Do Neighbor Buddies Make a Difference in Reblog Likelihood? An Analysis on SINA Weibo Data

Lumin Zhang[†] Jian Pei[‡] Yan Jia[†] Bin Zhou[†] Xiang Wang[†]
[†]School of Computer Science, National University of Defense Technology, China
zlm.nudt@gmail.com, {jiayan,binzhou,wangxiang}@nudt.edu.cn
[‡]School of Computing Science, Simon Fraser University, Burnaby, BC, Canada
jpei@cs.sfu.ca

***Abstract*—Reblogging, also known as retweeting in Twitter parlance, is a major type of activities in many online social networks. Although there are many studies on reblogging behaviors and potential applications, whether neighbors who are well connected with each other (called "buddies" in our study) may make a difference in reblog likelihood has not been examined systematically. In this paper, we tackle the problem by conducting a systematic statistical study on a large SINA Weibo data set, which is a sample of $135,859$ users, $10,129,028$ followers, and $2,296,290,930$ reblog messages in total. To the best of our knowledge, this data set has more reblog messages than any data sets reported in literature. We examine a series of hypotheses about how essential neighborhood structures may help to boost the likelihood of reblogging, including buddy neighbors versus buddyless neighbors, traffic between buddy neighbors, activeness (i.e., the total number of blog messages a user sends), and the number of buddy triangles a user participates in. Our empirical study discloses several interesting phenomena that are not reported in literature, which may imply interesting and valuable new applications.**

***Keywords*—Reblog, retweet, online social networks, neighborhood**

I. Introduction

Reblogging, also known as retweeting in Twitter parlance, is a major type of activities in many online social networks. Many online social networks support reblogging effectively and extensively in one way or another. Reblogging has been extensively used by many users. It is well known that retweeting is popularly used by twitter users. For example, Obama's victory tweet was retweeted $298,318$ times in 30 minutes[1]. Yang *et al.* [1] reported that about 25.5% of tweets are retweeted from friends' blog spaces. Moreover, Facebook allows users to re-share posts from other users' walls, which is essentially reblogging. Reblogging has generated a tremendous amount of traffic in SINA Weibo, a microblog online social network in China and akin to a combination of Twitter and Facebook. Some third-party service providers, such as retweet.co.uk, provide indexes of currently trending posts, hashtags, users, and lists in order to facilitate the retweet and reblog protocol.

Reblogging is an important and effective mechanism to boost dynamics in online social networks. It facilitates online social network analysis and applications dramatically. For example, LeBleu [2] modeled reblogging as virtual currency in online social networks. A user reposting can be viewed as gaining influence currency. Arrington [3] from the web search engines point of view regarded in the other way, that is, a user reposting means a loss of influence currency. Using such models, important users can be identified accordingly.

It is essential to understand reblogging in online social networks. As to be reviewed in Section II, a series of studies in literature focused on characterizing reblogging patterns, such as content, network influence, temporal decay factor, and user reputation. However, an important aspect, neighborhoods, has not been explored systematically.

In an online social network, users are connected by links formed by, for example, friendship or follower relations. Consequently, every user has its neighborhood, and sends blog messages and reblog messages to its neighborhood. It is well recognized that neighborhood is an important structural feature in online social networks. Neverthless, the effect of neighborhood on reblog likelihood has not been systematically examined.

In this paper, we tackle this interesting and important problem. We focus on "buddy users" who are also known as reciprocities in sociology. Two users are **buddies** (or in **buddy relation**) if they follow each other. Two buddy users represent close and two-way strong connections between them. Heuristically, using the buddy relation we can reduce the ill-effect of spamming users, fake users, and inactive users substantially, as active real users are unlikely to build active buddy relations with many spamming, fake, or inactive users.

In a social network where a link between two users indicates that the two users are buddies, we are particularly interested in several fundamental properties of neighborhoods formed by buddies and their effect on reblog likelihoods. Particularly, we investigate four questions.

- (Buddy neighbors) For a user u and its neighbors v and w, does v and w being buddies help to improve the likelihood that v and w reblog messages from u?
- (Traffic) For a user u and its neighbors v and w who are buddies, does the amount of traffic between v and w, i.e., the total number of blog messages sent

The research was supported in part by National Basic Research Program of China (973 Program, No. 2013CB329601), National Natural Science Foundation of China (No. 91124002), an NSERC Discovery Grant and a BCIC NRAS Team Project. All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

[1] http://www.examiner.com/article/justin-bieber-retweet-record-falls-to-president-barack-obama-s-victory-tweet

between v and w, affect the likelihood that v and w reblog messages from u?

- (Activeness) For a user u, does the activeness of u, i.e., the total number of blog messages sent by u, affect the probability that u's neighbors reblog messages from u?
- (Triangles) For a user u, does u and its neighbors forming many buddy triangles affect the likelihood that u's neighbors reblog messages from u?

Investigating the above questions is far from trivial. Collecting a large sample data set about reblogging from a popular online social network is challenging. Fortunately, we are able to obtain a large sample of SINA Weibo, a microblog online social network in China and akin to a combination of Twitter and Facebook. Our data set contains $135,859$ users, $10,129,028$ followers, and $2,296,290,930$ reblog messages in total. Moreover, in this data set, only "buddy user pairs" are recorded, that is, two users are connected if and only if they follow each other. To the best of our knowledge, this data set contains more reblog messages than any data sets reported in literature.

