Discovering Patterns of Advertisement Propagation in Sina-Microblog

Zibin Yin
Shanghai Jiao Tong University
Shanghai, China
zibinyin@sjtu.edu.cn

Weiyuan Chen
Shanghai Jiao Tong University
Shanghai, China
cwyalpha@sjtu.edu.cn

Ya Zhang*
Shanghai Jiao Tong University
Shanghai, China
ya_zhang@sjtu.edu.cn

Richard Zong
SINA Corporation
Beijing, China
ruixing@staff.sina.com.cn

ABSTRACT

The explosive growth of microblogs has attracted many corporations and organizations. Microblogging has been considered as a high-quality advertising platform. In this study, we attempt to reveal the patterns of advertisement propagation in Sina-Microblog through analyzing a selected set of message cascades. Each message cascade is represented by a propagation tree and 33 features were extracted, which cover mainly three aspects of a cascade: the volume of the participants, the topology of the propagation paths, and the promptness of the propagation in term of time. To reveal the propagation patterns, We then group these message cascades using K-means clustering algorithm. Analysis of the resulted clusters reveals the patterns of advertisement propagation, based on which we further propose several metrics to measure the effectiveness of advertisement in microblogs.

1. INTRODUCTION

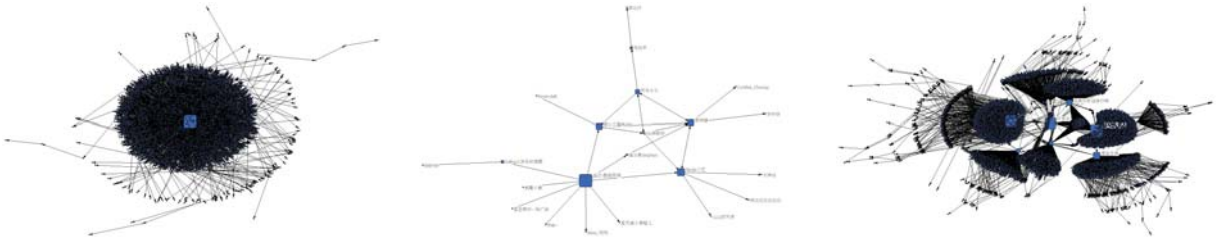
Nowadays people are increasingly engaged in publishing, sharing and commenting on information through Microblogging services. Several popular microblogging sites have grown dramatically in terms of their user volume since their emergence. Sina-Microblog, the most popular Chinese microblogging site that went online in fall 2009, has attracted more than 200 million registered users in two years. Microblogging services enable individuals and corporations to act like media and publish their status and thoughts in real time. With its significant user volume, microblogging has now been recognized as one of the most important types of media. For example, a Sina-Microblog user named Yao Chen (姚晨), a famous Chinese actress, has more than 19 million followers up to now. Yao's followers in fact outnumber the audience of most newspapers or web sites by several magnitudes. Many organizations and individuals have attempted to market themselves through microblogging, which is considered as an important advertising platform. To be successful in marketing with microblogging, it is important to understand how advertisements propagate and why some advertisements reach more users in the network than others.

The content of messages and the influence of authors are generally considered as two major factors impacting the information dissemination in microblogs. Previous studies revealed that the content[11, 8, 3] and the sentiment[1, 17] of messages play important roles in the propagation. Some other studies focus on scoring the user influence by analyzing message cascades or user relationships [13, 2].

The advertisement propagation in microblogosphere is very different from that of other advertising platforms (e.g. TV and radio) due to the following characteristics of microblogging services. A underlying directed social network is associated with each microblogging sites, where each user selectively follows other users and subscribes to their messages. This follower-followee relationship embeds shared interests among the users. Furthermore, any user may broadcast information to his/her followers. To some extent, advertising in microblog sites may be more targeted than other sites. In this study, we focuses on discovering the patterns of advertisements propagation. What are the propagation patterns for widely spread messages? Does 'celebrity effect' contribute to advertisements propagation? Which types of advertisements are able to attract more users? In seeking answers to the above questions, we selected three sets of advertising message cascades for our analysis: advertisements showing products, advertisements with celebrity, and advertisements through sales promotion. We also collected some other kinds of message cascades including news and informational messages as a control group. We represent each message cascade with a propagation tree. By eyeballing the propagation trees, we identified some obvious topology patterns such as the star type, the constellation type, and the nebular type(Figure 1). We attempt to group the above selected message cascades to reveal their propagation patterns. For each message cascade, 33 features are designed, mainly representing three sets of properties: the volume of participants, the topology of the propagation paths, and the promptness of the information dissimilation in terms of time.

*Author for correspondence.

ADKDD '12, August 12, Beijing, China



(a) The star type (b) The constellation type (c) The nebular type

Figure 1: The Topology Patterns of Message Propagation

These message cascades are grouped with K-means clustering methods. Analyzing the grouped message cascades reveals the patterns of advertisement propagation, based on which we further propose a set of metrics to quantitatively measure the effectiveness of advertising in microblogs.

2. RELATED WORK

Recent theoretical works provide a rich set of models that explain information flow and the social network structure. Nowell et al. [5] examined the internet chain letter on the Iraq petition. They described the dissemination tree by three metrics: the median node depth, the width, and the fraction of nodes with exactly one child. They found that the progress of these chain letters proceeds in a narrow but very deep tree-like pattern, continuing for several hundred propagation hops. Rodrigues et al. [15] compared the shapes of cascades on different content, finding that twitter cascades are wider, unlike those of the Internet chain letters. Yang et al. [18] analyzed how the retweet behavior on twitter is influenced by factors such as user, message and time. Based on their observations, they propose a factor graph model to predict users' retweeting behaviors and the range of spread for a new tweet. Ratkiewicz et al. [9] demonstrate a web service that tracks politicalmemes in Twitter. They combined sophisticated network analysis with content and time series mining to analyze why some memes go viral.

