Who Influenced You? Predicting Retweet via Social Influence Locality

JING ZHANG, Tsinghua University
JIE TANG, Tsinghua University
JUANZI LI, Tsinghua University
YANG LIU, Tsinghua University
Chunxiao Xing, Tsinghua University

Social influence occurs when one's opinions, emotions, or behaviors are affected by others in a social network. However, social influence takes many forms and its underlying mechanism is still unclear. For example, how is one's behavior influenced by a group of friends who know each other and by the friends from different ego friend circles?

In this paper, we study the social influence problem in a large microblogging network. Particularly, we consider users' (re)tweet behaviors and focus on investigating how friends in one's ego network influence her retweet behaviors. We propose a novel notion of *social influence locality* and develop two instantiation functions based on pairwise influence and structural diversity. The defined influence locality functions have strong predictive power. Without any additional features, we can obtain a F1-score of 71.65% for predicting users' retweet behaviors by training a logistic regression classifier based on the defined influence locality functions. We incorporate the social influence locality into a factor graph model, which can further leverage the network-based correlation. Our experiments on the large microblogging network show that the model significantly improves the precision of retweet prediction.

Our analysis also reveals several intriguing discoveries. For example, if you have six friends retweeting a microblog, the average likelihood that you will also retweet it strongly depends on the structure among the six friends: the likelihood will significantly drop (only $\frac{1}{6}$) when the six friends do not know each other comparing with the case when the six friends know each other.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Sciences**]: Sociology; H.1.2 [**User/Machine Systems**]: Human factors

General Terms: Human Factors, Measurement, Experimentation,Performance

Additional Key Words and Phrases: social network, social influence, microblog network, retweet prediction

ACM Reference Format:
Zhang, J., Tang, J., Li, J., Liu, Y., Xing C. 2014. Who Influenced You? Predicting Retweet via Social Influence Locality ACM Trans. Knowl. Discov. Data. V, N, Article A (January 2014), 27 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

Social influence occurs when one's behaviors (or opinions and emotions) are affected by others in a social network. Understanding the mechanism of influence in a social

The work is supported by the National High-tech R&D Program (No. 2014AA015103), National Basic Research Program of China (No. 2014CB340506, No. 2011CB302302), Natural Science Foundation of China (No. 61222212), a research fund supported by Huawei Inc. and Beijing key lab of networked multimedia.
Author's addresses: J. Zhang, J. Tang, J. Li, Y. Liu, and C. Xing, Department of Computer Science and Technology, Tsinghua University.
Corresponding Author: Jie Tang, jietang@tsinghua.edu.cn

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

network can help capture the complex patterns that govern the dynamics of the social network [Tang et al. 2009].

At the high level, social influence has global patterns and local patterns. The former means that one's behavior is influenced by the global pattern in a social network. Examples of the global patterns include influence by a global culture [Robertson 1992] and conform to the majority opinion in a community [Tang et al. 2013], and influence from external network [Myers et al. 2012]. The latter, i.e., local patterns of social influence, means that one's behavior is influence by friends in her ego-network (refer to Section 2 for a formal definition). Examples of local patterns include pairwise influence [Goyal et al. 2010; Saito et al. 2008], indirect influence [Shuai et al. 2012], and topic-level influence [Liu et al. 2012; Tang et al. 2009]. Social influence is a fundamental issue in social network analysis and it can benefit many real applications. For example, based on the influence probability, Kempe et al. presented the notion of influence maximization [Kempe et al. 2003], an essential problem for viral marketing in the social network. Baskshy et al. [Bakshy et al. 2012] conducted two very large field experiments in Facebook and verified the strong effect of social influence on consumer responses to ads. Leskovec et al. [Leskovec et al. 2006] leveraged social influence to help improve the recommendation performance in the social network. Despise much research has been conducted in this field, the underlying mechanism of social influence is still unclear. One important reason is that social influence takes many forms and each form may be determined by different factors.

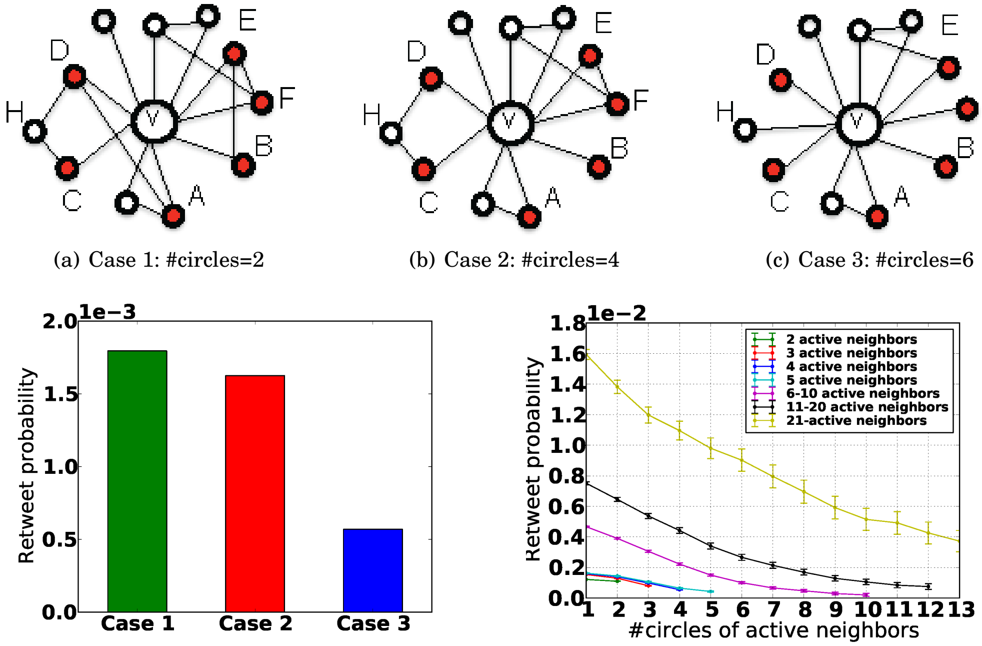
In this paper, we study the social influence problem in a large microblogging network, Weibo.com[1]. Specifically, we consider users' (re)tweet behaviors and focus on investigating how friends in one's ego network influence her retweet behaviors.

Illustrative example To clearly motivate this work, we first give a real example of retweet influence derived from the microblogging network (Weibo). We consider how a user's retweet behavior is influenced by her friends. Figure 1(a)-1(c) shows three similar cases when user v has six friends already retweeted a microblog (referred to as active neighbors and denoted as red nodes). A white node indicates the corresponding user (friend) does not retweet the microblog (referred to as inactive neighbor). The difference of the three cases is that the inner structure of the six friends is different. In Figure 1(a), the active neighbors A, C, and D are in one connected component, while B, E, and F are in another connected component. Thus the six active neighbors form two connected circles[2]. In Figure 1(b) and 1(c), the number of the connected circles formed by the six active neighbors are respectively four and six. In the three cases, we have totally collected 5,736,320 user samples in case1, 984,230 user samples in case2, and 805,889 user samples in case3.

We then study how likely user v will also retweet the microblog in the three different situations. We use a large microblogging network from Weibo to estimate the average probability that user v retweets the microblog in the three cases. We found several interesting patterns from the result as shown in Figure 1(d). The average likelihood that user v retweets the microblog strongly depends on the inner structure among the six friends: the probability that user v retweets the microblog in case 1 is three times ($3\times$) higher than that in case 3. We further expand the analyses to more cases by varying the number of active neighbors and the number of formed connected circles. Figure 1(e) shows the retweet probability by user v in various cases. It is very interesting that the retweet probability of user v is clearly negatively correlated with the number of connected circles formed by the active neighbors, no matter what kind of messages that

[1] The most popular Chinese microblogging service.

[2] The term *circle* comes from sociology to represent a group of socially interconnected people.



(a) Case 1: #circles=2

(b) Case 2: #circles=4

(c) Case 3: #circles=6

(d) Retweet probability in case 1,2 and 3

(e) Retweet probability in various cases

Fig. 1. Illustration of social influence locality for user v in her ego network. In (a) - (c), the node in the center of each network represents user v. Given a microblog m, red nodes represent "active" neighbors of user v who have retweeted m, while the white nodes denote those neighbors in v's ego network who did not retweet. Those active neighbors construct (a) 2 circles (b) 4 circles (c) 6 circles. Figure (d) shows the retweet probability of user v in the three cases. Figure (e) shows the retweet probability in more various cases.

have been retweeted by the neighbors. The retweet probabilities in this paper are all per user and are calculated by (#retweet users satisfying given conditions) / (#users satisfying given conditions). For example, to calculate the retweet probability in Figure 1(a), we first find all users with six followees who already retweeted one same microblog and meanwhile the six followees form 2 circles. Then among all these users, we denote the number of the users who also retweeted the same microblog as N^+ and the number of the users who did not retweet the microblog as N^-. Finally, the retweet probability is calculated as $N^+/(N^+ + N^-)$. The retweet probability is usally quite low because N^+ is often much smaller than N^-.

From the above example, it seems that users' retweet behaviors are strongly influenced by friends in her ego network. However, what are fundamental factors that trigger the phenomenon and why? In this paper, we formalize the problem as social influence locality and try to conduct a systematical investigation on the problem. The challenges of the problem are as follows:

— First, a straightforward question is: is there really influence between users for the retweet behavior? Figure 1 presents some intuitive explanation; however how to provide a theoretical proof.
— Second, how to formally define this type of influence using a principled function? It is necessary to give a formal definition of the phenomenon and provide its instantiation.
— Third, How to design a predictive model so that we can leverage the influence to predict users' retweet behaviors?

Results To address the above challenges, we first employ Weibo data as the basis in our study and present a debias sampling method to prove the existence of social influence for the retweet behaviors. We then propose a formal definition of social influence locality and develop two instantiation functions of social influence locality for modeling the retweet behaviors. The defined influence locality function has a strong predictive power. We employ it for modeling and predicting users' retweet behaviors. Predicting users' retweet behaviors have many applications. For example, it can help recommending important tweets to users. Since with the rapid increasing number of tweets, information overload becomes a serious problem, which makes it urgent to recommend most interesting tweets to users [Chen et al. 2012a; Feng and Wang 2013]. In addition, understanding how retweeting works can provide insight into how information spreads through large user communities [Petrovic et al. 2011] and also have applications in marketing, such as influence maximization [Kempe et al. 2003].

