Expert Systems With Applications 108 (2018) 143–158

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Exploiting user-to-user topic inclusion degree for link prediction in social-information networks

Zhiqiang Wang, Jiye Liang*, Ru Li

Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Shanxi, Taiyuan 030006, China

ARTICLE INFO

Article history:
Received 22 October 2017
Revised 26 April 2018
Accepted 26 April 2018
Available online 28 April 2018

Keywords:
Link prediction
Fusion model
Topic inclusion degree
Network data analysis

ABSTRACT

As one kind of typical network big data, social-information networks (such as Weibo and Twitter) include both the complex network structure among users and the rich microblog/tweets information published by users. Understanding the interplay of rich content and social relationships is potentially valuable to the fundamental network mining task, i.e. the link prediction. Although some of the link prediction methods have been proposed by combining topological and non-topological information simultaneously, the in-depth analysis of the rich content still being in a minority, and the rich content in the social-information networks is still underused in solving link prediction. In this paper, we approach the link prediction problem in social-information network by combining network structure and topic information which is extracted from users' rich content. We first define a kind of user-to-user topic inclusion degree (TID) based on the dissemination mechanism of the published content in the social-information networks, and then construct a TID-based sparse network. On the basis, we build a fusion probabilistic matrix factorization model which solves the link prediction problem by fusing the information of the original following/followed network and the TID-based network in a unified probabilistic matrix factorization framework. We conduct link prediction experiments on two types of real social-information network datasets, i.e. Twitter and Weibo. The experimental results demonstrate that the proposed method is more effective in solving the link prediction problem in social-information networks.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

Network is an important organizational form of real-world data. Analyzing on network data is essential to help us explore the law of network evolution (Juszczyszyn, Musial, & Budka, 2011; Zhang, Fang, Chen, & Tang, 2015a), and understand the mechanism of complex systems (Li et al., 2015; Pastor-Satorras, Castellano, Van Mieghem, & Vespignani, 2015). Among the many tasks in network data analysis, link prediction (Getoor & Diehl, 2005) is the most fundamental one, and its solution is of great significance for many applications, such as finding like-minded friends in social networks (Aiello et al., 2012), recommending items in user-item networks (Xie et al., 2015), finding experts in academic networks (Pavlov & Ichise, 2007), and discovering unknown interactions in biological networks (Lu, Guo, & Korhonen, 2017).

It still remains a challenge in networks to predict the node-to-node relations with rich content. For instance, in social-information networks (Romero & Kleinberg, 2010; Rowe, Stankovic, & Alani, 2012) (like Twitter and Weibo) with both social and informational properties, as the name implies. Formally, a social-information network can be modeled as $G(V, E, \{T_u\}_{u \in V})$ where V denotes the set of users, E is the set of following/followed links between users, and $T_{u \in V}$ correspond to the set of published microblogs/tweets of user u. As shown in Fig. 1, where a directed network is formed when some users begin to follow others, and such structures expose the generalized social relations among people; besides the following/followed relations in the network, rich published content, i.e. many tweets published by users, are also existed. As is well-known to those who familiar with the platforms of social-information networks (such as Twitter and Weibo), the dissemination of published content is entirely dependent on the network structure, where a tweet is usually propagated from its publisher to his/her followers. However, the formation of the network structure is probably due to many complex factors. One factor goes like this: during the process of the content dissemination, if any content appeals to some users, they would like to create following links to the information publisher/mediator. Although users' interests seem to play an apparent role in producing the following/followed links, both the quantity and the exact contents of the factors that manipulate the formation of the links in the

* Corresponding author.
E-mail addresses: zhiq.wang@163.com (Z. Wang), ljy@sxu.edu.cn (J. Liang), liru@sxu.edu.cn (R. Li).

https://doi.org/10.1016/j.eswa.2018.04.034
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

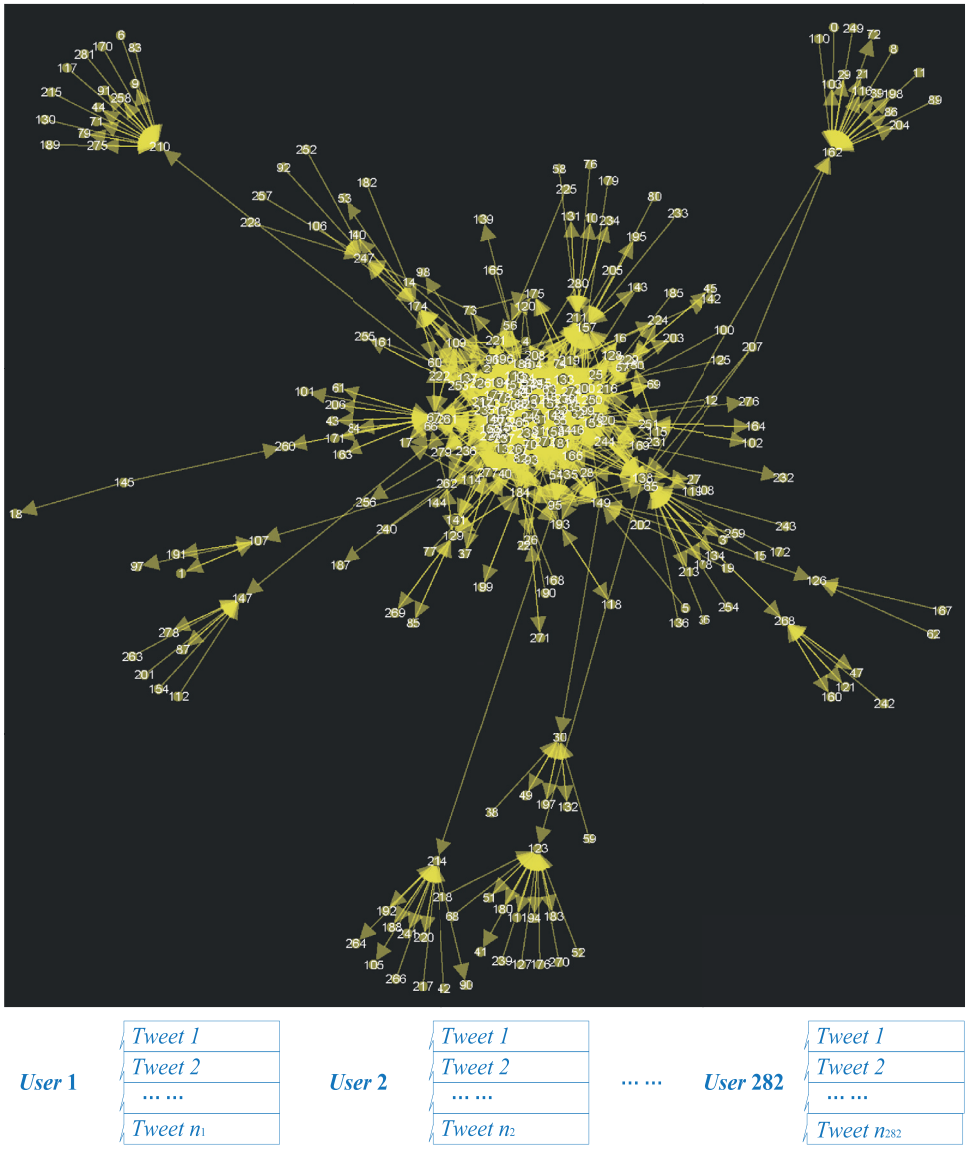


Fig. 1. An example of social-information network. The upper part of the figure shows a visual result of the Twitter network with 282 nodes (see the introduction of the Datasets in Section 6.1), where each node is denoted as a number from 1 to 282, and the node-to-node relations are denoted as directed links. The bottom of the figure is the sketch of users' published content in the social-information network.

social-information networks are still not clear. Here comes the challenge: how to build the relationships between the rich published content and the formation of the following/followed network in a social-information network. Dealing with the challenge is essential to understand the evolution of the network structure and the dissemination mechanism of the published content in social-information networks, and is certainly the key to efficiently solve the link prediction problem in this kind of network. In this paper, we on the one hand focus on effectively exploiting the rich content in the social-information networks, and on the other hand, aim to establish a fusion model which can build the relationships between the information of the following/followed network and the rich content and then to improve the link prediction performance in the social-information networks.

For link prediction, many methods have been proposed by researchers from physics, biology, sociology, and computer science, through focusing on physical networks, biological networks, social networks, and information networks (Clauset, Moore, & Newman, 2008; He, Liu, Hu, & Wang, 2015; Luo, Wu, & Li, 2017; Martnez, Berzal, & Cubero, 2016; Moradabadi & Meybodi, 2017; Rowe et al., 2012; Soares & Prudêncio, 2013; Wang, Liang, Li, & Qian, 2016). The existing metric-based methods, including neighbor-based metrics (Adamic & Adar, 2003; Ravasz, Somera, Mongru, Oltvai, & Barabási, 2002; Zhu, Lü, Zhang, & Zhou, 2012), path-based metrics (Katz, 1953; Lü, Jin, & Zhou, 2009; Papadimitriou, Symeonidis, & Manolopoulos, 2012), random walk-based metrics (Brin & Page, 1998; Fouss, Pirotte, Renders, & Saerens, 2007; Jeh & Widom, 2002; Lichtenwalter, Lussier, & Chawla, 2010) and auxiliary information-based metrics (Aiello et al., 2012; Anderson, Huttenlocher, Kleinberg, & Leskovec, 2012; Dong et al., 2012; Wang, Liao, Cao, & Qi, 2015), are taken into consideration in topological or non-topological information which can reflect users personal interests and social behaviors. Compared to the metric-based methods, the network models such as hierarchical network model (Clauset et al., 2008; Ravasz et al., 2002), stochastic block model (Airoldi, Blei, Fienberg, & Xing, 2008; Holland, Laskey, & Leinhardt, 1983; Nowicki & Snijders, 2001) and latent-feature model (Miller, Jordan, & Griffiths, 2009; Palla, Knowles, & Ghahramani, 2012; Zhu, 2012) have expanded the scope of application to a certain extent. Despite these significant advances, current state-of-the-art methods

may not be good enough for solving the following/followed link prediction problem in social-information networks. Of the existing metric-based and the learning-based methods, some methods have combined both the topological and non-topological information to solve link prediction problem. However, the in-depth analysis of the rich content in solving link prediction problem still being a minority, and the rich content is still underused in the existing link prediction methods. The depth mining and exploiting of the rich content may be great potential to improve the performance of link prediction in the social-information networks. Based on these considerations, we focus on addressing the following problems and dealing with link prediction task in social-information networks.

- How to in-depth analysis and exploit the rich content effectively in social-information networks.
- How to build a fusion model which can fuse the information of the network structure and the rich published content simultaneously and to deal with the link prediction task in social-information networks.

Concerning with these problems, this paper defines a kind of user-to-user topic inclusion degree based on the dissemination mechanism of the published content in social-information networks and constructs a topic inclusion degree-based network. On this basis, the paper builds a fusion probabilistic matrix factorization model which solves the link prediction problem by fusing the information of the original following/followed network and the topic inclusion degree-based network in a unified probabilistic matrix factorization framework. Finally, the linking probability between network nodes can be obtained based on the learning results of the model. The method provides a new way to solve the link prediction problem by fusing the two different types of semantic between users.

The rest of the paper is organized as follows. Section 2 introduces the related work, Sections 3 and 4 introduce a topic-based network construction and a fusion probabilistic matrix factorization model, respectively. Section 5 presents the link prediction algorithm based on the fusion model, and Section 6 evaluates the proposed methods with different social-information network datasets. Section 7 summarizes the whole text.

2. Related work

Research on link prediction has won increasing attention in recent years, and various link prediction methods have been proposed. Furthermore, there are also some surveys (Hasan & Zaki, 2011; Lü & Zhou, 2011; Martnez et al., 2016; Wang, Xu, Wu, & Zhou, 2014) for the link-prediction problem. The existing link prediction methods can be roughly divided into two parts, i.e. the Metric-based methods and the learning based methods.

2.1. Metric-based link prediction

A considerable part of metric-based link prediction methods is based on the topological information of networks, and it can be classified into three groups. One is to develop a neighbor-based similarity for link prediction since neighbors can indirectly reflect users' social behavior and directly affect users' social choice (Wang et al., 2014), such as Common Neighbors (Lorrain & White, 1971), Jaccard Coefficient (Jaccard, 1901), Adamic-Adar Coefficient (Adamic & Adar, 2003) and Preferential Attachment (Barabási & Albert, 1999). Another kind of similarity metrics include Local Path (Lü et al., 2009), Katz (Katz, 1953), Relation Strength Similarity (Chen, Gou, Zhang, & Giles, 2012), FriendLink (Papadimitriou et al., 2012), Vertex Collocation Profile (Papadimitriou et al., 2012) etc, and they take some local topological information or global information into consideration. Random walk-based similarity metrics is the other kind of link prediction methods, which defines the similarity between nodes by using random walk methods, such as Hitting Time (Fouss et al., 2007; Göbel & Jagers, 1974), Commute Time (Fouss et al., 2007), Rooted PageRank (Brin & Page, 1998), SimRank (Jeh & Widom, 2002; Zhang, Hu, He, Gao, & Sun, 2015b) and Blondel Index (Blondel, Gajardo, Heymans, Senellart, & Van Dooren, 2004). Compared with the neighbor-based and path-based similarities, the random walk-based methods usually have high complexity.