Besides the shape and structure of a cascade, time is considered as a crucial factor for message propagation. Ye et al. [14] measured message propagation by the number of hops away from the originators and the lifespan of retweets and mentions. Kwak et al. [7] examined the propagation of 106 million tweets by analyzing their popularity over time and information diffusion through retweet trees. They showed that 73% of the topics have a single active period. Another observation they made is that the distribution of the users in a retweet tree follows power-law distribution. Xie et al. [10] proposed an interest-driven model to simulate basic user communicating behaviors and processes, found that individual behaviors are burst of rapidly occurring events separated by long periods of inactivity, and the collective behaviors follow heavy-tailed power-law distribution.

Furthermore, with the use of hashtag and URL links, a lot of researchers measured the patterns of information propagation. Romero et al. [4] studied the sub-graph structure of the initial adopters for different widely-adopted hashtags, finding significant variation in the ways that hashtags on different topics spread. Galuba et al. [16] and Rodrigues et al. [15] tracked the word-of-mouth exchange of URLs among internet users. Based on the analysis in the user activity and social graph, Galuba et al. [16] proposed a propagation model that predicts user preferences on URLs.

Bakshy et al. [6] found that the widely-spread message cascades tend to involve influential users and/or those with a large number of followers. This result reflects the 'celebrity effect'. The celebrities might have posted lots of messages, but only a small portion of which receive a large number of retweets. Similar to Kwak et al. [7] and Cha et al. [12], they discovered that the number of followers is not related to the number of retweets and mentions.

3. SINA-MICROBLOG DATA

Sina-Microblog is a very popular social network site in China. It currently receives more than 75 million messages each day, which span a wide range of topics, including advertisements, personal updates, and news. We attempt to discover the propagation patterns for messages that reach a large volume of audience in Sina-Microblog.

As the majority of microblogging messages do not get any reply, retweet or comment, we manually select a list of messages which got more than 500 replies, retweets or comments in total. Besides advertisements, we intentionally collect some other kinds of messages that cover a variety of topics, including news and informational messages. In order to study the effect of different factors in message propagation, we include in our analysis messages of the same content but originated from different users. The Sina-Microblog API is used to trace the dissimilation of each message by following its replies, retweets, and comments using the tag '//@'. We collected 261 message cascades in total, containing 749,384 messages and 656,903 users. For each message, the data we fetched includes the following fields.

- User_id: The ID of a user
- User_name: The name of a user
- Content: The content of a message
- In_reply_to_user: The ID of a user that the message replied to
- Time_stamp: The time when the message posted
- Level_of_retweet: The level of cascade where the message located
- #_of_followers: The number of followers for a user

For the purpose of our analysis, we manually labeled the set of message cascades using the following six categories: news, advertisement (product), advertisement(celebrity), advertisement(sales promotion), knowledge and others. Table 1 shows the distribution of the labels.

4. REPRESENTATION OF MESSAGE CASCADE

Table 1: The Distribution of the Labels for the Set of Message Cascades

Label	# of original posts
News	73
Advertisement(product)	40
Advertisement(celebrity)	31
Advertisement(sales promotion)	23
Knowledge	15
Others	79

The propagation path of each message cascade is represented as a tree. Each node corresponds to a user. The root node of the tree represents the originator of a cascade. A directed edge from A to B means that the user B retweeted/commented the user A's tweets. A bidirectional edge means the two users retweeted/commented each other. As a user may retweet his own several times, each node is associated with a number (in brackets) which is the frequency of self-retweets. Considering two users may chat (retweet) with each other and a user may retweet/comment another user many times, we associate each edge with the frequency of retweeting and commenting between the two corresponding users. Figure 2 illustrates a propagation tree, where the node O represents the original user, n_i is the frequency of retweets and comments between the two neighboring nodes, and m_i represents the frequency of self-commenting.

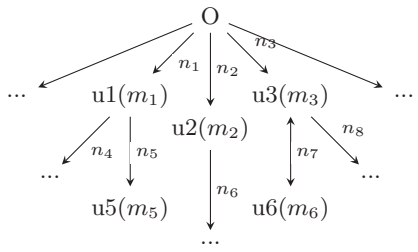
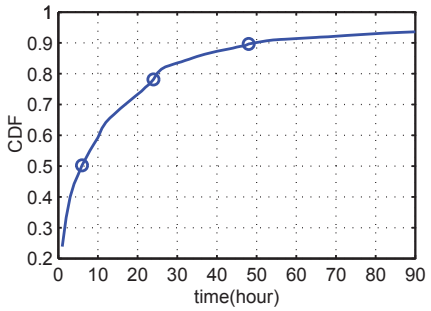


Figure 2: An Example of the Message Propagation Tree in Sina-Microblog

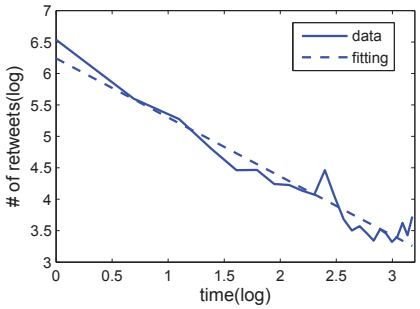
4.1 The Topology Pattern of Propagation Trees

As Figure 1 shows, three interesting topology patterns were identified when eyeballing the topology of propagation graphs. The star type is the most popular topology pattern in Sina-Microblog. Most message cascades in this category originate from users with more than 1 million followers, often representing celebrities and well-known organizations. The constellation type is relatively rare. This type usually occurs as a 'chat' among a small group of users who frequently reply to each other. Messages of this type usually do not get many retweets/comments. Because the cascades we selected have more than 500 retweets, replies and comments in total, there are only two cascades of this type in this data set. The nebular type has many retweeters at every level and is also considered as a successful propagation pattern. In our study, most of cascades demonstrating this pattern are news, especially emerging news.

