IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 10, 2018, accepted February 6, 2018, date of publication February 20, 2018, date of current version March 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2808158

Understanding User Behavior in Sina Weibo Online Social Network: A Community Approach

KAI LEI[1], YING LIU[1], SHANGRU ZHONG[1], YONGBIN LIU[1], KUAI XU[2], YING SHEN[1], AND MIN YANG[3]

[1]K. Lei, Y. Liu, S. Zhong, Y. Liu, and Y. Shen are with the Shenzhen Key Laboratory for Cloud Computing Technology and Applications, School of Electronic and Computer Engineering, Institute of Big Data Technologies, Peking University, Shenzhen 518055, China
[2]K. Xu is with the School of Mathematical and Natural Sciences, Arizona State University, Tempe, AZ 85281, USA
[3]M. Yang is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

Corresponding author: Min Yang (min.yang1129@gmail.com)

This work was supported by the Shenzhen Key Fundamental Research Projects under Grant JCYJ20170412150946024 and Grant JCYJ20170412151008290.

ABSTRACT Sina Weibo, a Twitter-like microblogging Website in China, has become the main source of different kinds of information, such as breaking news, social events, and products. There is great value to exploiting the actual interests and behaviors of users, which creates opportunity for better understanding of the information dissemination mechanisms on social network sites. In this paper, we focus our attention to characterizing user behaviors in tweeting, retweeting, and commenting on Sina Weibo. In particular, we built a *Shenzhen Weibo community graph* to analyze user behaviors, clustering the coefficients of the community graph and exploring the impact of user popularity on social network sites. Bipartite graphs and one-mode projections are used to analyze the similarity of retweeting and commenting activities, which reveal the weak correlations between these two behaviors. In addition, to characterize the user retweeting behaviors deeply, we also study the tweeting and retweeting behaviors in terms of the gender of users. We observe that females are more likely to retweet than males. This discovery is useful for improving the efficiency of message transmission. What is more, we introduce an information-theoretical measure based on the concept of entropy to analyze the temporal tweeting behaviors of users. Finally, we apply a clustering algorithm to divide users into different groups based on their tweeting behaviors, which can improve the design of plenty of applications, such as recommendation systems.

INDEX TERMS Sina Weibo, online social network, user behavior, bipartite graphs, entropy, clustering.

I. INTRODUCTION

Sina Weibo is one of the most popular social network sites in China and has become the crucial medium for over 600 million registered users to share information such as breaking news, social events and products [1]. Despite its popularity, there are few systematic studies exploring the characteristics of Sina Weibo.

In this paper, we propose a novel *community* approach for exploiting and understanding the Sina Weibo online social network site. Communities are the dense groups of the vertices, which are tightly coupled to each other inside the group and loosely coupled to the rest of the vertices in the network. Community approach plays a key role in understanding the complex networks. A systematic study is presented to characterize the tweeting activities of the users, which explores continuous public status streams and user status streams using Weibo open platform APIs. The public status streams can be used to collect the up-to-date tweets published by all Weibo users and the user status streams can be used to collect tweets of specified users with corresponding user ids. We focus our attention to characterizing the user behaviors in tweeting, retweeting and commenting, which creates opportunity for better understanding of the information dissemination mechanisms on social network sites. First, we collect all Weibo users in Shenzhen beginning from our seed users which are obtained by manually selecting the users in Shenzhen who have a large number of followers, and then tracking relationships among these users, and their original tweets, retweets and comments. We also use public status streams to collect real-time status information, which we divide into two independent groups of Weibo users – a subset of users from the global Weibo user population, called the global

cluster; and a subset of users from a local Weibo community registered with Shenzhen as the primary geographic location, called the local cluster.

Second, we use all collected data to build a *Shenzhen Weibo community graph* based on their following relationships. This graph can be used to exploit the influences and interactions among users. Given the graph structure, we measure the tightness of user relationships by exploring the clustering coefficients and use the edges information to explore the distribution of user popularity. Then we study the correlations between user popularity and the clustering coefficient. Furthermore, we discuss the influence of user popularity on retweeting and commenting activities, and characterize temporal patterns of retweeting and commenting activities on the original tweets posted by a variety of users with different levels of popularity. Moreover, based on the bipartite graphs and one-mode projections, we model the interactions of Shenzhen Weibo users and analyze the similarity of retweeting and commenting activities among these users. In addition, we apply an information-theoretic measure, based on the concept of entropy to characterize the distribution of inter-tweet time intervals of Weibo users in the global and local clusters, to reveal their diverse temporal tweeting patterns. Furthermore, we discuss the correlation between the temporal tweeting behavior of Weibo users, their tweeting frequency and user popularity, and analyze gender differences in tweeting and retweeting, which reveals different intraday behavioral trends in tweeting and retweeting between males and females.

The diverse behavior in tweeting sheds light on the clustered patterns of users with similar temporal tweeting patterns and user popularity. We use a k-means clustering algorithm to partition Weibo users into different groups. The results show that the users in different clusters exhibit distinct tweeting behavioral characteristics, thus reflecting different interests and objectives of Weibo users in exploiting the Weibo microblogging service.

We summarize our main contributions as follows:

- This paper introduces a novel approach to characterize the graphical structure and small-world nature of a community online social network – Sina Weibo;
- Based on the *Shenzhen Weibo community graph*, we discuss the impact of user popularity based on the frequency of retweeting and commenting activities on their tweeted messages and temporal patterns of information spread over Sina Weibo. We also analyze the similarity of retweeting and commenting behavior among users in the community and reveal the weak correlations between following, retweeting and commenting behaviors.
- We characterize the temporal tweeting patterns of Sina Weibo in terms of entropy to reveal a diversity of tweeting patterns of Weibo users.
- Based on the study of temporal tweeting patterns, we further analyze the different behaviors in tweeting and retweeting between males and females.
- This paper applies a clustering algorithm to partition Weibo users into different groups with distinct tweeting behavioral characteristics to identify Weibo users with similar interests and objectives.

The remainder of this paper is organized as follows. Section II describes our data collection effort and community approaches for big data analytics. Section III introduces the social network graph, while Section IV is devoted to discussing the impact of user popularity on their tweeted messages. Section V analyzes the similarity between retweeting and commenting behavior among users in the Shenzhen Weibo community. In section VI, we study the temporal tweeting patterns of Weibo users to characterize user tweeting behavior and then we discuss gender differences in tweeting and retweeting in section VII. In light of the diversity in tweeting behavior, section VIII applies clustering algorithms to allocate Weibo users into groups with distinct tweeting behavioral characteristics. In Section IX, we discuss related work and Section X concludes this paper and outlines our future work.

II. BACKGROUND

A. BRIEF INTRODUCTION TO WEIBO

Since its launch in August 2009, Sina Weibo has grown boomingly and rapidly to become an influential site for millions of internet users in China. It is used to disseminate news and information, promote new products, and express opinions and comments on popular events or controversial issues. Like other online social medium which have attracted significant attention from the research community [2]–[5], many studies have gradually focused on Sina Weibo. Guo *et al.* [1] studied the topology of social network graph on Sina Weibo and discovered an asymmetric property in the following relationships among users, as well as a densely connected topology of network formed by active users.

By comparing with Renren which is a Facebook-like social network, [6] revealed that Sina Weibo is more efficient for information diffusion. They also presented an analysis of ''Hall of Fame'' on Sina Weibo, which is a group of popular users verified and recommended by Weibo, indicating that relationships between users on Weibo are much looser than other friendship-based online social networks such as Facebook or Renren [7]. However, studies focused on user behavior on Sina Weibo are still at an early stage. Reference [8] characterized user behavior by analyzing changes in the number of V-users followers and the correlation between the number of followers and the frequency of tweeting. Differently, in this paper, we concentrate on user interactions and the features of user behavior in tweeting, retweeting and commenting to understand user behavior on Sina Weibo.

B. DATA COLLECTION

We collected our data via Sina Weibo's open platform APIs, which are developed to allow third-party application programmers to explore a huge amount of data such as tweets,

retweets, comments, user profiles and following/follower relationships on Sina Weibo. There are two stages in our data collection process. In the first stage, we focus on two types of information from Sina Weibo: relationships and content. The relationship information can capture following and follower activities among users, then can generate a social network graph of users. Meanwhile, the content constitutes of tweeted messages and the corresponding retweets and comments, capturing the process of information cascading over this social network as well as the influence of its users.

We use Weibo's API to crawl data and select Shenzhen as the community due to its technology-savvy population. In Sina Weibo, we can obtain the user's location from their profile, thus we use this property to determine whether a user is based in Shenzhen. Users follow other users who they are interested in. In this paper, we consider the act of following as a forward link. If user A follows user B, we say that A is a *follower* of B, meanwhile B is a *friend* of A, and the forward link extends from A to B. Because of the experiments in [9], which show that it is very cost-effective to crawl online social networks using the forward-link approach, we adopt this method for crawling Shenzhen users, which was illustrated in Algorithm 1 [10]. With this algorithm, we successfully obtained data on over 2 million Weibo users from Shenzhen and over 75 million direct following relationships among these users.

