Jia YT, Wang YZ, Cheng XQ. Learning to predict links by integrating structure and interaction information in microblogs. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 30(4): 829–842 July 2015. DOI 10.1007/s11390-015-1563-9

Learning to Predict Links by Integrating Structure and Interaction Information in Microblogs

Yan-Tao Jia (贾岩涛), *Member, CCF, ACM*, Yuan-Zhuo Wang (王元卓), *Member, CCF, ACM, IEEE* and Xue-Qi Cheng (程学旗), *Senior Member, CCF, Member, ACM, IEEE*

Key Laboratory of Network Science and Technology, Institute of Computing Technology Chinese Academy of Sciences, Beijing 100190, China

E-mail: {jiayantao, wangyuanzhuo, cxq}@ict.ac.cn

Received December 2, 2014; revised April 11, 2015.

Abstract Link prediction in microblogs by using unsupervised methods has been studied extensively in recent years, which aims to find an appropriate similarity measure between users in the network. However, the measures used by existing work lack a simple way to incorporate the structure of the network and the interactions between users. This leads to the gap between the predictive result and the ground truth value. For example, the $F1$-measure created by the best method is around 0.2. In this work, we firstly discover the gap and prove its existence. To narrow this gap, we define the retweeting similarity to measure the interactions between users in Twitter, and propose a structural-interaction based matrix factorization model for following-link prediction. Experiments based on the real-world Twitter data show that our model outperforms state-of-the-art methods.

Keywords link prediction, microblog, structure-interaction, retweeting similarity, matrix factorization

1 Introduction

The link prediction in microblogs such as Twitter has been extensively studied during recent years due to the applications in viral marketing[1], friendship recommendation[2], community detection, etc. Link prediction in microblogs is more difficult to handle and is a more challenging field. First, the microblog is a hybrid social-information and directed network, where people are connected not only by explicit following relations, but also by implicit information propagation process. It is desired to find a framework to unify and balance the social aspect and the information aspect of microblogs. Second, microblogs are evolving with a rapid speed over time, and the traditional methods for link prediction on static network cannot afford this challenges. Moreover, the size of the complete networks in microblogs is so huge that it is impossible to crawl all of them. A good sampling method to obtain one or more tropical snapshots of the network needs more consideration.

Although there are many challenges in link prediction process in microblogs, the common methodology used in social networks is still instructive. Generally, the methods in this research area can be classified into two parts: the supervised methods and the unsupervised methods. Supervised methods treat the link prediction as a classification problem, whereas many classic machine learning algorithms can be introduced, such as the supervised random walk algorithm in [3] and the logistic regression model in [4]. Although the supervised methods are the state-of-the-art methods in link prediction, they often suffer from the so-called imbalance and feature selection problem. In contrast, the unsupervised methods do not need to know the prior knowledge of the distribution of the dataset and can avoid the

Regular Paper

Special Section on Data Management and Data Mining

This work is supported by the National Basic Research 973 Program of China under Grant Nos. 2013CB329602 and 2014CB340405, the National Natural Science Foundation of China under Grant Nos. 61173008, 61232010, 60933005, 61402442, 61402022, and 61303244, Beijing Nova Program under Grant No. Z121101002512063, and the Natural Science Foundation of Beijing under Grant No. 4154086.

drawbacks of the supervised methods. The unsupervised methods intend to define statistics to measure the similarity between two users, such as common neighbors, Jaccard coefficients, the Adamic-Adar measure[5], preferential attachment[6], Katz measure[7], and so on. More recently, more various similarities are considered, such as the geographical location based similarity of the users[8], the geographical distance or the time zone based similarity[9-10], the social theory based similarity such as the link homophily, status homophily and the structure balance theory[9-10], the text similarity of the tweets posed by the users[8], and so on. Very recently, Yin *et al.*[11] defined the structure similarity measures between two users with respect to another user, and proposed a structure-based matrix factorization model (S-Model) for link prediction in microblogs. They discovered that the model achieved higher $F1$-measure than that obtained by other seven measures such as the PropFlow method ([12]), the Jaccard coefficient and so on. For example, the S-Model obtains $F1$-measure 0.197 in dynamic setting with an increase of 0.03 compared with the Jaccard coefficient method. In this sense, S-Model can be regarded as the best method for unsupervised methods.

Although S-Model gets a higher $F1$-measure, we discover that there still exists a gap between the predictive value and the ground truth for link prediction. To narrow this gap, in this paper, we propose an unsupervised method, the so-called structural-interaction model (SI-Model), which integrates the structural information and the interaction information between users to predict the future links of following. This idea comes from the observation that the interaction between users correlates with the formation of the following links on Twitter. For example, Akasaka *et al.*[13] pointed out this from the aspect of the differences between retweeting and following. Akasaka *et al.* examined the most popularly retweeted tweets from celebrities and noted that a surprising number of individuals retweeted those they actually did not follow. Furthermore, they found many of the users who were not following have become followers within the following year. More precisely, we define a similarity measure, called retweeting similarity, to measure the similarity of two users in terms of their interactions with another user respectively.

Our contributions are listed as follows.

• We firstly find the gap between the predictive result and the ground truth by using the matrix factorization method and prove its existence. The purpose of proving the gap is to show that there is still room for improving the state-of-the-art work. Meanwhile, during the proof, we further discover the necessary condition under which the gap exists, and explain one possible reason of the existence of the gap.

• To compensate the gap, we define the retweeting similarity to measure the interactions between users in Twitter, and propose a structural-interaction based matrix factorization model for following-link prediction. Theoretical analysis demonstrates that it can be used to narrow the gap if the parameters in the model are appropriately chosen.

• Experiments based on the real-world Twitter data show that under the same computational complexity, our model outperforms the state-of-the-art methods. For example, it reduces the RMSE value by about 70% compared with that of the best method.

The organization of this paper is as follows. In Section 3, we find the gap between the predictive value and the ground truth of S-Model experimentally. Section 4 is devoted to the theoretical analysis of the gap between the predictive result and the ground truth. In Section 5, we define retweeting similarity, explain the proposed model and present a theoretical analysis of the model's effectiveness. Section 6 presents experimental results that validate the effectiveness of our methodology. Finally, we conclude our work in Section 7.

2 Related Work

In this section, we shall briefly recall three types of methods used in link prediction problem, the content-based methods, the behavior-based methods, and the matrix factorization methods.

Link prediction in Twitter by integrating the tweets information and the structure information is a new way to improve the performance of prediction. The methods used in this direction are two-folded: one is content-based method by investigating the text of the tweets and the other is behavior-based by collecting the behaviors on tweets such as the retweet, reply or mention actions. For example, Sadilek *et al.*[8] combined the content of the tweets, users' location and the network structure features for friendship prediction problem. The text similarity of two tweets of two users is defined to be the amount of overlap in the vocabularies used by two users. Then a regression decision tree model is used to unify the three features. Such kind of methods needs to tackle the text of the tweets, and sometimes is time-consuming. In contrast, the behavior-based methods only care about the interaction information between users. This method makes

full advantage of the property of Twitter as a social media (see [14]) or an information diffusion channel. One Twitter user A can address tweets of user B, and then mentions B obliquely in his or her tweets, which is syntaxed as "$RT@B$". Another common practice is that A "retweets" or rebroadcasts B's message, which is syntaxed by $@B$. For a tweet message, the behavior-based methods extract the usernames after the symbol @, and consider that A and B have an interaction relationship. Hopcroft *et al.*[9] considered these interaction relationships and defined four features to represent the number of retweets or replies from user A to user B and from user B to user A, respectively. By integrating other features, they proposed a supervised method, i.e., the Triad Factor Graph model, to predict the reciprocity link. Similar work can refer to that by Lou *et al.*[10] Our model is also behavior-based. The difference is that we integrate the structure and the interaction behavior into a simpler matrix factorization framework.

As for the matrix factorization method used in the link prediction problem, it is motivated by the successful application of matrix factorization used in recommender systems, where the model aims to find latent features for users and items by factorizing the observed matrix, see [15-17]. Converting the user-item pair to the user-user pair leads to the link prediction problem as a link recommendation problem. Related work can be found in the work of Menon and Elkan[18] and Yin *et al.*[11] Their models learned the latent features just from the topological structures of the network. For example, Yin *et al.*[11] analyzed the role of the intermediate user between two users, and divided its contribution into two parts: one is the recommendation of the intermediate user, and the other is the acceptance of the recommendation of the intermediate user. Very recently, Zhang *et al.*[19] enhanced Yin *et al.*'s work to find the real intermediate users and studied how they contribute to the link formation process. To better predict new links in time-evolving social networks, Gao *et al.*[20] integrated three types of information: the global network structure, the content of nodes in the network and the local information of a given vertex to derive a matrix factorization model. Similar work by using the matrix factorization method or the tensor factorization method can refer to [8, 21-22], etc. However, these methods lack the consideration of the impact of interactions between users on the link prediction. Our work mixes the interaction information between users with the structure of the network.