Besides the topological-based metrics, many non-topological-based metrics are defined by using the non-topological information in social networks (Dong et al., 2012; Wan, Lan, Guo, Fan, & Cheng, 2013; Xie, 2010). Specifically speaking, the existing non-topology-based link prediction methods are commonly dependent on similarity; if users have some similar attributes, like age, school, and interest in the social network, they will be more likely to become friends. Wang et al. (2015) defined a similarity metric based on users' topic, which is extracted by the Latent Dirichlet Allocation model (Blei, Ng, & Jordan, 2003). Aiello et al. (2012) measured the similarity between users by using the tagging information in social networks. Leroy, Cambazoglu, and Bonchi (2010) presented a new similarity measurement based on the users' group features. Also, some other non-topology-based metrics defined the similarities by using users' interests (Anderson et al., 2012), keywords (Bhattacharyya, Garg, & Wu, 2011; Rowe et al., 2012) or tags (Rowe et al., 2012). Although the existing metrics consider topological or non-topological information which can reflect users personal interests and social behaviors, the effectiveness of the metrics depends on the domain, the specific network, and the available information.

2.2. Learning-based link prediction

A branch of learning methods are based on the classification models, where the link prediction task can be considered as a binary classification one. In a classification-based link prediction model, the features are defined on each pair of nodes, and they can be constructed in topological or non-topological. The topological features (such as the neighbors-based metrics and the path-based features) are the commonly used features in a classification-based link prediction model (Chiang, Natarajan, Tewari, & Dhillon, 2011; De Sá & Prudêncio, 2011; Leskovec, Huttenlocher, & Kleinberg, 2010). Besides the topological features, the non-topological features (such as users' location, interests, and education) are often selected to improve a classification-based link prediction model (Rowe et al., 2012; Scellato, Noulas, & Mascolo, 2011; Wohlfarth & Ichise, 2008). Although various features can be fused in a classification model to solve the link prediction problem, the class imbalance problem would be difficult to be dealt with and the models are prone to yield biased results.

Another kind of learning approaches are based on probabilistic graphical model (PGM), which builds a statistical network model to solve link prediction problem. The hierarchical network model (Clauset et al., 2008) models a network as a hierarchical random graph and the linking probability between nodes can be calculated by the probability expectation. Stochastic block models (Airoldi et al., 2008; Holland et al., 1983) assume that the network nodes can be partitioned into some blocks, and the linking probability between any two nodes depends on which blocks they belong to. Latent-feature models (Kim & Leskovec, 2011; Miller et al., 2009; Palla et al., 2012; Zhu, 2012) belong to the kinds of probabilistic generative model, where the nodes' latent-features and the edges in a network are all generated based on some distribution. Although the existing PGM-based models provide a deep insight into network structure, the algorithms usually have high complexity and do not scale to large networks.

Comparing with the PGM-based methods, the factorization-based methods usually solve the link-prediction problem by finding the low-rank approximation of the network adjacency matrix (Menon & Elkan, 2011; Zhai & Zhang, 2015). Menon and Elkan (2011) proposed a matrix factorization-based method to address the class imbalance problem by directly optimizing for a ranking loss, and the model is optimized with stochastic gradient descent and scales to large graphs. Zhai and Zhang (2015) investigated both Matrix Factorization (MF) and Autoencoder (AE) application to link prediction problem. They applied dropout to training both the MF and AE parts and showed that it can significantly prevent overfitting by acting as an adaptive regularization. Song, Meyer, and Min (2014) proposed a rank-one alternating direction method of multiplier (ADMM) for nonnegative matrix factorization, and the experiment results demonstrated that rank-one ADMM is more efficient and effective than multiplicative update rule (MUR), alternating least square (ALS), and traditional ADMM.

There are also several of methods aim to solve the link prediction problem by optimizing ranking-based models. Song, Meyer, and Tao (2015b) proposed a top-k link recommendation algorithm by incorporating both the latent features and the explicit features of the network, where the latent features are extracted from the network by optimizing a ranking-based matrix factorization model, and their experiments demonstrated that the algorithm outperforms several state-of-the-art methods. Rendle, Freudenthaler, Gantner, and Schmidt-Thieme (2009) presented a generic optimization criterion BPR-OPT for personalized ranking, and the experiments indicated that the optimization method outperforms the standard learning techniques for MF and k nearest-neighbor. Song, Meyer, and Tao (2015a), Song and Meyer (2015b) proposed a generalized AUC (GAUC) to quantify the ranking performance of potential links (including positive, negative and unknown status links) in partially observed signed social networks. They (Song & Meyer, 2015a) also presented a link ranking approach by optimizing the AUC function, where a log-likelihood of sigmoid function is utilized as a convex surrogate for the indicator function of AUC.

It is notable that in both the metric-based and the learning-based link methods, there are some methods have considered the topological information and non-topological information simultaneously. Bliss, Frank, Danforth, and Dodds (2014) provided an approach to solve link prediction problem by incorporating network topological features and node attributes, and the method exhibited fast convergence and high levels of precision for the top twenty predicted links. Valverde-Rebaza and Lopes (2013a) presented some metric-based link prediction methods by combining structural with community information, and their experiments showed its effectiveness in directed and asymmetric large-scale networks. Munasinghe and Ichise (2012) introduced a new time-aware feature, called 'time score' for link prediction using supervised machine learning methods, and the experimental results verify the effectiveness of time score for link prediction. Murata and Moriyasu (2007) defined a new weighted graph proximity measures which outperforms previous approaches. Pecli, Cavalcanti, and Goldschmidt (2017) reported the effects of three different automatic variable selection strategies (Forward, Backward and Evolutionary) when applied to supervised link prediction, and the results showed that the use of these strategies does lead to better classification models, and combining topological and non-topological data may improve link prediction. Some classical social theories, such as homophily and weak ties, are also used to solve the link prediction problem since they can help to capture the useful interaction patterns. Yang et al. (2011) exploited homophily to predict not only links between a user and his interested services, but also links between users who have common interests. L and Zhou (2009) verified that the weak ties play a significant role in the link prediction problem, especially to remarkably enhance the predicting accuracy. The above work implies a great potential for improving link prediction by combining the topological and non-topological information in social networks.

Table 1
Notations.

Symbol	Explanation
$n \in R$	The number of network nodes.
$N \in R^{n \times n}$	The adjacency matrix of the following/followed network.
$C \in R^{n \times n}$	The adjacency matrix of the constructed TID-based network.
$U \in R^{n \times L}$	The latent-feature matrix of network nodes.
$U_i \in R^{1 \times L}$	The latent-feature vector of user u_i.
$W^0 \in R^{L \times L}$	The linking parameter of network N.
$W^1 \in R^{L \times L}$	The linking parameter of network C.
$\vec{t_i} \in R^{1 \times K}$	The vectorized topic representation of user u_i.
$\vec{A_i} \in R^{1 \times K}$	The dominant topic vector of user u_i.

3. Topic inclusion degree-based network construction

To exploiting the rich content in social-information networks, we first define a user-to-user relation measurement from a perspective of the topic which refers to topic inclusion degree; then construct a network which encodes the information of the topic inclusion degree between users. The mainly used notations are listed in Table 1 before we introduce the method of this paper.

The topic inclusion degree is defined based on the dissemination mechanism of the published content in social-information networks. In a social-information network, the dissemination of the published content totally depends on the following/followed network structure, and the content usually disseminates from followees to followers. Because of this, the dissemination of the published content will lead to the overlap of certain topics between two users u_i and u_j, and also the overlapping part of the topics will account for a certain proportion of the topics of user u_i and user u_j respectively.

Before we start to define the topic inclusion degree, let's take a brief introduction of the "topic". In text mining, a "topic" is generally seen as a cluster of words that frequently occur together, which is implemented by topic model; topic models are the commonly used methods to discovery the hidden semantic structures in large volumes of unlabeled text. Of the existing topic models, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is the most popular one in modeling the topic of a document collection. Intuitively, given a document collection, each document can be represented as a distributed topic vector, and each topic is represented as a distributed word vector based on the LDA model. In a social-information network, the published microblogs/tweets of one user can be seen as one document, and all the users' published microblogs/tweets can be seen as one document collection. Based on the LDA model, we can get each user's distributed topic vector in the social-information network.

Formally, let $\vec{t_i} = < t_{i,1}, \ldots, t_{i,k}, \ldots, t_{i,K} > (t_{i,k} \in (0, 1), \sum_{k=1}^{K} t_{i,k} = 1)$ and $\vec{t_j} = < t_{j,1}, \ldots, t_{j,k}, \ldots, t_{j,K} > (t_{j,k} \in (0, 1), \sum_{k=1}^{K} t_{j,k} = 1)$ denote the K-dimension topic distribution vectors of u_i and u_j, respectively (details of the data preprocessing are introduced in Section 6.2.2). We argue that the topic inclusion degree is not affected by their topic vectors as a whole, but by the most important topics, i.e. the elements of the vector with larger probability values, which are also known as the dominant topics.

Definition 1. (Users' dominant topic vector) Let $\vec{t_i} = < t_{i,1}, \ldots, t_{i,k}, \ldots, t_{i,K} > (t_{i,k} \in (0, 1), \sum_{k=1}^{K} t_{i,k} = 1)$ be the K-dimension topic distribution representations of u_i, and let t_i^h be the top h topic elements of $\vec{t_i}$. The dominant topic vector $\vec{A_i}$ of user u_i is defined as

$$\vec{A}_i = < A_{i,1}, \ldots, A_{i,k}, \ldots, A_{i,K} > \quad (1)$$

where $A_{i,k} = \begin{cases} t_{i,k}, & t_{i,k} \geq t_i^h \\ 0, & t_{i,k} < t_i^h \end{cases}$. The determination of the threshold t_i^h is based on the statistics of the topic elements in datasets, which will be described in the (f) step of the "Data preprocessing" (see Section 6.2.2).

Based on the definition of the dominant topic vector, the topic inclusion degree between user u_i and user u_j is defined as follows.

Definition 2. (Topic inclusion degree (TID)) Let $\vec{A}_i = < A_{i,1}, \ldots, A_{i,k}, \ldots, A_{i,K} >$ and $\vec{A}_j = < A_{j,1}, \ldots, A_{j,k}, \ldots, A_{j,K} >$ be the dominant topic vector of user u_i and user u_j respectively. The topic inclusion degree from user u_i to user u_j can be conveniently defined as

$$C_{ij} = \frac{\sum_{k=1}^{K} min(A_{i,k}, A_{j,k})}{\sum_{k=1}^{K} A_{ik}} \quad (2)$$

where $\sum_{k=1}^{K} \min(A_{ik}, A_{jk})$ denotes the topic overlapping part between u_i and u_j, and $\sum_{k=1}^{K} A_{ik}$ denotes the dominant topic volume of u_i.

The topic inclusion degree can also be seen as a kind of asymmetry topic similarity. That is to say, the weight of the common topic between user u_i and user u_j relative to user u_i or user u_j is different. If $C_{ij} = 0$, it indicates that there is no common topic between user u_i and user u_j; if $C_{ij} = 1$, it indicates that the weight of the common topic between user u_i and user u_j relative to user u_j is 1, which means the topic of user u_j is all included in the topic of user u_i. Based on the definition of the topic inclusion degree, we further construct a topic sparse network as follows.

Definition 3. (Topic inclusion degree (TID)-based network) $N_C = (V_C; E_C; C; C_{thr})$ denotes the network based on topic inclusion degree, where V_C is the set of nodes, E_C is the set of edges, C is the adjacency matrix of the network and each element represents the topic inclusion degree between any pair of nodes u_i and u_j (i and j are the row number and the column number of matrix C respectively), and C_{thr} is a threshold to make the topic inclusion degree matrix C sparse and when the topic inclusion degree C_{ij} is greater than C_{thr}, there is a directed link from user u_i to user u_j.

Note that the TID-based network N_C is a directed network and the corresponding adjacency matrix C is an asymmetric matrix. Fig. 2 presents the construction procedures of the network in a Twitter dataset, and the sparse threshold C_{thr} is 0.92 in the visualization network. Why is it necessary to make the TID-based network sparse? There are two reasons. The rich published content in the social-information network will inevitably bring some noise and the noise will affect the measurement of the TID and then model computing afterward. To relieve the impact of noise, the necessity to make the TID-based network sparse will arise because the sparse procedure will only retain the relations with certain TID between the pair of users in the social-information networks. The other reason is that the sparse of the TID-based network will reduce the complexity of model computing afterward.