Algorithm 1 The Algorithm for Crawling All Shenzhen Users of Sina Weibo Online Social Network

Require: a set of seed users S: the top N users of Sina Weibo social network; all Shenzhen users U, initially is empty;

```
push S into Q_tasks;
while Q_tasks is not NULL do
    u = pop Q_tasks;
    find all the friends of u: friend_u;
    for v ∈ friend_u do
        if v is from Shenzhen and v hasn't been crawled then
            U = U + u;
            push v into Q_tasks;
        end if
    end for
    mark u as being crawled.
end while
```

After collecting the Weibo user's relationship data, we collected over 47 million original messages tweeted by users from Shenzhen and obtained over 4 million retweets as well as over 10 million comments on these tweets, written by other users in the Shenzhen Weibo community.

In the second stage, we were particularly interested in three types of data: public status streams, complete tweeting activities of selected users, and social network graphs of these selected users. We used the `public_timeline` API to crawl the real-time and sampled status streams of users, which returned a maximum of the 200 most recent

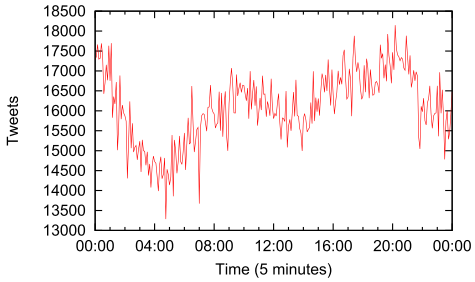


FIGURE 1. The number of sampled tweets in public status streams from February 1, 2016 to February 22, 2016.

randomly-selected tweets so-called statuses or weibos. Throughout this paper, we will use the three words, *status*, *tweet*, and *weibo*, interchangeably. In other words, these public status streams serve as a sample of all statuses posted by millions of Weibo users. As our empirical experiment found that making API calls every second often leads to the same set of status streams, we set the intervals between consecutive API calls to 5 seconds. This does not overload the API servers while obtaining fresh sets of status streams from each API call. During the data collection, we removed duplicate statuses that have been returned in previous API calls.

Fig. 1 illustrates the number of statuses, every 5 minutes, captured by our continuous `public_timeline` API calls during a 24-hour window. These real-time status streams, although they are a sample of all Weibo statuses, contain a rich set of valuable information on user tweeting behavior and information cascading patterns over the Weibo online social network. We collected a total of over 42 million statuses during a 22-day data collection period.

Due to the sheer size of the Sina Weibo population, it is impractical to collect all statuses of all Weibo users. Instead, we adopt a simple sampling approach to collect the complete tweeting activities of two independent groups of Weibo users — a subset of users from the global Weibo user population and a subset of users from a local Weibo community.

To extract a subset of global users, we selected the top 1,000 Weibo users, based on the number of tweets captured in public status streams during the data collection. For simplicity, we use "*global cluster*" to refer to users in this group. We divide the Weibo user population into local communities based on geographical location. In this stage, we selected the top 1,000 users of the Shenzhen Weibo community based on statuses captured in public status streams as the second group of Weibo users. For simplicity, we use "*local cluster*" to refer to the subset of users from the Shenzhen Weibo community.

For each user in the global and local clusters, we make Weibo `user_timeline` API calls and we obtain other API links by analyzing the Sina Weibo data access patterns to harvest its user status streams, i.e., complete tweeting activities of the user during the same time period as the public status streams. For each user in these two clusters, we also collect their number of friends and followers via API links we obtain after analyzing the Sina Weibo data access patterns.

C. COMMUNITY APPROACHES FOR BIG DATA ANALYTICS

The increasing use of big data has brought huge opportunities for data analysis. The research community has proposed many solutions for example, cloud computing, sampling, community approaches, NoSQL, etc. Network and graph theories are often used in social network analysis to investigate social structures. Here, network structures are characterized in terms of nodes (individuals or things within the network) and edges (relationships or interactions) that connect the nodes. However, the explosive growth in the number of users and tweets makes it impossible to map the whole network and derive an insight understanding of the entire online social network. Here we use a representative community to reflect the whole network's characteristics.

III. SOCIAL NETWORK GRAPHS

We build the community graph that consists of all local users from the Sina Weibo online social network in this section. Particulary, we present an analysis of the community graph based on symmetric property of the following relationships and study the statistical distribution of user popularity measured by the number of followers. Then, we use the clustering coefficients of the community graph to validate the small-world nature of such community. Finally we discuss the correlation between user popularity and the clustering coefficient.

For the Shenzhen users, relationships are initiated by following other users. We consider the relationship between two users as an edge, and each user as a vertex, in building a social network graph, namely the ''Shenzhen Weibo community graph''. Based on the graph structure, we can use graph theory to analyze user interactions, such as retweeting and commenting.

We use $\mathcal{N}$ to denote all the users from Shenzhen in the community graph, while $\mathcal{E}$ denotes the following relationships among $\mathcal{N}$ users. Thus the community graph can be represented as $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$. In our dataset, $|\mathcal{N}| = 2234591$, and $|\mathcal{E}| = 75974458$, which forms a useful sample for analyzing user interactions and the features of user tweeting behavior.

A. FOLLOWERS AND FRIENDS

As we mentioned above, user A follows user B makes A become a follower of B while B is a friend of A. Users can also follow each other, which will create two relationships between two users, referred to as a *reciprocal* following relationship. Thus, in the Shenzhen user community graph, it is common to observe two edges existing between two vertices. There are 43, 557, 116 pairs of users who follow each other in the community graph. Hence, the number of *reciprocal* following relationships is 43, 557, 116. *Reciprocal rate* $rr = \frac{|\mathcal{E}_{\mathcal{R}}|}{|\mathcal{E}|}$ can be applied for measuring the degree of reciprocal following in a social network, where $\mathcal{E}_{\mathcal{R}}$ represents the set of edges in $\mathcal{G}$ that have corresponding reciprocal edges. In the Shenzhen user community graph, the reciprocal rate is 0.57,

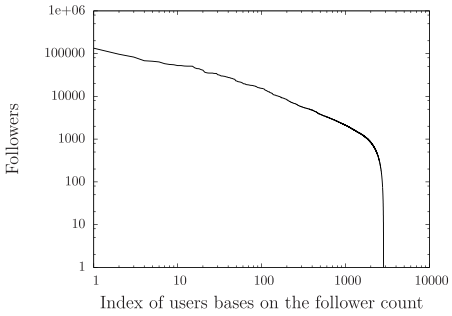


FIGURE 2. The distribution of followers for Shenzhen users.

which is obviously higher than that of the Sina Weibo [8] and Twitter [9] as a whole.

B. USER POPULARITY

In online social networks, popular users are defined as those who have a large number of followers. This makes it very effective to advertise and distribute news by popular users. Thus, it is very useful to be able to identify the popular users in an online social network. In order to quantify *user popularity*, we use the count of followers. In the community graph, if a user has a large number of followers then he or she is very popular. As illustrated in Fig. 2 [10], there are a small amount of popular users, whose followers number greater than 50, 000. Out of 2, 234, 591 Shenzhen users, less than 1% have more than 1000 followers and the vast majority have less than 10 followers.

C. CLUSTERING COEFFICIENT

In the community graph, there are some users whose neighbors follow each other closely. In order to find these close groups, we adopt the clustering coefficient approach, which is a measure of the degree to which nodes in a graph tend to cluster together. The formula for calculating the clustering coefficient of user u is illustrated as:

$$CC_u = \frac{number\ of\ pairs\ of\ neighbors\ connected\ by\ edges}{number\ of\ pairs\ of\ neighbors}. \tag{1}$$

where neighbors of user u are the users who follow user u. Clearly, $0 \leq CC_u \leq 1$. If any two users are the neighbors of user u and all follow each other, then CC_u is 1. Conversely, if there is no relationship in the following behavior of neighbors, the value of CC_u is 0. In Fig. 3 [10], most clustering coefficients are between 0 and 0.2. In other words, most users are not in close groups with their neighbors. However, there are about 190000 users whose clustering coefficient is 1. This means that there are some close groups within the Shenzhen user community. The presence of these distinct ''small-world'' implies that there are different types of user group in the Shenzhen user community and since the relationships within different groups vary, the progress of message dissemination is different. In the close groups, the spread of the messages can be very rapid.

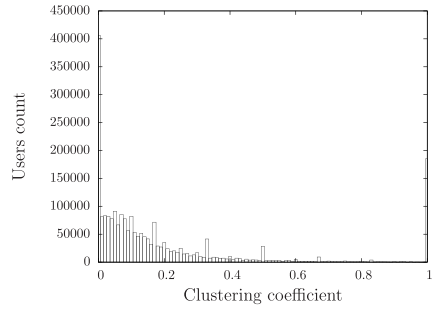


FIGURE 3. The distribution of clustering coefficients for Shenzhen users.

