SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

November 2018, Vol. 61 112103:1–112103:19
https://doi.org/10.1007/s11432-017-9334-3

Correlations multiplexing for link prediction in multidimensional network spaces

Yunpeng XIAO*, Xixi LI, Yuanni LIU, Hong LIU & Qian LI

Chongqing Engineering Laboratory of Internet and Information Security, Chongqing University of Posts and Telecommunications, Chongqing 400065*, China*

Received 16 May 2017/Revised 8 October 2017/Accepted 28 December 2017/Published online 8 June 2018

Abstract In social networks, link establishment among users is affected by diversity correlations. In this paper, we study the formation of links, map correlations into multidimensional network spaces and apply their behavioral and structural features to the problem of link prediction. First, by exploring user behavioral correlation and network structural correlation, we map them into three network spaces: following space, interaction space and structure space. With a hierarchical process, the coupling relationship between the spaces can be reduced and we can analyze the correlation in each space separately. Second, by taking advantage of the latent Dirichlet allocation (LDA) topic model for dealing with the polysemy and synonym problems, the traditional text modeling method is improved by Gaussian weighting and applied to user behavior modeling. In this way, the expression ability of topics can be enhanced, and improved topic distribution of user behavior can be obtained to mine user correlations in the following space and the interaction space. Moreover, the method can be extended using the hidden naive Bayesian algorithm which is good at reducing attribute independence. By quantifying the dependencies between common neighbors, we can analyze user correlations in the structure space and multiplex the correlations of the other two spaces to link prediction. The experimental results indicate that the method can effectively improve link prediction performance.

Keywords social networks, multidimensional network spaces, link prediction, LDA, hidden naive Bayes

Citation Xiao Y P, Li X X, Liu Y N, et al. Correlations multiplexing for link prediction in multidimensional network spaces. Sci China Inf Sci, 2018, 61(11): 112103, https://doi.org/10.1007/s11432-017-9334-3

1 Introduction

With the development of the Internet, social networks have gradually become a crucial way for people to communicate and share information. Mining the contents of these massive data has become a research hotspot. Among the various researches, the study of user relationships in networks can help better explain the evolution of network structures.

The link prediction problem [1] can summarize the above research and it has attracted particular interest. From Refs. [2, 3], link prediction can be understood as inferring feasible missing links and future links based on observable user information and network structures. The missing link problem has important ramifications because missing links can alter estimates of network-level statistics [4] , and our research focus is missing link prediction. Link prediction not only can help us understand the evolution mechanism of networks [5], but also has crucial application value in many fields [6–9]. Although link prediction has important research significance and positive research results have been achieved, some challenges still remain.

* Corresponding author (email: xiaoyp@cqupt.edu.cn)

On the one hand, user relationships in social networks constitute complex network structures. In 2007, several similarity indices were summed up by Nowell and Kleinberg [1], and the existence of links was predicted through common neighbors or network paths. Subsequently, a variety of link prediction methods based on network structures were proposed [10, 11], and the prediction results were improved. However, owing to the complexity of user relationships, simple networks cannot adequately reflect the true situation of social networks. As an example, the relationships among users are not independent of each other, and the aforementioned methods are not fully applicable to current social networks. The performance of link prediction needs to be improved.

On the other hand, the wide use of social networks generates massive user attribute information. In recent researches, to improve link prediction performance, methods combined with user attribute information were introduced by scholars, such as interest keywords [12] and text analysis [13, 14]. Compared to the direct characterization of network structures, user attributes can indirectly reflect potential user relationships. However, because of huge amount of information in networks, it is time-consuming to compute the similarity of information or establish models [15]. The efficiency of link prediction needs to be improved.

In addition, a lot of methods for link prediction in social networks consider only topological features and attributes, few studies take social features into consideration. The social features are useful for explaining the mechanisms of social activities. Incorporating social features into the link prediction methods would be promising [16]. Therefore, integrating network structure features and social features is an effective method of link prediction.

In order to analyze network structure features and social features respectively and integrate them effectively, a social network is mapped into multidimensional network spaces. This hierarchical process can reduce the coupling relationship between social features and network structure features. Taking the advantage of LDA (latent Dirichlet allocation) [17, 18] topic model in dealing with the problem of "polysemy" and "synonym", the traditional text modeling method is improved by Gaussian weighting and applied into user behavior modeling. Meanwhile, network structure features can be analyzed by hidden naive Bayesian algorithm which is good at reducing attribute independence. By extended with hidden naive Bayesian algorithm, our method can analyze user correlations in multidimensional network spaces and multiplex them to link prediction. We use two real datasets, Sina micro-blog and Twitter, to verify the effectiveness of our method. Experimental results indicate that our method can improve the performance of link prediction.

Our contribution can be summarized as follows.

(1) In order to analyze network structure features and social features respectively and integrate them effectively, a social network is mapped into multidimensional network spaces: following space, interaction space and structure space. By analyzing and multiplexing user correlations in these network spaces, we can effectively predict links among users.

(2) Owing to the power-law characteristics of user behavior, the traditional LDA text modeling method is improved by Gaussian weighting and applied to user behavior modeling. In this way, the improved topic distribution of user behavior can be obtained to mine user correlations in the following space and the interaction space.

(3) Given the insufficiency of LDA topic model on reflecting the contribution of network structures and the advantage of hidden naive Bayesian algorithm on reducing attribute independence, the method can be extended using hidden naive Bayesian algorithm. By quantifying the dependencies between common neighbors, we can analyze user correlations in the structure space and multiplex the correlations of the other two spaces to link prediction.

Organization. The rest of this paper is organized as follows. Section 2 introduces the related work. Section 3 formulates the problems and gives the necessary definitions. In Section 4, we explain the proposed method and describe the method learning algorithm. Section 5 presents and analyzes the experimental results of the method. Finally, Section 6 concludes the paper.

2 Related work

In social networks, link establishment among users depends mainly on user correlations. These correlations can be achieved by learning network structures and user attribute information, and we can analyze user correlations for link prediction. Various methods such as similarity computation [19, 20], random walks [21, 22], topic model, and probability model [23, 24] are used for prediction. This section focuses on the research on link prediction and application of the LDA topic model in recent years.

Prediction based on network structures is achieved by analyzing user similarity with partial or integral information about network structures. Such methods [1] assume that the more similarity between users, the greater posibility they will establish links. Sett et al. [20] proposed a weighted model based on an existing similarity index, and corresponding weights were added to improve link prediction performance. Li et al. [25] presented a new prediction algorithm by combining the various roles of users with common neighbor similarity indices. The algorithm could improve link prediction performance and effectiveness. Martínez et al. [26] proposed an adaptive degree penalization link prediction method to exploit the existence of a relationship between the best-performing degree of penalization for shared neighbors and the network clustering coefficient. Although these methods use network structures well, they do not consider the interdependence between common neighbors and ignore user attribute information, such as user behavior information, which can be used to measure user relationships.

Prediction based on user attribute information aims to find related friends by establishing a relevant model with user attribute information. Such methods [27] are based on the view that people who have similar hobbies, language, culture or geographical information are more likely to become friends. Scellato et al. [28] proposed a link prediction system. By analyzing user location information, they found that users who accessed the same places were more likely to become friends. Shahmohammadi et al. [29] proposed three new algorithms that employed collaborative filtering methods by weighting activities (e.g., comments, information shared and forwarded) to existing networks. These algorithms can recommend users with user activities to the target users. Liu et al. [30] presented a simple but effective similarity-based prediction strategy based on label propagation, which imitated natural communication between people to link prediction. Although these methods improve link prediction performance, they are time-consuming in terms of analyzing user attribute information and establishing models. Therefore, the efficiency of link prediction must be optimized.

In addition, link prediction methods leverage behavior analysis [31, 32] gradually becomes a new research perspective. Cha et al. [33] applied LDA to deal with community discovery problems and recommended friends. Chang et al. [34] presented a relational topic models (RTM) with the text data and analyzed the topic distribution of texts to predict links among the texts. By analyzing the user's own attributes and behavior, Cho et al. [18] established a potential spatial model based on LDA to predict user relationships. LDA topic model is widely applied with dimension reduction to reduce the complexity and traditional LDA text modeling method can be introduced into user behavior modeling to link prediction. Meanwhile, based on naive Bayesian algorithm, the attribute independences can be reduced [35] and the ideas can be introduced in link prediction. Jiang et al. [36] proposed weighted average of one-dependence estimators (WAODE), which is an improved naive Bayes algorithm. By assigning different weights to these one-dependence classifiers, attribute independence is weakened and the classification effect has been improved. Based on it, they proposed structure extended multinomial naive Bayes (SEMNB) [37], which can learn algorithm effectively without structure searching. The above-mentioned researches are based on LDA or Bayesian algorithm, and they have been applied in increasing scenes with its favorable generalization ability and strong scalability. If we can make use of the complementarity of them, it will helpful to analyze the correlation of different features.

3 Problem definition

3.1 Related definitions

The problem at hand here is predicting links among users by analyzing user behavior and user relationships. In this paper, a social network can be represented as an undirected network $G = (U, E)$, where U

represents the set of users and $E \subset U \times U$ represents the set of undirected links. The cardinality $|U| = N$ is used to denote the total number of whole network users and $|E| = M$ is used to denote the total number of whole network links. For predicting the existence of links among users, some basic concepts and related definitions are introduced.

Definition 1. Following network space $G^f = (U, E^f)$. G^f is a following network space that represents relationships based on users' following behavior. It is also called following space. $E^f \subset U \times U$ represents the set of directed following edges, and $|E^f| = M^f$ (i.e., total number of links in following network space).