In practice, there are no clear boundaries between these types. A unique characteristics for message cascades in the constellation type is that the propagation tree has more edges than nodes. The ratio of edges to nodes for the constellation type is usually larger than 2, while the ratio for the star type and the nebular type is often close to 1.



(a) Cumulative Distribution of Retweets



(b) Distribution of Retweets(log-log)

Figure 3: Timeline of Propagation

4.2 Propagation Time

In addition to the topology, propagation time is another important factor that may be highly correlated with the success of the message propagation. Figure 3a shows the distribution of the messages by propagation time (i.e. the difference in publishing time between the parent node and the child node) in the dataset we collected. According to Figure 3a, there are very few retweets/comments after 90 hours for all of the cascades. 50% messages are posted within the first 6 hours, 78% in the first day and 90% in the first two days. Figure 3b shows that the retweet distribution over time follows power-law distribution.

5. CLUSTERING THE PROPAGATION GRAPHS

We attempt to reveal the patterns of message propagation in Sina-Microblog by clustering the propagation graphs, both the shape and the propagation time are important characteristics of message propagation. We hence extract the features of propagation trees in the following three aspects: the size of propagation tree, the topology of the tree, and the distribution of messages along time.

Seven features are extracted to represent the size of propagation tree. Four features, the total number of the edges in the graph ($npEdge$), the total number of nodes ($totalNode$), the number of leaf nodes ($leafNode$), and the number of parent nodes ($parentNode$), directly reflect the size of the

propagation tree, where $parentNode$ equals to $totalNode$ minus $leafNode$. Because an edge in the graph may represent multiple interactions between two users, $weightEdge$ is proposed as the weighted summation of the number edges according to the number of interactions. Furthermore, the total number of self-retweets (i.e. one replies to his own messages) is represented by $selfRetweet$.

In terms of the topology of propagation tree, considering almost all the messages propagate within 10 hops of the originator, we extract the proportion of messages at each level ($level1p$-$level10p$), the max level ($maxLevel$) and the depth ($depth$) of the graph as features. The difference between $maxLevel$ and $depth$ is similar to that between $npEdge$ and $weightEdge$. The variable $depth$ takes into account the self-retweeting and iterative-retweeting in the graph.

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 messages, and name it $bigNode$. We first sort the contribution of each user in term of the number of messages published in a descending order, and then choose the top users that account for 70% of the messages. A larger value for $bigNode$ means that more users contribute to the message propagation. Furthermore, we calculated the ratio of $parentNode$ to $leafNode$ ($plRatio$), the ratio of $parentNode$ to $totalNode$ ($ptRatio$), and the ratio of $bigNode$ to $totalNode$ ($bigNodeRatio$).

We define the time span (hours) of a message cascade as $timeLength$, the average number of hourly message rate as $avgRt$, and the maximum number of hourly message rate as $maxRt$. The time span of a message cascade is divided into three segments and the length of the segments are denoted as $silenceTime$, $expTime$ and $followupTime$, respectively. The $silenceTime$ and $expTime$ are defined as the segments at beginning and in the end of the entire time span, for which the average hourly message rate is less than 20% of $maxRt$. The $expTime$ is the segment in between. In other words, $silenceTime$ measures the length of time span when the topic is heating up, $expTime$ is the time span when the majority of messages are posted, and $followupTime$ 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 denoted them as $silenceTimeRatio$, $expTimeRatio$ and $followupTimeRatio$ respectively.

In total, 33 features are proposed to represent each message cascade. Table 2 summarizes the definition of the features. We find that some of these features have high correlation with each other. Table 3 shows the Pearson correlation coefficient among the size-related features. These highly correlated variables may disrupt the clustering. We hence use Factor Analysis to describe the 33 variables above in terms of a lower number of factors which are uncorrelated with each other. The 33 variables are modeled as linear combinations of the factors. Choosing the factors corresponding to the eigenvalue $\lambda \geq 1$, we got 9 factors. We use K-means clustering to group the propagation trees. Euclidean distance is employed as the distance measure. We experiment with a range of values for the number of clusters and choose the one that maximizes the gain of mutual information. Figure 4 shows the mutual information gain with different number of clusters. Based on our experimental results, the number of clusters is set to be 20. By ignoring the clusters with less than 5 message cascades, we get 8 clusters. Table 4 shows the basic characteristics of each cluster, including the size of each cluster and the mean and variance of the intra-cluster distance. Setting the number of clusters to be 20, the mean inter-cluster distance of all clusters is 2.87, and the mutual information gain is 1.4527.

