter platform, our work focuses on revealing the higher level knowledge of microblogging services, that is, discovering patterns of information propagation.

As resharing/retweeting becomes the key mechanism for spreading information in online social networks (i.e., Twitter and Facebook), there have been a number of studies on people's retweeting behaviors to explore how information is diffused and whether the future trajectory of a cascade can be predicted. For example, the propagation graph and statistics are studied in Kwak et al. [2010] and [Galuba et al. 2010]. Boyd et al. [2010] examined retweeting as a conversational practice, and highlighted how authorship, attribution, and communicative fidelity are negotiated in various ways. To understand why some tweets spread more widely than others, Suh et al. [2010] investigated a number of content and contextual features that have potential relationships with the retweet ability of tweets, based on which a predictive retweet model was proposed using Generalized Linear Model. Yang et al. [2010] found that almost 25.5% of the tweets posted by users are actually retweeted from the blogs of their friends and proposed a factor graph model to predict the retweeting behaviors. Zarrella et al. [2009] observed that retweets and tweets are different in several dimensions, for example, the inclusion of URL. Yang et al. [2010] studied the underlying mechanism of retweeting behaviors. Lin et al. [2013] found that the underlying reasons of Twitter events include not only social influence inside the network but also external trends in the "real world". Therefore, they proposed to improve the learning of information diffusion models by extracting social events from data streams in Twitter network. Eftekhari et al. [2013], on the other hand, proposed a fine-grained model of information diffusion from the group perspective together with a coarse-grained model that inspects the network at group level. Cheng et al. [2014] developed a framework to address the cascade prediction problem, and achieved good performance in predicting whether a cascade will continue to grow in the future. They found that while temporal and structural features are key predictors of cascade sizes, the initial breadth rather than depth in a cascade is a better indicator of larger cascades. In general, these studies mainly focus on analyzing the related factors of retweeting from the information spreading perspective

rather than the user's perspective. However, the information flows in social networks usually carry rich information about user behaviors, therefore an open issue is what are the information propagation patterns from the user's perspective (e.g., how does the message propagate between users with different influences) and why information spreads in such ways? Wang et al. [2014a] tried to address this issue by inferring multi-aspect diffusion patterns with multi-pattern cascades, and studied the effects of various diffusion patterns on the information diffusion process by analyzing users' retweeting behaviors. However, we have totally different purposes and adopt distinct methods.

So far, few studies have been conducted on Sina Weibo. Wang et al. [2012] conducted a research to gain an in-depth understanding of what kinds of users are more active and what kinds of contents are favored the most for retweeting. Qu et al. [2011] studied the roles played by microblogging services in response to major disasters. Zhang et al. [2012] built a retweet model to predict the number of possible-views of a certain tweet. The differences between Sina Weibo and Twitter have been studied [Chen et al. 2011; Wang and Wu 2011; Wang et al. 2012]. For example, Yu et al. [2011] studied the trending topics in Sina Weibo and found that there are numerous differences between the content shared on Sina Weibo and Twitter. In this article, we mainly focus on exploiting different information propagation patterns and revealing how information spreads among different categories of users on Sina Weibo.

3. INFORMATION CASCADE REPRESENTATION

3.1. Definition of Weibo Propagation Tree

Once a piece of information is generated in the microblog network, it spreads in a complex cascaded way, forming an information flow tree [Leskovec et al. 2007a]. First, as one of the most important functions of microblog services, retweetings enable users to rebroadcast someone else's messages to their own followers. It means that once a message is retweeted by a user, it first spreads to the retweeter's followers, and then followers who are interested in the message may further retweet the message. By representing the original tweet as a root vertex, and placing the boosting retweets as children or grandchildren of the root, we can build a multi-level tree-like information diffusion structure for each original tweet.

Secondly, users in the microblog sphere would receive messages published by their followees in a single reverse-chronological order. Therefore, we can construct the propagation tree based on the follower/followee relationship among the involved users in the ordered retweet list. Specifically, the children in each depth of the tree are ranked in a chronological order, where the left vertices represent retweets that occur earlier than the right ones.

Finally, in order to keep the constructed tree manageable, we represent vertices in the propagation tree with a small number of labels ($L = \{l_1, l_2 \dots\}$). Specifically, a labeling function $l: V \rightarrow L$ is introduced using the follower number as the labeling criterion, which has been proved to be an important factor that influences the popularity of microblog messages [Ma et al. 2013]. Intuitively, a message from a user with millions of followers is much more influential than that from a user with tens of followers, and a retweet by a popular user may increase the popularity of the original message. In this article, we use five different labels ($L = \{a, b, c, d, e\}$) to indicate that the follower number of a Weibo user is larger than 100K, between 10K and 100K, between 1K and 10K, between 100K and 1K, and less than 100, respectively.

Particularly, in case that users retweet on Weibo messages published by themselves, loops may exist in the diffusion structure. To cope with this, we define that a user could appear more than once in the diffusion structure (i.e., there could be more than one

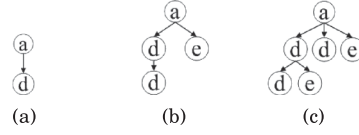


Fig. 1. Examples of weibo message propagation tree.

nodes that correspond to the same user). The reason why a new node rather than a loop should be generated in such cases is that the propagation tree is used to describe the diffusion of messages, and retweeting on the message published by oneself also contributes to the propagation process. Consequently, we are able to represent the propagation cascade of any Weibo message as a tree, which can be formally defined as follows.

Definition 1-Weibo Propagation Tree. A propagation tree $T = \langle V, E \rangle$ is **directed**, **ordered** (i.e., sequential), and **labeled**. In detail, $V = \{0, 1, \dots, n\}$ is the set of vertices representing Weibo users, and a distinguished vertex $v_r \in V$ is designated as the root node. $E = \{(v_i, v_j) \mid v_i, v_j \in V, v_i \neq v_j\}$ is the set of edges among users, in which a directed edge $(v_i, v_j) \in E$ means that v_i is the parent of v_j , that is, v_j retweeted a message from v_i .

Figure 1 illustrates the propagation trees of three Weibo messages, in which the arrow denotes that one user (the start of an arrow) has been retweeted by another user (the end of the arrow). For example, the user labeled as a in Figure 1(b) had a message retweeted by two other users who are labeled as d and e . Furthermore, d retweeted earlier than e , and d had been retweeted by another user who is also labeled as d .

Compared with other representation methods, such as **array based representation**, the earlier tree-based representation has several advantages. First, as the tree-based representation allows the emergence of a structure by means of the rightmost expansion, we are able to incrementally discover frequent tree-like patterns. Specifically, the attribute (e.g., the support value) of a candidate subtree T_s expanded from T_x can be incrementally calculated based on the corresponding attribute of T_x , which can significantly reduce the computational complexity. Second, as the directed, ordered, and labeled tree perfectly characterizes the diffusion process of microblog messages, and the obtained frequent propagation patterns correspond to subtrees with similar forms, we are capable of clearly interpreting the meanings of these frequent subtrees.

3.2. Construction of Weibo Propagation Tree

The relationship among Weibo users may change dynamically, as it is common for users to add or remove their followees. Meanwhile, it is difficult to obtain the complete structure of a certain retweet process due to the limitation of Sina Weibo APIs³. Therefore, we formulate the Weibo propagation tree construction in Algorithm 1 to get the optimal approximation of the information diffusion path.

