Front. Comput. Sci.
DOI 10.1007/s11704-016-5198-y

RESEARCH ARTICLE

The power of comments: fostering social interactions in microblog networks

Tianyi WANG[1,2,3], Yang CHEN (✉)[4], Yi WANG[5], Bolun WANG[3], Gang WANG[3], Xing LI[1,2], Haitao ZHENG[3], Ben Y. ZHAO[3]

1 Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
2 Tsinghua National Laboratory for Information Science and Technology, Beijing 100084, China
3 Department of Computer Science, University of California, Santa Barbara, CA 93106-5110, USA
4 School of Computer Science, Fudan University, Shanghai 201203, China
5 Department of Mathematics, Syracuse University, NY 13244-1150, USA

Abstract Today's ubiquitous online social networks serve multiple purposes, including social communication (Facebook, Renren), and news dissemination (Twitter). But how does a social network's design define its functionality? Answering this would need social network providers to take a proactive role in defining and guiding user behavior.

In this paper, we first take a step to answer this question with a data-driven approach, through measurement and analysis of the Sina Weibo microblogging service. Often compared to Twitter because of its format, Weibo is interesting for our analysis because it serves as a social communication tool and a platform for news dissemination, too. While similar to Twitter in functionality, Weibo provides a distinguishing feature, comments, allowing users to form threaded conversations around a single tweet. Our study focuses on this feature, and how it contributes to interactions and improves social engagement. We use analysis of comment interactions to uncover their role in social interactivity, and use comment graphs to demonstrate the structure of Weibo users interactions. Finally, we present a case study that shows the impact of comments in malicious user detection, a key application on microblogging systems. That is, using properties of comments significantly improves the accuracy in both modeling and detection of malicious users.

Keywords microblogs, comments, social and interaction graph, user behavior

1 Introduction

Today's online social networks (OSNs) pervade through all aspects of our daily lives, serving a variety of functions from social communication (Facebook, Renren), to information and news dissemination (Twitter), and to professional development (LinkedIn).

While some networks are designed with specific usage scenarios in mind, e.g., LinkedIn and Pinterest, others are more general and support a variety of usage-agnostic features such as making friends, messaging, and content sharing. Even within these general frameworks, recent work has shown that networks can evolve and become more specialized along specific usage scenarios. For example, Twitter is often considered more as a news media platform than a social network [1].

While a variety of factors clearly contribute to the formation of user behavior, can we determine what role is specific as user features are played in this process? The answer to this question can reveal the potential impacts of social networking features on user behavior, and whether OSN providers can

Received May 20, 2015; accepted October 29, 2015

E-mail: chenyang@fudan.edu.cn

proactively "guide" user behavior by introducing features or interface modifications.

In this paper, we first present efforts to answer this question using an empirical, data-driven approach. More specifically, we examine this question through detailed measurements and analysis of Sina Weibo, a social "microblogging network". The analysis of Sina Weibo provides an interesting case study because it is very similar to Twitter in nearly all aspects of its basic functionality, but is often viewed by its users as a hybrid network for news dissemination and social interactions [2]. This is somewhat surprising, since social media tools in China largely mirror the functionality and usage patterns of their relevant international counterparts. For example, Renren [3] provides near-identical functionality to Facebook.

Through our work, the first goal is to understand how Sina Weibo and Twitter actually differ via quantitative metrics. We then dive deeper to determine if these differences can be partially attributed to any individual user feature (or design choice). More specifically, our analysis focuses on Sina Weibo's comment feature[1)], which is considered as one of the most distinguishing features between Sina Weibo and Twitter by many analysts[2)3)]. To demonstrate how comments are correlated with such differences, we show how it contributes to social interactions in Sina Weibo at the level of individual users, and how it shapes user interaction patterns similar to those of Facebook at macroscopic network levels. Finally, we also evaluate the importance of the comment feature at the application level.

Our work makes four key contributions.

First, we use a large-scale dataset to quantify the difference in network structure between Sina Weibo and Twitter. Sina Weibo exhibits significant structural differences: not only are each Sina Weibo user's incoming and outgoing links more balanced, but a much larger portion of each user's relationships are bidirectional, i.e., consistent with friendships in social networks such as Facebook.

Second, we analyze the correlation between user comments and users' social interactions at the individual user level. We find comment is a more prevalent type of interactions in Sina Weibo than reposts and mentions. We also find that users who comment (commentors) are frequently friends (bi-directional social link) with the tweet author, and they usually form concentrated conversations. The observations demonstrate that comments contribute to social interactions among users, and are highly correlated with users' bidirectional friendships.

Third, we further explore this issue at the macroscopic level, by analyzing and comparing the structures of Sina Weibo's comment graph, Sina Weibo's repost graph and Facebook and Twitter interaction graphs. We find that Sina Weibo's comment graph most closely resembles Facebook's interaction graph despite the different user populations on the two sites, which implies that comment interactions are strongly indicative of bidirectional friendship interactions. In addition, we find significant correlations (overlaps) between Sina Weibo's comment graph and social graph, and comment graph acts as a good predictor of social links. These confirm our intuition that comments play an important role in bidirectional friendships at the macroscopic level.

Finally, at the application level, we use a case study to quantify the impact of comments on analyzing and modeling user behavior. Specifically, we look at malicious user detection, a critical application for all online social networks. We apply machine-learning tools to build a detector of malicious user activity, and apply it to ground-truth data from Sina Weibo. We find that compared to commonly used features, features based on comment activity offer much higher discriminatory power in modeling the difference between malicious and normal users. Not only are comment-based features sufficient to produce an accurate detector, but when added to common features, they significantly elevate the accuracy of the resulting detector.

To the best of our knowledge, our work is the first one to explore the impact of comments on user behavior in social networks. The results of our work show that single feature like comments can contribute to significantly higher levels of interaction between different users.

2 Background and methodology

We briefly describe Sina Weibo and our dataset to provide the background information for our analysis. We first discuss common features that Sina Weibo shares with Twitter, and highlight its unique features. We then explain our data collection methodology and present high-level statistics of our dataset. Finally we introduce the methodology of this paper.

2.1 Sina Weibo

[1)] Comment allows users to create threads of comments centered around a single tweet or microblog. For simplicity, we refer to microblogs in both platforms as tweets

[2)] Twitter vs Sina Weibo: eight things Twitter can learn from the latter. http://www.hongkiat.com/blog/ things-twitter-can-learn-from-sina-weibo/

[3)] Twitter vs Sina Weibo: differences. http://www.juanmarketing. com/twitter-vs-weibo-differences/2011/05/24/

Sina Weibo is the largest microblogging service in China, and the second largest microblogging service in the world. As of March 2013, it has more than 500 million registered users[4] and generates more than 100 million tweets per day[5]. Sina Weibo is popular around the world with many international users like Brazilian football player Pele and organizations like the Uinted States (UN).

Sina Weibo shares many features with Twitter. Users can post tweets with up to 140 Chinese characters or 280 English characters, and repost (retweet) others' tweets. Each tweet can tag specific topics, mention other users by using an "@", and post short uniform resource locations (URLs), geographic information and even pictures. A user can subscribe to other users' tweets by following these users. If user A follows user B, we say that A is B's follower, and B is A's followee. Sina Weibo tweets are public to its registered users, although the platform only places each user's followees' tweets on her timeline. By default, Sina Weibo users can manually visit any user's home page, which contains the user's profile and published tweets.

The most notable feature that distinguishes Sina Weibo from Twitter is the comment feature[6]. By default, a Sina Weibo user can comment directly on any published tweet, and reply to any comment. This is a broad type of interaction, since users do not need to follow the author and commentors of a tweet before interacting directly with them. A comment does not create new tweet, but is associated with the tweet in a comment list, which includes all comments and replies sorted by time. The threaded comments make it easier for users to have concentrated conversations with both tweet authors and other commentors within the same tweet. As described in the next section, the threaded comment list also allows us to crawl comments efficiently.

The comment function in Sina Weibo is significantly different from the reply function in Twitter. In Sina Weibo, all comments on a tweet are associated with the original tweet. Thus users can easily view past comments and add follow-on comments. In contrast, the reply function in Twitter will generate an independent tweet, which makes it more difficult to trace back all replies from a new reply[7]. The result is a much more strongly threaded sequence of messages into conversations.

2.2 Datasets

To crawl Sina Weibo, we use its open application programming interface (API) to access user profiles, tweets and comments. The API provides a user's complete list of followees, the latest 2 000 tweets, and up to 5 000 followers. Our crawls created two Sina Weibo datasets.

Crawling the social graph Obtaining an unbiased sample of the Sina Weibo social graph is nontrivial. An unbiased sample is desired because it would capture the graph properties (e.g., degree distribution) while making the graph size small. In an unbiased dataset, each node in the graph is sampled with the same probability. Conventional algorithms like breadth-first sampling (BFS) and random walk are known to be biased towards high degree nodes. That is, the users with high degree are more likely to be sampled. Existing unbiased sampling methods [4, 5] require the complete follower/followee set for each user, which is limited by the fact that Sina Weibo API returns at most 5 000 followers.

Instead, we seed our crawl using a large number of randomized user IDs. We leverage the fact that Sina Weibo's API provides fast access to public tweets, and each tweet contains IDs of its author and mentioned users. We have performed an API call once every three seconds for one month and obtained roughly 60 million unique user IDs [8]. We then crawl the Sina Weibo network using these IDs as seeds, and obtain 57.1 million user profiles, each profile containing the number of the user's followers, followees and friends[9].

Crawling reposts and comments We also crawl a smaller set of the tweets for detailed repost and comment data by accessing to the comment and repost lists of a tweet. Since crawling reposts or comments requires multiple API requests per tweet, we have to reduce the size of the targeted crawl to limit load on Sina Weibo servers. In addition, we need a connected subgraph of users in order to analyze interactions among them. Thus we use BFS to obtain a connected subgraph. While BFS can introduce bias in node selection, it is attractive under our scenario because it is efficient and provides a direct comparison to prior work on user interactions [1, 6] that also use BFS. In total, we have obtained 61.5 million tweets from 723 thousand users, with 118.1 million comments and 86.2 million reposts.

