An Approach to Cold-Start Link Prediction: Establishing Connections between Non-Topological and Topological Information

Zhiqiang Wang, Jiye Liang, Ru Li, and Yuhua Qian, *Member, IEEE*

Abstract—Cold-start link prediction is a term for information starved link prediction where little or no topological information is present to guide the determination of whether links to a node will form. Due to the lack of topological information, traditional topology-based link prediction methods cannot be applied to solve the cold-start link prediction problem. Therefore, an effective approach is presented through establishing connections between non-topological and topological information. In the approach, topological information is first extracted by a latent-feature representation model, then a logistic model is proposed to establish the connections between topological and non-topological information, and finally the linking possibility between cold-start users and existing users is calculated. Experiments with three types of real-world social networks Weibo, Facebook, and Twitter show that the proposed approach is more effective in solving the cold-start link prediction problem and establishing connections between topological and non-topological information.

Index Terms—Social network, link prediction, predictive model, latent feature

1 Introduction

Link prediction is a fundamental problem in network researches, and its solution is of great significance to network completion [1], [2], [3] and network evolution [4], [5]. The purpose of link prediction is to predict the missing links in current networks [6], [7] and new links that will appear in future networks [8]. Most relevant studies have not taken link prediction for isolated nodes in social network into consideration. Some studies explored the topological information of social networks [6], and some other studies combined topological information with some auxiliary information [9], [10], [11].

Different from general link prediction tasks, this paper focuses on an information-starved link prediction and attempts to predict the possible links between cold-start users (the isolated nodes in a social network) and existing users (the other nodes). Formally, $G(V_1, E, V_2, A)$ is regarded as the social network, where V_1 is a set of $n_1 = |V_1|$ existing users with network structure, $E \subseteq V_1 \times V_1$ is a set of edges, V_2 is a set of $n_2 = |V_2|$ cold-start users who are isolated nodes in the network, and we use $n = |n_1| + |n_2|$ to denote the total number of social network users. A (see Fig. 1) is an $n \times m$ user-attribute matrix extracted from users' auxiliary information; its rows represent users and its columns represent attributes respectively; m denotes the size of attribute dimension. Besides, $n \times n$ matrix (see Fig. 2) denotes the linked data, and the nodes from u_1 to u_{n_1} in this matrix are known as existing users in social network. In the adjacency matrix, the value in the corresponding position of the matrix will be 1 if there is a link between existing users, and the value will be 0 if there is not a link. The users from u_{n_1+1} to $u_{n_1+n_2}$ are what we call cold-start users in this network, and the linking information of these users is unobserved or missing. Discovering the unobserved links between existing users V_1 and cold-start users V_2 is the main purpose of cold-start link prediction. Most studies on link prediction are based on two backgrounds [12]: "non-temporal" link prediction [1], [3], [6], [13], [14], [15], which predicts the unobserved status of links for pairs of nodes, and "temporal" link prediction [8], [16], which predicts new links in the future. Both of them are very important in various fields such as network completion and network evolution. This paper focuses on the "non-temporal" cold-start link prediction problem, and we use the term "cold-start link prediction" to refer to a "non-temporal" version of the problem.

It is known that real social networks usually have rich structure characters such as homophily, heterophily and core-periphery. Additionally, various auxiliary information such as users' profiles, tags, publications, etc. are also an important part of the networks. Naturally, we would think about how the network structure interacts with the auxiliary information. If we can find the connections between network structure and auxiliary information, the link information of the cold-start nodes could be recovered to some extent. To be specific, this paper addresses the following two problems:

i. How to extract and represent the topological information of a network.
ii. How to establish the connection between the topological information and non-topological information to solve the cold-start link prediction problem.

• *The authors are with the Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China. E-mail: zhiq.wang@163.com, {ljy, liru}@sxu.edu.cn, jinchengqyh@126.com.*

Manuscript received 8 Jan. 2016; revised 24 July 2016; accepted 25 July 2016. Date of publication 3 Aug. 2016; date of current version 3 Oct. 2016. Recommended for acceptance by N. Chawla. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TKDE.2016.2597823

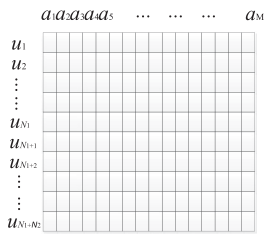


Fig. 1. User-attribute matrix.