3 Test the Existence of the Gap by Experiment

In this section, we first examine the different performances of the S-Model by Yin *et al.*[11] on datasets with different sparseness and get its best predictive performance, i.e., the maximum $F1$-measure obtained by S-Model. Experiments show that this maximum $F1$-measure does not take its theoretical maximum 1. This leads to a hypothesis that there exists a gap between the predictive performance and the ground truth for S-Model. To narrow this gap, we propose the idea to use the interaction information between users in the dataset.

Before reconstructing the experiment of Yin *et al.*[11], let us simply recall S-Model as follows. The idea of S-Model is to predict new follower v_i (called the target user) of the source user v_u via the contributions of some intermediate user v_k. The contributions of v_k can be divided into two parts: one is the recommendation of v_i to v_u, and the other is v_u's acceptance of the recommendation of v_k for v_i. Then S-Model studies the influence of the network structures on v_k's contributions by introducing the structure similarity between users. After using the matrix factorization technique, S-Model can predict new link formation for one static network as well as two snapshots of the network in Twitter. To find the best performance of S-Model, we conduct the experiment on the static dataset with different sparseness. Here the sparseness, denoted by nf, means the average number of non-followers for a number of users. We tune the sparseness of the dataset recursively from one original dataset by randomly converting some number of followers to non-followers. In other words, if we construct a rating-like matrix with the row corresponding to the source users and the column corresponding to the target users, denoted by $\boldsymbol{R}_{n\times m} = (r_{ui})$, where n is the number of source users and m is the number of target users, $r_{ui} = 1$ if v_u follows v_i and $r_{ui} = 0$ otherwise, the tuning process is to randomly replace some number of 1's for each row with 0 respectively. The initial dataset corresponds to the matrix with all elements being 1 except the diagonal elements, with the sparseness $nf = 1$. It is easy to see that $1 \leqslant nf \leqslant m-1$. The experiment is carried out by fixing both the smoothing factor and the structural factor being 0.01 in S-Model and setting $m = 10\,000$ to find the relation between the $F1$-measure and the value nf. We depict the relation for $nf = 1, \ldots, 31$ as follows, since for the rest part, the tendency of the curves is similar.

From Fig.1, we can see that the maximum $F1$-measure obtained by S-Model is 0.007 when $nf = 29$.

We also explore the performance of the curve when $nf = 9\,999$, and find the $F1$-measure takes zero. It is obvious that this value fails to obtain the theoretical maximal $F1$-measure 1. In this sense, we propose the hypothesis that there is a gap between the predictive performance and the ground truth for S-Model.

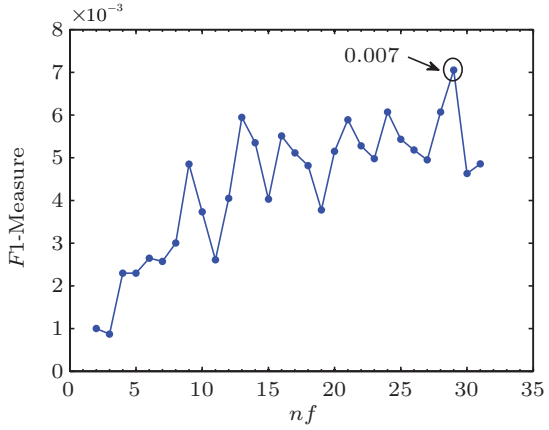


Fig.1. Relation between nf and the $F1$-measure of S-Model.

To narrow this gap, we turn attention to the interaction information between users in the dataset. Specifically, we find that the consideration of v_k's contributions in S-Model only focuses on the link structural information between v_u and v_k, and between v_k and v_i. In fact, there is another important ingredient to measure the contributions of v_k's recommendation of v_i to v_u, namely, the interaction information between v_k and v_i. This interaction information in microblogs such as Twitter is referred to as the retweet and reply behaviors between users. In this paper, we mainly focus on the retweeting behaviors. We illustrate the influence of these interactions on the link prediction problem as follows. This example can also be regarded as the triad closure process, which has been studied by many authors, for example, see the work of Romero and Kleinberg[23].

Suppose that before a specific time t_0, v_u follows three intermediate users v_k, $v_{k'}$ and $v_{k''}$, and each of these three users follows the target user v_i, see Fig.2. The goal is to predict the link $v_u \to v_i$ at time t_1. To better achieve this goal, we consider the interaction information between any pairs (v_k, v_i), $(v_{k'}, v_i)$ and $(v_{k''}, v_i)$. If v_k retweets v_i, which is depicted in the picture, then compared to the other intermediate users $v_{k'}$ and $v_{k''}$, its contribution is larger. Hence, our method not only is based on the link structures of the network, but also concentrates on the interaction information between users. By considering the interaction information between users, our link prediction problem can be formulated as follows: for a fixed user set U, given the network structure of the Twitter network for users in U at time t_0 and the interaction network between users in the time interval $(t_0, t_1]$, where two users in the interaction network are connected if they had interactions during $(t_0, t_1]$, we aim to predict new followers of one given user during the time interval $(t_0, t_1]$.

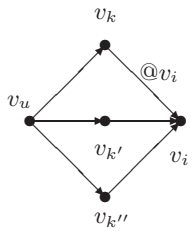


Fig.2. Picture to characterize the structures and interactions between users.

4 Theoretical Proof of the Existence of the Gap

In this section, we shall accept the hypothesis posed in Section 3 by verifying the existence of the the gap between the predictive result and the ground truth of S-Model. To minimize this gap, we shall propose SI-Model in Section 5.

Before proving the existence of the gap, we introduce some notations and definitions. Let $\boldsymbol{R}_{n\times m} = (r_{ui})$ denote the rating-like 0-1 matrix, where n is the number of source users and m is the number of target users, $r_{ui} = 1$ if v_u follows v_i and $r_{ui} = 0$ otherwise. The purpose of S-Model is to factorize the matrix $\boldsymbol{R}$ into two latent matrices $\boldsymbol{A}_{n\times K}$ and $\boldsymbol{B}_{K\times m}$ with $K \ll n$ and $K \ll m$. In terms of the elements of the matrices, this factorization is stated as $r_{ui} = \sum_{k=1}^{K} a_{uk} b_{ki}$, where r_{ui}, a_{uk}, b_{ki} are the elements of $\boldsymbol{R}$, $\boldsymbol{A}$, $\boldsymbol{B}$ respectively. The meanings of a_{uk} and b_{ki} are that a_{uk} measures the extent of v_u's acceptance of the recommendation of v_k, and b_{ki} scores the strength of v_k's recommendation for v_i to v_u. To avoid the over-fitting of the factorization, S-Model also considers the Gaussian prior distribution of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, as posed by Salakhutdinov and Mnih[24]. Moreover, to reveal the network structures, S-Model introduces the structural regulation restrictions during the factorization process. Specifically, S-Model first defines the effective structure

set $S^e = \{\Longrightarrow, \Longleftarrow, \Longleftrightarrow\}$ among the structures of the network as the elementary structures between two users one-hop away. Let $S_{u,k}$ be the set of structures from v_u to v_k, and $S^e_{u,k}$ be the set of effective structures from v_u to v_k. It is clear that $S^e_{u,k} = S_{u,k} \cap S^e$. Based on the effective structure set S^e, the structural similarity between two intermediate users v_k and $v_{k'}$ with respect to the source user v_u, denoted by $W_u(k,k')$ is defined as $W_u(k,k') = 1$ if $S^e_{u,k} = S^e_{u,k'}$ and $W_u(k,k') = 0$ otherwise. Similarly, the structural similarity between two target users v_i and v_j with respect to the intermediate user v_k, denoted by $W_k(i,j)$ can be also defined. Next, S-Model defines the structure regulation functions $S(\boldsymbol{A}), S(\boldsymbol{B})$ for the matrices $\boldsymbol{A}$, $\boldsymbol{B}$ respectively as follows:

$$S(\boldsymbol{A}) = \frac{\sum\limits_{u=1}^{n}\sum\limits_{k=1}^{K}\sum\limits_{k'=1}^{K} W_u(k,k')(a_{uk} - a_{uk'})^2}{\sum\limits_{u=1}^{n}\sum\limits_{k=1}^{K}\sum\limits_{k'=1}^{K} W_u(k,k')},$$

$$S(\boldsymbol{B}) = \frac{\sum\limits_{k=1}^{K}\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{m} W_k(i,j)(b_{ki} - b_{kj})^2}{\sum\limits_{k=1}^{K}\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{m} W_k(i,j)}.$$

With the structure regulation functions, S-Model aims to minimize the objective function

$$Y(\boldsymbol{A},\boldsymbol{B}) = \frac{1}{2}\sum_{\boldsymbol{A},\boldsymbol{B}} I_{ui}(r_{ui} - \sum_{k=1}^{K} a_{uk}b_{ki})^2 + \frac{\lambda_1}{2}\|\boldsymbol{A}\|^2_{\mathrm{Fro}} + \frac{\lambda_1}{2}\|\boldsymbol{B}\|^2_{\mathrm{Fro}} + \lambda_2 S(\boldsymbol{A}) + \lambda_2 S(\boldsymbol{B}), \quad (1)$$

where $\|\cdot\|_{\mathrm{Fro}}$ is the Frobenius norm, I_{ui} is the indicator function such that $I_{ui} = 1$ if (v_u, v_i) is an observed data, and $I_{ui} = 0$ otherwise. The nonnegative parameter λ_1 is often called the smoothing parameter, and the nonnegative parameter λ_2 is called the structural regulation factor.

