Knowledge-Based Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A fusion probability matrix factorization framework for link prediction

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, Taiyuan, Shanxi 030006, China

ARTICLE INFO

Keywords:
Network data analysis
Probability matrix factorization
Link prediction
Fusion model

ABSTRACT

Link prediction is a fundamental research problem in network data analysis. Networks usually contain rich node-to-node topological metrics and their effective use is crucial to solve the link prediction problem. Despite significant advances, the existing metric-based link prediction methods usually only consider one single topological metric and thus show some limitations in different types of networks; the existing matrix factorization-based models mainly focus on modeling the adjacent matrix of a network, and this is hard to ensure the modeling of those topological metrics that can play an important role in link prediction. This study develops effective approaches by fusing the adjacent matrix and some key topological metrics in a unified probability matrix factorization framework. In these approaches, we consider not only the symmetric metrics but also the asymmetric metrics which are usually not taken into consideration in the related work. In our probability matrix factorization framework, we first present two fusion models by fusing two kinds of metrics respectively, and based on the fusion models, we put forward the final fusion models which fuse the two kinds of metrics simultaneously. To verify the performance of all the fusion models, we conduct the experiments with six directed networks and six undirected ones, and the extensive experiments show that the proposed models provide impressive predicting performance for link prediction.

1. Introduction

Link prediction is a fundamental and important problem in network data analysis [1]. The solution to the problem is essential to explain the reason of network structure generation, to help us explore the law of network evolution [2], and to understand the mechanism of complex systems [3,4]. Furthermore, it is also of great significance for many applications, such as finding friends in social networks [5], recommending items in user-item networks [6], finding experts in academic networks [7], and discovering unknown interactions in protein-protein networks [8].

Research on link prediction has draw increasing attention in recent years. Many methods have been proposed by researchers from physics, biology, sociology, and computer science [9–14], including the metric-based methods, the classification-based methods, the probabilistic graph model (PGM)-based methods and the matrix factorization (MF)-based methods. As one kind of important link prediction methods, MF-based methods solve the link prediction problem mainly through focusing on modeling the low-rank approximation of the adjacent matrix of a network. Some existing work [15,16] has shown the advantages of MF-based methods in solving link prediction problem such as its robustness to the networks from different domains and its scalability to large datasets.

Though powerful, existing MF-based methods still have some problems that could limit their applicability and prediction accuracy. The core of MF is to get the low-rank approximation of the adjacent matrix of a network, but compared with the network itself, the presented information of the adjacent matrix is insufficient. This is because the adjacent matrix only present the observed links of a network, but actually a network still contains rich topological metrics like the common neighbors between nodes and the path length between nodes. No evidence indicating that the existing MF-based models can give dual attention to modeling the observed network links and those topological metrics which can play an important role in link prediction.

Based on the above considerations, we have the following motivations:

- Whether a MF-based model can be built to fuse the adjacent matrix and the topological metrics between nodes in a network?
- Whether a MF-based model should consider both the symmetric metrics and the asymmetric metrics between nodes in a network?
- How can we take the two factors into account in one MF-based model if both the symmetric metrics and the asymmetric metrics are related to the links formation in a network.

* 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.knosys.2018.06.005
Received 27 November 2017; Received in revised form 4 June 2018; Accepted 7 June 2018

Therefore, the objective of this study is to develop fusion models which fuse the adjacent matrix with some topological metrics together in one unified probability matrix factorization framework. In the framework, we first present two fusion models by fusing the symmetric metrics and the asymmetric metrics respectively, and based on the fusion models, we put forward the final fusion models which fuse the two kinds of metrics simultaneously.

This paper makes the following contributions:

- We propose fusion models to fuse two sides of network information, i.e. the adjacent matrix and some key topological metrics, in one unified probabilistic matrix factorization framework.
- Our fusion models consider not only the symmetric metrics but also the asymmetric metrics which are usually not taken into consideration in the related work.
- We conduct experimental evaluations with various directed and undirected network datasets, and our models get impressive predicting performance for link prediction.

The rest of the paper is organized as follows: Section 2 introduces the related work; Section 3 presents the building of the fusion models in the probabilistic matrix factorization framework, where the symmetric metrics and asymmetric metrics are fused respectively, and a final fusion model is proposed to fuse the two kinds of metrics simultaneously. In Section 4, we conduct a series of experiments to evaluate the proposed methods on various directed and undirected network datasets. Finally we conclude our work in Section 5.

2. Related work

For link prediction, there have been several excellent surveys [17–20] from different research perspectives. Liben-Nowell and Kleinberg [19] provided useful information for the prediction problem, especially some classical prediction measures based on topological information of networks. Lü and Zhou [21] summarized recent progress about link prediction algorithms, emphasizing the contributions from physical perspectives and approaches. Wang et al. [17] investigated the link prediction from the perspective of computer science, and systematically summarized all typical work on the link prediction in social networks. Martínez et al. [22] discussed a large number of proposed techniques focusing on undirected and unweighed networks. In this section, we will first give a brief summary of the related work following the researching line of MF-based methods which are related to the methodology of this paper, and after that we will provide a brief review of the work done in the area of link prediction.

2.1. Matrix factorization based link prediction

Matrix factorization is a type of technique to get the low-rank approximation (LAR) and global information of the adjacent matrix of a network. The classical matrix factorization methods such as singular value decomposition (SVD) [23], non-negative matrix factorization (NNMF) [24] and probabilistic matrix factorization (PMF) [25] can be directly used for solving the link prediction problem. Liben-Nowell and Kleinberg [19] investigated various types of link prediction methods, and among them the LAR-based link prediction methods are implemented by using SVD. Chen et al. [26] put forward a link prediction algorithm based on NNMF. Zhu et al. [27] proposed a scalable temporal link prediction model via NNMF. Yang et al. [28] combined link prediction method by convex NNMF with block detection to predict potential links using both of global and local information. In a word, the extensive experiments showed that the classical MF-based methods were effective in solving the link prediction problem.

In addition to the classical MF-based methods, some factorization methods also have been designed according to the characteristics of the link prediction problem to improve the prediction performance. Menon and Elkan [15] proposed a MF-based method to address the class imbalance problem by directly optimizing for a ranking loss, which is optimized with stochastic gradient descent and scales to large graphs. Zhai and Zhang [16] attempted to solve the link prediction problem by combining MF and Autoencoder (AE), and they utilized dropout to train both the MF and AE parts and the results showed that it could significantly prevent overfitting by acting as an adaptive regularization. Song et al. [29] proposed a rank-one alternating direction method of multiplier (ADMM) for nonnegative matrix factorization, and their experiment results demonstrated that rank-one ADMM is more effective than multiplicative update rule, alternating least square, and traditional ADMM.

Despite these significant advances, current state-of-the-art MF-based models mainly focus on modeling the adjacent matrix of a network, and this is hard to ensure the modeling of those topological metrics that can play an important role in link prediction. Intuitively, it is of certain potential to improve the performance of link prediction if the MF-based methods can give dual attention to modeling the adjacent matrix and some key topological metrics. Therefore, this paper intends to deal with the problem, which is the starting point of the study.

2.2. The other link prediction methods

Apart from the MF-based methods, the metric-based methods, the classification-based methods, and the PGM-based methods are also the mainstream methods in link prediction.

The metric-based methods address the link prediction problem by measuring the similarity between nodes, such as the neighbors-based metrics [30–37], path-based metrics [38–41], and random walk-based metrics [42–46]. David Liben-Nowell and Kleinberg [19] tested several topological metrics on social collaboration networks, and the results showed that the Katz [41] metric and its variants performed consistently well, and that some of the very simple metrics including common neighbors and the Adamic-Adar metric [33] also performed surprisingly well. Zhou et al. [31] compared a number of topological metrics on disparate networks which included the protein-protein interaction network, the electronic grid, the Internet, and the US airport network. The extensive experimental results showed that the Resource Allocation [33] metric performed best, while common neighbors and Adamic–Adar metric [33] had the second-best performance. Also, other topology-based metrics were proposed to solve the link-prediction problem [38–40,42,47–49]. Despite those significant advances, the effectiveness of the metrics depends on the domain, the specific network, and the available information.

The classification-based methods treat link prediction as a binary classification problem. In a classification-based link prediction model, the features are defined on each pair of nodes, and these features 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 [50–52]. Except for the topological features, the non-topological features (such as users' location, interests, and educational backgrounds) are often selected to improve the classification-based link prediction models [53–55]. In the classification-based methods, it is still a challenge to predict links because the class imbalance problem can be difficult to deal with and most models are prone to yield biased results.

The PGM-based methods solve the link prediction problem by building a statistical network model. The hierarchical network model [56] models a network as a hierarchical random graph and the linking possibility between nodes can be calculated by the probability expectation. Stochastic block models [57,58] assume that the network nodes can be partitioned into some blocks, and that the linking probability between any two nodes depends on which block the nodes belong to. Latent-feature models [59–62] are kinds of probabilistic generative model, where the nodes' latent-features and the edges in a

network are all generated based on some distribution, and the linking possibility can be obtained by using the model parameters which have been estimated by a maximum likelihood estimation method. Although the existing PGM-based modes provide a deep insight into network structure, the algorithms usually have high complexity and are not suitable for large networks.

3. Fusion probability matrix factorization models

In this section, we focus on building the fusion probability matrix factorization (FPMF) models which can fuse the adjacent matrix with some key topological metrics in a unified probability matrix factorization framework. It should be noted that the topological metrics we aim to fuse in our models are divided into two parts: the symmetric metrics and the asymmetric metrics. The symmetric metrics mean that the metrics from node u_i to node u_j is equal to that from node u_j to node u_i. However, asymmetric metrics need to distinguish the direction between node u_i and node u_j.

Since the strategies for fusing the symmetric metrics and the asymmetric metrics are different in the probability matrix factorization framework, we will first present two fusion models by fusing two kinds of metrics respectively, and based on the fusion models, we put forward the final fusion models which fuse the two kinds of metrics simultaneously in the upcoming sections.

For convenience, we list out the mainly-used notations of the document in Table 1.

3.1. FPMF model by fusing symmetric metrics

The FPMF model we aim to build in this section is to fuse the adjacent matrix with the symmetric metrics in a network. The adjacent matrix is denoted as $A_{n\times n}$, the symmetric metrics are denoted as a symmetric matrix $S_{n\times n}$, and each of the matrix element S_{ij} represents a symmetric metric between node u_i and node u_j.

In an undirected network, many existing metrics (such as the neighbor-based metrics and the path-based metrics) can be directly used to measure the topological metric between network nodes. Here we use the typical metrics including Common Neighbors [36], Jaccard Coefficient [37] and Preferential Attachment [35] as the symmetric measurements in an undirected network. However, in a directed network, more variations of the measurements will appear because of the link direction. For measuring the node-to-node relations in a directed network, several symmetric metrics are defined by taking the existing metrics [35,37,55] for reference.

To describe the symmetric metrics clearly, we give the following standard notations. In an undirected network, $\Gamma(u_i)$ denotes the set of neighbors of node u_i, and $|\Gamma(u_i)|$ denotes the number of elements of set $\Gamma(u_i)$. In a directed network, $\Gamma^+(u_i)$ denotes the set of neighbors pointing to node u_i, $\Gamma^-(u_i)$ denotes the set of neighbors directed at node u_i, $|\Gamma^+(u_i)|$ denotes the number of elements of set $\Gamma^+(u_i)$, and $|\Gamma^-(u_i)|$ denotes the number of elements of set $\Gamma^-(u_i)$.

Table 1
Notations.

Symbol	Explanation
$n\in R$	The number of network nodes.
$A\in R^{n\times n}$	The adjacent matrix of a network.
$S\in R^{n\times n}$	The symmetric metric matrix.
$C\in R^{n\times n}$	The asymmetric metric matrix.
$U\in R^{n\times L}$	The latent-feature matrix of a network nodes.
$U_i\in R^{1\times L}$	The latent-feature vector of node u_i.
g_A	The binary relation function of A.
g_S	The binary relation function of S.
g_C	The binary relation function of C.
$W^A\in R^{L\times L}$	The parameter matrix of the function g_A.
$W^C\in R^{L\times L}$	The parameter matrix of the function g_C.