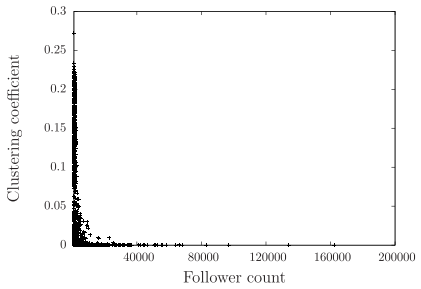


FIGURE 4. The correlation between follower counts and clustering coefficients.

D. CORRELATIONS BETWEEN USER POPULARITY AND THE CLUSTERING COEFFICIENT

Besides analyzing the user popularity and the clustering coefficient in isolation, we are also interested in the correlations between them. In this part, we discuss whether close groups exist in both highly popular and less popular users. As illustrated in Fig. 4 [10], most of the users whose follower counts are low have a larger clustering coefficient than popular users. Specially, the clustering coefficient of users whose follower count is larger than 40000 is almost 0. This means that popular users have no close groups; their followers are just interested in the popular user and do not know each other. However, less popular users whose follower count is low tend to have a high clustering coefficient. Their followers may also be friends in the real world or share common interests, which make them a close group.

IV. IMPACT OF USER POPULARITY ON THEIR TWEETS

Based on the community graph including all Shenzhen users, and the public status streams from Sina Weibo, we explore the impact of user popularity on tweeted messages and capture the temporal patterns of retweeting and commenting behaviors for tweets from Sina Weibo users, and subsequently perform the correlation analysis in this section.

In Sina Weibo, tweeted messages can be retweeted and commented on by others. The behavior of retweeting and commenting on tweeted messages posted by other users having a high follower count may be different from those posted by less popular users. In this part, we focus on three issues to understand the influence of user popularity on their tweeted message. First, we discuss whether tweets from users with high popularity result in significant retweeting and commenting behaviors. Then we further analyze whether most of the retweeting and commenting behaviors come from the followers. Finally, we analyze the dispersal speeds of tweets from different users to better understand the impact of user popularity.

We consider users who have more than 1, 000 followers as popular users and analyze the correlations between their follower count and the average number of retweets and comments received by their tweeted messages. As illustrated in Fig. 5[a][b] [10], most users have a small average amount of retweets and the high user popularity may not always result in high levels of retweeting and commenting actions.

However, there are a few users, who have relatively fewer numbers of followers but actually obtain a surprisingly large amount of retweets or comments. For example, the official Weibo account for Shenzhen University, has only 9,627 followers, but each tweet posted by this account receives an average of 78 retweets, achieving the highest retweet amount over all users in the Shenzhen Weibo community. After sampling survey on the users who retweet the Shenzhen University Weibo account's tweet. We found that most of them come from college students, who are very active on Weibo and other online social networks [10]. According to the above analysis, we can find that there is no correlation between user popularity and the average number of retweets and comments received by their tweeted messages. Popular users do not always have a high number of retweets and comments on their tweets.

Furthermore, we explore whether the retweeting and commenting behaviors mainly come from followers of the users who tend to post original tweets. As shown in Fig. 6[a] [10], in retweets, both followers and non-followers retweet popular user's tweets, but the majority of retweets actually result from the followers of the users. As for comments on user tweets, Fig. 6[b] [10] shows similar results. Comments on tweets nearly always come from followers.

Users can retweet other user's tweets to interact with them and the messages carried by tweets are spread in this way. Tweets carrying different messages and posted by different users have different spreading speeds. In order to understand the influnce of user popularity on the spreading speeds of retweets and comments, ''we characterize the temporal patterns of retweeting and commenting on tweets from users with a variety of follower counts. In particular, we select five users who rank 1st, 5th, 10th, 15th and 50th based on their follower count to measure the percentage growth of retweeting and commenting activities over time'' [10]. As illustrated in Fig. 7[a][b] [10], within 10 hours, over 50% of retweets and comments have been posted on the tweets originally posted by the users ranked 1st, 5th and 10th, while half of retweets of the tweets from the 15th and 50th happened after nearly two days. In general, higher ranking users have their retweets and comments made faster. In other words, the experimental results indicate that the retweeting and commenting behaviors of tweets from users with higher user popularity are more

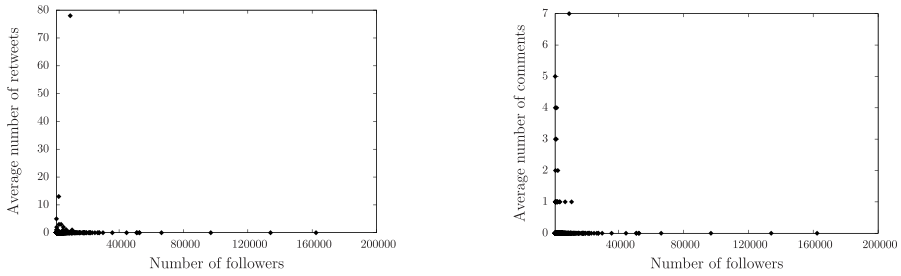


FIGURE 5. The correlation between follower counts and the volume of retweeting and commenting activities. (a) Retweeting activities. (b) Commenting activities.

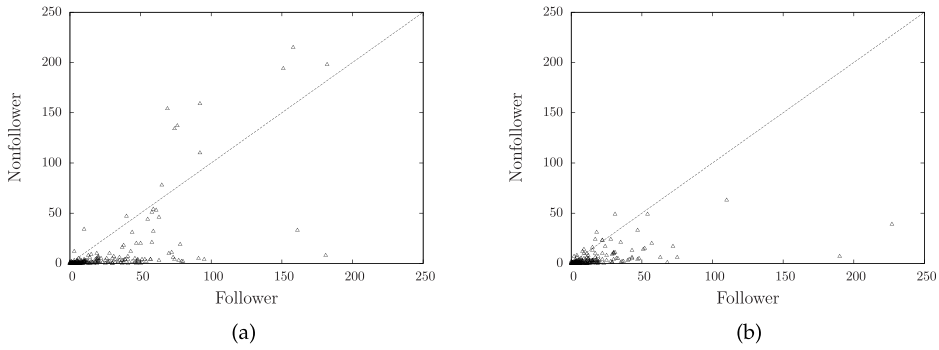


FIGURE 6. The difference between followers and non-followers who retweet and comment on the tweeted messages from popular users in the Shenzhen Weibo community. (a) Retweeting activities. (b) Commenting activities.

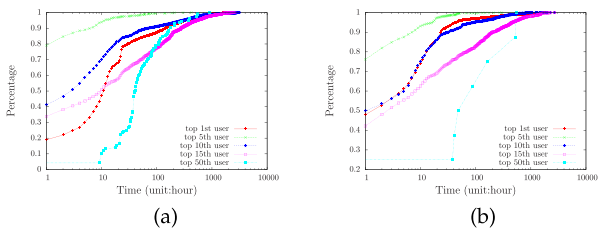


FIGURE 7. Temporal patterns of retweeting and commenting activities. (a) Temporal pattern of retweets. (b) Temporal pattern of comments.

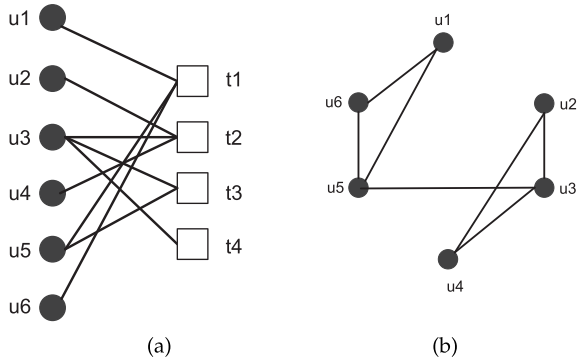


FIGURE 8. Modeling user interactions in the Shenzhen Weibo community with bipartite graphs and one-mode projection: (a) the retweeting interactions between six Weibo users in the Shenzhen Weibo community ($u_1, \ldots, u_6$) and four tweets ($t_1, \ldots, t_4$); (b) the similarity of retweeting activities among four users. (a) Bipartite graph. (b) One-mode projections.

likely to spread much faster than those posted by users with lower popularity.

V. SIMILARITY ANALYSIS OF INFLUENCED BEHAVIORS

In this section, we analyze the similarity between users in terms of retweeting and commenting in order to study whether Weibo users who share friends have similar interests in retweeting and commenting upon common original tweets. Since bipartite graphs have been widely used to model bipartite interactions such as authors and publications [11], [12], we also use this approach to build a bipartite graph between Weibo users and the tweets they have retweeted. As illustrated in Fig. 8[a] [10], if a user retweets a tweet, there will be an edge connecting the user and the tweet. Furthermore, based on the bipartite graph, we adopt the one-mode projection approach to build a one-mode projection graph to connect Weibo users. As illustrated in Fig. 8[b] [10], any two users who retweet the same tweets will have an edge connecting them in the one-mode projection graph. The weight of the edge captures the number of shared tweets both users have retweeted or commented upon. By using the bipartite and one-mode projection graphs, we can calculate the number of shared retweets in the Shenzhen user community. Adapting the same approach, we can obtain the number of shared comments and shared friends.