Definition 2. Interaction network space $G^i = (U, E^i)$. G^i is an interaction network space that represents relationships based on users' interaction behavior. It is also called interaction space. $E^i \subset U \times U$ represents the set of directed interaction edges, and $|E^i| = M^i$ (i.e., total number of links in interaction network space).

Definition 3. Structure network space $G^s = (U, E^s)$. G^s is a structure network space that represents users' real relationships (a real relationship is represented by bidirectional following edges). It is also called structure space. In other words, the structure space denotes a network G. $E^s \subset U \times U$ represents the set of undirected edges, and $|E^s| = M^s$ (i.e., total number of links in structure network space).

3.2 Problem formulation

To formally formulate the problem of our research, let $G = (U, E)$ be the whole network, and let $A = \{(a, u_i)|u_i \in U\}$ represent the behavior information of all users. Based on the relevant definitions in Subsection 3.1, the network can be mapped into multidimensional network spaces: G^f, G^i, G^s, and we can use our method to predict missing links E^* in network $G = (U, E)$. Specifically, the problem is defined as follows:

$$\left.\begin{array}{c} G \to G^f, G^i, G^s \\ A \end{array}\right\} \Rightarrow f : (G^f, G^i, G^s) \to E^*.$$

3.2.1 *Problem input*

Given the related definitions, the input to this problem can be defined as follows:

(1) Whole network $G = (U, E)$;

(2) Behavior information of all users $A = \{(a, u_i)|u_i \in U\}$, which represents an action a of a user u_i (action in this study refers to publishing or forwarding information).

3.2.2 *Problem output*

Based on the above description, the problems to be solved are as follows.

How can we model and predict links among users? A whole network G is mapped into multidimensional network spaces: G^f, G^i, G^s. We use the proposed method to mine user correlations from every space. A parameter set λ is introduced in our link prediction method. By multiplexing user correlations and finding the optimal set of parameters λ^*=argmax$_\lambda P_\lambda(E|G^f, G^i, G^s, A)$, the missing links in network G can be predicted: E^*=argmax$_{\lambda^*} P_{\lambda^*}(E|G^f, G^i, G^s, A)$.

4 Proposed method

To solve the above problems, we propose a link prediction method based on user information, behavior, and relationships. The details of this method are introduced in three modules: correlation quantification, user behavior and network structure modeling and link prediction, as shown in Figure 1. In the first module, related attributes are considered for correlation quantification in every space, and multiple driving mechanisms are proposed to represent them. In the second module, the prediction features of the correlations are obtained based on LDA and the hidden naive Bayesian algorithm. In the third module, multiplexing prediction features and the link prediction method can be proposed to predict links among users.

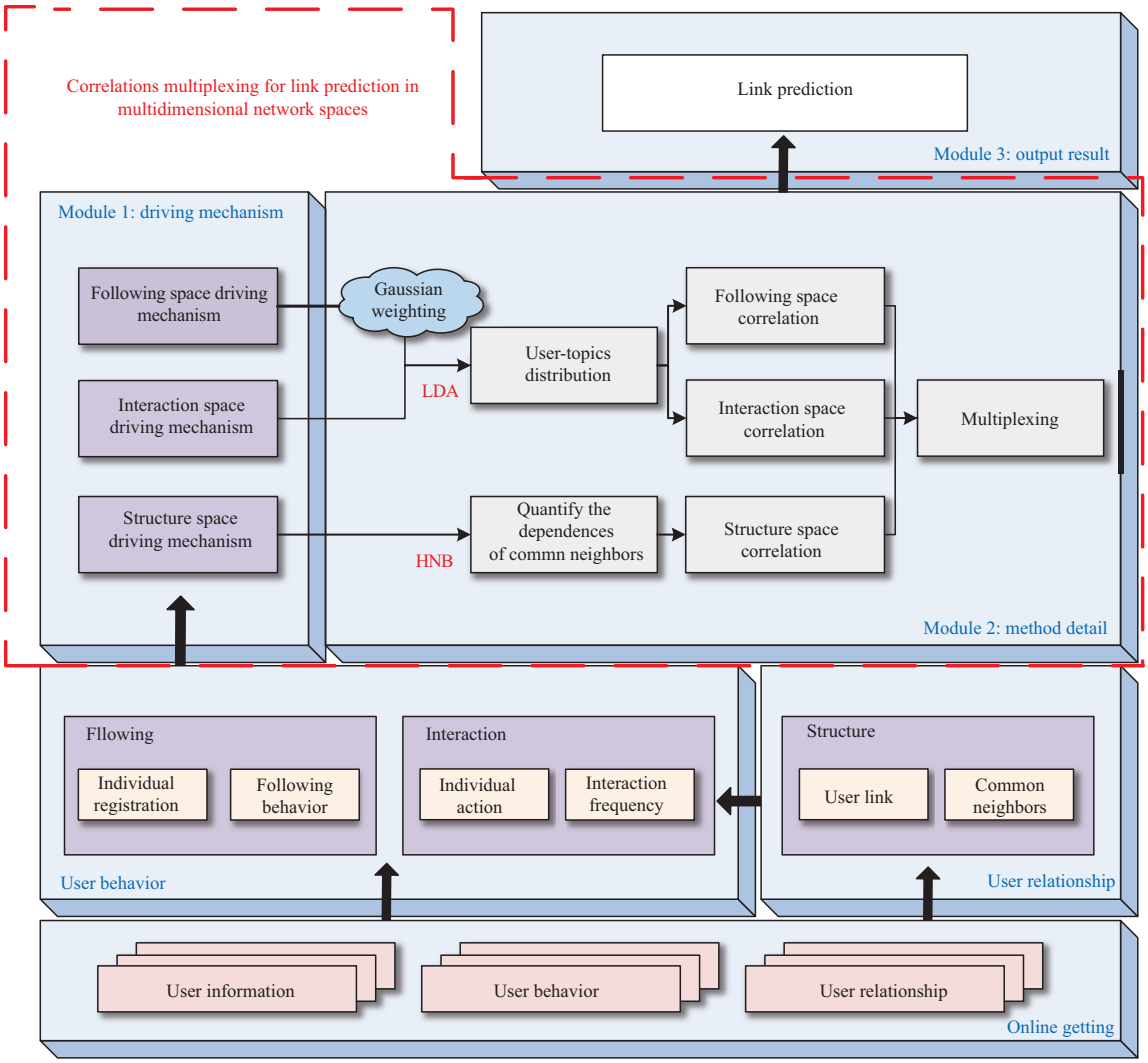

Figure 1 (Color online) Link prediction framework.

4.1 Correlation quantification

In the first module, a social network is mapped into multidimensional network spaces: following space, interaction space and structure space, as shown in Figure 2. To analyze user correlations in every space, we extract a few attributes for correlation quantification. Moreover, we define multiple driving mechanisms to represent the classes of the correlations.

4.1.1 *Following space driving mechanism*

In the following space, link establishment is affected by user following behavior. Generally, we hold the view that user following behavior can reflect user interests. Namely, users are more likely to follow another user if they share a greater number of common interests. The user's following set is defined as follows:

$$\boldsymbol{F}(u_x) = \boldsymbol{f_x} = [f_{x,1}, f_{x,2}, \ldots, f_{x,N_x}], \tag{1}$$

where $f_{x,n} \in U$ denotes a user followed by user u_x, also called followed user (each followed user in this set has no particular order). N_x is the number of followed users. Taking into account that user following behavior is one-way, that is, user A follows user B, and user B does not necessarily follow user A. This is also the reason why the edges in the following space are directed edges. As an example, there are some followed users $b, c, d, e, \ldots$ for user g, and the following set of user g is $\boldsymbol{F}(g) = [b, c, d, e, \ldots]$.

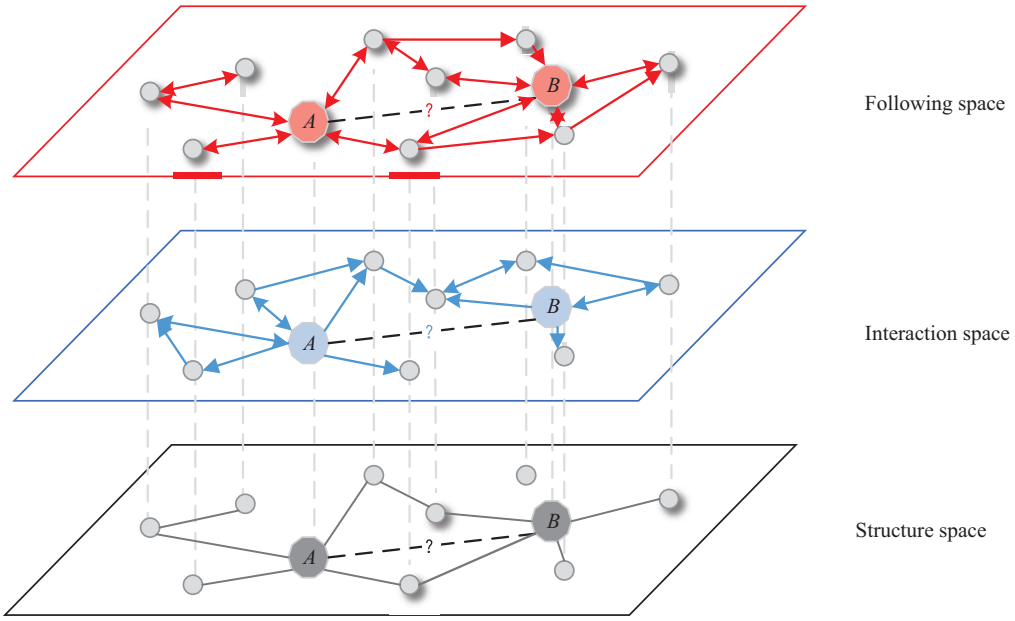

Figure 2 (Color online) Multidimensional network spaces.