To the best of our knowledge, our datasets are the most

[4] Sina Weibo has more than 500 million registered users. http://news.xinhuanet.com/newmedia/2013-02/21/c-124369896. htm
[5] Sina Weibo has more than 100 milion tweets per day. http://www.washingtonpost.com/blogs/worldviews/wp/2013/03/08/how-china-censors-100-million-tweets-per-day/
[6] The comment feature has been in Sina Weibo since its inception
[7] Twitter supports the function of displaying the full conversations in Aug., 2013. https://blog.twitter.com/2013/keep-up-with-conversations-on-twitter
[8] Each API request returns 200 public tweets
[9] When users A and B follow each other in a bidirectional social link, we call them friends

comprehensive sample of Sina Weibo social network to date. More importantly, while the prior efforts focused on the social graph and tweet content [2, 7–10], our datasets contain all reposts and comments for 61.5 million tweets.

2.3 Our methodology

We seek potential factor(s) that drive Sina Weibo, a microblogging network, into a hybrid platform for both news dissemination (like Twitter) and social friending and communication (like Facebook). We reveal this by demonstrating how comments contribute to social interactions and friending in Sina Weibo.

We start our analysis by quantifying the differences between Sina Weibo and Twitter in this section. Particularly, by comparing the follower-followee relationship between users, we find Sina Weibo users not only have more balanced incoming and outgoing links, but maintain a much larger portion of bidirectional relationships, i.e., resembling the friendships in traditional social networks such as Facebook.

These results motivate us to understand what function (or design choice) is correlated with the difference. We draw our attention to the comment function in Sina Weibo. We focus on comments for two reasons. First, comment is a key feature that distinguishes Sina Weibo from Twitter, and it is also a common function supported by traditional social networks, e.g., Facebook. Second, comment is one of the most heavily used communication channels by Sina Weibo users. Users generate an order of magnitude more comments than tweets [11].

We study Sina Weibo comment and how it contributes to social interactions and friending from three levels: individual user level (Section 4), macroscopic network level (Section 5) and application level (Section 6). We aim to give a comprehensive view on how this single design choice impacts on user's way of using Sina Weibo.

Individual user level We start from individual user's perspective to understand the usage patterns of Sina Weibo comments. More specifically, we analyze the popularity of comment as a communication channel in comparing with other channels such as tweet and repost. In addition, we analyze the relationship between users who participant in comment interactions, and explore whether users who comment are likely to be friends (bidirectional social link).

Macroscopic network level We then move to a macroscopic view, by building network-wide interaction graphs, and comparing them to interaction graphs on Facebook and Twitter. We seek whether it is the comment (or repost) function that defines users' interaction patterns (regardless of user background and culture). We compare Sina Weibo's comment graph with that of Facebook, and Sina Weibo's repost graph with that of Twitter, to explore their network-level similarities. In addition, we explore the correlation between comments and friendship at the network level by analyzing the overlaps between Sina Weibo's comment graph and social graph. We also study the ability of comment graph to predict potential social links.

Application level Following observations that comments significant impact user behavior both at individual and network levels, we further study the importance of comment actions in user's overall behavioral profiles, and explore using comments to augment user-behavior based applications. In this paper we study two applications. First, we study machine-learning (ML) detectors of malicious users. Second, we study the influence maximization in social graph. For both applications, we do not focus on proposing any new methods or algorithms, instead. We introduce comment, and evaluate how comment improves existing methods.

3 Sina Weibo versus Twitter

We start from a high-level comparison of Sina Weibo and Twitter in terms of the social graph, focusing on the follower-followee relationship between users. We examine user degree distributions and the ratio of relationship links that are bidirectional.

Our analysis uses our crawled and anonymized Sina Weibo social graph, and an anonymized Twitter graph from Ref. [1] (41.7 million user profiles and 1.47 billion following relationships). Note that the Twitter graph is a complete crawl, and is unbiased and comparable to our Sina Weibo data.

Degree distribution In both Twitter and Sina Weibo social graphs, a node's in-degree represents the number of followers of the user, and the out-degree represents the number of followees. In Figs. 1 and 2 we plot the complementary cumulative distribution function (CCDF) of the follower and followee counts respectively. We make two key observations.

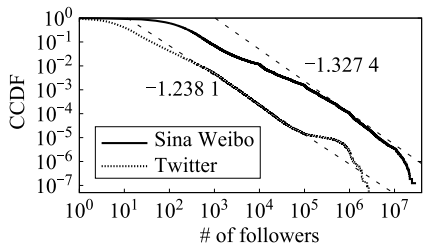


Fig. 1 Followers distribution

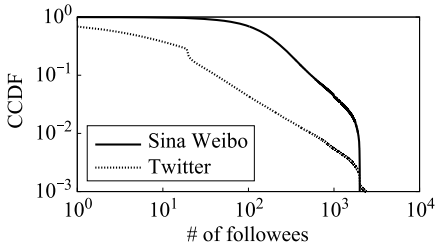


Fig. 2 Followees distribution

First, in terms of the follower count, the two networks display a similar pattern: the lines follow a commonly observed power-law distribution ($P(k) \propto k^{-\alpha}$) [1, 12], with similar scaling parameters [13] (α = 1.327 4 for Sina Weibo and 1.238 1 for Twitter). Second, more interestingly, the two networks differ significantly in terms of the followee count (out-degree). In Sina Weibo, most users follow 10^1–10^2 users, while a significant portion of Twitter users follow no more than one user. One intuitive explanation is that Sina Weibo users use it to communicate with friends, and hence every user has the minimum number of users/friends who follow. To verify this intuition, we will check the balance between followers and followees and reciprocity in later sections.

Balance between followers and followees To further study each individual user's follower/followee behavior, we define a balance metric, which is the ratio of a user's follower count to his followee count, i.e., a user's in-degree divided by her out-degree. Since a fully symmetric social network has a balance value of 1 for all users, for our analysis we define well balanced users to be those whose balance ratio is between 0.5 to 2.

We find that a higher fraction of users in Sina Weibo belong to the balanced user category. Specifically, 57.4% of Sina Weibo users are well balanced, while the ratio drops to 49% in Twitter. More broadly speaking, 80% of Sina Weibo users have a balance ratio between 0.1 and 2 while only 60% of Twitter users have. This again indicates that Sina Weibo as a whole is more similar towards symmetric social networks.

Reciprocity We also consider reciprocity, another widely used metric for quantifying the symmetric relationship between users. It is defined as the ratio of a user's friend count (the number of bidirectional links) to his followee count (out-degree). It holds a value between 0 and 1 — the higher the reciprocity is, the higher the fraction of friends in a user's relationship is.

Figure 3 plots the cumulative distribution function (CDF) of reciprocity for both Sina Weibo and Twitter. Sina Weibo has a higher level of reciprocity than Twitter as a whole. For Sina Weibo (Twitter), 0.55% (40%) users howe a zero reciprocity, and less than half (more than 80%) users have a reciprocity smaller than 0.5.

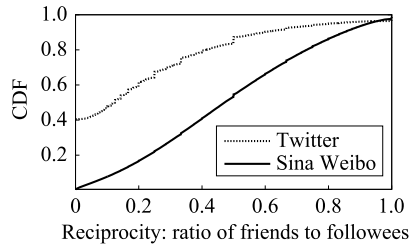


Fig. 3 Reciprocity distribution

From these results, we conclude that Sina Weibo has a much higher level of reciprocity than Twitter, and shows more signs of supporting symmetric social relationships. The results motivate us to seek pontential factors which correlate with such differences.

4 Comment analysis

With the differences between Twitter and Sina Weibo, our next goal is to find whether and how any design choice(s) in Sina Weibo lead(s) to such differences. As discussed earlier, we focus on "comment" feature, and look at the role of comments in the Sina Weibo network. In this section, we focus on analysis at the individual user level and answer the following questions.

- Are comments a popular channel of user interactions?
- What are the temporal properties of comments, e.g., when do comments arrive following tweets?
- Who posts comments? What are the relationships between the author of a tweet and users who comment on them (commentors)?
- Do comments (and their replies) form intense social interactions, e.g., conversations among users?

4.1 User interactions in Sina Weibo

There are three types of user interactions in Sina Weibo: repost, mention and comment. A user can repost an existing tweet, similar to retweet in Twitter, mention another user in a tweet, or comment on an existing tweet (or other comments). Among the three, the comment feature is unique in Sina Weibo. Our analysis in this subsection demonstrates that comment is the dominating form of user interactions in Sina Weibo.

We present two key results that demonstrate the popularity of comments among Sina Weibo users. The first result

shows, for each user, the total number of comments, reposts and mentions for his latest 2 000 tweets. Out of 723 thousand users, the majority (65.8%) received more comments than reposts, and more than 55% users received at least twice more comments than reposts. A more detailed result is in Table 1, which lists the statistics of each feature in terms of the 80th, 50th and 20th percentile values across all the users. For all three metrics, the value corresponding to comments is significantly higher than that of reposts and mentions, often by more than 50% (e.g., $157 > 40 \times (1 + 50\%)$). These results show that users are more inclined to use comments.

Table 1 Quantitative comparison among interactions

Percentile of user interaction distributions	Comment	Repost	Mention
The 80th percentile	157	40	11
The 50th percentile	21	13	3
The 20th percentile	3	2	0

The second result examines interactions for each tweet. We plot the number of reposts versus that of comments in a heat map (Fig. 4). For each point in the 2D plane, we count the number of the corresponding tweets and use the color to represent the tweet count. The darker the point is, the larger the number of tweets. For better visualization, we display only the significant part of the heat map by truncating it at 30 reposts and 50 comments[10]. We see that the black areas stretch widely along the x axis where there are considerably more comments than reposts.