The rest of the paper is organized as follows. We review the related work in Section II. In Section III, we define some preliminaries and describe the SINA Weibo data set to be used in our study. In Section IV, we investigate whether and how much buddy neighbors may be more likely to reblog. In Section V, we examine how the activeness of a user may affect how likely the user's neighbors may reblog the messages from the user. In Section VI, we analyze the effect of participation in buddy triangles on reblog likelihood. We conclude the paper in Section VII.

II. Related Work

This paper is related to the existing studies on three aspects: structural characteristics of online social networks, reblogging and retweeting behaviors, and applications of reblogging and retweeting. A thorough and detailed review of those aspects is unfortunately far beyond the capacity of this paper. Instead, we provide here a brief review of some state-of-the-art results, and discuss the differences between this study and the existing ones.

A. Structural Characteristics of Online Social Networks

Due to the emerging popularity of online social networks, such as Twitter, Facebook, and LinkedIn, many existing studies were dedicated to the characteristics of online social networks in general and in specific, such as [4], [5], [6], [7]. Since in this study we are mainly concerned about the relation between the neighborhood structures formed by links and reblog likelihood in online social networks, we briefly review some existing works about link formation in online social networks.

Links, also known as social ties in some studies, play an important role in the structures of networks. Romero and Kleinberg [8] systematically studied the directed closure process in information networks. They found that directed links, such as from c to a, are frequently formed by "short-cutting" a length-2 path between the source and the destination, such as c to b and b to a, which is an implicit "link copying" analogous to the process of triadic closure in social networks. Romero *et al.* [9] further analyzed the competing factors and their interplay on the relationship between two users in an online social network when they develop mutual relationships with third parties, and associated the underlying issues to the classical principles in sociology, including the theories of balance, exchange, and betweenness.

Yin *et al.* [10] studied the formation of the follower relationship in Twitter and found that 90% of new links are to people of just two hops. Hopcroft *et al.* [11] focused on the problem of reciprocal relationship prediction, and developed a Triad Factor Graph (TriFG) model, which incorporates social theories into a semi-supervised leaning model, to infer whether user A will follow back user B after B creating a following link to A. Kwak *et al.* [12] studied the "unfollowing" (i.e., removing a previous follower relation) dynamics in Twitter, and discovered that the major factors affecting the decision to unfollow are reciprocity of the relationships, the duration of a relationships, the followee's informativeness and the overlap of the relationship. Meeder *et al.* [13] inferred social link creation times in Twitter using a single static snapshot of network edges and user account creation times.

Many studies investigated various aspects of structural characteristics. In this study, we explore the relation between neighborhood structures and reblog likelihood. To the best of our knowledge, this issue has not been addressed in any previous studies in a systematic manner.

B. Reblogging and Retweeting Behaviors

A series studies analyzed reblogging and retweeting behaviors. For example, Yang *et al.* [1] found that about 25.5% of the tweets are actually retweeted from friends' blog spaces. Kwak *et al.* [4] found that a retweeted tweet in Twitter on average reaches $1,000$ users, independent to the number of followers of the original tweet. On average, a tweet, once retweeted, gets retweeted almost instantly on next hops. This phenomenon of fast diffusion of information after the first retweet is significant.

Peng *et al.* [14] used conditional random fields to model the retweeting patterns. They considered the content influence, the network influence and the temporal decay factor. Suh *et al.* [15] performed large scale analyses on various factors impacting retweeting in Twitter. They found that URLs and hashtags have strong relationships with retweetability. The number of followers and followees as well as the age of the account seem to also affect retweetability. Luo *et al.*[16] found that followers who retweeted or were mentioned before and have common interests with the author are more likely to be retweeters.

In this paper, we also reveal some characteristics about the retweeting behaviors with respect to the friend-follower relationships in SINA Weibo, a Twitter like platform in China. Different from the previous works, we focus on the reblog likelihood with respect to different patterns of neighborhood structures, which to the best of our knowledge has not been explored in any previous works.

C. Applications of Reblogging and Retweeting

Reblogging and retweeting, as an common type of interaction in online social networks, also have many applications, such as information diffusion and influence maximization.

Based on retweeting behaviors, Macskassy and Michelson [17] developed four information diffusion models to discover what information is being spread and why. Yang and Counts [18] developed a method to predict the speed, scale and range of information propagation in Twitter using retweeting and reply behaviors. The number of messages retweeted by others is often regarded as an important factor in evaluating the influence of a user [19], [20] and a message [21].

Yang *et al.* [22] focused on discovering interesting posts in Twitter by mining a retweet graph, where users and posts are nodes and retweeting relations between the nodes are edges. Ota *et al.* [23] discovered interesting users by leveraging overlapping propagation paths of retweets. They built an overlap graph, which contains users sharing same retweets, and validated users according to the frequencies and content of the retweets. Gupta *et al.* [24] studied the fake images during Hurricane Sandy and found that 86% of tweets spreading the fake images were retweets, and the top 30% users out of $10,215$ resulted in 90% of the retweets of fake images.

The previous studies on applications of reblogging and retweeting did not consider neighborhoods. Our study focuses on the buddy relationship and reveals interesting findings about how neighborhoods may affect reblog likelihood. The new findings imply some interesting and valuable new applications.

III. Preliminaries and the SINA Weibo Data Set

In this section, we first define the preliminaries, particularly the buddy relation. Then, we describe the SINA Weibo data set used in our study.

A. Preliminaries

We consider an online social networks of users where a user can **follow** another user. The follower relation is directed. As discussed in Section I, we focus on buddy users. Two users u and v are called **buddies** or in **buddy relation** if u follows v and v follows u. The **buddy relation** is undirected. Hereafter, we denote by $\mathcal{B}$ the buddy relation. The buddy relation $\mathcal{B}$ can be represented in a **buddy graph** G_B, where each node is a user, and there is an edge (u, v) between users u and v if $(u, v) \in \mathcal{B}$.