Now we are in a position to define the gap $G(\boldsymbol{A},\boldsymbol{B})$ between the predictive value and the ground truth of S-Model. It is measured by the predictive error, i.e., the Frobenius norm of the difference between the matrix $\boldsymbol{R}$ and the matrix $\boldsymbol{AB}$, since $(\boldsymbol{AB})_{ui}$ is a prediction of the value r_{ui}. Denote the gap by

$$Gap(\boldsymbol{A},\boldsymbol{B}) = \|\boldsymbol{R} - \boldsymbol{AB}\|^2_{\mathrm{Fro}} = \sum_{v=1}^{n}\sum_{i=1}^{m}\left(r_{ui} - \sum_{k=1}^{K} a_{uk}b_{ki}\right)^2.$$

It should be noted that this gap is similar to the training error such as squared errors for linear regression. If it becomes to be 0, the model can fit the data exactly, but this may suffer from the potential over-fitting problem. This can be seen in the section of experimental results.

In the following part, we shall show the existence of the gap $G(\boldsymbol{A},\boldsymbol{B})$. Set

$$W = \sum_{u=1}^{n}\sum_{k=1}^{K}\sum_{k'=1}^{K} W_u(k,k'),$$

$$W' = \sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{m} W_k(i,j),$$

and $T_k = \{k' | S^e_{uk'} = S^e_{uk} \neq \emptyset\}$ and $T_i = \{j | S^e_{kj} = S^e_{ki} \neq \emptyset\}$. Our statement is as follows.

Theorem 1. *Assume that $Y(\boldsymbol{A},\boldsymbol{B})$ gets the local minimum. If for u and k, there exists at least one u and one i in which a_{uk} and b_{ki} satisfy:*

$$a_{uk}W\sum_{j\in T_i}(b_{ki} - b_{kj}) \neq b_{ki}W'\sum_{k'\in T_k}(a_{uk} - a_{uk'}), \quad (2)$$

then $Gap(\boldsymbol{A},\boldsymbol{B}) \neq 0$.

The idea of our proof is by contradiction. We assume that $Gap(\boldsymbol{A},\boldsymbol{B}) = 0$ and aim to prove that $Y(\boldsymbol{A},\boldsymbol{B})$ cannot reach the local minimum unless $\lambda_1 = 0$ and $\lambda_2 = 0$.

Proof. Suppose that $Gap(\boldsymbol{A},\boldsymbol{B}) = 0$. To minimize the function $Y(\boldsymbol{A},\boldsymbol{B})$, we differentiate the two sides of (1) with respect to a_{uk} and b_{ki} respectively to obtain the linear equations:

$$\frac{\partial Y}{\partial a_{uk}} = e_{ui}(-b_{ki}) + \lambda_1 a_{uk} + \lambda_2\frac{\sum\limits_{k'=1}^{K} W_u(k,k')(a_{uk} - a_{uk'})}{W} = 0, \quad (3)$$

$$\frac{\partial Y}{\partial b_{ki}} = e_{ui}(-a_{uk}) + \lambda_1 b_{ki} + \lambda_2\frac{\sum\limits_{j=1}^{m} W_k(i,j)(b_{ki} - b_{kj})}{W'} = 0, \quad (4)$$

where

$$e_{ui} := r_{ui} - \sum_{k=1}^{K} a_{uk}b_{ki},$$

for all $1 \leqslant u \leqslant n$ and $1 \leqslant i \leqslant m$. Since $Gap(\boldsymbol{A},\boldsymbol{B}) = 0$, we deduce

$$e_{ui} = 0,$$

for all $1 \leqslant u \leqslant n$ and $1 \leqslant i \leqslant m$. Plugging this equation into (3) and (4), we obtain

$$\lambda_1 a_{uk} + \lambda_2\frac{\sum\limits_{k'=1}^{K} W_u(k,k')(a_{uk} - a_{uk'})}{W} = 0, \quad (5)$$

$$\lambda_1 b_{ki} + \lambda_2 \frac{\sum_{j=1}^{m} W_k(i,j)(b_{ki} - b_{kj})}{W'} = 0. \quad (6)$$

Next we will show that (5) and (6) do not hold unless $\lambda_1 = \lambda_2 = 0$. We compute the determinants of (5) and (6) and obtain

$$\begin{vmatrix} a_{uk} & \frac{\sum_{k'=1}^{K} W_u(k,k')(a_{uk} - a_{uk'})}{W} \\ b_{ki} & \frac{\sum_{j=1}^{m} W_k(i,j)(b_{ki} - b_{kj})}{W'} \end{vmatrix} = \frac{a_{uk} W \sum_{j=1}^{m} W_k(i,j)(b_{ki} - b_{kj})}{WW'} - \frac{b_{ki} W' \sum_{k'=1}^{K} W_u(k,k')(a_{uk} - a_{uk'})}{WW'}.$$

Next we make some simplifications of the equation

$$a_{uk} W \sum_{j=1}^{m} W_k(i,j)(b_{ki} - b_{kj}) - b_{ki} W' \sum_{k'=1}^{K} W_u(k,k')(a_{uk} - a_{uk'}).$$

By definition of the structural similarity, $W_u(k,k') = 1$ when $S^e_{u,k} = S^e_{u,k'}$. By convention, we set $a_{uk} = 0$ if $S^e_{u,k} = \emptyset$. Thus, the index k' which makes contributions to the sum $\sum_{k'=1}^{K} W_u(k,k')(a_{uk} - a_{uk'})$ belongs to the set

$$T_k = \{k' | S^e_{uk'} = S^e_{uk} \neq \emptyset\}. \quad (7)$$

Similarly, the index j which makes contributions to the sum $\sum_{j=1}^{m} W_k(i,i')(b_{ki} - b_{ki'})$ belongs to the set

$$T_i = \{j | S^e_{kj} = S^e_{ki} \neq \emptyset\}. \quad (8)$$

Under this symbol, we find that

$$\begin{aligned} & a_{uk} W \sum_{j=1}^{m} W_k(i,j)(b_{ki} - b_{kj}) - b_{ki} W' \sum_{k'=1}^{K} W_u(k,k')(a_{uk} - a_{uk'}) \\ = {} & a_{uk} W \sum_{j \in T_i} (b_{ki} - b_{kj}) - b_{ki} W' \sum_{k' \in T_k} (a_{uk} - a_{uk'}). \end{aligned}$$

It follows from the condition

$$a_{uk} W \sum_{j \in T_i} (b_{ki} - b_{kj}) \neq b_{ki} W' \sum_{k' \in T_k} (a_{uk} - a_{uk'})$$

that

$$\begin{vmatrix} a_{uk} & \frac{\sum_{k'=1}^{K} W_u(k,k')(a_{uk} - a_{uk'})}{W} \\ b_{ki} & \frac{\sum_{j=1}^{m} W_k(i,j)(b_{ki} - b_{kj})}{W'} \end{vmatrix} \neq 0.$$

Therefore, it follows from the knowledge of linear algebra that (5) and (6) have solutions $\lambda_1 = \lambda_2 = 0$. This completes the proof. □

To understand (2)

$$a_{uk} W \sum_{j \in T_i} (b_{ki} - b_{kj}) \neq b_{ki} W' \sum_{k' \in T_k} (a_{uk} - a_{uk'}),$$

we investigate some of its special cases and try to use them to explain what this inequality means in the real dataset. We have the following corollary.