Table 2
Symmetric metrics.

Number	formula	References
①	$\lvert\Gamma(u_i)\cap\Gamma(u_j)\rvert$	[36]
②	$\frac{\lvert\Gamma(u_i)\cap\Gamma(u_j)\rvert}{\lvert\Gamma(u_i)\cup\Gamma(u_j)\rvert}$	[37]
③	$\lvert\Gamma(u_i)\rvert\times\lvert\Gamma(u_j)\rvert$	[35]
④	$\lvert\Gamma^-(u_i)\cap\Gamma^-(u_j)\rvert$	[55]
⑤	$\frac{\lvert\Gamma^-(u_i)\cap\Gamma^-(u_j)\rvert}{\lvert\Gamma^-(u_i)\cup\Gamma^-(u_j)\rvert}$	[37,55]
⑥	$\lvert\Gamma^+(u_i)\cap\Gamma^+(u_j)\rvert$	[55]
⑦	$\frac{\lvert\Gamma^+(u_i)\cap\Gamma^+(u_j)\rvert}{\lvert\Gamma^+(u_i)\cup\Gamma^+(u_j)\rvert}$	[37,55]
⑧	$\lvert(\Gamma^-(u_i)\cap\Gamma^+(u_i))\cap(\Gamma^-(u_j)\cap\Gamma^+(u_j))\rvert$	[55]
⑨	$\frac{\lvert(\Gamma^-(u_i)\cap\Gamma^+(u_i))\cap(\Gamma^-(u_j)\cap\Gamma^+(u_j))\rvert}{\lvert(\Gamma^-(u_i)\cap\Gamma^+(u_i))\cup(\Gamma^-(u_j)\cap\Gamma^+(u_j))\rvert}$	[37,55]
⑩	$\lvert(\Gamma^-(u_i)\cup\Gamma^+(u_i))\cap(\Gamma^-(u_j)\cup\Gamma^+(u_j))\rvert$	[55]
⑪	$\frac{\lvert(\Gamma^-(u_i)\cup\Gamma^+(u_i))\cap(\Gamma^-(u_j)\cup\Gamma^+(u_j))\rvert}{\lvert(\Gamma^-(u_i)\cup\Gamma^+(u_i))\cup(\Gamma^-(u_j)\cup\Gamma^+(u_j))\rvert}$	[37,55]
⑫	$\lvert\Gamma^-(u_i)\rvert\times\lvert\Gamma^-(u_j)\rvert$	[35,55]
⑬	$\lvert\Gamma^+(u_i)\rvert\times\lvert\Gamma^+(u_j)\rvert$	[35,55]
⑭	$\lvert(\Gamma^-(u_i)\cap\Gamma^+(u_i))\rvert\times\lvert(\Gamma^-(u_j)\cap\Gamma^+(u_j))\rvert$	[35,55]
⑮	$\lvert(\Gamma^-(u_i)\cup\Gamma^+(u_i))\rvert\times\lvert(\Gamma^-(u_j)\cup\Gamma^+(u_j))\rvert$	[35,55]

Table 2 presents 15 symmetric metrics. Among them, the metrics ①, ② and ③ are the neighbor-based metrics for an undirected network, and the other 12 metrics are defined for a directed network. The metrics ④, ⑥, ⑧ and ⑩ are based on [55], and ⑤, ⑦ ⑨ and ⑪ are the extension of four normalized metrics. As the formulas in the Table 2 show, the metrics ④, ⑥, ⑧ and ⑩ focus on the number of different types of common neighbors between node u_i and node u_j, and the metrics ⑤, ⑦, ⑨ and ⑪ concentrate on the proportion of different types of common neighbors between u_i and u_j. Besides, we also extend the Preferential Attachment [35] metric into four metrics in a directed network, i.e. ⑫, ⑬, ⑭ and ⑮. By using any one of the symmetric metrics list in Table 2, we construct a symmetric matrix $S=|S_{ij}|_{n\times n}$ by computing the metrics between all of the network nodes.

Given the adjacent matrix $A_{n\times n}$ and the symmetric matrix $S_{n\times n}$ of a network, the FPMF model aims to fuse them in a unified probability matrix factorization framework. More specifically, the FPMF model is based on the following three assumptions:

- In a network, each node is represented as an L-dimension latent-feature vector $U_i\in R^{1\times L}$ ($i\in\{1,\ldots,n\}$), and $U\in R^{n\times L}$ is the $n\times L$ latent-feature matrix of the n network nodes. We suppose that U_i obeys an L-dimension Gaussian distribution with mean [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

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

- For modeling the observed network links, i.e. the adjacent matrix A, a binary relation function is defined as

$$g_A(U_i,U_j)=U_i W^A U_j^{\mathrm{T}} \tag{2}$$

where $W^A\in R^{L\times L}$ is the parameter matrix of the relation function g_A. The reason we introduce the parameter matrix $W^A\in R^{L\times L}$ in the binary relation function g_A is that $U_i W^A U_j^{\mathrm{T}}$ represents a generalized measurement from node u_i to node u_j and need to be learned from

[1] $\vec{0}$ is a zero vector.
[2] I is a identity matrix.

the specific network data. If there is a symmetry network, the learned linking parameter matrix W^A will be a symmetry parameter matrix or vice versa. If $W^A = I$, the binary relation function g_A just corresponds to the inner product between node u_i and node u_j in Euclid space.

We assume that the value of the function $g_A(U_i, U_j)$ obeys a Gaussian distribution with mean 1 and variance σ_A^2 if there is a link from node u_i to node u_j. Otherwise, the value of the function $g_A(U_i, U_j)$ obeys the Gaussian distribution with mean 0 and variance σ_A^2 if there is not a link from node u_i to node u_j. Formally, $U_i W^A U_j^\top \sim N(A_{ij}, \sigma_A^2)$. As for the parameter matrix W^A, we also suppose that W_l^A obeys the L-dimension Gaussian distribution with and covariance matrix $\sigma_{W^A}^2 I$, i.e. $W_l^A \sim N\left(\vec{0}, \sigma_{W^A}^2 I\right)$. Based on these assumptions, the probability density of the matrix W^A and the value of $UW^A U^\top$ can be denoted as

$$p\left(W^A \middle| \sigma_{W^A}^2\right) = \prod_{l=1}^{L} N\left(W_l^A \middle| \vec{0}, \sigma_{W^A}^2 I\right) \tag{3}$$

$$p(UW^A U^\top | A, U, W^A, \sigma_A^2) = \prod_{i=1}^{N}\prod_{j=1}^{N} N(U_i W^A U_j^\top | A_{ij}, \sigma_A^2) \tag{4}$$

- For the modeling of the symmetric matrix S, we introduce a symmetric binary relation function

$$g_S(U_i, U_j) = U_i U_j^\top \tag{5}$$

where the inner product $U_i U_j^\top$ can be seen as the symmetric similarity metric between node u_i and node u_j, and $U_i U_j^\top$ just corresponds to the defined symmetric metrics (see Table 2).

Given the nodes' latent-feature matrix U, we assume that the value of the function $g_S(U_i, U_j)$ between node u_i and node u_j obeys the Gaussian distribution with mean S_{ij} and variance σ_S^2. The probability density of the value of $UU^\top$ can be denoted as

$$p(UU^\top | S, U, \sigma_S^2) = \prod_{i=1}^{N}\prod_{j=1}^{N} N(U_i U_j^\top | S_{ij}, \sigma_S^2) \tag{6}$$

Fig. 1 shows the relations among the matrices and parameters of the FPMF model which fuse the symmetric matrix S. By utilizing the product rule of the directed probabilistic graphical model, the joint probability density distribution over the variables U, W^A, $UW^A U^\top$ and $UU^\top$ can be represented as

$$p\left(UW^A U^\top, U, W^A \middle| A, S, \sigma_U^2, \sigma_{W^A}^2, \sigma_S^2, \sigma_A^2\right) = p(U|\sigma_U^2)\, p\left(W^A \middle| \sigma_{W^A}^2\right) p(UU^\top | S, U, \sigma_S^2)\, p(UW^A U^\top | A, U, W^A, \sigma_A^2) \tag{7}$$

The goal of the fusion model is to learn the nodes' latent-feature representation U and the matrix parameter W^A of the adjacent matrix A by maximizing the joint probability density distribution. The problem can be deduced as the optimization problem $\underset{U,W^A}{argmin} E_S$, and the objective function E_S is denoted as

$$E_S = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(A_{ij} - U_i W^A U_j^\top)^2 + \frac{\lambda_S}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(S_{ij} - U_i U_j^\top)^2 + \frac{\lambda_U}{2}\sum_{i=1}^{n} U_i U_i^\top + \frac{\lambda_{W^A}}{2}\sum_{l=1}^{L} W_l^A W_l^{A\top} \tag{8}$$

where $\lambda_S = \frac{\sigma_A^2}{\sigma_S^2}$, $\lambda_U = \frac{\sigma_A^2}{\sigma_U^2}$, and $\lambda_{W^A} = \frac{\sigma_A^2}{\sigma_{W^A}^2}$. Also, the objective function (8) can be more briefly rewritten as

$$E_S = \frac{1}{2}\left\|A - UW^A U^\top\right\|_F^2 + \frac{\lambda_S}{2}\left\|S - UU^\top\right\|_F^2 + \frac{\lambda_U}{2}\left\|U\right\|_F^2 + \frac{\lambda_{W^A}}{2}\left\|W^A\right\|_F^2 \tag{9}$$

$\underset{U,W^A}{argmin} E_S$ can be solved by using gradient methods, and the Eq. (10) shows the gradients of the function E_S against variables U and W^A.

$$\begin{aligned}\frac{\partial E_S}{\partial U} &= (UW^{A\top}U^\top UW^A + UW^{A\top}U^\top UW^{A\top} - A^\top UW^A - AUW^{A\top}) + \lambda_S(2UU^\top U - SU - S^\top U) + \lambda_U U \\ \frac{\partial E_S}{\partial W^A} &= U^\top UW^A U^\top U - U^\top AU + \lambda_{W^A} W^A\end{aligned} \tag{10}$$

Intuitively, in the fusion model, the adjacent matrix A is approximated to $UW^A U^\top$, and the symmetric matrix S is approximated to $UU^\top$.

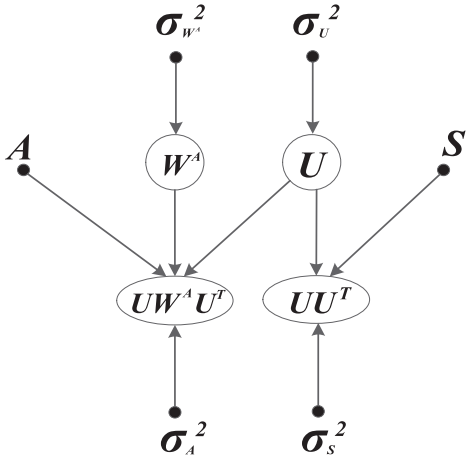


Fig. 1. A directed probabilistic graphical model representing the relations among the matrices and parameters in the FPMF model which fuse the symmetric matrix S.

3.2. FPMF model by fusing asymmetric metrics

In this section, we aim to build the FPMF model for fusing the adjacent matrix A with the node-to-node asymmetric metrics in a unified probability matrix factorization framework. The asymmetric metrics is denoted as an $n \times n$ asymmetric matrix C, and each element C_{ij} of the asymmetric matrix C corresponds to the asymmetric metric from node u_i to node u_j.