Specifically, given a retweet list RL , the algorithm first labels each user based on one's corresponding number of followers as mentioned earlier and sorts the retweet list in the chronological order (lines 1–4), followed by the initialization process (lines 5–7). Afterwards, for each user u in RL , we iterate to select the appropriate parent node p from P_i and then add u to the propagation tree (lines 9–16). The construction of next-level propagation tree continues by adding 1 to i (line 17) until either RL or P_i is empty. If RL is not empty, the remaining users in RL are added as child nodes of the root (lines 19–21). These users are not followers of any users in the propagation tree, as they retweet the message through other approaches, such as trending topics and search engine. Finally, we store the generated propagation tree in a xml file (line 22).

³<http://open.weibo.com/wiki/2/>.

ALGORITHM 1: *PTC(RL)* - Propagation Tree Construction**Input:**

— retweet list for a given original tweet, RL

Output:


— propagation tree in the xml format

```

1: for each retweet user  $u \in RL$  do
2:   label  $u$  using the labeling function  $l$ 
3: end for
4: sort  $RL$  using the retweet time from earlier to late
5: init root node of the propagation tree as  $r$  using author of the tweet
6:  $i \leftarrow l$ 
7: add  $r$  to the parent nodes set  $P^i$ 
8: repeat
9:   for each retweet user  $u \in RL$  do
10:    for each candidate parent node  $p$  in  $P^i$  do
11:      if  $p$  is the followee of  $u$  and  $u$  retweet from  $p$  then
12:        add  $u$  as the right most child of  $p$ , and then remove it from  $RL$ 
13:         $P^{i+1} \leftarrow P^i \cup u$ 
14:      end if
15:    end for
16:   end for
17:    $i \leftarrow i + 1$ 
18: until  $RL = \emptyset$  and  $P^i = \emptyset$ 
19: if  $RL \neq \emptyset$  then
20:   add the remaining users in  $RL$  as child nodes of  $r$ , ranking based on the retweet time
21: end if
22: record the propagation tree using xml

```

4. MACRO PROPAGATION PATTERN EXTRACTION

In  section, we attempt to reveal the macro level patterns of information propagations in Sina Weibo by **clustering the propagation trees**. We first give the definition of **macro propagation pattern**, and then present the features used for propagation pattern extraction.

Definition 2 (Macro Propagation Pattern). The macro propagation pattern refers to common information dissemination patterns of Weibo propagation trees.

As the **topology** and the **propagation time** are important characteristics of message propagation, we extract prominent features of propagation trees from three aspects: **the size of the propagation tree, the topology of the tree, and the distribution of messages along time**. The extracted features are listed in Table I.

To characterize the size of a propagation tree, we extract 4 features. On one hand, the total number of nodes (*totalNode*), the number of leaf nodes (*leafNode*), and the number of parent nodes (*parentNode*) directly reflect the size of a propagation tree. On the other hand, we define the ratio between *parentNode* and *totalNode* as *ptRatio*.

In terms of the topology of propagation trees, **considering that most of the messages propagate within 5 hops** from the originator [Kwak et al. 2010], we extract features including the proportion of retweets at each level (i.e., *level2p-level5p*), the proportion of *level > 5* (*levelDeep*), as well as the *depth* and *width* of propagation trees. Meanwhile, to **measure the number of key contributors for a message cascade**, we calculate the smallest number of users who accounts for more than 70% of the total retweets, and name it as *bigNode*. Specifically, we first sort the contribution of each user in term of

Table I. Variables to Represent a Message Cascade

Variable	Description
totalNode	number of total nodes
leafNode	number of leaf nodes
parentNode	number of parent nodes
ptRatio	ratio of parentNode to totalNode
bigNode	number of users who retweeted 70% of the total retweets
bigNodeRatio	ratio of bigNode to totalNode
overlapNode	number of non-first time retweets
overlapNodeRatio	ratio of overlapNode to totalNode
depth	depth of the propagation tree
width	width of the propagation tree
level2p - level5p	The proportions of nodes with each level (2-5) in the structure
levelDeep	The sum proportions of nodes with level > 5 in the structure
timeLength	the time span
maxRate	the max number of retweets per hour
avgRate	the average number of retweets per hour
beginTime	the time span with 10% of total retweets at beginning
experienceTime	the time span with 80% of total retweets at duration
finalTime	the time span with 10% of total retweets after experienceTime
beginTimeRatio	the ratio of beginTime to timeLength
experienceTimeRatio	the ratio of experienceTime to timeLength
finalTimeRatio	the ratio of finalTime to timeLength

Table II. Pearson Correlation Coefficient of the Variables

	totalNode	leafNode	parentNode	width	bigNode
totalNode	1.000	0.998	0.898	0.982	0.928
leafNode	0.998	1.000	0.866	0.987	0.939
parentNode	0.898	0.866	1.000	0.837	0.752
width	0.982	0.987	0.837	1.000	0.962
bigNode	0.928	0.939	0.752	0.962	1.000

retweet amounts in a descending order, and then choose the top users that account for 70% of the retweets. A larger value for *bigNode* indicates that more users contribute to the message propagation. The *overlapNode* is the number of retweets that retweeted by the already-retweeted users. Furthermore, we calculate the ratio of *bigNode* to *totalNode* (i.e., *bigNodeRatio*) as well as the ratio of *overlapNode* to *totalNode* (i.e., *overlapNodeRatio*).

To characterize the **temporal distribution** of messages, we introduce another nine features. Specifically, we define the time span of a message cascade as *timeLength*, the average number of retweets per hour as *avgRate*, and the maximum number of retweet per hour as *maxRate*. The time span of a message cascade is divided into three segments and the length of the segments are denoted as *beginTime*, *experienceTime* and *finalTime*, respectively. The *beginTime* and *finalTime* are defined as the segments at the beginning and in the end of the entire time span, where the number of retweets accounts for 10% of the total amounts. In other words, *beginTime* measures the length of time when the topic is heating up, *experienceTime* is the time span when the majority of retweets are posted, and *finalTime* is the length of the cooling off period. We also represent the length of these three segments in terms of their proportion to *timeLength* and denote them as *beginTimeRatio*, *experienceTimeRatio* and *finalTimeRatio*.

In total, we have identified 24 features to represent the message cascade, as summarized in Table I. However, some of these features are highly correlated with each other. For example, Table II shows the Pearson correlation coefficient among five features.

Table III. Symbols

Notation	Description
t	A tree
TS	A tree set consists of propagation trees
t^k	A subtree with k nodes, that is, k -subtree
C^k	A set of candidates with k nodes
F^k	A set of frequent k -subtrees
σ	A support threshold $minSupp$

As highly correlated variables might disrupt the clustering, we use factor analysis to describe the earlier 24 variables using a smaller number of uncorrelated new factors. Specifically, these 24 variables are modeled as a linear combination of the new factors. By choosing factors corresponding to larger eigenvalues ($\lambda \geq 1$), we obtain **seven factors**.

We adopt the hierarchical clustering algorithm [Bandyopadhyay and Coyle 2003] to group propagation trees, where the **Euclidean distance** is used as the distance measure. We first generated a dendrogram for the dataset used in this article, based on which the best sub-group structure could be obtained. Specifically, by introducing different cut thresholds, we got the most appropriate inter-cluster distance when the number of clusters is **35**. **We ignored clusters that consist of less than five cascades** (i.e., clusters that do not represent significant propagation patterns) and as a result eight significant clusters are obtained. The detailed results will be described in **Section 6**.