Corollary 1. *Assume that $Y(\boldsymbol{A}, \boldsymbol{B})$ gets the local minimum and for each u and k, we have*

$$W|T_i| - W'|T_k| \geqslant 0,$$

where T_k and T_i are defined in (7) *and* (8)*, then $Gap(\boldsymbol{A}, \boldsymbol{B}) \neq 0$.*

Proof. We again proceed by contradiction. Suppose that $Gap(\boldsymbol{A}, \boldsymbol{B}) = 0$. We aim to prove that

$$a_{uk} W \sum_{j \in T_i} (b_{ki} - b_{kj}) \neq b_{ki} W' \sum_{k' \in T_k} (a_{uk} - a_{uk'}) = 0,$$

for at least one u and one i. If this inequality holds, we can use Theorem 1 to find that $Y(\boldsymbol{A}, \boldsymbol{B})$ cannot get the local minimum. By definition of $Gap(\boldsymbol{A}, \boldsymbol{B})$, this means

$$\sum_{v=1}^{n} \sum_{i=1}^{m} \left(r_{ui} - \sum_{k=1}^{K} a_{uk} b_{ki} \right)^2 = 0. \quad (9)$$

Recall that $r_{ui} = 1$ if v_u follows v_i and $r_{ui} = 0$ otherwise. If for some u and i, $r_{ui} = 0$, by (9), we obtain $\sum_{k=1}^{K} a_{uk} b_{ki} = 0$. Furthermore, for the nonnegative property of $\boldsymbol{A}$ and $\boldsymbol{B}$, we have $a_{uk} \geqslant 0$ and $b_{ki} \geqslant 0$. Hence, when $r_{ui} = 0$, we deduce $a_{uk} = b_{ki} = 0$. In this case, we have the following equation trivially

$$a_{uk} W \sum_{j \in T_i} (b_{ki} - b_{kj}) = b_{ki} W' \sum_{k' \in T_k} (a_{uk} - a_{uk'}) = 0.$$

We do not consider these kinds of u and i such that $r_{ui} = 0$. Without loss of generality, we suppose that there exist just one u and i such that $r_{ui} = 1$. We intend to prove for such u and i,

$$a_{uk} W \sum_{j \in T_i} (b_{ki} - b_{kj}) \neq b_{ki} W' \sum_{k' \in T_k} (a_{uk} - a_{uk'}).$$

Set

$$L = a_{uk}W\sum_{j\in T_i}(b_{ki}-b_{kj}) - b_{ki}W'\sum_{k'\in T_k}(a_{uk}-a_{uk'}).$$

Rewriting the above equation, we find

$$\begin{aligned}
&\sum_k\left(W\sum_{j\in T_i}a_{uk}b_{ki} - W'\sum_{k'\in T_k}a_{uk}b_{ki}\right)\\
&=\sum_k(W|T_i|-W'|T_k|)a_{uk}b_{ki}\\
&=W\sum_k\sum_{j\in T_i}a_{uk}b_{kj} - W'\sum_k\sum_{k'\in T_k}a_{uk'}b_{ki}+L\\
&=W\sum_{j\in T_i}\sum_k a_{uk}b_{kj} - W'\sum_k\sum_{k'\in T_k}a_{uk'}b_{ki}+L\\
&=-W'\sum_k\sum_{k'\in T_k}a_{uk'}b_{ki}+L.
\end{aligned}$$

The last equality holds since for other u and i, we assume $r_{uj}=\sum_k a_{uk}b_{kj}=0$. It follows from the condition $W|T_i|-W'|T_k|\geqslant 0$ that

$$\sum_k(W|T_i|-W'|T_k|)a_{uk}b_{ki}\geqslant 0.$$

We turn our attentions to the expression $-W'\sum_k\sum_{k'\in T_k}a_{uk'}b_{ki}+L$. Since $r_{ui}=1$, we conclude that $b_{ki}\neq 0$ for all k and i. Meanwhile, the condition $k'\in T_k$ guarantees that $a_{uk'}\neq 0$. Hence, if $L=0$, we have

$$-W'\sum_k\sum_{k'\in T_k}a_{uk'}b_{ki}+L<0,$$

a contradiction. Thus, we find that $L\neq 0$ for such u and i. This completes the proof. □

We illustrate the condition $W|T_i|-W'|T_k|\geqslant 0$ in the network of Twitter. In Fig.3, we can compute $W=3\times 2=6$, $|T_k|=2$, $W'=2$ and $|T_i|=1$ such that $W|T_i|-W'|T_k|=2>0$. In this case, if $Y(\boldsymbol{A},\boldsymbol{B})$ gets its minimum, from the corollary, the gap $Gap(\boldsymbol{A},\boldsymbol{B})\neq 0$.

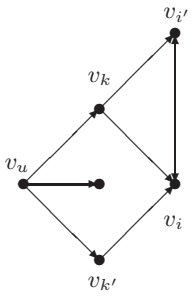


Fig.3. Example of subgraph satisfying $W|T_i|-W'|T_k|\geqslant 0$.

To narrow the gap $G(\boldsymbol{A},\boldsymbol{B})$, one possible way is to define a similarity measure between the user v_k and the user $v_{k'}$ with respect to v_i based on their interactions with v_i respectively. For instance, we can record the number of tweets of v_i that v_k and $v_{k'}$ retweet respectively. If v_k and $v_{k'}$ retweet the same number of tweets of v_i, we regard them as similar users. In Subsection 5.1, we shall define this interaction similarity as retweeting similarity.

5 SI-Model for Narrowing the Gap

In this section, to narrow the gap $G(\boldsymbol{A},\boldsymbol{B})$, we shall propose a nonnegative matrix factorization based model, called SI-Model to unify the structure of the network and the interactions between users. Given the source users v_u, the intermediate user set V_k and the target user set V_i, we firstly define the retweeting similarity between two intermediate users $v_k\in V_k$ and $v_{k'}\in V_k$ based on their interactions with one target user $v_i\in V_i$. Then we define an objective function $F(\boldsymbol{A},\boldsymbol{B})$ in connection with the retweeting similarity. Our SI-Model is devoted to minimizing the function $F(\boldsymbol{A},\boldsymbol{B})$.

5.1 Retweeting Similarity

Firstly, we define a similarity measure to characterize the similarity between two intermediate users $v_k\in V_k$ and $v_{k'}\in V_k$ based on their interactions with one target user $v_i\in V_i$ respectively in the time interval $(t_0,t_1]$, denoted by $R_i(k,k')$. The interactions are referred to as the retweeting behaviors. Suppose that v_i posted a list of m tweets $\{tw_1,tw_2,\ldots,tw_m\}$ in the time interval $(t_0,t_1]$. There are two ways to define $R_i(k,k')$. The first one is to compare the number of retweets of v_k and $v_{k'}$. Assume that v_k retweets n_k tweets of v_i and $v_{k'}$ retweets $n_{k'}$ tweets of v_i. Then $R_i(k,k')$ can be defined in a binary way: $R_i(k,k')=1$ if $n_k=n_{k'}$ and $R_i(k,k')=0$ otherwise. On the other hand, we can obtain a refined vector to record whether v_k retweets each of the tweets of v_i as $\boldsymbol{r}_k=(r_{k1},r_{k2},\ldots,r_{km})$, where $r_{ki}=0$ if v_k does not retweet the i-th tweet, and $r_{ki}=1$ otherwise. Similarly, we can get the refined vector for $v_{k'}$ as $\boldsymbol{r}_{k'}=(r_{k'1},r_{k'2},\ldots,r_{k'm})$. Next the retweeting similarity $R_i(k,k')$ can be defined as the cosine similarity of the two vectors $\boldsymbol{r}_k$ and $\boldsymbol{r}_{k'}$.

$$R_i(k,k')=\cos(\boldsymbol{r}_k,\boldsymbol{r}_{k'})=\frac{\boldsymbol{r}_k\cdot\boldsymbol{r}_{k'}}{\|\boldsymbol{r}_k\|\cdot\|\boldsymbol{r}_{k'}\|}.$$

5.2 Interaction Regulation

In this part, we shall propose SI-Model which unifies the structural information and retweet knowledge into a framework. We introduce an interaction regulation factor $R(\boldsymbol{A})$ defined in connection with the retweeting similarity as follows.