Although many symmetric metrics have been designed to measure the topological relations between network nodes, little attention has been paid to accounting for measuring the asymmetric relations. Here we define one asymmetric metric for an undirected network and four asymmetric metrics for a directed network (as shown in Table 3). Among them, the metric ① is defined for the undirected network, and it represents the proportion of the common neighbors of u_i and u_j in the neighbors of u_i. In a directed network, the metric ② represents the proportion of common neighbors of u_i and u_j pointing to the other nodes in the neighbors of u_i pointing to the other nodes, and the metric ③ represents the proportion of common neighbors of u_i and u_j directed at the other nodes in the neighbors of u_i directed at the other nodes. The metric ④ denotes the proportion of common friends of u_i and u_j in the friends of u_i, and the metric ⑤ denotes the proportion of common

Table 3
Asymmetric metrics.

Number	formula
①	$\frac{\lvert\Gamma(u_i)\cap\Gamma(u_j)\rvert}{\lvert\Gamma(u_i)\rvert}$
②	$\frac{\lvert\Gamma^-(u_i)\cap\Gamma^-(u_j)\rvert}{\lvert\Gamma^-(u_i)\rvert}$
③	$\frac{\lvert\Gamma^+(u_i)\cap\Gamma^+(u_j)\rvert}{\lvert\Gamma^+(u_i)\rvert}$
④	$\frac{\lvert(\Gamma^-(u_i)\cap\Gamma^+(u_i))\cap(\Gamma^-(u_j)\cap\Gamma^+(u_j))\rvert}{\lvert(\Gamma^-(u_i)\cap\Gamma^+(u_i))\rvert}$
⑤	$\frac{\lvert(\Gamma^-(u_i)\cup\Gamma^+(u_i))\cap(\Gamma^-(u_j)\cup\Gamma^+(u_j))\rvert}{\lvert\Gamma^-(u_i)\cup\Gamma^+(u_i)\rvert}$

neighbors of u_i and u_j in the neighbors of the u_i.

Given the observed network matrix $A_{n\times n}$ and the asymmetric matrix $C_{n\times n}$, the FPMF model aims to fuse the two kinds of information in a unified probability matrix factorization framework. Similar to the fusion model in Section 3.1, the first two assumptions are the same, but the third assumption is different because it aims to fuse the asymmetric node-to-node metrics.

- For modeling the asymmetric matrix C, we also define a binary relation function

$$g_C(U_i, U_j) = U_i W^C U_j^\top \tag{11}$$

where W_l^C obeys an L-dimension Gaussian distribution with mean and covariance matrix $\sigma_{W^C}^2 I$, i.e. $W_l^C \sim N\left(\vec{0}, \sigma_{W^C}^2 I\right)$. Here we also introduce a matrix parameter $W^C \in R^{L\times L}$, which makes the binary function g_C represents a generalized relation measurement from node u_i to node u_j, and $U_i W^C U_j^\top$ corresponds to the asymmetry matrix C.
Given the nodes' latent-feature matrix U and matrix parameter W^C, we assume that the value of the binary function $g_C(U_i, U_j)$ from node u_i to node u_j obeys the Gaussian distribution with mean C_{ij} and variance σ_C^2. Therefore, the probability density of W^C and the value of $UW^CU^\top$ can be denoted as

$$p\left(W^C\middle|\sigma_{W^C}^2\right) = \prod_{l=1}^{L} N\left(W_l^C\middle|\vec{0}, \sigma_{W^C}^2 I\right) \tag{12}$$

$$p(UW^CU^\top|C, U, W^C, \sigma_C^2) = \prod_{i=1}^{N}\prod_{j=1}^{N} N(U_i W^C U_j^\top|C_{ij}, \sigma_C^2) \tag{13}$$

Fig. 2 shows the relations among the matrices and parameters in the FPMF model. By utilizing the product rule of the directed probabilistic graphical model, the joint probability density distribution over the variables U, W^A, W^C, $UW^AU^\top$ and $UW^CU^\top$ can be represented as

$$\begin{aligned} &p\left(UW^AU^T, UW^CU^\top, U, W^A, W^C\middle|A, C, \sigma_U^2, \sigma_{W^A}^2, \sigma_S^2, \sigma_A^2\right) = \\ &p(U|\sigma_U^2)p\left(W^A\middle|\sigma_{W^A}^2\right)p\left(W^C\middle|\sigma_{W^C}^2\right) \\ &p(UW^AU^\top|A, U, W^A, \sigma_A^2) \\ &p(UW^CU^\top|C, U, W^C, \sigma_C^2) \end{aligned} \tag{14}$$

The goal of the fusion model is to learn the nodes' latent feature representation U, the parameters W^A of the adjacent matrix A and the parameters W^C of the asymmetric matrix C by maximizing the joint probability density distribution. The problem can be deduced as the optimization problem $\underset{U,W^A,W^C}{argmin} E_C$, and the objective function E_C is denoted as

$$\begin{aligned} E_C = &\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(A_{ij} - U_i W^A U_j^\top)^2 + \frac{\lambda_C}{2}\sum_{i=1}^{N}\sum_{j=1}^{n}(C_{ij} - U_i W^C U_j^\top)^2 \\ &+ \frac{\lambda_U}{2}\sum_{i=1}^{n} U_i U_i^\top + \frac{\lambda_{W^A}}{2}\sum_{l=1}^{L} W_l^A W_l^{A\top} + \frac{\lambda_{W^C}}{2}\sum_{l=1}^{L} W_l^C W_l^{C\top} \end{aligned} \tag{15}$$

where $\lambda_C = \frac{\sigma_A^2}{\sigma_C^2}$, $\lambda_U = \frac{\sigma_A^2}{\sigma_U^2}$, $\lambda_{W^A} = \frac{\sigma_A^2}{\sigma_{W^A}^2}$, and $\lambda_{W^C} = \frac{\sigma_A^2}{\sigma_{W^C}^2}$. Also, the objective function (15) can be more briefly rewritten as

$$\begin{aligned} E_C = &\frac{1}{2}\left\|A - UW^AU^\top\right\|_F^2 + \frac{\lambda_C}{2}\left\|C - UW^CU^\top\right\|_F^2 \\ &+ \frac{\lambda_U}{2}\left\|U\right\|_F^2 + \frac{\lambda_{W^A}}{2}\left\|W^A\right\|_F^2 + \frac{\lambda_{W^C}}{2}\left\|W^C\right\|_F^2 \end{aligned} \tag{16}$$

Intuitively, the asymmetric matrix C is approximated to $UW^CU^\top$ in the model. $\underset{U,W^A,W^C}{argmin} E_C$ can be solved by using gradient methods, and the Eq. (17) shows the gradients of function E_C against variables U, W^A and W^C.

$$\begin{aligned} \frac{\partial E_C}{\partial U} = &(UW^{A\top}U^\top UW^A + UW^AU^\top UW^{A\top} - A^\top UW^A - AUW^{A\top}) \\ &+ \lambda_C(UW^{C\top}U^\top UW^C + UW^{C\top}U^\top UW^{C\top} - C^\top UW^C - CUW^{C\top}) \\ &+ \lambda_U U \\ \frac{\partial E_C}{\partial W^A} = &U^\top UW^AU^\top U - U^\top AU + \lambda_{W^A}W^A \\ \frac{\partial E_C}{\partial W^C} = &\lambda_C(U^\top UW^CU^\top U - U^\top CU) + \lambda_{W^C}W^C \end{aligned} \tag{17}$$

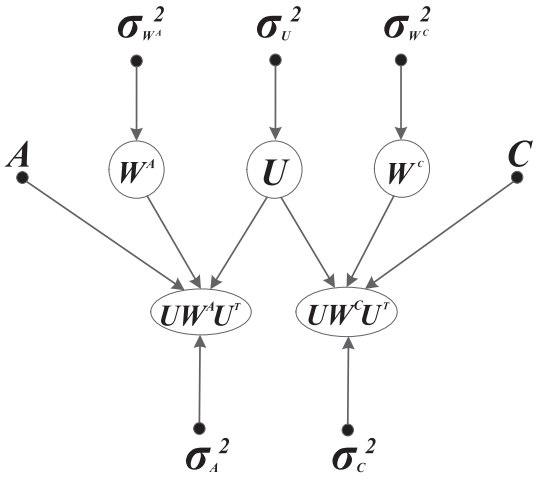


Fig. 2. A directed probabilistic graphical model representing the relations among the matrices and parameters in the FPMF model which fuse the asymmetric matrix C.

3.3. Final FPMF model fusing both symmetric and asymmetric semantic

In the above description, we have presented the fusion models which respectively fuse the symmetric metrics and asymmetric metrics. In this section, we will present the final fusion model which fuse the two kinds of metrics simultaneously.

The final fusion model is based on the two above-mentioned models. Fig. 3 shows the relations among the matrices and parameters in the final FPMF model. The joint probability density distribution over the variables U, W^A, W^C, $UW^AU^\top$, $UU^\top$ and $UW^AU^\top$ can be represented as

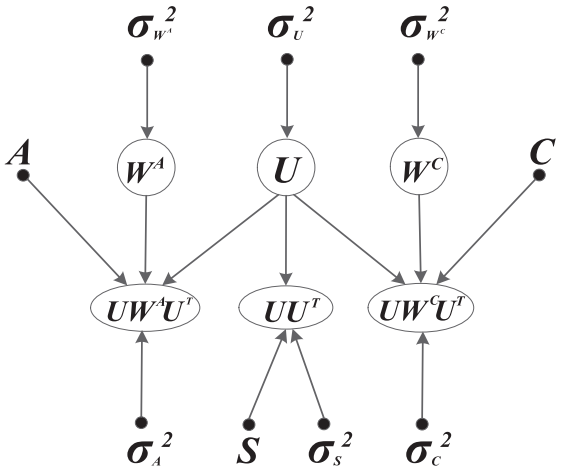


Fig. 3. A directed probabilistic graphical model representing the relations among the matrices and parameters in the FPMF model which fuse both the symmetric matrix S and asymmetric matrix C.

$$
\begin{aligned}
&p\left(UW^AU^\top, UU^\top, UW^AU^\top, U, W^C, W^A \middle| A, S, C, \sigma_U^2, \sigma_{W^A}^2, \sigma_{W^C}^2, \sigma_A^2, \sigma_S^2, \sigma_C^2\right)\\
&= p(U|\sigma_U^2)p\left(W^A\middle|\sigma_{W^A}^2\right)p\left(W^C\middle|\sigma_{W^C}^2\right)\\
&p(UW^AU^\top|A, U, W^A, \sigma_A^2)\\
&p(UW^CU^\top|C, U, W^C, \sigma_C^2)\\
&p(UU^\top|S, U, \sigma_S^2)
\end{aligned}
\tag{18}
$$

By maximizing the joint probability distribution, we can deduce the optimization problem $\underset{U,W^A,W^C}{argmin} E$, and the objective function E is denoted as

$$
\begin{aligned}
E &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(A_{ij} - U_iW^AU_j^\top)^2 + \frac{\lambda_S}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(S_{ij} - U_iU_j^\top)^2\\
&+ \frac{\lambda_C}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(C_{ij} - U_iW^CU_j^\top)^2 + \frac{\lambda_U}{2}\sum_{i=1}^{N}U_iU_i^\top\\
&+ \frac{\lambda_{W^A}}{2}\sum_{l=1}^{L}W_l^{A\top}W_l^A + \frac{\lambda_{W^C}}{2}\sum_{l=1}^{L}W_l^{C\top}W_l^C
\end{aligned}
\tag{19}
$$

where $\lambda_S = \frac{\sigma_A^2}{\sigma_S^2}$, $\lambda_C = \frac{\sigma_A^2}{\sigma_C^2}$, $\lambda_U = \frac{\sigma_A^2}{\sigma_U^2}$, $\lambda_{W^A} = \frac{\sigma_A^2}{\sigma_{W^A}^2}$, and $\lambda_{W^C} = \frac{\sigma_A^2}{\sigma_{W^C}^2}$. Similarly, the objective function (19) can be rewritten as

$$
\begin{aligned}
E &= \|A - UW^AU^\top\|_F^2 + \lambda_S\|S - UU^\top\|_F^2 + \lambda_C\|C - UW^CU^\top\|_F^2\\
&+ \lambda_U\|U\|_F^2 + \lambda_{W^A}\|W^A\|_F^2 + \lambda_{W^C}\|W^C\|_F^2
\end{aligned}
\tag{20}
$$

$\underset{U,W^A,W^C}{argmin} E$ can be solved by using the gradient methods as $\underset{U,W^A}{argmin} E_S$ and $\underset{U,W^A,W^C}{argmin} E_C$.

$$
\begin{aligned}
\frac{\partial E}{\partial U} &= (UW^{A\top}U^\top UW^A + UW^AU^TUW^{A\top} - A^\top UW^A - AUW^{A\top})\\
&+ \lambda_S(2UU^\top U - SU - S^\top U)\\
&+ \lambda_C(UW^{C\top}U^\top UW^C + UW^CU^\top UW^{C\top} - C^\top UW^C - CUW^{C\top})\\
&+ \lambda_U U\\
\frac{\partial E}{\partial W^A} &= (U^\top UW^AU^\top U - U^\top AU) + \lambda_{W^A}W^A\\
\frac{\partial E}{\partial W_C} &= (U^\top UW^CU^\top U - U^\top CU) + \lambda_{W^C}W^C
\end{aligned}
\tag{21}
$$

Note that the symmetric matrix S and the asymmetric matrix C fused in our fusion models need to be normalized to keep the two matrices and the adjacent matrix A stay at the same magnitude so as to facilitate adjusting the fusion model's parameters.