However, it is well known that the hierarchical clustering is **computationally expensive**, especially when there are a large number of items to be clustered. Fortunately, in this article we don't need to deal with a large amounts of propagation trees. Specifically, a reasonable assumption is that there is only a finite number of possible information propagation patterns in the microblog sphere. Therefore, **in order to extract all the possible information propagation patterns** (i.e., tree clusters), we gradually perform hierarchical clustering on different amounts of trees (e.g., 100, 1000, 10,000, ...) until the number of significant clusters reaches the maximum value and stops to increase (i.e., the pattern training/extraction phase). **Afterwards, given any new propagation tree, we can assign it to the closest cluster by calculating its distance between all the extracted propagation patterns (i.e., the pattern testing/using phase).**

Moreover, some recent studies shown that the performance of **hierarchical clustering** can be notably improved by introducing appropriate modifications [Krishnamurthy et al. 2012; Wang et al. 2014b]. Thereby, even though the hierarchical clustering is relatively computation expensive, it is suitable for the extraction of macro information propagation patterns as it has the following advantages. **On one hand, it is capable of revealing cluster structures regarding all the possible granularity, which assure that all the significant propagation patterns can be extracted. On the other hand, compared with other methods, the hierarchical clustering algorithm is more versatile and able to handle any forms of similarity or distance.**

5. MICRO PROPAGATION PATTERN MINING

In this section, we aim to discover the micro level information propagation patterns in Sina Weibo based on **frequent pattern mining techniques**. We first give the definitions of *subtree*, *tree set size*, *support* of a tree/subtree, and *micro propagation pattern*. Afterwards, we present the proposed incremental mining algorithm. Particularly, a list of used symbols is summarized in Table III.

Definition 3 (Subtree). Given a tree $T = (V, E)$, tree $T' = (V', E')$ is called a subtree of T , denoted as $T' \subseteq T$, if

- a) $V' \subseteq V$ and
- b) $E' \subseteq E$ and

c) $\forall v_i \in V, L(v_i) = L'(v_i)$ and

d) $\forall v_i, v_j \in V$, there is a path between v_i and v_j with all edges in E' regarded as undirected.

For example, the propagation tree shown in Figure 1(a) is a subtree of the two trees shown in Figure 1(b) and 1(c).

Definition 4 (Tree Set Size). We use $|TS|$ to represent the numbers of nodes in a tree set named TS.

Definition 5 (Support). Given a subtree T and a tree set TS , the support of T is defined as:

$$\text{supp}(T) = \frac{\text{number of occurrences of } T}{\text{total number of vertices in } TS}. \quad (1)$$

If the value of $\text{supp}(T)$ exceeds the predefined threshold value minSupp (e.g., 1%), T is called a *frequent subtree*.

Definition 6 (Micro Propagation Pattern). A micro propagation pattern is a frequent subtree in the tree set.

Given a minimum support threshold σ , we aim to find all the subtrees that appear at least $\sigma \times |TS|$ times in the set.

Based on the graph mining [Chakrabarti and Faloutsos 2006] and frequent pattern mining algorithms [Asai et al. 2002; Yu et al. 2012b], we put forward a tree mining approach which is able to incrementally discover frequent tree-like patterns within the tree set. To get all frequent subtrees according to the minSupp , the *support* of each subtree is calculated. The procedure of the mining process is presented in Algorithm 2.

The algorithm first calculates the corresponding *support* for nodes with distinct labels, and then selects the nodes whose *supports* are larger than σ to form the set of frequent nodes F^1 (lines 1–3). Afterwards, it invokes the *Right_Most_Expand*

ALGORITHM 2: $FPSPM(TS, \sigma)$ - Frequent Propagation SubTree Pattern Mining

Input:

- a tree set consists of propagation trees, TS
- support threshold, σ

Output:

- all frequent tree patterns with respect to σ

```

1:  $i \leftarrow 1$ 
2: scan  $TS$ , calculate the support for each labeled node
3: select the nodes whose support are larger than  $\sigma$  to form  $F^i$ 
4: repeat
5:   for each tree  $t^i$  in  $F^i$  do
6:      $\text{expandResult} \leftarrow \text{Right\_Most\_Expand}(t^i)$ 
7:      $C^{i+1} \leftarrow C^{i+1} \cup \text{expandResult}$ 
8:   end for
9:   for each pattern  $T \in C^{i+1}$  do
10:    if  $\text{supp}(T) > \sigma$  then
11:       $F^{i+1} \leftarrow F^{i+1} \cup T$ 
12:    end if
13:  end for
14:   $i \leftarrow i + 1$ 
15: until  $F^i = \emptyset$ 
16: output all frequent subtrees in  $F_k (1 \leq k < i)$  whose supporting values are larger than  $\sigma$ 

```

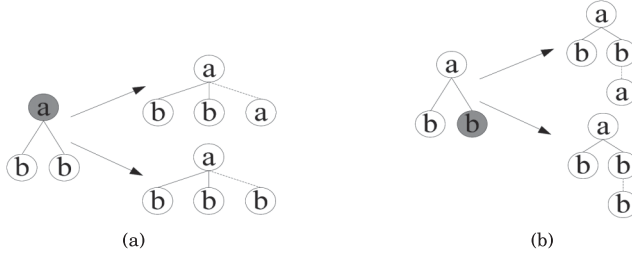


Fig. 2. Examples of the right most expansion.

subprocess for each existing frequent i -subtrees from F^i to generate the set of candidate subtrees with $i + 1$ nodes C^{i+1} (lines 5–8). If there are trees whose *supports* are larger than σ , it selects them to form F^{i+1} (lines 9–13). Such a procedure is repeated by adding 1 to i (line 14) until F^i is empty. Finally, it outputs all the subtrees whose *supports* are larger than σ from F^1 to F^{i-1} (line 16). Specifically, the *support* value of a candidate subtree T_s expanded from T_x is incrementally calculated based on $\text{supp}(T_x)$, which can significantly reduce the computational complexity.

A key concept in the previous algorithm is the rightmost expansion. Particularly, a subprocess namely *Right_Most_Expand* is invoked in line 6 to get the expanding result of t^i , which is used to grow a tree by attaching new nodes only on its rightmost branch, as shown in Algorithm 3. It first gets the right most branch of t^k (line 1), and then for each node r in the branch a new node n with different labels is iteratively added as its right most child (lines 2–5). In line 6, a newly generated pattern c^{k+1} is added to S^{k+1} (i.e., all the expansion results of t^k). The generated candidate patterns will be returned in line 9.

ALGORITHM 3: *Right_Most_Expand*(t^k)

Input:

— a tree pattern of k nodes, t^k

Output:

— candidate frequent subtree patterns, each including $k+1$ node, S^{k+1}

```

1:  $rmb \leftarrow$  the right most branch of  $t^k$ 
2: for each node  $r \in rmb$  do
3:   for each  $l \in L$  do
4:     create a new node  $n$  labeled as  $l$ 
5:      $c^{k+1} \leftarrow$  add  $n$  as the right most child of  $r$ 
6:      $S^{k+1} \leftarrow S^{k+1} \cup c^{k+1}$ 
7:   end for
8: end for
9: return all the generated candidate patterns  $S^{k+1}$ 

```

We define the number of labels in L as $|L|$ and depth of the right most branch of T as $|D|$, then according to Algorithm 3 there would be $|L| \times |D|$ right most expansion results for T . An example is given in Figure 2 to illustrate the expansion process, where $L = \{a, b\}$. As there are two nodes on the right most branch, that is, the two gray nodes in Figure 2(a) and 2(b), we can obtain four different expansion results. Particularly, in Figure 2(a) two new nodes labeled as a and b are added as the right most child of the gray node a , and another two nodes are added as the right most child of the gray node b in Figure 2(b).