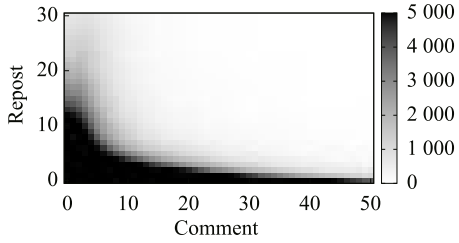


Fig. 4 Comment vs. repost

These results confirm that comments are much more popular than reposts and mentions in Sina Weibo. Next, we perform the detailed analysis on this unique feature to gain a better understanding of users' comment behavior.

4.2 Frequency and response

We begin by examining the temporal properties of comments. Specifically, we quantify the frequency, i.e., how often a user receives comments, and response, i.e., how fast the comments arrive after a tweet.

To measure the comment frequency, we calculate for each user the total number of comments he has received, normalized by the time span between his first and last tweets. We organize the results in terms of the average number of comments per week and show the CDF across all the users in Fig. 5. Among all the users, 46.1% receive at least 1 comment per week and 17.5% receive more than 10 comments per week (when $x = 1, y = 53.9\%; x = 10, y = 82.5\%$). As a reference, we also plot the CDF of the average number of tweets per week for each user. The fact that the two CDF curves are close to each other indicates that Sina Weibo users, by using comments, interact with each other almost as frequently as they tweet.

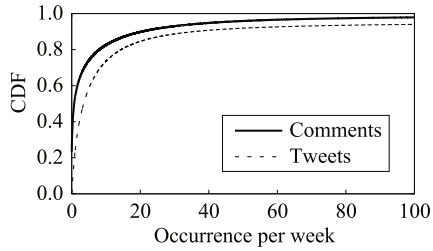


Fig. 5 User comment frequency

Next, we measure how quickly users interact via comments after a tweet. In Fig. 6 we plot a histogram on the time span between each tweet and its first comment, excluding those without any comment. We observe that users act quickly to make comments. The large majority of tweets (95% (7% for < 1min, 64% for (1min, 1h), and 24% for (1h,1d))) received their first comment within a day, and 71% (7% for < 1min, 64% for (1min,1h)) got a comment within just an hour. Thus we can conclude that in Sina Weibo, the comment feature enables fast (and concentrated) social interactions among users.

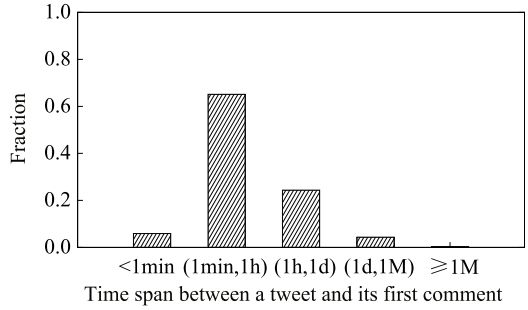


Fig. 6 First comment arrival time (min: minute, h: hour, d: day, M: month)

4.3 Composition of commentors

For each user, we define a commentor as any other user who

[10] This covers more than 99% of the tweets, because less than 1% of the tweets have more than 30 reposts or 50 comments

has posted at least one comment on his tweets. We now study the composition of commentors with three key questions: 1) how many commentors does a user have, 2) what portion of these commentors are the user's friends and 3) what are the relationships among the commentors?

Figure 7 plots the CDF of the number of commentors for each user. We see that 50% users have more than 10 (when $y = 0.5, x > 10$) commentors, and 10% received comments from more than 100 users (when $y = 0.9, x > 100$). We can also calculate that the average number of commentors per user is 39.1, which is significantly larger than the average number of comments per tweet (1.92). We can infer that a commentor usually involves in commenting on multiple tweets.

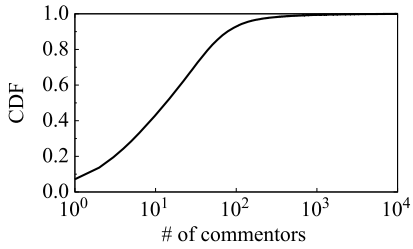


Fig. 7 Distribution of commentors for a user

To answer the second question, we plot in Fig. 8 the CDF of the fraction of friend commentors, who are the commentors that are also the user's friends (bidirectional links). In our case, for 60% users, their friends contribute to more than half of their commentors (when $y = 0.4, x > 0.5$). We see that the commentors are mostly friends with the tweet author.

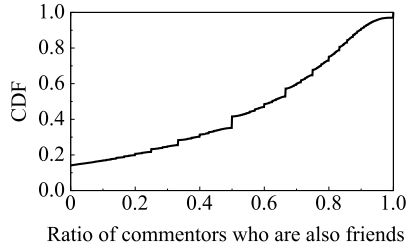


Fig. 8 Fraction of commentors who are also friends

For the third question, we define CC_u, a user u's commentor clustering coefficient, which measures the extent to which the commentors follow each other:

$$CC_u = \begin{cases} \dfrac{|F_{v,w}|}{c_u(c_u-1)}, & \text{if } c_u > 1; \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where c_u is the number of commentors and $F_{v,w}$ is the set of all following links between v and w (either direction) such that v, w are both u's commentors. CC_u helps us to evaluate how likely a user's commentors are friends (follow each other). A higher value of CC_u means user u's commentors are more likely friends with each other.

For our dataset, the average commentor clustering coefficient is 0.180. It is higher than the (general) clustering coefficient in Sina Weibo (0.130 as we measure), Renren (0.063) [3], Facebook (0.164) [6] and Twitter (0.106) [14]. It indicates that a user's commentors are more likely to follow each other, reaffirming the fact that comments are indicative of strong social connections between users.

4.4 Conversations

The comment feature allows Sina Weibo users to interact with each other at ease. In particular, users can reply to each other's comments. These replies (i.e., replying comments), if exist, arrive quickly, usually within an hour (Fig. 9). Because comments and their replies reveal a unique type of (concentrated) interactions among users under the original tweet, we characterize them via conversations and study them in detail. Intuitively, a conversation contains a series of comments and the subsequent replies. To be more exactly, a conversation is a chain of comments that each comment replies to the former one.

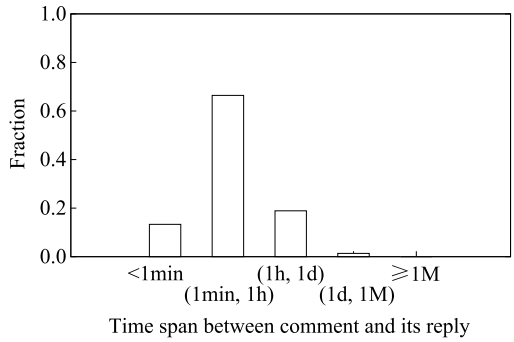


Fig. 9 Time span between comment and its reply (min: minute, h: hour, d: day, M: month)

In the following, we study the conversations in Sina Weibo from three perspectives: 1) how often do commentors form conversations? 2) how long does each conversation last? and 3) how often does a conversation involve the tweet author himself?

We start by investigating, for each tweet with at least one comment, the portion of commentors involved in at least one conversation:

$$\text{ratio} = \frac{\text{\# of commentors in at least one conversation}}{\text{\# of commentors}}. \quad (2)$$

Here we consider a tweet author as a commentor if she also participates in at least one conversation. Figure 10 shows

the CDF of this metric across all the qualified tweets. We see that 60% tweets produced conversations (ratio>0) (when $x = 0, y = 0.4$). More specifically, in 50%+ tweets, more than 50% commentors are involved in conversations (when $y = 0.5, x > 0.5$). This result is not surprising, preseating the fact that users tend to respond to comments, effectively forming conversations.

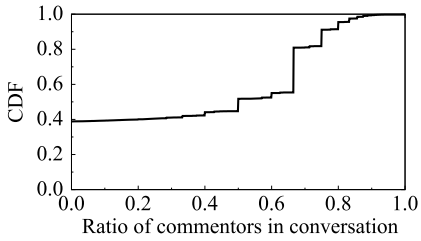


Fig. 10 Ratio of commentors who join in conversations

Next, we quantify the length of a conversation by the number of comments and replies this conversation contains. Since a conversation requires at least one comment and reply, the minimum length is 2. Figure 11 plots the CDF of the average length of conversations in each tweet. We see that the majority of conversations are short, e.g., 80% conversations contain 4 or fewer comments (when $y = 0.8, x = 4$), and almost all the conversations have less than 15 threads. Note that a tweet may contain multiple conversations. Based on our analysis, 98% tweets have less than 6 conversations (Fig. 12 , when $x = 6, y = 0.98$). These results show that the comment feature increases concentrated interactions among users.

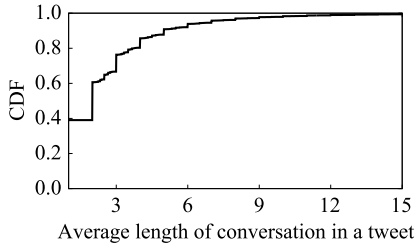


Fig. 11 Average length of conversations in a tweet

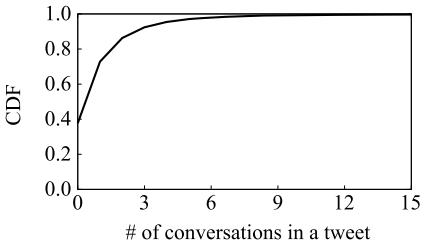


Fig. 12 Number of conversations per tweet

Finally, we look at the participants in each conversation, and examine how often the original tweet author gets involved. Interestingly, out of all the tweets that have any conversation, we observe 92% whose conversations all involve the tweet author, and 3% whose conversations never involve the tweet author. The later case maps to tweets posted by celebrities and organizations, where the tweet authors just broadcast the news and the commentors initiate the subsequent conversations. From these results, we can conclude that the comment feature is highly effective for a tweet author to interact (bidirectionally) with other users.

4.5 Summary of observations

Our detailed analysis provides four key observations.