4. Fusion probabilistic matrix factorization model

Given the adjacency matrixes N and C of the following/followed network and the TID-based network, the fusion probabilistic matrix factorization (FPMF) model is built to fuse the two kinds of

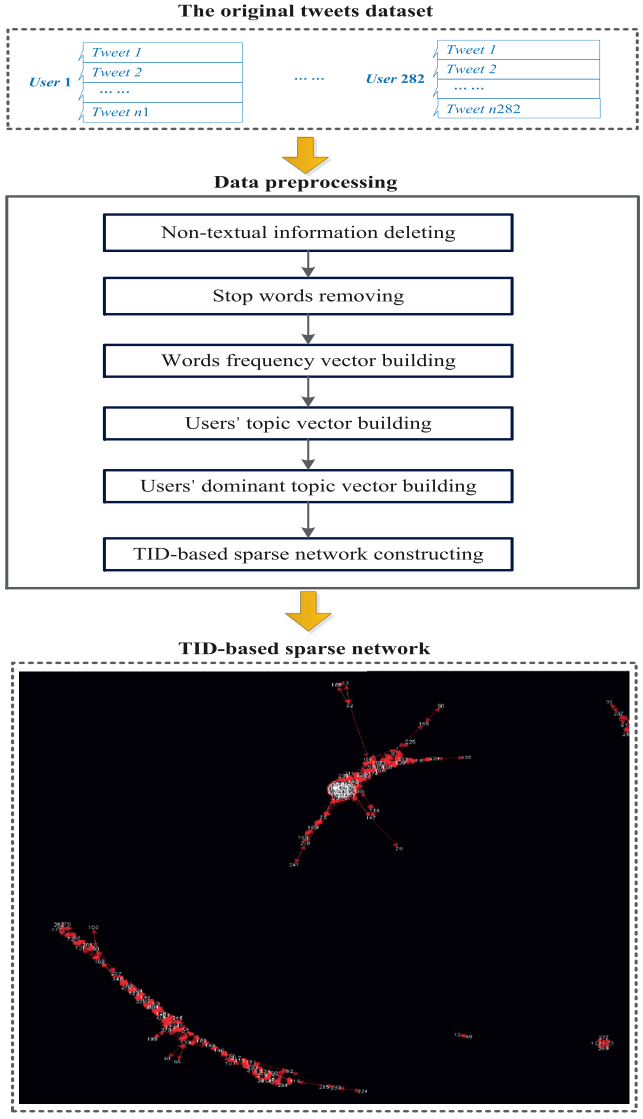


Fig. 2. The construction procedures of the topic inclusion degree based sparse network. The top of the figure is the original tweets dataset of users. The middle of the figure describes the process of the tweets data preprocessing which mainly involves 6 steps, i.e. the non-textual information deletion, stop words removing, words frequency vector building, users topic vector building, users' dominant topic vector building, and the TID-based sparse network constructing (details of the data preprocessing are introduced in Section 6.2.2). The bottom of the figure presents the finally constructed TID-based sparse network, and the sparse threshold C_{thr} is set to 0.92 in the dataset.

network information in a unified probabilistic matrix factorization framework. Specifically, the FPMF model is based on the following assumptions

1. Each network node is represented as a L-dimension latent-feature vector U_i ($i \in \{1, \ldots, n\}$), and U is the $n \times L$ latent-feature matrix of the n nodes in the network. We suppose that U_i obeys a L-dimension Gaussian distribution with mean $\vec{0}$[1] and covariance matrix $\sigma_U^2 I$[2], i.e. $U_i \sim N(\vec{0}, \sigma_U^2 I)$. The probability density of the matrix U can be denoted as:

$$p(U|\sigma_U^2) = \prod_{i=1}^{n} N(U_i|\vec{0}, \sigma_U^2 I) \quad (3)$$

[1] $\vec{0}$ is a zero vector.

[2] I is an identity matrix.

2. For measuring the relation between any two nodes in the following/followed network, a binary relation function is defined as

$$g_0(U_i, U_j) = U_i W^0 U_j^\top \tag{4}$$

where $W^0 \in R^{L\times L}$ is the matrix parameter of the binary relation function g_0.

We assume that the value of the binary function $g_0(U_i, U_j)$ from node u_i to node u_j obeys the Gaussian distribution with mean 1 and variance $\sigma_N^2 I$ if there is a link from node u_i to node u_j. Otherwise, the value of the binary function $g_0(U_i, U_j)$ obeys the Gaussian distribution with mean 0 and variance $\sigma_N^2 I$ if there is not a link from node u_i to node u_j. Formally, $U_i W^0 U_j^\top \sim N(N_{ij}, \sigma_N^2)$. Where N is the n^*n adjacency matrix of the following/followed network. In the adjacency matrix N, the value N_{ij} will be 1 if there is a link from node u_i to node uj, or the value will be 0 if there is no link from node u_i to node u_j. For the matrix parameter W^0, we also suppose that W_l^0 obeys a L-dimension Gaussian distribution with $\vec{0}$ and covariance matrix $\sigma_{W^0}^2 I$, i.e. $W_l^0 \sim N(\vec{0}, \sigma_{W^0}^2 I)$. Based on the assumption, the probability density of the parameter matrix W^0 and the value of $UW^0U^\top$ can be denoted as

$$p(W^0|\sigma_{W^0}^2) = \prod_{l=1}^{L} N(W_l^0|\vec{0}, \sigma_{W^0}^2 I) \tag{5}$$

$$p(UW^0U^\top|N, U, W^0, \sigma_N^2) = \prod_{i=1}^{n}\prod_{j=1}^{n} N(U_i W^0 U_j^\top|N_{ij}, \sigma_N^2) \tag{6}$$

3. Similar to the above assumptions, a binary relation function is defined as

$$g_1(U_i, U_j) = U_i W^1 U_j^\top \tag{7}$$

where W_l^1 obeys a L-dimension Gaussian distribution with mean $\vec{0}$ and covariance matrix $\sigma_{W^1}^2 I$, i.e. $W_l^1 \sim N(\vec{0}, \sigma_{W^1}^2 I)$.

Given the nodes' latent-feature matrix U and parameter matrix W^1, we assume that the value of the binary function $g_1(U_i, U_j)$ from node u_i to node u_j obeys the Gaussian distribution with mean C_{ij} and variance $\sigma_C^2 I$. Where C is the n^*n adjacency matrix of the TID-based network (see Definition 3). Therefore, the probability density of the parameter matrix W^1 and the value of $UW^1U^\top$ can be denoted as

$$p(W^1|\sigma_{W^1}^2) = \prod_{l=1}^{L} N(W_l^1|\vec{0}, \sigma_{W^1}^2 I) \tag{8}$$

$$p(UW^1U^\top|C, U, W^1, \sigma_C^2) = \prod_{i=1}^{n}\prod_{j=1}^{n} N(U_i W^1 U_j^\top|C_{ij}, \sigma_C^2) \tag{9}$$

The reason we introduce the matrix parameter $W^0 \in R^{L\times L}$ or $W^1 \in R^{L\times L}$ in the binary relation function g_0 or g_1 is that $U_i W^0 U_j^\top$ or $U_i W^1 U_j^\top$ represents a generalized measurement from node u_i to node u_j and need to be learned from the specific network data. If there is a symmetry network, the learned linking parameter matrix W^0 or W^1 will be a symmetry parameter matrix or vice versa. If $W^0 = I$ or $W^1 = I$, the binary relation function g_0 or g_1 just corresponds to the inner product between node u_i and node u_j in Euclid space. More important, comparing with the way of the common used low-rank approximation $A \approx UV^\top$ in the collaborative filtering field (where A is the user-item matrix, U is the users' latent-feature matrix, and V is the items' latent-feature matrix), the way of the factorization $N_{ij} \approx U_i W^0 U_j^\top$ or the $C_{ij} \approx U_i W^1 U_j^\top$ is able to model the transitivity in the networks with the same set of nodes, which have been validated by Zhu, Yu, Chi, and Gong (2007).

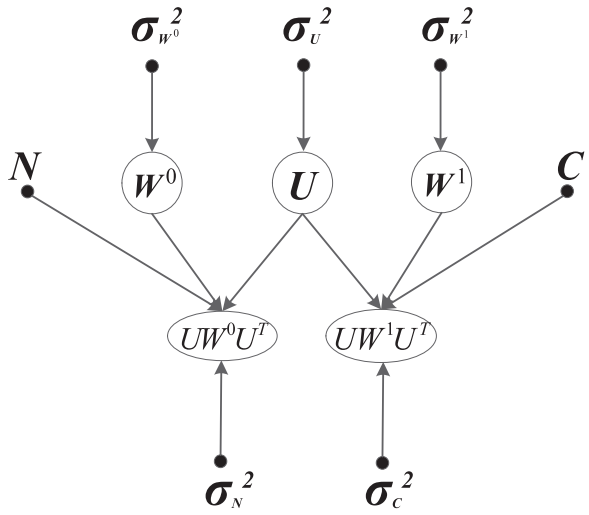


Fig. 3. A directed probabilistic graphical model representing the relations between the matrixes and parameters.

Fig. 3 shows the relations among the matrixes and parameters by using a directed probabilistic graphical model. Through applying the product rule of the directed probabilistic graphical model, the joint probability density distribution over the variables U, W^0, W^1, $UW^0U^\top$ and $UW^1U^\top$ can be represented as the decomposition on the right-hand side of the following equation.

$$\begin{aligned} &p(UW^0U^\top, UW^1U^\top, U, W^0, W^1|N, C, \sigma_U^2, \sigma_{W^0}^2, \sigma_C^2, \sigma_N^2) \\ &= p(UW^0U^\top|N, U, W^0, \sigma_N^2)\, p(UW^1U^\top|C, U, W^1, \sigma_C^2) \\ &\quad p(U|\sigma_U^2)\, p(W^0|\sigma_{W^0}^2)\, p(W^1|\sigma_{W^1}^2) \end{aligned} \tag{10}$$

The goal of the FPMF model is to learn the nodes' latent feature representation U, linking parameters W^0 of the original following/followed network and linking parameters W^1 of the topic inclusion degree-based network by maximizing the joint probability density distribution, which can be deduced as the optimization problem $\underset{U, W^0, W^1}{argmin}\ E_C$, and the objective function E_C is denoted as

$$\begin{aligned} E_C = &\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(N_{ij} - U_i W^0 U_j^\top)^2 + \frac{\lambda_C}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(C_{ij} - U_i W^1 U_j^\top)^2 \\ &+ \frac{\lambda_U}{2}\sum_{i=1}^{n} U_i U_i^\top + \frac{\lambda_{W^0}}{2}\sum_{l=1}^{L} W_l^0 W_l^{0\top} + \frac{\lambda_{W^1}}{2}\sum_{l=1}^{L} W_l^1 W_l^{1\top} \end{aligned} \tag{11}$$

where $\lambda_C = \frac{\sigma_N^2}{\sigma_C^2}$, $\lambda_U = \frac{\sigma_N^2}{\sigma_U^2}$, $\lambda_{W^0} = \frac{\sigma_N^2}{\sigma_{W^0}^2}$, and $\lambda_{W^1} = \frac{\sigma_N^2}{\sigma_{W^1}^2}$. In the end, the objective function E_C can be briefly rewritten as

$$\begin{aligned} E_C = &\frac{1}{2}\| N - UW^0U^\top \|_F^2 + \frac{\lambda_C}{2}\| C - UW^1U^\top \|_F^2 \\ &+ \frac{\lambda_U}{2}\| U \|_F^2 + \frac{\lambda_{W^0}}{2}\| W^0 \|_F^2 + \frac{\lambda_{W^1}}{2}\| W^1 \|_F^2 \end{aligned} \tag{12}$$

The objective function E_C can be solved by using gradient methods, Eq. (13) shows the gradients of function E_C against variables U, W^0 and W^1.

$$\begin{aligned} \frac{\partial E_C}{\partial U} &= (UW^{0\top}U^\top UW^0 + UW^{0\top}U^\top UW^{0\top} - N^\top UW^0 - NUW^{0\top}) \\ &\quad + \lambda_C(UW^{1\top}U^\top UW^1 + UW^{1\top}U^\top UW^{1\top} \\ &\quad - C^\top UW^1 - CUW^{1\top}) + \lambda_U U \\ \frac{\partial E_C}{\partial W^0} &= U^\top UW^0U^\top U - U^\top NU + \lambda_{W^0}W^0 \\ \frac{\partial E_C}{\partial W^1} &= \lambda_C(U^\top UW^1U^\top U - U^\top CU) + \lambda_{W^1}W^1 \end{aligned} \tag{13}$$