4.1.2 *Interaction space driving mechanism*

In the interaction space, link establishment is affected by user interaction behavior. In other words, the frequency of interaction between two users indirectly influences link establishment between them. User interaction can be understood as forwarding information among users. In this paper, the user's interaction set is defined as follows:

$$\boldsymbol{I}(u_x) = \boldsymbol{i_x} = [i_{x,1}, i_{x,2}, \ldots, i_{x,N'_x}], \tag{2}$$

where $i_{x,n} \in U$ denotes a user interacted with user u_x, also called interacted user (each interacted user in this vector has no particular order). N'_x is the number of interacted users. The user interaction behavior is also one-way, that is, user A forwards information from user B, and user B does not necessarily forward information from user A. Therefore the edges in the interaction space are directed edges. As an example, user h forwards information from user j 2 times and forwards information from user k 3 times, so the interaction set of user h is $\boldsymbol{I}(h) = [j, j, k, k, k, \ldots]$.

4.1.3 *Structure space driving mechanism*

In the structure space, link establishment is influenced by network structures in social networks. The more common friends you have, the greater is the possibility of the existence of links. The structure set is defined as follows:

$$\boldsymbol{S}(u_x, u_y) = \boldsymbol{s_{xy}} = [s_1, s_2, \ldots, s_{Q_{xy}}], \tag{3}$$

where $s_q \in U$ denotes a common friend of user u_x and user u_y, also called common neighbor (each common neighbor in this vector has no particular order). Q_{xy} is the number of common neighbors. We believe that only following each other can be called friends, that is to say, the edges in the structure space are undirected edges. As an example, $l, o, p, v, \ldots$ are the common neighbors of users r and t, and the structure set of them is $\boldsymbol{S}(r, t) = [l, o, p, v, \ldots]$.

4.2 Link prediction

Given the three driving mechanisms defined in the previous module, the problem to be solved is how to incorporate those driving mechanisms into user behavior and network structures modeling. The next module presents the process of modeling, and it includes four steps: following space correlation analysis,

Table 1 Description of symbols in graphic model

Symbols	Description	Symbols	Description
N	Number of the users in G	$f_{x,n}, i_{x,n}$	The n-th followed or interacted user of u_x
N_x, N'_x	Number of followed or interacted users about u_x	l_{xy}	Existence of link between u_x and u_y
N_f, N_i	Number of followed or interacted users in G	s_q	The q-th common neighbor between u_x and u_y
K	Number of interest topics	η_q	Independent dependency weight of the q-th common neighbor between u_x and u_y
γ_i, γ'_i	Gaussian weight of the i-th followed or interacted user in G	π_q	Joint dependency weight of the q-th common neighbor between u_x and u_y
Q_{xy}	Number of common neighbors between u_x and u_y	R_1	Set of correlation in following space G^f
$\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\alpha}', \boldsymbol{\beta}'$	Dirichlet priors	R_2	Set of correlation in interaction space G^i
$\boldsymbol{\theta_x}, \boldsymbol{\theta'_x}$	A topic distribution of u_x	R_3	Set of correlation in structure space G^s
$\boldsymbol{\psi_k}, \boldsymbol{\psi'_k}$	A behavior distribution of topic k	τ	Learning rate for driving mechanism
$z_{x,n}, z'_{x,n}$	The n-th interest topic of u_x	Y	Set of the existence of links in G

interaction space correlation analysis, structure space correlation analysis and multiplexing. Corresponding to the different driving mechanisms, the prediction features of the correlations are obtained. The existence of links among users can be predicted by multiplexing these prediction features in the last module.

4.2.1 *Following space correlation analysis*

In the following space, we focus on the analysis of following behavior. Taking advantage of the LDA topic model in dealing with polysemy and synonym problems, the traditional text modeling method is improved by Gaussian weighting and applied to user behavior modeling. In other words, a user is regarded as a document and user behavior is regarded as vocabulary. By introducing interests as topics, we can mine potential interest relationships.

Assume that the set of users is $U = \{u_1, u_2, \ldots, u_N\}$, where N is the number of users in a whole network G. Each user can be understood as the component of followed users, which can also be expressed as its following vector. Each followed user obeys a multinomial distribution of interest topic $z_{x,n}$ and each interest topic $z_{x,n}$ obeys a multinomial distribution of user u_x.

Owing to the power-law characteristics of users' following behavior, the topic distribution will be inclined toward high-frequency users. Meanwhile, Gaussian weighting filtering is commonly used to remove noise. Thus, the standard LDA is improved by Gaussian weighting. For each user, the Gaussian weighted formula is used to weight the followed users, and it can be written as follows:

$$\gamma_{f_{x,n}} = \exp\left(-\frac{(f_{f_{x,n}} - f_i)^2}{2\sigma^2}\right), \tag{4}$$

where $f_{f_{x,n}}$ is the frequency of followed user $f_{x,n}$. f_i is the average frequency of all followed users. Given parameter K as the number of interest topics, improvement of the generation process of the LDA model by using Gaussian weighting can be described as follows:

(1) For each interest topic k, draw $\boldsymbol{\psi_k} \sim \mathrm{Dir}(\boldsymbol{\beta})$;

(2) Given the x-th user u_x, in whole network G, draw $\boldsymbol{\theta_x} \sim \mathrm{Dir}(\boldsymbol{\alpha})$;

(3) For the n-th followed user in the x-th user $f_{x,n}$: (a) draw an interest topic $k = z_{x,n} \sim \mathrm{Mult}(\boldsymbol{\theta_x})$; (b) draw a followed user $f_{x,n} \sim \mathrm{Mult}(\boldsymbol{\psi_{z_{x,n}}}|\gamma_{f_{x,n}})$.

In this study, $\mathrm{Dir}(\cdot)$ denotes Dirichlet distribution and $\mathrm{Mult}(\cdot)$ denotes multinomial distribution. $\boldsymbol{\theta_x}$ is the topic distribution of user u_x and $\boldsymbol{\psi_{z_{x,n}}}$ is the following behavior distribution of interest topic $z_{x,n}$. The graphic model is shown in Figure 3 and the symbols are described in Table 1.

In Figure 3, the gray circles represent the adjustable parameters, the white circles represent the calculated variables and double circles represent observable variables. In the gray circles, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\alpha}'$, $\boldsymbol{\beta}'$ are the parameters that determine the topic distribution $\boldsymbol{\theta_x}$, $\boldsymbol{\theta'_x}$ and behavior distribution $\boldsymbol{\psi_k}, \boldsymbol{\psi'_k}$. Since the topic distribution and behavior distribution are multinomial distributions, we select conjugate priors

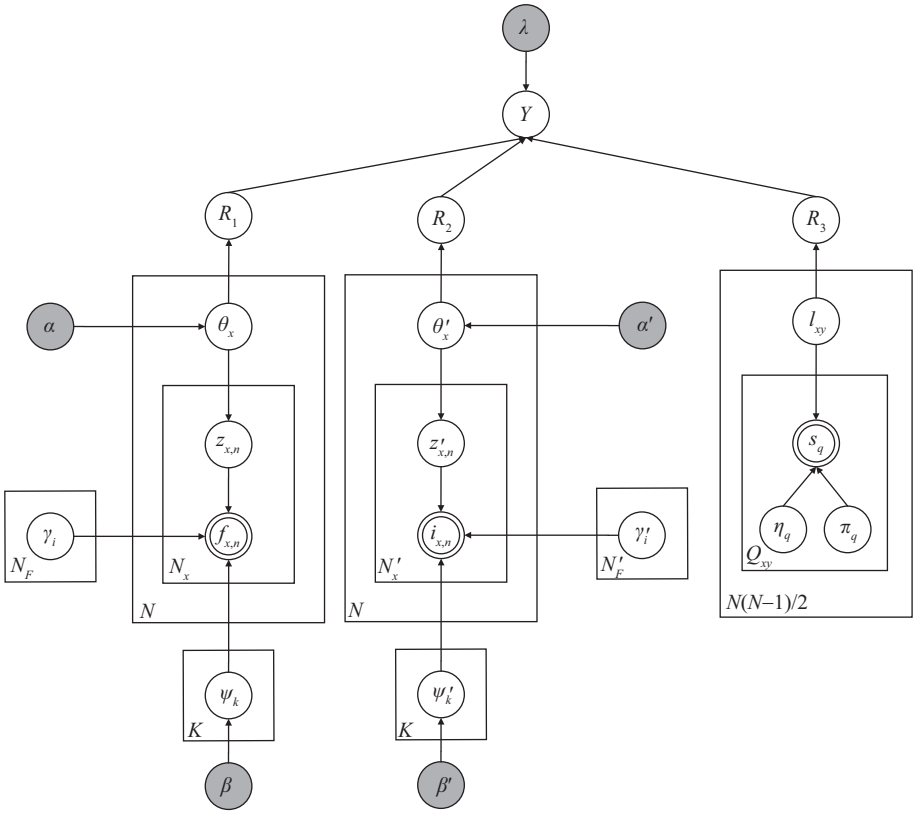

Figure 3 Graphic model.

of multinomial distributions as prior parameters for the convenience of computation. In other words, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\alpha}'$, $\boldsymbol{\beta}'$ are Dirichlet priors. τ is the learning rate for driving mechanism. By adjusting the learning rate, the weight of the driving mechanism can be converged.