Heuristically, using the buddy relation we can reduce the ill-effect of spamming users, fake users, and inactive users substantially. The rationale is that many active real users are unlikely to build active buddy relations with many spamming, fake, or inactive users.

Hereafter, by default our discussion is based on a buddy graph.

For a user u, user v is u's **neighbor** if (u, v) is an edge in the buddy graph. Let $N(u)$ be the set of neighbors of u. We consider two essential types of neighbors.

- For a user u, a user v is called a **buddy neighbor** of u if v is a neighbor of u and there exists another neighbor w of u such that v and w are buddies.
- For a user u, a user v is called u's **buddyless neighbor** if v is a neighbor of u and there does not exist another neighbor w of u such that v and w are buddies.

Let $N_b(u)$ be the set of buddy neighbors of u, and $N_{bl}(u)$ the set of buddyless neighbors of u. Clearly, we have $N_{bl}(u) \cap N_b(u) = \emptyset$ and $N(u) = N_{bl}(u) \cup N_b(u)$.

We denote by $M(u)$ the number of blog messages that u posts, and $R_u(v)$ the number of blog messages posted by u and reblogged by v. Then, the likelihood that v reblogs u is calculated by

$$Pr_u(v) = \frac{R_u(v)}{M(u)}. \tag{1}$$

The average likelihood that a message from u is reblogged by a neighbor of u is calculated by

$$Pr_u(N(u)) = \frac{1}{|N(u)|} \sum_{v \in N(u)} Pr_u(v). \tag{2}$$

Similarly, the average likelihood that a message from u is reblogged by a buddy neighbor of u is

$$Pr_u(N_b(u)) = \frac{1}{|N_b(u)|} \sum_{v \in N_b(u)} Pr_u(v).$$

Last, for a user u, let $\Delta(u)$ be the number of **buddy triangles** that u is involved, that is,

$$\Delta(u) = |\{(v, w) \mid v, w \in N_b(u), (v, w) \in \mathcal{B}\}|.$$

B. SINA Weibo Data

In the rest of this paper, we use a large corpus collected from SINA Weibo. As there are many spamming users in online social networks, we used the following strategy to collect data for our experiments.

We started from user *Kaifu Lee*, a famous and active user in SINA Weibo, and used him to initiate the seed user set U. Iteratively, we crawled the buddies for each user $u \in U$ and formed the set $N(u)$. We then formed a follower set $V = \cup_{u \in U} N(u)$. We set $U = U \cup V$, and repeated the above process. The iterative process stop after 3 iterations.

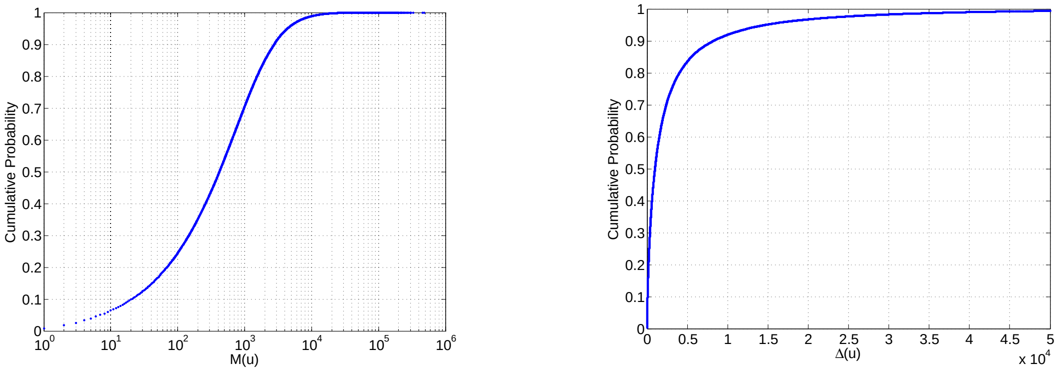
The data set we obtained contains $135,859$ users and $10,129,028$ followers. Please note that for the users only in the follower set obtained at the last iteration, our data set does not contain any follower information about them. This is the reason we have to separate the set of followers. We further extract the buddy neighbors and buddyless neighbors for every user in the data set.

We scanned all the messages posted by those users to extract the reblog messages, and identified $2,296,290,930$ reblog messages in total. To the best of our knowledge, this data set has more reblog messages than any data sets reported in literature.

Figure 1 shows the cumulative distribution of $M(u)$ (the number of messages posted by user u) and $\Delta(u)$ (the number of buddy triangles participated by u) for the 135,589 users. 92.32% users posted less than $10,000$ messages, and 90% users participated in less than $10,000$ buddy triangles.

IV. Buddy Neighbors

In general, we are interested in whether buddy neighbors may be more likely to reblog than buddyless neighbors. In this section, we pursue this investigation in two steps. The first step is qualitative and the second step is quantitative.



(a) Number of messages.

(b) Number of buddy triangles.

Fig. 1. Cumulative distribution of $M(u)$ and $\Delta(u)$.

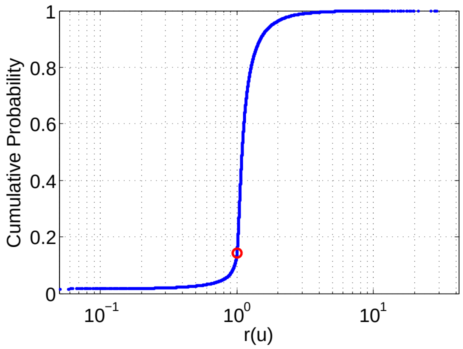


Fig. 2. Cumulative distribution of $r(u)$.