- The comment feature, which is a key difference between user features in Sina Weibo and Twitter, is also the most prevalent interaction mode in Sina Weibo.
- Sina Weibo users interact via comments nearly as frequently as they tweet, and usually give comments very quickly within an hour of the original tweet.
- Each user receives comments mainly from friends.
- Comments on each tweet often form conversations, allowing users to interact intensively with each other in a short period. The mass majority of the conversations involve the original tweet authors.

Together, these key findings also confirm that the comment feature is a significant enabler (and contributor) to social interactions in Sina Weibo. It makes Sina Weibo be essentially different from Twitter and much closer to classical social networks like Facebook.

5 Comment and interaction graphs

Next, we take our analysis on comments to a macroscopic level and examine the graph (or network) structure of comment activity in Sina Weibo. We construct a comment graph and a repost graph from our Sina Weibo data (Section 5.1), and compare their structures to different types of interaction graphs on Facebook and Twitter (Section 5.2). The comparisons clearly reveal two patterns of interaction graphs: comment graph resembles that of Facebook and repost graph is close to that of Twitter. Then, we explore the correlations between user's interactions with their social relationships, by comparing Sina Weibo's comment and repost graphs with Sina Weibo's social graph from a statistical perspective (Section 5.3) and deploying the experiments of link prediction through an experimental perspective (Section 5.4).

5.1 Building interaction graphs

The prior works have examined social interactions in a number of social networks, focusing on both visible interactions such as wall posts and photo tags [6, 15, 16], and latent interactions such as social profile browsing [3, 17, 18]. For each form of interaction, we can build an interaction graph capturing the activity across the network [6,19]. The resulting graph represents each user as a node, and each interaction taken by user A onto user B as a directed edge from A to B. If a user does not perform or receive any interactions, he becomes a singleton and is removed from the graph. For our analysis, the interaction graphs only need to capture the existence of interactions between users rather than the number of interactions. Thus the resulting graphs are directed and unweighted. In the following, we describe how we build interaction graphs in Sina Weibo, Facebook and Twitter. Table 2 lists the basic properties of these graphs.

Sina Weibo interaction graphs To build Sina Weibo's interaction graphs, we first obtain a connected subgraph (Sina Weibo's social graph) of 723 thousand users (see Section 2). We crawl the latest 2000 tweets from these users, along with all reposts and comments created by these users. These include comments or reposts made by users in our dataset on other users outside of our dataset. To build our graph, we only consider those interactions where both endpoints are within our user set. We denote each user as a node and a comment/repost from user A to B as a directed edge from A to B. In this way, we construct two interaction graphs from our dataset: Sina Weibo comment graph and repost graph.

Facebook's and Twitter's interaction graphs For our Facebook and Twitter graphs, we contact the authors of the prior papers on Facebook [6] and Twitter [19], and receiv permission to use their anonymized graphs as bases for comparison in our work. Wilson et al. built the visible interaction graph based on Facebook's wall posts [6] and compared it with the social graph of Facebook. We use the same dataset[11], and built an anonymous interaction graph of Facebook. Unlike [6], our interaction graph is directed: when user A posts on user B's wall, we create a directed edge from A to B. For Twitter, we use the dataset from the previous work [19], which contains about 3 million users' profiles with social links and all of their tweets. We identify retweet interactions and build an interaction graph for Twitter from those events. More specifically, if user A publishes a tweet that includes "RT @" followed by user B's name, we create a directed edge from A to B in the interaction graph. We call this Twitter interaction graph.

5.2 Comparing interaction graphs

We compare and contrast the four interaction graphs in the context of graph metrics, including degree distribution, clustering coefficient, reciprocity and balance, assortativity, and tie strength. Note that the data for all graphs are obtained through the same BFS algorithm, and thus metrics of the three graphs (Sina Weibo, Facebook and Twitter) are comparable.

Degree distribution The prior studies [13, 20] show that for most social networks, node degree follows a power-law distribution $P(k) \propto k^{-\alpha}$. We found, however, the power-law distribution with an exponential cutoff ($P(k) \propto k^{-\alpha}e^{-\lambda k}$), a generalized version of the power-law distribution, lowers fitting errors and is a better fit for our graphs.

With the probability distribution of power-law, we can estimate the expectation of the value of maximum degree. If the maximum degree is K, and the number of nodes in the graph is N, we have the following formula:

$$\int_K^\infty P(x)dx \approx \frac{1}{N}.$$

It means: the expected number of nodes with degree $> K$ should be less than 1. If we set $P(x) = (\alpha - 1)x^{-\alpha}$, we can calculate an expected maximum degree:

$$K = N^{\frac{1}{\alpha-1}}.$$

In practice, the value of α is usually between 2 and 3 [13]. If we have $\alpha = 2.5$, and we have 1 billion users, the maximum degree would be 1 million. This is normal for the number of Twitter followers, but would be too large for a user's Facebook friends.

Table 2 Basic properties of our interaction graphs, including the comment and repost graphs from Sina Weibo, plus the Facebook's and Twitter's interaction graphs

Graph	Node	Edge	Power-law fit α in/out	Exponent cutoff λ in/out	RMSE in/out	Clustering coefficient	Assortativity	Tie strength
Comment	382 713	2 009 948	0.980/0.804	0.079 6/0.090 2	0.124/0.111	0.110	–0.011	7.74
Repost	352 557	1 529 967	2.090/1.130	**0.001 5**/0.076 9	0.193/0.145	0.070	–0.100	2.20
Facebook	273 497	1 627 253	0.927/0.956	0.047 6/0.059 6	0.112/0.167	0.081	0.230	6.54
Twitter	4 175 695	28 599 550	1.910/1.092	**0.002 8**/0.057 8	0.163/0.466	0.092	–0.027	3.14

[11] http://sandlab.cs.ucsb.edu/facebook/

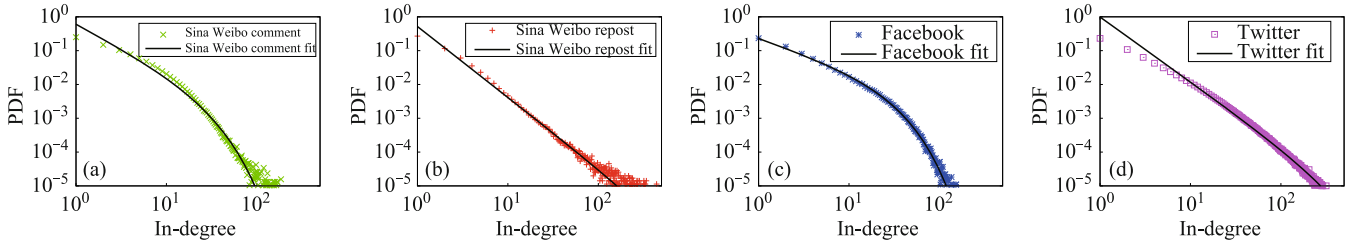


Fig. 13 In-degree distribution and fits using power-law with exponent cutoff. (a) Sina Weibo comment; (b) Sina Weibo repost; (c) Facebook; (d) Twitter

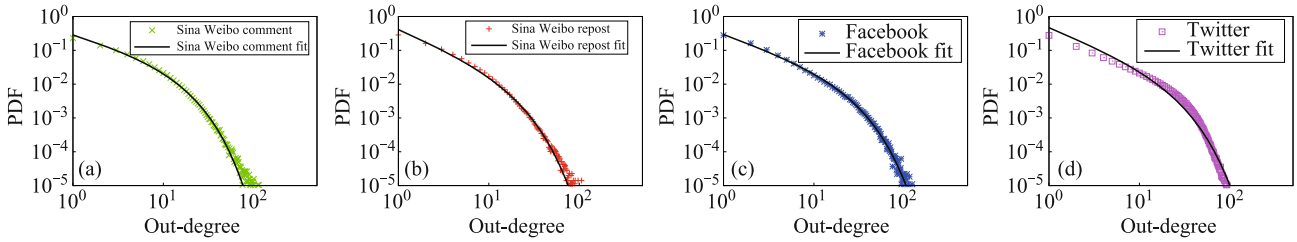


Fig. 14 Out-degree distribution and fits using power-law with exponent cutoff. (a) Sina Weibo comment; (b) Sina Weibo repost; (c) Facebook; (d) Twitter

Larger α would get smaller K. In the extreme case where $\alpha = 3$, we would get the smallest $K \approx 30\,000$. This would still be a little bit too large for a user's friends list. The prior studies on the Dunbar number [21, 22] shows that one can only maintain stable social relationships with a small set of people. As a result, a refinement with exponent cutoff ($P(k) \propto k^{-\alpha}e^{-\lambda k}$) needs to be introduced to the pure power-law distribution.

Table 2 lists the fitting parameters (α and λ) and the root mean square error (RMSE) for these four interaction graphs. We use the Matlab function cftool to estimate the power-law fit parameter with the least mean square error. We also plot the in- and out-degree distributions together with their fitted curves in Figs. 13 and 14 respectively. From these results, we make two key observations.

First, the in-degree and out-degree distributions are similar for both Sina Weibo comment graph and Facebook interaction graph. This means that like interactions in Facebook, comments in Sina Weibo display a symmetric graph structure. On the other hand, for Twitter interaction graph and Sina Weibo repost graph, the in-degree and out-degree distributions are significantly different, indicating a strong asymmetry in user relationships.

Second, the in-degree distributions of Twitter interaction and Sina Weibo repost graphs have very small λ values. A small λ means that the percentage of users with high node degrees drops slowly (or there are a relatively large number of users with high node degrees). This is because in Twitter and Sina Weibo, celebrities and popular organizations generate many tweets that are also highly retweeted, and thus Twitter interaction and Sina Weibo repost graphs have many nodes with high in-degrees. However, users are unlikely to maintain a large number of (symmetric) social interactions. In fact, this result aligns with the prior studies on the Dunbar number [21, 22], which suggest that one can only maintain stable social relationships with a small set of people.