Table 2: The Variables to Represent a Message Cascade

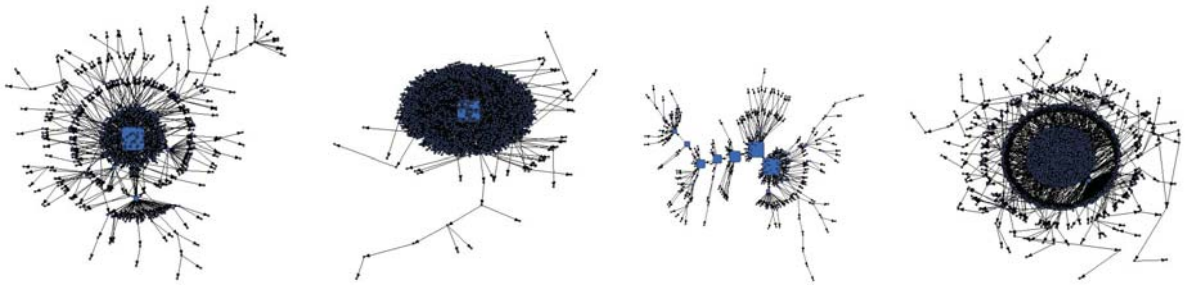
Variable	Description
totalNode	# of total nodes
leafNode	# of leaf nodes
parentNode	# of parent nodes
plRatio	the ratio of parentNode to leafNode
ptRatio	the ratio of parentNode to totalNode
bigNode	# of users who were replied more than 70% of the total retweeters
bigNodeRatio	the ratio of bigNode to totalNode
weightEdge	# of edges including duplicate edges
npEdge	# of distinct edges
nptEdgeRatio	the ratio of nptEdgeRatio to npEdge
selfRetweet	# of retweets replied by oneself
timeLength	the time span(Hour)
silenceTime	the time span(Hour) with little retweets at beginning
expTime	the time span(Hour) with many retweets per hour
followupTime	the time span(Hour) which follow expTime with little retweets
silenceTimeRatio	the ratio of silenceTimeRatio to timeLength
expTimeRatio	the ratio of expTimeRatio to timeLength
followupTimeRatio	the ratio of followupTimeRatio to timeLength
avgRt	the average retweet hour
maxRt	the max number of retweets per hour
maxLevel	the max number of level
depth	the depth of the structure
level1p - level10p	the proportions of each level(1-10) in the structure
tweetLength	the length of message

6. RESULT OF CLUSTERING

In this section, we analyze the above eight clusters in detail. For each cluster, we plot the propagation tree that is closest to the cluster centroid and use it as the representative for the corresponding cluster (Figure 5). We use the blue square nodes to represent users involved in the corresponding cascade and the size of each node is proportion to the betweenness of the corresponding user in the graph. The ID of the original posters, the number of retweets, the depth of the tree, and the proportion of messages at level-1 are marked at the bottom of each figure. We also show the topic of each cascade, but it does not necessarily mean that the cluster contains only message of this topic.

According to Figure 5(b)(d)(e)(f), the cascades are mainly contributed by the immediate neighbor (i.e. level-1 neighbor) of the original posters, accounting for more than 70% of the total messages. On the contrary, the cascade as shown in Figure 5(a)(c)(h) only have less than 50% of the messages from the level-1 neighbor. The other retweets/comments are mostly from a set of other key users.

As the factors resulted from the Factor Analysis are not

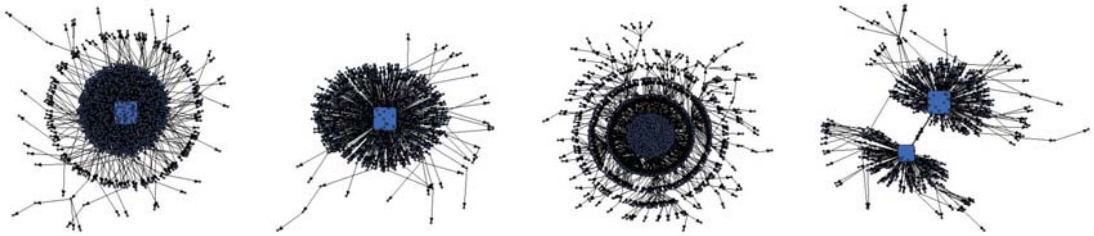


(a) Cluster A (Personal Update, PFG骑士, 994, 8, 43%)

(b) Cluster B (Advertisement about an Ice Cream brand, 可爱多, 2987, 7, 98%)

(c) Cluster C (News, 人在耶鲁, 405, 9, 27%)

(d) Cluster D (Joke, 姚晨, 9225, 9, 78%)



(e) Cluster E (Personal Update, 蔡卓妍, 1640, 5, 85%)

(f) Cluster F (Sales Promotion, 团美妆, 1224, 4, 92%)

(g) Cluster G (Scandal about celebrity, 头条新闻, 4324, 11, 61%)

(h) Cluster H (Live Commentary, 新浪视频, 639, 4, 33%)

Figure 5: Representative Propagation Trees of the Eight Selected Clusters. The blue square nodes represent the users involved in the corresponding message cascade and the size of each node is proportion to the betweenness of the corresponding user in the graph. The topic of the representative message cascade, its original poster, the number of messages, the depth of the tree, and the proportion of level-1 messages are noted in the parenthesis.