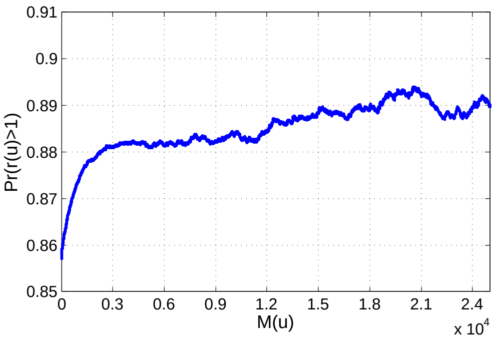


Fig. 3. Distribution of $Pr(r(u) > 1)$ with respect to users of $M(u) \geq \sigma$.

A. A Qualitative Study

We start with a qualitative question, that is, whether $Pr_u(N_b(u)) > Pr_u(N_{bl}(u))$. This is equivalently to examine whether $Pr_u(N_b(u)) > Pr_u(N(u))$.

One natural conjecture behind this question is that buddy neighbors may share common interest and interactions. Consequently, given a pair of buddy neighbors v and w, when v reblogs a message sent by u, the other neighbor w receives the message twice at least, one time from u and the other time from v, and may be more likely to reblog the message due to w's common interest with u and v.

To examine whether buddy neighbors may be more likely to reblog, we propose the following hypothesis.

Hypothesis 1 (Buddy neighbors): Buddy neighbors are more likely to reblog than buddyless neighbors.

For each user u, we define

$$r(u) = \frac{Pr_u(N_b(u))}{Pr_u(N(u))}.$$

When $Pr_u(N(u)) = 0$ and thus $Pr_u(N_b(u)) = 0$, we set $r(u) = \frac{0}{0} = 1$. Please note that this is the only possible situation where $Pr_u(N(u)) = 0$. In our data set, the expectation of $r(u)$, $E(r(u)) = 1.169$. Figure 2 shows the cumulative distribution for ratio $r(u)$.

Treating each user as an independent experiment of two possible outcomes: $r(u) \leq 1$ and $r(u) > 1$, we set the null hypothesis to "*The probability of* $r(u) \leq 1$ *is the same as that of* $r(u) > 1$, *i.e.,* 0.5."

Among $135,895$ users, $116,458$ have $r(u) > 1$. Therefore, the p-value of the null hypothesis is less than 2.2×10^{-16}. The null hypothesis is rejected. In other words, with strong confidence Hypothesis 1 holds. Buddy neighbors are more likely to reblog than buddyless neighbors.

We sort all users according to their $M(u)$ values, and calculate $Pr(r(u) > 1)$ for users whose $M(u) \geq \sigma$, that is, the percentage of users of $r(u) > 1$ among the users who posted at least σ messages, where σ is a threshold. Figure 3 shows the results. Interestingly, in general, $Pr(r(u) > 1)$ increases as $M(u)$ increases. For a user who posted many blog messages, the user's buddy neighbors are more likely to reblog those messages than buddyless neighbors. Please note that $Pr(r(u) > 1)$ is not a monotonically increasing function of $M(u)$. Thus, $Pr(r(u) > 1)$ shows a decreasing trend in some interval with the increase in $M(u)$ in Figure 3.

Knowing buddy neighbors are more likely to reblog, an obvious application is friend recommendation. We can recommend a user u to a pair of users (v, w) who follow each other, instead of just recommending user u to v but not w or the other way.

B. A Quantitative Analysis

Hypothesis 1 qualitatively tests whether buddy neighbors may be more likely to reblog. Next, we quantitatively examine

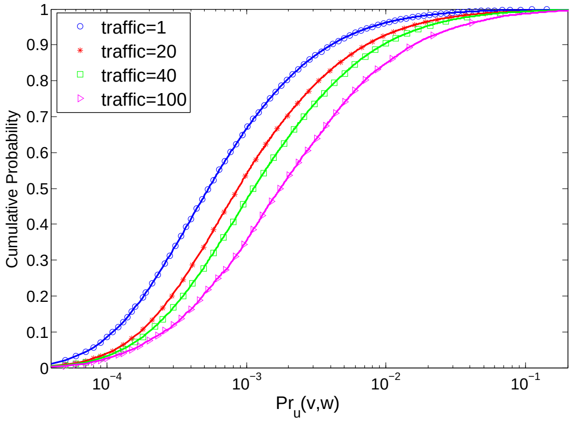


Fig. 4. Cumulative distribution of repost probability for different value of traffics.

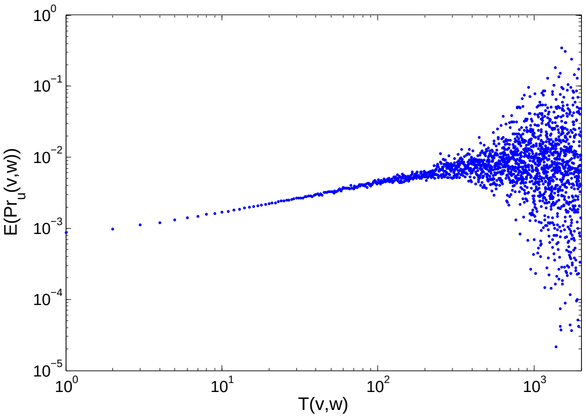


Fig. 5. Reblog likelihood versus traffic.

how the strength of connections between buddy neighbors may affect the reblog likelihood.