Clustering coefficient In Fig. 15 we plot local clustering coefficients [23] for the four interaction graphs, and the average values are also shown in Table 2. The local clustering coefficient is defined as:

$$C_u = \begin{cases} \dfrac{|E_{v,w}|}{k_u(k_u-1)}, & \text{if } k_u > 1; \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where $E_{v,w}$ is the set of all edges between v and w (either direction) such that $v, w \in N_u$. Here N_u is the set of u's neighbors, i.e., all nodes that are directly connected to or from u. Then $k_u = |N_u|$. The clustering coefficient is a fraction between 0 and 1 and characterizes the connectivity among one node's neighbors, where 0 represents a star shape around the local node, and 1 represents a full clique.

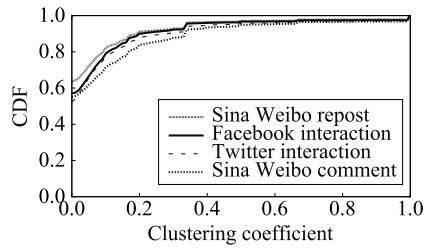


Fig. 15 Clustering coefficient distribution of interaction graphs

Interestingly, Sina Weibo comment graph has a larger local clustering coefficient than Sina Weibo repost graph. This means that the comment graph is more densely connected in its local structure, which confirms that comments are more prevalent than reposts.

Reciprocity and balance Figure 16 plots the reciprocity of our interaction graphs. The reciprocities for Sina Weibo comment graph and Facebook interaction graph are close, and both are significantly higher than graphs based on information dissemination events, i.e., Twitter interaction graph and Sina Weibo repost graph. Bidirectional edges make up more than half of all edges for more than 70% of all nodes in Sina Weibo's comment graph and Facebook's interaction graph (when $x = 0.5, y < 0.3$). This again demonstrates the impact of comments as a mechanism for bidirectional social interactions.

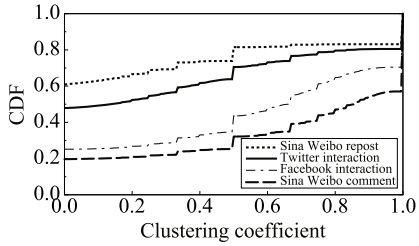


Fig. 16 Reciprocity comparison among interaction graphs

Figure 17 shows the balance of the interaction graphs, which represents the ratio of in-degree to out-degree in each graph. Again, we find that Sina Weibo comment graph and Facebook interaction graph are similar, and are more balanced than the other two. Surprisingly, the ratio of balanced users is 72.8% in Sina Weibo comment graph (y ranges between $x = [0.5, 2]$), which is slightly higher than that of Facebook (66%). This indicates that comments in Sina Weibo are slightly more balanced and symmetrical than social interactions in Facebook.

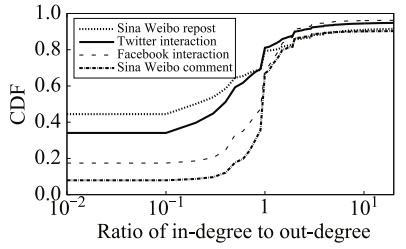


Fig. 17 Balance comparison among interaction graphs

Assortativity Assortativity measures the homophily of users in a social network. A positive assortativity coefficient means nodes tend to connect with other nodes with similar degrees, which is usually considered as a property of social network. Our result shows Sina Weibo comment and Facebook have higher assortativity than Sina Weibo repost and Twitter. Sina Weibo comment does not produce very high assortativity, because comments partially serve as an interaction channel for users and celebrities and organizations, as we find in Section 4.

Tie strength Tie strength measures how frequently users interact with each other. Our result shows users interact more frequently via comments than reposts. This is consistent with the intuition. Friends comment to each other frequently, while users repost only when there is a valuable tweet.

5.3 Interaction graphs versus social graph

We compare Sina Weibo's interaction graphs (comment and repost) with Sina Weibo's social graph. Our goal is to better understand, at the network level, whether user's interactions are correlated to whom users make friends with.

We start by analyzing the overlaps between comment graph and social graph. Specifically, we examine how many edges in comment graph connect users with established social relationships (e.g., friend, follower or followee)[12]. The results are shown in the upside of Table 3. A quick observation is that users who have comment interactions (either one-way or bidirectional comment) have a high probability to be friends (76.8%), while the chance they are followers or followees is only 9.2%. This indicates that comment interaction has a strong correlation with the bidirectional social relationships in Sina Weibo. Particularly, if two users comment to each other (bi-comment), the chance that they are friends is 93.5%.

Table 3 Overlaps between interaction and social graphs

Interaction edges	Total #	Friend/%	Follower(e)/%	None/%
One-way comment	558 883	57.4	16.6	26.0
Bi-comment	654 500	93.5	2.8	3.7
All comment	1 213 383	**76.8**	**9.2**	14.0
One-way repost	1 351 534	28.6	20.3	51.1
Bi-repost	26 603	95.9	3.1	1.0
All repost	1 378 137	29.9	**19.9**	**50.2**

Next, we repeat the same analysis between repost graph and social graph. The results are shown in the bottom of Table 3. For all cases, users who have repost interactions are not necessarily to be socially connected: 70.1% (100%−29.9% = 70.1%) of the chance that they are not friends, and 50.2% of the chance that they do not have any kind of social re-

[12] Edges that indicate self-interaction, i.e., user comments to himself, are not considered in this analysis

lationships with each other. This indicates that repost is an interaction that also happens among strangers. Note that if two users repost each other's tweet (bi-repost), they still have a high probability to be friends (95.9%). However, bi-reposts are very rare in Sina Weibo, with only 26 603 out of 1 378 137 repost edges bidirectional (1.9%).

Our analysis shows Sina Weibo's comment graph has a bigger overlap with social graph than repost graph does, especially on bidirectional social relationships (76.8% to 29.9%). In other words, users who participate in comment interactions are typically friends, while repost is a interaction that often happens between non-friends or even strangers. This result helps to explain why Sina Weibo (with comment function) has more balanced and symmetric social relationships than Twitter (with repost but no comment function).

5.4 Link prediction

In this section, we further confirm the strong correlation between comments and strength of social links via link prediction experiments. We follow experiments from Ref. [24] with widely-used metrics in the prior work [25].

From the social graph we built, we randomly select 10% of edges as "missing" edges. We remove these edges from the social graph and comment graph. The link prediction problem is to find out the deleted links based on the remaining graph structure, i.e., remaining social graph or comment graph after the edges are deleted. The idea is that two nodes with higher similarity are more likely to establish a link. We use three widely used metrics, i.e., common neighbors (CN), jaccard coefficient (J), and adamic (A) to measure node similarity [25]. If we use $\Gamma(x)$ to represent the neighbors of user x, then the three metrics can be calculated by the following three formulas: $\mathrm{CN} = |\Gamma(x) \cap \Gamma(y)|$, $\mathrm{J} = |\Gamma(x) \cap \Gamma(y)|/|\Gamma(x) \cup \Gamma(y)|$, and $\mathrm{A} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} 1/log|\Gamma(z)|$. For all three metrics, a larger value means a higher similarity between the two nodes. To avoid unnecessary computational complexity, we only consider the nodes from the deleted edges. We then rank node pairs by the similarity and get a list of top M edges. Here M is the number of the deleted edges. We define accuracy as the fraction of the correctly predicted edges in the list out of all deleted edges.

We repeat the random selection of the deleted edges 10 times, and show the average accuracy in Table 4. We have two observations. First, both graphs (comment graph and social graph) give a much better prediction than the baseline, i.e., random guessed edges from node pairs. Second, comment graph is more accurate at predicting social links than the social graph with an improvement of at least 30% (e.g., $(0.035 - 0.026)/0.026 > 30\%$). This implies that comments better capture the strength of social ties, and those properties can be leveraged for more effective link prediction.

Table 4 Accuracy of link prediction

Graph	Common neighbors (CN)	Jaccard coefficient (J)	Adamic (A)
Random		$8.33 * 10^{-6}$	
Comment graph	0.035	0.021	0.041
Social graph	0.026	0.014	0.031

5.5 Summary of observations

We have two key observations in our network-wised analysis.

- Sina Weibo's comment graph resembles Facebook's interaction graph, with high bidirectionality; while Sina Weibo's repost graph is similarity to Twitter's interaction graph, which represents one-way, asymmetric interactions in information dissemination events.
- Sina Weibo's comment graph has a much bigger overlap with social graph than repost graph. Users are more likely to be friends if they have comment interactions.

These results indicate the comment function has played a significant role in enabling and improving bidirectional interactions and friending among users. It helps transform Sina Weibo into a platform not only just for information dissemination (like Twitter), but also for social friending and interactions (like Facebook).

6 Comments in user modeling

Thus far, our analysis has confirmed that Sina Weibo's comment function has significantly changed how users interact. This in turn makes comment be a potentially important dimension for modeling user behaviors. In this section, we take a deeper look at comment and user-behaviors based applications with two key questions. First, how can we apply comment features in user-behavior based applications; second, how significant the role do comments play in such applications?

To answer these questions, we consider two practical application cases of comment feature. First, we use machine-learning detectors to identify malicious users. We aim to leverage comment features to build (or augment) user behavior models for detection. Second, we investigate the influence maximization problem with our comment and repost graph.

We want to evaluate how information disseminates over different types of interaction links.

6.1 Malicious user detection

In the following, we first label ground-truth legitimate and malicious accounts. Then we describe our comment features to model user behavior. Finally, we use several ML techniques to build behavioral models for malicious user detection. As we illustrate, our results show that the comment feature is a key factor in defining user behavior, which significantly boosts the accuracy across all of our machine learning based detectors.

6.1.1 Labeling users

For our case study, we identify both malicious and legitimate accounts from the crawled dataset as follows.

Malicious accounts Sina Weibo relies in part on its users to report suspicious accounts, by submitting screenshots of their malicious behavior. Sina Weibo administrators manually check the reported accounts, and immediately block the confirmed malicious accounts, making them inaccessible by Sina Weibo's APIs. We leverage this to generate a ground-truth dataset of banned accounts. We perform a second round of crawls in September 2013 (seven months after the original crawl), and discover 4 639 accounts are blocked. We label these 4 639 accounts as malicious accounts.