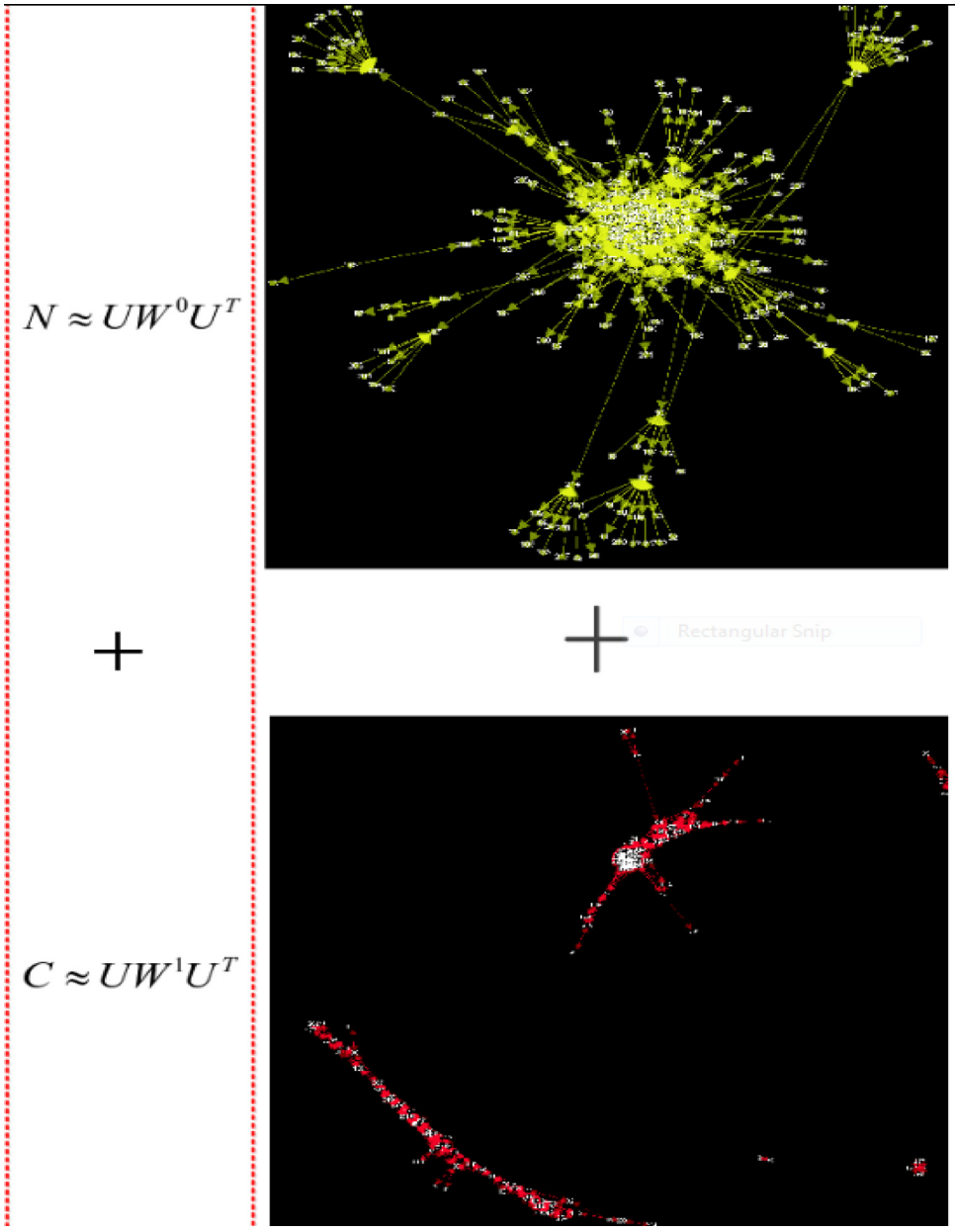


Fig. 4. The sketch of the FPMF model. The upper half of the figure shows the following/followed network, where the adjacency matrix N of the network is approximated as $N \approx UW^0U^\top$. The lower half of the figure shows the constructed TID-based network, where the adjacency matrix C of the network is approximated as $C \approx UW^1U^\top$.

Intuitively, the FPMF model can be visually described as Fig. 4, where the matrix N of the following/followed network is approximated as $N \approx UW^0U^\top$ (see the upper half part of Fig. 4), U_{n^*L} is the L-dimension vector representation of network nodes, and W^0 is a L^*L parameter matrix to measure the following/followed relations; the matrix C of the topic inclusion degree-based network is approximated as $C \approx UW^1U^\top$ (see the lower half part of Fig. 4), where W^1 is a L^*L parameter matrix that used to measure relations based on the topic inclusion degree. Combining the two different approximations, the model aims to learn the low-dimension representations U, and two kinds of matrix parameters W^0 and W^1.

5. Link prediction based on FPMF model

5.1. Link prediction algorithm

We have presented the FPMF model which provides a strategy to fusion the information of the following/followed network and the topic inclusion degree-based network in a unified probabilistic matrix factorization framework. In the FPMF model, the basic part of the model is the approximation $UW^0U^\top$ of the following/followed network N. Supposing we have learned any two users' low-dimension vector representation U_i and U_j and the linking parameter matrix W^0, the linking probability density p_{ij} from user u_i to user u_j can finally be calculated.

$$p_{i,j} = \quad p(U_iW^0U_j^\top | N_{ij} = 1, \sigma_N^2) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(U_iW^0U_j^\top - 1)^2}{2\sigma^2}} \tag{14}$$

The pseudo code of the link prediction procedures based on the FPMF model is shown in Algorithm 1.

5.2. Computational complexity analysis

The computational overhead of the model's learning process is mainly from the calculation of the gradients of the function E_C against variables U, W^0 and W^1. Because of the sparsity of the

Algorithm 1 Link prediction based on FPMF model.

Input: The adjacency matrix N of the following/followed network, and the adjacency matrix C of the constructed TID-based network.

Output: *AUC* value and *Accuracy* result of link prediction.

1: **Initialize:** Users' L-dimension representation U_{n*L}, matrix parameters W^0_{L*L} of N, and matrix parameters W^1_{L*L} of C, $L = integer \ll n$.
2: **repeat**
3: $U^{new} := U^{old} - \gamma \frac{\partial E_C}{\partial U}$. (see Eq. (13))
4: $W^{0new} := W^{0old} - \gamma \frac{\partial E_C}{\partial W^0}$. (see Eq. (13))
5: $W^{1new} := W^{1old} - \gamma \frac{\partial E_C}{\partial W^1}$. (see Eq. (13))
6: **until** Convergence
7: **for** each pair $<u_i, u_j>$ **calculate**
8: $p_{i,j} = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(U_i W^0 U_j^\top - 1)^2}{2\sigma^2}}$. (see Eq. (14))
9: **end for**
10: **Calculate** $AUC = \frac{r' + 0.5r''}{r}$. (see Eq. (15))
11: **Calculate** $Accuracy = \frac{k}{P}$. (see Eq. (16))
12: **Return:** *AUC* value and *Accuracy* results.

matrixes N and C, the computational complexity of multiplication of N and U is $O(\mu_N L)$, where μ_N is the number of nonzero entries in N. Similarly, the computational complexity CU is $O(\mu_C L)$. The computational complexity of the rest multiplications in the gradients is $O(nL^2)$. Therefore, the total computational complexity of the model in one iteration is $O(\mu_N L + \mu_C L + nL^2)$. In other words, the computational complexity is linearly as the increase of the nonzero entries in the matrixes N and C.

6. Experiments

In this section, we conduct the experiments for the following purposes: (1) find out whether the proposed fusion model is superior to baseline methods in link prediction, (2) find out whether our method is superior to other methods in link prediction, (3) analyze the impacts of the sparseness of the constructed topic inclusion degree-based networks on the performance of link prediction, (4) find out the impacts of the weight parameter λ_C on link prediction.

6.1. Datasets

In our experiments, we adopt two kinds of social-information network datasets, i.e. Weibo and Twitter. Each of them contains information on two aspects, the following/followed network and users' published content. Twitter and Weibo are well-known social media platforms for people to create, spread, and share information. In such platforms, the following/followed user-to-user relations formed directed networks. The Weibo and Twitter datasets[3] we used in this study are derived from Zhang, Liu, Tang, Chen, and Li (2013a), which was crawled from Weibo and Twitter platforms. Since many users have published little or no contents, we select the users who have more than 50 microblogs/tweets to meet the experiment requirement. Finally, we extract five sub-datasets, i.e. Weibo 1, Weibo 2, Twitter 1, Twitter 2 and Twitter 3 as the experimental datasets. Details of the Weibo and Twitter datasets are shown in Table 2. The in-degree distribution of the nodes is also shown in each network (see Fig. 5(a)–(e)), where the significant characteristics of power-law distribution are manifested.

Table 2
Statistics of the datasets.

Datasets	Tweets number	Links	Nodes	Link density (%)
Weibo 1	51,597	3182	977	0.33
Weibo 2	70,654	5877	1378	0.31
Twitter 1	13,792	1475	282	1.84
Twitter 2	16,187	1551	337	1.37
Twitter 3	543,230	26,243	11,328	0.21

6.2. Experiment setup and data pre-processing

6.2.1. Experiment setup

Following the widely used evaluation method in link prediction literatures (Backstrom & Leskovec, 2011; Clauset et al., 2008; Lü & Zhou, 2011; Zhai & Zhang, 2015), we split the set of the original following/followed links E into the training set E^T and the testing set E^P. Specifically, we take randomly one-tenth of the links in E as the testing set E^P, and the remaining links constitute the training set E^T.

For the parameter selection of our FPMF models, we conduct 10-fold cross-validation in the training set E^T for each set of parameter values. Specifically, we divide the training set E^T into 10 equal subsets, each subset is served as the validation set for testing the model, and the remaining 9 subsets are used as training data. For each set of parameter values, the cross-validation process is repeated 10 times, with each of the 10 subsets used exactly once as the validation data. Then we average the results of the 10-fold cross validation as the estimation of the model under the set of parameter values. We finally select the set of parameter values corresponding to the best average results, and the selected parameters are $(K = 30, L = 60, C_{thr} = 0.95, \lambda_C = 0.80)$ for Twitter 1, $(K = 20, L = 60, C_{thr} = 0.70, \lambda_C = 0.40)$ for Twitter 2, $(K = 100, L = 80, C_{thr} = 0.50, \lambda_C = 0.40)$ for Twitter 3, $(K = 20, L = 80, C_{thr} = 0.97, \lambda_C = 1.20)$ for Weibo1, and $(K = 40, L = 80, C_{thr} = 0.70, \lambda_C = 2.0)$ for Weibo2, in our FPMF models. In addition, the regularization parameters λ_U, λ_{W^0} and λ_{W^1} are set to 0.05.

It is notable that if there's no special specifying, the experiment setup including the train-test set partition and the parameter configuration are the same in the following experiments.

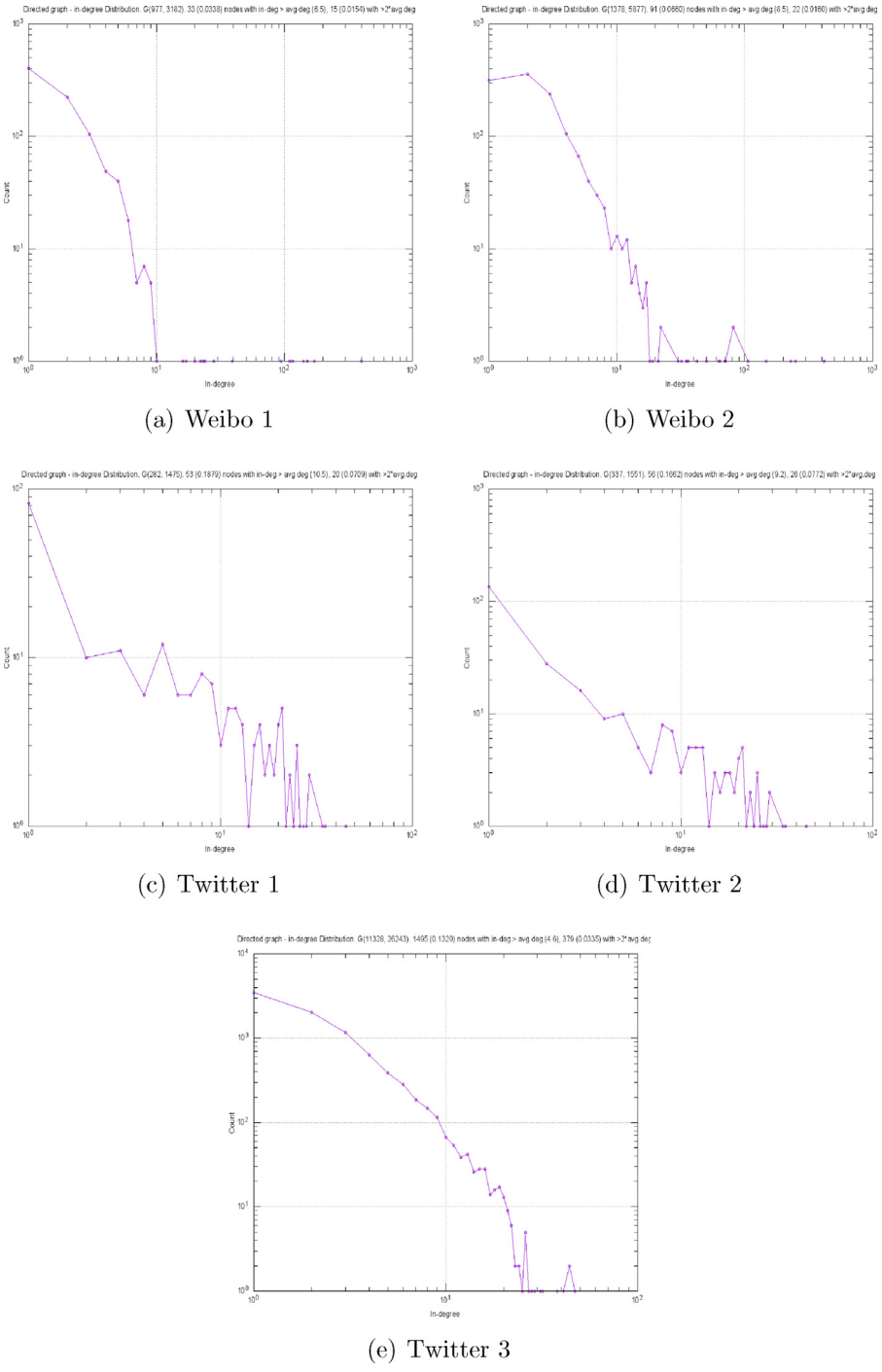
6.2.2. Data preprocessing

Before the experiment, we need to process the published microblogs/tweets in the five social-information network datasets and further build the TID-based sparse networks.