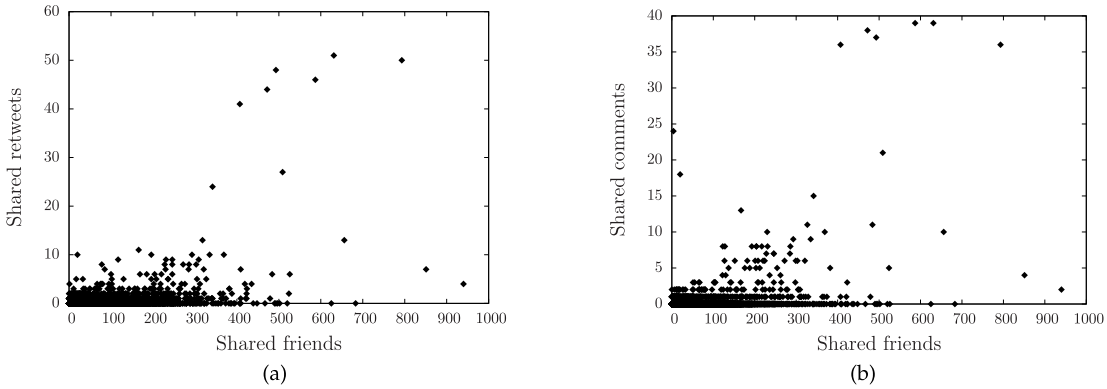


FIGURE 9. Analysis of shared friends and common retweeting and commenting behavior for users in the Shenzhen Weibo community. (a) Shared retweeting behavior. (b) Shared commenting behavior.

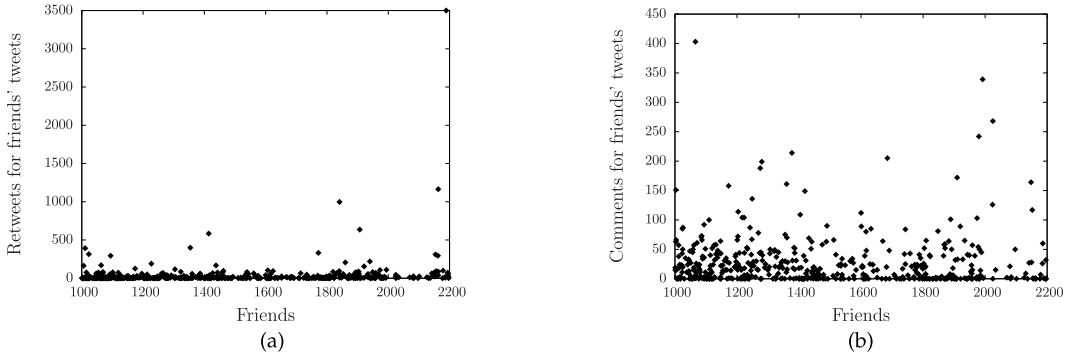


FIGURE 10. Following behavior, retweeting and commenting activities for users with a large number of friends in the Shenzhen Weibo community. (a) Retweeting activities. (b) Commenting activities.

Fig. 9[a][b] [10] illustrate the correlation between the number of shared friends and similar retweeting and commenting behavior for Shenzhen users. From these figures, we find that many users share hundreds of friends but they have few retweets or comments on a single common tweet. This means that a high number of shared friends does not often lead to similar behavior in retweeting and commenting activities. According to this result, we can suggest that although users may follow the same users, they do not interact with each other. They just passively view the original users tweets.

In order to validate the above conjecture, we further study the correlation between the number of friends and the number of retweets and comments on the tweets of friends to explore whether Weibo users with a large number of friends actively retweet or comment on his or her friend's tweets. Our experimental results show that most Weibo users simply read or ignore their friend's tweets, and so do not participate in the process of spreading information around Sina Weibo. As illustrated in Fig. 10 [10], most users make a small number of retweets and comments on their friend's tweets, and this is the case for users who have a small number of friends as well as for users who have many friends. However, although the number of retweets and comments on friend's tweets are similar, comparing with users who have many friends, we find that some users with fewer friends actively participate in information distribution via retweeting and commenting on the tweeted messages of their friends.

VI. TEMPORAL TWEETING BEHAVIOR

In this section, we first characterize the temporal patterns of tweeting in Sina Weibo users and subsequently perform a correlation analysis between the temporal tweeting patterns of Weibo users and their tweeting volume as well as their user popularity as measured by the number of their followers.

A. CHARACTERIZING THE TEMPORAL PATTERNS OF TWEETING BEHAVIORS

The public status stream from Sina Weibo not only provides information on the sampled tweeting activities of Weibo users but also reveals other interesting tweeting characteristics for individual users. Fig. 11 illustrates the distribution of the numbers of sampled tweets for all Weibo users whose tweets were captured in public status streams during the data collection period. It is very interesting to see that there are a number of users who are very active in posting tweets, as evidenced by the large number of their tweets in randomly selected public status streams.

To gain a deeper understanding of the tweeting activities of these active users, we analyze the time *when* they post these tweet messages. Fig. 12[a][b] illustrate time-series

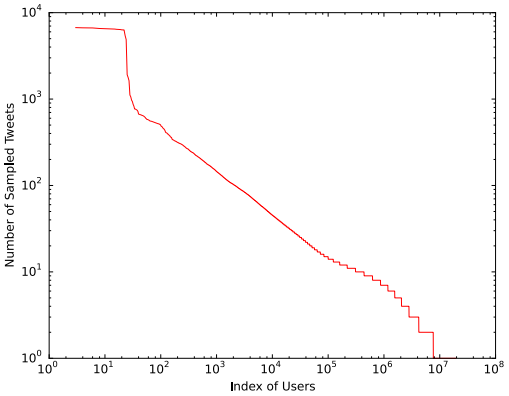


FIGURE 11. Distribution of the numbers of sampled tweets for all Weibo users.

patterns of tweeting activities from the most active users in the global and local clusters, based on the number of complete tweets. We discover that some Weibo users submit tweets with a certain fixed frequency, i.e., posting one or more tweets in a given time period. For example, one Weibo account tweets advertisements about online products in a very regular manner. On the other hand, some other users appear to post tweet messages in a completely random fashion, reflecting the underlying random tweeting behavior.

The uncovering of interesting temporal patterns in tweeting behavior led us to apply the concept of entropy to measure the distributions of inter-tweet time intervals. Entropy is a widely-used technique to characterize the distribution of observations. For a given Weibo user u in the global or local cluster, we first extract the time-stamps of all complete tweets. Let N_u denote the number of all tweet messages during the data collection period, and let t_i represent the time-stamp for the i-th tweet, where $1 \leq i \leq N_u$. Thus, the inter-tweet interval $\delta_{j,j+1}$ is derived as $t_{j+1} - t_j$, where $1 \leq j \leq (N_u - 1)$. For simplicity, we use δ_j to denote $\delta_{j,j+1}$, and accumulate tweets into 5-minute time bins. In other words, a δ of 1 means the time interval between two consecutive tweets is 5 minutes.

To calculate the entropy of inter-tweet time intervals for a user u, we continue to find the probability, p_k, of each unique time interval k, $1 \leq k \leq m_u$, where m_u denotes the total number of unique inter-tweet intervals for user u. Then, the entropy of the inter-tweet time intervals for user u is measured as the following equation:

$$e_u = -\sum_{i=1}^{m_u} p_i \log p_i. \tag{2}$$

Considering the diversity in the total number of tweets posted by different Weibo users, we further calculate the standardized entropy, e'_u, by normalizing the entropy values over the maximum entropy, $log(N_u - 1)$, for each user u, i.e.,

$$e'_u = \frac{e_u}{log(N_u - 1)}. \tag{3}$$

A standardized entropy, e'_u of 0 indicates that there exists a single value of the inter-tweet interval, suggesting the user tweets at the same regular time interval. On the other hand, a standardized entropy, e'_u of 1 indicates that the user tweets at random time intervals. Using the above equation, we calculate the standardized entropies of the inter-tweet time intervals for all users in the global and local cluster. Fig. 13[a][b] show the histogram of the standardized entropies for the two clusters. In the global cluster, most of the users tweet regularly because about 85% of users have standardized entropies that are less than 0.1. Conversely, in the local cluster, the distribution of entropy is relatively uniform. However, there are some users who tweet at regular intervals for example, there are about 200 users whose standardized entropies are about 0.1. After the analysis, we found that most of these users are advertising or news Weibo accounts. They tweet advertisement information about online products or news at regular intervals.

B. CORRELATION ANALYSIS OF TWEETING BEHAVIORAL CHARACTERISTICS

The diversity in the temporal tweeting patterns of Weibo users motivates us to explore the correlations between temporal tweeting patterns and other behavioral characteristics such as their tweeting frequency as measured by the number of tweets and user popularity, measured by the number of followers.