With merely a few features based on the influence locality functions, we could learn a simple classifier which results in good predictive performance, and is even better than existing methods that employ various features by +0.6% in terms of F1-measure. We further propose a factor graphic model for modeling and predicting users' retweet behaviors. The model not only considers traditional features for modeling users' retweet behaviors, but also incorporates social influence locality and network-based correlation into a probabilistic graphical model. Our experiments on the large microblogging network show that the model significantly improves the precision of retweet prediction. In addition, we also have several interesting findings:

- There is a strong evidence for the existence of social influence locality. The fraction of active users (retweeted a microblog) with 2 active neighbors (followees who have retweeted the same microblog) is about 2 times greater than the fraction of active users with only one active neighbors (Cf. Figure 3).
- Though the probability of a user retweeting a microblog is positively correlated with the number of active neighbors, it is negatively correlated with the number of *connected circles* that are formed by those neighbors. When there are six active neighbors, the likelihood will significantly drop (only $\frac{1}{6}$) when the six friends do not know each other comparing with case when the six friends know each other (Cf. Figure 1(e)).
- Pairwise influence differs from users. The retweet probability generally increases about 10% per 0.05 increase of the average pairwise influence from the active neighbors (Cf. Figure 5).

Compared with the previous conference version, the major improvements lie in that we conduct deeper analysis about the parameter τ of ego network, the whole data set distribution, the correlation feature and the basic features that influence the prediction performance. In addition, we explicitly give problem definition and add more concrete related work.

Organization. Section 2 formulates the problem of social influence locality. Section 3 describes the investigated data. Section 4 performs an investigation to test the existence of influence locality on retweet behaviors. Section 5 explains the instantiation functions for influence locality. Section 6 proposes the methods of influence locality based classification model to predict retweet behaviors. Section 7 presents experimental results of retweet behavior prediction. Finally, Section 8 reviews the related work and Section 9 concludes.

2. PROBLEM DEFINITION

In this section, we first give several necessary definitions and then present a formal definition of the problem.

A social network can be represented as $G = (V, E)$, where V is a set of $|V| = N$ users, $E \subseteq V \times V$ is a set of directed/undirected relationships between users, and $e_{ij} \in E$ represents a relationship between v_i and v_j. From the definition, we see that each user has a set of neighbors in the network. To make it more general, we give the definition of ego network.

Definition 2.1. **τ-ego network** For a user $v_i \in V$, we use $G_i^\tau \subseteq G$ to denote v_i's τ-ego network, which means a subnetwork formed by v_i's τ-degree friends in the network G, where $\tau \in \mathbb{N}$.

In the definition, τ is a tunable integer parameter to control the scale of the ego network. If $\tau = 1$, the τ-ego network is equal to the a subnetwork formed by user v_i's direct neighbors. On the other hand, according to the theory of six-degree separation [Milgram 1967], any two persons can be connected in a maximum of six steps. Therefore, the value of τ should not be too large, otherwise, the τ-ego network G_i^τ will be equal to the complete network G. In the definition, we can consider either bi-directional relationships or directional relationships. For example, for modeling the retweet behaviors in microblogging networks, we consider directed relationships between users. In addition, we give the following definition of retweet action.

Definition 2.2. **Retweet Action** We use a triple (v_i, t, m) to represent that user v_i retweets a microblog m at time t. For the microblog m, we denote all users' retweet actions as the action history $Y = \{v_i, t, m\}_{i,t}$. Further we denote $y_{i,m}^t$ as the action status of user v_i at time t for the given microblog m.

Without loss of generality, for the microblog m, we consider the binary action, i.e, $y_i^t \in \{0, 1\}$, where $y_i^t = 1$ indicates that user v_i performed a retweet action at time t, and $y_i^t = 0$ indicates that the user did not perform the retweet action. We call the user who performed a retweet action as active user, otherwise inactive user. Such an action log can be available in all the microblogging systems.

As one important goal of this work is to understand how users' retweet behaviors influence (or are influenced by) friends in their τ-ego network, we further define the notion of social influence locality.

Definition 2.3. **Social Influence Locality**
Suppose at time t, for the microblog m, user v_i has a set of active neighbors $N_{i,m}^t = \{v_j | v_j \in G_i^\tau \wedge y_{j,m}^t = 1\}$ in her ego network G_i^τ, social influence locality is defined as a function to quantify the degree that user v_i's retweet action at time t' ($t' > t$) is influenced by the active neighbors $N_{i,m}^t$, i.e.,

$$Q(N_{i,m}^t, G_i^\tau), \quad \text{with } \tau \in \mathbb{N}^+ \tag{1}$$

Here we only give a general definition of social influence locality, which can be instantiated in different ways. Finally, given all the above definitions, we can define the problem we are going to address in this work.

Problem 1. Given a network $G = (V, E)$ and users retweet action history $Y = \{v_i, t, m\}_{i,t}$, for user $v_i \in V$ at time t, our goal is to infer 1) whether the active neighbors of v_i have a influence on v_i's retweet action? 2) how to quantify the influence (i.e., how to instantiate the social influence locality $Q(N^t_{i,m}, G^\tau_i)$? 3) how to incorporate the influence into a principled model for predicting user v_i's retweet behavior?

3. DATA DESCRIPTION

The microblogging network we used in this study was crawled from Sina Weibo.com, which, similar to Twitter, allows users to follow each other. Particularly, when user A follows B, B's activities such as (tweet and retweet) will be visible to A. A can then choose to retweet a microblog that was tweeted (or retweeted) by B. User A is also called the follower of B and B is called the followee of A.

As of December 2013, the total number of Sina Weibo users is about 560 million, a similar number to Twitter. This is obviously too large for a few tens of crawlers to collect the entire user space within a short period of time. We had to choose a sampling strategy. To begin with, 100 random users were selected as seed users, and then their followees and followees' followees were collected. In total 1,787,443 users were included in the core network. The crawling of the followees can make the core network more cohesive than crawling the followers, because many users are commonly followed by massive users. Then we monitor the dynamic changes of the "following" relationships for the 1,787,443 users from 8/28/2012 to 9/29/2012. Averagely there are 364,600 new "following" links and 267,515 "unfollow" links created per day. At the end of the crawling, we produced in total 4 billion following relationships among them, with average 200 followees per user. Notice that part of the final crawled users being followed may not be in the core network. Readers can refer to the Sina API of retrieving friend list for details [3].

After crawling the network structure, for each one in the 1,787,443 core users, the crawler collected her 1,000 most recent microblogs. The process resulted in 1 billion microblogs in total. Each miscroblog can be either an original or a retweeted microblog and contains id, original microblog id, user id, content, time and so on. Readers can refer to the Sina API of retrieving user's microblogs for details [4]. We also crawled all the users' profiles, which include name, gender, verification status, #bi-followers, #followers, #followees, creating time and so on. The specific format can refer to the Sina API of retrieving user profile [5].

We conducted some high level analyses on the crawled data set. Figure 2 shows several interesting statistics obtained from the data set. All the distributions are all drawn in log-log scale. Figure 2(a) shows the follower distribution per user. We see that a small portion of users has a huge number of followers. For example, 0.05% users have more than 2 million followers. The phenomenon coincides with the discoveries that 0.05% of the user population attracts almost 50% of all attention within Twitter [Wu et al. 2011] and 1% of the Twitter users control 25% of the information diffusion in Twitter [Lou and Tang 2013]. Figure 2(b) shows the followee distribution per user. It seems that the number of the followees is almost averagely distributed in different ranges except the largest number of the followees (Sina weibo sets up a limitation that each user can only follow at most 3000 followees). Figure 2(c) shows the tweet distribution per user, i.e., the distribution of the number of microblogs posted by each user. We see that the distribution satisfies the power law distribution followed by an exponential tail, which is similar to the discoveries about human communications found by Wu et al. [Wu et al. 2010]. Figure 2(d) shows the retweet distribution per original microblog, which presents an explicit power law trend.

[3] http://open.weibo.com/wiki/2/friendships/friends

[4] http://open.weibo.com/wiki/2/statuses/user_timeline

[5] http://open.weibo.com/wiki/2/users/show

Table I. Data statistics.

Dataset	#Users	#Follow-relationships	#Original-microblogs	#Retweets
Weibo	1,776,950	308,489,739	300,000	23,755,810

In this paper, we focus on studying the influence from the followees on one's retweet behavior in microblogging networks. For each investigated user sample, we need to make the followees in her ego network and the action status of those followees complete enough. Thus, we only consider the core network comprising of 1,787,443 users. We also select 300,000 popular microblog diffusion episodes from the above crawled data set. Each diffusion episode contains the original microblog and all its retweets. On average each microblog has been retweeted 80 times. The total number of retweets associated to the 300,000 diffuion episodes are 23,755,810. The sampled data set ensures that for each diffusion episode, the active (retweet) status of followees in one's τ-ego network is completed. In this way, we can better investigate the influence effect from the active followees of one user. Table I lists statistics of the data set used in this paper. All data set and codes used in this work are publicly available [6].

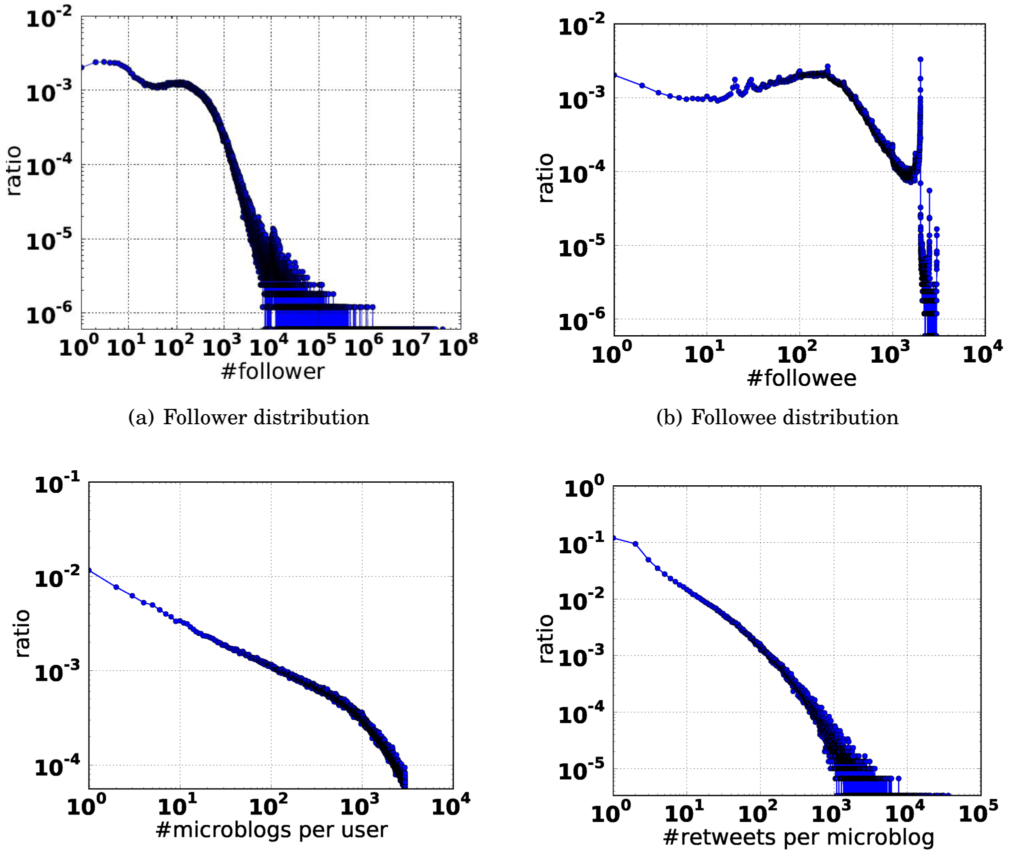
4. TEST FOR EXISTENCE OF INFLUENCE LOCALITY

We first engage into a sampling test to verify the existence of social influence locality for the retweet behaviors. This problem can be connected to the causality inference problem [Pearl 2009]. For this purpose, randomized experiment is the preferred golden method. The basic idea is to partition users into two groups: *treatment group* V_T and *control group* V_C. For users in the treatment group, we assign some treatment of interest, and for users in the control group, we do not assign the treatment. In our test, the treatment of interest is defined as the social influence one would receive in her ego network. We associate a status for each user. If a user retweets a microblog posted by her friend, we say her status becomes active, otherwise inactive. Finally, we compare the activation statuses of all users between the two groups.