For a user u and its buddy neighbors v and w, that is, $(v, w) \in \mathcal{B}$, we define the **traffic volume** (**traffic** for short) between v and w as

$$T(v, w) = R_v(w) + R_w(v).$$

Here, we use $R_v(w)$ and $R_w(v)$ to measure the amounts of messages sent by one node and well received (in fact, reblogged) by the other. Heuristically, those numbers approach the effective traffic volume between the two nodes.

In order to investigate the relation between traffic and reblog likelihood, we need to measure the reblog likelihood of a pair of buddy neighbors. Specifically, for a user u and a pair of buddy neighbors v and w, i.e., $(v, w) \in \mathcal{B}$, we define the **average reblog likelihood** of v and w as

$$Pr_u(v, w) = \frac{Pr_u(v) + Pr_u(w)}{2}.$$

With large traffic volume $T(v, w)$, v and w may share more similar interest. Consequently, we conjecture that the average reblog likelihood of a pair of buddy neighbors is positively correlated to the traffic. Formally, we propose the following hypothesis.

Hypothesis 2 (Traffic): The average reblog likelihood of buddy neighbors is positively correlated with the traffic between buddy neighbors.

To examine the hypothesis, we consider every tuple $(u, (v, w))$ where v and w are neighbors of u, and v and w are buddies. We obtain $(T(v, w), Pr_u(v, w))$ as an independent sample.

First, we examine the distribution of $Pr_u(v, w)$ on users with the same traffic $T(v, w)$. Figure 4 plots the cumulative percentage of $Pr_u(v, w)$ for buddy user pairs (v, w) of traffic $T(v, w) = 1$, 20, 40, and 100. In each curve, we sort the $Pr_u(v, w)$ values in ascending order, and plot at position γ at the horizontal axis the cumulative percentage of points falling in range $(0, \gamma)$.

The four curves follow similar trends. Interestingly, the smaller the traffic $T(v, w)$, the faster the cumulative percentage curve increases. This indicates that a pair of buddy neighbors of larger traffic tend to be more likely to reblog. This is consistent with Hypothesis 1.

Figure 5 plots $E(Pr_u(v, w))$ with respect to $T(v, w)$. That is, if multiple tuples $(u, (v, w))$ have the same traffic $T(v, w)$, we calculate the expectation $E(Pr_u(v, w))$ and plot in the figure. Please note that the figure is in log-log scale.

We observe that 99% of the tuples have traffic no more than 200. When the traffic is in the range of $[0, 200]$, the trend is close to a line in the log-log scale. Therefore, we conduct a regression using function

$$\ln E(Pr_u(v, w)) = \beta \ln T(v, w) + c$$

in the traffic range $[0, 200]$, where β and c are parameters to be determined from data. This is equivalent to:

$$E(Pr_u(v, w)) = \alpha T(v, w)^{\beta} \tag{3}$$

with $\alpha = e^c$. The regression result is $\alpha = 6.85 \times 10^{-4}$ (SE: 2.55×10^{-5}), $\beta = 0.4017$ (SE: 0.0078), and $R^2 = 0.9562$.

Since the coefficient of multiple determination R^2 is close to 1, there is a significant correlation between the reblog likelihood and traffic. As $\beta = 0.4017 > 0$, the dependent variable $E(Pr_u(v, w))$ increases as traffic increases.

Hypothesis 2 provides a hint in choosing the pair users when recommending a given user u. It is better to recommend u to a pair of buddy users (v, w) with high traffic in between.

Before we leave this section, let us consider one more question related. Does $M(u)$, the number of messages posted by user u, affect the correlation? That is, do users of different $M(u)$ values hold different correlation patterns? If the answer is yes, we should adopt different strategies for different users.

According to the distribution of $M(u)$ in Figure 1(a), we divide the users with different $M(u)$ values in to 6 groups:

1) those of $M(u) \leq 100$;
2) those of $100 < M(u) \leq 500$;
3) those of $500 < M(u) \leq 1,000$;
4) those of $1000 < M(u) \leq 5,000$;
5) those of $5000 < M(u) \leq 10,000$; and
6) those of $M(u) > 10,000$.

$M(u) \leq 100$			$100 < M(u) \leq 500$		$500 < M(u) \leq 1000$		$1000 < M(u) \leq 5000$		$5000 < M(u) \leq 10000$		$M(u) > 10000$	
	Value	SE	Value	SE	Value	SE	Value	SE	Value	SE	Value	SE
α	4.24E-02	3.25E-03	3.57E-03	3.43E-04	1.24E-03	1.13E-04	4.98E-04	2.67E-05	3.07E-04	1.25E-05	2.02E-04	9.62E-06
β	0.1740	0.0166	0.4019	0.0201	0.4984	0.0190	0.4843	0.0111	0.4152	0.0085	0.4149	0.0100
R^2	0.6235		0.7698		0.8586		0.9418		0.9515		0.9342	

TABLE I. REGRESSION RESULT FOR DIFFERENT $M(u)$.

