Social Influence Locality for Modeling Retweeting Behaviors

Jing Zhang, Biao Liu, Jie Tang, Ting Chen, and Juanzi Li
Department of Computer Science and Technology, Tsinghua University, China
{zhangjing0544,liubiao2638,iamtingchen}@gmail.com, {jietang, ljz}@tsinghua.edu.cn

Abstract

We study an interesting phenomenon of *social influence locality* in a large microblogging network, which suggests that users' behaviors are mainly influenced by close friends in their ego networks. We provide a formal definition for the 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 functions. Our analysis also reveals several intriguing discoveries. For example, though the probability of a user retweeting a microblog is positively correlated with the number of friends who have retweeted the microblog, it is surprisingly negatively correlated with the number of *connected circles* that are formed by those friends.

1 Introduction

Social influence captures the ways in which people affect each others' opinions, emotions, and behaviors. Roughly speaking, social influence has global patterns and local patterns. Examples of the global patterns include the influence from an opinion leader and the influence by a hot topic. Examples of local patterns include pairwise influence and community influence. Much research has been conducted in this field including pairwise influence [Goyal *et al.*, 2010; Saito *et al.*, 2008], topic influence [Liu *et al.*, 2012; Tang *et al.*, 2009], indirect influence [Shuai *et al.*, 2012], external influence [Myers *et al.*, 2012], and community influence [Belak *et al.*, 2012]. However, there is still lack of a formal definition of the global (or local) effect of influence that a user receives in the social network.

In this paper, we study an interesting problem on how users' behaviors are influenced by friends in their ego network. In particular, we focus on studying retweet behaviors in a large microblogging network, Weibo.com[1]. We try to understand the underlying mechanism that users' retweet behaviors influence with each other. We formulate the notion of social influence locality and verify its effects in the microblogging network.

Definition 1 Social Influence Locality. *Let $G = (V, E)$ denote a social network, where V is a set of users and $E \subset V \times V$ is a set of directed relationships between users. For a user $v \in V$, we use $G_v^\tau \subseteq G$ to denote v's τ-ego network, where τ-ego network means a subnetwork formed by v's τ-degree friends in the network G and $\tau \geq 1$ is a tunable integer parameter to control the scale of the ego network. Assume each user is associated with an action status $s_v \in \{0, 1\}$, with $s_v = 1$ indicating active and $s_v = 0$ inactive.*

Given $S_v = \{v_i | v_i \in G_v^\tau \wedge s_{v_i} = 1\}$ as the collection of active neighbors in v's ego network G_v^τ, social influence locality is defined as a function to quantify how user v's behavior (action status) is influenced by other users in her τ-ego network,

$$Q(S_v, G_v^\tau), \quad \textit{with } \tau \in \mathbb{N}^+ \tag{1}$$

Here we only give a general definition of social influence locality, which can be instantiated in different ways. In the definition, we define the action status as binary for simplicity, but in principle it can be extended to multiple values. Also, for the τ-ego network, we can consider either bi-directional relationships or directional relationships. For example, for modeling the retweet behaviors in the microblogging network, as users have directed (following) relationships with each other, we consider all the following relationships between users in the τ-ego network. Figure 1 shows an example of the τ-ego network ($\tau = 2$) of user v (centered in the network), with six directed neighbors active and five inactive. By taking a close look at the inner structure formed by those neighbors, we could find that users A and B are not directly connected, E and F are connected with each other, while C and D also are not directly connected, but they are connected via H, except v. In this sense, the six active neighbors form four connected circles[2].

It is non-trivial to instantiate the function $Q(S_v, G_v^\tau)$ of social influence locality. As shown in the example of Figure 1,

[1] The most popular Chinese microblogging service.

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

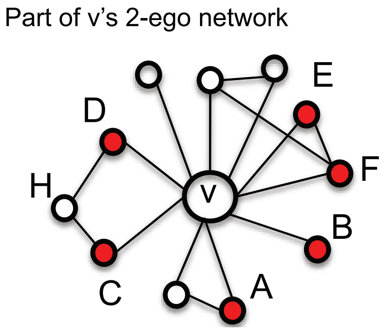


Figure 1: Illustration of social influence locality for user v in her 2-ego network ($\tau = 2$). Given a microblog m, red nodes represent "active" users who have retweeted m, while the white nodes denote those users in v's 2-ego network who did not retweet.

the influence not only depends on the number of users who have already become active, but may be also correlated with the inner structure formed by the "active" users.