1. For the Weibo datasets, we process the microblogs as following procedures.
   (a) Non-textual information deletion: The messy codes and tags are first deleted by scanning the content of the original microblogs.
   (b) Chinese words segmentation (Chang, Galley, & Manning, 2008; Sproat, Gale, Shih, & Chang, 1996): It is a basic procedure for Chinese language processing. In this step, we process the texts using the words segmentation tool "jieba"[4].
   (c) Stop words removing: It is also a necessary step in text mining and is usually to remove the nonsense words and the extremely common words. We remove the stop words based on the stop words list we've collected.
   (d) Words frequency vector building: Based on the above preprocessing, we can build the word frequency vector. Specifically, we first determine the dimension of the vector for each dataset according to the number of words in the corresponding dataset; then we treat all the microblogs that one

[3] https://aminer.org/Influencelocality.

[4] https://github.com/fxsjy/jieba.



(a) Weibo 1 (b) Weibo 2

(c) Twitter 1 (d) Twitter 2

(e) Twitter 3

Fig. 5. The in-degree distribution of nodes in the five networks.

user has posted as one text document and vectoring the text document in the words frequency space; finally, the dimensions of the users' frequency vector corresponding to the two Weibo datasets (Weibo 1, Weibo 2) are 3740 and 6011 respectively.

(e) Users' topic vector building: In text mining, Latent Dirichlet allocation (LDA) (Blei et al., 2003) is a generative statistical model which aims to mine the latent topic representation. Based on the users' words frequency vector in the last step, we obtain each user's topic vector representation by using the LDA toolkit from scikit-learn (Pedregosa et al., 2011).

(f) Users' dominant topic vector building: Based on the definition of the dominant topic vector (see Definition 1.), we can get each user's dominant topic vector. The selection of the top h topic elements in the dominant topic vector is based on the statistics of the topic elements weight. We have counted the proportion of the top h topic elements in the different value of h $(1 \leq h \leq L)$. We found when $h = 5$, the proportion of the top h topic elements is over 90% in all the datasets. Thus we take the value of $h = 5$ to build users' dominant topic vectors.

(g) TID-based sparse network constructing: Based on Definition 2 and Definition 3, we can further construct the TID-based sparse network for each dataset.

2. For the Twitter datasets, the same preprocessing steps need be adopted in the tweets, except for the words segmentation pro-

cedure. Note that the tweets data are already preprocessed in the original data resource, thus we extract the users' words frequency vector directly, and the dimensions of the users' frequency vector corresponding to the three Twitter datasets (Twitter 1, Twitter 2 and Twitter 3) are 919, 1877, and 102,033 respectively; the remaining preprocessing steps are the same as the preprocessing steps of Weibo datasets.

6.3. Evaluation measures

For evaluating the experimental results, we adopt two widely used measures 'AUC' (Area under the receiver operating characteristic) and 'Accuracy' to quantify the link prediction performance in the link prediction literature (Lü & Zhou, 2011; Valverde-Rebaza & Lopes, 2013a; 2013b; Yin, Hong, & Davison, 2011; Zhang, L, Wang, Zhu, & Zhou, 2013b).

1. *AUC*. Given the ranking of all non-observed links in the present network, the *AUC* value can be seen as the probability that a randomly chosen non-observed link is given a higher score than a randomly chosen non-existent link (Hanley & McNeil, 1982). Specifically, we can use the following equation to compute the *AUC* value.

$$AUC = \frac{r' + 0.5r''}{r} \tag{15}$$

where r is the number of independent comparisons, r' is the times for the non-observed links are given higher scores than non-existent links, and r'' is the times for the scores of the non-observed links are equal to the scores of the non-existent links. The value of r is set to 10,000 in our experiments.
2. *Accuracy*. It is defined as the ratio between the k accurately predicted links and the top-P predicted links. In our experiments, the value of P is the number of the links in the testing set.

$$Accuracy = \frac{k}{P} \tag{16}$$

6.4. Comparison methods

To verify the performance of the proposed method, we conduct two parts of comparison experiments. One is to compare the proposed method with two baseline methods, and the other is to compare our method with 11 typical link prediction methods.

6.4.1. The baseline methods

1. *Basic model*. The basic probabilistic matrix factorization model for link prediction, which only considers the following/followed network.

$$E = \frac{1}{2} \parallel N - UW^0U^\top \parallel_F^2 + \frac{\lambda_U}{2} \parallel U \parallel_F^2 + \frac{\lambda_{W^0}}{2} \parallel W^0 \parallel_F^2 \tag{17}$$

The optimization problem $\underset{U,W^0}{argmin}\, E$ can be solved by using gradient methods, Eq. (18) shows the gradients of the function E against variables U and W^0.

$$\begin{aligned} \frac{\partial E}{\partial U} &= (UW^{0\top}U^\top UW^0 + UW^0U^\top UW^{0\top} \\ &\quad - N^\top UW^0 - NUW^{0\top}) + \lambda_U U \\ \frac{\partial E}{\partial W^0} &= (U^\top UW^0U^\top U - U^\top NU) + \lambda_{W^0} W^0 \end{aligned} \tag{18}$$

2. Topic inclusion degree (TID)-based link prediction. The method indicates that two users are more likely to have a link if they have a larger topic inclusion degree (see Definition 3).

6.4.2. Popular link prediction methods

1. Katz is mentioned as the best link prediction method in Katz (1953), which computes the similarity between nodes u_i and u_j by summing over all possible paths from u_i to u_j.

$$\sum_{l=1}^{m} \beta^l \cdot |paths^l_{u_i,u_j}| \tag{19}$$

where $paths^l_{u_i,u_j}$ is the set of all length-l path from user u_i to user u_j, and β is the weight to the contribution of the paths with different lengths (we use $\beta = 0.001$ in our experiments).
2. Common Neighbors (CN). The CN metric is one of the most widely-used measurements in link prediction mainly due to its simplicity(Newman, 2001), and it did give better performance in many link prediction tasks for networks. Because there are different types of neighbors (followers and followees) in directed social-information network, we base our calculation of common neighbors on the number of different types of neighbors between user u_i and user u_j.
   - CN1(Common followers):
     $$|\Gamma^-(u_i) \cap \Gamma^-(u_j)| \tag{20}$$
   - CN2(Common followees):
     $$|\Gamma^+(u_i) \cap \Gamma^+(u_j)| \tag{21}$$
   - CN3(Common friends):
     $$|(\Gamma^-(u_i) \cap \Gamma^+(u_i)) \cap (\Gamma^-(u_j) \cap \Gamma^+(u_j))| \tag{22}$$
   - CN4(Common followees and followers):
     $$|(\Gamma^-(u_i) \cup \Gamma^+(u_i)) \cap (\Gamma^-(u_j) \cup \Gamma^+(u_j))| \tag{23}$$

   where $\Gamma^+(u_i)$ denotes the set of neighbors following user u_i, $\Gamma^-(u_i)$ denotes the set of neighbors followed user u_i, and $|\Gamma^+(u_i)|$ and $|\Gamma^-(u_i)|$ separately denote the number of elements of set $\Gamma^+(u_i)$ and set $\Gamma^-(u_i)$.
3. Preferential Attachment (PA)-based method (Barabási & Albert, 1999). The preferential attachment has received considerable attention as a model of the growth of networks (Barabâsi et al., 2002). Given the number of neighbors of user u_i and user u_j, this index is defined as the product of the neighbor numbers. As mentioned in the definition of *CN* metrics above, we also define four *PA*-based metrics based on the different types of neighbors between user u_i and user u_j in a directed network.
   - PA1(Followers-based):
     $$|\Gamma^-(u_i)| \times |\Gamma^-(u_j)| \tag{24}$$
   - PA2(Followees-based):
     $$|\Gamma^+(u_i)| \times |\Gamma^+(u_j)| \tag{25}$$
   - PA3(Friends-based):
     $$|\Gamma^-(u_i) \cap \Gamma^+(u_i)| \times |\Gamma^-(u_j) \cap \Gamma^+(u_j)| \tag{26}$$
   - PA4(Followers and followees based):
     $$|\Gamma^-(u_i) \cup \Gamma^+(u_i)| \times |\Gamma^-(u_j) \cup \Gamma^+(u_j)| \tag{27}$$
4. Low-rank approximation (LRA)-based methods (Liben-Nowell & Kleinberg, 2007). Low-rank approximation, a common technique to find an approximate matrix with low-rank of an original matrix. For link prediction, LRA is to compute the rank-k matrix N_k that best approximates the network adjacency matrix N and to use entry $<i, j>$ in the matrix N_k as the linking score between user u_i and user u_j. This can be done efficiently using many low-rank approximation techniques, and we use two common *LRA* techniques as the comparison methods in our experiments: Singular Value Decomposition (SVD) (Golub & Reinsch, 1970) and Non-negative Matrix Factorization (NNMF) (Lee & Seung, 1999). We denote the two LRA-based methods as LRA-SVD and LRA-NNMF separately.

Table 3
AUC results of baseline methods.

Data sets	Baseline methods		Our method
	Basic model	TID	FPMF model
Twitter 1	0.8506	0.6883	0.9502
Twitter 2	0.7676	0.5932	0.9447
Twitter 3	0.9055	0.8164	0.9128
Weibo 1	0.9758	0.5848	0.9886
Weibo 2	0.9489	0.5180	0.9573

6.4.3. The methods considering topological and non-topological information simultaneously

1. WIC. It has been validated to be the best link prediction method in literature (Valverde-Rebaza & Lopes, 2013a), which considers community membership information.

$$s_{u_i,u_j}^{WIC} = \begin{cases} |\Lambda_{u_i,u_j}^{W}|, & if\ \Lambda_{u_i,u_j}^{W} = \Lambda_{u_i,u_j} \\ \dfrac{\Lambda_{u_i,u_j}^{W}}{\Lambda_{u_i,u_j}}, & otherwise \end{cases} \tag{28}$$

where Λ_{u_i,u_j} is the number of the all common neighbors between u_i and u_j, and Λ_{u_i,u_j}^{W} is the number of the within-community common neighbors between u_i and u_j.

2. RA-W. It is also from literature (Valverde-Rebaza & Lopes, 2013a) considering the community membership information, which is the second best link prediction method.

$$s_{u_i,u_j}^{RA-W} = \sum_{z \in \Lambda_{u_i,u_j}^{W}} \frac{1}{|\Lambda_{u_i,u_j}|} \tag{29}$$

3. CMA-ES1 (Bliss et al., 2014). It is a linear combination of similarities, which contains both topological and non-topological similarities, and the combination weights is optimized by applying the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) (Hansen & Ostermeier, 2001), and we refer to CMA-ES1 in this paper.
4. CMA-ES2 (Bliss et al., 2014). For fair comparison with our method, we also implement an another CMA-ES based method, which replaces the original key-words based similarity with our TID-based metric (see Definition 3).