Actually, the aim of user following behavior modeling is to compute topic distribution $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N]$ and following behavior distribution $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_K]$. Owing to the coupling of $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$, we cannot compute them directly and Gibbs sampling [38] is applied to indirectly get $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$. The principle of Gibbs sampling in terms of extracting topic z_i is as follows:

$$p(z_i = k|\boldsymbol{z}_{\neg \boldsymbol{i}}, \boldsymbol{f}) \propto p(z_i = k, f_i = t|\boldsymbol{z}_{\neg \boldsymbol{i}}, \boldsymbol{f}_{\neg \boldsymbol{i}}) = \widehat{\theta}_{x,k} \times \widehat{\psi}_{k,t} = \frac{n^{(k)}_{x,\neg i} + \alpha_k}{\sum_{k=1}^{K} n^{(k)}_{x,\neg i} + \alpha_k} \times \frac{n^{(t)}_{k,\neg i} + \beta_t}{\sum_{t=1}^{N} n^{(t)}_{k,\neg i} + \beta_t}, \quad (5)$$

where $\boldsymbol{z}_{\neg \boldsymbol{i}}$ represents the topic of followed users except for the current followed user; $\boldsymbol{f}_{\neg \boldsymbol{i}}$ represents followed users except for the current followed user; $n^{(t)}_{k,\neg i}$ is the number of obtained Gaussian weighting for followed user t assigned to interest topic k with Gaussian weighting except for the current followed user; and $n^{(k)}_{x,\neg i}$ is the number of interest topic k assigned to user u_x except for the current followed user. When the sampling converges, $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$ can be obtained.

Then, the uncertainty in user interest topics is measured by entropy, and it is computed as follows:

$$H_x = -\sum_{k=1}^{K} \theta_{x,k} \log \theta_{x,k}, \quad (6)$$

where $\boldsymbol{\theta}_{\boldsymbol{x}} = [\theta_{x,1}, \theta_{x,2}, \ldots, \theta_{x,K}]$ is a topic distribution of user u_x. If a user's interests obey uniform distribution, the user's topic entropy is large and it reflects that the user's interests are wide. By contrast, if the user's interests focused on one or a few topics, the user's topic entropy is small and it reflects that the user's interests are concentrated. To compare the interest difference between users, we use the absolute value of topic entropy difference to measure the interest difference. The formula for

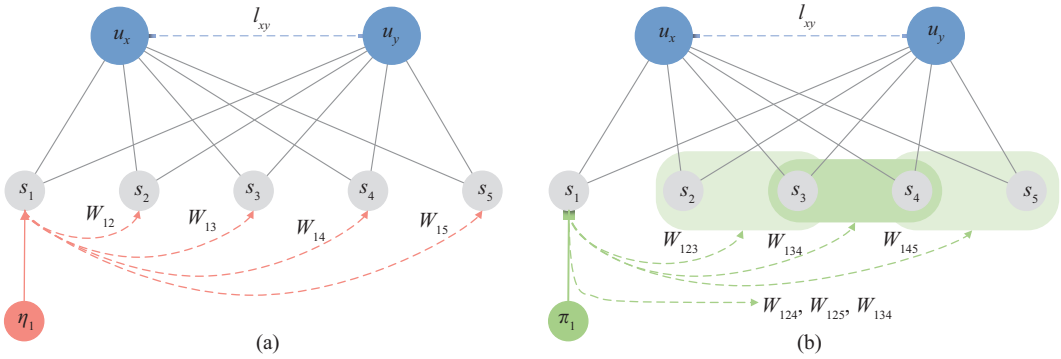

Figure 4 (Color online) Example of dependence. (a) Independent dependence; (b) joint dependence.

computing user following space correlation is as follows:

$$r_{1-xy} = 1 - |H_x - H_y|. \tag{7}$$

4.2.2 *Interaction space correlation analysis*

Based on the description in Subsection 4.2.1, we use a similar approach to analyze user interaction behavior. Given the same number of topics, we can also get $\boldsymbol{\Theta}'$ and $\boldsymbol{\Psi}'$ in the interaction space. Then, cosine similarity is used to measure interaction space correlation. The probability of two users' topics comes closer when their interaction behavior are more similar and the formula is as follows:

$$r_{2-xy} = \cos(u_x, u_y) = \frac{\sum_{k=1}^{K} \theta'_{x,k} \times \theta'_{y,k}}{\sqrt{(\sum_{k=1}^{K} (\theta'_{x,k})^2)(\sum_{k=1}^{K} (\theta'_{y,k})^2)}}, \tag{8}$$

where $\boldsymbol{\theta}'_{\boldsymbol{x}} = [(\theta'_{x,1}), (\theta'_{x,2}), \ldots, (\theta'_{x,K})]$ and $\boldsymbol{\theta}'_{\boldsymbol{y}} = [(\theta'_{y,1}), (\theta'_{y,2}), \ldots, (\theta'_{y,K})]$ represent the topic distribution of user u_x and u_y. The probability of two users' topic become closer when there is more similar interaction behavior between them.

4.2.3 *Structure space correlation analysis*

Hidden naive Bayes [35] is a naive Bayes algorithm with an added implicit factor for each attribute to represent the dependence between other attributes. In this paper, hidden naive Bayes is introduced to measure the contribution of common neighbors in the structure space. We assume that there are two types of dependence relationships: independent dependence and joint dependence. Independent dependence refers to the individual influence of common neighbor s_q affected by a common neighbor, and joint dependence refers to the joint influence of s_q affected by multiple common neighbors. We use l and $\overline{l}$ to represent the existence and absence of links respectively. The implicit factor η here is used to represent the summation of independent dependence, and the implicit factor π is used to represent the summation of joint dependence. For example, we assume that there are five common neighbors s_1, s_2, s_3, s_4, s_5 between the user u_x and u_y, and the structure set of them is $\boldsymbol{S}(u_x, u_y) = [s_1, s_2, s_3, s_4, s_5]$. The dependence between them as shown in Figure 4.

In Figure 4(a), common neighbor s_1 will be affected by single common neighbor s_2, s_3, s_4, s_5. The importance of independent influence are W_{12}–W_{15} and implicit factor η_1 is the summation of W_{12}–W_{15}. In Figure 4(b), common neighbor s_1 will also be affected by common neighbor pair $[s_2, s_3]$, $[s_2, s_4]$, $[s_2, s_5]$, $[s_3, s_4]$, $[s_3, s_5]$, $[s_4, s_5]$. The importance of joint influence are W_{123}–W_{145} and implicit factor π_1 is the summation of W_{123}–W_{145}.

Taking the structure set as a prior condition, the conditional probability can be computed as follows:

$$\begin{cases} P(l|\boldsymbol{s_{xy}}) = \dfrac{P(l)}{P(\boldsymbol{s_{xy}})} \prod\limits_{q=1}^{Q_{xy}} P(s_q|\eta_q, l)P(s_q|\pi_q, l), \\ P(\bar{l}|\boldsymbol{s_{xy}}) = \dfrac{P(\bar{l})}{P(\boldsymbol{s_{xy}})} \prod\limits_{q=1}^{Q_{xy}} P(s_q|\eta_q, \bar{l})P(s_q|\pi_q, \bar{l}). \end{cases} \tag{9}$$

In this paper, we use the ratio of conditional probabilities to measure structure space correlation. The formula for this computation is as follows:

$$r_{3-xy} = \log_2 \frac{P(l|\boldsymbol{s_{xy}})}{P(\bar{l}|\boldsymbol{s_{xy}})} = \log_2 \frac{P(l)}{P(\bar{l})} \prod_{q=1}^{Q_{xy}} \frac{P(s_q|\eta_q, l)P(s_q|\pi_q, l)}{P(s_q|\eta_q, \bar{l})P(s_q|\pi_q, \bar{l})}, \tag{10}$$

where $P(l)$ and $P(\bar{l})$ represent the probability of link existence and absence respectively, which are computed as follows:

$$\begin{cases} P(l) = \dfrac{2M}{N(N-1)}, \\ P(\bar{l}) = 1 - \dfrac{2M}{N(N-1)}. \end{cases} \tag{11}$$

In (10), $P(s_q|\eta_q, l)$ and $P(s_q|\eta_q, \bar{l})$ represent independent dependence under the existence and absence of links, respectively; $P(s_q|\pi_q, l)$ and $P(s_q|\pi_q, \bar{l})$ represent joint dependence under the existence and absence of links. The formulas for computing the dependence between common neighbors are as follows:

$$\begin{cases} P(s_q|\eta_q, l) = \sum\limits_{j=1, j\neq q}^{Q_{xy}} W_{qj} \times P(s_q|s_j, l), \\ P(s_q|\pi_q, l) = \sum\limits_{j=1, j\neq q}^{Q_{xy}} \sum\limits_{k\neq j, k\neq q}^{Q_{xy}} W_{qjk} \times P(s_q|[s_j, s_k], l). \end{cases} \tag{12}$$

In (12), $P(s_q|s_j, l)$ and $P(s_q|[s_j, s_k], l)$ represent the contribution of user s_j or pair $[s_j, s_k]$, and they can be expressed as the reciprocal of user degree $P(s_q|s_j, l)=\frac{1}{D_{s_j}}$, $P(s_q|[s_j, s_k], l)=\frac{1}{D_{s_j} \times D_{s_k}}$. D_{s_j} and D_{s_k} represent the degree of common neighbor s_j and s_k. W_{qj} and W_{qjk} represent the importance of independent dependence and joint dependence among common neighbors, respectively. Conditional mutual information weighted summation is used to represent these dependencies, and they can be computed as follows:

$$\begin{cases} W_{qj} = \dfrac{I_p(s_q, s_j|l)}{\sum_{j=1, j\neq q}^{Q_{xy}} I_p(s_q, s_j|l)}, \\ W_{qjk} = \dfrac{I_p(s_q, [s_j, s_k]|l)}{\sum_{j=1, j\neq q}^{Q_{xy}} \sum_{k\neq j, k\neq q}^{Q_{xy}} I_p(s_q, [s_j, s_k]|l)}. \end{cases} \tag{13}$$