Results. In order to instantiate the influence locality function $Q(S_v, G_v^\tau)$. We first perform an investigation to test whether influence locality exists in the micrologging network and whether it significantly influences users' retweet behaviors. Then we focus on studying the effects from the pairwise influence and structure influence. Based on the study, we give instantiation functions of social influence locality for modeling the retweet behaviors. We have several interesting discoveries from the study:

- There is 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 2).
- Though the probability of a user retweeting a microblog is positively correlated with the number of active neighbors, it is surprisingly negatively correlated with the number of *connected circles* that are formed by those neighbors. Especially when the number of active neighbors is larger than 10, the probability will decrease about 10% from 1 circle to 2 circles (Cf. Figure 4).
- 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 3).

The defined influence locality function has strong predictive power. We employ it for modeling and predicting users' retweet behaviors. With merely a few features defined based on the influence locality functions, we could learn a simple classifier which results in good prediction performance, which is even better than existing methods which employ various features by +0.6% in terms of F1-measure.

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

Table 1: Data statistics.

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

2 Data Description

The microblogging network we used in this study was crawled from Sina Weibo.com, which, similar to Twitter, allows users to follow with 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 called the follower of B and B is called the followee of A.

The data set was crawled in the following ways. To begin with, 100 random users were selected as seed users, and then their followees and followees' followees were collected. The crawling process produced in total 1.7 million users and 4 billion following relationships among them, with average 200 followees per user. For each user, the crawler collected her 1,000 most recent microblogs (including tweets and retweets). The process resulted in totally 1 billion microblogs. We also crawled all the users' profiles which contain name, gender, verification status, #bi-following, #followers, #followees, and #microblogs.

We focus on the retweet behaviors in the microblogging network. Thus we select 300,000 popular microblog diffusion episodes from the data set. Each diffusion episode contains the original microblog and all its retweets. On average each microblog has been retweeted for about 80 times. The sampled data set ensures that for each diffusion episode, the active (retweet) statuses of followees in one's τ-ego network is completed. Table 1 lists statistics of the crawled network.

3 Sampling Test for 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 followee 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 with 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 followee retweeted a given 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:

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

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 hour. For each user who has retweeted m, we view her as active at the 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 inactive instance. Finally, we can determine the instances in the original treatment and control groups, and learn the logistic regression model based on them.

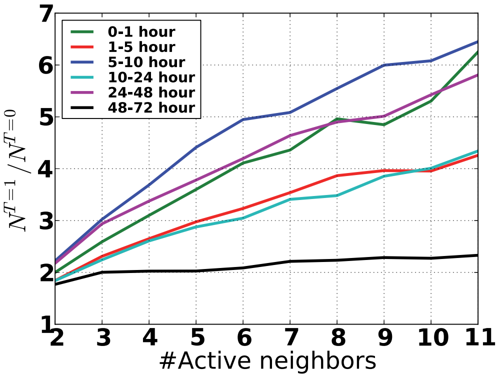


Figure 2: The result of matching-based sampling test for influence locality. $N^{T=1}$ is the average number of active users in the treatment group, and $N^{T=0}$ is the average number of active users in the control group.

Results. The test results are shown in Figure 2. In this test and the following experiments in the paper, we set the parameter τ as 1 and hence focus on the 1-ego network. 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.