As shown in Fig. 14[a][b], a user who generates a large number of tweeting messages tends to exhibit small standardized entropies, suggesting that these users are likely to post messages at approximately similar time intervals. Meanwhile, users who tweet less frequently exhibit higher entropies, reflecting their random tweeting temporal patterns. Fig. 15[a][b] demonstrate the correlation between the number of followers and the standardized entropy of inter-tweet time intervals for users in the global and local clusters. It is very interesting to observe the clustering patterns of points for both clusters. Note that each point represents one Weibo user, thus clustered points suggest a group of users sharing similar levels of follower counts and standardized entropies.

In summary, our proposed information-theoretic measure, based on the concept of entropy, captures interesting temporal patterns in the tweeting behavior of Sina Weibo users. In addition, our correlation analysis shows interesting relationships between the temporal tweeting patterns of Weibo users, their tweeting frequency and user popularity.

VII. GENDER ANALYSIS IN TWEETING AND RETWEETING

Based on the analysis of information entropy, we have found temporal patterns in the tweeting of Sina Weibo users. We also understand the relationship between the user popularity and the temporal tweeting patterns, which were shown in the previous section. In this section, we will further study the properties of tweeting and retweeting by gender. To do this, we count the number of females and males in the local and global clusters. As illustrated in Fig. 16, in both clusters, the number of men is slightly higher than the number of women, particularly in the local cluster. This indicates that tweeting is more popular for males than for females. Next, we calculate

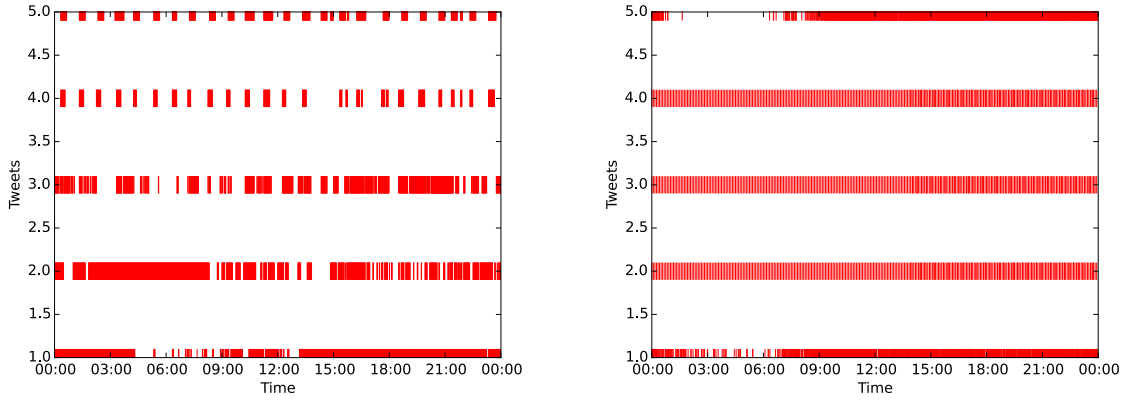


FIGURE 12. Time-series patterns of tweeting activities for the top 5 users in the global and local clusters. (a) top 5 users of the global cluster. (b) top 5 users of the local cluster.

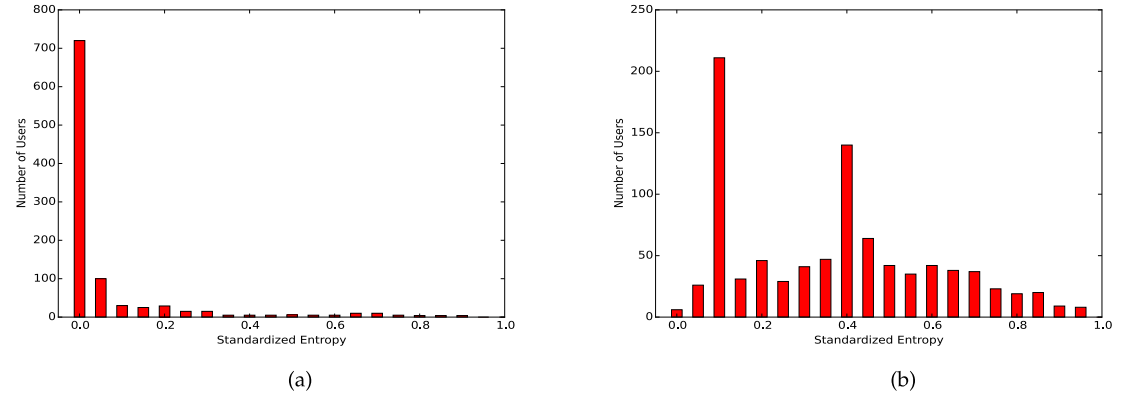


FIGURE 13. The distribution of standardized entropy for Weibo users in the global and local clusters. (a) Global cluster. (b) Local cluster.

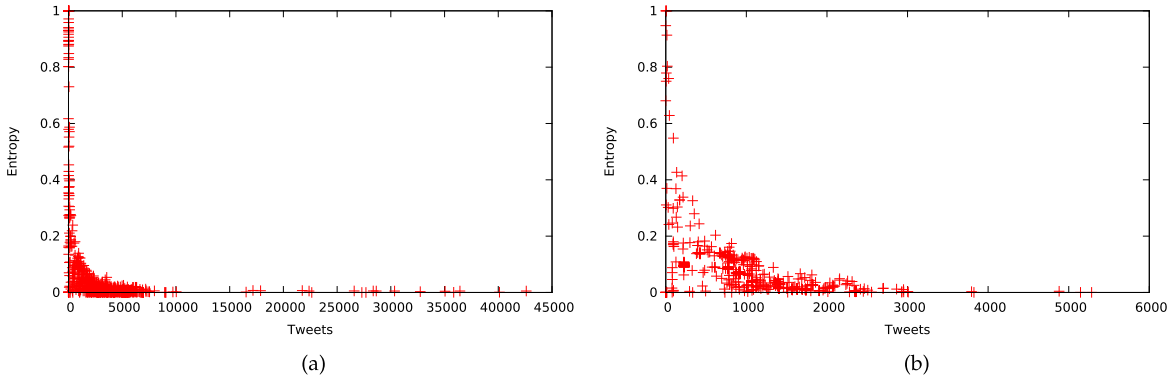


FIGURE 14. The correlation between the number of tweets and the standardized entropy of inter-tweet intervals. (a) Global cluster. (b) Local cluster.

the average number of tweets and retweets for females and males in the two clusters.

After the analysis, we obtain Fig. 17. It is clear that both genders are more likely to tweet than retweet in both the global and local clusters. In the local cluster, users tend to retweet more than users in the global cluster. This is particularly the case for females. Finally, we further study the daily variation in tweeting and retweeting for females and males. Here, we use a time bin of half an hour, which means that every half hour we count the number of tweets and retweets.

The results are shown in Fig. 18, from which we can see that during a day, the most common times for tweeting and retweeting for males and females is very similar. The preferred time is between 9am and 12 midnight because

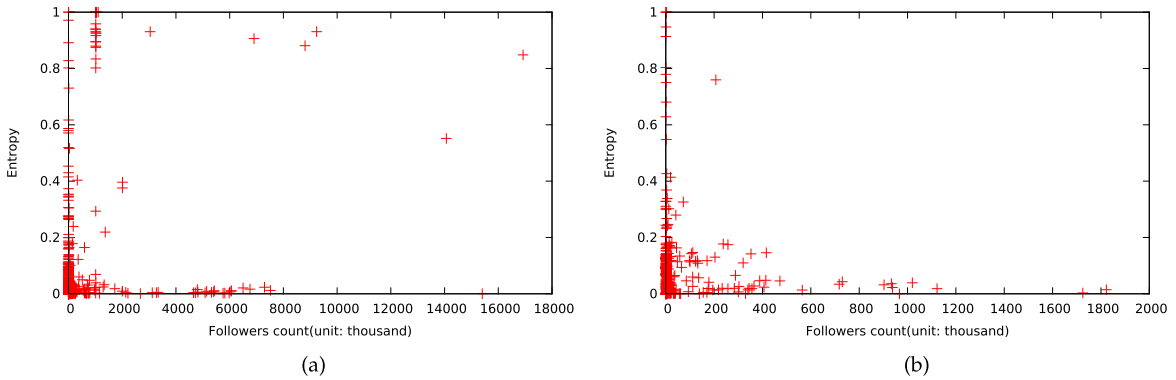


FIGURE 15. The correlation between the number of followers and the standardized entropies of inter-tweet intervals. (a) Global cluster. (b) Local cluster.