One problem in the sampling test is how to *randomly* assign users to the treatment and the control groups. Straightforwardly, given a microblog, we could view users who have followees already retweeted the microblog as users in the treatment group, and assign users who do not have any followees retweeted the microblog to the control group. However, in practice, it is highly infeasible. This is because, in the microblogging network, if a user does not have any followees retweeted the microblog, she will have no chance to see the microblog and thus will not be possible to retweet it. To address this, we assign users who have only one followee retweeted the microblog to the control group and users who have more than one followees retweeted the microblog to the treatment group. In this sense, we try to evaluate the correlation between the probability of a user performing the retweet behavior and her active neighbors. Another trouble we are facing is the selection bias, that is users who were treated would have a higher activation probability than those who were not treated even though the treated users were not treated. This problem was also reported in the study on the influence of product adoption [Arala et al. 2009]. Another bias is the confounding bias, e.g., popular microblogs make users more likely to retweet and be treated, and recently posted microblogs seem to be more likely to be retweeted.

Methodologies To deal with the above problems, we use a matching-based sampling method for testing the influence. The intuition behind this method is to first fix users in the treatment group as those who have more than one followees retweeted a given

[6] http://arnetminer.org/Influencelocality



(a) Follower distribution

(b) Followee distribution

(c) Tweet distribution

(d) Retweet distribution

Fig. 2. Data statistics

microblog, and then for each user in the treatment group, we try to find the most matched user from the original control group, and finally construct a new control group by all the matched users. Specifically, we use a logistic regression model to learn a probabilistic classification model, and then apply the model to estimate the posterior probability of each user belonging to the treatment group. Finally, for a particular user $u \in V_T$ in the treatment group, we select user $v \in V_C$ who results in the minimal difference of the posterior probability with user u as u's matched user, i.e.,

$$v = \arg\min_{v' \in V_C} \|p_u - p_{v'}\| \tag{2}$$

To learn the logistic regression model, we aim to maximize the following likelihood objective function:

$$\begin{aligned} \mathcal{O}(\alpha, \beta) &= \prod_{v \in V_T} P(T = 1|X_v) \prod_{v \in V_C} P(T = 0|X_v), \\ P(T = 1|X_v) &= \frac{1}{1 + e^{-(\alpha X_v + \beta)}} \end{aligned} \tag{3}$$

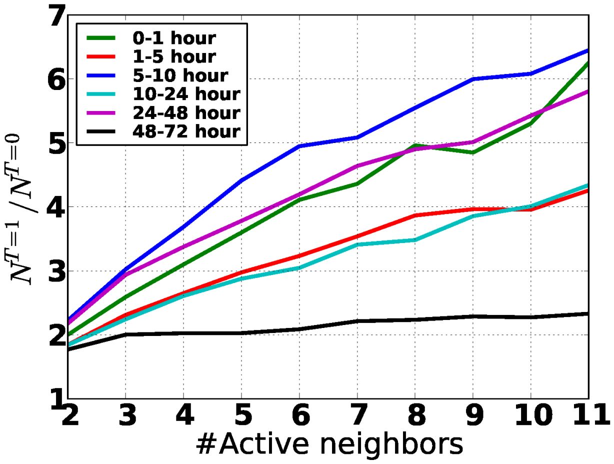


Fig. 3. The result of matching-based sampling test for influence locality. $N^{T=1}$ is the average number of active users in the treatment group, and $N^{T=0}$ is the average number of active users in the control group.

where X_v is the feature vector describing attributes of user v; α are weights of the attributes and β is a bias, both of which are learned by maximizing the objective function $\mathcal{O}$.

In learning the logistic regression model, for each microblog m, we consider various time spans after it has been published, i.e., 0-1, 1-5 , 5-10, 10-24, 24-48, and 48-72 hours. For each user who has retweeted m, we view her as active at a specific time span when she retweeted, and we also treat her as inactive instances at other time spans before she really retweeted. For each follower of an active user, we treat her as an inactive instance at every time span. Then we count the number of previous active neighbors for each active and in-active instance. Finally, we can determine the instances in the original treatment and control groups, and learn the logistic regression model based on them.

Results The test results are shown in Figure 3. From the figure, we can see that for all the time spans, the fraction of active users with 2 active neighbors is about 2 times greater than the fraction of active users with only one active neighbor, i.e. $\frac{N^{T=1}}{N^{T=0}} \approx 2$. Meanwhile, the fraction of active users in the treatment group increases with the number of active neighbors. The test results show strong evidence for the existence of the social influence locality on user's retweet behaviors. However, we also observe that after 48 hours when the original microblog has been published, the increasing rate slows down with the number of active neighbors, which suggests that the influence decays over time.

In the figure, $N^{T=1}$ is the average number of active users in the treatment group, and $N^{T=0}$ is the average number of active users in the control group. We calculate the ratio of the fractions for the two numbers and can conclude that the influence locality exerts positive effect on users' retweet behaviors if $\frac{N^{T=1}}{N^{T=0}} > 1$.

Discussion of τ In the above test and the following experiments in the paper, we set the parameter τ as 1 and hence focus on the 1-ego network. We conduct some analysis to verify the reason why we select τ as 1.

When $\tau = 1$, the influence actually only comes from the direct connected neighbors. When $\tau > 1$, the influence from the indirect connected neighbors is also included.

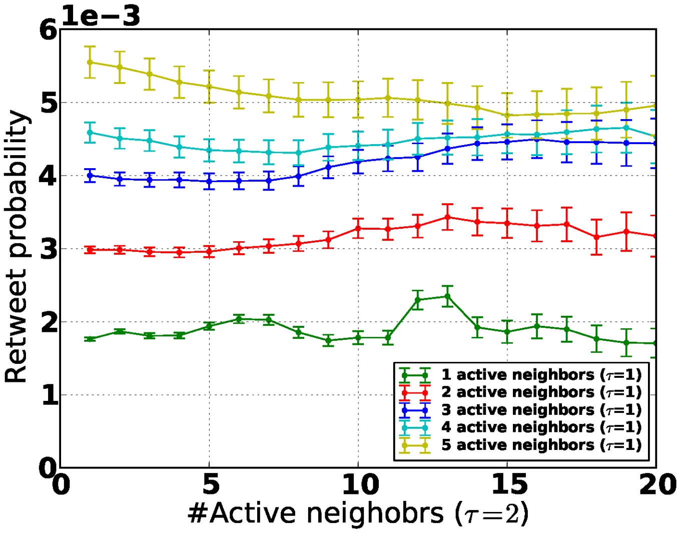


Fig. 4. The effect of influence locality ($\tau = 2$) for users.

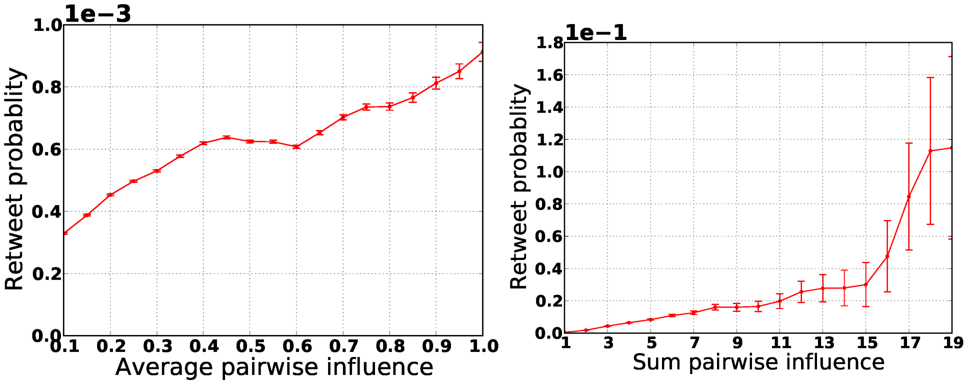
The larger value of τ is, the more indirect influence is considered within the influence locality. The question is, whether the indirect influence is evident or not.

For different kinds of online social networks, the form of the relationships between users is different, which may indicate different effect of influence. For example, the relationships in Facebook are undirected friend relationships and the indirect friends sometimes present evident influence, while the relationships in microblogging networks are directed “following” relationships and the influence from those indirect “following” relationships may be relatively weak. In this paper, we investigate whether the indirect influence from 2-ego network in Weibo network is evident. Specifically, we fix the number of active neighbors in one’s 1-ego network, and then analyze the effect from active neighbors in 2-ego network. Here the neighbors are the users that a user directly ($\tau = 1$) or indirectly ($\tau > 1$) follows. The results are shown in Figure 4. We can see from the figure that the effect from the active neighbors in 2-ego network almost remains unchanged when fixing the number of active neighbors in 1-ego network (i.e., from 1 to 5). We also find that the retweet probability increases with the number of active neighbors in 1-ego network. The results indicate that the retweet behaviors from the indirect followees in Weibo network exert little influence on users’ retweet behaviors. The phenomenon can be explained as that users can not be easily exposed by the messages from their indirect followees, and thus the influence from those indirect followees is trivial. According to the analyses , we set the parameter τ as 1 and hence focus on the 1-ego network.