Legitimate accounts One way for users to verify their identities to Sina Weibo is to bind or associate their accounts with either their Chinese national ID or cell phone number. These users can be identified by a special Authenticated label in their profiles. By providing their real-world identities, these users can be held legally for their actions, and are unlikely to behave maliciously. We have identified 71 890 authenticated accounts in our crawls, and use them as legitimate users in our dataset.

6.1.2 Characterizing comment patterns

We first study the per-user comment activities of the malicious and legitimate user groups. In previous analysis (Sections 4 and 5) we have mentioned some general metrics to characterize user comment interactions, but here we focus on four metrics below that are more indicative of malicious users. The results are consistent, indicating that malicious accounts tend to make a significantly smaller number of comments than legitimate users.

Tweets with comments Our first metric is, for each user, the percentage of tweets that have user comments. As shown in Fig. 18(a), legitimate users clearly are more likely to attract comments. For 71.9% of legitimate accounts, at least 20% of their tweets have comments (when $y = 0.281, x = 0.2$). In contrast, only 24.4% of malicious accounts have comments on 20% or more of their tweets (when $y = 0.756, x = 0.2$).

Incoming comments per tweet We plot the average number of comments users received per tweet in Fig. 18(b). Malicious accounts have fewer comments on their tweets. Only 16.1% of malicious accounts have one or more incoming comments for their tweets (when $x = 1, y = 0.839$). The analogous number for legitimate accounts is 92.4% (when $x = 1, y = 0.076$).

Ratio of bi-commentors Figure 18(c) plots the ratio of *bi-commentors* associated with each user. Two users are called bi-commentors if they have commented on each other's tweets or on the same tweet. The figure shows a significant difference between malicious and legitimate users. Nearly 80% of malicious accounts have zero bi-commentors (when $x = 0, y = 0.8$), while bi-commentors are quite common for legitimate users.

Comment h-index Our last metric is comment h-index. A user has a comment h-index of h if he has at least h tweets with no less than h comments. This metric is inspired by the "h-index" research publication impact metric [26]. Here we use it to measure user influence. As shown in Fig. 18(d), malicious accounts have significantly smaller comment h-index values relative to legitimate accounts.

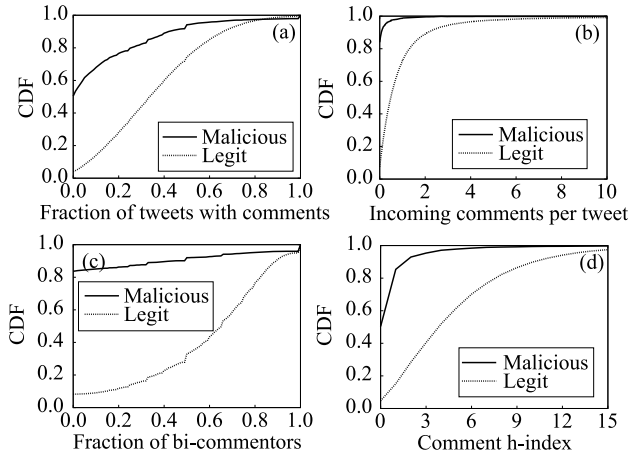


Fig. 18 Comment patterns of malicious and legitimate accounts. (a) Ratio of tweet with comments; (b) in-comments per tweet; (c) ratio of bi-commentors; (d) comment h-index

These results show that malicious accounts have much less comment interactions. There are two possible reasons. First, comments are a form of interaction that involves frequent exchanges with friends (as shown in Section 4). These interactions come with a heavy overhead in time and energy, making

them too costly for most malicious users. Second, comment does not really help malicious users spam their followers, because unlike repost and tweet, comment does not generate new "events" on users' timeline, i.e., none impact on their followers. Thus comment is not an attractive activity for malicious users.

6.1.3 Detecting malicious accounts

To measure the extent to which comment interaction features can help malicious user detection, we conduct three experiments, each applying a set of machine learning techniques with different features. The first experiment uses ten common user features already used in the previous works [27, 28], including the number of followers and followees, the ratio of follower to followee, the reciprocity, the average, the minimum and the maximum number of tweets per day, the number of mentions per tweet, the ratio of tweets with mentions and ratio of tweets with URLs. We refer to these as the "existing features (EF)" set. For the second experiment, we use nine features solely based on comment interactions, i.e., the "comment features (CF)" set, including the ratio of tweets with comments, the ratio of outgoing to incoming comments, the comment h-index, the number of commentors, the ratio of bi-commentors, the ratio of friend commentors, the ratio of conversations with tweet authors, the number of incoming comments per tweet and the number of outgoing commentors. Third, we experiment with the combination set of "existing and comment-based features (ECF)".

For detection, we use the banned 4 639 malicious accounts and 4 639 legitimate accounts randomly selected from the legitimate accounts set, and apply several widely used classification algorithms[13)], including Naive Bayes [30], support vector machine (SVM) [31], random forests [32] and logistic regression [33]. We conduct the experiments with ten fold cross-validation using EF, CF and ECF respectively. Accuracy is defined as the ratio of the correctly predicted users to all users. Since we are using the same number of items for both categories, the accuracy is equal to the weighted average recall and F-measure [34].

A summary of the experiment results are listed in Table 5. Our results show that using CF alone can accurately distinguish malicious accounts from normal users with about 90% accuracy. In addition, adding CF to EF significantly boosts accuracy across all techniques, with an average improvement of more than 4%, i.e., $((88.2-75.9)+(93.1-92.2)+(95.0-93.6)+(90.6-86.4))/4$.

[13)] Algorithm implementation by Weka toolkits [29]

Table 5 Detection accuracy using different algorithms

Classification model	EF accuracy/%	CF accuracy/%	ECF accuracy/%
Naive Bayes	75.9	87.1	88.2
SVM	92.2	90.0	93.1
Random forests	93.6	90.2	95.0
Logistic regression	86.4	89.3	90.6

Feature importance Further, we examine the relative role comment-based features play in defining these detectors of malicious activities. Table 6 lists ECF's top ten features ranked by χ^2 (Chi square) statistic [35], which is a widely-used metric to measure feature's discriminative power. As shown, comment-based features account for all top five features and seven out of the top ten features. This indicates that comment-based features have stronger discriminative power than the existing features.

Table 6 Top ten features based on χ^2 statistic

Rank	Feature	χ^2	Type
1	# of out commentors	6 144.9	CF
2	Ratio of friend commentors	5 996.9	CF
3	Ratio of out to in comments	5 905.2	CF
4	Average # of in comments per tweet	5 902.4	CF
5	Bi-commentors/all-commentors	5 795.8	CF
6	# of followers	5 215.5	EF
7	Comment h-index	4 934.8	CF
8	Reciprocity	4 932.8	EF
9	# of commentors	4 675.7	CF
10	Ratio of tweets with mentions	4 116.6	EF

However, despite CF's higher ranking, EF still outperforms CF on certain algorithms, i.e., random forests and SVM (Table 5). A possible explanation is that comment features are individually stronger but tend have an overlapping effect when combined in a classifier. Figure 19 confirms this intuition — as we add more features to random forests classifier, EF provides higher incremental values than CF for boosting the overall accuracy. This indicates that the diversity of the feature set is still important for accurate user behavior modeling. Finally, Fig. 19 also demonstrates the possibility to further shrink the feature set. In fact, if we only use the top-4 features from EF plus the top-1 feature from CF, random forests classifier can still produce accuracy as high as 94.1%.

Summary The above analysis shows that comment interaction is a key factor in defining user behaviors. Particularly, it is a strong indicator for malicious behaviors of social network accounts. To this end, using comment-based features for malicious user detection is a big favor for higher detection accuracy. In addition, comment-based features also make the

detector more robust in practice — it is very difficult for malicious accounts to evade these features, because generating organic comment interactions takes a significant amount of effort, time and even costs (e.g., getting comments via paid crowdsourcing [36]).

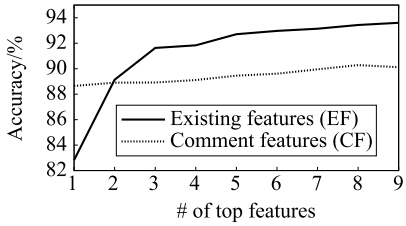


Fig. 19 Accuracy versus the number of top features on random forests classifier

6.2 Influence maximization

Understanding how information disseminates in OSNs is a critical problem for better marketing and user experience. The influence maximization problem is to decide a set of the most influential people who can maximize the information diffusion in a social network. Formally, in a graph G and a random process which defines the information dissemination, the influence maximization problem is to optimize a set of seeds S, which maximizes the information propagation in G.

However, the influence maximization problem is proved to be NP-hard [37]. We can only use a greedy algorithm to get the best possible approximation. Varied algorithms are proposed to improve the efficiency of the greedy algorithm [38–40]. For example, cost-effective lazy forward (CELF) [38] optimizes the simple greedy algorithm based on the submodularity. In each round, CELF does not need to re-evaluate the incremental influence. As reported in Ref. [38], CELF is 700 times faster than the simple greedy algorithm empirically.

Goals and methodology In this experiment, we aim to evaluate how information disseminates in two different graphs: comment graph and repost graph. We do not aim to propose any new algorithms for influence maximization. As a result, we leverage an existing and widely used algorithm CELF. This algorithm also serves as a baseline comparison in other works [39, 40]. For the information propagation, we use independent cascade model (IC).

For our experiments, we run the CELF algorithm and consider the size of seeds set S varying from 1 to 100. For each S with n seeds, we evaluate the number of influenced nodes.