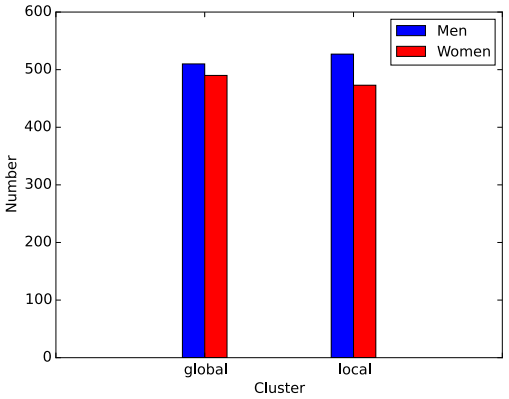


FIGURE 16. The number of males and females in the global and local clusters.

the remaining time is used for sleeping. While Fig. 18[a] illustrates that the tweeting and retweeting behavior for males and females is very similar over the day, Fig. 18[b] shows that after 18:00, females are more likely to retweet than males, and the peak time for retweeting for females is between 21:00 and 22:00.

This finding can improve the efficiency of message transmission, which has potential for enhancing the influence of advertising and to implementing a high-efficiency news diffusion system. As is well known, it is an efficient method to tweet advertisement information or news via Weibo users with many followers. However, if we also select users who have many female followers, the effect will be even greater. This is because female followers are more likely to forward tweets compared with male followers, which makes the influence of tweets more extensive.

VIII. CLUSTERING ANALYSIS OF WEIBO COMMUNITIES

The observations on the clustering patterns of user popularity and temporal tweeting lead us to explore clustering algorithms to partition Weibo users into distinct groups. In this study, we apply the k-means clustering algorithm due to its simplicity and computational efficiency [13], [14]. K-means is a widely used partitioning algorithm that groups N data points in k predefined subsets by minimizing the sum of the squared errors (SSE), i.e.,

$$\min \sum_{i=1}^{k} \sum_{x_{ij} \in c_i} dist(x_{ij} - \mu_i) \tag{4}$$

where k represents the total number of clusters and μ_i represents the geometric centroid of all data points in the cluster c_i. The distance measure, $dist(x_{ij} - \mu_i)$, denotes the distance between a data point x_{ij}, $1 \leq j \leq |c_i|$, in the cluster c_i and its cluster centroid μ_i. In other words, the algorithm attempts to group data points in nearby geometric space into the same clusters for minimizing their distances to the cluster centroids.

Before running k-means clustering algorithms, we first project the observations of tweeting behavioral patterns measured by standardized entropies, and user popularity, measured by the number of followers into a 2-D geometric space. Due to the heterogeneity of the units, we normalize user popularity into the range of [0, 1] with min-max normalization [14] to align with the units of standardized entropy. Thus the distance measure between the data points of two Weibo users a and b in this 2-D space becomes

$$dist(a, b) = \sqrt{|e_a - e_b|^2 + |p_a - p_b|^2}, \tag{5}$$

where e_a and e_b denote the standardized entropies of the temporal tweeting patterns of two users, while p_a and p_b denote the normalized user popularity of these two Weibo users.

For two independent user groups, i.e., the global cluster and the local cluster, we run the k-means clustering algorithm to search for k distinct partitions. Fig. 19 illustrates the decreasing trend of SSE as the value of k increases. However, when k reaches 9, the marginal benefits of SSE reduction is insignificant. Hence, we choose 9 as the optimal number of clusters for the global cluster. As illustrated in Fig. 19, the k clusters are able to partition data points in close geometric space into the same clusters, represented by different colors.

To gain a deeper understanding of these sub-clusters of Weibo users, we analyze a variety of tweeting behavioral

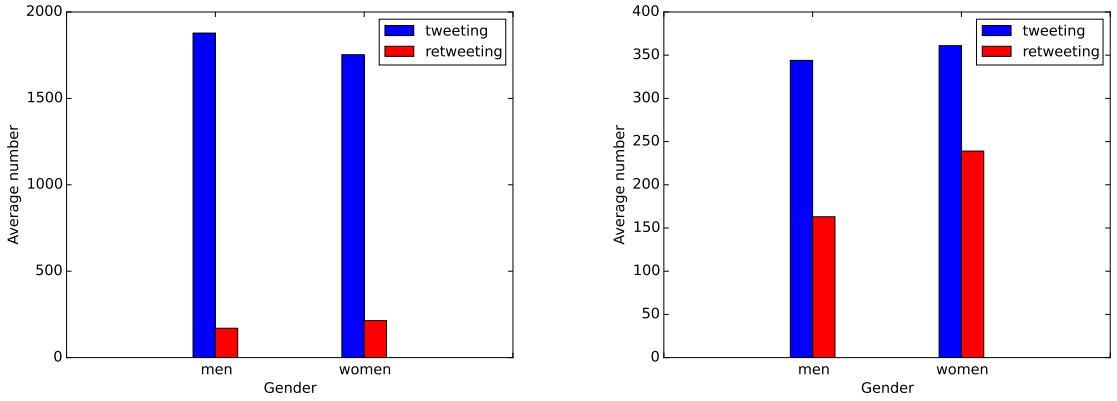


FIGURE 17. The average number of tweets and retweets in the global and local clusters. (a) Global cluster. (b) Local cluster.

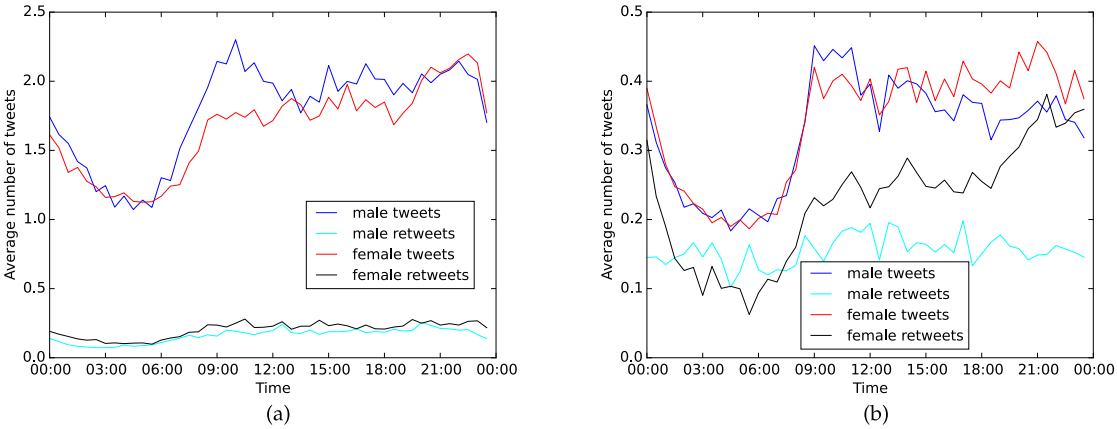


FIGURE 18. Trend for different genders in tweeting and retweeting in the global and local clusters. (a) Trend for different genders in tweeting and retweeting in the global cluster. (b) Trend for different genders in tweeting and retweeting in the local cluster.

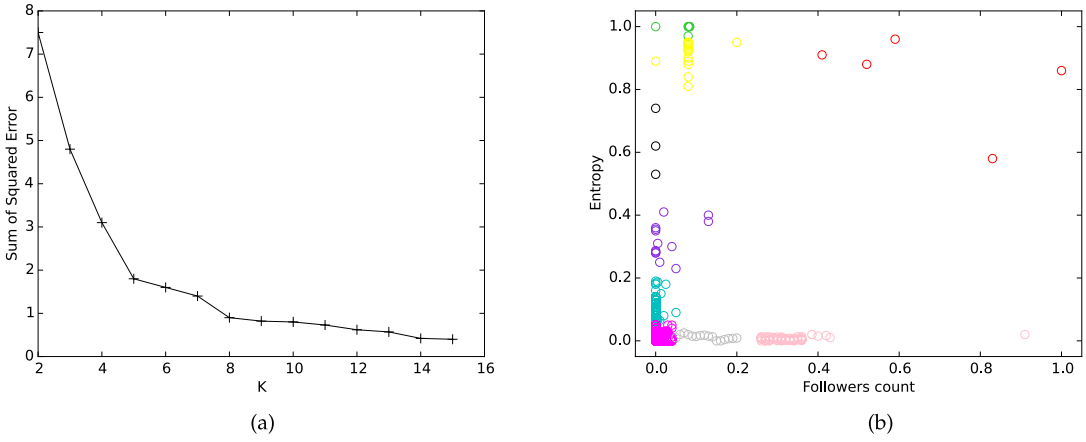


FIGURE 19. Determining the optimal k as the number of clusters and the clustering results of 9 clusters for the global cluster. (a) The quality of clustering measured by SSE. (b) Clustering results in the global cluster.

characteristics for each sub-cluster including user popularity, temporal tweeting patterns, following activities, tweeting frequency and retweeting behavior. Table 1 summarizes these tweeting behavioral characteristics for each of the 9 sub-clusters within the global cluster. As shown in Table 1, the clustering algorithm indeed partitions

TABLE 1. Tweeting behavior characteristics for 9 clusters for the global cluster.