$$R(\boldsymbol{A}) = \frac{\sum\limits_{u=1}^{n}\sum\limits_{k=1}^{K}\sum\limits_{k'=1}^{K} R_i(k,k')(a_{uk}-a_{uk'})^2}{\sum\limits_{u=1}^{n}\sum\limits_{k=1}^{K}\sum\limits_{k'=1}^{K} R_i(k,k')},$$

where $R_i(k,k')$ is the retweeting similarity of the intermediate users v_k and $v_{k'}$ with respect to the target user v_i. SI-Model aims to minimize the following objective function

$$\begin{aligned} F(\boldsymbol{A},\boldsymbol{B}) = {} & \frac{1}{2}\sum_{\boldsymbol{A},\boldsymbol{B}} I_{u,i}(R_{u,i}-\sum_{k=1}^{K} a_{uk}b_{ki})^2 + \\ & \frac{\lambda_1}{2}\|\boldsymbol{A}\|_{\mathrm{Fro}}^2 + \frac{\lambda_1}{2}\|\boldsymbol{B}\|_{\mathrm{Fro}}^2 + \lambda_2 S(\boldsymbol{A}) + \\ & \lambda_2 S(\boldsymbol{B}) + \lambda_3 R(\boldsymbol{A}), \end{aligned} \tag{10}$$

where λ_3 is a nonnegative parameter called the interaction regulation parameter. Remark that except for $R(\boldsymbol{A})$, we do not add another interaction regulation $R(\boldsymbol{B})$ which has a similar definition of $R(\boldsymbol{A})$, like the pair of $S(\boldsymbol{A})$ and $S(\boldsymbol{B})$. This is because $R(\boldsymbol{A})$ has already modeled the different contributions of the intermediate users when they recommended user v_i to user v_u from the interaction aspect. Whether two of the intermediate users interacted with v_i has been considered in the term $R(\boldsymbol{A})$, and their different interaction behaviors determine whether v_u will follow v_i. Therefore, it is enough to define $R(\boldsymbol{A})$.

To minimize the function $F(\boldsymbol{A},\boldsymbol{B})$, it is easily seen that the objective function $F(\boldsymbol{A},\boldsymbol{B})$ is not convex in either $\boldsymbol{A}$ or $\boldsymbol{B}$, and thus it is unrealistic to obtain the global minimum of $F(\boldsymbol{A},\boldsymbol{B})$. However, there are many techniques from numerical optimization that can be applied to find the local minimum of $F(\boldsymbol{A},\boldsymbol{B})$. Gradient descent is perhaps the simplest technique to implement, but the convergence can be slow. We follow the multiplicative update rule by Lee and Seung[25] and provide the similar update rule for the elements of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

The reader may find that the difference of the objective functions $Y(\boldsymbol{A},\boldsymbol{B})$ and $F(\boldsymbol{A},\boldsymbol{B})$ is the addition of the interaction regulation factor $\lambda_3 R(\boldsymbol{B})$. A natural question is what is the essential difference of our SI-Model with others, such as that established by Yin *et al.*[11] We point out that the addition of the interaction regulation factor $R(\boldsymbol{A})$ is important not only because it considers more information than other models, but also because adding this new factor, we can reduce the gap $G(\boldsymbol{A},\boldsymbol{B})$ between the predictive value and the ground truth if we choose the appropriate value of λ_3. The following proposition is devoted to the choices of λ_2 and λ_3, given the value of λ_1.

Proposition 1. *If $F(\boldsymbol{A},\boldsymbol{B})$ obtains the minimum for the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, and $Gap(\boldsymbol{A},\boldsymbol{B})=0$. Then for a given factor $\lambda_1 \geqslant 0$, we have*

$$\lambda_2 = \frac{-\lambda_1 b_{ki} W'}{\sum\limits_{j=1}^{m} W_k(i,j)(b_{ki}-b_{kj})},$$

$$\lambda_3 = \frac{-\lambda_1 (a_{uk}-b_{ki})R}{\sum\limits_{k'=1}^{K} R_i(k,k')(a_{uk}-a_{uk'})},$$

where

$$R = \sum_{i=1}^{m}\sum_{k=1}^{K}\sum_{k'=1}^{K} R_i(k,k').$$

Proof. Suppose that $Gap(\boldsymbol{A},\boldsymbol{B})=0$. To minimize the function $F(\boldsymbol{A},\boldsymbol{B})$, we differentiate both sides of (10) with respect to a_{uk} and b_{ki} respectively and obtain the linear equations:

$$\begin{aligned} \frac{\partial F}{\partial a_{uk}} = {} & e_{ui}(-b_{ki}) + \lambda_1 a_{uk} + \\ & \lambda_2 \frac{\sum\limits_{k'=1}^{K} W_u(k,k')(a_{uk}-a_{uk'})}{W} + \\ & \lambda_3 \frac{\sum\limits_{k'=1}^{K} R_i(k,k')(a_{uk}-a_{uk'})}{R} = 0, \end{aligned} \tag{11}$$

$$\begin{aligned} \frac{\partial F}{\partial b_{ki}} = {} & e_{ui}(-a_{uk}) + \lambda_1 b_{ki} + \\ & \lambda_2 \frac{\sum\limits_{j=1}^{m} W_k(i,j)(b_{ki}-b_{kj})}{W'} \\ = {} & 0, \end{aligned} \tag{12}$$

where

$$e_{ui} = r_{ui} - \sum_{k=1}^{K} a_{uk} b_{ki},$$

for all $1 \leqslant u \leqslant n$ and $1 \leqslant i \leqslant m$ and

$$R = \sum_{u=1}^{n}\sum_{k=1}^{K}\sum_{k'=1}^{K} R_i(k,k').$$

Since $Gap(\boldsymbol{A},\boldsymbol{B})=0$, we deduce

$$e_{ui} = 0,$$

for all $1 \leqslant u \leqslant n$ and $1 \leqslant i \leqslant m$. Plugging this equation into (11) and (12), we obtain

$$\lambda_1 a_{uk} + \lambda_2 \frac{\sum_{k'=1}^{K} W_u(k,k')(a_{uk} - a_{uk'})}{W} + \lambda_3 \frac{\sum_{k'=1}^{K} R_i(k,k')(a_{uk} - a_{uk'})}{R} = 0,$$

$$\lambda_1 b_{ki} + \lambda_2 \frac{\sum_{j=1}^{m} W_k(i,j)(b_{ki} - b_{kj})}{W'} = 0.$$

Suppose that λ_1 is a free variable, then a routine calculation leads to the solutions for the parameters λ_2 and λ_3 as stated in the proposition, which completes the proof. □

So far, we have known how to choose the parameters λ_2 and λ_3 through the parameter λ_1 under the condition that the gap $G(\boldsymbol{A}, \boldsymbol{B})$ is equal to zero. Although the analysis is theoretical, it provides the evidence that when adding the interaction regulation factor $\lambda_3 R(\boldsymbol{A})$ into the objective function where λ_3 is determined by Proposition 1, we can indeed reduce $Gap(\boldsymbol{A}, \boldsymbol{B})$ to better approximate the value zero.

5.3 Personalized SI-Model

SI-Model is a global model used for the prediction task for all users in the Twittersphere. However, many users, due to some reason, would not like the system to recommend new users for them. In other words, their friend lists turn out to be stable. If we directly reduce SI-Model to fit the local structures of these kinds of users, the model faces the problem of over-fitting. Therefore, we should construct a personalized SI-Model for the users who are really willing to follow new users. This local model is more efficient than SI-Model for link prediction for the specific source user v_u, and can overcome the problem of over-fitting. In this sense, the matrices $\boldsymbol{A}$ and $\boldsymbol{R}$ reduce to the row vectors $\boldsymbol{A}_u$ and $\boldsymbol{R}_u$. The task is to predict new followers of v_u. Notice that Yin *et al.*[11] also proposed a local S-Model for prediction in this case, whereas they called the prediction in ego-centric networks. They introduced a measure β_{v_u,v_i,v_k} to approximately characterize the probability of the the intermediate user v_k to recommend the target user v_i to the source user v_u, given the local structure of v_k. More precisely, β_{v_u,v_i,v_k} is defined as follows:

$$\beta_{v_u,v_i,v_k} = \frac{\sum_{v_{k'} \in N_{v_u}} W_k(v_i, v_{k'})}{\sum_{v_{k'} \in V} W_k(v_i, v_{k'})},$$

where N_{v_u} is the set of v_u's friends. In fact, β_{v_u,v_i,v_k} calculates the number of v_u's friends who share similar structures with v_i divided by the number of all users who share similar structures with v_i. The larger the value β_{v_u,v_i,v_k} is, the more likely v_u will follow v_i through the recommendation of v_k. Yin *et al.*[11] considered the extreme case when only one target user v_u hopes the system to recommend new friends, and proposed the personalized S-Model which aims to minimize the following objective function $Y_{v_u}(\boldsymbol{A}, \boldsymbol{B})$,

$$Y_{v_u}(\boldsymbol{A}, \boldsymbol{B}) = \frac{1}{2} \sum_{\boldsymbol{A},\boldsymbol{B}} I_{u,i}(R_{u,i} - \sum_{k=1}^{K} a_{uk} b_{ki})^2 + \frac{\lambda_1}{2} \|\boldsymbol{A}_{v_u}\|_{\mathrm{Fro}}^2 + \frac{\lambda_1}{2} \|\boldsymbol{B} - \boldsymbol{\beta}_{v_u}\|_{\mathrm{Fro}}^2 + \lambda_2 S(\boldsymbol{A}) + \lambda_2 S(\boldsymbol{B}),$$

where $\boldsymbol{\beta}_{v_u} = (\beta_{v_u,v_i,v_k})_{K \times m}$ is the matrix with the rows corresponding to the intermediate users, and the columns corresponding to the target users.