$P(s_q|\eta_q, \bar{l})$ and $P(s_q|\pi_q, \bar{l})$ are computed using formulas similar to the ones presented above. To reduce the computational complexity of conditional mutual information computation, the following decision criterion is employed. If the influence of the implicit factor π is larger than that of the implicit factor η, we consider the joint influence of π and η. Otherwise, we consider only the influence of η, that is

$$r_{3-xy} = \begin{cases} \log_2 \dfrac{P(l)}{P(\bar{l})} \prod_{q=1}^{Q_{xy}} \dfrac{P(s_q|\eta_q, l)P(s_q|\pi_q, l)}{P(s_q|\eta_q, \bar{l})P(s_q|\pi_q, \bar{l})}, & I_{qjk} > \max\{I_{qj}, I_{qk}\}, \\ \log_2 \dfrac{P(l)}{P(\bar{l})} \prod_{q=1}^{Q_{xy}} \dfrac{P(s_q|\eta_q, l)}{P(s_q|\eta_q, \bar{l})}, & \text{otherwise.} \end{cases} \tag{14}$$

4.2.4 *Multiplexing*

Based on the above description, the correlations in every space can be obtained. By multiplexing these correlations, the existence of links among the users can be judged. In other words, we can predict the missing links E^* in network G by multiplexing correlations in multidimensional network spaces: G^f, G^i, G^s. We introduce the parameter set $\lambda = \{\lambda_1,\lambda_2,\lambda_3\}$, and we use logistic regression for multiplexing correlations. The parameter set λ is updated using the gradient descent algorithm, and the update process is as follows:

$$\lambda_{j-\text{new}} = \lambda_{j-\text{old}} + \tau \left(y - \sum_{j=1}^{3} \lambda_j r_j \right) \times r_j, \tag{15}$$

where τ is the learning rate, and y denotes the existence of links among users. The algorithm convergences when the value of each parameter is less than a threshold value. The parameters are updated until convergence, and they are output as λ^*. Finally, the logistic regression definition $P(e^*|r)$ can be used for link prediction, and only if the value of the probability is greater than the specified threshold value, the links can be deemed to exist. Assuming e^* denotes the existence of each missing link in E^*, the definition is as follows:

$$P(e^*|r) = \frac{\exp(\sum_{j=1}^{3} {\lambda_j}^* r_j)}{1 + \exp(\sum_{j=1}^{3} {\lambda_j}^* r_j)}, \tag{16}$$

$$e^* = \begin{cases} 1, & P(e^*|r) \geqslant \xi, \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

4.3 Learning algorithm

The input of the algorithm includes the whole network $G = (U, E)$ and behavior of the all users $A = \{(a, u_i)|u_i \in U\}$, as explained in Section 3. Then, the number of interest topics K, Dirichlet priors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\alpha}'$, $\boldsymbol{\beta}'$ and learning parameter τ are input to get the optimal parameter set λ^* and link prediction result E^*. Meanwhile, the learning algorithm can be divided into training and testing. In other words, we use our method train to get the optimal parameter set λ^* and test to get link prediction result E^*. The training algorithm is as Algorithm 1.

In the training algorithm, parameter set $\lambda = \{\lambda_1,\lambda_2,\lambda_3\}$ is initialized and three correlation vectors $\boldsymbol{F}$, $\boldsymbol{I}$, $\boldsymbol{S}$ are established. Then, LDA improved by Gaussian weighting is used to obtain user topic distribution. By extending it with the hidden naive Bayesian algorithm, correlations in three spaces can be computed. Finally, by multiplexing correlations with logistic regression, we can get the converged parameter set λ^*. After training, we can use the trained parameters for link prediction. In other words, we can get the prediction result E^*=argmax$_{\lambda^*} P_{\lambda^*}(E|G^f, G^i, G^s, A)$.

In addition, the complexity of the algorithm is considered. Assuming that n denotes the number of users, the extraction of user topic distribution is $T_{\text{extract}} = O(nN_x) + O(nN'_x) \sim O(n)$; the computation of correlations in every space is $T_{\text{compute}} = O(n^2)$; and the multiplexing of the correlations is $T_{\text{multiplex}} = O(n)$. Based on the above analysis, the total time complexity of the training algorithm is $T_{\text{train}} = T_{\text{extract}} + T_{\text{compute}} + T_{\text{multiplex}} \sim O(n^2)$ and the complexity of using the trained parameters for link prediction is $T_{\text{test}} = T_{\text{predict}} = O(n)$. Meanwhile, we explain the network size that is applicable to this method. This method can be used to analyze the small networks and medium networks directly. For large networks, because of the time complexity of training is high and the time complexity of testing is low, we can make the training process offline and take distributed computing into account to alleviate the problem of time complexity.

Algorithm 1 Training algorithm

Input: Whole network $G = (U, E)$; activities $A = \{(a, u_i)|u_i \in U\}$; parameter $\boldsymbol{\alpha},\ \boldsymbol{\beta},\ \boldsymbol{\alpha}',\ \boldsymbol{\beta}',\ K,\ \tau$;
Output: Parameter set λ^*=argmax$_\lambda P_\lambda(E|G^f, G^i, G^s, A)$.
// Initialization
Initialize $\lambda = \{\lambda_1,\lambda_2,\lambda_3\}$;
Map network G into G^f, G^i, G^s;
Construct three correlation sets $\boldsymbol{F},\ \boldsymbol{I},\ \boldsymbol{S}$ from Eqs. (1)–(3);
// Sample topic (take the followed user as an example).
Repeat
  // Similar approach to followed user.
  For $f_i \leftarrow 1$ **to** N_x **do**
    Sample topic z_i from Eq. (5);
  Endfor
Until Convergence;
// Compute correlations.
For Each user pair (u_x, u_y) of U **do**
  For Common neighbor $s_q \leftarrow 1$ **to** Q_{xy} **do**
    Compute $P(s_q|\eta_q, l),\ P(s_q|\pi_q, l),\ P(s_q|\eta_q, \bar{l}),\ P(s_q|\pi_q, \bar{l})$ from Eq. (12);
  Endfor
  Compute correlations $r_{1-xy},\ r_{2-xy},\ r_{3-xy}$ from Eqs. (7), (8), (14);
Endfor
// Multiplex correlations.
Repeat
  Update parameters λ in G from Eq. (15);
Until Convergence;
Get optimal parameters λ^*=argmax$_\lambda P_\lambda(E|G^f, G^i, G^s, A)$.

5 Experiments and analysis

5.1 Experimental settings

In this subsection, the experimental settings are described in detail. First, the experimental data is introduced. Then, the baseline methods used in the experiments are presented. Finally, evaluation metrics are proposed to evaluate the performance of our method.

5.1.1 *Experimental data*

The datasets used in this paper are collected from Sina micro-blog and Twitter. Sina micro-blog is one of the most popular social networking platforms in China. In the process of data collection, we randomly selected a user as the starting point. Some users and their micro-blogs were captured based on breadth-first-search. Then, the data of 49556 users and 61880 user relationships for the 2011/08/21–2012/02/22 were collected. Twitter dataset [39] is the public dataset from SNAP[1)]. It has been built after monitoring the spreading processes on Twitter before, during and after the announcement of the discovery of a new particle with the features of the elusive Higgs boson. The messages of it are considered during 2011/07/01–2012/07/07, and it had been updated on Mar 31, 2015. We selected 97632 users and 6883144 relationships from it to do the experiment. The statistics of the datasets are shown in Table 2.

Figure 5 shows the degree distributions in the following space, interaction space and structure space. We can see that there are a large number of edge nodes and a small number of central nodes in the network. Therefore, every network space has power-law characteristics.

1) http://snap.stanford.edu/index.html.

Table 2 Statistics of datasets

Dataset	Sina micro-blog	Twitter
Users	49556	97632
Relationships	61880	6883144
Activity	3057635	167420
Forward	506765237	91142
Review	185079821	62986