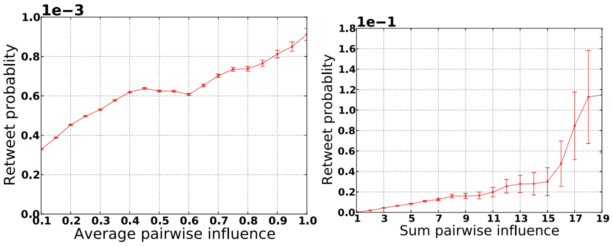
In the figure, $N^{T=1}$ is the average number of active users in the treatment group, and $N^{T=0}$ is the average number of active users in the control group. We calculate the ratio of the fractions for the two numbers and can conclude that the influence locality exerts positive effect on users' retweet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

4 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. In addition, the influence may exist between either directly connected users or users with indirected relationships. To quantify this, we cast the problem 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 a user v's τ-ego network



(a) Average pairwise influence (b) Sum pairwise influence

Figure 3: 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.

G_v^τ and calculate the random walk probability p_{v_i} for each active neighbor v_i. The random walk probability can be explained as how the influence of an active neighbor can finally reach the given user v via the network connection between them. For example, as shown in Figure 1, 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 3 shows the probability that a user retweets a microblog conditioned on the average random walk probability (a) and sum of the random walk probability (b) 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 behaviors.

Structure Influence. As observed in Figure 1, 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* 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 4 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.

Please note that when calculating the number of circles, we only consider reciprocal (bi-directional) following relationships between users. This is because, we find that directional links 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].

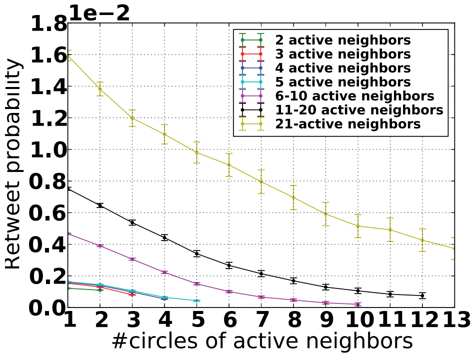


Figure 4: The effect of structure influence. Structural diversity is represented by the number of circles formed by the active neighbors.

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

$$Q(S_v, G_v^\tau) = w \times g(S_v, G_v^\tau) + (1 - w) \times f(S_v, G_v^\tau) \quad (4)$$

where $g(S_v, G_v^\tau)$ denotes the pairwise influence and $f(S_v, G_v^\tau)$ 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.,

$$g(S_v, G_v^\tau) = \sum_{v_i \in S_v} p_{v_i} \quad (5)$$

where p_{v_i} is the random walk probability from the active user v_i to the given user v. 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,

$$g(S_v, G_v^\tau) = \sum_{v_i \in S_v} h_{v_i} p_{v_i} \quad (6)$$

where h_{v_i} is the difference between the time when v_i retweeted the microblog and the time when we try to predict v'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 4, the influence does not linearly decrease. Thus we give two definitions. The first one uses the exponential function, i.e.,

$$f(S_v, G_v^\tau) = e^{-\mu|C(S_v)|} \tag{7}$$

where $C(S_v)$ is the collection of circles formed by the active neighbors and μ is a decay factor. The other function additionally considers the influence from the number of the active neighbors:

$$f(S_v, G_v^\tau) = a\log(|S_v| + 1) + be^{-\mu|C(S_v)|} \tag{8}$$

where a and b are two balance parameters. This definition linearly combines the logarithm function for the number of the active neighbors and the exponential function for the number of the circles formed by them.

5 Retweet Behavior Prediction

The defined influence locality function has strong predictive power and can be used for many applications such as retweet behavior prediction and social recommendation. In this section, we introduce how to apply the influence locality function to retweet behavior prediction.

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

$$P(s_{v,m,t} = 1|X_{v,m,t}) = \frac{1}{1 + e^{-(\alpha X_{v,m,t} + \beta)}} \tag{9}$$

where $X_{v,m,t}$ is the feature vector of user v associated with m at time t, and α are weights of the features and β is a bias, both of which are learned by maximizing an likelihood objective function that can be similarly defined as Eq. 3.

6 Experimental Results.

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

6.1 Experimental Setup

Data Preparation. We use the data set described in Section 2 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 hour 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.

Additional Features. Besides the influence locality based features, we can also consider other basic features that usually used in the traditional methods for retweet prediction. We define three kinds of basic features including personal attributes, instantaneity and topic propensity. Specifically, we use six personal attributes including gender (0 indicates male and 1 indicates female), verification status (0 indicates being verified as a celebrity and 1 indicates not being verified), the number of followers, parasocial following relationships, reciprocal following relationships, and historical microblogs. Instantaneity is defined as the elapsed time from when the original microblog m published. 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.