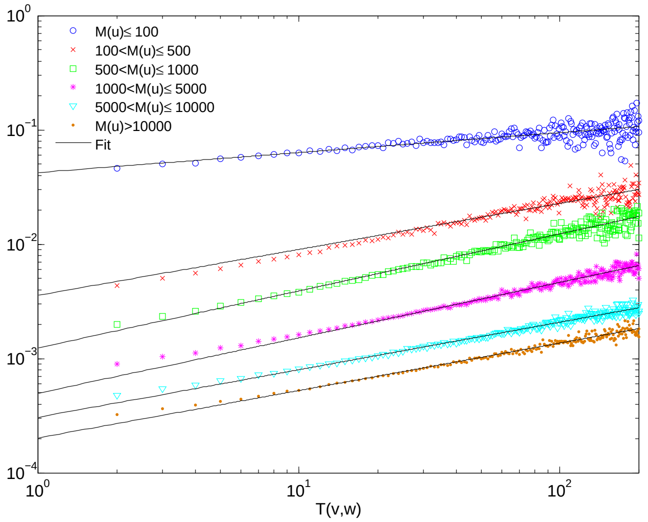


Fig. 6. Reblog likelihood versus traffics for different groups of users in $M(u)$.

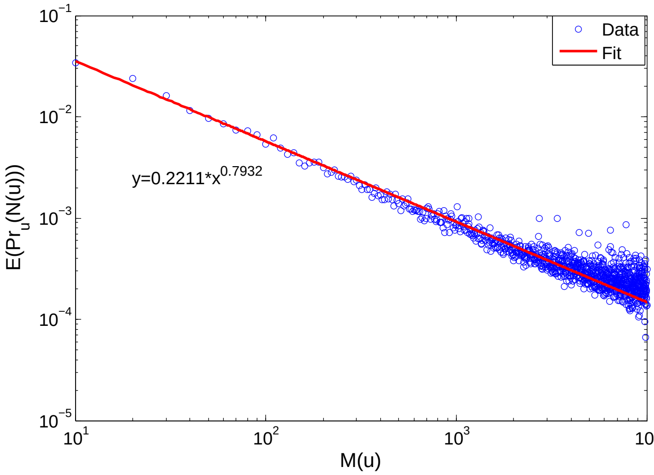


Fig. 7. Expectation of reblog likelihood with respect to $M(u)$.

For each group, we use the regression model in Equation 3 to study the correlation between reblog likelihood and traffic.

Table I shows the regression results for different groups. Figure 6 shows the reblog likelihood with respect to traffic in different groups in log-log scale. All groups with different $M(u)$ ranges follow similar patterns. They only differ in parameters. As can be seen from Figure 6, users with larger $M(u)$ values tend to have a smaller reblog likelihood in expectation. We will systematically study this problem in the next section.

Interestingly, the parameter $\beta = 0.1740$ for $M(u) \leq 100$ is much less than those in the other groups. The β values in other groups are all larger than 0.4. That is, the reblog likelihood increases slower with respect to traffic for users with $M(u) \leq 100$ than other user groups. A possible explanation is that users posting a small number of messages are less active than those posting many messages. Those less active users may have less interactions with other users. Therefore, the other users may not be sensitive to their messages.

Variable	Value	Standard Error	t-ratio	R^2	F-ratio
λ	0.2211	1.86E-03	119.00	0.9882	84238.58
γ	0.7932	2.26E-03	350.84		

TABLE II. REGRESSION RESULT ON $E(Pr_u(N(u)))$ VERSUS $M(u)$.

V. ACTIVENESS OF USERS

It is natural to ask whether a more active user, who posts more blog messages, may have a higher likelihood that her/his messages are reblogged by her/his neighbors. Technically, we propose the following hypothesis.

Hypothesis 3 (Activeness): For a user u, $Pr_u(N(u))$ is positively correlated with $M(u)$.

Figure 7 shows the expectation of reblog likelihood $E(Pr_u(N(u)))$ with respect to $M(u)$ in log-log scale. Surprisingly, there is a significant negative correlation between $E(Pr_u(N(u)))$ and $M(u)$. We use the following power-law formula to conduct a regression analysis.

$$E(Pr_u(N(u))) = \lambda M(u)^{-\gamma} \tag{4}$$

where λ and γ are positive parameters.

Table II shows the result of the regression. The coefficient of multiple determination $R^2 = 0.9882$, showing a significant correlation between $E(Pr_u(N(u)))$ and $M(u)$. Moreover, the parameter $\gamma = 0.7932 < 1$, which means $E(Pr_u(N(u)))$ decreases slower than $\frac{1}{M(u)}$.

For a user u, let

$$C_u(N(u)) = \frac{1}{|N(u)|} \sum_{v \in N(u)} R_u(v)$$

be the average number of times that u's messages are reblogged by a neighbor of u. Using Equations 1 and 2, we have

$$Pr_u(N(u)) = \frac{C_u(N(u))}{M(u)} \tag{5}$$

Since $E(Pr_u(N(u)))$ decreases slower than $\frac{1}{M(u)}$, when $M(u)$ increases, $E(C_u(N(u)))$ increases, too. In other words, expectation $E(C_u(N(u)))$ is also positively correlated with $M(u)$. This explains our observation that more messages posted by a user u, more times u's messages are reblogged on average. Importantly, the expected number of times

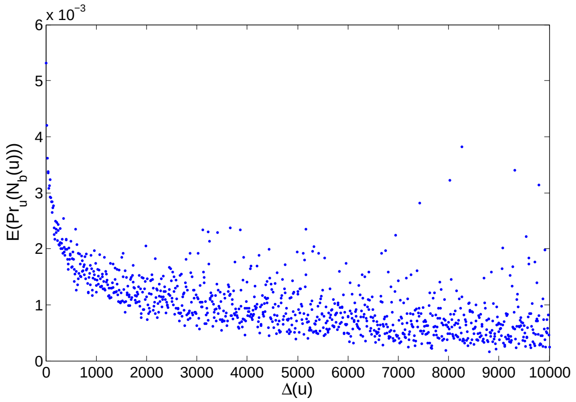


Fig. 8. Expectation of reblog likelihood with respect to the number of buddy triangles participated in.