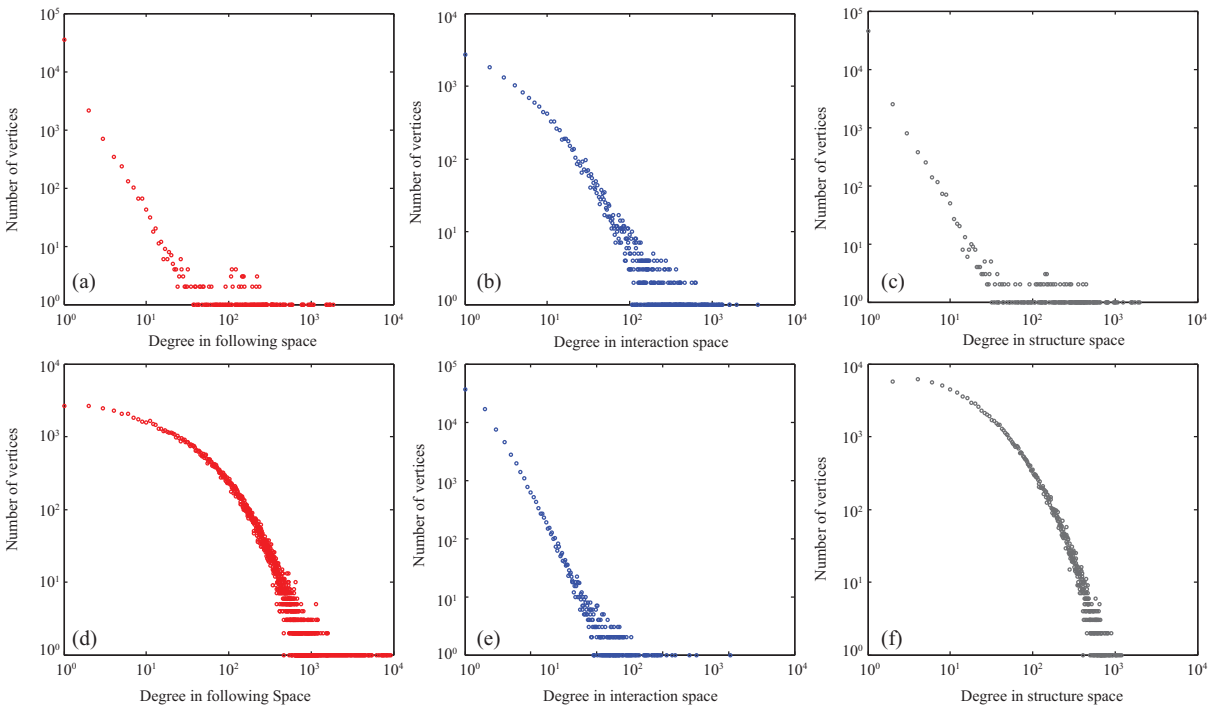

Figure 5 (Color online) Degree distribution in three network spaces. (a) Following space, (b) interaction space, (c) structure space in Sina; (d) following space, (e) interaction space, (f) structure space in Twitter.

5.1.2 *Baseline methods*

In the experiments, our method is compared with a few baseline methods.

LDA [32]. This is a topic model, which is applied in the field of natural language processing. It introduces the topic space on the basis of the traditional vector space and employs dimension reduction to reduce complexity. By computing the similarity of the user topic space, the probability of link establishment can be obtained.

VSM [40]. The vector space model is used for computing similarity in the field of natural language processing. It transforms the similarity between two documents into two vectors and uses the term frequency-inverse document frequency (TF-IDF) algorithm to measure the weight of every feature. By computing the similarity of the user feature vectors, the probability of link establishment can be obtained.

CF [41]. Collaborative filtering algorithm comes from the recommendation system, which infers possible preference or interests based on user historical behavior. Owing to its easy implementation and interpretability, it can be introduced to link prediction.

RW [21]. Random walks is a commonly link prediction algorithm and it suggests that a random walker is more likely to visit the nodes to which new links will be created in the future.

HNB [35]. Hidden naive Bayes is a Naive Bayes algorithm with an addition of implicit factor for each attribute to represent the dependence between other attributes. The link is considered as a class node and the common neighbors are regarded as the attribute nodes, it can be introduced to link prediction.

WAODE [36]. Weighted average of one-dependence estimators is an improved naive Bayes algorithm.

By assigning different weights to these one-dependence classifiers, attribute independence is weakened. Similarly, the common neighbors are treated as attribute nodes, the algorithm can be used to link prediction.

CN & RA. Common neighbor algorithm and resource allocation algorithm are based on network structure. They focus mainly on the number of common neighbors, and the probability of link establishment is large if a user has more common neighbors. These types of algorithms can be used alone or in combination with other methods.

LR & SVM. The use of classifiers is also an effective method of link prediction, and such methods depend on attribute extraction. They use associated attributes to train a classification, and employ the trained classification for link prediction.

5.1.3 *Evaluation metrics*

In this paper, accuracy, precision, recall, F-Measure, and ROC were used to verify the prediction results. We assumed that a user pair with a link between them is a positive example "1", and another user pair is a negative example "0". Meanwhile, the dataset needs to be partitioned into training set and testing set. The user pair set is randomly divided into training set and testing set. The ratio of the training set and the test set is set to be 7:3. We assume that the links in testing set are not exist, that is to say, the links in testing set are missing links. Using the proposed link prediction method, the training set can be used to fit parameters and the missing links in testing set can be predicted. According to the definitions of these five metrics, the experimental results are measured. The better prediction results have great accuracy, precision, recall and F-Measure, and their ROC curves are close to the upper left corner.

5.2 Prediction performance analysis

In this subsection, the performance of our method is evaluated from three viewpoints. First, we verify the impact of interest topic number on prediction result by changing the multiplexing process, and SVM is used to change in it. Next, in comparing our method with its sub-methods, the relationship between the proportion of training sets and the prediction performance can be shown. Finally, we evaluate the performance of our method by comparing it with other baseline methods. According to the above three viewpoints, the superiority of our method can be verified.

Firstly, by changing the multiplexing process, the impact of interest topic number on the prediction result can be verified. Considering that SVM is a better binary classifier, we choose it to change the multiplexing process. The effects of interest topic number on link prediction are shown in Figure 6. The x-axis represents the interest topic number and the y-axis represents the values of the four metrics: accuracy, precision, recall, and F1-measure.

As shown in Figure 6(a)–(d), when K is 10–15, the values of the metrics are greater than those when K takes other values in Sina. As shown in Figure 6(e)–(h), the results are the same in Twitter. In other words, the best range of K is 10–15. Compared with SVM, logistic regression has better performance.

Secondly, by reducing the attribute characteristics in our method, three sub-methods are obtained: Sub-FI, Sub-IS, and Sub-FS. The sub-methods are built by considering partial spaces. Comparing our method with the sub-methods, the relationship between the proportion of training sets and the prediction performance can be shown. Taking the Sina dataset as an example, we choose appropriate values of K=10 and K=15, and a comparison of the prediction effects is shown in Figure 7. The x-axis represents the proportion of training sets and the y-axis represents the values of the four metrics.

As shown in Figure 7, the recall performance of the proposed method is not the greatest, but it is optimal in other metrics. Owing to the inversion of recall and precision, recall will be sacrificed in the case of higher precision, so the recall of the proposed method is lower than that of the sub methods. As the proportion of training sets increases, the prediction effect of the method improves.

Meanwhile, by comparing with LDA derivative methods or its similar methods, the effective of the improved LDA can be verified. We choose LDA and it derivative methods or similar methods: VSM,

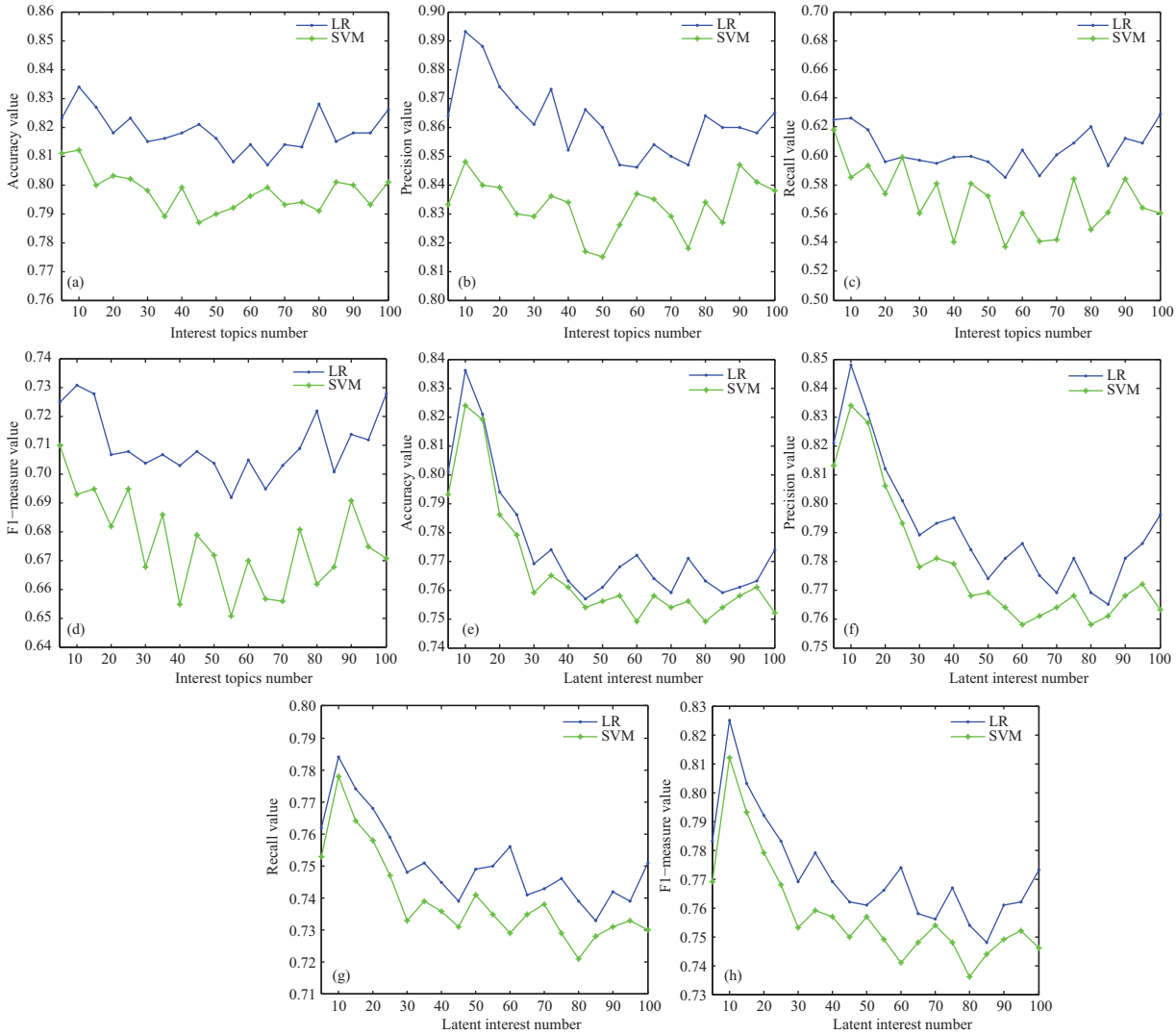

Figure 6 (Color online) Effect of interest topic number. (a) Accuracy, (b) precision, (c) recall, (d) F1-measure in Sina; (e) accuracy, (f) precision, (g) recall, (h) F1-measure in Twitter.