Table 3: Pearson Correlation Coefficient of the Variables about Size

weight Edge	total Node	leaf Node	parent Node	np Edge	self Retweet	max Rt
1.000	**.961**	**.963**	**.816**	**.965**	**.758**	**.795**
.961	**1.000**	**.995**	**.882**	**1.000**	**.732**	**.792**
.963	**.995**	**1.000**	**.832**	**.995**	**.724**	**.786**
.816	**.882**	**.832**	**1.000**	**.880**	**.669**	**.708**
.965	**1.000**	**.995**	**.880**	**1.000**	**.742**	**.794**
.758	**.732**	**.724**	**.669**	**.742**	**1.000**	**.633**
.795	**.792**	**.786**	**.708**	**.794**	**.633**	**1.000**

Table 4: The Basic Characteristics of the Clusters

	Cluster Size (Percentage)	Average intra-cluster distance	Variance of intra-cluster distance
Cluster A	32(13.67%)	1.617	0.256
Cluster B	12(5.13%)	1.807	0.379
Cluster C	21(8.97%)	1.806	0.375
Cluster D	10(4.27%)	1.096	0.092
Cluster E	80(34.19%)	1.054	0.122
Cluster F	19(8.12%)	1.523	0.244
Cluster G	19(8.12%)	1.579	0.436
Cluster H	41(17.52%)	1.191	0.292

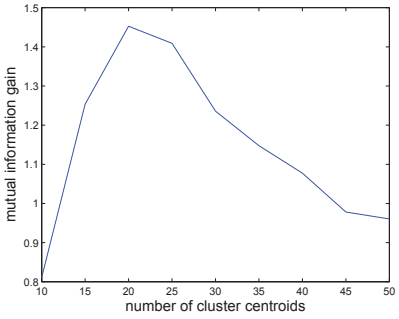


Figure 4: The Mutual Information Gain for Different Number of Clusters

easily interpretable, we selected to analyze the clusters based on the 16 features with coefficients in Factor Analysis larger than 0.5. As Figure 6 shows, the clusters have very distinct patterns with regarding to the 16 features. Cluster A has more replies, retweets or comments from level-2 to level-10 nodes than from level-1 nodes. Cluster B has the longest time span, indicating the corresponding messages have more vitality. Cluster C has the largest value of *bigNodeRatio*, suggesting the users contribute more evenly to the messages compared to other clusters. Cluster D reaches the largest volume of audiences, but its *bigNodeRatio* is the smallest one, and about 80% of the original posters in the cluster have more than 500,000 followers. This may be considered as a typical example of celebrity effects. Cluster E includes around 30% of all messages. The pattern of it is

similar to the overall one. Message cascades in Cluster F have the shortest period of life span in time and a relatively larger $expTimeRatio$, which implies that the messages are very time-sensitive. Investigation the content of the messages reveals that the messages are mainly time-restricted promotion or event by business accounts. Cluster G has a similar level structure and $bigNodeRatio$ to Cluster C. But compared to Cluster C, it has the second largest volume of audiences among all clusters which may be attribute to the fact that 73% of the original posters have more than 500,000 followers. Cluster H has much more replies, retweets and comments from level-2 nodes than the other clusters.

According to Figure 6d, Clusters B, D, E and F are close to the star type, while Clusters A, C, G and H are close to the nebular type. The propagation tree of Cluster H shows a two-polar phenomenon, where the replies, retweets and comments mostly attributed to two major users. Cluster C has a large value for $bigNodeRatio$ and a small value for $totalNode$, suggesting that the propagation of information is limited to a small scale. We can see from the structure of Cluster C that there are several key users with almost equal contributions for the messages.

7. FINDINGS

The resulted clusters shows clear patterns of propagation in Sina-Microblog. In this section, we focus our analysis on the following three aspects, the patterns of advertisement propagation, the role of celebrities in propagation, and effectiveness metrics for advertising.

7.1 Pattern of Advertisement Propagation

In our dataset, we find that $silenceTime$ of 85.4% message cascades (223 of 261) are zero. Further investigation reveal that the average silence time is only 17 minutes. We think this phenomenon is a result due to the microblogging service characteristics. Most of the users only see the messages at the first page. Considering the amount of tweets published per minute in microblog, if an advertisement does not get retweet in the first 20 minutes, it is very likely to be ignored by most users. However, even if an advertisement gets retweet in time, whether it will be propagated further depends on many other factors, including the celebrity effect.

In general, a successful advertisement not only attracts a lot of audience but also reaches users further away in the social graph. As Table 5 shows that most of advertisement using celebrity belong to the Cluster E, which attracted less audience and propagated in shallower levels than average. So the effect of advertisements using celebrity may not be good. In addition, the advertisements through sales promotion are mostly in the Cluster F. Although it has the shortest lifespan due to its time-limited property, the audience size of the Cluster F is twice as many as that of the Cluster E according to Figure 6(a). This phenomenon shows that the users in microblogging service are more interested in discount than products and celebrities, although celebrities commonly have a large number of fans. We can see from Table 5 that most of the news belong to Clusters C, E, G and H, which indicates that news are more likely to attracted influential users than messages of other topics in propagation. The percentage about topic in Table 5 is one of the total cascades.

We pick up three cascades which are all the advertisements about a same product "LYNX perfume" posted by an advertiser user "LYNX". These three ads messages use different strategies including showing products, leveraging celebrity, and sales promotion. We found that the advertisement with showing products belongs to the Cluster H, the advertisement using celebrity belongs to the Cluster E, and the one with sales promotion belongs to the Cluster F. Figure 7 shows some variables of these ads, where the variables are normalized by the max value of each of them. The ads (sales promotion) seem to be the most successful, because it attracted more users than the other two. Considering the cost of inviting a celebrity is quite high, the return on investment (ROI) of ads (celebrity) seems to be low. In another word, the users in microblogging services are more interested in discount rather than products and celebrities. Hence, advertising with sales promotions may be a better choice in Sina-Microblog.