3.4. Link prediction

We have presented the FPMF models which provide strategies to fuse the adjacent matrix and some key topological metrics in a unified probability matrix factorization framework. In the fusion models, the basic part of the model is the approximation of the adjacent network A, i.e. $UW^AU^\top$. Suppose that we have learned any two nodes' vectorized latent-feature representation U_i and U_j and the matrix parameters W^A, then the linking possibility p_{ij} from node u_i to node u_j can be computed as

$$
p_{i,j} = p(U_iW^AU_j^\top|A_{ij} = 1, \sigma_A^2) = \frac{1}{\sqrt{2\sigma_A^2\pi}}e^{-\frac{\left(U_iW^AU_j^\top - 1\right)^2}{2\sigma_A^2}}
\tag{22}
$$

The pseudo code of the link prediction mechanism based on the final fusion model is shown in Algorithm 1.

3.5. Computational complexity analysis

The computational overhead of the model's learning process mainly comes from the calculation of the gradients of function E against variables U, W^A and W^C. Because of the sparsity of the matrices A, S and C,

Input: The adjacent matrix A, the symmetric matrix S, and the asymmetric matrix C.
Output: AUC result of link prediction.
1: **Initialize:** Nodes' L-dimension representation U_{n*L},matrix parameters W_{L*L}^A of A, and matrix parameters W_{L*L}^C of C, $L = integer \ll n$.
2: **repeat**
3: $U^{new} := U^{old} - \gamma\frac{\partial E}{\partial U}$
4: $W^{Anew} := W^{Aold} - \gamma\frac{\partial E}{\partial W^A}$
5: $W^{Cnew} := W^{Cold} - \gamma\frac{\partial E}{\partial W^C}$
6: **until** Convergence
7: **for** each unknown node pair $<u_i, u_j>$ **calculate**
8: $p_{i,j} = \frac{1}{\sqrt{2\sigma_A^2\pi}}e^{-\frac{(U_iW^AU_j^\top - 1)^2}{2\sigma_A^2}}$
9: **end for**
10: **Sort** all node pairs $<u_i, u_j>$ according to $p_{i,j}$
11: **Calculate** $AUC = \frac{r' + 0.5r''}{r}$. (see Equation (23))
12: **Return** AUC result of the fusion model.

Algorithm 1. Link prediction based on the final fusion model.

Table 4
Statistics of the six directed networks.

Datasets	Links	Nodes	Density (%)
cit	13,039	1499	0.58
email	37,098	1349	2.04
gplus	52,504	1300	3.11
weibo	102,750	1600	4.01
soclive	63,102	1375	3.33
socpokec	10,752	1396	0.55

Table 5
Statistics of the six undirected networks.

Datasets	Links	Nodes	Density(%)
eroad	1417	1175	0.20
hf	12,534	1859	0.72
facebook	9626	534	6.76
yeast	11,693	2375	0.41
power	6593	4941	0.05
router	6258	5022	0.05

the computational complexity of multiplication of A and U is $O(\mu_A L)$, where μ_A is the number of nonzero entries of A. Similarly, the computational complexity CU is $O(\mu_C L)$, and SU is $O(\mu_S L)$, where μ_C and μ_S are the number of nonzero entries of C and S, respectively. The computational complexity of the rest multiplications in the gradients is $O(nL^2)$. Therefore, the total computational complexity of the fusion model which fuses the symmetric matrix S in one iteration is $O(\mu_A L + \mu_S l + nL^2)$, of the fusion model which fuses the asymmetric matrix C is $O(\mu_A L + \mu_C L + nL^2)$, and of the final fusion model is $O(\mu_A L + \mu_S L + \mu_C L + nL^2)$. In other words, the computational complexity is scale linearly as the increase of the nonzero entries in the matrices A, S, and C.

4. Experiments

4.1. Data sets

We conduct our experiments with six directed networks and six undirected networks. The statistic details of the network datasets are summarized in Tables 4 and 5.

In Table 4, *cit* is a citation network, and if a paper *i* cites paper *j*, the graph will contain a directed edge from *i* to *j; email* is a communication network, and if an email address *i* sends at least one message to address *j*, a directed edge between address *i* and address *j* will exist. All the rest four datasets *gplus, weibo, soclive*, and *socpokec* are online directed social networks. *Weibo* is a well-known online social network service (SNS) in China, and its dataset in our study is from [63] which crawled the datasets from *weibo* SNS platform. Based on the original *weibo* dataset in [63], we further extract a subgraph for our experiments. Apart from the *weibo* dataset, the other five datasets are available to download at the platform of Stanford Network Analysis Project (SNAP)[3].

Table 5 presents six undirected networks, where *eroad* is a road network mostly located in Europe, and *hf* (hamsterster friendships) is the friendship network of the website "hamsterster.com". The two datasets are from the project of the Koblenz Network Collection[4]. *Facebook* is the user-user undirected network downloaded from the SNAP platform. The rest three undirected networks *power, router*, and *yeast* are the networks of USA power, internet route and protein interaction respectively; the three datasets are the open datasets of the link prediction group[5].

4.2. Experiment setup and evaluation measure

4.2.0.1. Experiment setup. Following the general experimental protocol for link prediction [16,21,56,64], we split the observed node-to-node links into training data and testing data and adopt the partition 90%/10% and carry out 10 times independently.

4.2.0.2. Evaluation measure. Like many existing link prediction studies [21], our study adopts the most frequently-used metrics "Area under the receiver operating characteristic" (AUC) to measure the performance of link prediction. This metric is viewed as a robust measure in the presence of imbalance [17]. 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 which is given a higher score than a randomly chosen non-existent link [65]. Specifically, we can use the following equation to compute the *AUC* value.

$$AUC = \frac{r' + 0.5r''}{r} \tag{23}$$

where r is the number of independent comparisons, r' is the times for the non-observed links which are given higher scores than non-existent links, and r'' is the times for the scores of the non-observed links which are equal to the scores of the non-existent links. The value of r is set to 10,000 in our experiments.

4.3. Comparison methods

Before we present the comparison methods, we denote the proposed fusion models in this paper as FPMF_S, FPMF_C, and FPMF_S_C, where FPMF_S represents the models fusing the symmetric matrix S, FPMF_C denotes the models fusing the asymmetric matrix C, and FPMF_S_C refers to the models fusing both of S and C. The specific matrices S and C fused in our fusion models will be introduced in the following experiments.

4.3.1. Baseline methods

In Table 6, we list two types of link prediction methods as the baseline methods, i.e. the metric-based methods and the MF-based methods. They are relevant to the methodology in the paper. There are 15 symmetric metrics in Table 6, where SCN, SJ, and SP are the metric-based methods for the undirected networks, and the other 12 symmetric metrics SO, SI, SF, SN, SPO, SPI, SPF, SPN, SJO, SJI, SJF, and SJN are for the directed networks. Besides, CJ, CO, CI, CF, and CN are the asymmetric metrics, where CJ is for the undirected networks, and the other 4 asymmetric metrics are for the directed networks. Note that the listed metrics in Table 6 are from the labeled references or the variants of the references.

Low-rank approximation (LRA)-based methods [19] is a common technique to find an approximate matrix with low-rank of an original matrix. For link prediction, LRA-based methods usually compute the rank-k matrix A_k that best approximates the network adjacent matrix A and scores the link possibility between u_i and u_j via using the entry $< i, j >$ of the matrix A_k. In our experiments, we will use the two common LRA techniques as the comparison methods, i.e. SVD [66] and NNMF [67]. We denote the two LRA-based methods as LRA-SVD and LRA-NNMF respectively.

4.3.2. Popular link prediction methods

Apart from the baseline methods, we also compare our methods with six popular link prediction methods.

1. RA (Resource Allocation) [68], which is motivated by the resource allocation dynamics on complex networks, is mentioned as the best

[3] http://snap.stanford.edu
[4] http://konect.uni-koblenz.de
[5] http://www.linkprediction.org

Table 6
Baseline methods.

Type	Abbreviate	Formula	References
Symmetric metrics	SCN	$S_{ij}^{CN} = \|\Gamma(u_i) \cap \Gamma(u_j)\|$	[36]
	SJ	$S_{ij}^{J} = \frac{\|\Gamma(u_i) \cap \Gamma(u_j)\|}{\|\Gamma(u_i) \cup \Gamma(u_j)\|}$	[37]
	SP	$S_{ij}^{P} = \|\Gamma(u_i)\| \times \|\Gamma(u_j)\|$	[35]
	SO	$S_{ij}^{O} = \|\Gamma^{-}(u_i) \cap \Gamma^{-}(u_j)\|$	[55]
	SJO	$S_{ij}^{JO} = \frac{\|\Gamma^{-}(u_i) \cap \Gamma^{-}(u_j)\|}{\|\Gamma^{-}(u_i) \cup \Gamma^{-}(u_j)\|}$	[37,55]
	SI	$S_{ij}^{I} = \|\Gamma^{+}(u_i) \cap \Gamma^{+}(u_j)\|$	[55]
	SJI	$S_{ij}^{JI} = \frac{\|\Gamma^{+}(u_i) \cap \Gamma^{+}(u_j)\|}{\|\Gamma^{+}(u_i) \cup \Gamma^{+}(u_j)\|}$	[37,55]
	SF	$S_{ij}^{F} = \|(\Gamma^{-}(u_i) \cap \Gamma^{+}(u_i)) \cap (\Gamma^{-}(u_j) \cap \Gamma^{+}(u_j))\|$	[55]
	SJF	$S_{ij}^{JF} = \frac{\|(\Gamma^{-}(u_i) \cap \Gamma^{+}(u_i)) \cap (\Gamma^{-}(u_j) \cap \Gamma^{+}(u_j))\|}{\|(\Gamma^{-}(u_i) \cap \Gamma^{+}(u_i)) \cup (\Gamma^{-}(u_j) \cap \Gamma^{+}(u_j))\|}$	[37,55]
	SN	$S_{ij}^{N} = \|(\Gamma^{-}(u_i) \cup \Gamma^{+}(u_i)) \cap (\Gamma^{-}(u_j) \cup \Gamma^{+}(u_j))\|$	[55]
	SJN	$S_{ij}^{JN} = \frac{\|(\Gamma^{-}(u_i) \cup \Gamma^{+}(u_i)) \cap (\Gamma^{-}(u_j) \cup \Gamma^{+}(u_j))\|}{\|(\Gamma^{-}(u_i) \cup \Gamma^{+}(u_i)) \cup (\Gamma^{-}(u_j) \cup \Gamma^{+}(u_j))\|}$	[37,55]
	SPO	$S_{ij}^{PO} = \|\Gamma^{-}(u_i)\| \times \|\Gamma^{-}(u_j)\|$	[35,55]
	SPI	$S_{ij}^{PI} = \|\Gamma^{+}(u_i)\| \times \|\Gamma^{+}(u_j)\|$	[35,55]
	SPF	$S_{ij}^{PF} = \|(\Gamma^{-}(u_i) \cap \Gamma^{+}(u_i))\| \times \|(\Gamma^{-}(u_j) \cap \Gamma^{+}(u_j))\|$	[35,55]
	SPN	$S_{ij}^{PN} = \|(\Gamma^{-}(u_i) \cup \Gamma^{+}(u_i))\| \times \|(\Gamma^{-}(u_j) \cup \Gamma^{+}(u_j))\|$	[35,55]
Asymmetric metrics	CJ	$C_{ij}^{J} = \frac{\|\Gamma(u_i) \cap \Gamma(u_j)\|}{\|\Gamma(u_i)\|}$	[37,55]
	CO	$C_{ij}^{O} = \frac{\|\Gamma^{-}(u_i) \cap \Gamma^{-}(uj)\|}{\|\Gamma^{-}(u_i)\|}$	[37,55]
	CI	$C_{ij}^{I} = \frac{\|\Gamma^{+}(u_i) \cap \Gamma^{+}(u_j)\|}{\|\Gamma^{+}(u_i)\|}$	[37,55]
	CF	$C_{ij}^{F} = \frac{\|(\Gamma^{-}(u_i) \cap \Gamma^{+}(u_i)) \cap (\Gamma^{-}(u_j) \cap \Gamma^{+}(u_j))\|}{\|(\Gamma^{-}(u_i) \cap \Gamma^{+}(u_i))\|}$	[37,55]
	CN	$C_{ij}^{N} = \frac{\|(\Gamma^{-}(u_i) \cup \Gamma^{+}(u_i)) \cap (\Gamma^{-}(u_j) \cup \Gamma^{+}(u_j))\|}{\|\Gamma^{-}(u_i) \cup \Gamma^{+}(u_i)\|}$	[37,55]
MF-based method	LRA-SVD	Singular Value Decomposition	[19,66]
	LRA-NNMF	Non-negative Matrix Factorization	[67]