CN+LDA, RA+LDA, which are applied to link prediction by behavioral modeling. And the comparison of the prediction effects about these methods are shown in Figure 8.

From Figure 8, we can see that the ROC curve of our method is closest to the upper left corner which means the effect of prediction is the best. In other words, the improvements to the LDA model are effectively.

Finally, the performance of our method is evaluated by comparing with some baseline methods. We choose some classical link prediction methods to compare: VSM, LDA, CN+LDA, RA+LDA, CF, RW, HNB and WAODE. The performances of them are shown in Tables 3 and 4.

The experimental results show that the method extended with hidden naive Bayesian algorithm can effectively improve the prediction precision. And the proposed method plays optimal performance compared with baseline methods. Therefore, our method can predict the links among the users effectively.

6 Conclusion

In this study, a method of correlation multiplexing for link prediction is proposed, and it can effectively predict links among users by analyzing user behavior and user relationships. Firstly, we mapped users into three network spaces: following space, interaction space and structure space. With a hierarchical process,

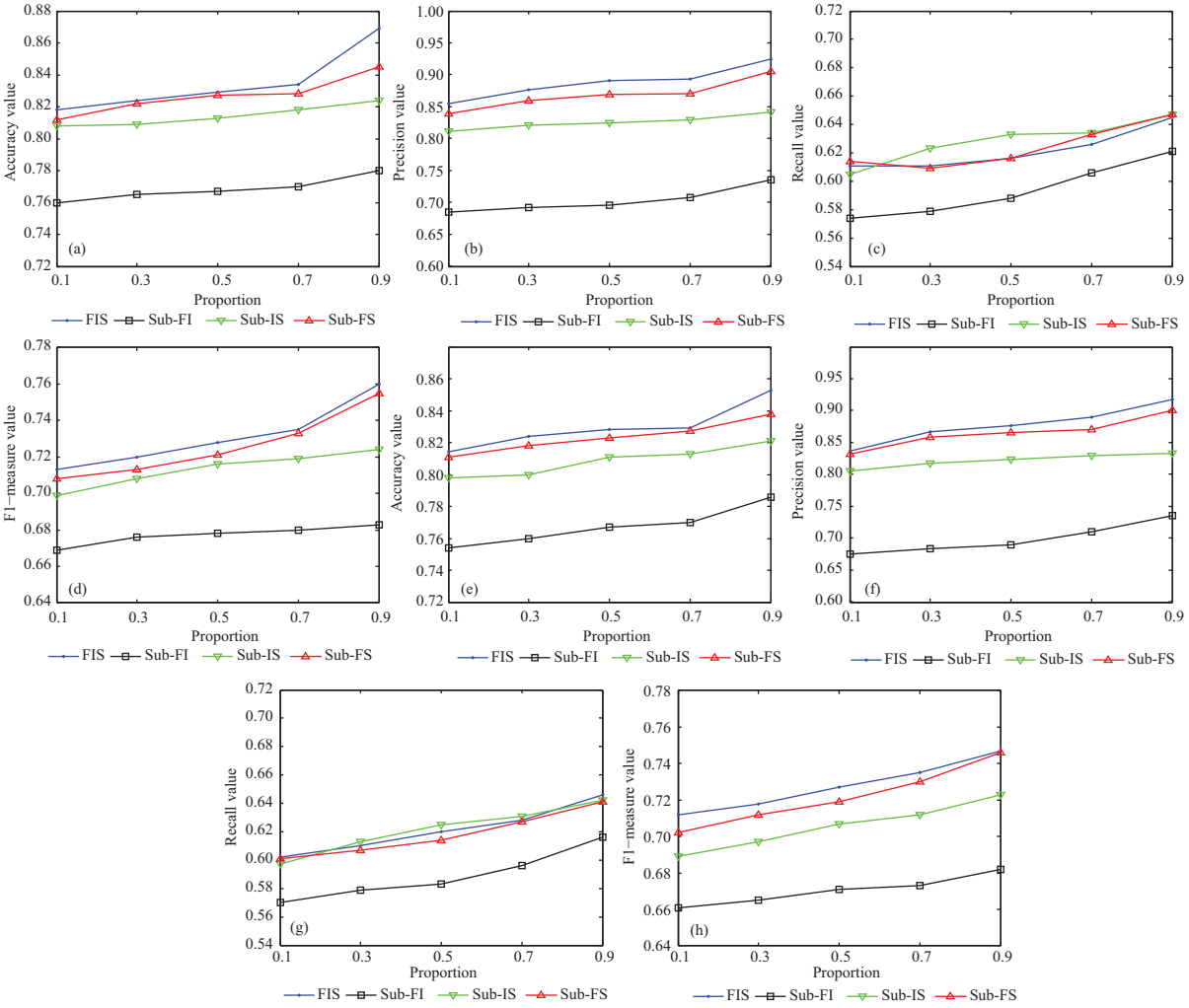

Figure 7 (Color online) Comparison of prediction effects between proposed method and sub methods. (a) Accuracy (K=10); (b) precision (K=10); (c) recall (K=10); (d) F1-measure (K=10); (e) accuracy (K=15); (f) precision (K=15); (g) recall (K=15); (h) F1-measure (K=15).

Table 3 Comparison of different methods in Sina

Algorithm	Accuracy	Precision	Recall	F1-measure
LDA	0.716	0.719	0.626	0.669
CN+LDA	0.820	0.857	0.634	0.729
RA+LDA	0.826	0.861	0.638	0.733
VSM	0.834	0.839	0.605	0.703
CF	0.831	0.836	0.592	0.693
RW	0.843	0.877	0.612	0.721
HNB	0.847	0.881	0.608	0.719
WAODE	0.851	0.893	0.629	0.738
Proposed method (K=10)	**0.874**	**0.913**	**0.649**	**0.759**
Proposed method (K=15)	0.858	0.895	0.636	0.744

we analyzed the correlations in each space separately. Secondly, the traditional LDA text modeling method was improved by Gaussian weighting and applied to user behavior modeling. Finally, the method was extended with the hidden naive Bayesian algorithm and we multiplexed user correlations in three spaces for link prediction.

We used data from a social network (Sina micro-blog) in the experiments. The experimental results