6.5. Experimental results

6.5.1. Results of baseline methods

Table 3 shows the performance of our proposed method compared with two baseline methods, i.e. the Basic model, and the TID-based method in the datasets of Weibo and Twitter separately. From the results, we can see that the proposed FPMF model outperforms the Basic model significantly. The Basic model underperforms our method because another side of information, i.e. the published content by users in the social-information networks, is ignored in the Basic model. The metric-based baseline method TID is lower not only than our FPMF model but also than the Basic model. This indicates that a single metric has a very limited role in predicting links in social-information networks. The FPMF model considers both the information of the following/followed network and the TID information, and therefore, our link prediction method is superior to the baseline methods.

6.5.2. Results of comparison methods

Tables 4 and 5 show the AUC values and Accuracy results of our FPMF model compared with 15 link prediction methods in the five Twitter and Weibo datasets separately. In most cases, the results show that the proposed FPMF model get better Accuracy results and AUC values than the compared methods. Table 6 shows the P value of the statistical T test between our method and the comparison methods. We can see all the statistical P-value is less than 0.05, which indicates that our method is significantly better than the comparison methods.

As for the four common neighbors-based metrics, CN1, CN2, CN3 and CN4, their performances are almost equal to pure chance in the very sparse Weibo datasets but better in Twitter; which indicates that the neighbors-based metrics are susceptible to the sparseness of the networks (see the Link Density in Table 2). The four Preferential Attachment-based metrics, PA1, PA2, PA3 and PA4,

Table 4
AUC results of the experimental comparison.

Methods	Weibo 1	Weibo 2	Twitter 1	Twitter 2	Twitter 3
CN1	0.4566	0.5191	0.8490	0.8356	0.7820
CN2	0.4688	0.5477	0.8955	0.8878	0.8050
CN3	0.4742	0.5422	0.7751	0.7447	0.6745
CN4	0.5240	0.5832	0.9045	0.9105	0.8621
PA1	0.9150	0.8847	0.9231	0.9141	0.8297
PA2	0.8936	0.8664	0.9405	0.9416	0.8537
PA3	0.6990	0.6810	0.8998	0.8917	0.7709
PA4	0.7881	0.8010	0.9502	0.9418	0.8736
Katz	0.8466	0.8778	0.9416	0.9363	0.9159
LRA-NNMF	0.8527	0.9028	0.9354	0.9396	0.9061
LRA-SVD	0.7594	0.8394	0.7526	0.7751	0.8703
WIC	0.4995	0.5553	0.6394	0.6313	0.6073
RA-W	0.4979	0.4860	0.6407	0.6304	0.6122
CMA-ES1	0.8933	0.8639	0.9412	0.9317	0.8141
CMA-ES2	0.9042	0.8531	0.9436	0.9382	0.8635
FPMF	0.9886	0.9597	0.9517	0.9494	0.9187

Table 5
Accuracy results of the experimental comparison.

Methods	Weibo 1	Weibo 2	Twitter 1	Twitter 2	Twitter 3
CN1	0.0014	0.0034	0.0879	0.0897	0.0541
CN2	0.0035	0.0034	0.06757	0.0769	0.0392
CN3	0.0015	0.0102	0.1254	0.1090	0.0194
CN4	0.0013	0.0034	0.0946	0.0577	0.0396
PA1	0.0564	0.0460	0.1081	0.1090	0.0027
PA2	0.0125	0.0018	0.1014	0.0577	0.0103
PA3	0.0251	0.0153	0.1081	0.0833	0.0061
PA4	0.0031	0.0187	0.1284	0.0833	0.0328
Katz	0.0001	0.0068	0.1014	0.0769	0.1036
LRA-NNMF	0.0897	0.0886	0.0932	0.0615	0.0942
LRA-SVD	0.0646	0.1014	0.0811	0.0577	0.1170
WIC	0.0001	0.0004	0.0270	0.0449	0.0324
RA-W	0.0002	0.0004	0.0622	0.0446	0.0467
CMA-ES1	0.0243	0.0172	0.1051	0.1240	0.1225
CMA-ES2	0.0627	0.0751	0.1203	0.0943	0.1342
FPMF	0.1210	0.1452	0.1554	0.1179	0.1665

Table 6
The P value of the statistical T test between our method and the comparison methods.

Evaluation metrics	AUC	Accuracy
CN1	0.0135	0.0025
CN2	0.0321	0.0005
CN3	0.0008	0.0139
CN4	0.0486	0.0009
PA1	0.0187	0.0087
PA2	0.0345	0.0011
PA3	0.0085	0.0031
PA4	0.0494	0.0078
Katz	0.0469	0.0095
LRA-NNMF	0.0420	0.0011
LRA-SVD	0.0003	0.0046
WIC	0.0000	0.0001
RA-W	0.0000	0.0001
CMA-ES1	0.0363	0.0409
CMA-ES2	0.0400	0.0283

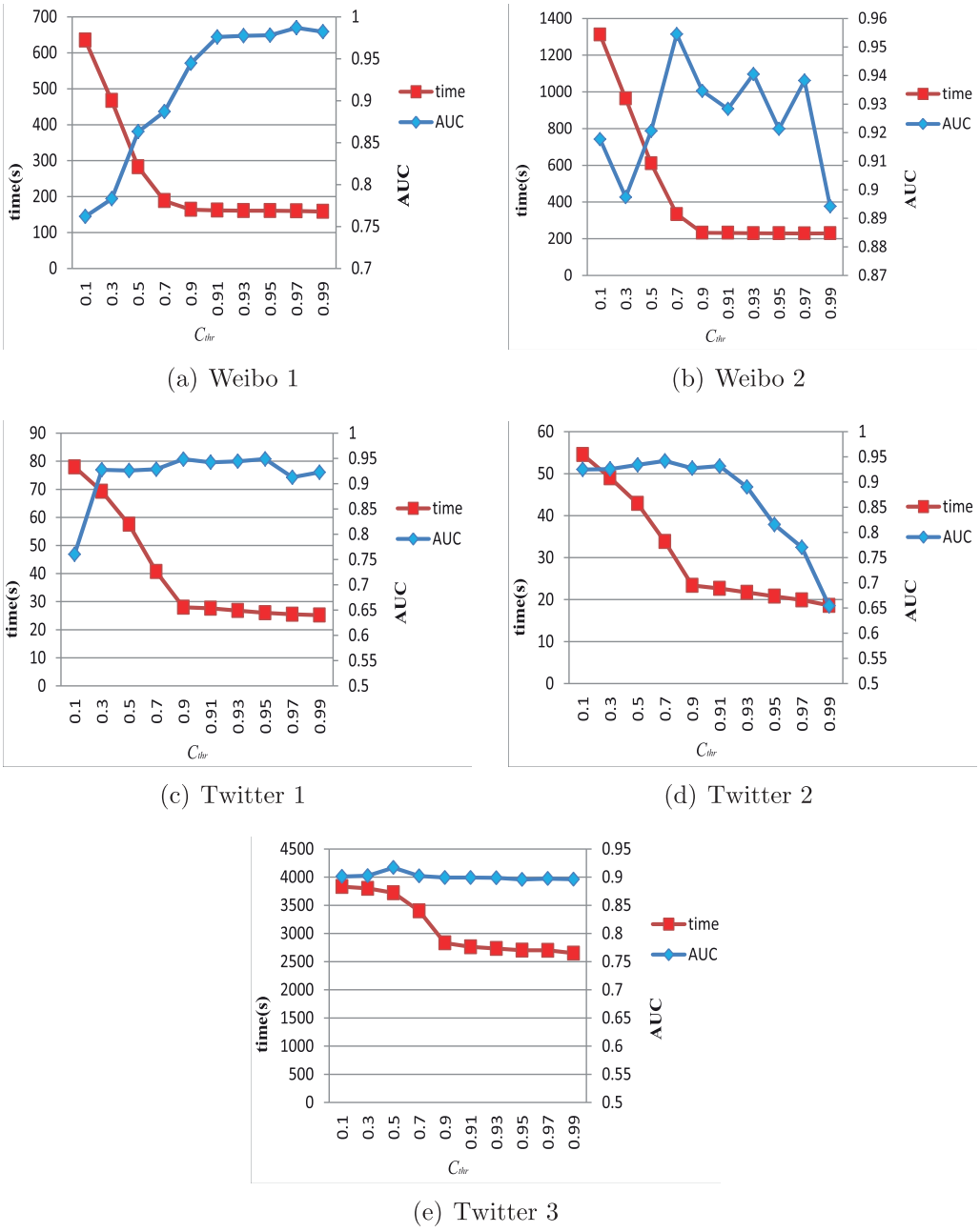


Fig. 6. Sparse parameter C_{thr} analysis. The five figures from (a) to (e) respectively show the performances of the FPMF model on data sets Weibo 1, Weibo 2, Twitter 1, Twitter 2 and Twitter 3, when the sparse parameter C_{thr} changes. The horizontal axis in each figure represents the values of C_{thr}, the vertical axis on the left in each figure represents the running time (s) of the model, and the vertical axis on the right in each figure represents the *AUC* values of link prediction.

get better results than the four common neighbors-based metrics since the obvious power law distribution of the node in-degree in the five network datasets (see Fig. 5(a)–(e)). However, their results are significantly inferior to our method. The performances of the Katz-based method are also inferior to our model. In short, the advantage of the compared metric-based methods is that they have low complexity and only need to use little information to achieve the link prediction; while each metric has certain limitation and one single metric usually cannot get very high results because it exploits little network information.

In terms of the LAR-based methods LAR-SVD and LRA-NNMF, the AUC value and the Accuracy results of LRA-NNMF are better than LRA-SVD, and they are still inferior to our FPMF model. The two matrix factorization-based methods in essence are learning-based method by fitting the network adjacent matrix. Their advantage is that they can fit the network adaptively in learning process, while overfitting is usually the big problem in the methods.

The methods WIC and RA-W are also belonging to the common neighbor-based methods, and they consider the common neighbors from the inside and outside communities. They show poor results in the five social-information networks, which indicates that the community information is insufficient and even pull down the results of the original common neighbors-based methods. The reason for this is that the community information may limit the weak ties prediction in the social-information networks. The CMA-ES1 and CMA-ES2 methods get the second best link prediction result in all comparison methods because they consider the combination of many metrics in topological and non-topological. That is to say, the more features the better results in link prediction can be

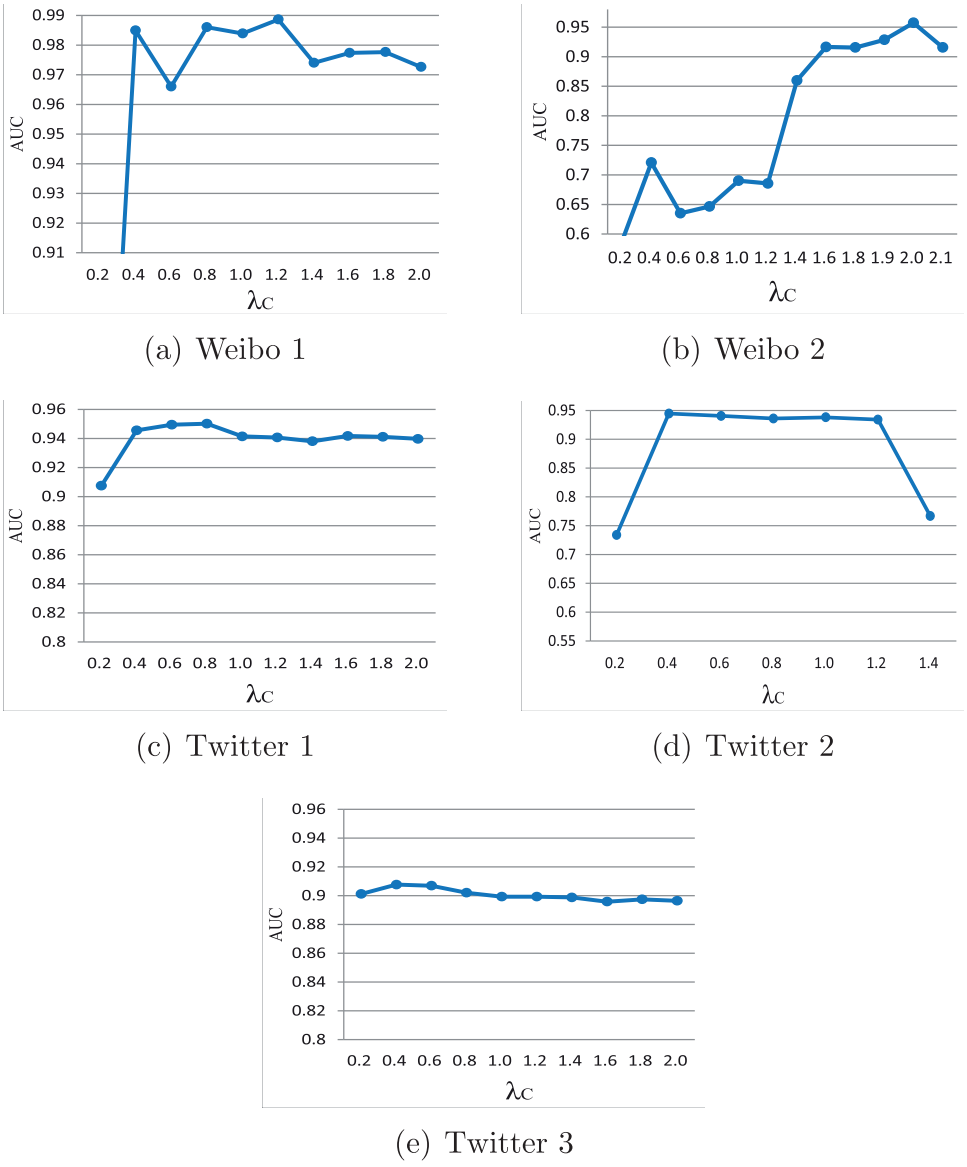


Fig. 7. Parameter λ_C analysis. The five figures from (a) to (d) respectively show the AUC values on data Weibo 1, Weibo 2, Twitter 1, Twitter 2 and Twitter 3 when the weight parameter λ_C changes in the FPMF model. The horizontal axis in each figure represents the value of λ_C, and the vertical axis in each figure represents the AUC values of link prediction.

obtained if the features are combined properly; while the more features mean a higher complexity.