To obtain the topic distributions for all the microblogs and users, we firstly treat every microblog as a document and utilize Latent Dirichlet Allocation [Heinrich, 2004] to estimate the probability of generating a microblog m from each topic k. Then we estimate the probability of generating a user u from each topic k by averaging the probabilities of all her historical microblogs associated to topic k.

Comparison Methods and Evaluation Metrics. Our method (named as LRC-Q) only uses the influence locality function $Q(S_v, G_v^\tau)$ as features to train the logistic regression classifier and to predict retweet behaviors. We compare with the classifier using the above defined basic features (LRC-B). We also incorporate the defined influence locality functions into the baseline LRC-B method, which results in a new comparison method named LRC-BQ. We set w as 0.5 for Q function and select g_6 and f_2 presented in Table 3. For f function, we empirically set μ as 1, a and b as 0.5.

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.

6.2 Performance and Analysis.

Table 2 shows the performance of the comparison methods. The results show that using only the influence locality function to predict retweet behaviors (LRC-Q) can result in a performance comparable with (even better than) that 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).

Table 2: 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 3: Performance of LRC-Q by using different $Q = w \times g + (1 - w) \times f$ 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
$f_1 = e^{-\mu \lvert C(S_v) \rvert}$	68.26	68.33	68.30	68.28
$f_2 = a \log(\lvert S_v \rvert + 1) + be^{-\mu \lvert C(S_v) \rvert}$	**68.64**	**68.96**	**68.80**	**68.48**

Influence Locality Functions. We further try different $Q = w \times g + (1 - w) \times f$ functions for predicting retweet behaviors. Specifically, we first set $w = 1$ and try seven g functions for pairwise influence and then set $w = 0$ and try two kinds of f functions for structure influence defined in Section 4. The evaluation results are shown in Table 3. We can see that for pairwise influence, g_6, which averages the time weighted pairwise influence by using geometric mean, performs the best. The result suggests that the followees with different retweet time actually exert different influence on retweet behaviors. Besides, we also find that arithmetic mean performs poorly comparing with 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 estimates of the pairwise influence from the active neighbors are not normally distributed but right-skewed. That is, the majority of pairwise influence are low and a minority of pairwise influence are scattered in a fat right tail. We can see that for structure influence, f_2 function considering both the number of active neighbors and the number of circles formed by them performs better than only considering the number of circles, which indicates that there in deed exists strong correlation between the two factors.

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 5 plots the accuracy of LRC-Q with various values of w, where g is set as g_6 and f is set as f_2 according to the best performance presented in Table 3. We see that the highest accuracy is obtained when w is 0.5.

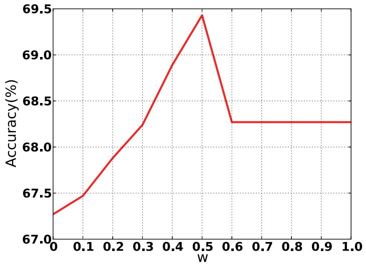


Figure 5: Performance of LRC-Q under different values of w.

7 Related Work.

Existing social influence research studies different forms of influence. For example, Tang et al. [Tang *et al.*, 2009] and Liu et al. [Liu *et al.*, 2012] propose measuring the influence on different topics. Goyal et al. [Goyal *et al.*, 2010] and Saito et al. [Saito *et al.*, 2008] measure 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] study the indirect influence using the theory of quantum cognition. Myers et al. [Myers *et al.*, 2012] propose a probabilistic model to quantify the external influence out-of-network sources. Belak et al. [Belak *et al.*, 2012] investigate and measure the influence between two communities. In this work, we study the influence from a user's ego network and formally defines it as social influence locality.

A bulk of studies try to understand why and how people retweet. Boyd et al. [Boyd *et al.*, 2010] give an interesting investigation on the reasons why they retweet. Some other researches try to explain the retweet behaviors from different perspectives, for example, popularity of the topics, strength of the social ties, and the status of the publisher [Chen *et al.*, 2012; Duan *et al.*, 2010; Suh *et al.*, 2010; Yang *et al.*, 2010].