local similarity in literature [21].

2. AA (Adamic–Adar) [33], which refines the simple counting of common neighbors by assigning the less-connected neighbors more weight, is mentioned as the second best local similarity in literature [21].
3. Katz [41], which is based on the ensemble of all paths, is a popular link prediction method and usually performs good subject to the AUC value [21].
4. SR (SimRank) [44] is a popular random-walk metric, which measures how soon two random walkers, respectively starting from nodes u_i and u_j, are expected to meet at a certain node.
5. WIC, which considers community membership information, is validated as the best link prediction method in literature [69].
6. RA-W is validated as the second best link prediction method in literature [69].

4.4. Experimental results and analysis

To verify the performance of the proposed fusion models, we conduct two parts of comparison experiments. One is to compare the proposed fusion models (FPMF_S, FPMF_C, and FPMF_S_C) with the baseline methods (see Section 4.4.1), and the other one is to compare the final fusion models (FPMF_S_C) with six popular link prediction methods (see Section 4.4.2). Moreover, we analyze the parameters L and λ (λ_S, λ_C) in Section 4.4.3.

All the experiments are conducted on an Intel E5-2660 20 Core 2.6 GHz with 96 GB memory. The underlined results denote that our methods are superior to the compared methods.

4.4.1. Comparing with baseline methods

(1)Results of the fusion models (FPMF_S)

For the directed networks, we build two fusion models FPMF_SN and FPMF_SJN by fusing the two symmetric metrics SN and SJN because the two metrics perform better in link prediction. Analogously, we also build a fusion model FPMF_SCN for the undirected networks by fusing the symmetric metric SCN.

As shown in Tables 7 and 8, the results indicate that the fusion models FPMF_SN, FPMF_SJN and FPMF_SCN for the most parts obtain better link prediction results than the baseline methods in the six directed and the six undirected networks.

As for the 15 symmetric metric-based methods, the link prediction problem is solved by directly using the metrics to score the link possibility, and the majority of them show the unstable performance in different networks. Instead of directly using the metrics for scoring the link possibility, our methods provide a fusion strategy to fuse the metrics with the adjacent matrix in a unified probability matrix framework. The results confirm that, in the fusion models, a more proper practical correlation can be established between the network links and the metrics, rather than the strong assumptions behind the metric-based methods in which the link state is directly related to the size of the metric. Therefore, the fusion models are superior to the way of predicting links directly based on a metric. Taking the *gplus* network as an example, the AUC value of the SJN-based method is 0.396 (lower than 0.5), which means that SJN has a certain opposite correlation with the forming of links. However, after fusing the metric in our model, the fusion model FPMF_SJN obtains impressive results.

As far as the two MF-based methods (LAR-SVD and LAR-NNMF) are concerned, the performance is better than most of the other metric-based methods. However, they are still inferior to our fusion models.

Table 7
Comparing with baseline methods in directed networks (mean ± std-err).

	cit	*email*	*gplus*	*weibo*	*soclive*	*socpokec*
SO	0.785 ± 0.009	0.905 ± 0.003	0.697 ± 0.003	0.863 ± 0.003	0.914 ± 0.004	0.778 ± 0.004
SJO	0.784 ± 0.010	0.876 ± 0.004	0.639 ± 0.003	0.811 ± 0.005	0.903 ± 0.004	0.772 ± 0.005
SI	0.799 ± 0.004	0.903 ± 0.004	0.614 ± 0.003	0.855 ± 0.003	0.912 ± 0.004	0.791 ± 0.007
SJI	0.794 ± 0.005	0.866 ± 0.002	0.507 ± 0.004	0.790 ± 0.005	0.908 ± 0.003	0.782 ± 0.008
SF	0.500 ± 0.000	0.870 ± 0.005	0.695 ± 0.002	0.820 ± 0.003	0.889 ± 0.003	0.698 ± 0.006
SJF	0.500 ± 0.000	0.850 ± 0.003	0.673 ± 0.003	0.814 ± 0.004	0.891 ± 0.005	0.695 ± 0.005
SN	0.905 ± 0.004	0.928 ± 0.002	0.730 ± 0.005	0.882 ± 0.003	0.929 ± 0.003	0.832 ± 0.006
SJN	0.901 ± 0.004	0.879 ± 0.005	0.396 ± 0.005	0.843 ± 0.002	0.917 ± 0.002	0.822 ± 0.005
SPA	0.660 ± 0.011	0.846 ± 0.003	0.679 ± 0.004	0.784 ± 0.004	0.815 ± 0.007	0.805 ± 0.007
SPI	0.604 ± 0.009	0.853 ± 0.007	0.583 ± 0.006	0.781 ± 0.004	0.820 ± 0.006	0.808 ± 0.009
SPF	0.501 ± 0.001	0.843 ± 0.005	0.685 ± 0.003	0.779 ± 0.003	0.812 ± 0.005	0.786 ± 0.012
SPN	0.741 ± 0.005	0.857 ± 0.004	0.798 ± 0.004	0.815 ± 0.004	0.823 ± 0.004	0.806 ± 0.009
LAR-SVD	0.942 ± 0.007	0.923 ± 0.004	0.956 ± 0.004	0.918 ± 0.004	0.920 ± 0.002	0.860 ± 0.005
LRA-NNMF	0.942 ± 0.005	0.932 ± 0.003	0.952 ± 0.002	0.915 ± 0.003	0.917 ± 0.003	0.876 ± 0.008
FPMF_SN	0.949 ± 0.006	0.932 ± 0.001	0.984 ± 0.002	0.920 ± 0.001	0.923 ± 0.004	0.891 ± 0.007
FPMF_SJN	0.937 ± 0.008	0.930 ± 0.005	0.983 ± 0.005	0.920 ± 0.007	0.922 ± 0.007	0.895 ± 0.007

The MF-based methods solve the link prediction problem by using the low-rank approximation techniques [19,66,67] and model the network by considering the information of the observed network links (i.e. the adjacent matrix). This is hard to ensure the modeling of those topological metrics that can play an important role in link prediction. While our fusion models (FPMF_SN, FPMF_SJN, and FPMF_SCN) can give dual attention to modeling the observed network links and the metrics. Therefore, the fusion models almost consistently outperform the MF-based methods.

(2)Results of the fusion models (FPMF_C)

Like the fusion models mentioned above, we build two fusion models FPMF_CN and FPMF_CI for the directed networks by fusing the two asymmetric metrics CN and CI. Meanwhile, we also build the fusion model FPMF_CJ for the undirected network by fusing the asymmetric metric CJ.

The experiment results (see Tables 9 and 10) indicate that the fusion method FPMF_CN, FPMF_CI, and FPMF_CJ compared with the baseline methods, in most cases, get better results. The results verify that our fusion models by fusing the asymmetric metrics are also effective for link prediction.

(3)Results of the fusion models (FPMF_S_C)

In this section, we build the final fusion models by fusing both the symmetric metrics and the asymmetric metrics simultaneously. Specifically, the metrics fused in our model are constructed by the various combination between the symmetric metrics and the asymmetric metrics. For the directed networks, the symmetric metrics fused are SN and SJN, and the asymmetric metrics are CN and CI. Based on the four metrics, we construct four fusion models, and they are denoted as FPMF_SJN_CN, FPMF_SJN_CI, FPMF_SN_CN, and FPMF_SN_CI. For the undirected networks, we build the fusion model FPMF_SCN_CJ by fusing the symmetric metric SCN and the asymmetric metric CJ.

As shown in Tables 11 and 12, the final fusion models (i.e. FPMF_SJN_CN, FPMF_SJN_CI, FPMF_SN_CN, FPMF_SN_CI and FPMF_SCN_CJ) almost consistently outperform the other approaches in the directed and the undirected networks. Note that the final fusion models are mostly superior to the version of fusion models (FPMF_SN, FPMF_SJN, FPMF_CN, FPMF_CI, FPMF_SCN, and FPMF_CJ) which only fuse one side of the symmetric or the asymmetric metrics. This illustrates that the two kinds of metrics are complementary to each other in the final fusion models. The improvements verify that our models are more effective by fusing both sides of the symmetric and the asymmetric metrics.

4.4.2. Comparing with popular methods

Tables 13 and 14 show the link prediction results of the proposed final FPMF models (FPMF_SJN_CN, FPMF_SJN_CI, FPMF_SN_CN, FPMF_SN_CI and FPMF_SCN_CJ) compared with the popular methods in the directed and the undirected networks respectively. In most cases, the results show that the proposed fusion models get better AUC values than the comparison methods.

As for the six comparison methods, the RA method gets the best prediction results either in the directed networks or the undirected networks, which is consistent with the validation results in literature [21]. While the other comparison methods AA, Katz, SR, WIC, and RA-W perform less well, and their results show instability in different datasets. By contrast, our fusion methods achieve better AUC values and can give relatively stable prediction results in different networks as well.

To sum up, the reasons why our fusion models are superior to the comparison methods can be elaborated from two aspects: (1) The fusion models, in essence, are learning-based methods by fitting the adjacent matrix, the symmetric metrics, and the asymmetric metrics. Their advantage is that they can fit the network adaptively in the learning process, and thus they have better adaptability and robustness. (2) The fusion models can give dual attention to modeling the observed network links and some key metrics between nodes, which are superior to the traditional MF-based methods.

4.4.3. Model parameters analysis

There are two kinds of parameters in our model, the one is dimension parameter L, and the other is weighting parameters (λ_S, λ_C). The parameters are analyzed separately as follows.