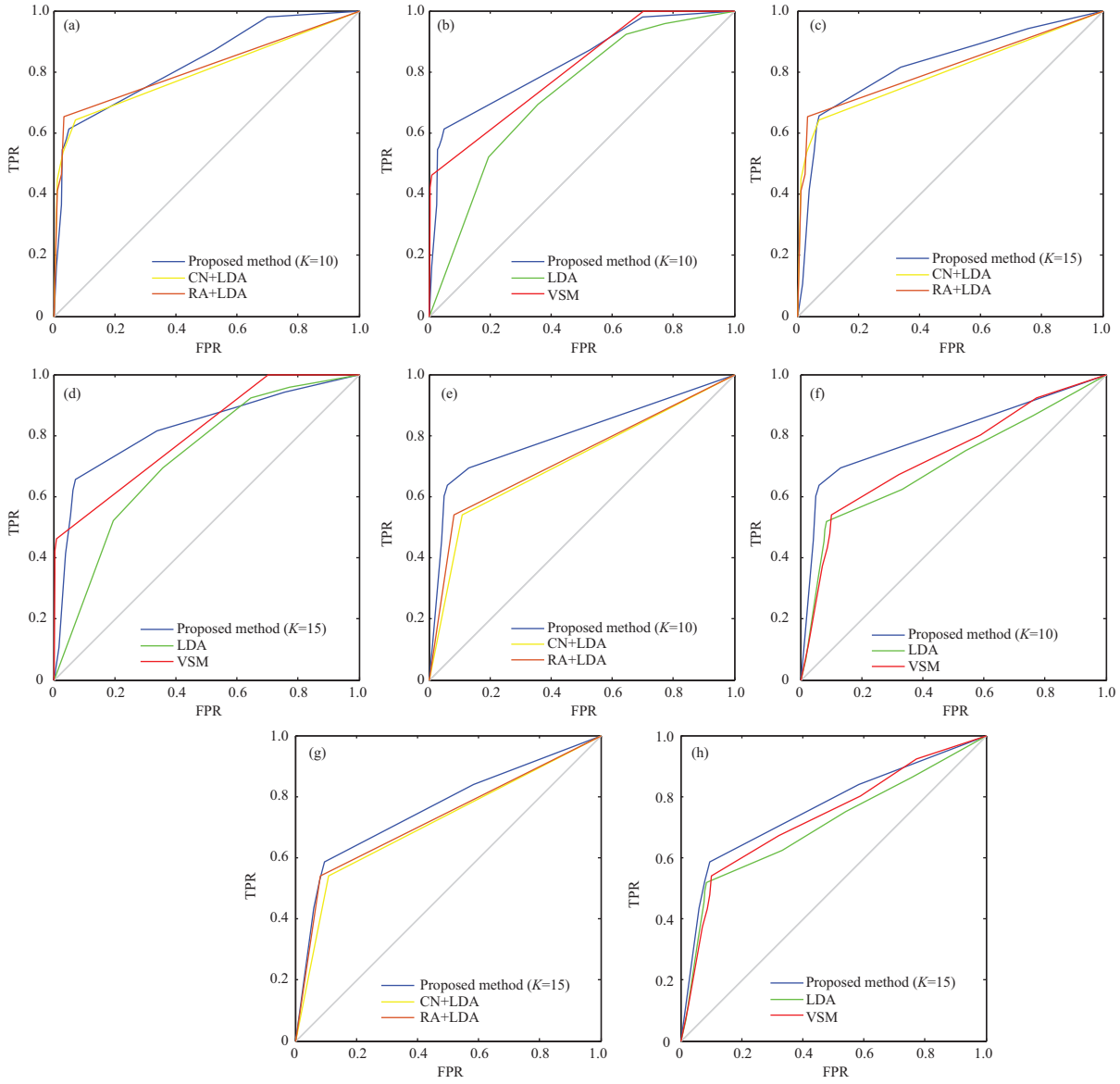

Figure 8 (Color online) Comparison of different methods in ROC. (a) ROC1, (b) ROC2, (c) ROC3, (d) ROC4 in Sina; (e) ROC1, (b) ROC2, (c) ROC3, (d) ROC4 in Twitter.

Table 4 Comparison of different methods in Twitter

Algorithm	Accuracy	Precision	Recall	F1-measure
LDA	0.709	0.725	0.719	0.722
CN+LDA	0.775	0.801	0.743	0.771
RA+LDA	0.783	0.809	0.751	0.779
VSM	0.768	0.792	0.735	0.762
CF	0.761	0.784	0.726	0.754
RW	0.796	0.813	0.754	0.782
HNB	0.809	0.828	0.763	0.794
WAODE	0.818	0.834	**0.793**	0.813
Proposed method (K=10)	**0.836**	**0.848**	0.784	**0.825**
Proposed method (K=15)	0.821	0.831	0.774	0.803

showed that the proposed method can improve link prediction performance in comparison to other baseline prediction methods. By studying link prediction in social networks, we can predict links among users

effectively, and the method can support studies on the evolution and discovery of network structures. In future work, we mainly focus on the applicability of methods in large-scale networks. The improvement of link prediction method includes how to train data in distributed environment and how to alleviate the time complexity of training algorithms.

Acknowledgements This work was supported by National Basic Research Program of China (Grant No. 2013CB329606), National Natural Science Foundation of China (Grant No. 61772098), Chongqing Science and Technology Commission Project (Grant No. cstc2017jcyjAX0099), and Foundation of Ministry of Education of China and China Mobile (Grant No. MCM20130351).

References

1 Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. J Assoc Inf Sci Technol, 2007, 58: 1019–1031
2 Getoor L, Diehl C P. Link mining: a survey. ACM SIGKDD Explor Newslett, 2005, 7: 3–12
3 Lü L, Zhou T. Link prediction in complex networks: a survey. Phys A Stat Mech Appl, 2011, 390: 1150–1170
4 Gong N Z, Talwalkar A, Mackey L, et al. Joint link prediction and attribute inference using a social-attribute network. ACM Trans Intell Syst Technol, 2014, 5: 1–20
5 Ding J Y, Jiao L C, Wu J S, et al. Prediction of missing links based on community relevance and ruler inference. Knowl-Based Syst, 2016, 98: 200–215
6 Wang H H, Raza A A, Lin Y B, et al. Behavior analysis of low-literate users of a viral speech-based telephone service. In: Proceedings of the 4th Annual Symposium on Computing for Development, Cape Town, 2013
7 Agarwal V, Bharadwaj K K. A collaborative filtering framework for friends recommendation in social networks based on interaction intensity and adaptive user similarity. Soc Netw Anal Min, 2013, 3: 359–379
8 Martínez V, Cano C, Blanco A. ProphNet: a generic prioritization method through propagation of information. BMC Bioinformatics, 2014, 15: 1506–1526
9 Wang J, Sun J Q, Lin H F, et al. Convolutional neural networks for expert recommendation in community question answering. Sci China Inf Sci, 2017, 60: 110102
10 Ermis B, Acar E, Cemgil A T. Link prediction in heterogeneous data via generalized coupled tensor factorization. Data Min Knowl Discov, 2015, 29: 203–236
11 Sherkat E, Rahgozar M, Asadpour M. Structural link prediction based on ant colony approach in social networks. Phys A Stat Mech Appl, 2015, 419: 80–94
12 Hannon J, Bennett M, Smyth B. Recommending twitter users to follow using content and collaborative filtering approaches. In: Proceedings of the 4th ACM Conference on Recommender Systems, Barcelona, 2010. 199–206
13 Jiang B, Liang J G, Sha Y, et al. Domain dictionary-based topic modeling for social text. In: Proceedings of International Conference on Web Information Systems Engineering, Shanghai, 2016. 109–123
14 Wagner C, Singer P, Strohmaier M, et al. Semantic stability and implicit consensus in social tagging streams. IEEE Trans Comput Soc Syst, 2014, 1: 108–120
15 Nguyen L, Vo B, Hong T P, et al. Interestingness measures for classification based on association rules. In: Proceedings of International Conference on Computational Collective Intelligence, Ho Chi Minh City, 2012. 383–392
16 Wang P, Xu B W, Wu Y R, et al. Link prediction in social networks: the state-of-the-art. Sci China Inf Sci, 2015, 58: 011101
17 Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. J Mach Learn Res, 2003, 3: 993–1022
18 Cho Y S, Ver Steeg G, Ferrara E, et al. Latent space model for multi-modal social data. In: Proceedings of the 25th International Conference on World Wide Web, Montréal, 2016. 447–458
19 Bliss C A, Frank M R, Danforth C M, et al. An evolutionary algorithm approach to link prediction in dynamic social networks. J Comput Sci, 2014, 5: 750–764
20 Sett N, Singh S R, Nandi S. Influence of edge weight on node proximity based link prediction methods: an empirical analysis. Neurocomputing, 2016, 172: 71–83
21 Liu Y C, Tong H H, Xie L, et al. Supervised link prediction using random walks. In: Proceedings of Chinese National Conference on Social Media Processing, Guangzhou, 2015. 107–118
22 Gong J B, Gao X X, Cheng H, et al. Integrating a weighted-average method into the random walk framework to generate individual friend recommendations. Sci China Inf Sci, 2017, 60: 110104
23 Wang T, Krim H, Viniotis Y. A generalized Markov graph model: application to social network analysis. IEEE J Sel Top Signal Process, 2013, 7: 318–332
24 Mossel E, Sly A, Tamuz O. Asymptotic learning on bayesian social networks. Probab Theory Relat Fields, 2014, 158: 127–157
25 Li F H, He J, Huang G Y, et al. Node-coupling clustering approaches for link prediction. Knowl-Based Syst, 2015, 89: 669–680
26 Martínez V, Berzal F, Cubero J C. Adaptive degree penalization for link prediction. J Comput Sci, 2016, 13: 1–9
27 Pizzato L, Rej T, Akehurst J, et al. Recommending people to people: the nature of reciprocal recommenders with a case study in online dating. User Model User-Adapt Interact, 2013, 23: 447–488

28 Scellato S, Noulas A, Mascolo C. 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, San Diego, 2011. 1046–1054
29 Shahmohammadi A, Khadangi E, Bagheri A. Presenting new collaborative link prediction methods for activity recommendation in Facebook. Neurocomputing, 2016, 210: 217–226
30 Liu J, Xu B M, Xu X, et al. A link prediction algorithm based on label propagation. J Comput Sci, 2016, 16: 43–50
31 Althoff T, Jindal P, Leskovec J. Online actions with offline impact: how online social networks influence online and offline user behavior. In: Proceedings of the 10th ACM International Conference on Web Search and Data Mining, Portland, 2016. 537–546
32 Li L, He J P, Wang M, et al. Trust agent-based behavior induction in social networks. IEEE Intell Syst, 2016, 31: 24–30
33 Cha Y, Cho J. Social-network analysis using topic models. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, Portland, 2012. 565–574
34 Chang J, Blei D M. Relational topic models for document networks. AIStats, 2009, 9: 81–88
35 Jiang L X, Zhang H, Cai Z H. A novel Bayes model: hidden naive Bayes. IEEE Trans Knowl Data Eng, 2009, 21: 1361–1371
36 Jiang L X, Zhang H, Cai Z H, et al. Weighted average of one-dependence estimators. J Exp Theor Artif Intell, 2012, 24: 219–230
37 Jiang L X, Wang S S, Li C Q, et al. Structure extended multinomial naive Bayes. Inf Sci, 2016, 329: 346–356
38 Zhao F, Zhu Y J, Jin H, et al. A personalized hashtag recommendation approach using LDA-based topic model in microblog environment. Future Gener Comput Syst, 2016, 65: 196–206
39 De Domenico M, Lima A, Mougel P, et al. The anatomy of a scientific rumor. Sci Rep, 2013, 3: 2980
40 Waitelonis J, Exeler C, Sack H. Linked data enabled generalized vector space model to improve document retrieval. In: Proceedings of NLP& DBpedia 2015 Workshop at the 14th International Semantic Web Conference CEUR-WS, Bethlehem, 2015
41 Jiang B, Liang J G, Sha Y, et al. Retweeting behavior prediction based on one-class collaborative filtering in social networks. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, 2016. 977–980