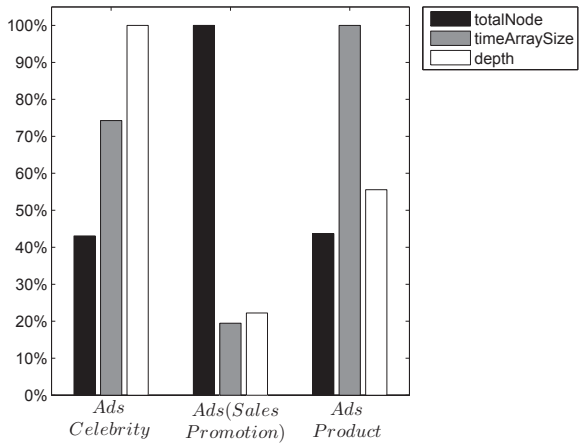
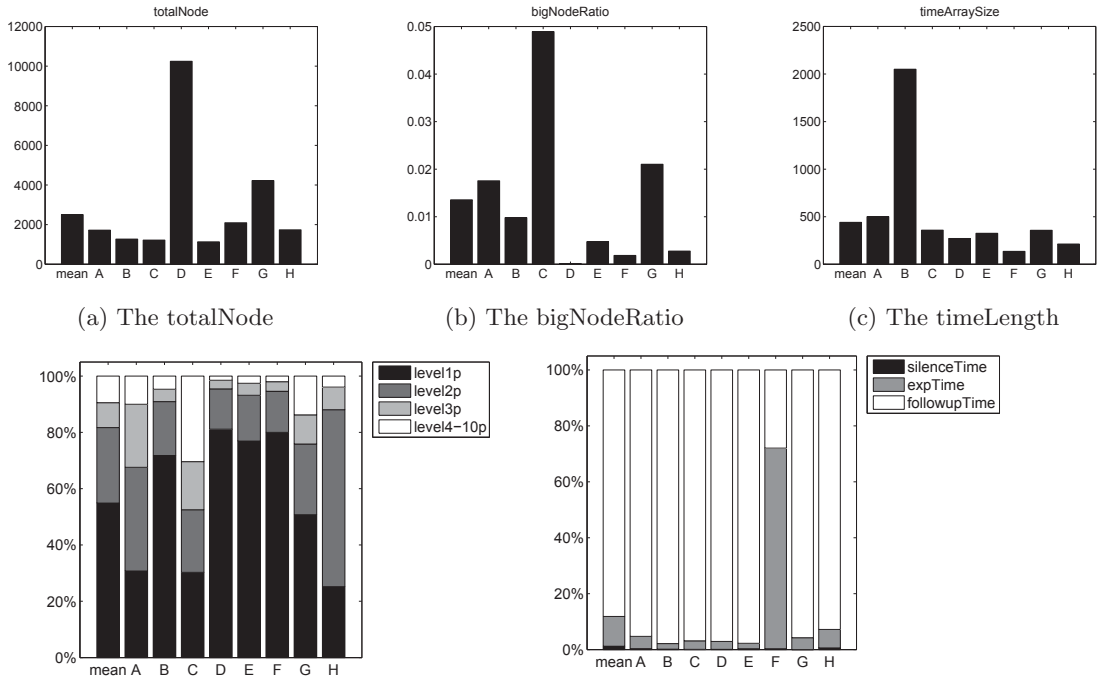


Figure 7: Example of Advertisements

7.2 The Role of Celebrities in Message Propagation

Celebrities have big influence on individual's decision making and this fact is widely leveraged in advertising. In the case of Sina-Microblog, the term 'celebrity effect' means a message will get more replies, retweets or comments if the original poster or some repliers are users with a large volume of followers. It is general accepted that the celebrity effect does exist and is the major type of influence in the social network.

The users in the Clusters B and F were attracted significantly by the originator of the message cascades. This phenomenon is a typical instance of celebrity effect. Similarly, the Cluster D has the largest number of replies or retweets among all clusters. The example shown in Figure 5(d) is originally posted by Yao Chen who have the largest volume of followers in Sina-Microblog. By checking the data, we can easily find that most of the replies, retweets, and comments are attributed to users whose followers are over 10 million. There is a high correlation between $totalNode$ and the proportion of the number of users whose fans are more than 500,000 in each message cascade(with a correlation coefficient of 0.7866). Furthermore, our data set is composed of 749,384 messages from 656,903 users, while the



(a) The totalNode (b) The bigNodeRatio (c) The timeLength

(d) The Proportion of Levels (e) The Composition of Time

Figure 6: The Characteristics of Each Cluster by Selected Features

Table 5: The Classification of Messages in Each Cluster

	News	Ads (product)	Ads (celebrity)	Ads(sales promotion)	Knowledge	Others
Cluster A	2.99%	2.99%	1.71%	0.43%	0.00%	**5.56%**
Cluster B	0.00%	0.00%	0.85%	0.43%	0.00%	**3.85%**
Cluster C	**5.13%**	0.43%	1.71%	0.00%	0.43%	1.28%
Cluster D	0.00%	0.00%	0.00%	1.28%	0.85%	2.14%
Cluster E	**7.26%**	**6.84%**	**5.56%**	1.71%	1.71%	**11.11%**
Cluster F	0.85%	0.85%	1.28%	**4.27%**	0.43%	0.43%
Cluster G	**5.98%**	0.43%	0.43%	0.43%	0.00%	0.85%
Cluster H	**6.84%**	3.42%	0.85%	0.43%	1.71%	4.27%

4785 users contribute to more than 70% replies, retweets and comments. The proportion of them is only 7.284×10^{-3}. Clearly, celebrity effect exists in Sina-Microblog in general.