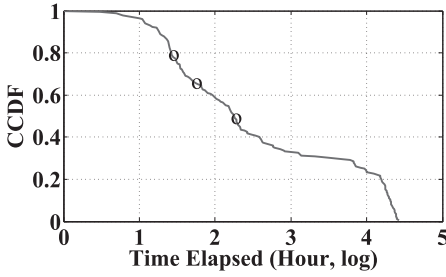


Fig. 3. CCDF of the elapsed propagation time.

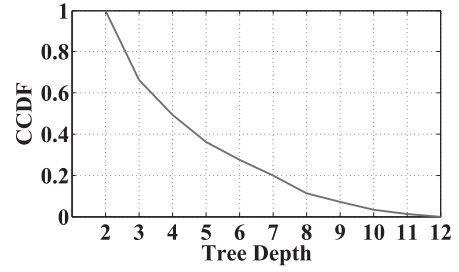


Fig. 4. CCDF of the tree depth.

6. EXPERIMENTAL RESULTS AND FINDINGS

In this section, we first describe the dataset used in our experiment, and then present the experimental results and findings.

6.1. Sina Microblog Dataset

6.1.1. Data Collection. To trace the dissemination of tweets, we implemented a crawler tool and collected 241 message cascades based on Sina Weibo open APIs⁴. As some of these messages do not have any retweets, we select 172 messages which got more than 10 replies. These messages cover a variety of topics, containing 749,384 retweets and 656,903 users. We also collected the retweet list of these messages and the corresponding follower/followee relationships of related Weibo users. To ensure that our social graph contains most of the follower/followee relationships, we collected the followee IDs of the users who are involved in the retweeting process. For each cascade, the data fields we fetched mainly include:

- Tweet_ID: The ID of a tweet.
- User_ID: The ID of a user.
- Followees_ID: The followees ID of a user.
- Number_of_Followers: The number of followers for a user.
- Content: The content of a message.
- Timestamp: The time when the message posted.
- Retweet_List: List of retweets in chronological order, each retweet includes a pointer to the original tweet.
- Retweet_Count: Number of retweets for a message.

6.1.2. Statistics of the Propagation Time and Depth. We explored the complementary cumulative distribution function (CCDF) of the propagation time (i.e., the *timeLength* feature) and tree depth (i.e., the *depth* feature) in this section. Specifically, propagation time is an important factor that might highly correlate with the success of the message propagation. Figure 3 shows the distribution of the propagation time elapsed (hours) of tweets within the dataset we collected. While the longest propagation lasted for more than three years, about 50% messages rapidly decayed within the first week, 66% in the first two days and 80% in the first day (referring the circles).

Afterwards, we analyze the distribution of the depth of the propagation trees. As shown in Figure 4, we find that the tree depth varies from 2 to 11, and the number of propagation trees decrease gradually as the depth increases, for example, there are about 34% propagation trees of depth 2 and only 3% of depth 11. We also observe that a deeper propagation tree always spreads more widely in the Weibo sphere and

⁴<http://open.weibo.com/wiki/2/>.

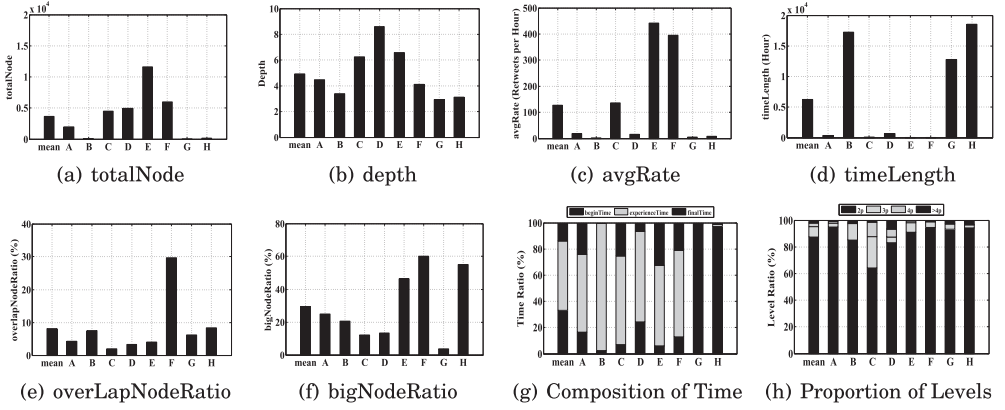


Fig. 5. Characteristics of the 8 Clusters based on the Selected Features.

retweeted by more users. In other words, there is a high correlation between the depth of a propagation tree and its total number of retweets (with a correlation coefficient of 0.81).

6.2. Results of Macro Propagation Pattern Extraction

In Section 4, we obtained eight significant clusters of macro propagation patterns. A thorough analysis of these clusters will be given in this section, followed by detailed interpretation of the characteristics of two significant propagation patterns: popular message propagation and artificial propagation.

6.2.1. Experiment Setup and Results. For brevity, we name the obtained eight clusters as Cluster A, B, C, D, E, F, G, and H, respectively. As the results of factor analysis cannot be easily interpreted, we select 13 features to analyze the clusters with coefficients larger than 0.5. According to the result shown in Figure 5, these clusters have very distinct patterns regarding to the selected features.

Specifically, Cluster A includes around 30% of all messages, and its characteristic is similar to the overall one (i.e., *mean*). Cluster B has the smallest *beginTime* and the biggest *experienceTime*, which means that the messages take a short time to be retweeted by a certain amount of users and a long time to spread. Cluster C has the smallest *level2p*, the biggest *level3p*, and *level4p* than other clusters. Cluster D has the biggest depth, sum of *level5p* and *levelDeep*, which means that the original messages spread far away from the originator. Cluster E reaches the largest volume of audiences, the maximal *avgRate*, and the shortest *timeLength*, indicating that the messages are very time-sensitive. By investigating the content of these messages, we find that they are mainly time-restricted messages published by celebrities who have a great number of followers. A further study of this cluster, that is, popular message propagation, will be presented in subsection 6.2.2. Cluster F has the largest value of *bigNodeRatio*, suggesting that users contribute more evenly to the messages compared to other clusters. Meanwhile, its *overlapNodeRatio* is especially bigger than other clusters, which indicates that there exist many users who repeatedly retweet the same message. This could be explained as the artificial inflation and a further study is presented in subsection 6.2.3. Cluster G has the smallest *totalNode* and *bigNodeRatio* while its *beginTimeRatio* is close to 100%, which indicates that the messages take a particularly long time to be retweeted by a certain number of users. These information cascades are usually conversions between a clique. While cluster H has relatively small *totalNode* and *depth*, its *timeLength* is the longest, indicating the corresponding

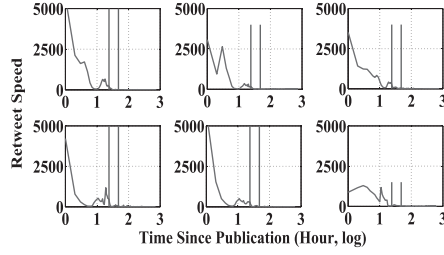


Fig. 6. Distribution of the retweet speed of six example tweets.

messages have higher vitality. Meanwhile, it has the second biggest *bigNodesRatio*, which suggests that the propagation is limited to a small scale. By examining the contents, we find that most of these messages are proverbs or personal experiences (e.g., comments to a famous sight). Therefore, there would still be users retweeting such messages that were published a long time ago.