5. INSTANTIATION FOR INFLUENCE LOCALITY

We present the instantiation functions of influence locality for modeling retweet behaviors. In particular, we focus on studying the effects of pairwise influence and structure influence.

Pairwise influence Most existing literatures on social influence focus on analyzing influence between users, i.e., pairwise influence. The pairwise influence can be defined based on social ties and interactions between users. To quantify this, we cast the prob-



(a) Average pairwise influence

(b) Sum pairwise influence

Fig. 5. The effect of random walk based pairwise influence (a) calculated by averaging the random walk probabilities of active neighbors. (b) calculated by adding up the random walk probabilities of active neighbors.

lem as measuring the relatedness between nodes in a graph and use the theory of random walk with restart (RWR) [Lovasz 1993; Sun et al. 2005] to achieve it.

Specifically, we conduct RWR in user v_i's τ-ego network G_i^τ and calculate the random walk probability p_j for each active neighbor v_j using Eq. 4.

$$\vec{P_i} = (1-c)\mathbf{A}\vec{P_i} + c\vec{I_i} \tag{4}$$

where $\vec{P_i} = (P_i(1), P_i(2), ..., P_i(|G_i^\tau|))$ is the steady state probability vector, with each dimension $P_i(j)$ denotes the steady state probability that our random walker will find herself at node v_j. $|G_i^\tau|$ is the size of the ego network of v_i. $\vec{I_i}$ is a column vector with all its elements zero, except for the entry that corresponds to node v_i; set this entry to 1. We call $\vec{I_i}$ the "restart vector". $\mathbf{A}$ is the adjacency matrix of the ego network. c is the probability of returning to the node v_i.

The random walk probability can be explained as how the influence of an active neighbor can finally reach the given user v_i via the network connection between them. For an example, as shown in Figure 1(b), user B only has one path to reach v, while F has a number of different paths to connect v through E and another two users. Figure 5 shows the probability that a user retweets a microblog conditioned on (a) the average random walk probability and (b) the sum of the random walk probability of all active neighbors in her ego network. From both figures, we can observe that the random walk based pairwise influence score can be used as a good indicator of the retweet behavior.

Structure influence As observed in Figure 1(b), user v has six active neighbors, A, B, C, D, E, and F, who form four connected circles. How is the influence locality correlated with the inner structure of active neighbors? A more specific question is: comparing with A and B who distribute into different circles, will the pair of users C and D who reside in the same circle have the same influence effect on v's retweet behavior? Literature [Ugander et al. 2012] reports that *structural diversity* (captured by the presence of multiple compnents of the local structure) can be used as a positive predictor of user engagement. They simply consider the number of connected components (circles) as the indicator to analyze its correlation with the probability of user engagement to some activity, and find significantly positive correlation there. Will the structural diversity has the same effect on the retweet behavior? How to define an utility function to capture this effect?

Figure 1(e) plots the curves of retweet probability versus the number of connected circles formed by the active neighbors. Specifically, we analyze the results by varying the number of active neighbors by 2,3,4,5,6-10, 11-20, and 21-30 respectively. We see that, surprisingly, the retweet probability is negatively correlated with the number of circles, which is opposite to the discovery in [Ugander et al. 2012]. This phenomenon might be explained from the purpose of retweet. Boyd et al. [Boyd et al. 2010] found that one important purpose for people to retweet is to influence others. According to this, people may quickly lose interests to retweet when they find that many of their social circles are already aware of the message. We may also explain it by using the theory of peer pressure [Durkin 1996]. Peer pressure is the influence that a peer group, observers or individual exerts that encourages others to change their attitudes, values, or behaviors to conform the group norms. If one user observes a microblog having been retweeted by many friends from one same group, she will be very likely to also retweet it under the peer pressure.

Note that when calculating the number of circles, we only consider reciprocal (bi-directional) “following” relationships between users in one’s ego network. This is because, we find that directional relationships are meaningless from an interaction point of view. Huberman et al. also empirically prove that a sparser and simpler network of actual friends is a more influential network in driving the microblogging usage [Huberman et al. 2009]. Furthermore, we also limit circles in k-brace (k is set as 2) as proposed by Ugander et al. [Ugander et al. 2012], since two actual circles may be easily connected by an arbitrary local bridge, which however, should not be treated as in the same circle. k-brace circle is defined as the circle with all the edges of embeddedness less than k being removed, where embeddedness of an edge is the number of common neighbors shared by the two endpoints.

We can use the same matched sampling method introduced in section 4 to test the structure influence from ego network. Specifically, we assign the users with only one connected component formed by the active neighbors to the control group, and the users with more than one connected component formed by the active neighbors to the treatment group. To reduce the selection bias caused by the assigning process, we find the most matched users from the original control group to each user in the treatment group to construct a new control group. We train a logistic regression model $P(Y|X)$ to match users, where Y is the posterior probability of a user belonging to the treatment group (i.e., having more than one, e.g., five, active connected component), and X are all observed features except the number of the connected components. However, the matched sampling method is time consuming. It takes about more than seven days to get all the test results shown in Figure 3 with the scale of 1 million users. Actually, we tried the matched sampling method to test the structure influence on one setting, where users with 1 connected component formed by 5 active neighbors were assigned to the control group and users with 5 connected components formed by 5 active neighbors were assigned to the treatment group. We found from the test result that the ratio of the active users between the two groups was less than 1, which indicated that the users with less connected components formed by active neighbors are more likely to retweet. The result is actually consistent with the empirical analysis shown in Figure 1(e). Considering the time efficiency, we empirically analyze the retweet probability under different number of circles formed by active neighbors.

Instantiation functions Based on the above observations, we give a definition of the influence locality function. More precisely, we define it as,

$$Q(N_{i,m}^t, G_i^\tau) = w \times g(N_{i,m}^t, G_i^\tau) + (1 - w) \times f(N_{i,m}^t, G_i^\tau) \tag{5}$$

where $g(N_{i,m}^t, G_i^\tau)$ denotes the pairwise influence and $f(N_{i,m}^t, G_i^\tau)$ denotes the structure influence. Briefly, we abbreviate them as Q, g, and f, respectively. Notation w denotes a tunable parameter to balance the two terms.

For the pairwise influence, we have tried different definitions, for example, the sum of the random walk probabilities of all active neighbors, i.e.,

$$g(N_{i,m}^t, G_i^\tau) = \sum_{v_j \in N_{i,m}^t} p_j \tag{6}$$

where p_j is the random walk probability from the active user v_j to the given user v_i. We also tried other definitions by replacing the sum with the average functions (arithmetic mean and geometric mean).

In addition, in the definition, we should consider the temporal information (the time that a user retweets a microblog). By adding the time into the above equation, we obtain,

$$g(N_{i,m}^t, G_i^\tau) = \sum_{v_j \in N_{i,m}^t} h_j p_j \tag{7}$$

where h_j is the difference between the time when v_j retweeted the microblog and the time when we try to predict v_i's retweet behavior. The function sum can be also replaced by other functions such as arithmetic mean, geometric mean, and max.

For the structure influence, we can simply use a linear combination of the number of connected circles to quantify the influence function. However, as we see from Figure 1(e), the influence does not linearly decrease. Thus we uses the exponential function to describe the effect of the structure influence:

$$f(N_{i,m}^t, G_i^\tau) = e^{-\mu |C(N_{i,m}^t)|} \tag{8}$$

where $C(N_{i,m}^t)$ is the collection of circles formed by the active neighbors and μ is a decay factor.

6. RETWEET BEHAVIOR PREDICTION

The defined influence locality function has a strong predictive power and can be used for many applications such as retweet behavior prediction and social recommendation. In this section, we first introduce the defined features, and then describe two methods that will be used to predict the retweet behaviors.

6.1. Feature definition

To predict retweet behaviors, in addition to the influence locality based features, we also investigate several other basic features that may affect the retweet probability. We define three kinds of basic features, including personal attributes, topic propensity and instantaneity. Specifically, we try six personal attributes, including the number of followees, the number of followers, the number of bi-followers (i.e., the reciprocal “following” relationships), the longevity (age of the account), gender (0 indicates male and 1 indicates female) and verification status (0 indicates being verified as a celebrity and 1 indicates not being verified).

Table II. Personal attributes for retweet behavior prediction.

Basic Feature	Value Range	Ave	Median
#Followers	0-54,164,442	9,849	493
#Followees	0-3,000	467	280
#Bi-followers	0-2,983	210	114
Longevity	49-1,234	472	760
Gender	0,1	-	-
Verification status	0,1	-	-

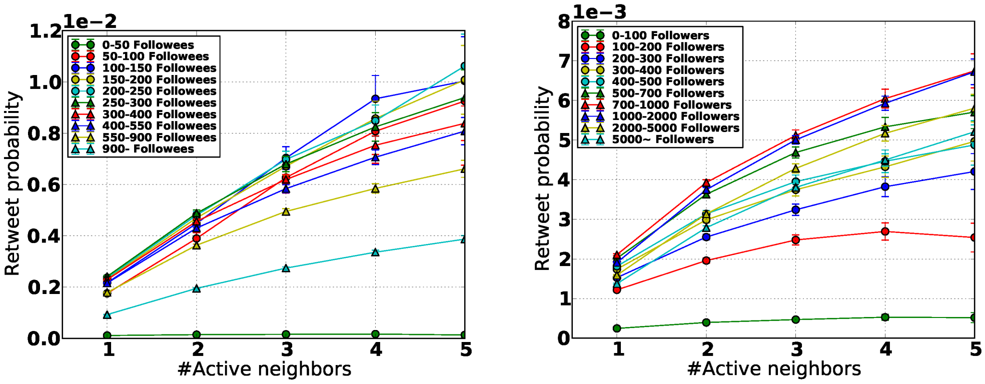
We summarize the statistics for the six personal attributes in Table II. In order to investigate the correlation between the retweet probability and the influence from the active neighbors for users with different values of personal attributes, we divide each personal attribute into different ranges. For example, for the number of followees, we count the number of users with different number of followees and find the ten fractile respectively, and then for each ten fractile point, we investigate the retweet probabilities under the influence from 1 to 5 active neighbors. We analyze the retweet probabilities for users with different values of followers, bi-followers and longevity in the same way as followees. For gender and verification status, the values are naturally classified into 0 and 1.

Personal attributes In Figure 6(a), we find that when the number of followees is too few (0-50), the whole retweet probability (the retweet probabilities under different number of active neighbors) is very low and the influence from active neighbors (the increase rate of the retweet probabilities from 1 to 5 active neighbors) almost exerts no effect. When the number of followees increases to 100-150, the whole retweet probability increases and the influence from active neighbors becomes evident. However, when the number of followees continues to increase, the whole retweet probability and the influence begin to decrease. The phenomenon may be explained as that when people follow a lot of users, the information source becomes overloaded, that makes people difficult to find valuable information to retweet. For the follower attribute in Figure 6(b) and bi-follower attribute in Figure 7(a), the patterns are similar to the followee attribute that the whole retweet probability and influence first rise then fall. The peak point for follower attribute is 700-1000 followers and for bi-follower, the peak point is 100-160. The phenomenon may be explained that when the social relationships grow up to a moderate status, the users are most active in retweeting and are also most easily influenced by the active neighbors in order to develop their social circle.