Cluster ID	number of nodes	average followers	average entropy	average friends	average daily tweets	average daily retweets
1	15	464059	0.3083	975	4	2
2	43	66188	0.0989	1213	40	4
3	20	970136	0.9979	318	0	0
4	588	49791	0.0081	1296	210	19
5	3	17742	0.6212	535	0	1
6	5	11187749	0.8236	1235	1	0
7	16	1826881	0.0125	230	213	2
8	14	1088204	0.8988	443	0	0
9	19	6237478	0.0086	768	193	0

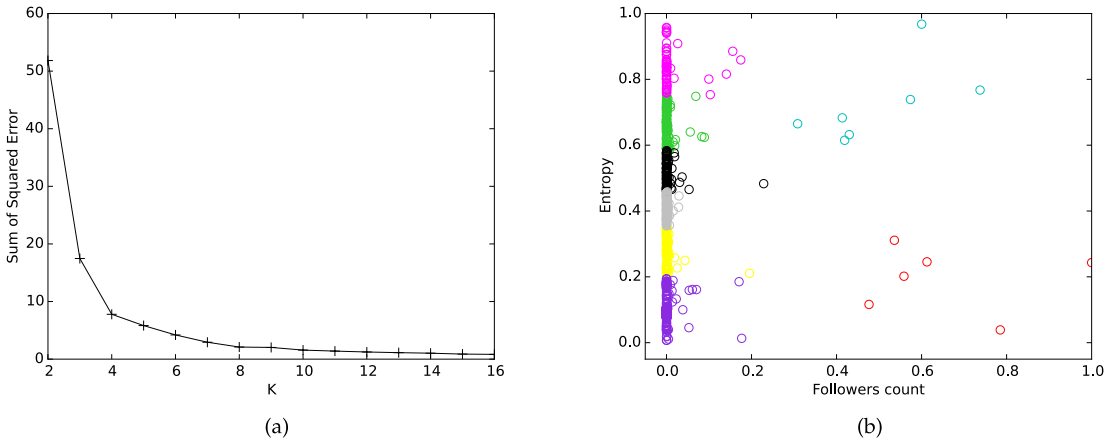


FIGURE 20. Determining the optimal k as the number of clusters and the clustering results of 8 clusters for the local cluster. (a) The quality of clustering measured by SSE. (b) Clustering results in the local cluster.

Weibo users into groups with distinct tweeting behavioral characteristics.

Similarly, we find the optimal number of sub-clusters within the local cluster is 8, based on the same method of varying values of k and monitoring the corresponding SSE values in Fig. 20[a]. The 8 sub-clusters are also illustrated in Fig. 19[b] using different colors. Table 2 summarizes the distinct tweeting behavioral characteristics of the 8 sub-clusters within the local cluster. As with the experimental results from the global cluster, these clusters also show very distinct tweeting behavioral characteristics. These results confirm that clustering successfully partitions Weibo users into separate groups with distinct tweeting behaviors. More importantly, partitioning users into small clusters could improve our understanding of the similarity and differences among Weibo users' tweeting behavior.

Clustering Weibo users into different groups creates opportunities for in-depth studies of user interests. For example, we can divide users into groups who are interested in food or fashion clothes etc. This will be helpful in recommending products of interest to the appropriate people. Further, by combining with the analysis of the small-world, we can enhance the acceptability and interpretability of recommendations by displaying the welcome result of other users in the small-world.

IX. RELATED WORK

In recent years, popular microblogging sites such as Twitter and Sina Weibo have become influential social media for information distribution, viral marketing and political campaigns. Thanks to its wide popularity in the United States and Europe, Twitter has received significant attention from the research community. A rich body of research has been devoted to study the topological characteristics and social graphs [15], [16]. References [17]–[19] utilized nonnegative matrix factor(NMF) combined topological characteristics and context to identify of community structures while [20], [21] utilized probabilistic graphical model(PGM). References [22] and [23] proposed a method to detect overlapping communities based on information of nodes and edges. Reference [24] leveraged a semi-supervised deep learning method to detect community and [25] further incorporated prior knowledge into the community discovery process.

Sina Weibo has been used by millions of internet users in China on a daily basis, and there has been some work that has studied the topological characteristics [1], user verification [26], [27], rumor detection [28], and information cascading over Weibo online social network for breaking earthquake news [29]. References [1] and [8] characterizes information spreading and Weibo user profiles in terms of

TABLE 2. Tweeting behavior characteristics for 8 clusters for the local cluster.

Cluster ID	number of nodes	average followers	average entropy	average friends	average daily tweets	average daily retweets
1	6	3763127	0.19281654	748	27	0
2	118	21964	0.66424584	633	11	3
3	109	18268	0.2863817	674	28	17
4	109	25618	0.5133685	600	15	7
5	303	14213	0.10462508	922	13	10
6	7	2830367	0.7241048	1130	16	4
7	63	66505	0.84034234	351	7	1
8	199	9302	0.40505964	476	17	8

genders and geographical location, the social attributes of followers, friends and reciprocal relationships. Reference [6] modeled user interactions on Sina Weibo and Renren, revealing that Sina Weibo is more efficient for information diffusion compared with Renren. However, there is relatively little work that has studied the behavior of its users. Reference [8] mainly analyzed users' reciprocal rate, dynamic changes, tags, collected tweets and [30] conducted a measurement study on the patterns of video tweeting over the Sina Weibo system and observed frequent flash crowds for popular videos due to massive numbers of retweets of the videos.

Sampling is a widely used technique to analyze and process vast amounts of data such as internet traffic [31] and online social networks [32]–[35]. Given the sheer number of users and their relationships as well as the massive data streams consisting of real-time tweets, retweets, and comments, several recent studies have advocated the adoption of sampling for data collection in online social networks. For example, [32] characterizes social network activity using a graph with a streaming time node sampling algorithm while [33] develops a Metropolis-Hastings random walk and a re-weighted random walk algorithm to collect unbiased samples from Facebook. Similarly, [34] uses a heuristics approach based on a stratified weighted random walk to sample a node set from large unknown graphs such as online social networks to estimate certain metrics of interest, while the study in [35] introduces sampling-based algorithms to obtain near-uniform random samples of a given user's local neighborhood.

A unique feature of Sina Weibo, user verification, has been studied in [26], which finds that the verification mechanism builds trust between users and their followers, and establishes the authenticity of the sources of original posts, thus encouraging more user interactions. In addition, [27] analyses the social network graphs of over 37 thousand verified Weibo users and reveals power-law distributions in the following relationships among these users. Reference [29] analyzed various aspects of information diffusion over Sina Weibo based on a case study of the 2010 Yushu earthquake event. To detect rumors on Sina Weibo, [28] develops an automated approach based on the client programs used and the event locations. Different from these studies on Sina Weibo, this paper focuses on user interactions and understanding user behavior in tweeting, retweeting and commenting.

X. CONCLUSIONS AND FUTURE WORK

As Sina Weibo has become an influential social media platform for millions of internet users in China, it has become extremely important to understand the tweeting behavior of its users, which is very useful for understanding the mechanisms of information dissemination. In this paper, according to the following relationships, we crawl all the Weibo data from Shenzhen Weibo users and use the Sina Weibo API to collect real-time sampled statuses of all its users. With the Shenzhen Weibo user data, we establish a community social network graph to reveal Weibo users' following and follower relationships and analyze the graphical structure. In order to better understand the impact of user relationships on tweeting, retweeting and commenting behaviors, we conduct a correlation analysis on user's popularity and tweet influence. We also characterize the similarity of following patterns and user interactions within the local Weibo community, which helps us to reveal the influence of user popularity on information spreading over online social networks and discover the similarity of users' retweeting and commenting activities via exploiting the interactions between users and tweets. In order to further research into the temporal tweeting behavior of Weibo users, we develop a community approach and an inference algorithm to analyze the public sampled status of two independent groups and their complete tweets. Then we study the properties of tweeting and retweeting by gender and find that females have higher possibility to retweet than males, which is useful for improving the efficiency of message transmission. By basing on an information-theoretic measure of entropy, we characterize the temporal tweeting behavior and analyze its correlation with user popularity. This reveals interesting patterns of clustering among Weibo users and leads us to apply a k-means clustering algorithm to partition users into distinct clusters with diverse tweeting behavioral characteristics including user popularity, temporal tweeting patterns, tweeting frequency and retweeting activities, which can improve the efficiency of recommendations by combining with the analysis of the small-world.

Our future work lies in analyzing the unstructured contents of tweeting messages posted by Weibo users for gaining an

in-depth understanding of the behavior, interests and intents of Weibo users.

REFERENCES

[1] Z. Guo, Z. Li, and H. Tu, "Sina microblog: An information-driven online social network," in *Proc. Int. Conf. Cyberworlds*, Oct. 2011, pp. 160–167.

[2] J. Cheng, D. M. Romero, B. Meeder, and J. Kleinberg, "Predicting reciprocity in social networks," in *Proc. IEEE Conf. Social Comput.*, Oct. 2011, pp. 49–56.

[3] J. Yang and J. Leskovec, "Community-affiliation graph model for overlapping network community detection," in *Proc. Int. Conf. Data Mining (ICDM)*, Dec. 2012, pp. 1170–1175.