6.2.2. Patterns of Popular Message Propagation. Among all the eight significant clusters extracted in Section 4, cascades in cluster *E* have the largest average number of audiences, where each cascade has been retweeted more than 10 thousand times on the average and could be regarded as a popular message. To obtain the properties of these popular messages, we select the top six propagation trees and analyze them from three aspects, that is, retweet speed, retweet time delay, and celebrity effect.

Retweet Speed. We aim to investigate the process of persistence and decay for the trending messages on Sina Weibo and measure the retweet speed using the number of retweets that each message gets in each time interval (1 hour). We sum up the number of retweets over time, and obtain the cumulative retweet amounts $N_w(t_i)$ for tweet w at any time frame t_i , which is given as:

$$N_w(t_i) = \sum_{\tau=1}^i n_w(t_\tau), \quad (2)$$

where $n_w(t_\tau)$ is the number of retweets of tweet w in time interval t_τ . The ratio for tweet w between time t_i and t_j is defined as:

$$C_w(t_i, t_j) = \frac{N_w(t_j) - N_w(t_i)}{t_j - t_i}, \quad (3)$$

where t_j is one hour later than t_i .

Figure 6 shows the distribution of $C_w(t_i, t_j)$ where the two vertical lines correspond to the 24th and 48th hour of each tweet propagation respectively. The results suggest that the retweet speed reached the maximum at the first hour and then decayed with some ups and downs, except for the last one, whose retweet speed increased to the top at first and then decayed. It is obvious that after 48 hours the retweet speed of all these six messages became quite low and did not get many retweets till the end. Therefore, if a tweet is not retweeted anymore, it will be pushed out of the first page of the user's timeline and become forgotten. However, some tweets may occasionally be retweeted after a long period of time due to spamming, robots, or other reasons in real life microblogging services. Thereby, we further investigate the content of such tweets and find that they are either time-restricted messages or related to instant events (e.g., the Birthday wishes). Meanwhile, according to the earlier analysis, message cascades in Cluster *E* have the smallest *timeLength*. Similarly, Yu et al. [2012a] also discovered that about 59% popular tweets (whose number of retweets is larger than 1000) have

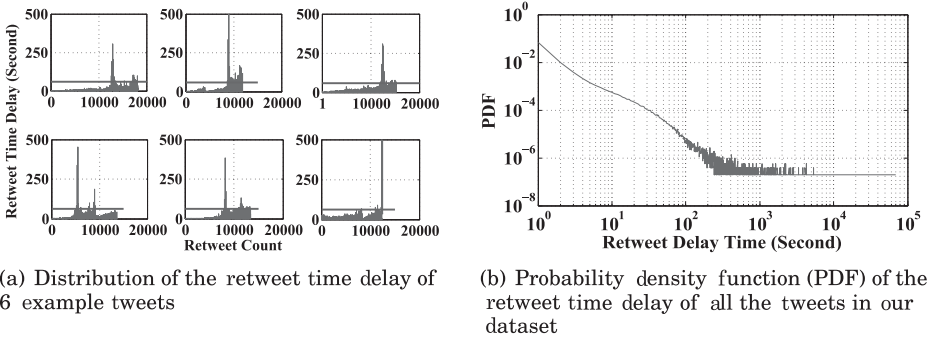


Fig. 7. Distribution of the retweet time delay.

completed 80% of their retweet within 24 hours, while 76% have completed such a percentage within 48 hours.

Retweet Time Delay. We conduct experiments to investigate the retweet time delay of each pair of consecutive retweeting behaviors in the first 24 hours for the six tweets as shown in Figure 7(a), in which the horizontal lines correspond to the 1 minute limit (i.e., 60 seconds).

Although most of the time delays between consecutive retweets are within 1 minute, there do exist some long delays (i.e., the peak value in Figure 7(a)). Specifically, according to the study of Barabasi, deliberate human activities inherently follow the heavy tailed distribution, rather than Poisson statistics. While a Poisson distribution decreases exponentially, forcing the consecutive events to follow each other at relatively regular time intervals and forbidding very long waiting times, the slowly decaying heavy-tailed processes allow for very long periods of inactivity that separate bursts of intensive activity [Barabasi 2005]. We present the probability density function (PDF) of the retweet time delay of all the tweets in our dataset in Figure 7(b), which illustrates that the time delays between consecutive retweets follow approximately a power law statistics, which is one of typical characteristics of heavy-tailed distributions. However, as shown in Figure 7(a), the heavy tails of Weibo messages intensively appeared in the final phase of retweeting, which differs from other online human behaviors [Barabasi 2005; Zhang et al. 2011].

Specifically, by taking a closer look at these long delays, we find that they correspond to intervals of midnight or daybreak when users hardly retweet messages compared with the daytime.

Celebrity Effect. It is widely accepted that the celebrity effect does exist and acts as one of the major types of influence in the social network [Kwak et al. 2010]. For example, Yin et al. [2012] studied the patterns of advertisement propagation in Sina Weibo, and discovered that celebrities have significant influence on people's decision making which is widely leveraged in advertising. In the Sina Weibo sphere, the term *celebrity effect* means a message gets more retweets if its originator is a user with a large number of followers. We find that the celebrity effect works in some of the extracted clusters. For example, the correlations between *totalNode* and the proportion of users who have more than 10,000 fans are 0.753 and 0.67 for clusters *E* and *G* respectively, while it is not so obvious for the other clusters. Specifically, for cluster *E*, we select users in the message cascades who have been retweeted more than once and analyze the relationship between the retweeter's number of followers and the resulting number of retweets. As shown in Figure 8, although the retweet amounts of most users are less than 500, there do exist some discrete points corresponding to users who have more than 500 retweets and 1 million followers. These users are the influential users

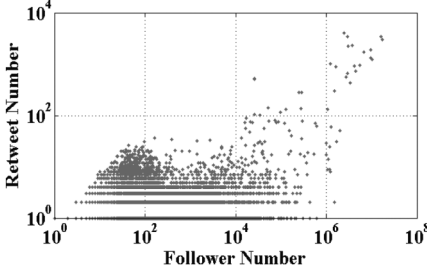


Fig. 8. Follower number versus retweet number.

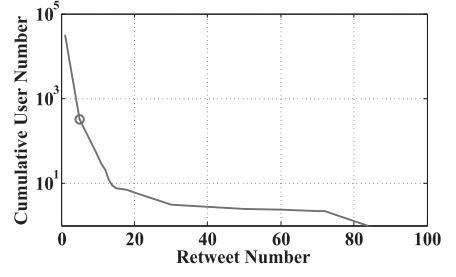


Fig. 9. Retweet number versus cumulative user number.

for the propagation process, who play important roles for the popularity of the original message, reflecting the celebrity effect.

While the celebrity effect is relatively significant in the Sina Weibo sphere, the propagation of Twitter messages is governed by a much larger set of active users, and each of them has relatively less influence [Yang and Leskovec 2010]. Moreover, in the Twitter sphere, users with the most followers are not the most influential ones in propagating messages.