Results We plot the evaluation results in Fig. 20. The x-axis is the number of seeds, and the y-axis is the number of users influenced by the seeds. From the figure, we find more and more users are influenced in comment graph with the same size of seeds set. With 10 seeds, around 1 800 users are influenced in comment graph. The corresponding value for repost graph is only 600. The result implies that comment graph is a more efficient representation for information diffusion in Sina Weibo. This is not surprising, because comment is the dominating interaction type in Sina Weibo. It reveals a closer relationship among users.

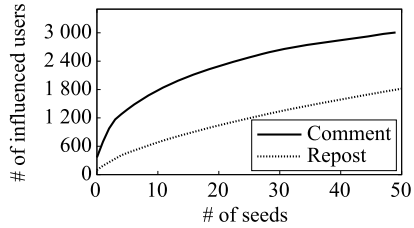


Fig. 20 Influence maximization in comment and repost graphs

In Fig. 21, we evaluate the overlap of selected seeds between comment and repost graphs. We find the selected nodes are quite different for different graphs. Out of 50 seeds, there are only 3 overlaps between the two graphs, the union set size is 97, while the intersection set size is 3. That is, the representation of user interactions can lead to varied results in practical applications. The result suggests we should be careful when selecting the model of user interactions.

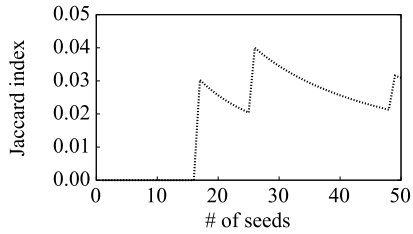


Fig. 21 Overlap of selected seeds

7 Related work

Behaviors of microblogging networks Recent studies have examined the behaviors of microblogging systems in detail, focusing primarily on Twitter. The work by Hwak et al. is the first to show that Twitter is a news media rather than a conventional social network [1]. A recent study [41] shows that Twitter has envolved into a hybrid of information network and social network, and it is based on only active users. Subsequent efforts have studied Twitter from different perspectives, ranging from information diffusion [42, 43], user influence [44, 45], to opinion mining [46] and user demo-

graphics [47]. Our work differs from these efforts by focusing on the unique feature of Sina Weibo, a different (social) microblogging network. Using detailed data analysis, our study discovers viable structure differences between Sina Weibo and Twitter, and identifies the key user feature that causes such significant differences.

There have also been data-driven studies on Sina Weibo, including video tweeting analysis [10], tweets deletion behavior [48, 49] and social influence [50]. Specifically, there are studies comparing Sina Weibo and Twitter — one study compares Sina Weibo and Twitter from several aspects of user behaviors (tweets in particular) [2], but do not consider social interactions, i.e., comments, among Sina Weibo users; another study examines information propagation in Sina Weibo and Twitter [7], concluding that Twitter's information propagation is much faster and more frequent than Sina Weibo. This conclusion is consistent with ours, which shows Sina Weibo is not only news media but also supports social interactions among users. We note that both studies are based on a very small fraction of the Sina Weibo users ($< 1\ 400$) and tweets (1.5 million). Our work differs from these by focusing on the comment feature. And our data in use is a much larger collection of both tweets and comments, i.e., 61.5 million tweets from 723 thousand users, and analysis of 57 million users.

Several works have examined Sina Weibo from an application-centric perspective. Qu et al. studied the user behavior after a major earthquake, demonstrating the effectiveness of Sina Weibo in providing quick responses to disasters [51]. Yu et al. tried to detect the sleeping times of users in Sina Weibo according to their activities [52]. And Liao et al. studied the rumor propagation in Sina Weibo and explored the information dynamics [9] while Yang et al. proposed automatic detection algorithms of rumors in Sina Weibo [53]. These studies all conclude that in-depth understanding of Sina Weibo is critical in developing successful applications.

User interactions in social networks User interaction is a unique and critical feature of OSNs, and has attracted attentions from the research community, including the visible interaction in Facebook [6] and mention-based interactions in Twitter [54]. Our work is motivated by these studies and their insights. Our work differs from these existing works by focusing on the impact of interactions and how these interactions are shaped. We have found that Sina Weibo users interact mostly via comments, not reposts, which form interesting social interactions among users similar to those of Facebook. We then have built a series of Sina Weibo interaction graphs to further understand such social interactions.

Malicious account detection Researchers have devoted significant efforts to detect malicious accounts (e.g., spammers and sybils) in large OSNs and microblogging systems, including Facebook [55], Twitter [27, 28] and Renren [56–58]. One category of these works [27, 28, 59] takes advantage of different classification algorithms of machine learning techniques. These works mostly focus on features of malicious accounts' (suspected) attacking behavior such as mentions, blacklist URLs and spam key words. Our work differs from these works by proposing effective new features of comments (user interactions).

8 Conclusion

This paper raises the question. Can the design of a single feature affect user behavior in microblogging networks? To answer this question, we perform a detailed measurement and analysis of the Sina Weibo microblogging network. Sina Weibo is similar in all respects to Twitter, except for its support for comments on tweets. We believe that this key feature has led to significant amount of symmetric social interactions on Sina Weibo, partially transforming it from primarily a news dissemination platform (like Twitter) into a hybrid social network that supports two primary use models: news dissemination and social interactions.

We test our hypothesis from a variety of perspectives using datasets from Sina Weibo, Twitter and Facebook. Our results show that users of Sina Weibo display much more symmetric social connectivity than Twitter. We also show that Sina Weibo's user comments are similar in many respects to Facebook's social interactions, and mark a clear departure from the asymmetric interactions typically found in Twitter. While limited to a single feature, our study demonstrates a strong correlation between a single design feature (comments) and a dramatic difference in user behavioral patterns.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant Nos. U1405254, 61472092, 61402115, and 61271392). Any opinions, findings, and conclusions or recommendations expressed in this material were those of the authors and did not necessarily reflect the views of any funding agencies.

References

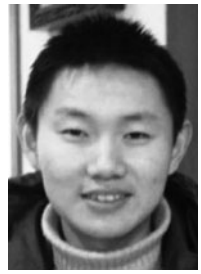
1. Kwak H, Lee C, Park H, Moon S. What is Twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web. 2010, 591–600
2. Gao Q, Abel F, Houben G J, Yu Y. A comparative study of user's microblogging behavior on Sina Weibo and Twitter. In: Proceedings of

the 20th Conference on User Modeling, Adaptation, and Personalization. 2012, 88–101

3. Jiang J, Wilson C, Wang X, Sha W P, Huang P, Dai Y F, Zhao B Y. Understanding latent interactions in online social networks. ACM Transactions on the Web (TWEB), 2013, 7(4): 18
4. Gjoka M, Kurant M, Butts C T, Markopoulou A. Practical recommendations on crawling online social networks. IEEE Journal on Selected Areas in Communications, 2011, 29(9): 1872–1892
5. Ribeiro B, Towsley D. Estimating and sampling graphs with multidimensional random walks. In: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement. 2010, 390–403
6. Wilson C, Boe B, Sala A, Puttaswamy K P N, Zhao B Y. User interactions in social networks and their implications. In: Proceedings of the 4th ACM European Conference on Computer Systems. 2009, 205–218
7. Gao Q, Abel F, Houben G J, Yu Y. Information propagation cultures on Sina Weibo and Twitter. In: Proceedings of the 4th Annual ACM Web Science Conference. 2012, 157–162
8. Fu K, Chau M. Reality check for the Chinese microblog space: a random sampling approach. PloS one, 2013, 8(3): e58356
9. Liao Q Y, Shi L. She gets a sports car from our donation: rumor transmission in a Chinese microblogging community. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work. 2013, 587–598
10. Guo Z D, Huang J, He J, Hei X J, Wu D. Unveiling the patterns of video tweeting: a Sina Weibo-based measurement study. In: Proceedings of Passive and Active Measurement. 2013, 166–175
11. Chen L, Zhang C, Wilson C. Tweeting under pressure: analyzing trending topics and evolving word choice on Sina Weibo. In: Proceedings of the 1st ACM Conference on Online Social Networks. 2013, 89–100
12. Sala A, Zheng H T, Zhao B Y, Gaito S, Rossi G P. Brief announcement: revisiting the power-law degree distribution for social graph analysis. In: Proceedings of the 29th ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing. 2010, 400–401
13. Clauset A, Shalizi C R, Newman M E J. Power-law distributions in empirical data. SIAM Review, 2009, 51(4): 661–703
14. Java A, Song X D, Finin T, Tseng B. Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. 2007, 56–65
15. Chun H, Kwak H, Eom Y H, Ahn Y Y, Moon S, Jeong H. Comparison of online social relations in volume vs interaction: a case study of cyworld. In: Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement. 2008, 57–70
16. Viswanath B, Mislove A, Cha M, Gummadi K P. On the evolution of user interaction in Facebook. In: Proceedings of the 2nd ACM Workshop on Online Social Networks. 2009, 37–42
17. Benevenuto F, Rodrigues T, Cha M, Almeida V. Characterizing user behavior in online social networks. In: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference. 2009, 49–62
18. Schneider F, Feldmann A, Krishnamurthy B, Willinger W. Understanding online social network usage from a network perspective. In: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference. 2009, 35–48
19. Xu T Y, Chen Y, Jiao L, Zhao B Y, Hui P, Fu X M. Scaling microblogging services with divergent traffic demands. In: Proceedings of the 12th International Middleware Conference. 2011, 20–39
20. Mislove A, Marcon M, Gummadi K P, Druschel P, Bhattacharjee B. Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement. 2007, 29–42
21. Dunbar R I M. Neocortex size as a constraint on group size in primates. Journal of Human Evolution, 1992, 22(6): 469–493
22. Dunbar R I M. The social brain hypothesis. Foundations in Social Neuroscience, 2002, 5(71): 69
23. Watts D J, Strogatz S H. Collective dynamics of “small-world” networks. Nature, 1998, 393(6684): 440–442
24. Zhao X H, Chang A, Sarma A D, Zheng H T, Zhao B Y. On the embeddability of random walk distances. VLDB Endowment, 2013, 6(14): 1690–1701
25. Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019–1031
26. Hirsch J E. An index to quantify an individual’s scientific research output. National Academy of Sciences of the United States of America, 2005, 102(46): 16569–16572
27. Benevenuto F, Magno G, Rodrigues T, Almeida V. Detecting spammers on Twitter. In: Proceedings of Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference. 2010, 12
28. Wang A H. Don’t follow me: spam detection in Twitter. In: Proceedings of the 2010 International Conference on Security and Cryptography. 2010, 1–10
29. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I H. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter, 2009, 11(1): 10–18
30. Lewis D D. Naive (Bayes) at forty: the independence assumption in information retrieval. In: Proceedings of the 10th European Conference on Machine Learning. 1998, 4–15
31. Chang C C, Lin C J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 27
32. Breiman L. Random forests. Machine Learning, 2001, 45(1): 5–32
33. Le Cessie S, Van Houwelingen J C. Ridge estimators in logistic regression. Applied Statistics, 1992, 191–201
34. Billsus D, Pazzani M J. Learning collaborative information filters. In: Proceedings of International Conference on Machine Learning. 1998, 46–54
35. Yang Y M, Pedersen J O. A comparative study on feature selection in text categorization. In: Proceedings of International Conference on Machine Learning. 1997, 412–420
36. Wang G, Wilson C, Zhao X H, Zhu Y B, Mohanlal M, Zheng H T, Zhao B Y. Serf and turf: crowdturfing for fun and profit. In: Proceedings of the 21st International Conference on World Wide Web. 2012, 679–688
37. Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2003, 137–146
38. Leskovec J, Krause A, Guestrin C, Faloutsos C, VanBriesen J, Glance N. Cost-effective outbreak detection in networks. In: Proceedings of