Similar to the argument of the difference between S-Model and SI-Model, the local S-Model only considers the structure information in the link prediction process, ignoring the indispensable impact of the interaction between the intermediate user v_k and the source user v_i. To compensate this insufficiency, we define the following measure γ_{v_u,v_i,v_k},

$$\gamma_{v_u,v_i,v_k} = \frac{\sum_{v_{k'} \in N_{v_u}} R_i(v_k, v_{k'})}{\sum_{v_{k'} \in V} R_i(v_i, v_{k'})},$$

where N_{v_u} is defined as before. Actually, γ_{v_u,v_i,v_k} calculates the number of v_u's friends who retweet v_i divided by the number of all users who retweet v_i. Similarly, the larger the value γ_{v_u,v_i,v_k} is, the more likely v_u will follow v_i, since the fact that more and more v_u's friends retweet v_i increases the probability that user v_u becomes aware of v_i and v_i appears frequently in the tweets of v_u's friends. In other words, the set of such intermediate users v_k provides enough social proof (see Cialdini[26]) that it is beneficial for v_u to follow v_i.

Then we introduce the personalized SI-Model which aims to minimize the following objective function

$$F_{v_u}(\boldsymbol{A}, \boldsymbol{B}) = \frac{1}{2} \sum_{\boldsymbol{A},\boldsymbol{B}} I_{u,i} \left(R_{u,i} - \sum_{k=1}^{K} a_{uk} b_{ki} \right)^2 + \frac{\lambda_1}{2} \|\boldsymbol{A}_{v_u}\|_{\mathrm{Fro}}^2 + \frac{\lambda_1}{2} \|\boldsymbol{B} - \boldsymbol{\beta}_{v_u} - \boldsymbol{\gamma}_{v_u}\|_{\mathrm{Fro}}^2 + \lambda_2 S(\boldsymbol{A}) + \lambda_2 S(\boldsymbol{B}) + \lambda_3 R(\boldsymbol{A}),$$

where $\boldsymbol{\beta}_{v_u} = (\beta_{v_u,v_i,v_k})_{K\times m}$ is defined as before, $\boldsymbol{\gamma}_{v_u} = (\gamma_{v_u,v_i,v_k})_{K\times m}$ is the matrix with the rows corresponding to the intermediate users, and the columns corresponding to the target users.

6 Experimental Results and Analysis

In this section, we describe the experimental results. We present the detailed information of our dataset, and the experimental results in comparison with other methods. We find that our SI-Model can reduce the RMSE value by about 70% compared with S-Model.

The dataset in the experiments is crawled by Twitter API in the way of randomly selecting 10 000 Twitter users, updating their immediate neighbors per day from the period of Oct. 1st, 2012 to Nov. 19th, 2012. This leads to the user networks. Meanwhile, we also extract the tweets of these users per day and use them to construct the retweeting network, where user A has relations with user B if A's tweet contains the syntax $@B$ or $RT@B$, or equivalently, A retweets B or mentions B in his or her tweets. In total, there are 140 000 users and 400 000 000 tweets. In this part, we carry out two types of settings for evaluation similar to the work of [27]. We firstly carry out the experiments in dynamic setting, where we randomly choose 1 000 pairs of snapshots with one week interval and for each pair, and we use the first snapshot network to predict the following links in the second snapshot network. The interval between these two snapshots is one week. Note that the interval can be chosen differently, for instance, two weeks, three weeks and so on. For example, we use the snapshot network on Nov. 08, 2012 to predict new following links at the snapshot network on Nov. 13, 2012. During these two snapshots, the number of new followers of the users is added up to 8 090. To evaluate the final performance for these 1 000 snapshot pairs, we compute their average performance. Secondly, we work out the experiment in sparse static setting to test the performance of SI-Model on sparse data, where we use the sparse network by randomly deleting the links of the snapshot network on Nov. 08, 2012 such that the average number of followers for each user is 2. This evaluation is widely used in link prediction problem, for example, see [11, 28]. To use SI-Model for link prediction, we remove three followers for each user of the sparse network and intend to find these missing links.

Our model runs on the matrix $\boldsymbol{R}$ with the number of rows $n = 10\,000$, and the number of column $m = 10\,000$. Two evaluation criteria for the predictive result are used. One is the root mean square error (RMSE) defined as

$$RMSE = \sqrt{\frac{\sum_{u=1}^{n}\sum_{i=1}^{m}(r_{ui} - \hat{r}_{ui})^2}{mn}},$$

where $\hat{r}_{ui} = \sum_{k=1}^{K} a_{uk}b_{ki}$. RMSE measures the error between the predictive value and the ground truth. The smaller the RMSE value is, the better the predictive result will be. In other words, the predictive result is more accurate. The other criterion is the $F1$-measure based on the break-even point. These two types of measures are widely used in link prediction problem, for example, see [11, 28].

To evaluate the performance of SI-Model, we tune the parameters λ_1, λ_2 and λ_3 in the full grid, where λ_1 ranges from 0 to infinity, and the other two parameters are determined by Proposition 1. We find that when $\lambda_1 = 0.01$, $\lambda_2 = 0.001$, and $\lambda_3 = 0.005$, the optimal RMSE value of SI-Model is equal to 0.033. Meanwhile, after a full search, we find that when $\lambda_1 = 0.01$ and $\lambda_2 = 0.01$, the optimal RMSE value of S-Model is equal to 0.102. It can be seen that SI-Model obtains smaller RMSE value than S-Model.

6.1 Experimental Results

To demonstrate the effectiveness of our method, we compare it with the other three methods in recent years for link prediction in Twitter. One is S-Model proposed by Yin *et al.*[11] Another is the Jaccard coefficient based unsupervised method, which has been proven by Yin *et al.*[11] to have a good performance on the dynamic link prediction setting. The last one is the common neighbors method. These two measures are commonly used in link prediction task. For other similar measures, readers can refer to [29]. Note that for the Jaccard coefficient and the common neighbors methods, there is no RMSE by definition. In the dynamic setting, the results are listed as in Table 1.

Table 1. Comparison of Different Methods

Method	RMSE	$F1$-Measure
SI-Model	0.033	0.278
S-Model	0.102	0.252
Jaccard coefficent	–	0.125
Common neighbors	–	0.091

From Table 1, we see that SI-Model achieves a smaller RMSE value and a bigger $F1$-measure than any

of the other three models. Especially, the RMSE value is reduced by about 0.07 compared with that obtained by S-Model. Note that S-Model obtains the RMSE value 0.102. If we want to reduce this value, the maximal reduction is 0.102 (the corresponding RMSE value is 0). Our SI-Model obtains the reduction of 0.07. In other words, we get 70% reduction by using SI-Model. On the other hand, our SI-Model gets the $F1$-measure 0.278, with the increase 0.02 compared to the S-Model. Furthermore, Table 1 is devoted to the average performance. For detailed comparison, we also conduct the experiment. For instance, as for the $F1$-measure, we illustrate the $F1$-measure of S-Model and SI-Model for 50 different snapshot pairs.

From Fig.4, we see that SI-Model performs better than S-Model for 72% snapshot pairs in which the second column is higher than the first one. Especially, when $t = 8$, SI-Model gets 0.124 improvement. For the rest 28% snapshot pairs, we find SI-Model is not better because in these pairs, the retweeting behavior between users does not correlate so much with the link formation process. Meanwhile, in this situation, although SI-Model still uses the structure information to predict like S-Model, the interaction information reduces the performance instead. In this sense, the interaction information seems to bring in some overfitting in the prediction process.

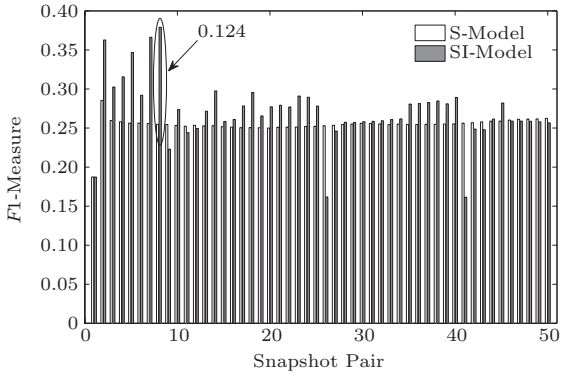


Fig.4. $F1$-measure of S-Model and SI-Model for different snapshot pairs.