Table 8
Comparing with baseline methods in undirected networks (mean ± std-err).

	eroad	*facebook*	*yeast*	*hf*	*power*	*router*
SCN	0.539 ± 0.009	0.967 ± 0.003	0.914 ± 0.011	0.814 ± 0.008	0.625 ± 0.007	0.651 ± 0.008
SJ	0.539 ± 0.010	0.969 ± 0.002	0.912 ± 0.009	0.799 ± 0.007	0.627 ± 0.007	0.651 ± 0.008
SP	0.455 ± 0.028	0.766 ± 0.013	0.863 ± 0.032	0.887 ± 0.005	0.577 ± 0.012	0.955 ± 0.003
LAR-SVD	0.756 ± 0.041	0.942 ± 0.006	0.910 ± 0.031	0.897 ± 0.007	0.670 ± 0.027	0.829 ± 0.016
LRA-NNMF	0.619 ± 0.012	0.942 ± 0.008	0.909 ± 0.026	0.909 ± 0.008	0.606 ± 0.008	0.892 ± 0.010
FPMF_SCN	0.916 ± 0.013	0.961 ± 0.016	0.963 ± 0.008	0.930 ± 0.017	0.683 ± 0.015	0.956 ± 0.011

Table 9
Comparing with baseline methods in directed networks (mean ± std-err).

	cit	*email*	*gplus*	*weibo*	*soclive*	*socpokec*
CO	0.784 ± 0.009	0.871 ± 0.004	0.643 ± 0.004	0.814 ± 0.003	0.902 ± 0.003	0.778 ± 0.005
CI	0.802 ± 0.005	0.858 ± 0.003	0.691 ± 0.005	0.841 ± 0.003	0.911 ± 0.003	0.788 ± 0.007
CF	0.500 ± 0.000	0.853 ± 0.005	0.671 ± 0.003	0.811 ± 0.004	0.883 ± 0.003	0.700 ± 0.005
CN	0.901 ± 0.004	0.889 ± 0.004	0.527 ± 0.005	0.840 ± 0.002	0.903 ± 0.005	0.826 ± 0.006
LAR-SVD	0.942 ± 0.007	0.923 ± 0.004	0.956 ± 0.004	0.918 ± 0.004	0.920 ± 0.002	0.860 ± 0.005
LRA-NNMF	0.942 ± 0.005	0.932 ± 0.003	0.952 ± 0.002	0.915 ± 0.003	0.917 ± 0.003	0.876 ± 0.008
FPMF_CN	0.948 ± 0.006	0.931 ± 0.004	0.982 ± 0.002	0.921 ± 0.005	0.923 ± 0.004	0.886 ± 0.002
FPMF_CI	0.949 ± 0.006	0.924 ± 0.001	0.984 ± 0.003	0.922 ± 0.003	0.921 ± 0.003	0.890 ± 0.003

Table 10
Comparing with baseline methods in undirected networks (mean ± std-err).

	eroad	*facebook*	*yeast*	*hf*	*power*	*router*
CJ	0.539 ± 0.009	0.960 ± 0.004	0.916 ± 0.010	0.799 ± 0.006	0.626 ± 0.007	0.651 ± 0.007
LAR-SVD	0.756 ± 0.041	0.942 ± 0.006	0.910 ± 0.031	0.897 ± 0.007	0.670 ± 0.027	0.829 ± 0.016
LRA-NNMF	0.619 ± 0.012	0.942 ± 0.008	0.909 ± 0.026	0.909 ± 0.008	0.606 ± 0.008	0.892 ± 0.010
FPMF_CJ	0.917 ± 0.022	0.959 ± 0.017	0.960 ± 0.043	0.937 ± 0.020	0.684 ± 0.016	0.957 ± 0.011

Table 11
Comparing with baseline methods in directed networks (mean ± std-err).

	cit	*email*	*gplus*	*weibo*	*soclive*	*socpokec*
SN	0.905 ± 0.004	0.928 ± 0.002	0.730 ± 0.005	0.882 ± 0.003	0.929 ± 0.003	0.832 ± 0.006
SJN	0.901 ± 0.004	0.879 ± 0.005	0.396 ± 0.005	0.843 ± 0.002	0.917 ± 0.002	0.822 ± 0.005
CN	0.901 ± 0.004	0.889 ± 0.004	0.527 ± 0.005	0.840 ± 0.002	0.903 ± 0.005	0.826 ± 0.006
CI	0.802 ± 0.005	0.858 ± 0.003	0.691 ± 0.005	0.841 ± 0.003	0.911 ± 0.003	0.788 ± 0.007
LAR-SVD	0.942 ± 0.007	0.923 ± 0.004	0.956 ± 0.004	0.918 ± 0.004	0.920 ± 0.002	0.860 ± 0.005
LRA-NNMF	0.942 ± 0.005	0.932 ± 0.003	0.952 ± 0.002	0.915 ± 0.003	0.917 ± 0.003	0.876 ± 0.008
FPMF_SN	0.949 ± 0.006	0.932 ± 0.001	0.984 ± 0.002	0.920 ± 0.001	0.923 ± 0.004	0.891 ± 0.007
FPMF_SJN	0.937 ± 0.008	0.930 ± 0.005	0.983 ± 0.005	0.920 ± 0.007	0.922 ± 0.007	0.895 ± 0.007
FPMF_CN	0.948 ± 0.006	0.931 ± 0.004	0.982 ± 0.002	0.921 ± 0.005	0.923 ± 0.004	0.886 ± 0.002
FPMF_CI	0.949 ± 0.006	0.924 ± 0.001	0.984 ± 0.003	0.922 ± 0.003	0.921 ± 0.003	0.890 ± 0.003
FPMF_SJN_CN	0.956 ± 0.002	0.933 ± 0.002	0.985 ± 0.001	0.924 ± 0.004	0.926 ± 0.005	0.944 ± 0.004
FPMF_SJN_CI	0.952 ± 0.006	0.935 ± 0.003	0.985 ± 0.001	0.926 ± 0.003	0.923 ± 0.002	0.953 ± 0.004
FPMF_SN_CN	0.951 ± 0.004	0.932 ± 0.004	0.984 ± 0.000	0.923 ± 0.001	0.927 ± 0.006	0.947 ± 0.005
FPMF_SN_CI	0.953 ± 0.005	0.935 ± 0.003	0.984 ± 0.003	0.924 ± 0.002	0.925 ± 0.002	0.952 ± 0.002

Table 12
Comparing with baseline methods in undirected networks (mean ± std-err).

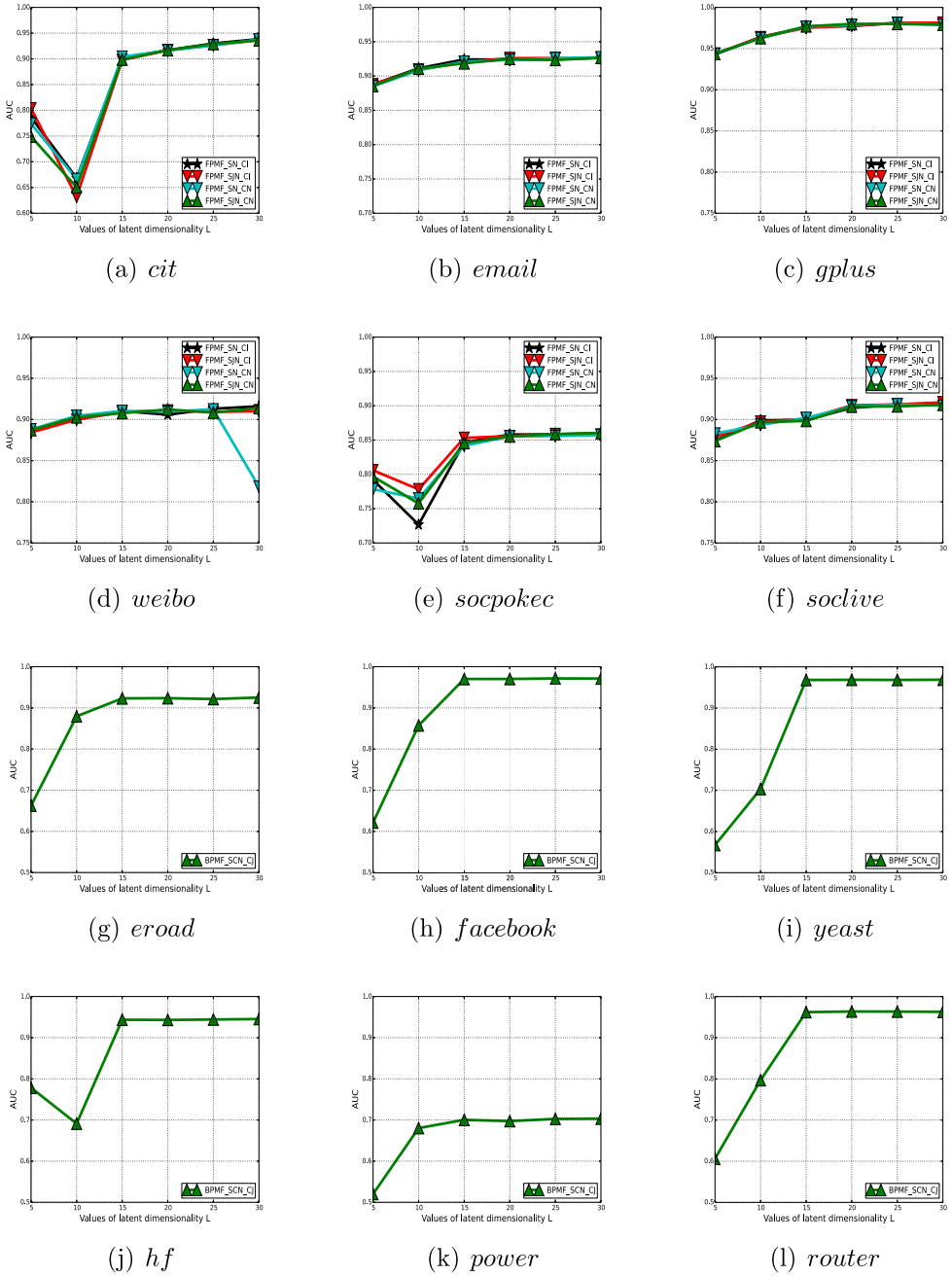
	eroad	*facebook*	*yeast*	*hf*	*power*	*router*
SCN	0.539 ± 0.009	0.967 ± 0.003	0.914 ± 0.011	0.814 ± 0.008	0.625 ± 0.007	0.651 ± 0.008
CJ	0.539 ± 0.009	0.960 ± 0.004	0.916 ± 0.010	0.799 ± 0.006	0.626 ± 0.007	0.651 ± 0.007
LAR-SVD	0.756 ± 0.041	0.942 ± 0.006	0.910 ± 0.031	0.897 ± 0.007	0.670 ± 0.027	0.829 ± 0.016
LRA-NNMF	0.619 ± 0.012	0.942 ± 0.008	0.909 ± 0.026	0.909 ± 0.008	0.606 ± 0.008	0.892 ± 0.010
FPMF_SCN	0.916 ± 0.013	0.961 ± 0.016	0.963 ± 0.008	0.930 ± 0.017	0.683 ± 0.015	0.956 ± 0.011
FPMF_CJ	0.917 ± 0.022	0.959 ± 0.017	0.960 ± 0.043	0.937 ± 0.020	0.684 ± 0.016	0.957 ± 0.011
FPMF_SCN_CJ	0.923 ± 0.050	0.970 ± 0.030	0.968 ± 0.010	0.944 ± 0.014	0.700 ± 0.040	0.964 ± 0.009

Table 13
Comparing with popular methods in directed networks (mean ± std-err).