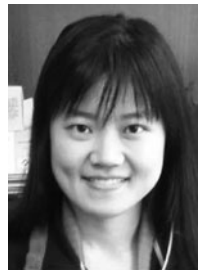
the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2007, 420–429

39. Chen W, Wang Y J, Yang S Y. Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2009, 199–208
40. Goyal A, Lu W, Lakshmanan L V S. CELF++: optimizing the greedy algorithm for influence maximization in social networks. In: Proceedings of the 20th International Conference Companion on World Wide Web. 2011, 47–48
41. Myers S A, Sharma A, Gupta P, Lin J. Information network or social network?: the structure of the Twitter follow graph. In: Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion. 2014, 493–498
42. Yang J, Counts S. Predicting the speed, scale, and range of information diffusion in Twitter. In: Proceedings of the International AAAI Conference on Web and Social Media. 2010, 355–358
43. Lumezanu C, Feamster N, Klein H. #bias: Measuring the tweeting behavior of propagandists. In: Proceedings of the International AAAI Conference on Web and Social Media. 2012
44. Cha M, Haddadi H, Benevenuto F, Gummadi P K. Measuring user influence in Twitter: The million follower fallacy. In: Proceedings of the International AAAI Conference on Web and Social Media. 2010, 10–17
45. Bakshy E, Hofman J M, Mason W A, Watts D J. Everyone's an influencer: quantifying influence on Twitter. In: Proceedings of the 4th ACM International Conference on Web Search and Data Mining. 2011, 65–74
46. Boutet A, Kim H, Yoneki E. What's in your tweets? I know who you supported in the UK 2010 general election. In: Proceedings of the International AAAI Conference on Web and Social Media. 2012
47. Mislove A, Lehmann S, Ahn Y Y, Onnela J P, Rosenquist J N. Understanding the demographics of Twitter users. In: Proceedings of the International AAAI Conference on Web and Social Media. 2011
48. Bamman D, O'Connor B, Smith N. Censorship and deletion practices in Chinese social media. First Monday, 2012, 17(3)
49. Zhu T, Phipps D, Pridgen A, Crandall J R, Wallach D S. The velocity of censorship: high-fidelity detection of microblog post deletions. In: Proceedings of the 22nd USENIX Conference on Security. 2013, 227–240
50. Zhang J, Liu B, Tang J, Chen T, Li J. Social influence locality for modeling retweeting behaviors. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence. 2013, 2761–2767
51. Qu Y, Huang C, Zhang P Y, Zhang J. Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake. In: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work. 2011, 25–34
52. Yu H R, Sun G Z, Lv M. Users sleeping time analysis based on microblogging data. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing. 2012, 964–968
53. Yang F, Liu Y, Yu X H, Yang M. Automatic detection of rumor on Sina Weibo. In: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. 2012, 13
54. Macskassy S A. On the study of social interactions in Twitter. In: Proceedings of the International AAAI Conference on Web and Social Media. 2012
55. Gao H Y, Hu J, Wilson C, Li Z C, Chen Y, Zhao B Y. Detecting and characterizing social spam campaigns. In: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement. 2010, 35–47
56. Wang G, Konolige T, Wilson C, Wang X, Zheng H, Zhao B Y. You are how you click: Clickstream analysis for sybil detection. In: Proceedings of the 22nd USENIX Conference on Security. 2013, 1–15
57. Wang G, Mohanlal M, Wilson C, Wang X, Metzger M, Zheng H T, Zhao B Y. Social turing tests: crowdsourcing sybil detection. In: Proceedings of the 20th Annual Network and Distributed System Security Symposium. 2013
58. Yang Z, Wilson C, Wang X, Gao T, Zhao B Y, Dai Y. Uncovering social network sybils in the wild. ACM Transactions on Knowledge Discovery from Data, 2014, 8(1): 2
59. Stringhini G, Kruegel C, Vigna G. Detecting spammers on social networks. In: Proceedings of the 26th Annual Computer Security Applications Conference. 2010, 1–9



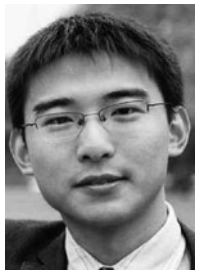
Tianyi Wang is a PhD student in the Department of Electronic Engineering, Tsinghua University, China. He was a visiting student at the Department of Computer Science, University of California, Santa Barbara, USA. His research interests include data mining, the analysis and modeling of online social networks.

Yang Chen is a pre-tenure associate professor within the School of Computer Science at Fudan University, China. Before that, he was a postdoctoral associate within the Department of Computer Science at Duke University from 2011 to 2014. He received his BS and PhD degrees from the Department of Electronic Engineering, Tsinghua University, China in 2004 and 2009, respectively. His research interests include online social networks, Internet architectures, and cloud computing. He is a senior member of the IEEE.



Yi Wang received her diploma in mathematics from Huazhong University of Science and Technology, China in 2005. From 2006 to 2012, she was with the University of Minnesota, USA, where she received her MS in statistics and PhD in mathematics. Between 2012 and 2015, she was a postdoctoral researcher at the Statistical and Applied Mathematical Sciences Institute (SAMSI), USA, and a vis-

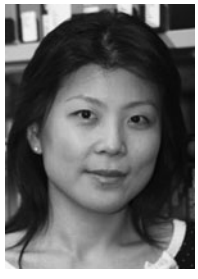
iting assistant professor at the Mathematics Department, Duke University, USA. Since 2015, she joined as an assistant professor the Department of Mathematics at Syracuse University, USA. Her research interests lie in applied harmonic analysis, machine learning, signal and image processing, as well as applications to real data.



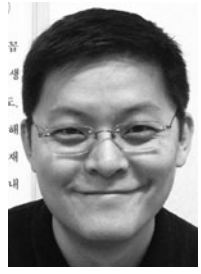
Bolun Wang is a PhD student in the Computer Science Department at the University of California, Santa Barbara, USA. He received his BS degree from Tsinghua University, China in 2009. His research interests are online social networks, data security, and privacy.

Gang Wang received his BE degree in electrical engineering from Tsinghua University, China in 2010. He is currently pursuing the PhD degree in computer science in the University of California, Santa Barbara, USA. His research interests are security and privacy, online social networks, mobile networks and crowdsourcing systems.

Xing Li received his PhD in Electrical Engineering from Drexel University, USA. He is a professor with the Electronic Engineering Department, Tsinghua University, China and the deputy director of China Education and Research Network (CERNET) Center.



Haitao Zheng received her BS degree from Xi'an Jiaotong University, China in July 1995, and her MS and PhD degrees in electrical and computer engineering from University of Maryland, College Park, USA in May 1998 and July 1999, respectively. She joined wireless research lab, Bell-Labs, Lucent Technologies as a member of technical staff in August 1999, and moved to Microsoft Research Asia as a project leader and researcher, in March 2004. Since September 2005, she has been a faculty member in Computer Science Department, University of California, Santa Barbara, USA, where she is now a professor. She was named as the 2005 Massachusetts Institute of Technology (MIT) Technology Review Top 35 Innovators under the age of 35 for her work on cognitive radios. Her work was selected by MIT Technology Review as one of the 10 Emerging Technologies in 2006. She also received 2002 Bell Laboratories President's Gold Award from Lucent Bell-Labs, and 1998–1999 George Harhalakis Outstanding Graduate Student Award from Institute of System Research, University of Maryland, College Park, USA. She was elected IEEE fellow in 2014. Her recent research interests include wireless systems and networking and social networks.



Ben Y. Zhao is a professor at the Computer Science department, University of California, Santa Barbara, USA. He completed his MS and PhD degrees in computer science at University of California-Berkeley, USA (2000, 2004), and his BS from Yale University, USA (1997). He is a recipient of the National Science Foundation's CAREER award, Massachusetts Institute of Technology (MIT) Technology Review's TR-35 Award (Young Innovators Under 35), and ComputerWorld Magazine's Top 40 Technology Innovators award. His work has been covered by media outlets such as New York Times, Boston Globe, MIT Technology Review, and Slashdot. He has published over 100 publications in areas of security and privacy, networked and distributed systems, wireless networks and data-intensive computing. Finally, he has served as program chair for top conferences (WOSN, WWW 2013 OSN track, IPTPS, IEEE P2P), and is a co-founder and steering committee member of the ACM Conference on Online Social Networks (COSN).