6.2.3. Patterns of Artificial Propagation. For all the clusters obtained based on macro propagation pattern extraction, we notice that the *overlapNodeRatio* of cluster F is significantly bigger than that of others, which might be due to continuous retweeting of the same tweet by some fraudulent accounts. Yu et al. [2012a] gave the definition of spamming accounts as ones that are set up for the purpose of repeatedly retweeting certain messages, thus giving these messages artificially inflated popularity. According to the definition, users who retweet abnormally large amounts are more likely to be spam accounts. By examining each individual's retweet amounts to messages in cluster F , we find that there are a number of users who retweeted the same message continuously in a very short time interval.

Particularly, Figure 9 illustrates the relationship between the number of retweets and the cumulative number of users for messages in cluster F . We observe that there are totally 670 users who retweet more than five times of the same message. To make a further study, we pick up a message cascade which belongs to cluster F and is posted by a public content sharing account. We find that the maximum retweet frequency of a single account to this tweet reaches 24 times per minute, and there are 2.5% users who retweet more than five times and produce 10.3% of the total retweets. It is suggested that although this message had been retweeted for about 6,000 times, it was actually due to the artificially inflation of fake users.

Meanwhile, according to Figure 10, we find that the distribution of this message's retweet speed is quite different from that of popular messages. Specifically, there are four spikes in the distribution which are caused by the consecutive and instantaneous retweets of the fraudulent accounts, and there also is a long time period between the 1st and 2nd spikes when the retweet speed turns down to zero, which is quite abnormal.

However, if there is an account that retweets the same message continuously for multiple times, it does not necessarily mean that it is a spamming account. For example, in cluster F , we find a tweet posted by a verified account (a famous online video sites named *Tudou* in China) whose maximum number of retweets from a single account reaches 84 times. Although there are many users repeatedly retweeting this message, it is different from the pattern of artificial propagation. With the topic of retweeting for prize, this tweet was posted at the beginning of the national day of China when

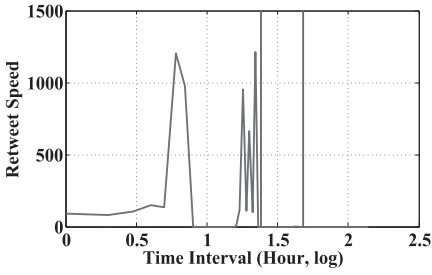


Fig. 10. Distribution of the retweet speed.

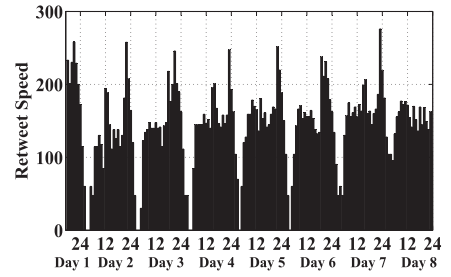


Fig. 11. Retweet speed of a prizing message.

Table IV. TreeSet Constitution

TreeSet	Number of Trees	Number of Nodes
C_E	31	360540
C_F	8	47794
C_Total	172	1022205

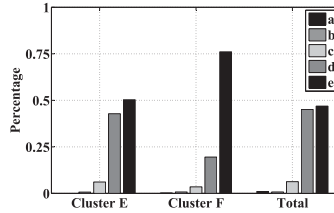


Fig. 12. Distribution of different kind of nodes in the TreeSets.

people have a seven-day holiday. In particular, Figure 11 shows the hourly retweet amounts after the release of the message. We find that the retweet speed follows a 8-hour periodic variation, and its value at the afternoon or evening is always higher than that of other times. Specifically, the highest retweet speed tends to appear around 10:00 p.m., when users are most likely to be surfing the Internet. Thereby, the popularity of this message is mainly due to temporal social events.

6.3. Results of Micro Propagation Pattern Mining

Based on the pattern mining algorithm presented in Section 5, we study the information flow patterns of message cascades in this section. Specifically, we will first give the experimental results of micro propagation pattern mining, and then discuss some interesting and distinct information flow patterns obtained based on the results.

6.3.1. Experiment Setup and Results. According to the results of macro propagation pattern extraction presented in Section 6.2, clusters E and F have very distinct features. Therefore, a further study on the information flow patterns of these two clusters is conducted from the micro perspective. To this end, we first introduce three different tree sets, as shown in Table IV.

Since nodes in the tree set are labeled, the label distribution is then analyzed. According to Figure 12, we can see that nodes with labels d and e account for a greater proportion than the others for all these three sets, which indicates that most of the users in our dataset have a small number of followers.

Based on the *FPSPM* algorithm proposed in Section 5, we discover the frequent patterns of C_E , C_F , and C_Total (consisting of all the clusters), respectively. As shown

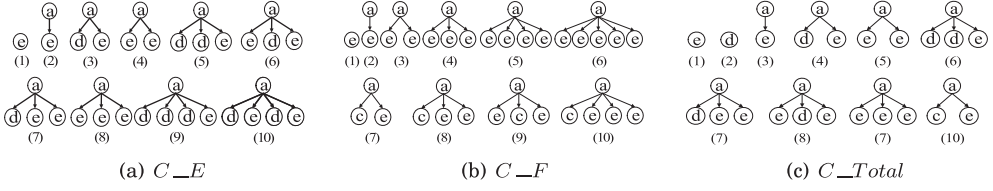


Fig. 13. Top 10 frequent patterns for the three TreeSets.

in Figure 13, we select the top 10 frequent patterns, which are ranked by their support values. All the patterns present the most frequent paths between different types of users for information propagation in the Weibo sphere. By comparing the patterns of different tree sets, we find that messages tend to propagate in certain paths.

In Figure 13, the most frequent pattern for all sets is an isolated node e . It indicates that most Sina Weibo users who express their attitudes by retweeting have very few followers and seldom get responses from their followers. Therefore, most of the information propagation processes are interrupted and sliced. The most frequent pattern with two nodes is $a \rightarrow e$, which means that a user labeled as a is retweeted by another user labeled as e . The pattern shown in Figure 13(a)–(3) and 13(c)–(4) indicates that one user labeled as a is retweeted by two users who are labeled as d and e , and the left node d is prior to the right node e according to the retweeting time. All these frequent patterns of information flows are in trivial propagating ways which is the nature of online social network.

6.3.2. Patterns of Information Flows. By comparing the generated flow patterns of different TreeSets, it can be observed that messages tend to propagate in certain paths. While some of them are similar to traditional media such as newspapers, the others are inherited in the online social network structure.

- Information tends to propagate from advanced-users to low-level users.* It indicates that the advanced-users get a lot of retweets and promote the information spreading to a deeper level in the Weibo sphere. For example, information flows of all frequent patterns start with users labeled as a .
- There is a polarization phenomenon among users in the Weibo sphere.* On one hand, all advanced-users are labeled as a while label b does not appear in any of the extracted patterns. On the other hand, the most frequent low-level users are e , d , and c , respectively. In other words, some of the information flows are less frequent, for example, $a \rightarrow b$ and $b \rightarrow e/d/c$. Therefore, the polarization phenomenon is quite significant among Weibo users. Specifically, while both the high-level and low-level users play important roles in information propagation, the middle-level users (i.e., those who are labeled as b and c) are less determinative for the popularity of Weibo messages.
- Frequent propagation patterns tend to be wider rather than deeper.* It means that the probability of retweeting gets lower when the information spreads to a deeper level. For instance, all the patterns in Figure 13 are trees of depth two, and as the number of nodes increases these trees tend to become wider rather than deeper. A similar observation had been obtained in Leskovec et al. [2007b], where the authors found that information cascades tend to be wide, and not too deep.
- Frequent information flows tend to take certain shapes preferentially.* For instance, the top six information flows of C_F are all ended with nodes labeled as e (i.e., users with followers between 1 and 100). This is in accordance with the artificial propagation phenomenon, indicating that there are a number of fraudulent accounts who have very few followers and retweet repeatedly. Meanwhile, we also notice that

the cascade frequency rank does not simply decrease as a function of the cascade size. For example, a six-node pattern (i.e., the 6th subtree in Figure 13(b)) is more common than a pattern of three nodes (i.e., the 7th subtree in Figure 13(b)). Leskovec et al. [2007b] had acquired similar findings from the traditional blog networks.