Variable	Value	Standard Error	t-ratio	R^2	F-ratio
λ_b	0.2321	1.85E-03	125.77	0.9890	90235.71
γ_b	0.7706	2.07E-03	371.57		

TABLE III. REGRESSION RESULTS ON $E(Pr_u(N_b(u)))$ VERSUS $M(u)$.

$E(C_u(N(u)))$ that u's messages are reblooged by u's neighbors increases slower than $M(u)$, resulting in the decrease of $E(Pr_u(N(u)))$ when $M(u)$ increases.

Analytically, we can estimate the correlation between $E(C_u(N(u)))$ and $M(u)$. Specifically, we assume a function h such that

$$E(C_u(N(u))) = h(M(u)).$$

According to Equation 5, we have

$$h(M(u)) = M(u) * \lambda M(u)^{-\gamma} = \lambda M(u)^{1-\gamma}$$

It means that $h(M(u))$ should grow as fast as $O(M(u)^{\kappa})$, where $\kappa = 1 - \gamma = 0.2068$.

We further study the correlation between the expected likelihood that u's messages are reblogged by a buddy neighbor, i.e., $E(Pr_u(N_b(u)))$, and the number of messages posted by u. The correlation follows a pattern similar to that between $E(Pr_u(N(u))$ and $M(u))$. Specifically, we denote the correlation between $E(Pr_u(N_b(u)))$ and $M(u)$ by

$$E(Pr_u(N_b(u))) = \lambda_b M(u)^{-\gamma_b},$$

where λ_b and γ_b are positive parameters.

Table III shows the regression results. Interestingly, comparing the results in Tables II and III, the exponent parameters $\gamma_b < \gamma$ and the parameters $\lambda_b > \lambda$. By taking derivative of Equation 4, we have

$$\frac{\mathrm{d}E(Pr_u(N(u)))}{\mathrm{d}M(u)} = -\lambda\gamma M(u)^{-(1+\gamma)}$$

As $\lambda_b\gamma_b > \lambda\gamma$ and $-(1+\gamma_b) > -(1+\gamma)$, we have

$$\frac{\mathrm{d}E(Pr_u(N(u)))}{\mathrm{d}M(u)} > \frac{\mathrm{d}E(Pr_u(N_b(u)))}{\mathrm{d}M(u)}$$

Therefore, $E(Pr_u(N_b(u)))$ decreases slower than $E(Pr_u(N(u)))$, which means that the reblog likelihood of buddy neighbors decreases slower than that of all neighbors.

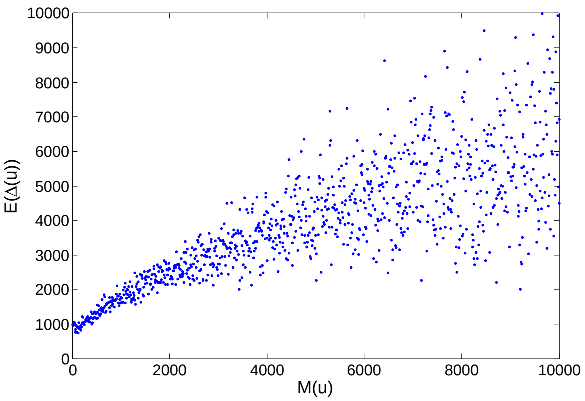


Fig. 9. Expectation of buddy triangles participated in with respect to number of messages.

VI. The Effect of Buddy Triangles

Interestingly, for a user u and neighbors v and w such that v and w are buddies, due to the construction of the buddy graph, where each edge represents a pair of buddies, u and v as well as u and w are buddies, too. Therefore, u, v, and w form a **buddy triangle**.

It is well recognized that triangles play an important role in social networks. For example, clustering coefficient, a popularly used feature in graph theory and social network analysis, is based on the number of triangles. If Hypotheses 1 and 2 hold, it is interesting to examine whether the more buddy triangles a user participates in, the more likely the user's blog messages are reblogged by her/his buddy neighbors. We propose the following hypothesis.

Hypothesis 4 (Triangles): For a user u, $P_u(N_b(u))$ is positively correlated with $\Delta(u)$.

Figure 8 shows the distribution of the expectation of reblog likelihood for buddy neighbors $E(Pr_u(N_b(u)))$ with respect to the number of buddy triangles $\Delta(u)$. Clearly, there is a negative correlation between $E(Pr_u(N_b(u)))$ and $\Delta(u)$. The more buddy triangles that u participates in, the less likelihood that u's messages are reblogged in expectation. In other words, Hypothesis 4 does not hold in our data set.

To investigate the possible reason that results in the negative correlation between $E(Pr_u(N_b(u)))$ and $\Delta(u)$, we further test the correlation between the number of messages u posts, $M(u)$, and the number of buddy triangles $\Delta(u)$ that u participates in. Figure 9 shows the distribution of $\Delta(u)$ with respect to $M(u)$. It is interesting that there is a positive correlation between these two variables. Users who post a large number of messages are more likely to participate in many buddy triangles. Consequently, users with a large $\Delta(u)$ value are often have a large $M(u)$ value, which leads to a small $Pr_u(N_b(u))$ value in expectation according to negative correlation between $Pr_u(N_b(u))$ and $M(u)$ studied in section V.