We choose some messages from the data set with the same content and compare the propagation trees of them. In this experiment, we choose six message cascades which report the same event (i.e. a subway accident). The originators of these cascades are 头条新闻 (*Top News*), 土豆网 (*Tudou.com*), *TBOne*, 成都晚报 (*Chengdu Evening*), 南都周刊 (*Southern Weekly*) and *Andy*. *Top News* and *Tudou.com* are two media sites; *Chengdu Evening* is a newspaper; *Southern Weekly* is a journal; *TBOne* and *Andy* are two individual users. The propagation of *Top News* belongs to Cluster E; the propagation patterns of *Tudou.com*, *TBOne* and *Chengdu Evening* belong to Cluster G. The propagation patterns of *Southern Weekly* and *Andy* belong to Cluster H. Figure 8 plots the variables $followers$, $totalNode$, $timeLength$ and $bigNodeRatio$ of these six cascades. Each variable is normalized with the corresponding maximum value. As it shows, *Top News* has the most followers, but the $totalNode$ is not the largest one and even among the smallest ones. However, *Chengdu Evening* got the most replies, retweeters and comments although its followers are much less than *Top News*. We found, unlike $followers$, $bigNodeRatio$ is more relevant to $totalNodes$. The message propagations of *TBOne* and *Chengdu Evening* are much more successful despite of their smaller number of followers. By checking the detailed propagation pattern of these six cascades, we find that the participation of famous users in the process plays a significant role. The more famous users participated, the more the replies, retweets or comments there would be. So whether the original post user is a famous user does not matter. The participation of famous users is the key to message propagation.

Here, we need to note that the concept of celebrity used by advertisement is not necessarily the same as that of 'celebrity' effect. The term 'celebrity' in celebrity effect means the user who has numerous followers in a microblog site, while the celebrity of advertisement is a famous person in normal life, especially in entertainment or sports.

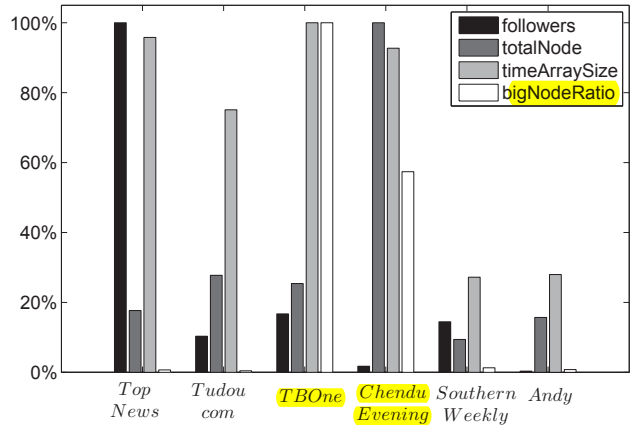


Figure 8: Message Cascades of the Same Content May Demonstrate Different Propagation Patterns Due to Different Users Involved.

7.3 Effectiveness Metrics for Advertisement

Word-of-mouth is a usual form of advertising. The mechanism of advertisement propagation in microblog sites can be considered as a restricted word-of-mouth diffusion. In traditional advertisement, ad ratings, product sales, and click rate are commonly used to measure the effectiveness of an advertisement. However, the traditional advertising evaluation criteria are not directly applicable to the advertisements in microblog. The reaction to an advertisement in microblogs is limited to reply, retweet or comments, but how do they correspond to the effectiveness of the advertisement is unclear. In order to quantitatively measure the effectiveness for advertisement in microblogspace, we attempt to propose a set of new metrics in this section.

First, we examine the following five measures as possible metrics for advertisement effects: *Max Reach*, *FPM* (*Forward Per M*), the variable *totalNode*, *bigNodeRatio* and *timeLength*. (The scores of these three variables in Factor Analysis are all larger than 0.7, and these variables are easily get too.) *Max Reach* is an indicator of quantity, measured by the sum of followers of total repliers. *Max Reach* represents the total number of users to whom the advertisement may be shown. *FPM* is an indicator of content quality, measured by the average number of replies, retweets and comments from every 1000 followers. A good advertisement means the advertisement has been retweeted/commented by as many users as possible. Users' response to an advertisement implies their interests in it. The variable $totalNode$ is the number of users involved in the propagation. We used the variable $bigNodeRatio$ as a measure because it correlates negatively to the degree of celebrity effect in the propagation. Bigger value of the $bigNodeRatio$ suggests smaller effect of celebrity. The $timeLength$ implies the vitality of the message. Table 6 shows the Pearson correlation coefficient matrix among these five measures. The variable $totalNode$ has a high degree of correlation with $MaxReach$ and FPM. We hence remove it from the metrics system. The remaining four variables correlate little with each other. Hence we recommend to user these four metrics to evaluate the effectiveness of an advertisement.

Table 7 shows the mean values of three kinds of advertisements in our metric system. It is clear that no matter which variables, the values of ads with sales promotion are the best. So the ads (sales promotion) is much more suitable to microblogs and get better effects than other kinds of ads. Corporations and organizations may want to advertise their products through sales promotion in microblog, which is much different from traditional media platform.

The four variables in our metrics system are independent of each other, and they represent different aspects of the advertisement effect. As Table 8 shows, we pick up some advertisements of the same products. We choose three ads about tourism. Tourism #2 is a sales promotion which is much more popular than the others. However it has the least duration. On the contrary, Tourism #1 and #3, which showed products and used celebrity respectively, have nearly two times the duration of Tourism #2. Considering that the microblogging is an advertising platform which topics change rapidly, the performance of sales promotions is better.