To sum up, the reasons why our FPMF model is superior to the other methods can be elaborated from two aspects: From the perspective of information fusion, our model can be seen as the fusion of the network information with the topic-based information in one unified factorization-based model, and the topic-based information (the TID information) is a useful supplement to network topological information; from the perspective of machine learning, the adopted of the TID information in our FPMF model can be seen to add a regularization in the factorization-based objective function from the topic-based semantic space. Therefore, our method has good generalization ability and shows better results.

6.5.3. *Sparse parameter* C_{thr} *analysis*

In our model, the sparse parameters C_{thr} correspond to the constructed topic inclusion degree-based network, and they control the sparseness of the network. The bigger the sparse parameter is, the sparser the network will be.

The figures from 6(a) to (e) show the performance of the FPMF model on Twitter and Weibo datasets when the sparse parameter α_C changes. The horizontal axis of each figure represents the value of C_{thr} from 0.1 to 0.99, and the two vertical axes in each figure represent the *AUC* results of link prediction and the running time (s) of the 1000 iterations in the FPMF model separately. From the red curves of these figures, we can observe when the sparse parameter C_{thr} become bigger, the iterating time of the model is significantly reduced. The changes of the red curves indicate that the sparser of the network is, and the less time of the model runs. From the blue curves in these figures, we can observe when the sparse parameter C_{thr} increases, the *AUC* values of link prediction also increase at first, but when α_C surpasses a certain threshold, the prediction results will stop increasing and even decrease with further increase of the C_{thr} value. This phenomenon indicates that the reasonable use of the sparseness of the constructed underlying network is effective to improve the performance of link prediction. While over-sparse or under-sparse of the network in the FPMF model will produce lower results in link prediction. In-

tuitively, over-sparse of the constructed network means that the FPMF model cannot fuse the underlying information enough to get good results for link prediction, and under-sparse of the underlying network means that the FPMF model fuse much more user-to-user TID-based information, but at the same time high noise will be brought to the model, which affects model's performance.

6.5.4. Weight parameter λ_C analysis

The weight parameter λ_C balances the information between the original following/followed network and the topic inclusion degree-based network. In the FPMF model, if $\lambda_C = 0$, the model only considers the information from the following/followed network, and if $\lambda_C = inf$ (*inf* represent that the value of λ_C is infinity), the model only extracts information from the topic inclusion degree-based network.

The figures from 7(a) to (e) show the performances of the FPMF model on Twiter and Weibo datasets when λ_C is changed separately. We observed that the value of λ_C impacts the link prediction results significantly, which demonstrates that fusing the information of the underlying network greatly improves the prediction results. As λ_C increases, the prediction results also increase at first, but when λ_C surpasses a certain threshold, the prediction results stop increasing and even decrease. The phenomenon coincides with the intuition that overmuch usage of such information, the observed network, underlying network, rather than reasonable fusing these resources together, cannot generate best performances.

6.6. Method discussions

In this section, we summarize the characteristics and advantages of the proposed method and discuss the scalability of the method.

1. Constructing the topic inclusion degree-based network is one of the characteristics of our method.
   The construction of the topic inclusion degree-based network provides a way to structure the rich published content in social-information networks. Topic inclusion degree is uniquely user-to-user semantic in social-information networks, and the TID-based network encodes the user-to-user semantic relations in topic space.
2. The constructed TID-based network and the original following/followed network are fused in one unified probabilistic matrix factorization framework.
   In the unified probabilistic matrix factorization framework, the FPMF model fuses the two networks into a single, consistent, and compact feature representation of the network nodes. The learning outcome of the FPMF model is the result of information fusion, which not only benefits for performing on the link prediction task in social-information networks but also helps to perform on many other network data mining tasks.
3. The FPMF model provides a general modeling strategy to fuse network information.
   The FPMF model we built in this paper focuses on fusing the information of the following/followed links and the user-to-user semantic relations based on topic. Actually, the FPMF model is not limited to fuse the information of the topic-based network. It provides a general strategy to model the multi-network information via probability matrix factorization.

7. Conclusions and future work

The study of how to accurately infer the node-to-node relations in social-information networks still remains a challenge. This study presents a fusion model, in which the information of the original following/followed network and a topic-based network are fused in one unified probabilistic matrix factorization framework. Based on the learned latent-feature representation and the learned matrix linking parameters of the fusion model, the linking probability between any pair of the network nodes is obtained. To assess the performance of the proposed model for link prediction, we compare our approach with nine metric-based methods and two traditionally matrix factorization-based methods. Experiments with two types of real-world social-information networks Twitter and Weibo show that the proposed approach is effective for predicting the links in the social-information networks.

This work has many potential directions in the future. For example, we could study how to conduct incremental learning on the proposed fusion model so that the model could be adapted to a dynamic circumstance. Besides, it is worth pointing out that studying on the explanation of the observed links by exploiting the rich published content will be a very interesting direction.

Acknowledgments

This work was supported by the State Key Program of National Natural Science Foundation of China (No.61432011, No.U1435212), the Key Scientific and Technological Project of Shanxi Province (MQ2014-09), and the 1331 Engineering Project of Shanxi Province, China.

Appendix. The derivation process of the objective function

In the appendix, we will derive the following objective function of the FPMF model.

$$E_C = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(N_{ij} - U_i W^0 U_j^\top)^2 + \frac{\lambda_C}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(C_{ij} - U_i W^1 U_j^\top)^2 + \frac{\lambda_U}{2}\sum_{i=1}^{n} U_i U_i^\top + \frac{\lambda_{W^0}}{2}\sum_{l=1}^{L} W_l^0 W_l^{0\top} + \frac{\lambda_{W^1}}{2}\sum_{l=1}^{L} W_l^1 W_l^{1\top} \quad (30)$$

The joint probability density distribution over the variables U, W^0, W^1, $UW^0U^\top$ and $UW^1U^\top$ can be expressed as

$$\begin{aligned} &p(UW^0U^\top, UW^1U^\top, U, W^0, W^1 | N, C, \sigma_U^2, \sigma_{W^0}^2, \sigma_C^2, \sigma_N^2) \\ &= p(UW^0U^\top | N, U, W^0, \sigma_N^2)\, p(UW^1U^\top | C, U, W^1, \sigma_C^2) \\ &\quad p(U|\sigma_U^2)\, p(W^0|\sigma_{W^0}^2)\, p(W^1|\sigma_{W^1}^2) \end{aligned} \quad (31)$$

Maximizing the joint probability density distribution is equivalent to maximizing the following logarithmic function.

$$\begin{aligned} &\ln p(UW^0U^\top, UW^1U^\top, U, W^0, W^1 | N, C, \sigma_U^2, \sigma_{W^0}^2, \sigma_C^2, \sigma_N^2) \\ &= \ln p(UW^0U^\top | N, U, W^0, \sigma_N^2) + \ln p(UW^1U^\top | C, U, W^1, \sigma_C^2) \\ &\quad + \ln p(U|\sigma_U^2) + \ln p(W^0|\sigma_{W^0}^2) + \ln p(W^1|\sigma_{W^1}^2) \end{aligned} \quad (32)$$

where the five condition probability density distribution functions can be expressed as Eqs. (33)–(37) because of the Gaussian distributional assumptions.

$$p(U|\sigma_U^2) = \prod_{i=1}^{n} (2\pi)^{-\frac{l}{2}} |\sigma_U^2 I|^{-\frac{1}{2}} e^{-\frac{1}{2} U_i (\sigma_U^2 I)^{-1} U_i^\top} \quad (33)$$

$$p(W^0|\sigma_{W^0}^2) = \prod_{i=1}^{n} (2\pi)^{-\frac{l}{2}} |\sigma_{W_l^0}^2 I|^{-\frac{1}{2}} e^{-\frac{1}{2} W_l^0 (\sigma_{W_l^0}^2 I)^{-1} W_l^{0\top}} \quad (34)$$

$$p(W^1|\sigma_{W^1}^2) = \prod_{i=1}^{n} (2\pi)^{-\frac{l}{2}} |\sigma_{W_l^1}^2 I|^{-\frac{1}{2}} e^{-\frac{1}{2} W_l^1 (\sigma_{W_l^0}^2 I)^{-1} W_l^{1\top}} \quad (35)$$

$$p(UW^0U^\top|N,U,W^0,\sigma_N^2) = \prod_{i=1}^{n}\prod_{j=1}^{n}(2\pi)^{-\frac{1}{2}}(\sigma_N^2)^2 e^{-\frac{1}{2}(\sigma_N^2)^{-1}(U_iW_0U_j^\top - N_{ij})^2} \quad (36)$$

$$p(UW^1U^\top|C,U,W^1,\sigma_C^2) = \prod_{i=1}^{n}\prod_{j=1}^{n}(2\pi)^{-\frac{1}{2}}(\sigma_C^2)^2 e^{-\frac{1}{2}(\sigma_C^2)^{-1}(U_iW_1U_j^\top - C_{ij})^2} \quad (37)$$

Based on Eqs. (33)–(37), the logarithmic function (30) can be written in the form

$$\begin{aligned}\ln p(UW^0U^\top, UW^1U^\top, U, W^0, W^1|N,C,\sigma_U^2,\sigma_{W^0}^2,\sigma_C^2,\sigma_N^2) &= \frac{-1}{2\sigma_N^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(N_{ij}-U_iW^0U_j^\top)^2 + \frac{-1}{2\sigma_C^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(C_{ij}-U_iW^1U_j^\top)^2 \\ &+ \frac{-1}{2\sigma_U^2}\sum_{i=1}^{n}U_iU_i^\top + \frac{-1}{2\sigma_{W^0}^2}\sum_{l=1}^{L}W_l^0W_l^{0\top} + \frac{-1}{2\sigma_{W^1}^2}\sum_{l=1}^{L}W_l^1W_l^{1\top} + M\end{aligned} \quad (38)$$

where M represents the constant term in the equation. And further the maximization problem can be expressed as the minimization problem $\underset{U,W^0,W^1}{argmin}\, E_C$, where E_C is denoted as

$$\begin{aligned}E_C &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(N_{ij}-U_iW^0U_j^\top)^2 + \frac{\lambda_C}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(C_{ij}-U_iW^1U_j^\top)^2 \\ &+ \frac{\lambda_U}{2}\sum_{i=1}^{n}U_iU_i^\top + \frac{\lambda_{W^0}}{2}\sum_{l=1}^{L}W_l^0W_l^{0\top} + \frac{\lambda_{W^1}}{2}\sum_{l=1}^{L}W_l^1W_l^{1\top}\end{aligned} \quad (39)$$

where $\lambda_C = \frac{\sigma_N^2}{\sigma_C^2}$, $\lambda_U = \frac{\sigma_N^2}{\sigma_U^2}$, $\lambda_{W^0} = \frac{\sigma_N^2}{\sigma_{W^0}^2}$, and $\lambda_{W^1} = \frac{\sigma_N^2}{\sigma_{W^1}^2}$.

References

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. *Social Networks, 25*, 211–230.

Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., & Menczer, F. (2012). Friendship prediction and homophily in social media. *ACM Transactions on the Web, 6*, 9.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research, 9*, 1981–2014.

Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2012). Effects of user similarity in social media. In *Proceedings of the fifth ACM international conference on web search and data mining* (pp. 703–712). Seattle, USA.

Backstrom, L., & Leskovec, J. (2011). Supervised random walks: Predicting and recommending links in social networks. In *Proceedings of the fourth ACM international conference on web search and data mining* (pp. 635–644).

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*, 509–512.

Barabâsi, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and Its Applications, 311*, 590–614.