As we know, if a user u has n neighbors, u may participate in up to $\frac{n(n-1)}{2}$ buddy triangles. Since the number of possible buddy triangles increases in a higher order than the number of neighbors, the increases of the reblog likelihood by buddy neighbors may not be able to catch up with the increase of the number of buddy triangles participated in. This is another possible explanation about why Hypothesis 4 does not hold.

We now turn to the positive correlation between $\Delta(u)$ and $M(u)$ shown in Figure 9. According to the study of Romero and Kleinberg [8], if v follows u and w follows v, then w is more likely to follow u to form a triangle. This is also true for buddy relationship. Our findings here indicate that buddy friends are frequently formed by a common buddy friend. Therefore, if u posts more messages, the number of messages v reblogs from u also increases. Then, w has a higher probability to see u's messages and becomes a buddy friend of u, resulting in a positive correlation between $\Delta(u)$ and $M(u)$.

VII. Conclusions

Reblogging is an important and popularly used mechanism in online social networks. In this paper, we examined whether buddy neighbors may make a difference in reblog likelihood. We used a large data set from SINA Weibo to systematically analyze this interesting problem.

We specifically studied several interesting questions. Our analysis suggested that buddy neighbors are more likely to reblog, and the more traffic between a pair of buddy neighbors, the more likely they reblog their common friends they both follow. Moreover, the more active a user and more messages the user posts, the more times that the user's message are reblogged yet the less reblog likelihood for each message. At the same time, it is surprising to find that the more buddy triangles that a user participates in, the less likely that the user's messages are reblogged by the buddy neighbors.

Our study serves as the first step towards utilizing buddy neighbors and reblogging in social network analysis and applications. As future work, we plan to develop more specific metrics considering various specific mechanisms, such as different types of friendship based on classmates or common interests. Developing recommendation methods based on buddy neighbors and features of reblogging statistics is also an interesting direction.

References

[1] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su, "Understanding retweeting behaviors in social networks," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 1633–1636.

[2] G. Lebleu, (2009, January) "Please ReTweet": RT as currency and Twitter social ad business model. [Online]. Available: http://lebleu.org/blog/2009/01/08/please-retwit-rt-as-currency-and-twitter-social-ad-business-model/.

[3] M. Arrington, (2009, May 26) (TechCrunch) Topsy Search: ReTweets Are The New Currency Of The Web. [Online]. Available: http://techcrunch.com/2009/05/26/topsy-search-launches-retweets-are-the-new-currency-of-the-web/.

[4] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 591–600.

[5] M. Cha, A. Mislove, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the flickr social network," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 721–730.

[6] A. Java, X. Song, T. Finin, and B. Tseng, "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*. ACM, 2007, pp. 56–65.

[7] A. Anagnostopoulos, R. Kumar, and M. Mahdian, "Influence and correlation in social networks," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 7–15.

[8] D. M. Romero and J. M. Kleinberg, "The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter." in *ICWSM*, 2010.

[9] D. M. Romero, B. Meeder, V. Barash, and J. M. Kleinberg, "Maintaining ties on social media sites: The competing effects of balance, exchange, and betweenness." in *ICWSM*, 2011.

[10] D. Yin, L. Hong, X. Xiong, and B. D. Davison, "Link formation analysis in microblogs," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 2011, pp. 1235–1236.

[11] J. Hopcroft, T. Lou, and J. Tang, "Who will follow you back?: reciprocal relationship prediction," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 1137–1146.

[12] H. Kwak, H. Chun, and S. Moon, "Fragile online relationship: a first look at unfollow dynamics in twitter," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, pp. 1091–1100.

[13] B. Meeder, B. Karrer, A. Sayedi, R. Ravi, C. Borgs, and J. Chayes, "We know who you followed last summer: inferring social link creation times in twitter," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 517–526.

[14] H.-K. Peng, J. Zhu, D. Piao, R. Yan, and Y. Zhang, "Retweet modeling using conditional random fields," in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. IEEE, 2011, pp. 336–343.

[15] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be retweeted? large scale analytics on factors impacting retweet in twitter network," in *Social computing (socialcom), 2010 ieee second international conference on*. IEEE, 2010, pp. 177–184.

[16] Z. Luo, M. Osborne, J. Tang, and T. Wang, "Who will retweet me?: finding retweeters in twitter," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2013, pp. 869–872.

[17] S. A. Macskassy and M. Michelson, "Why do people retweet? anti-homophily wins the day!" in *ICWSM*, 2011.

[18] J. Yang and S. Counts, "Predicting the speed, scale, and range of information diffusion in twitter." *ICWSM*, vol. 10, pp. 355–358, 2010.

[19] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, "Measuring user influence in twitter: The million follower fallacy." *ICWSM*, vol. 10, pp. 10–17, 2010.

[20] A. Pal and S. Counts, "Identifying topical authorities in microblogs," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 45–54.

[21] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," in *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*. IEEE, 2010, pp. 1–10.

[22] M.-C. Yang, J.-T. Lee, S.-W. Lee, and H.-C. Rim, "Finding interesting posts in twitter based on retweet graph analysis," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2012, pp. 1073–1074.

[23] Y. Ota, K. Maruyama, and M. Terada, "Discovery of interesting users in twitter by overlapping propagation paths of retweets," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on*, vol. 3. IEEE, 2012, pp. 274–279.

[24] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy," in *Proceedings of the 22nd international conference on World Wide Web companion*. International World Wide Web Conferences Steering Committee, 2013, pp. 729–736.