	cit	*email*	*gplus*	*weibo*	*soclive*	*socpokec*
RA	0.931 ± 0.003	0.928 ± 0.003	0.958 ± 0.002	0.913 ± 0.003	0.923 ± 0.003	0.774 ± 0.004
AA	0.777 ± 0.010	0.525 ± 0.006	0.427 ± 0.005	0.310 ± 0.006	0.383 ± 0.006	0.688 ± 0.006
Katz	0.913 ± 0.002	0.423 ± 0.004	0.328 ± 0.030	0.445 ± 0.020	0.401 ± 0.041	0.529 ± 0.003
SR	0.802 ± 0.005	0.858 ± 0.003	0.581 ± 0.008	0.762 ± 0.006	0.911 ± 0.003	0.788 ± 0.007
WIC	0.631 ± 0.007	0.646 ± 0.003	0.521 ± 0.002	0.515 ± 0.003	0.651 ± 0.002	0.602 ± 0.003
RA-W	0.624 ± 0.002	0.632 ± 0.008	0.542 ± 0.006	0.492 ± 0.001	0.655 ± 0.004	0.624 ± 0.001
FPMF_SJN_CN	0.956 ± 0.002	0.933 ± 0.002	0.985 ± 0.001	0.924 ± 0.004	0.926 ± 0.005	0.944 ± 0.004
FPMF_SJN_CI	0.952 ± 0.006	0.935 ± 0.003	0.985 ± 0.001	0.926 ± 0.003	0.923 ± 0.002	0.953 ± 0.004
FPMF_SN_CN	0.951 ± 0.004	0.932 ± 0.004	0.984 ± 0.000	0.923 ± 0.001	0.927 ± 0.006	0.947 ± 0.005
FPMF_SN_CI	0.953 ± 0.005	0.935 ± 0.003	0.984 ± 0.003	0.924 ± 0.002	0.925 ± 0.002	0.952 ± 0.002

Table 14
Comparing with popular methods in undirected networks (mean ± std-err).

	eroad	*facebook*	*yeast*	*hf*	*power*	*router*
RA	0.543 ± 0.007	0.969 ± 0.004	0.922 ± 0.001	0.814 ± 0.003	0.628 ± 0.002	0.658 ± 0.005
AA	0.537 ± 0.002	0.961 ± 0.004	0.910 ± 0.010	0.820 ± 0.007	0.627 ± 0.002	0.660 ± 0.002
Katz	0.919 ± 0.011	0.356 ± 0.020	0.340 ± 0.060	0.388 ± 0.050	0.698 ± 0.020	0.710 ± 0.021
SR	0.902 ± 0.004	0.925 ± 0.003	0.839 ± 0.004	0.839 ± 0.003	0.643 ± 0.006	0.801 ± 0.004
WIC	0.675 ± 0.003	0.727 ± 0.004	0.681 ± 0.007	0.623 ± 0.005	0.642 ± 0.004	0.792 ± 0.003
RA-W	0.762 ± 0.004	0.675 ± 0.002	0.697 ± 0.002	0.652 ± 0.002	0.607 ± 0.004	0.695 ± 0.008
FPMF_SCN_CJ	0.923 ± 0.050	0.970 ± 0.030	0.968 ± 0.010	0.944 ± 0.014	0.700 ± 0.040	0.964 ± 0.009



(a) *cit* (b) *email* (c) *gplus*

(d) *weibo* (e) *socpokec* (f) *soclive*

(g) *eroad* (h) *facebook* (i) *yeast*

(j) *hf* (k) *power* (l) *router*

Fig. 4. Impact of Parameter *L*.

4.4.3.1. Impact of the parameter L. In the fusion models, *L* controls the dimension of the latent-feature matrix *U* of network nodes. We evaluate the influence on the final results of link prediction by changing the value of *L*. As shown in Fig. 4(a)–4(l), the improvements of results become increasingly marginal as *L* increases and tend to converge to stable results quite rapidly. Hence, we should limit the value of *L* so that an acceptable compromise will be reached between the performance of link prediction and that of time consumption.

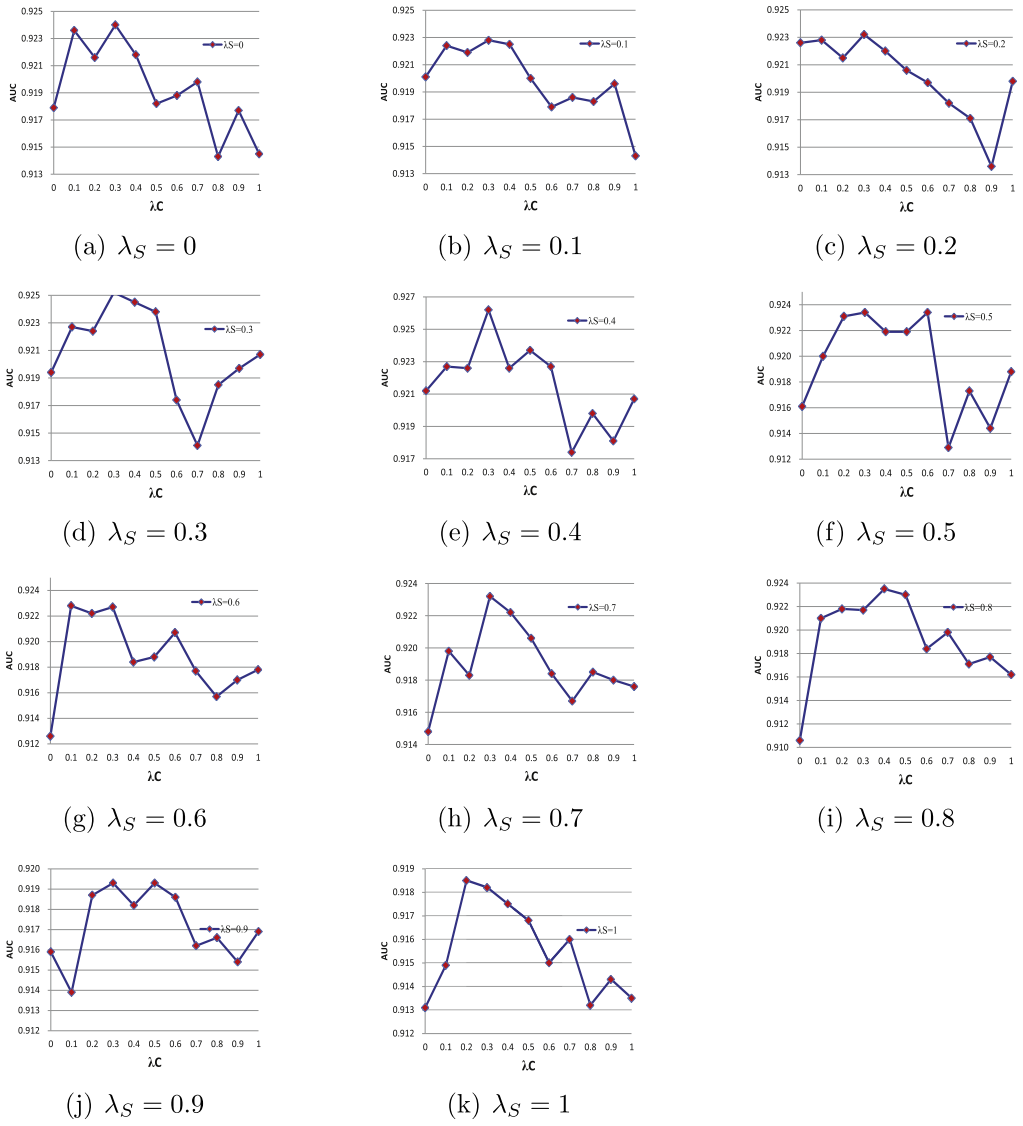


Fig. 5. Impact of Parameters λ (λ_S and λ_C) on data *soclive*. The 11 figures from (a) to (k) show the performance of fusion model FPMF_SJN_CN on *soclive* when parameters λ_S and λ_C change. The horizontal axis of each figure represents the value of λ_C from 0 to 1, and the vertical axis of each figure represents the *auc* results of link prediction. In the figures from (a) to (k), we fix the value of λ_S which corresponds to 0, 0.1, ... , 0.9, 1, while the value of λ_C in each figure is changes within [0,1] based on the interval of 0.1.

4.4.3.2. Impact of parameters (λ_S, λ_C). The main advantage of our fusion models is that it can give dual attention to modeling the observed network links and those topological metrics which can play an important role in link prediction. In the models, parameters (λ_S, λ_C) balance the information between the adjacent matrix and the metrics. If $\lambda_S = 0$ and $\lambda_C = 0$, the model only considers the information from the adjacent matrix, and if $\lambda_S = inf$ or if $\lambda_C = inf$, the model only extracts information from the symmetric metrics or the asymmetric metrics.

Take the experiments of fusion model FPMF_SJN_CN on data *soclive* for example, we observe that changes in parameters, either λ_S or λ_C, will affect the results (see Fig. 5). First, as λ_S is fixed to 0 (see Fig. 5(a)), and λ_C is fixed to 0, the model only mines the adjacent matrix A, and the corresponding AUC result is not very well. Second, from Fig. 5(a) to Fig. 5(k), the fusion model's performance is relatively better when $\lambda_C \in [0.2, 0.5]$. Moreover, as $\lambda_C \in [0.2, 0.5]$, and λ_S is separately fixed to 0.1, 0.2, 0.3, or 0.4 (see Figs. 5(b), 5(c), 5(d) and 5(e)), the model gets best results. The phenomenon coincides with the intuition that overmuch usage of such information, the adjacent matrix A, the symmetric matrix S or the asymmetric matrix C, rather than reasonable fusing these resources together, cannot generate best performance.

4.4.4. Relative discussions

In this section, we summarize the advantages of the *FPMF* models and analyze the reasons.

- Each of the fusion models (fusing the symmetric metrics or the asymmetric metrics, or both of them), for the most part, obtain better link prediction results.
  First, the fusion models not only consider the information of the observed network links but also some key topological metrics. Second, the topological metrics fused in our model do not directly play a role in link prediction but integrate with the network links in a probabilistic matrix factorization model. Using the model's final representation to solve the link prediction can avoid the inferiority of predicting links by directly using a single metric between nodes.
- The final fusion models not only consider the symmetric metrics but also the asymmetric metrics.
  Because the formation of the links in a network may be caused by some symmetric or asymmetric semantic between nodes, the proposed models (FPMF_S_C) provide a way to fuse both kinds of semantic in one unified probabilistic matrix factorization framework.

5. Conclusion and future work

Studying how to accurately infer links in networks is still a difficult problem in network data analysis. Despite significant advances, the existing metric-based link prediction methods usually only consider one single topological metric and thus show some limitations in different types of networks; the existing matrix factorization-based models mainly focus on modeling the adjacent matrix of a network, and this is hard to ensure the modeling of those topological metrics that can play an important role in link prediction.

This study presents the fusion models (FPMF_S, FPMF_C, and FPMF_S_C) to fuse the adjacent matrix and some key topological metrics in a unified probability matrix factorization framework. The final fusion models consider both the symmetric metrics and the asymmetric metrics as well. The asymmetric metrics are usually not taken into consideration in the related work. To verify the performance of the FPMF models for the link prediction, we compare our approaches with a number of relevant link prediction methods. Experiments with 12 real-world directed and undirected networks show that the proposed models give impressive predicting performance for link prediction.

This work has many potential directions in the future. For example, we can study how to conduct incremental learning on the fusion probability matrix factorization models so that the models could be adapted to a dynamic circumstance. Besides, it will be an interesting topic to fuse richer metrics between nodes in a network in the probability matrix factorization framework. One case in point is to fuse the metrics based on some non-topological information.