We also investigate the correlation between the retweet probability and longevity (age of an account). The results presented in figure 7(b) show that users with longer longevity are more inclined to retweet and easily influenced by their active neighbors. In Figure 8(a) and Figure 8(b), the results show that female and verified users are more likely to retweet and be influenced by their neighbors than the male and unverified users.

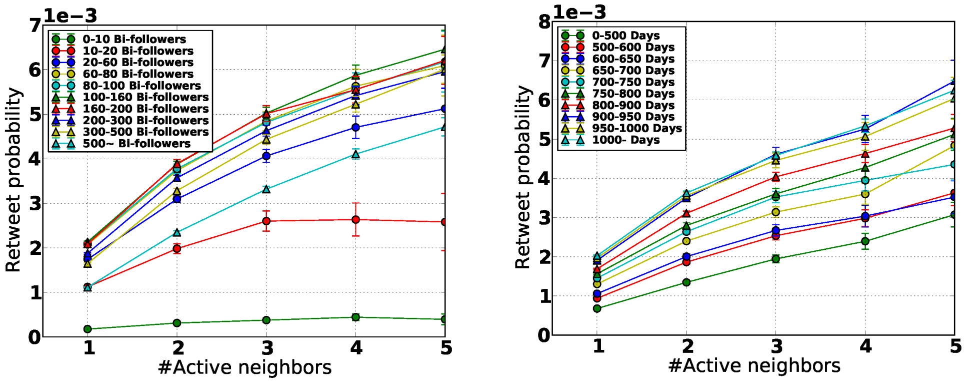
Instantaneity Instantaneity is defined as the elapsed time from when the original microblog m was published. We find from Figure 3 that users are mostly influenced by their active neighbors 5 to 10 hours after the original microblog being published. After that, the retweet propensity presents decreasing trend over time. Especially after 48 hours, the influence increases very slowly.

Topic propensity Topic propensity is defined as the Jensen-Shannon divergence [Heinrich 2004] between the topic distribution of the user v and the topic distribution of the microblog m.



(a) Followees (b) Followers

Fig. 6. The effect of influence locality for users with different number of (a) followees. (b) followers.



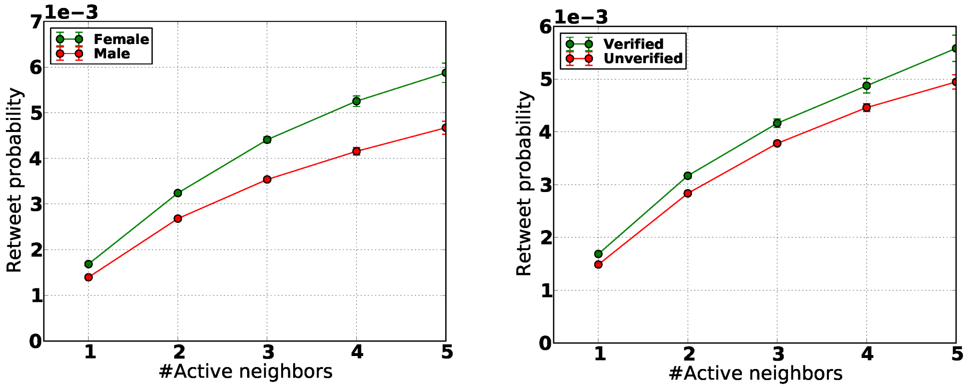
(a) Bi-followers (b) Longevity

Fig. 7. The effect of influence locality for users (a) with different number of bi-followers and (b) with different longevity in weibo.

$$JSD(P_v \parallel P_m) = \frac{1}{2}D_{KL}(P_v \parallel P_a) + \frac{1}{2}D_{KL}(P_m \parallel P_a) \tag{9}$$

where P_v is the topic distribution of user v, P_m is the topic distribution of microblog m, and P_a is the average result of P_v and P_m. D_{KL} is the KL divergence and is calculated by $D_{KL}(P_v \parallel P_a) = \sum_{k=1}^{K} \ln \frac{P_v(k)}{P_a(k)} P_v(k)$. To obtain the topic distributions for all the microblogs and users, we first treat each historical microblog as a document and utilize Latent Dirichlet Allocation [Heinrich 2004] to estimate the probability of generating a microblog m from each topic k, which is denoted as $P(m|k)$. Then we estimate the probability of generating a user v from each topic k by averaging the probabilities of all her historical microblogs associated to topic k.

$$P(v|k) = \frac{\sum_{m \in M_v} P(m|k)}{|M_v|} \tag{10}$$



(a) Gender

(b) Verification Status

Fig. 8. The effect of influence locality for (a) male and female users and (b) verified or unverified users.

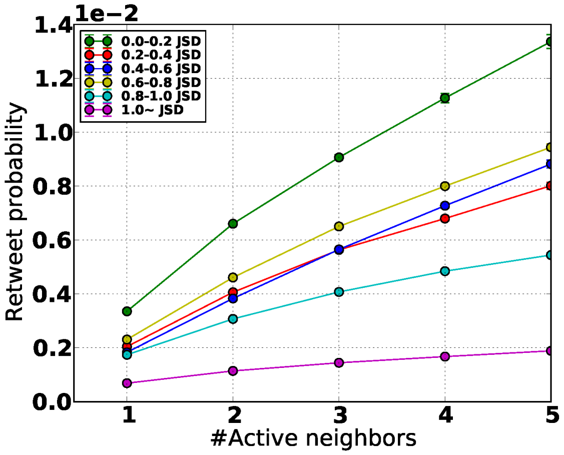


Fig. 9. The effect of influence locality for users with different topic propensity.

where M_v is the historical microblog collection of user v. Finally the topic distribution for one user is denoted as $P_v = \{P(v|k)\}_{k=1}^{K}$.

We show the correlation between retweet probability and topic propensity in Figure 9. From the result we can see that basically, when the calculated JSD is smaller, i.e., the topic propensity of a user to a microblog is greater, the user will be more likely to retweet the microblog, and will be more likely to be influenced by the active neighbors to retweet the microblog.

6.2. Logistic Regression Model

The retweet behavior prediction can be considered as a classification problem: given one microblog m, a user v_i and a timestamp t, the goal is to categorize user v_i's status at t. We denote the classification outcome as $y_{i,m}^t$. $y_{i,m}^t = 1$ indicates that v_i will retweet m before t, and $y_{i,m}^t = 0$ otherwise. We use the influence locality function $Q(N_{i,m}^t, G_i^\tau)$

and the above defined features as evidence to predict $y^t_{i,m}$. The advantage of the classification model is that we can integrate different combinations of the functions into the model conveniently.

To solve the classification problem, many machine learning models can be used, such as SVM and logistic regression classifier. In this paper, we use a logistic regression classifier to predict the value of $y^t_{i,m}$ for each given (v_i, t, m):

$$P(y^t_{i,m} = 1|x^t_{i,m}) = \frac{1}{1 + e^{-(\alpha x^t_{i,m} + \beta)}} \tag{11}$$

where $x^t_{i,m}$ is the feature vector of user v_i associated with microblog m at timestamp t, and α are weights of the features and β is a bias, both of which are learned by maximizing an objective function similar as Eq. 3.

6.3. Factor Graphic Model

In the above setting of the logistic regression model, each user to be predicted are independent with each other. For predicting the behavior of user v_i, the behaviors of v_i's neighbors before time t are viewed as the observable information and can be leveraged to predict the behavior of v_i. However, when we change the setting as after δ time interval when a microblog m was published, predicting the retweet behaviors of all the users who have seen the microblog m, the behaviors of all the users to be predicted are actually dependent. For example, when two friends v_i and v_j are both the users to be predicted, one's behavior can not be viewed as the observable information and used as features to predict the behavior of another user. However, v_i's action status will be very likely to influence the action status of v_j. In summary, the action status of one user not only depends on the predefined attributes associated to the user, but also may be influenced by the action statuses of the neighbors to be predicted. The logistic regression model views each instance as independent and cannot leverage the correlation between instances. Thus, we propose to use a factor graphic model to model the correlation between instances.

We construct a factor graph for each microblog m. We map each instance v_i related to m as a node in the factor graph and assign a label y_i for each node. $y_i = 1$ means v_i retweeted m and $y_i = 0$ means v_i did not retweet m. Each node is associated with a vector of attributes $\vec{x}_i$, of which each dimension comes from the influence locality features defined in Section 5 and the basic features defined in Section 6. For the influence locality features, we only consider the active neighbors in the training data. We define two kinds of factors. The first kind is the attribute factor which represents the posterior probability of the label y_i given the attribute vector $\vec{x}_i$. The second factor is the correlation factor which denotes the correlation between the labels of neighboring nodes. We only construct the correlation factors between two nodes with the reciprocal "following" relationships because the factor graphic model is an undirected probabilistic model.

To instantiate the factor graphic model, we still need to give the formal definition of the objective function and instantiate the feature definitions. Given a network $G = (V, E)$, the action history $Y = \{y^t_{i,m}\}$ and the corresponding feature vector $X = \{x^t_{i,m}\}$, with some known variable $y_i = 1$ or 0 and some unknown variables $y_i =?$, our goal is to infer values of those unknown variables. For simplicity, we remove the superscript t and subscript m for all variables if there is no ambiguity. We begin with the posterior probability of $P(Y|X, G)$. Directly solving the posterior probability is obviously intractable. Here, we instantiate the probabilities $P(Y|G)$ and $P(x_i|y_i)$ within

Markov random field and Hammersley-Clifford theorem [Hammersley and Clifford 1971]:

$$P(Y|X,G) = \frac{1}{Z}\exp\{\sum_{i=1}^{V}\sum_{d=1}^{D}\alpha_d \times h_d(x_{id}, y_i) + \sum_{e_{ij}\in E}\beta \times r(y_i, y_j)\} \quad (12)$$

where D is the number of attribute features, x_{id} is the d^{th} feature value of the i^{th} node; e_{ij} is a reciprocal relationship in the network G. $h(\cdot)$ represents the correlation between user v_i's action status and her own features. $r(\cdot)$ corresponds to the correlation between user v_i's action and her friend v_j's action. Finally Z is a normalization factor to guarantee that the resultant is a valid probability.