Table 6: Pearson Correlation Coefficient among Three Measures

	Max Reach	FPM	total Node	big Node Ratio	time Length
Max Reach	**1**	.166	.484	-.174	-.089
FPM	.166	**1**	**.825**	-.119	.070
totalNode	.484	**.825**	**1**	-.129	.031
bigNodeRatio	-.174	-.119	-.129	**1**	.003
timeLengh	-.089	.070	.031	.003	**1**

Table 7: The Mean Values of Ads in Metrics System

	Max Reach	FPM	bigNode Ratio	time Length
Ads(Products)	1787996	4971	0.01918	279
Ads(Celebrity)	2935083	3808	0.00878	379
Ads(Sales Promotion)	2289861	7236	0.00356	479

8. CONCLUSION

This paper presents patterns of advertisements propagation and analyzes the commercial messages in microblogging services. First, we extract variables from message flows in three aspects including the audience size of propagation, the shape of retweet structure and the composition of time. We use Factor Analysis to reduce the dimensionality of the 33 variables. By using K-means clustering we find that there are eight significant patterns in microblogging services, each of which is a combination of three major aspects of different kinds. Second, by comparing the variables of each cluster

Table 8: Examples Ads of Metrics System

	Max Reach	FPM	bigNode Ratio	time Length
Tourism#1	7460176	1927.71	0.000801	70
Tourism#2	14377893	22942.69	0.000059	43
Tourism#3	4724772	6425.43	0.000183	72

we find that news usually reach a wide audience with high efficiency. Further, we demonstrate the power of the celebrity effect in message propagation. Finally, by analyzing the influence of different patterns of advertisement message we find the effect of the advertisements using celebrity is not as good as what is expected. Furthermore, we propose a set of effectiveness metrics for advertisements in microblogs which describe the audience size of the advertisement, user acceptance of advertisement content, the participation of celebrities and the lifespan of the advertisement.

Acknowledgments

This work is partially supported by Shanghai Science and Technology Rising Star Program, Grant No. 11QA1403500, Shanghai Talent Development Fund, Grant No. 2010002, Shanghai Key Laboratory of Digital Media Processing and Transmission, Grant No. STCSM(12DZ2272600).

9. REFERENCES

[1] J. F. Asli Celikyilmaz, Dilek Hakkani-Tur. Probabilistic model-based sentiment analysis of twitter messages. *The 2010 IEEE International Conference on Spoken Language Technology Workshop (SLT)*, 2010.

[2] P. P. E. H. C. Bongwon Suh, Lichan Hong. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. *The 2010 IEEE International Conference on Social Computing/IEEE International Conference on Privacy, Security, Risk and Trust*, pages 177–184, August 2010.

[3] R. D. L. V. C. Dan Zhang, Yan Liu. Alpos: A machine learning approach for analyzing microblogging data. *The 2010 IEEE International Conference on Data Mining Workshops*, pages 1265–1272, 2010.

[4] J. K. Daniel M. Romero, Brendan Meeder. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. *The 2011 International World Wide Web Conference (WWW)*, June 2011.

[5] J. K. David Liben-Nowell. Tracing information flow on a global scale using internet chain-letter data. *The National Academy of Sciences*, 2008.

[6] W. A. M. D. J. W. Eytan Bakshy, Jake M. Hofman. Everyone's an influencer: Quantifying influence on twitter. *The 2011 International Conference on Web Search and Data Mining (WSDM)*, 2011.

[7] H. P. S. M. Haewoon Kwak, Changhyun Lee. What is twitter, a social network or a news media? *The 19th International World Wide Web Conference (WWW)*, pages 591–600, April 2010.

[8] T. Y. C. Q. L. L. He Yanxiang, Suwen. Summarizing microblogs on network hot topics. *The 2011 International Conference on Internet Technology and Applications (iTAP)*, pages 1–4, August 2011.

[9] M. M. B. G. S. P. A. F. F. M. Jacob Ratkiewicz, Michael Conover. Truthy: Mapping the spread of astroturf in microblog streams. *The 20th International World Wide Web Conference (WWW)*, 2011.

[10] M. W. Jianjun Xie, Chuang Zhang. Modeling microblogging communication based on human dynamics. *The 8th International Conference on Fuzzy System and Knowledge Discovery*, 2011.

[11] V. L. Marc Cheong. A study on detecting patterns in twitter intra-topic user and message clustering. *The 2010 20th International Conference on Pattern Recognition (ICPR)*, pages 3125–3128, 2010.

[12] F. B. K. P. G. Meeyoung Cha, Hamed Haddadi. Measuring user influence in twitter: The million follower fallacy. *The 2010 International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[13] Z. J. W. L. H. W. Pengyi Fan, Pei Li. Measurement and analysis of topology and information propagation on sina-microblog. *The 2011 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 396 – 401, July 2011.

[14] F. W. Shaozhi Ye. Measuring message propagation and social influence on twitter.com. *The 2010 International Social Informatics Conference*, 2010.

[15] M. C. K. P. G. V. A. Tiago Rodrigues, Fabricio Benevenuto. On word-of-mouth based discovery of the web. *ACM SIGCOMM Internet Measurement Conference (IMC'11)*, 2011.

[16] D. C. Z. D. W. K. Wojciech Galuba, Karl Aberer. Outtweeting the twitterers-predicting information cascades in microblogs. *The 2010 USENIX Conference on Online Social Networks (WOSN)*, 2010.

[17] F. R. Ye Wu. Learning sentimental influence in twitter. *The 2011 International Conference on Future Computer Sciences and Application (ICFCSA)*, 2011.

[18] J. T. K. C. J. L. Zi Yang, Jingyi Guo. Understanding retweeting behaviors in social networks. *ACM Conference on Information and Knowledge Management (CIKM)*, 2010.