7. APPLICATIONS

Several potential applications could be developed based on the discovered information propagation patterns.

7.1. Trending Prediction

The information diffusion patterns discovered can help us predict the future trends of online content. Bao et al. [2013] claimed that better prediction performance can be obtained by leveraging structural characteristics of the microblog network. Specifically, the authors found that a low link density and a deep diffusion usually lead to wide spreading, that is, a diverse group of individuals is able to disseminate a message to wider audience than a dense group. For example, with a better understanding of the discovered diffusion patterns, we can develop more fine grained prediction models (e.g., retweeting behavior prediction) to estimate the propagation of crucial messages [Lu et al. 2014; Yang et al. 2010].

7.2. Role Recognition

As social media has revolutionized the nature of influence and the role of influential people, the extracted patterns can also provide evidence to support the influence topology theory. As shown in Figure 8, most of the users in Sina Weibo have a small number of followers and tend to follow and retweet influential individuals with a large number of followers, such as movie stars and grassroots celebrities. Users with fewer followers are more likely to be information viewers while users with a lot of followers tend to be idea starters or information amplifiers. For example, based on Edelman's topology of influence (*TOI*) [Bentwood 2007], Tinati et al. [2012] developed a conversational model which enables topic analysis and role identification of different users.

7.3. Cognitive Analysis

Another potential application is to use the extracted patterns for cognitive analysis. For example, the information flow patterns are extremely useful for interpreting the information dissemination processes in the Weibo sphere. Specifically, messages about a certain popular event might be retweeted, distributed, and massively followed within a very short time, and then spark off large-scale discussions, which even affect people's daily lives [Leskovec et al. 2009]. Based on the obtained propagation patterns, researchers from cognitive science could study the psychology and behavior pattern (e.g., retweeting behaviors) of microblog users. For example, more and more companies choose to publish and advertise their new products using microblogging services. However, while some of the advertisements are successful, others may fail to spark off attentions from a large amounts of users. By analyzing the literal content together with the propagation patterns of different advertising messages, cognitive science researchers are able to study what kind of factors are more likely to affect people's psychology states and further determine the popularity of a microblog message (i.e., product).

8. CONCLUSION

In this article, we studied the information propagation patterns in microblogging services based on the data collected from Sina Weibo. By representing the cascade of microblog messages as trees according to the retweeting process, we explored the

information propagation patterns from both macro level and micro level. On one hand, the macro propagation patterns refer to general propagation modes, and eight significant clusters of message cascades were extracted by grouping the constructed propagation trees based on hierarchical clustering. On the other hand, the micro propagation patterns are frequent information flows, and a set of frequent propagation patterns were discovered using the proposed tree-based mining algorithm. Experimental results demonstrated that several interesting patterns were revealed, such as popular message propagation, artificial propagation, and typical information flows between different types of users.

The preliminary research can be used to facilitate various studies on social network modeling and analysis, and thus suggests several interesting problems that are worth further exploring. On one hand, we can utilize the propagation patterns obtained in this article to facilitate the study of information diffusion. For example, the extracted patterns can help construct more accurate and efficient trending prediction models, which can be used to predict and even impact the future trends of microblog messages. Moreover, we can also investigate the relationship between the information propagation patterns and the sentiments embodied in microblogs, that is, revealing whether sentiments would impact the diffusion process of messages. On the other hand, the proposed methods and obtained findings can also be used to facilitate the study of several fundamental issues in social network analysis, such as key node/user identification [Kimura et al. 2007] and community detection [Wang et al. 2014b]. Specifically, by analyzing the roles that users with different labels have played during the diffusion of messages, key users who dominate the propagation of information can be identified. Meanwhile, once we get the community structure of the microblog network, the proposed approach in this article can be used to study how information propagates within individual community and among different communities.

REFERENCES

- Tatsuya Asai, Kenji Abe, Shinji Kawasoe, Hiroki Arimura, Hiroshi Sakamoto, and Setsuo Arikawa. 2002. Efficient substructure discovery from large semi-structured data. In *Proceedings of SDM'02*, Robert L. Grossman, Jiawei Han, Vipin Kumar, Heikki Mannila, and Rajeev Motwani (Eds.). SIAM, Philadelphia, PA, 158–174.
- Seema Bandyopadhyay and Edward J. Coyle. 2003. An energy efficient hierarchical clustering algorithm for wireless sensor networks. In *Proceedings of INFOCOM'03*. IEEE, San Francisco California, USA, 1713–1723.
- Peng Bao, Hua-Wei Shen, Junming Huang, and Xue-Qi Cheng. 2013. Popularity prediction in microblogging network: A case study on sina weibo. In *Proceedings of WWW'13*. ACM, New York, NY, USA, 177–178.
- Albert-Laszlo Barabasi. 2005. The origin of bursts and heavy tails in human dynamics. *Nature* 435, 207–211.
- Jonny Bentwood. 2007. Distributed Influence: Quantifying the impact of Social Media. Edelman White Paper.
- Danah Boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, Tweet, Retweet: Conversational aspects of retweeting on Twitter. In *Proceedings of HICSS'10*. IEEE Computer Society, Washington, DC, USA, 1–10.
- Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. 2010. Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of ICWSM'10*. AAAI Press, Washington, DC, USA.
- Deepayan Chakrabarti and Christos Faloutsos. 2006. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.* 38, 1.
- Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. 2010. Short and Tweet: Experiments on recommending content from information streams. In *Proceedings of CHI'10*. ACM, New York, NY, USA, 1185–1194.
- Shaoyong Chen, Huanming Zhang, Min Lin, and Shuanghuan Lv. 2011. Comparison of microblogging service between Sina Weibo and Twitter. In *Proceedings of ICCSNT'11*. IEEE, Washington, DC, USA, 2259–2263.
- Justin Cheng, Lada Adamic, P. Alex Dow, Jon M. Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted?. In *Proceedings of WWW'14*. ACM New York, NY, USA, 925–936.