Next, we shall examine the performance of SI-Model in the sparse static setting. It should be noted that Table 1 is obtained on a dense dataset, in which the density of the network is equal to 491. This kind of dataset is usually used for link prediction in Twitter. For example, Yin *et al.*[11] tested S-Model with the dataset density 100. If the dataset is very sparse, for example, we use the sparse dataset mentioned in the beginning of this subsection, with the density only 2, we intend to examine the performance of SI-Model. The results are listed in Table 2. From Table 2, we can see that our SI-Model is also better than the others. Especially, the RMSE value is reduced by about 0.005 compared with that obtained by S-Model. Note that S-Model obtains the RMSE value 0.013. Compared to the reduction 0.005, SI-Model reduces about 38% RMSE value. On the other hand, SI-Model also increases the $F1$-measure. For detailed comparison, we also conduct the experiment. For instance, as for the RMSE value, we illustrate the RMSE value of S-Model and SI-Model for 50 different snapshot pairs.

Table 2. Comparison on Sparse Data

Method	RMSE	$F1$-Measure
SI-Model	0.008	6.4×10^{-5}
S-Model	0.013	6.0×10^{-5}
Jaccard coefficent	–	5.1×10^{-5}
Common neighbors	–	3.6×10^{-5}

From Fig.5, we see that SI-Model performs better than S-Model for 86% snapshot pairs in which the second column is lower than the first one. Especially, when $t = 20$, SI-Model gets a 0.005 8 decrease. Similarly, for the rest 14% snapshot pairs, we find SI-Model is not better because in these pairs, the retweeting behavior between users does not correlate so much with the link formation process. Meanwhile, in this situation, although SI-Model still uses the structure information to predict like S-Model, the interaction information reduces the performance instead. In this sense, the interaction information seems to bring in some overfitting in the prediction process.

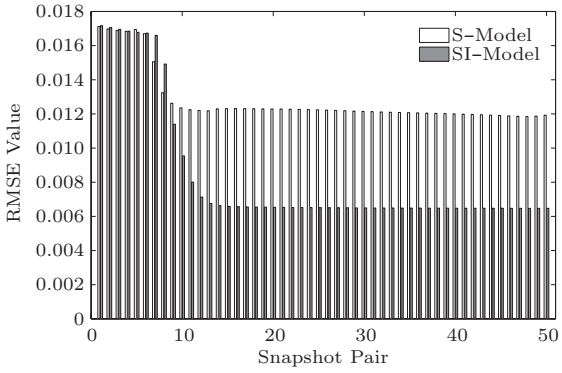


Fig.5. RMSE of S-Model and SI-Model for different snapshot pairs.

Finally, to check the performance of the personalized SI-Model, we also tune the parameters λ_1, λ_2

and λ_3 in the full grid, where λ_1 ranges from 0 to infinity, and the other two parameters are determined by Proposition 1. When $\lambda_1 = 0.01$, $\lambda_2 = 0.001$, and $\lambda_3 = 0.01$, the optimal $F1$-measure of the personalized SI-Model is equal to 8.77×10^{-5}. It can be seen that the personalized S-Model achieves larger $F1$-measure than S-Model, and the personalized SI-Model achieves larger $F1$-measure than SI-Model, which obtains the $F1$-measure 6.4×10^{-5} as shown in Table 2.

6.2 Influence of Dimension K and Sparseness of Data on the Performance of SI-Model

In this subsection, we shall analyze two factors that may influence the predictive performance of SI-Model. One is the choice of the latent dimension K during the matrix factorization process. The other is the dataset with different sparse degrees. In other words, we try to find the "best" dataset with an appropriate sparse degree on which our SI-Model achieves the best performance.

Let us first check whether the choice of the value of latent dimension K leads to different predictive performance. Recall that we factorize the matrix $\boldsymbol{R}_{n\times m}$ into two nonnegative matrices $\boldsymbol{A}_{n\times K}$ and $\boldsymbol{B}_{K\times m}$ in (10). To investigate the effect of the dimension K, we fix the parameters λ_1, λ_2 and λ_3 as $0.01, 0.001$ and 0.005 respectively, use the dataset as in Table 1, and range K from 1 to the number of users n (in fact, by the definition of the matrix factorization, $K \ll n$). The relations between the $F1$-measure and the value K, and the relations between the RMSE value and the value K are examined. We only depict the relations for $K = 1, \ldots, 18$ as follows, since for the rest part, the tendency of the curves is similar.

From Fig.6, we see that the values of RMSE and $F1$-measures vary with the increase of K. Especially, when $K = 2$, $F1$-measure obtains the maximum 0.278. This value is consistent with the $F1$-measure of SI-Model in Table 1. When $K \geqslant 2$, the $F1$-measure takes the value in the interval $[0.201, 0.278]$. On the other hand, for the RMSE value, when $K = 1$, it takes the minimum 0.031, and when $K \geqslant 7$, RMSE value tends to converge. To sum up, the performance of SI-Model is relatively stable as the dimension K increases.

Next, we are interested in the effect of the dataset with different sparseness. Recall that the sparseness, denoted by nf, is the average number of non-followers for a number of users. We aim to find how the performance of SI-Model varies with the dataset of different sparseness. To this end, we fix the parameters λ_1, λ_2 and λ_3 to be 0.01, 0.001, and 0.005 respectively as before, range nf from 1 to $m-1$ with $m = 10\,000$, and examine the relations between the $F1$-measure and the value of nf, and the relations between the RMSE value and the value of nf. We only depict the relation for $nf = 1, \ldots, 41$ as follows, since for the rest part, the tendency of the curves is similar.

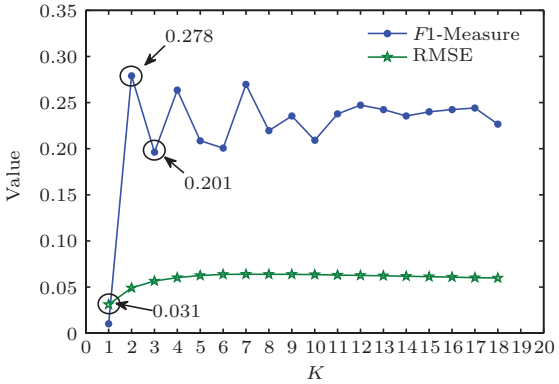


Fig.6. Relations between K and $F1$-measure on the top, and the relations between K and RMSE on the bottom for SI-Model.

From Fig.7, we see that for $nf \geqslant 3$, when $nf = 5$, the $F1$-measure achieves the maximum 0.50 and then the $F1$-measure decreases with the increase of nf. For the RMSE value, when $nf = 1$, it obtains the maximum 0.077, and then it slowly decreases. We also explore the performance of both curves when $nf = 9\,999$, and find they both turn to infinity, because in this situation, all the other users are non-followers for each user and the matrix $\boldsymbol{R}$ is 0-matrix whose elements are all equal to zero.

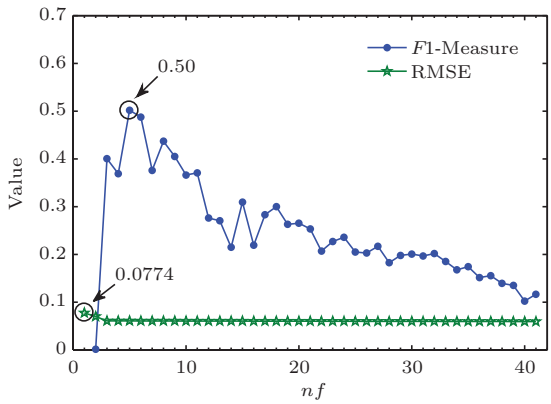


Fig.7. Relations between nf and $F1$-measure on the top, and the relations between nf and RMSE on the bottom for SI-Model.

7 Conclusions

In this paper, we proposed a structural-interaction matrix factorization model for the link prediction problem in microblogs. We firstly claimed the drawbacks of existing matrix factorization based methods by showing the gap between the predictive value and the ground truth. Then we introduced the retweeting similarity between users and mixed the interaction regulation factor with the structural regulation factor to obtain a hybrid model, SI-Model. To verify the efficiency of SI-Model, we made the experiments on real-world Twitter data and got its RMSE value reduced by 70% compared with that obtained by the state-of-the-art model. It should be noted that our SI-Model is a global model on the complete network. We similarly deduced a local, or personalized SI-Model which aims to recommend the target users to a given source user who has the requirement of following new users. It can be also experimentally verified that the local model's performance is competitive. The importance of our SI-Model is that we presented a framework to unify the structure of the network and the interaction behavior information between users together for the link prediction problem. Our model can also be applied to the other hybrid social-information networks except for the microblogs Twitter.