Acknowledgments

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

References

[1] L. Getoor, C.P. Diehl, Link mining: a survey, ACM SIGKDD Explor. Newsl. 7 (2) (2005) 3–12.
[2] K. Juszczyszyn, K. Musial, M. Budka, Link Prediction Based on Subgraph Evolution in Dynamic Social Networks, IEEE International Conference on Social Computing, IEEE, Boston, Massachusetts, USA, 2011, pp. 27–34.
[3] R. Pastor-Satorras, C. Castellano, P.V. Mieghem, A. Vespignani, Epidemic processes in complex networks, Rev. Mod. Phys. 87 (3) (2015) 925.
[4] D. Li, B. Fu, Y. Wang, G. Lu, Y. Berezin, H.E. Stanley, S. Havlin, Percolation transition in dynamical traffic network with evolving critical bottlenecks, Proc. Nat. Acad. Sci. 112 (3) (2015) 669–672.
[5] C. Ma, T. Zhou, H.F. Zhang, Playing the role of weak clique property in link prediction: a friend recommendation model, Sci. Rep. 6 (2016) 30098.
[6] F. Xie, Z. Chen, J. Shang, X. Feng, W. Huang, J. Li, A link prediction approach for item recommendation with complex number, Knowl. Based Syst. 81 (2015) 148–158.
[7] M. Pavlov, R. Ichise, Finding experts by link prediction in co-authorship networks, Proceedings of the Second International Conference on Finding Experts on the Web with Semantics, ACM, New York, 2007, pp. 42–55.
[8] H. Hu, C. Zhu, H. Ai, L. Zhang, J. Zhao, Q. Zhao, H. Liu, Lpi-etslp: lncrna-protein interaction prediction using eigenvalue transformation-based semi-supervised link prediction, Mol Biosyst 13 (9) (2017) 1781–1787.
[9] B.C. De, E.A. Power, D.B. Larremore, C. Moore, Community detection, link prediction, and layer interdependence in multilayer networks, Phys. Rev. E 95 (4-1) (2017) 042317.
[10] Y. Lu, Y. Guo, A. Korhonen, Link prediction in drug-target interactions network using similarity indices, BMC Bioinf. 18 (1) (2017) 39.
[11] T. Man, H. Shen, S. Liu, X. Jin, X. Cheng, Predict anchor links across social networks via an embedding approach, Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, ACM, New York, 2016, pp. 1823–1829.
[12] Z. Wang, J. Liang, R. Li, Y. Qian, An approach to cold-start link prediction: establishing connections between non-topological and topological information, IEEE Trans. Knowl. Data Eng. 28 (11) (2016) 2857–2870.
[13] F. Li, J. He, G. Huang, Y. Zhang, Y. Shi, R. Zhou, Node-coupling clustering approaches for link prediction, Knowl. Based Syst. 89 (2015) 669–680.
[14] J. Ding, L. Jiao, J. Wu, F. Liu, Prediction of missing links based on community relevance and ruler inference, Knowl. Based Syst. 98 (2016) 200–215.
[15] A.K. Menon, C. Elkan, Link prediction via matrix factorization, Proceedings of the Twenty-Second European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, Berlin, 2011, pp. 437–452.
[16] S. Zhai, Z. Zhang, Dropout training of matrix factorization and autoencoder for link prediction in sparse graphs, Proceedings of the Fifteenth International Conference on Data Mining, SIAM, Philadelphia, PA, 2015, pp. 451–459.
[17] P. Wang, B. Xu, Y. Wu, X. Zhou, Link prediction in social networks: the state-of-the-art, Sci. China Inf. Sci. 58 (1) (2014) 1–38.
[18] B. Ermis, E. Acar, A.T. Cemgil, Link prediction in heterogeneous data via generalized coupled tensor factorization, Data Min Knowl Discov 29 (1) (2013) 203–236.
[19] D. Liben-Nowell, J. Kleinberg, The link prediction problem for social networks, Journal of the American Society for Information Science and Technology 58 (7) (2007) 1019–1031.
[20] M.A. Hasan, V. Chaoji, S. Salem, M. Zaki, Link prediction using supervised learning, Proceedings of the SDM'06 Workshop on Link Analysis, Counter Terrorism and Security, ACM, Bethesda, USA, 2006.
[21] L. Lü, T. Zhou, Link prediction in complex networks: a survey, Phys. A 390 (6) (2011) 1150–1170.
[22] V. Martnez, F. Berzal, J.C. Cubero, A survey of link prediction in complex networks, ACM Comput. Surv. 49 (4) (2016) 69.
[23] G. Golub, W. Kahan, Calculating the singular values and pseudo-inverse of a matrix, Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis 2 (2) (1965) 205–224.
[24] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, Proceedings of the Fourteenth International Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, 2001, pp. 556–562.
[25] A. Mnih, R. Salakhutdinov, Probabilistic Matrix Factorization, Proceedings of the Twentieth International Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, 2007, pp. 1257–1264.
[26] B. Chen, F. Li, S. Chen, R. Hu, L. Chen, Link prediction based on non-negative matrix factorization, PLoS ONE 12 (8) (2017). E0182968
[27] L. Zhu, D. Guo, J. Yin, G.V. Steeg, A. Galstyan, Scalable temporal latent space inference for link prediction in dynamic social networks, IEEE Trans. Knowl. Data Eng. 28 (10) (2016) 2765–2777.
[28] Q. Yang, E. Dong, Z. Xie, Link prediction via nonnegative matrix factorization enhanced by blocks information, Proceedings of the International Conference on Natural Computation, (2014), pp. 823–827.
[29] D. Song, D.A. Meyer, M.R. Min, Fast nonnegative matrix factorization with rank-one admm, Proceedings of the NIPS Workshop on Optimization for Machine Learning, (2014).
[30] Y.X. Zhu, L. Lü, Q.M. Zhang, T. Zhou, Uncovering missing links with cold ends, Phys. A 391 (22) (2012) 5769–5778.
[31] T. Zhou, L. Lü, Y.C. Zhang, Predicting missing links via local information, Eur. Phys. J. B Condens. Matter Complex Syst. 71 (4) (2009) 623–630.
[32] E. Leicht, P. Holme, M.E. Newman, Vertex similarity in networks, Phys. Rev. E 73 (2) (2006) 026120.
[33] L.A. Adamic, E. Adar, Friends and neighbors on the web, Soc. Netw. 25 (3) (2003) 211–230.
[34] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.L. Barabási, Hierarchical organization of modularity in metabolic networks, Science 297 (5586) (2002) 1551–1555.
[35] A.L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.
[36] F. Lorrain, H.C. White, Structural equivalence of individuals in social networks, J. Math. Sociol. 1 (1) (1971) 49–80.
[37] P. Jaccard, Etude de la distribution florale dans une portion des alpes et du jura, Bulletin De La Societe Vaudoise Des Sciences Naturelles 37 (142) (1901) 547–579.
[38] H.H. Chen, L. Gou, X.L. Zhang, C.L. Giles, Discovering missing links in networks using vertex similarity measures, Proceedings of the Twenty-Seventhth Annual ACM Symposium on Applied Computing, ACM, Trento, Italy, 2012, pp. 138–143.
[39] A. Papadimitriou, P. Symeonidis, Y. Manolopoulos, Fast and accurate link prediction in social networking systems, J. Syst. Soft. 85 (9) (2012) 2119–2132.
[40] L. Lü, C.H. Jin, T. Zhou, Similarity index based on local paths for link prediction of complex networks, Phys. Rev. E 80 (4) (2009) 046122.
[41] L. Katz, A new status index derived from sociometric analysis, Psychometrika 18 (1) (1953) 39–43.
[42] F. Fouss, A. Pirotte, J.M. Renders, M. Saerens, Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation, IEEE Trans. Knowl. Data Eng. 19 (3) (2007) 355–369.
[43] V.D. Blondel, A. Gajardo, M. Heymans, P. Senellart, P.V. Dooren, A measure of similarity between graph vertices: applications to synonym extraction and web searching, SIAM Rev. 46 (4) (2004) 647–666.
[44] G. Jeh, J. Widom, Simrank: a measure of structural-context similarity, Proceedings of the Eighteenth International Conference on Knowledge Discovery and Data Mining, ACM, Edmonton, Canada, 2002, pp. 538–543.
[45] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Comput. Netw. ISDN Syst. 30 (1) (1998) 107–117.
[46] F. Göbel, A. Jagers, Random walks on graphs, Stoch Process Appl. 2 (4) (1974) 311–336.
[47] W. Liu, L. Lu, Link prediction based on local random walk, Europhys. Lett. 89 (5) (2010). 58007–58012(6)
[48] I. Günes, S. Gündüz-Ogüdücü, Z. Cataltepe, Link prediction using time series of neighborhood-based node similarity scores, Data Min. Knowl. Discov. 30 (1) (2015) 1–34.
[49] R.N. Lichtenwalter, N.V. Chawla, Vertex collocation profiles: subgraph counting for link analysis and prediction, Proceedings of the Twenty-First International

Conference on World Wide Web, ACM, Lyon, France, 2012, pp. 1019–1028.
[50] H.R. De Sá, R.B. Prudêncio, Supervised link prediction in weighted networks, Proceedings of the International Joint Conference on Neural Networks, IEEE, San Jose, USA, 2011, pp. 2281–2288.
[51] J. Leskovec, D. Huttenlocher, J. Kleinberg, Predicting positive and negative links in online social networks, Proceedings of the Nineteenth International Conference on World Wide Web, ACM, Raleigh, USA, 2010, pp. 641–650.
[52] K.Y. Chiang, N. Natarajan, A. Tewari, I.S. Dhillon, Exploiting Longer Cycles for Link Prediction in Signed Networks, Proceedings of the Twentieth ACM International Conference on Information and Knowledge Management, ACM, Glasgow, UK, 2011, pp. 1157–1162.
[53] S. Scellato, A. Noulas, C. Mascolo, Exploiting place features in link prediction on location-based social networks, Proceedings of the Seventeenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, California, USA, 2011, pp. 1046–1054.
[54] T. Wohlfarth, R. Ichise, Semantic and event-based approach for link prediction, Practical Aspects of Knowledge Management, Springer, Yokohama, Japan, 2008, pp. 50–61.
[55] M. Rowe, M. Stankovic, H. Alani, Who will follow whom? Exploiting semantics for link prediction in attention-information networks, Proceedings of the Eleventhth International Conference on the Semantic Web, Springer, Berlin, 2012, pp. 476–491.
[56] A. Clauset, C. Moore, M.E. Newman, Hierarchical structure and the prediction of missing links in networks, Nature 453 (7191) (2008) 98–101.
[57] P.W. Holland, K.B. Laskey, S. Leinhardt, Stochastic blockmodels: first steps, Soc Netw. 5 (2) (1983) 109–137.
[58] E.M. Airoldi, D.M. Blei, S.E. Fienberg, E.P. Xing, Mixed membership stochastic blockmodels, J. Mach. Learn. Res. 9 (2008) 1981–2014.
[59] K. Palla, D. Knowles, Z. Ghahramani, An infinite latent attribute model for network data, Proceedings of the Twenty-Ninth International Conference on Machine Learnin, Edinburgh, Scotland, 2012.
[60] K. Miller, M.I. Jordan, T.L. Griffiths, Nonparametric latent feature models for link prediction, Proceedings of the Twenty-Second International Conference on Neural Information Processing Systems, Vancouver, Canada, 2009, pp. 1276–1284.
[61] J. Zhu, Max-margin Nonparametric Latent Feature Models for Link Prediction, Proceedings of the Twenty-Ninth International Conference on Machine Learning, Edinburgh, Scotland, 2012.
[62] M. Kim, J. Leskovec, Modeling social networks with node attributes using the multiplicative attribute graph model, Proceedings of the Twenty-Seventhth Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, 2011.
[63] J. Zhang, B. Liu, J. Tang, T. Chen, J. Li, Social influence locality for modeling retweeting behaviors, Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, AAAI Press, Beijing, China, 2013, pp. 2761–2767.
[64] L. Backstrom, J. Leskovec, Supervised random walks: predicting and recommending links in social networks, Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, ACM, 2011, pp. 635–644.
[65] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (roc) curve, Radiology 143 (1) (1982) 29–36.
[66] G.H. Golub, C. Reinsch, Singular value decomposition and least squares solutions, Numerische Mathematik 14 (5) (1970) 403–420.
[67] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788–791.
[68] Q. Ou, Y.D. Jin, T. Zhou, B.H. Wang, B.Q. Yin, Power-law strength-degree correlation from resource-allocation dynamics on weighted networks, Phys. Rev. E Stat. Nonlinear Soft Matter Phys. 75 (2007) 021102.
[69] J. Valverde-Rebaza, A.D.A. Lopes, Exploiting behaviors of communities of twitter users for link prediction, Soc. Netw. Anal. Min. 3 (4) (2013) 1063–1074.