- Marc Cheong and Vincent Lee. 2010. A study on detecting patterns in Twitter intra-topic user and message clustering. In *Proceedings of ICPR'10*. IEEE Computer Society, Washington, DC, USA, 3125–3128.
- Milad Eftekhari, Yashar Ganjali, and Nick Koudas. 2013. Information cascade at group scale. In *Proceedings of KDD'13*. ACM New York, NY, USA, 401–409.
- Pengyi Fan, Pei Li, Zhihong Jiang, Wei Li, and Hui Wang. 2011. Measurement and analysis of topology and information propagation on Sina-Microblog. In *Proceedings of ISP'11*. IEEE, Washington, DC, USA, 396–401.
- Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. 2010. Out-tweeting the Twitterers - Predicting information cascades in microblogs. In *Proceedings of WOSN'10*. USENIX Association, Berkeley, CA, USA, 3–3.
- Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. 2009. Social networks that matter: Twitter under the microscope. *First Monday* 14, 1 (2009), Article 8.
- Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. Twitter Power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.* 60, 11 (Nov. 2009), 2169–2188.
- Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why we twitter: Understanding microblogging usage and communities. In *Proceedings of WebKDD/SNA-KDD'07*. ACM, New York, NY, USA, 56–65.
- Masahiro Kimura, Kazumi Saito, and Ryohei Nakano. 2007. Extracting influential nodes for information diffusion on a social network. In *Proceedings of AAAI'07*. AAAI Press, Washington, DC, USA, 1371–1376.
- Akshay Krishnamurthy, Sivaraman Balakrishnan, Min Xu, and Aarti Singh. 2012. Efficient active algorithms for hierarchical clustering. In *Proceedings of ICML'12*. 473–480.
- Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of WWW'10*. ACM, New York, NY, USA, 591–600.
- Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of KDD'09*. ACM, New York, NY, USA, 497–506.
- Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. 2007a. Cascading behavior in large blog graphs: Patterns and a model. In *Proceedings of SDM'07*. SIAM, Philadelphia, PA, USA, 551–556.
- Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. 2007b. Cascading behavior in large blog graphs: Patterns and a model. In *Proceedings of SDM'07*. SIAM, Philadelphia, PA, USA.
- Shuyang Lin, Fengjiao Wang, Qingbo Hu, and Philip S. Yu. 2013. Extracting social events for learning better information diffusion models. In *Proceedings of KDD'13*. ACM New York, NY, USA, 365–373.
- Xinjiang Lu, Zhiwen Yu, Bin Guo, and Xingshe Zhou. 2014. Predicting the content dissemination trends by repost behavior modeling in mobile social networks. *Journal of Network and Computer Applications* 42, 197–207.
- Haixin Ma, Weining Qian, Fan Xia, Xiaofeng He, Jun Xu, and Aoying Zhou. 2013. Towards modeling popularity of microblogs. *Front. Comput. Sci.* 7, 2 (April 2013), 171–184.
- Yan Qu, Chen Huang, Pengyi Zhang, and Jun Zhang. 2011. Microblogging after a major disaster in China: A case study of the 2010 Yushu Earthquake. In *Proceedings of CSCW'11*. ACM, New York, NY, USA, 25–34.
- Daniel Ramage, Susan T. Dumais, and Daniel J. Liebling. 2010. Characterizing microblogs with topic models. In *Proceedings of ICWSM'10*. The AAAI Press.
- Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. 2011. Influence and passivity in social media. In *Proceedings of WWW'11*. ACM, New York, NY, USA, 113–114.
- Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. 2010. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *Proceedings of SocialCom'10*. IEEE Computer Society, Washington, DC, USA, 177–184.
- Ramine Tinati, Leslie Carr, Wendy Hall, and Jonny Bentwood. 2012. Identifying communicator roles in Twitter. In *Proceedings of WWW'12*. ACM, New York, NY, USA, 1161–1168.
- Chenxu Wang, Xiaohong Guan, Tao Qin, and Wei Li. 2012. Who are active? An in-depth measurement on user activity characteristics in sina microblogging. In *Proceedings of GLOBECOM'12*. IEEE, Washington, DC, USA, 2083–2088.
- Senzhang Wang, Xia Hu, Philip S. Yu, and Zhoujun Li. 2014a. MMRate: Inferring multi-aspect diffusion networks with multi-pattern cascades. In *Proceedings of KDD'14*. ACM New York, NY, USA, 1246–1255.
- Wenhao Wang and Bin Wu. 2011. Comparing Twitter and chinese native microblog. In *Proceedings of EWT'11*. IEEE, Washington, DC, USA, 1–4.

- Zhu Wang, Daqing Zhang, Xingshe Zhou, Dingqi Yang, Zhiyong Yu, and Zhiwen Yu. 2014b. Discovering and profiling overlapping communities in location-based social networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44, 4, 499–509.
- Ye Wu and Fuji Ren. 2011. Learning sentimental influence in Twitter. In *Proceedings of ICFCSA'11*. IEEE, Washington, DC, USA, 119–122.
- Jiang Yang and Scott Counts. 2010. Predicting the speed, scale, and range of information diffusion in Twitter. In *Proceedings of ICWSM'10*. AAAI Press, Washington, DC, USA, 355–358.
- Jaewon Yang and Jure Leskovec. 2010. Modeling information diffusion in implicit networks. In *Proceedings of ICDM'10*. IEEE, Washington, DC, USA, 599–608.
- Zi Yang, Jingyi Guo, Keke Cai, Jie Tang, Juanzi Li, Li Zhang, and Zhong Su. 2010. Understanding retweeting behaviors in social networks. In *Proceedings of CIKM'10*. ACM, New York, NY, USA, 1633–1636.
- Zibin Yin, Ya Zhang, Weiyuan Chen, and Richard Zong. 2012. Discovering patterns of advertisement propagation in sina-microblog. In *Proceedings of ADKDD'12*. ACM, New York, NY, USA, Article 1, 9.
- Louis Lei Yu, Sitaram Asur, and Bernardo A. Huberman. 2011. What trends in chinese social media. In *Proceedings of SNA-KDD Workshop'11*. ACM, New York, NY, USA.
- Louis Lei Yu, Sitaram Asur, and Bernardo A. Huberman. 2012a. Artificial inflation: The real story of trends and trend-setters in sina weibo. In *Proceedings of SocialCom'12*. IEEE, Washington, DC, USA, 514–519.
- Zhiwen Yu, Zhiyong Yu, Xingshe Zhou, Christian Becker, and Yuichi Nakamura. 2012b. Tree-based mining for discovering patterns of human interaction in meetings. *IEEE Trans. on Knowl. and Data Eng.* 24, 4, 759–768.
- Dan Zarrella. 2009. The Science of Retweets. <http://danzarrella.com/thescience-of-retweets-report.html>.
- Daqing Zhang, Bin Guo, and Zhiwen Yu. 2011. The emergence of social and community intelligence. *Computer* 44, 7 (July 2011), 21–28.
- Dan Zhang, Yan Liu, Richard D. Lawrence, and Vijil Chenthamarakshan. 2010. ALPOS: A machine learning approach for analyzing microblogging data. In *Proceedings of ICDM'10 Workshop*. IEEE Computer Society, Washington, DC, USA, 1265–1272.
- Hongbo Zhang, Qun Zhao, Hongyan Liu, Ke xiao, Jun He, Xiaoyong Du, and Hong Chen. 2012. Predicting retweet behavior in weibo social network. In *Proceedings of WISE'12*. Springer-Verlag, Berlin, Heidelberg, 737–743.
- Dejin Zhao and Mary Beth Rosson. 2009. How and why people Twitter: The role that micro-blogging plays in informal communication at work. In *Proceedings of GROUP'09*. ACM, New York, NY, USA, 243–252.
- Zicong Zhou, Roja Bandari, Joseph Kong, Hai Qian, and Vwani Roychowdhury. 2010. Information resonance on Twitter: Watching Iran. In *Proceedings of SOMA'10*. ACM, New York, NY, USA, 123–131.

Received June 2014; revised November 2014; accepted March 2015