Microblogs provide an information platform for users to share their experience, ideas, etc, based on which people interact with each other for socialization and information diffusion. Our SI-Model actually reveals a part of this relationship between the information diffusion (including the interaction behavior) and the link formation process. However, there is still a long way to go if one aims to investigate this relationship more deeply. For example, one such problem is how and to what extent the interaction behaviors between users influence their link formation process. Furthermore, as is known to all, the content of the tweet of one target user is an essential factor to determine whether the intermediate user retweets him or her. We are all familiar with the fact that if one tweet is very positive, its diffusion process makes many people connected and they are prone to becoming friends. Hence, proposing a model fusing the content of the tweets, the interaction behavior and the structural information among users, is probably one promising area for future link prediction work. Notice that recently, Liu *et al.*[30] proposed HYDRA for user identity linkage across different social platforms, which integrates the behavior similarity among online users with multi-dimensional similarity vectors and users' core social network structure by means of a multi-objective optimization approach. This integration approach is helpful for readers for future integration strategy design in link prediction. Other important studies include [31-36]. They presented deep insight of the link formation process and the construction of an end-to-end link prediction pipeline, by integrating the knowledge of game theory, knowledge acquisition and reasoning into the predictive framework.

References

[1] Wasserman S, Faust K. Social Network Analysis: Methods and Applications. Cambridge University Press, Nov. 1994.

[2] Gu Q, Zhou J, Ding C. Collaborative filtering: Weighted nonnegative matrix factorization incorporating user and item graphs. In *Proc. SDM*, April 29-May 1, 2010, pp.199-210.

[3] Backstrom L, Leskovec J. Supervised random walks: Predicting and recommending links in social networks. In *Proc. the 4th WSDM*, Feb. 2011, pp.635-644.

[4] Leskovec J, Huttenlocher D, Kleinberg J. Predicting positive and negative links in online social networks. In *Proc. the 19th WWW*, Apr. 2010, pp.641-650.

[5] Adamic L, Adar E. Friends and neighbors on the web. *Social Networks*, 2003, 25(3): 211–230.

[6] Newman M E. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 2001, 64(2): Article No. 025102.

[7] Katz L. A new status index derived from sociometric analysis. *Psychometrika*, 1953, 18(1): 39-43.

[8] Sadilek A, Kautz H, Bigham J P. Finding your friends and following them to where you are. In *Proc. the 5th WSDM*, Feb. 2012, pp.723-732.

[9] Hopcroft J, Lou T, Tang J. Who will follow you back? Reciprocity relationship prediction. In *Proc. the 20th CIKM*, Oct. 2011, pp.1137-1146.

[10] Lou T, Tang J, Hopcroft J, Fang Z, Ding X. Learning to predict reciprocity and triadic closure in social networks. *ACM Transactions on Knowledge Discovery from Data*, 2013, 7(2): 5:1-5:25.

[11] Yin D, Hong L, Davison B D. Structural link analysis and prediction in microblogs. In *Proc. the 20th CIKM*, Oct. 2011, pp.1163-1168.

[12] Lichtenwalter R, Lussier J, Chawla N. New perspectives and methods in link prediction. In *Proc. the 16th KDD*, Jul. 2010, pp.243-252.

[13] Akasaka R, Grafe P, Kondo M. 'Me Too' 2.0: An analysis of viral retweets on the Twittersphere, 2010. http://snap.stanford.edu/class/cs224w-2010/proj2010/13_projectFinal.pdf, May 2015.

[14] Kwak H, Lee C, Park H, Moon S. What is Twitter, a social network or a news media? In *Proc. the 19th WWW*, Apr. 2010, pp.591-600.

[15] Koren Y. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proc. the 14th KDD*, Aug. 2008, pp.426-434.

[16] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender system. *Computer*, 2009, 42(8): 30-37.

[17] Zhou T, Shan H, Banerjee A, Sapiroy G. Kernelized probabilistic matrix factorization: Exploiting graphs and side information. In *Proc. the 12th SDM*, Aug. 2012, pp.403-414.

[18] Menon A K, Elkan C. Link prediction via matrix factorization. In *Proc. ECML PKDD*, Sept. 2011, pp.437-452.

[19] Zhang J, Wang C, Wang J, Yu P S. LaFT-tree: Perceiving the expansion trace of one's circle of friends in online social networks. In *Proc. WSDM*, Feb. 2013, pp.597-606.

[20] Gao S, Denoyer L, Gallinari P. Temporal link prediction by integrating content and structure information. In *Proc. the 20th CIKM*, Oct. 2011, pp.1169-1174.

[21] Acar E, Dunlavy D, Kolda T. Link prediction on evolving data using matrix and tensor factorizations. In *Proc. ICDM Workshops*, Dec. 2009, pp.262–269.

[22] Spiegel S, Clausen J, Albayrak S, Kunegis J. Link prediction on evolving data using tensor factorization. In *Proc. PAKDD Workshops*, May 2011, pp.100-110.

[23] Romero D M, Kleinberg J. The directed closure process in hybrid social-information networks, with an analysis of link formation on Twitter. In *Proc. the 4th ICWSM*, May 2010.

[24] Salakhutdinov R, Mnih A. Probabilistic matrix factorization. In *Proc. the 21st NIPS*, Dec. 2007, pp.1257-1264.

[25] Lee D, Seung H S. Algorithms for non-negative matrix factorization. In *Proc. NIPS*, Nov. 2000, pp.556-562.

[26] Cialdini R B. Influence: Science and Practice. Allyn and Bacon/Pearson, 2001.

[27] Qu Q, Liu S, Jensen C S, Zhu F, Faloutsos C. Interestingness-driven diffusion process summarization in dynamic networks. In *Proc. ECML PKDD*, *Part II*, Sept. 2014, pp.597-613.

[28] Yin Z, Gupta M, Weninger T, Han J. LINKREC: A unified framework for link recommendation with user attributes and graph structure. In *Proc. the 19th WWW*, Apr. 2010, pp.1211-1212.

[29] Liben-Nowell D, Kleinberg J. The link prediction problem for social networks. In *Proc. CIKM*, Nov. 2003, pp.556-559.

[30] Liu S, Wang S, Zhu F, Zhang J, Krishnan R. HYDRA: Large-scale social identity linkage via heterogeneous behavior modeling. In *Proc. SIGMOD*, June 2014, pp.51-62.

[31] Jia Y, Wang Y, Li J, Feng K, Cheng X, Li J. Structural-interaction link prediction in Microblogs. In *Proc. the 22nd WWW*, May 2013, pp.193-194.

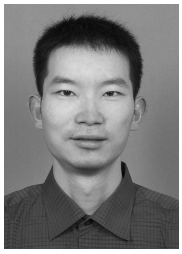
[32] Liu D, Wang Y, Jia Y, Li J, Yu Z. From strangers to neighbors: Link prediction in microblogs using social distance game. In *Proc. WSDM*, Feb. 2014.

[33] Liu D, Wang Y, Jia Y *et al.* LSDH: A hashing approach for large-scale link prediction in microblogs. In *Proc. the 28th AAAI*, July 2014, pp.3120-3121.

[34] Zhao Z, Jia Y, Wang Y, Cheng X. Content-structural relation inference in knowledge base. In *Proc. the 28th AAAI*, July 2014, pp.3154-3155.

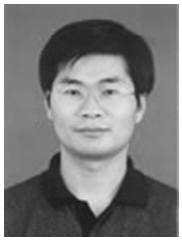
[35] Jia Y, Wang Y, Cheng X, Jin X, Guo J. OpenKN: An open knowledge computational engine for network big data. In *Proc. ASONAM*, Aug. 2014, pp.657-664.

[36] Jia Y, Wang Y, Jin X, Cheng X. TSBM: The temporal-spatial Bayesian model for location prediction in social networks. In *Proc. WI-IAT*, Aug. 2014, pp.194-201.



Yan-Tao Jia is an assistant professor at Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing. He received his Ph.D. degree in mathematics from Nankai University, Tianjin, in 2012. His main research interests include open knowledge network, social computing, and combinatorial algorithms.

Yuan-Zhuo Wang is an associate professor at Institute of Computing Technology, Chinese Academy of Sciences, Beijing. He received his Ph.D. degree in computer science from Tsinghua University, Beijing, in 2008. His current research interests include social computing, open knowledge network, etc.



Xue-Qi Cheng is a professor at Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing. He received his Ph.D. degree in computer science from ICT, CAS, in 2006. His current research interests include social computing, information security analysis, etc.