Bhattacharyya, P., Garg, A., & Wu, S. F. (2011). Analysis of user keyword similarity in online social networks. *Social Network Analysis and Mining, 1*, 143–158.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bliss, C. A., Frank, M. R., Danforth, C. M., & Dodds, P. S. (2014). An evolutionary algorithm approach to link prediction in dynamic social networks. *Journal of Computational Science, 5*, 750–764.

Blondel, V. D., Gajardo, A., Heymans, M., Senellart, P., & Van Dooren, P. (2004). A measure of similarity between graph vertices: applications to synonym extraction and web searching. *SIAM Review, 46*, 647–666.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*, 107–117.

Chang, P. C., Galley, M., & Manning, C. D. (2008). Optimizing chinese word segmentation for machine translation performance. Association for Computational Linguistics. Proceedings of the third workshop on statistical machine translation, 224–232,

Chen, H. H., Gou, L., Zhang, X. L., & Giles, C. L. (2012). Discovering missing links in networks using vertex similarity measures. In *Proceedings of the 27th annual ACM symposium on applied computing* (pp. 138–143). Trento, Italy.

Chiang, K., Natarajan, N., Tewari, A., & Dhillon, I. S. (2011). Exploiting longer cycles for link prediction in signed networks. In *Proceedings of the 20th ACM international conference on information and knowledge management* (pp. 1157–1162). Glasgow, UK.

Clauset, A., Moore, C., & Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature, 453*, 98–101.

De Sá, H. R., & Prudêncio, R. B. (2011). Supervised link prediction in weighted networks. In *The 2011 international joint conference on neural networks* (pp. 2281–2288). San Jose, USA.

Dong, Y., Tang, J., Wu, S., Tian, J., Chawla, N. V., Rao, J., & Cao, H. (2012). Link prediction and recommendation across heterogeneous social networks. In *IEEE international conference on data mining* (pp. 181–190). Brussels, Belgium.

Fouss, F., Pirotte, A., Renders, J. M., & Saerens, M. (2007). Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering, 19*, 355–369.

Getoor, L., & Diehl, C. P. (2005). Link mining: A survey. *ACM SIGKDD Explorations Newsletter, 7*, 3–12.

Göbel, F., & Jagers, A. (1974). Random walks on graphs. *Stochastic Processes and Their Applications, 2*, 311–336.

Golub, G. H., & Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik, 14*, 403–420.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology, 143*, 29–36.

Hansen, N., & Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation, 9*, 159.

Hasan, M. A., & Zaki, M. J. (2011). *A survey of link prediction in social networks*. Berlin: Springer US.

He, Y. L., Liu, J. N. K., Hu, Y. X., & Wang, X. Z. (2015). Owa operator based link prediction ensemble for social network. *Expert Systems with Applications, 42*, 21–50.

Holland, P. W., Laskey, K. B., & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks, 5*, 109–137.

Jaccard, P. (1901). Etude de la distribution florale dans une portion des alpes et du jura. *Bulletin De La Societe Vaudoise Des Sciences Naturelles, 37*, 547–579.

Jeh, G., & Widom, J. (2002). Simrank: A measure of structural-context similarity. In *Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 538–543). Edmonton, Canada.

Juszczyszyn, K., Musial, K., & Budka, M. (2011). Link prediction based on subgraph evolution in dynamic social networks. In *IEEE international conference on social computing* (pp. 27–34). Boston, Massachusetts, USA.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika, 18*, 39–43.

Kim, M., & Leskovec, J. (2011). Modeling social networks with node attributes using the multiplicative attribute graph model. *The 27th conference on uncertainty in artificial intelligence*. Barcelona, Spain.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*, 788–791.

Leroy, V., Cambazoglu, B. B., & Bonchi, F. (2010). Cold start link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 393–402). Washington, USA.

Leskovec, J., Huttenlocher, D., & Kleinberg, J. (2010). Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on world wide web* (pp. 641–650). Raleigh, USA.

Li, D., Fu, B., Wang, Y., Lu, G., Berezin, Y., Stanley, H. E., & Havlin, S. (2015). Percolation transition in dynamical traffic network with evolving critical bottlenecks. *Proceedings of the National Academy of Sciences, 112*, 669–672.

Liben-Nowell, D., & Kleinberg, J. (2007). The link prediction problem for social networks. *Journal of the American Society for Information Science and Technology, 58*, 1019–1031.

Lichtenwalter, R. N., Lussier, J. T., & Chawla, N. V. (2010). New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 243–252).

Lorrain, F., & White, H. C. (1971). Structural equivalence of individuals in social networks. *The Journal of Mathematical Sociology, 1*, 49–80.

Lü, L., Jin, C., & Zhou, T. (2009). Similarity index based on local paths for link prediction of complex networks. *Physical Review E, 80*, 046122.

Lü, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and Its Applications, 390*, 1150–1170.

Lu, Y., Guo, Y., & Korhonen, A. (2017). Link prediction in drug-target interactions network using similarity indices. *Bmc Bioinformatics, 18*, 39.

Luo, P., Wu, C., & Li, Y. (2017). Link prediction measures considering different neighbors effects and application in social networks. *International Journal of Modern Physics C, 28*, 1750033.

L, L., & Zhou, T. (2009). Role of weak ties in link prediction of complex networks. In *The first ACM international workshop on complex networks meet information and knowledge management* (pp. 55–58).

Martnez, V., Berzal, F., & Cubero, J. C. (2016). A survey of link prediction in complex networks. *ACM Computing Surveys, 49*, 69.

Menon, A. K., & Elkan, C. (2011). Link prediction via matrix factorization. In *Machine learning and knowledge discovery in databases* (pp. 437–452).

Miller, K., Jordan, M. I., & Griffiths, T. L. (2009). Nonparametric latent feature models for link prediction. In *Advances in neural information processing systems* (pp. 1276–1284). Vancouver, Canada.

Moradabadi, B., & Meybodi, M. R. (2017). Link prediction in fuzzy social networks using distributed learning automata. *Applied Intelligence, 47*, 837–849.

Munasinghe, L., & Ichise, R. (2012). Time score: A new feature for link prediction in social networks. *IEICE Transactions on Information & Systems, 95*, 821–828.
Murata, T., & Moriyasu, S. (2007). Link prediction of social networks based on weighted proximity measures. In *IEEE/WIC/ACM international conference on web intelligence* (pp. 85–88).
Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical Review E, 64*, 025102.
Nowicki, K., & Snijders, T. A. B. (2001). Estimation and prediction for stochastic block structures. *Journal of the American Statistical Association, 96*, 1077–1087.
Palla, K., Knowles, D., & Ghahramani, Z. (2012). An infinite latent attribute model for network data. In *Proceedings of the 29th international conference on machine learnin*. Edinburgh, Scotland.
Papadimitriou, A., Symeonidis, P., & Manolopoulos, Y. (2012). Fast and accurate link prediction in social networking systems. *Journal of Systems and Software, 85*, 2119–2132.
Pastor-Satorras, R., Castellano, C., Van Mieghem, P., & Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of Modern Physics, 87*, 925.
Pavlov, M., & Ichise, R. (2007). Finding experts by link prediction in co-authorship networks. In *FEWS: 290* (pp. 42–55).
Pecli, A., Cavalcanti, M. C., & Goldschmidt, R. (2017). Automatic feature selection for supervised learning in link prediction applications: A comparative study. *Knowledge & Information Systems*, 1–37. doi:10.1007/s10115-017-1121-6.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research, 12*, 2825–2830.
Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabási, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science, 297*, 1551–1555.
Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2009). Bpr: Bayesian personalized ranking from implicit feedback. In *Conference on uncertainty in artificial intelligence* (pp. 452–461).
Romero, D. M., & Kleinberg, J. (2010). The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. *International conference on weblogs and social media*.
Rowe, M., Stankovic, M., & Alani, H. (2012). Who will follow whom? exploiting semantics for link prediction in attention-information networks. In *The semantic web–ISWC* (pp. 476–491).
Scellato, S., Noulas, A., & Mascolo, C. (2011). Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1046–1054). San Diego, California, USA.
Soares, P. R., & Prudêncio, R. B. (2013). Proximity measures for link prediction based on temporal events. *Expert Systems with Applications, 40*, 6652–6660.
Song, D., & Meyer, D. A. (2015a). Link sign prediction and ranking in signed directed social networks. *Social Network Analysis & Mining, 5*, 1–14.
Song, D., & Meyer, D. A. (2015b). Recommending positive links in signed social networks by optimizing a generalized auc. In *Proceedings of the twenty-ninth AAAI conference on artificial intelligence* (pp. 25–30).
Song, D., Meyer, D. A., & Min, M. R. (2014). Fast nonnegative matrix factorization with rank-one admm. *NIPS 2014 workshop on optimization for machine learning*.
Song, D., Meyer, D. A., & Tao, D. (2015a). Efficient latent link recommendation in signed networks. In *ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1105–1114).
Song, D., Meyer, D. A., & Tao, D. (2015b). Top-k link recommendation in social networks. In *IEEE international conference on data mining* (pp. 389–398).
Sproat, R., Gale, W., Shih, C., & Chang, N. (1996). A stochastic finite-state word-segmentation algorithm for chinese. *Computational Linguistics, 22*, 377–404.
Valverde-Rebaza, J., & Lopes, A. D. A. (2013a). Exploiting behaviors of communities of twitter users for link prediction. *Social Network Analysis & Mining, 3*, 1063–1074.
Valverde-Rebaza, J., & Lopes, A. D. A. (2013b). Structural link prediction using community information on twitter. In *Fourth international conference on computational aspects of social networks* (pp. 132–137).
Wan, S., Lan, Y., Guo, J., Fan, C., & Cheng, X. (2013). Informational friend recommendation in social media. In *Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval* (pp. 1045–1048). Dublin, Ireland.
Wang, P., Xu, B., Wu, Y., & Zhou, X. (2014). Link prediction in social networks: The state-of-the-art. *Science China Information Sciences, 58*, 1–38.
Wang, Z., Liang, J., Li, R., & Qian, Y. (2016). An approach to cold-start link prediction: Establishing connections between non-topological and topological information. *IEEE Transactions on Knowledge and Data Engineering, 28*, 2857–2870.
Wang, Z., Liao, J., Cao, Q., & Qi, H. (2015). Friendbook: A semantic-based friend recommendation system for social networks. *IEEE Transactions on Mobile Computing, 14*, 538–551.
Wohlfarth, T., & Ichise, R. (2008). Semantic and event-based approach for link prediction. In *Practical aspects of knowledge management* (pp. 50–61). Yokohama, Japan.
Xie, F., Chen, Z., Shang, J., Feng, X., Huang, W., & Li, J. (2015). A link prediction approach for item recommendation with complex number. *Knowledge-Based Systems, 81*, 148–158.
Xie, X. (2010). Potential friend recommendation in online social network. In *Green computing and communications, 2010 IEEE/ACM international conference on & international conference on cyber, physical and social computing* (pp. 831–835). Hangzhou, China.
Yang, S. H., Long, B., Smola, A., Sadagopan, N., Zheng, Z., & Zha, H. (2011). Like like alike:joint friendship and interest propagation in social networks. In *International conference on world wide web* (pp. 537–546).
Yin, D., Hong, L., & Davison, B. D. (2011). Structural link analysis and prediction in microblogs. In *the 20th ACM international conference on information and knowledge management* (pp. 1163–1168).
Zhai, S., & Zhang, Z. (2015). Dropout training of matrix factorization and autoencoder for link prediction in sparse graphs. In *Proceedings of the SIAM international conference on data mining* (pp. 451–459).
Zhang, J., Fang, Z., Chen, W., & Tang, J. (2015a). Diffusion of “following” links in microblogging networks. *IEEE Transactions on Knowledge and Data Engineering, 27*, 2093–2106.
Zhang, J., Liu, B., Tang, J., Chen, T., & Li, J. (2013a). Social influence locality for modeling retweeting behaviors. In *Proceedings of the twenty-third international joint conference on artificial intelligence* (pp. 2761–2767). Beijing, China.
Zhang, M., Hu, H., He, Z., Gao, L., & Sun, L. (2015b). Efficient link-based similarity search in web networks. *Expert Systems with Applications, 42*, 8868–8880.
Zhang, Q. M., L, L., Wang, W. Q., Zhu, Y. X., & Zhou, T. (2013b). Potential theory for directed networks. *Plos One, 8*, e55437.
Zhu, J. (2012). Max-margin nonparametric latent feature models for link prediction. In *Proceedings of the 29th international conference on machine learning*. Edinburgh, Scotland.
Zhu, S., Yu, K., Chi, Y., & Gong, Y. (2007). Combining content and link for classification using matrix factorization. In *Proceedings of the international ACM SIGIR conference on research and development in information retrieval* (pp. 487–494).
Zhu, Y. X., Lü, L., Zhang, Q., & Zhou, T. (2012). Uncovering missing links with cold ends. *Physica A: Statistical Mechanics and Its Applications, 391*, 5769–5778.