The objective function is defined as $\mathcal{O}_{\alpha,\beta} = \log P_{\alpha,\beta}(Y|X,G)$. Learning the factor graphic model is to estimate a parameter configuration $\theta = (\{\alpha_d\}, \beta)$ from a given historical data, which is to maximize the log-likelihood objective function, i.e., $\theta = \arg\max \mathcal{O}(\theta)$.

Model learning We employ a gradient descent method (or a Newton-Raphson method) for model learning. Here we use α_d as the example to explain how we learn the parameters. Specifically, we first write the gradient of each α_d with regard to the objective function:

$$\frac{\mathcal{O}(\theta)}{\alpha_d} = \mathbb{E}[f(y_i, x_{id})] - \mathbb{E}_{P(y_i|X,G)}[f(y_i, x_{id})] \quad (13)$$

where $\mathbb{E}[f(y_i, x_{id})]$ is the expectation of the local factor function $f(y_i, x_{id})$ given the data distribution in the input network and $\mathbb{E}_{P(y_i|X,G)}[f(y_i, x_{id})]$ represents the expectation of $f(y_i, x_{id})$ under the distribution $P(y_i|X,G)$ learned by the model. Similar gradients can be derived for parameter β.

The graphical structure in the factor graphic model can be arbitrary and may contain cycles, which makes it intractable to directly calculate the marginal distribution $P(y_i|X,G)$. We choose Loopy Belief Propagation due to its ease of implementation and effectiveness. Specifically, we approximate the marginal distribution $P(y_i|X,G)$ using LBP. With the marginal probabilities, the gradient can be obtained by summing over all instances. It is worth noting that we need to perform the LBP process twice in each iteration, one time for estimating the marginal distribution of unknown variables, i.e., $y_i =?$, and another time for estimating the marginal distribution over all instances. In this way, the algorithm essentially performs a semi-supervised learning over the complete network. This idea was first proposed in [Tang et al. 2011] for learning to categorize social relationships. Finally with the obtained gradient, we update each parameter with a learning rate η.

Prediction With the estimated parameters θ, we can predict the label of unknown variables $y_i =?$ by finding a label configuration which maximizes the objective function, that is, $Y^\star = \arg\max \mathcal{O}(Y|X,G,\theta)$. To do this, again, we utilize the loopy belief propagation to approximate the solution, that is, to calculate the marginal distribution of each node with unknown variable, i.e.,$P(y_i|x_i,G)$, and assign each node the label with the maximal marginal probability.

Note that Logistic regression and factor graphic model are widely used in social prediction. For example, Yang et al. [Yang et al. 2011] tried different loss functions such as huber loss, lazy loss and logistic regression, and showed that the three loss functions perform the same on the task of social prediction. Leskovec et al. [Leskovec et al. 2010] leveraged logistic regression to verify the effects of the discovered positive and negative patterns on the task of link prediction. Tang et al. [Wu et al. 2013] utilized factor graphic model to recommend partners in patent collaboration. Although we can use other prediction algorithms such as matrix factorization to verify the effects of the discovered influence patterns, logistic regression provides a coefficient for each feature, which suggests how the feature is used by the model to provide weight for or against a retweet behavior and provides proposals for how subset of these features offers evidence for retweet behaviors. We also consider factor graphic model, because it can explicitly model the correlation between the retweet behaviors of two users to be predicted.

7. EXPERIMENTAL RESULTS

In this section, we validate the effectiveness of using influence locality functions for predicting retweet behaviors.

7.1. Experimental Setup

Data preparation We use the data set described in Section 3 for retweet prediction. Basically, for each user who retweeted a microblog in the collected data set, we treat her as a positive instance, the goal is to predict whether she will retweet before her real retweet time. For each follower of a positive instance, if the follower is never observed to retweet the microblog exposed by her followee, we treat her as a negative instance. The goal for each negative instance is to predict whether she will retweet before a randomly selected timestamp. We select from 6 timestamps including 0-1, 1-5, 5-10, 10-24, 24-48, and 48-72 hours after the original microblog being published.

We observe that the positive and negative instances are much unbalanced (about 1:300) in the constructed dataset. Thus we sample a balanced data set with equal number of positive and negative instances. Specifically, we sample a random negative instance for each positive instance to ensure the equal number in the dataset. Learning effective models from unbalanced data is an open problem in machine learning field. The balanced method is aimed to alleviate this problem. We use a similar balanced method as that in [Guha et al. 2004] in this work. This balance sampling method will not influence the comparison between different approaches.

Comparison methods We compare our model with several methods.

LRC-Q: it uses the logistic regression model and only uses the influence locality function $Q(N_{i,m}^t, G_i^\tau)$ defined in Section 5 as a feature to train the logistic regression classifier.

LRC-B: it uses the logistic regression model and only uses the basic features defined in Section 6 to train the logistic regression classifier.

LRC-BQ: it uses the logistic regression model and uses both the influence locality function $Q(N_{i,m}^t, G_i^\tau)$ and the basic features to train the logistic regression classifier.

FGM-BQ: it uses the factor graphic model and uses both the influence locality function $Q(N_{i,m}^t, G_i^\tau)$ and the basic features as the attribute features, and also considers the correlation feature between the labels of neighboring nodes to train the classifier.

For LRC-Q, LRC-BQ and FGM-BQ, we empirically set $w = 0.5$ in Q function and $\mu = 1$ in the structure influence function f.

Evaluation metrics We divide the constructed data set into training and test data, and perform 5-fold cross validation. We evaluate the performance of retweet behavior prediction in terms of Precision, Recall, F1-measure, and Accuracy.

Table III. Performance of retweet behavior prediction. (%)

Model	Prec.	Rec.	F1	Acc.
LRC-B	68.11	74.26	71.05	69.74
LRC-Q	66.82	77.22	71.65	69.44
LRC-BQ	69.89	77.06	73.30	71.93

Table IV. Performance of LRC-Q ($w = 1$) by using different g functions. (%)

Model	Prec.	Rec.	F1	Acc.
$g_1 = \sum p_{v_i}$	57.42	77.13	65.83	59.96
$g_2 = \frac{\sum p_{v_i}}{\lvert S_v \rvert}$	60.21	75.03	66.81	62.72
$g_3 = \sqrt[\lvert S_v \rvert]{\prod p_{v_i}}$	60.28	75.31	66.96	62.84
$g_4 = \sum h_{v_i} p_{v_i}$	58.85	92.68	71.99	63.94
$g_5 = \frac{\sum h_{v_i} p_{v_i}}{\lvert S_v \rvert}$	61.57	91.72	73.68	67.24
$g_6 = \sqrt[\lvert S_v \rvert]{\prod h_{v_i} p_{v_i}}$	**61.85**	**92.67**	**74.19**	**67.76**
$g_7 = \max h_{v_i} p_{v_i}$	61.15	91.13	73.19	66.61

7.2. Performance

Table III shows the performance of the comparison methods. The results show that using only the influence locality function to predict retweet behaviors (LRC-Q) can obtain a comparable performance with (even better than) the method using all the additional features (LRC-B) (+0.6% in terms of F1-measure, -0.3% in terms of accuracy). By combining the influence locality function and the additional features together, we can obtain a bit improvement on performance (+1.65% in terms of F1-measure, +2.49% in terms of accuracy).

Pairwise influence functions According to the various definitions of the pairwise influence functions in Section 5, we further try different g functions for predicting retweet behaviors. Specifically, we set $w = 1$ and try seven g functions for pairwise influence defined in Section 5. The evaluation results are shown in Table IV. We can see that, g_6, which averages the time weighted pairwise influence by using geometric mean, performs the best. The result suggests that the neighbors with different retweet time exert different influence. Besides, we also find that arithmetic mean performs poorer than geometric mean for both the time weighted pairwise influence (g_5 under-performs g_6) and the pairwise influence without time weighting (g_2 under-performs g_3). This is due to the reason that the pairwise influences from the active neighbors are not normally distributed but right-skewed, i.e., the majority of pairwise influences are low and a minority of pairwise influences are scattered in a fat right tail.

Parameter w There is one parameter w used in the $Q = w \times g + (1 - w) \times f$ function. We study how the parameter w affects the performance of retweet prediction. Figure 10 plots the accuracy of LRC-Q with various values of w, where g is set as g_6 according to the best performance presented in Table IV. We see that the highest accuracy is obtained when w is 0.5.

7.3. Analysis and Discussions

Feature contribution analysis We analyze the contribution of different features on retweet behavior prediction. Specifically, we respectively add the basic features and influence locality features into LRC-BQ one by one and evaluate the increase of the

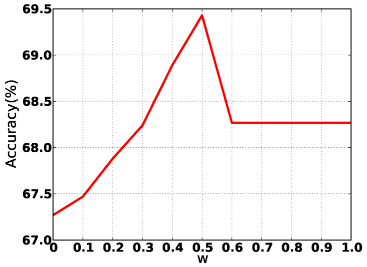


Fig. 10. Performance of LRC-Q under different values of w.

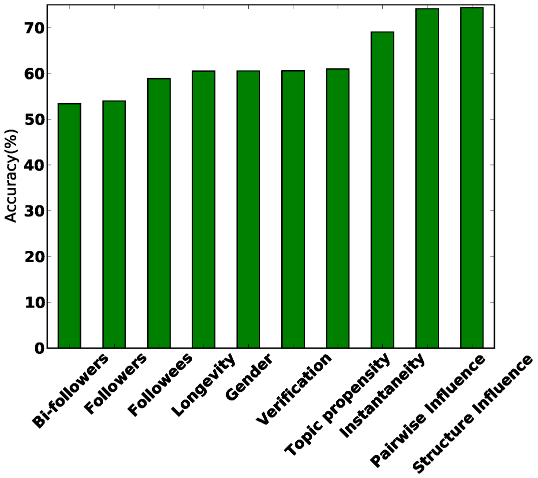


Fig. 11. Performance of different features.

predictive performance. A larger increase means a higher predictive power. We train and evaluate the predictive performance of the logistic regression model with different set of features. Figure 11 shows the accuracy of different versions of logistic regression models. We can observe clear increase on the performance when adding the basic features of bi-followers, followees and instantaneity, which indicates that the three kinds of basic features can contribute a lot on predicting retweet behavior. Other basic features do not present evident contributions, which indicates that their effects are counteracted by other features. We also observe a clear increase on the performance when adding the pairwise influence feature, which indicates that the aggregated pairwise influence locality from the 1-ego network indeed exerts significant effect on retweet behaviors. However, the structure influence presents insignificant effect.