[4] Z. Guo, Z. Li, H. Tu, and D. Xie, "Detecting and modeling the structure of a large-scale microblog," in *Future Information Technology, Application, and Service* (Lecture Notes in Electrical Engineering), vol. 164. Germany: Springer, Jan. 2012, pp. 151–160.

[5] F. Wang, H. Wang, K. Xu, J. Wu, and X. Jia, "Characterizing information diffusion in online social networks with linear diffusive model," in *Proc. IEEE Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jul. 2013, pp. 307–316.

[6] J. Lin, Z. Li, D. Wang, K. Salamatian, and G. Xie, "Analysis and comparison of interaction patterns in online social network and social media," in *Proc. Int. Conf. Comput. Commun. Netw. (ICCCN)*, Jul./Aug. 2012, pp. 1–7.

[7] G. Hao, L. Yu-Liang, W. Yu, and Z. Tong-Tong, "Measurement of the weibo hall of fame network," in *Proc. Int. Conf. Instrum., Meas., Comput., Commun. Control*, Oct. 2011, pp. 192–195.

[8] Z. Guo, Z. Li, H. Tu, and L. Li, "Characterizing user behavior in weibo," in *Proc. Int. Conf. Mobile, Ubiquitous, Intell. Comput.*, Jun. 2012, pp. 60–65.

[9] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proc. Internet Meas. Conf.*, 2007, pp. 29–42.

[10] K. Lei, K. Zhang, and K. Xu, "Understanding sina weibo online social network: A community approach," in *Proc. IEEE GLOBECOM*, Dec. 2013, pp. 3114–3119.

[11] J.-L. Guillaume and M. Latapy, "Bipartite graphs as models of complex networks," *Phys. A, Stat. Mech. Appl.*, vol. 371, no. 2, pp. 795–813, 2006.

[12] J. J. Ramasco, S. N. Dorogovtsev, and R. Pastor-Satorras, "Self-organization of collaboration networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 70, no. 3, p. 036106, 2004.

[13] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, 1967, pp. 281–297.

[14] H. Jiawei, K. Micheline, and P. Jian, *Data Mining: Concepts and Techniques*, 3rd ed. San Mateo, CA, USA: Morgan Kaufmann, 2011.

[15] S. K. Bisma and M. A. Niazi. (2017). "Network community detection: A review and visual survey." [Online]. Available: https://arxiv.org/abs/1708.00977

[16] H. Fani and E. Bagheri, "Community detection in social networks," in *Encyclopedia With Semantic Computing and Robotic Intelligence*, vol. 1. Singapore: World Scientific, 2017, p. 1630001.

[17] X. Wang, D. Jin, X. Cao, L. Yang, and W. Zhang, "Semantic community identification in large attribute networks," in *Proc. 13th AAAI Conf. Artif. Intell. (AAAI)*, 2016, pp. 265–271.

[18] X. Wang, P. Cui, J. Wang, J. Pei, W. Zhu, and S. Yang, "Community preserving network embedding," in *Proc. 31st AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 203–209.

[19] M. Qin, D. Jin, D. He, B. Gabrys, and K. Musial, "Adaptive community detection incorporating topology and content in social networks," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, 2017, pp. 675–682.

[20] Z. Xu, Y. Ke, Y. Wang, H. Cheng, and J. Cheng, "A model-based approach to attributed graph clustering," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2012, pp. 505–516.

[21] D. He, Z. Feng, D. Jin, X. Wang, and W. Zhang, "Joint identification of network communities and semantics via integrative modeling of network topologies and node contents," in *Proc. 31st AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 116–124.

[22] D. Jin, H. Wang, J. Dang, D. He, and W. Zhang, "Detect overlapping communities via ranking node popularities," in *Proc. 13th AAAI Conf. Artif. Intell. (AAAI)*, 2016, pp. 172–178.

[23] D. Jin, B. Gabrys, and J. Dang, "Combined node and link partitions method for finding overlapping communities in complex networks," *Nature Sci. Rep.*, vol. 5, Feb. 2015, Art. no. 8600.

[24] L. Yang,, X. Cao, D. He, C. Wang, X. Wang, and W. Zhang, "Modularity based community detection with deep learning," in *Proc. 25th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2016, pp. 2252–2258.

[25] M. Akbari and T.-S. Chua, "Leveraging behavioral factorization and prior knowledge for community discovery and profiling," in *Proc. 10th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2017, pp. 71–79.

[26] J. Chen and J. She, "An analysis of verifications in microblogging social networks—Sina weibo," in *Proc. Int. Conf. Distrib. Comput. Syst. Workshop Hot Topics Peer-Peer Comput. Online Social Netw. (HotPOST)*, Jun. 2012, pp. 147–154.

[27] G. Hao, L. Yu-Liang, W. Yu, and Z. Tong-Tong, "Measurement of the weibo Hall of fame network," in *Proc. Int. Conf. Instrum., Meas., Comput., Commun. Control*, Oct. 2011, pp. 192–195.

[28] F. Yang, Y. Liu, X. Yu, and M. Yang, "Automatic detection of rumor on sina weibo," in *Proc. ACM SIGKDD Workshop Mining Data Semantics*, Aug. 2012, Art. no. 13.

[29] Y. Qu, C. Huang, P. Zhang, and J. Zhang, "Microblogging after a major disaster in China: A case study of the 2010 Yushu earthquake," in *Proc. ACM Conf. Comput. Supported Cooperat. Work*, 2011, pp. 25–34.

[30] Z. Guo, J. Huang, J. He, X. Hei, and D. Wu, "Unveiling the patterns of video tweeting: A sina weibo-based measurement study," in *Proc. Passive Active Meas. Conf.*, Mar. 2013, pp. 166–175.

[31] N. Duffield, "Fair sampling across network flow measurements," in *Proc. 12th ACM SIGMETRICS*, Jun. 2012, pp. 367–378.

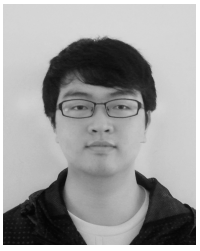
[32] N. K. Ahmed, F. Berchmans, J. Neville, and R. Kompella, "Time-based sampling of social network activity graphs," in *Proc. Mining Learn. Graphs Workshop*, Aug. 2010, pp. 1–9.

[33] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in Facebook: A case study of unbiased sampling of OSNs," in *Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–9.

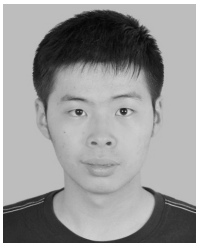
[34] M. Kurant, M. Gjoka, C. T. Butts, and A. Markopoulou, "Walking on a graph with a magnifying glass: Stratified sampling via weighted random walks," in *Proc. ACM SIGMETRICS*, Jun. 2011, pp. 281–292.

[35] M. Papagelis, G. Das, and N. Koudas, "Sampling online social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 662–676, Mar. 2013.

KAI LEI received the B.Sc. degree in computer science from Peking University, China, in 1998, the M.Sc. degree in computer science from Columbia University in 1999, and the Ph.D. degree in computer science from Peking University in 2015. He was with several companies, including IBM T. J. Waston Research Center, Citigroup, Oracle, and Google, from 1999 to 2004. He is currently an Associate Professor with the School of Electronic and Computer Engineering, Peking University. His research interests include knowledge graph, named data networking, and blockchain.



YING LIU received the B.S. degree in computer science from the Dalian University of Technology, China, in 2015. He is currently pursuing the M.S. degree in computer science with Peking University. His research interests include social networks, natural language processing, and data mining.



SHANGRU ZHONG received the B.S. degree in information engineering from the South China University of Technology, China, in 2015. He is currently pursuing the M.S. degree in computer science with Peking University. His research interests include data mining, recommender system, and social networks.

YONGBIN LIU received the B.S. degree in computer science from the South China University of Technology, China, in 2014. He is currently pursuing the M.S. degree in computer science with Peking University. His research interests include computer networks, named data networking, and data mining.

KUAI XU received the B.S. and M.S. degrees in computer science from Peking University, China, in 1998 and 2001, respectively, and the Ph.D. degree in computer science from the University of Minnesota in 2006. He is currently an Associate Professor of applied computing program with the School of Mathematical and Natural Sciences, New College of Interdisciplinary Arts and Sciences, Arizona State University. His research interests include network security, network measurement and analysis, cloud computing, home networks, and online social networks.

YING SHEN received the Erasmus Mundus M.S. degree in natural language processing from the University of Franche-Comt, France and University of Wolverhampton, England the another master's degree in linguistics from Peking University, China, and the Ph.D. degree from the University of Paris Ouest Nanterre La Dfense, France, specialized in medical and biomedical information science. She is currently an Assistant Professor with the School of Electronics and Computer Engineering, Peking University. Her research interests include clinical decision support, knowledge representation, knowledge reasoning, big data processing, and so on.

MIN YANG received the B.S. degree from Sichuan University in 2012 and the Ph.D. degree from The University of Hong Kong in 2017. She is currently an Assistant Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Science. Her current research interests include machine learning, deep learning, and natural language processing.