8 Conclusion

In this paper, we study a novel idea of social influence locality for modeling users' retweet behaviors in the microblogging network. We first conduct a sampling test to provide evidence for the existence of influence locality, and then formally define the influence locality function based on the observations of pairwise influence and structure influence on the retweet behaviors. 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.

As future work, it is interesting to study other functions to quantify the influence locality. It is also interesting to extend study on larger scale ego network ($\tau > 1$). For the retweet prediction problem, it is also helpful to design a better predictive model with higher accuracy.

Acknowledgements. The work is supported by the NSFC (No. 61222212, 61073073, 61170061, 60933013, 61035004), National Basic Research Program of China (No. 2011CB302302), and a research fund from Huawei company.

References

[Arala *et al.*, 2009] Sinan Arala, Lev Muchnika, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *PNAS*, 106(51):21544–21549, 2009.

[Belak *et al.*, 2012] Vclav Belak, Samantha Lam, and Conor Hayes. Cross-community influence in discussion fora. In *ICWSM'12*, 2012.

[Boyd *et al.*, 2010] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *HICSS '10*, pages 1–10, 2010.

[Chen *et al.*, 2012] Kailong Chen, Tianqi Chen, Guoqing Zheng, Ou Jin, Enpeng Yao, and Yong Yu. Collaborative personalized tweet recommendation. In *SIGIR '12*, pages 661–670, 2012.

[Duan *et al.*, 2010] Yajuan Duan, Long Jiang, Tao Qin, Ming Zhou, and Heung-Yeung Shum. An empirical study on learning to rank of tweets. In *COLING '10*, pages 295–303, 2010.

[Goyal *et al.*, 2010] Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan. Learning influence probabilities in social networks. In *WSDM'10*, pages 241–250, 2010.

[Heinrich, 2004] Gregor Heinrich. Parameter estimation for text analysis. Technical report, 2004.

[Huberman *et al.*, 2009] B. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under microscope. In *First Monday*, volume 14, pages 118–138, 2009.

[Kempe *et al.*, 2003] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD'03*, pages 137–146, 2003.

[Liu *et al.*, 2012] Lu Liu, Jie Tang, Jiawei Han, and Shiqiang Yang. Learning influence from heterogeneous social networks. *DataMKD*, 25(3):511–544, 2012.

[Lovasz, 1993] László Lovasz. Random walks on graphs: A survey. *Combinatorics*, 2(1):1–6, 1993.

[Myers *et al.*, 2012] Seth A. Myers, Chenguang Zhu, and Jure Leskovec. Information diffusion and external influence in networks. In *KDD '12*, pages 33–41, 2012.

[Pearl, 2009] Judea Pearl. *Causality: Models, Reasoning and Inference*. ambridge University Press, 2009.

[Saito *et al.*, 2008] Kazumi Saito, Ryohei Nakano, and Masahiro Kimura. Prediction of information diffusion probabilities for independent cascade model. In *KES '08*, pages 67–75, 2008.

[Shuai *et al.*, 2012] Xin Shuai, Ying Ding, Jerome Busemeyer, Shanshan Chen, Yuyin Sun, and Jie Tang. Modeling indirect influence on twitter. *IJSWIS*, 8(4), 2012.

[Suh *et al.*, 2010] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *SOCIALCOM '10*, pages 177–184, 2010.

[Sun *et al.*, 2005] Jimeng Sun, Huiming Qu, Deepayan Chakrabarti, and Christos Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *ICDM'05*, pages 418–425, 2005.

[Tang *et al.*, 2009] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social influence analysis in large-scale networks. In *KDD'09*, pages 807–816, 2009.

[Ugander *et al.*, 2012] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 2012.

[Yang *et al.*, 2010] Zi Yang, Jingyi Guo, Keke Cai, Jie Tang, Juanzi Li, Li Zhang, and Zhong Su. Understanding retweeting behaviors in social networks. In *CIKM '10*, pages 1633–1636, 2010.