How structure influence help? Through the analyses in Section 5, we find that the probability of a user retweeting a microblog is negatively correlated with the number of connected circles that are formed by the active neighbors. However, the negative effect is evident only when the number of active neighbors is relatively large, i.e., larger than 5 active neighbors as shown in Figure 1(e). The negative influence is insignificant when there are only a few active neighbors, i.e., less than 5 as shown in Figure 1(e). However,

Table V. The proportion of instances with different number of active neighbors. (%)

#Active neighbors	**1**	**2**	**3**	**4**	**5**	**>5**
Proportion of instances	79.47	11.96	3.82	1.69	0.89	2.17

Table VI. Performance of retweet behavior prediction with structure influence and without structure influence. (%)

Model	**Prec.**	**Rec.**	**F1**	**Acc.**
LRC-Q(w=1)	49.51	51.53	50.50	49.49
LRC-Q(w=0.5)	51.86	67.70	58.73	52.43

through analyzing the instances in our dataset for prediction, we find that almost 80% instances only have one active neighbors, and the instances with the number of active neighbors larger than 5 only occupies 2.17% (shown in Table V). Therefore, when we use those instances for training and test, the predictive performance is dominated by the instances with a few number of active neighbors and the effect of the number of circles can not be presented. To prove the effect of the structure influence, we sample the instances with the number of active neighbors larger than 5, and then use the sampled instances to conduct the training and test. We compare the performance of the logistic regression model with only the pairwise influence function g (LRC-Q($w = 1$)) and the logistic regression model with both the pairwise influence function g and the structure influence function f (LRC-Q($w = 0.5$)). The results presented in Table VI show that when adding the feature of structure influence, F1-measure is improved by about 8% and the accuracy is improved by about 3%. This indicates that structure influence can indeed improve the performance of retweet prediction when the number of active neighbors is large enough. We also notice that the whole performance in Table VI is lower than that in Table III. It may due to the reason that users with large number of active neighbors are very likely to have similar personal attributes and propensity to retweet a microblog, and thus the behaviors of those users are more difficult to predict.

In fact, in the current Sina Weibo system, when a user reads a microblog, the other neighbors who have already retweeted the microblog are also exposed to the user. This information is very useful to help the user determine whether the microblog is valuable or not. According to our experiments, if the system can cluster the active neighbors in different ego circles and tell the user in which circles the microblog has already been diffused, it will further benefit the user to identify valuable information, and meanwhile, the system can recommend the microblogs to the users according to the features more accurately.

How correlation affects the predcitive performance? For evaluating the performance of factor graphic model, we construct data set in different ways. Specifically, for each microblog m, the positive instances, i.e., the users who retweeted a microblog m, are divided into training data and test set. The training data are those who retweeted m within 10 hours after m being published. Test set are those who retweeted m 10 hours after m being published. The negative instances, i.e., the followers of positive instances that were never observed to retweet m, are also divided into training data and test set in the same way, where the timestamps of the negative instances are randomly generated.

Table VII shows the performance of LRC-BQ and FGM-BQ in the above experimental setting. The two methods both consider all the basic and influence locality features, while FGM-BQ also considers the correlations between instances additionally. We can see from the results that FGM-BQ outperforms LRC-BQ in terms of Precision but under-performs it in terms of Recall. This is because FGM-BQ uses additional features

Table VII. Performance of retweet behavior prediction with correlation feature and without correlation feature. (%)

Model	Prec.	Rec.	F1	Acc.
LRC-BQ	66.78	52.28	58.65	70.76
FGM-BQ	70.75	48.02	57.20	71.25

and the prediction conditions are more strict. In order to distinguish their difference, we perform sign test on their prediction results. The p-value of sign test is less than 0.001, which confirms that the factor graphic model considering the correlation feature significantly performs better than the logistic regression model without the correlation feature. We also find that the whole performance of this setting is lower than that in Table III. This is because we only use the active users within 10 hours after a microblog m being published to construct influence locality features. Thus the information can be leveraged is less, which makes the prediction under this setting more difficult.

8. RELATED WORK

Social influence analysis We investigate the related work from two aspects, social influence analysis and retweet behavior prediction.

Considerable work has been conduced for studying the effects of social influence. These studies can be roughly classified into three categories: influence testing [Anagnostopoulos et al. 2008; Arala et al. 2009; Bakshy et al. 2012; Bond et al. 2012; La Fond and Neville 2010], influence quantification [Barbieri et al. 2012; Belak et al. 2012; Dietz et al. 2007; Goyal et al. 2010; Gruhl et al. 2004; Kempe et al. 2003; Kimura et al. 2011; Saito et al. 2008; Tang et al. 2009; Tang et al. 2013; Tan et al. 2011; Liu et al. 2012; Myers et al. 2012; Weng et al. 2010; Shuai et al. 2012], and influence models and maximization methods [Barbieri et al. 2012; Chen et al. 2010; Chen et al. 2011; Chen et al. 2012c; Goyal et al. 2011; Kimura et al. 2011; Leskovec et al. 2007; Li et al. 2013].

Influence testing is to verify whether the influence indeed exists or not on certain behaviors. One kind of work focuses on statistical causal inference based on the observed data. For example, Arala et al. [Arala et al. 2009] distinguished the effect of influence from homophily based on a statistical propensity score matching method. Fond et al. [La Fond and Neville 2010] measured the gain in correlation and used a randomization technique to assess whether a significant portion of this gain is due to influence and/or homophily. Anagnostopoulos et al. [Anagnostopoulos et al. 2008] proposed a shuffle test to examine the existence of social influence. Another kind of work leverages the online social networks to conduct real controlled trials. For examples, Bakshy et al. [Bakshy et al. 2012] conducted randomized controlled trials to identify the effect of social influence on consumer responses to advertising, and Bond et al. [Bond et al. 2012] used a randomized controlled trial to verify the social influence on political voting behaviors.

Influence learning is to quantify influence. From the perspective of measured objects, we classify the studies into: quantifying influence from topic level [Barbieri et al. 2012; Tang et al. 2009; Liu et al. 2012; Weng et al. 2010], sentiment level [Tan et al. 2011], and so on. For example, Tang et al. [Tang et al. 2009] proposed a Topical Affinity Propagation (TAP) approach to model the topic-level social influence in large networks. Liu et al. [Liu et al. 2012] proposed a generative topic model to mine topic influence between users in heterogeneous networks. Goyal et al. [Goyal et al. 2010] and Saito et al. [Saito et al. 2008] measured the pairwise influence between two individuals based on the idea of independent cascade model [Kempe et al. 2003]. Xin et al. [Shuai et al. 2012] studied the indirect influence using the theory of quantum cognition. Myers et al. [Myers et al. 2012] and Lin et al. [Lin et al. 2013] proposed probabilistic models to quantify the external influence out-of-network sources. Tang et al. [Tang et al. 2013] proposed a probabilistic factor graphic model to quantify the individual, peer and group influence in a social network. Zhang et al. [Zhang et al. 2014] formalized conformity influence and measured conformity of different roles. Belak et al. [Belak et al. 2012] investigated and measured the influence between two communities. From the perspective of the measuring methods, Dietz et al. [Dietz et al. 2007], Liu et al. [Liu et al. 2012] and Zhang et al. [Zhang et al. 2014] used probabilistic topic models to learn the influential strength between papers or users. Tang et al. [Tang et al. 2009], Tan et al. [Tan et al. 2011] and Tang et al. [Tang et al. 2013] used probabilistic discriminative models to learn the weights of different influence factors. Some other work learns the influence based on the state-of-art influence models. For example, Gruhl et al. [Gruhl et al. 2004] proposed a time-decayed diffusion model for blogging writing, and used an EM-like algorithm to estimate the influence probabilities. Saito et al. [Kimura et al. 2011] learned influence by solving a likelihood function based on time-decayed IC model. Some heuristic methods have also been proposed to quantify the influence. For example, Goyal et al. [Goyal et al. 2010] effectively calculated the influence probabilities by directly counting the number of actions in the collected data set.

Influence models describe the process of how users influence each other. The state-of-art influence models include two fundamental diffusion models, Linear Threshold (LT) Model and Independent Cascade (IC) Model [Kempe et al. 2003]. Recent years, several new influence models considering different factors have been proposed, such as time-decayed Independent Cascade Model [Chen et al. 2012c; Kimura et al. 2011], topic sensitive Independent Cascade Model and Linear Threshold Model [Barbieri et al. 2012], influence models considering positive and negative opinions together [Chen et al. 2011], influence model considering friend and foe relationships together [Li et al. 2013], and the diffusion model distinguishing user roles [Yang et al. 2015]. Influence maximization is one of the most important applications of influence. The objective is to find K seeds in a network that can exert maximal influence. One kind of work is to propose more efficient maximization methods [Chen et al. 2010; Goyal et al. 2011; Leskovec et al. 2007]. Another kind of work is to propose the corresponding approximate solutions for variant influence models [Barbieri et al. 2012; Chen et al. 2011; Li et al. 2013]

In this work, we mainly study the problem of influence testing and influence learning. The kind of influence we study is the aggregated influence from a user's ego network. The testing method is the statistical causal inference based on the observed data and the learning method is a heuristic method to quantify the influence from the pairwise and the structural angle.

Retweet behavior study A bulk of studies try to understand why and how people retweet. Boyd et al. [Boyd et al. 2010] gave an interesting investigation on the reasons why they retweet. The study was mainly based on human interviews and the results are lack of verification on real large data. Some other studies try to explain the retweet behaviors from different perspectives, for example, some researches focus on analyzing the effect of the content of the tweets on retweet probability. Naveed et al. [Naveed et al. 2011] used a machine learning approach to learn the weights for different features extracted from the tweet content. They found that the tweets containing hashtags, URLs or usernames are more likely to be retweeted [Naveed et al. 2011]. Macskassy et al. [Macskassy and Michelson 2011] tagged the tweets with Wikipedia categories and aggregated these tags for a particular user to generate a topics-of-interest profile for that user. They came up with four models and found the profile model (one would be more likely to retweet another user if they share similar profiles) was the most likely model. Some other perspectives such as the popularity of the topics, strength of the social ties, and the status of the publisher are also investigated by different researchers [Chen et al. 2012b; Duan et al. 2010; Suh et al. 2010; Yang et al. 2010]. Despite the success of the previous work, it would be interesting to see the influence effect of retweet behaviors from the neighbors on the user. All these works do not consider how friends in one's ego network influence the individual's behavior. This paper finds that by merely using the influence locality factors, we can train a simple predictive model to forecast users' retweet behaviors with high accuracy.

9. CONCLUSION

In this paper, we study a novel phenomenon of social influence locality, which is also proposed as a challenge in [Fernau et al. 2014; Liu et al. 2014]. We first conduct a sampling test to provide evidence of the existence of influence locality, and then formally define the influence locality function based on the observations of pairwise influence and structure influence on retweet behaviors. One interesting discovery is that retweet probability is negatively correlated with the number of *connected circles* that are formed by the active neighbors.We evaluate the influence locality functions through retweet behavior prediction by using the logistic regression model and the factor graphic model. Our experiments on retweet behavior prediction show that merely using single influence locality function, we can obtain a F1-score that is comparable with existing methods with a bunch of various features. In addition, we investigate the effect of correlation feature between the neighbors to be predicted by using the factor graphic model. The results show that the factor graphic model performs better in Precision than the logistic regression model without the correlation feature.

As future work, it is interesting to study other functions to quantify the influence locality. In addition, traditional influence models only consider the pairwise influence between users, however, from the study in this paper, we know that the aggregated influence from the local ego network is different if the ego network structure constructed by the active neighbors is different. Thus, how to define an influence model by incorporating the structure influence is an interesting problem to be studied in the future.

REFERENCES

ANAGNOSTOPOULOS, A., KUMAR, R., AND MAHDIAN, M. 2008. Influence and correlation in social networks. In *KDD'08*. 7–15.

ARALA, S., MUCHNIKA, L., AND SUNDARARAJAN, A. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *PNAS 106,* 51, 21544–21549.

BAKSHY, E., ECKLES, D., YAN, R., AND ROSENN, I. 2012. Social influence in social advertising: evidence from field experiments. In *EC'12*. 146–161.

BARBIERI, N., BONCHI, F., AND MANCO, G. 2012. Topic-aware social influence propagation models. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 81–90.

BELAK, V., LAM, S., AND HAYES, C. 2012. Cross-community influence in discussion fora. In *ICWSM'12*.

BOND, R. M., FARISS, C. J., JONES, J. J., KRAMER, A. D. I., MARLOW, C., SETTLE, J. E., AND FOWLER, J. H. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature 489*, 295–298.

BOYD, D., GOLDER, S., AND LOTAN, G. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *HICSS '10*. 1–10.

CHEN, K., CHEN, T., ZHENG, G., JIN, O., YAO, E., AND YU, Y. 2012a. Collaborative personalized tweet recommendation. In *SIGIR'12*. 661–670.

CHEN, K., CHEN, T., ZHENG, G., JIN, O., YAO, E., AND YU, Y. 2012b. Collaborative personalized tweet recommendation. In *SIGIR '12*. 661–670.

CHEN, W., COLLINS, A., CUMMINGS, R., KE, T., LIU, Z., RINCN, D., SUN, X., WANG, Y., WEI, W., AND YUAN, Y. 2011. Influence maximization in social networks when negative opinions may emerge and propagate. In *SDM'11*. 379–390.

CHEN, W., LU, W., AND ZHANG, N. 2012c. Time-critical influence maximization in social networks with time-delayed diffusion process. In *AAAI'12*.

CHEN, W., WANG, C., AND WANG, Y. 2010. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *KDD'10*. 1029–1038.

DIETZ, L., BICKEL, S., AND SCHEFFER, T. 2007. Unsupervised prediction of citation influences. In *ICML'07*. 233–240.

DUAN, Y., JIANG, L., QIN, T., ZHOU, M., AND SHUM, H.-Y. 2010. An empirical study on learning to rank of tweets. In *COLING '10*. 295–303.

DURKIN, K. 1996. "peer pressure", in: Anthony s. r. manstead and miles hewstone (eds.). *The Blackwell Encyclopedia of Social Psychology*.

FENG, W. AND WANG, J. 2013. Retweet or not?: personalized tweet re-ranking. In *WSDM'13*. ACM, 577–586.

FERNAU, H., FOMIN, F. V., LOKSHTANOV, D., MNICH, M., PHILIP, G., AND SAURABH, S. 2014. Parameterized algorithmics for computational social choice: Nine research challenges. *Tsinghua Science and Technology 19,* 4, 358–373.

GOYAL, A., BONCHI, F., AND LAKSHMANAN, L. V. 2010. Learning influence probabilities in social networks. In *WSDM'10*. 241–250.

GOYAL, A., LU, W., AND LAKSHMANAN, L. V. S. 2011. Simpath: An efficient algorithm for influence maximization under the linear threshold model. In *ICDM'11*. 211–220.

GRUHL, D., GUHA, R., LIBEN-NOWELL, D., AND TOMKINS, A. 2004. Information diffusion through blogspace. In *WWW'04*. 491–501.

GUHA, R., KUMAR, R., RAGHAVAN, P., AND TOMKINS, A. 2004. Propagation of trust and distrust. In *WWW'04*. 403–412.

HAMMERSLEY, J. M. AND CLIFFORD, P. 1971. Markov field on finite graphs and lattices. *Unpublished manuscript*.

HEINRICH, G. 2004. Parameter estimation for text analysis. Technical report.

HUBERMAN, B., ROMERO, D. M., AND WU, F. 2009. Social networks that matter: Twitter under microscope. In *First Monday*. Vol. 14. 118–138.

KEMPE, D., KLEINBERG, J., AND TARDOS, E. 2003. Maximizing the spread of influence through a social network. In *KDD'03*. 137–146.

KIMURA, M., SAITO, K., OHARA, K., AND MOTODA, H. 2011. Learning information diffusion model in a social network for predicting influence of nodes. *Intell. Data Anal. 15*, 633–652.

LA FOND, T. AND NEVILLE, J. 2010. Randomization tests for distinguishing social influence and homophily effects. In *WWW'10*. 601–610.

LESKOVEC, J., HUTTENLOCHER, D., AND KLEINBERG, J. 2010. Predicting positive and negative links in online social networks. In *WWW'10*. 641–650.

LESKOVEC, J., KRAUSE, A., GUESTRIN, C., FALOUTSOS, C., VANBRIESEN, J., AND GLANCE, N. 2007. Cost-effective outbreak detection in networks. In *KDD'07*. 420–429.

LESKOVEC, J., SINGH, A., AND KLEINBERG, J. 2006. Patterns of influence in a recommendation network. In *PAKDD'06*. 380–389.

LI, Y., CHEN, W., WANG, Y., AND ZHANG, Z.-L. 2013. Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships. In *WSDM'13*. 657–666.

LIN, S., WANG, F., HU, Q., AND YU, P. S. 2013. Extracting social events for learning better information diffusion models. In *KDD'13*. ACM, 365–373.

LIU, L., TANG, J., HAN, J., AND YANG, S. 2012. Learning influence from heterogeneous social networks. *DataMKD 25,* 3, 511–544.

LIU, Y., WU, B., WANG, H., AND MA, P. 2014. Bpgm: A big graph mining tool. *Tsinghua Science and Technology 19,* 1, 33–38.

LOU, T. AND TANG, J. 2013. Mining structural hole spanners through information diffusion in social networks. In *WWW'13*. 825–836.

LOVASZ, L. 1993. Random walks on graphs: A survey. *Combinatorics 2,* 1, 1–6.

MACSKASSY, S. A. AND MICHELSON, M. 2011. Why do people retweet? anti-homophily wins the day! In *ICWSM*.

MILGRAM, S. 1967. The small world problem. *Psychology Today 2*, 60–67.

MYERS, S. A., ZHU, C., AND LESKOVEC, J. 2012. Information diffusion and external influence in networks. In *KDD '12*. 33–41.

NAVEED, N., GOTTRON, T., JÉRÔME, AND ALHADI, A. C. 2011. Bad news travel fast: A content-based analysis of interestingness on twitter. In *WebSci'11*.

PEARL, J. 2009. *Causality: Models, Reasoning and Inference*. ambridge University Press.

PETROVIC, S., OSBORNE, M., AND LAVRENKO, V. 2011. Rt to win! predicting message propagation in twitter. In *ICWSM'11*.

ROBERTSON, R. 1992. *Globalization: Social theory and global culture*. Vol. 16. Sage.

SAITO, K., NAKANO, R., AND KIMURA, M. 2008. Prediction of information diffusion probabilities for independent cascade model. In *KES '08*. 67–75.

SHUAI, X., DING, Y., BUSEMEYER, J., CHEN, S., SUN, Y., AND TANG, J. 2012. Modeling indirect influence on twitter. *IJSWIS 8,* 4.

SUH, B., HONG, L., PIROLLI, P., AND CHI, E. H. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *SOCIALCOM '10*. 177–184.

SUN, J., QU, H., CHAKRABARTI, D., AND FALOUTSOS, C. 2005. Neighborhood formation and anomaly detection in bipartite graphs. In *ICDM'05*. 418–425.

TAN, C., LEE, L., TANG, J., JIANG, L., ZHOU, M., AND LI, P. 2011. User-level sentiment analysis incorporating social networks. In *KDD'11*. 1049–1058.

TANG, J., SUN, J., WANG, C., AND YANG, Z. 2009. Social influence analysis in large-scale networks. In *KDD'09*. 807–816.

TANG, J., WU, S., AND SUN, J. 2013. Confluence: Conformity influence in large social networks. In *KDD'13*. 347–355.

TANG, W., ZHUANG, H., AND TANG, J. 2011. Learning to infer social ties in large networks. In *ECML/PKDD'11*. 381–397.

UGANDER, J., BACKSTROM, L., MARLOW, C., AND KLEINBERG, J. 2012. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*.

WENG, J., LIM, E.-P., JIANG, J., AND HE, Q. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM'10*. 261–270.

WU, S., HOFMAN, J. M., MASON, W. A., AND WATTS, D. J. 2011. Who says what to whom on twitter. In *WWW'11*. 705–714.

WU, S., SUN, J., AND TANG, J. 2013. Patent partner recommendation in enterprise social networks. In *WSDM'13*. ACM, 43–52.

WU, Y., ZHOU, C., XIAO, J., KURTHS, J., AND SCHELLNHUBER, H. J. 2010. Evidence for a bimodal distribution in human communication. *Proceedings of the national academy of sciences 107,* 44, 18803–18808.

YANG, S.-H., LONG, B., SMOLA, A., SADAGOPAN, N., ZHENG, Z., AND ZHA, H. 2011. Like like alike: joint friendship and interest propagation in social networks. In *WWW*. 537–546.

YANG, Y., TANG, J., LEUNG, C. W.-K., SUN, Y., CHEN, Q., LI, J., AND YANG, Q. 2015. Rain: Social role-aware information diffusion. In *AAAI'15*.

YANG, Z., GUO, J., CAI, K., TANG, J., LI, J., ZHANG, L., AND SU, Z. 2010. Understanding retweeting behaviors in social networks. In *CIKM '10*. 1633–1636.

ZHANG, J., TANG, J., ZHUANG, H., LEUNG, C. W.-K., AND LI, J. 2014. Role-aware conformity influence modeling and analysis in social networks. In *AAAI'14*. 958–965.