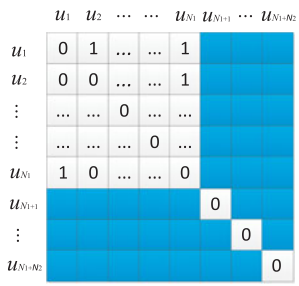


Fig. 2. Linked data.

The above problems are the main motivations of this study. In this paper, we propose a representation of latent features in a unified space for cold-start users and existing users. First, we extract the latent features for existing users based on a latent-feature representation (LFR) model; then a logistic model is proposed to establish the connections between topological and non-topological information. Based on the connections, the existing users and cold-start users are simultaneously represented in a unified latent-feature space, and the linking possibility between cold-start users and existing users could be calculated through their latent-feature representations.

The rest of the paper is organized as follows. The related work is introduced in Section 2; in Section 3, we present the cold-start link prediction method; we evaluate the proposed method with different social network datasets in Section 4; finally we make a conclusion in Section 5.

2 Related Works

In this section, we will review the related works from the perspective of link prediction. Since there are great similarities between cold-start link prediction and cold-start recommendation, relevant literatures on cold-start recommendation will be covered in this section.

2.1 Metric-Based Link Prediction

Metric-based methods, which address the problem by calculating the similarity between nodes, are very common because people usually create new relationship with people who have certain similarity with themselves in topological or non-topological features.

2.1.1 Topology-Based Metrics

Many existing topology-based metrics are defined by using various topological information of network. The metrics like Common Neighbors [17], Jaccard Coefficient [18], Sørensen Index, Salton Cosine Similarity (COS), Hub Promoted [19], Hub Depressed [2], Leicht-Holme-Nerman [20], Parameter-Dependent [21], Adamic-Adar Coefficient [22], Preferential Attachment [23], and Resource Allocation [22] etc. are defined by neighbors because neighbors can indirectly reflect users' social behavior and directly affect users' social choice [7]. There are also many other different metrics such as Katz [24], Local Path [25], Relation Strength Similarity [26], FriendLink [27], Hitting Time [28], Commute Time, SimRank [29], Rooted PageRank [30] and PropFlow [31] etc. Katz [24] counts all paths between two nodes; Local Path [25] makes use of information of local paths with lengths 2 and 3; Relation Strength Similarity [26] is a vertex similarity that could capture potential relationships of real world network structure; FriendLink [27], a new node similarity measure takes into account all ℓ-length paths between nodes; Hitting Time [28] is an asymmetric metric which is the expected number of steps required for a random walk between nodes, and Commute Time is a symmetric version of Hitting Time; SimRank [29] is defined through the assumption that two nodes are similar if they are connected to similar nodes; Rooted PageRank [30] is a modification of PageRank; PropFlow [31] is similar to Rooted PageRank, but it is more localized. For all of the topology-based metrics, the key is the topological information between nodes in a network. Due to the loss of topological information of cold-start users, the existing topology-based metrics cannot be used to deal with the cold-start link prediction.

2.1.2 Non-Topology Metrics

Non-topology metrics [32], [33], [34] focus on the information outside the network structure. For instance, in an online social network, each user has his/her profile which covers the description of age, interests and geographic location. In addition, large amount of shared content is important external information, and it is beneficial for social network data mining [35], [36], [37]. Non-topology link prediction methods are commonly dependent on similarity (two nodes are considered similar if they have many common attributes). Thus, the crux of the matter lies in extracting users' attributes and designing their attribute similarity. Wang et al. [38] proposed a lifestyle-based friend recommendation method, and in this method his/her lifestyles are first extracted based on Latent Dirichlet Allocation model [39], and then a similarity metric is designed based on the lifestyle of users. Aiello et al. [40] found that users' label could reflect their interest, so they proposed label-similarity-based method to predict links in Flickr, Last.fm, and aNobill. Different from a general metric-based method, the two-phase bootstrap probabilistic method proposed by Leroy et al. [41] utilized users' group features to measure the similarity between nodes in the first phase, and employed graph-based measure to produce the final prediction in the second phase. Besides, there are also many other non-topology metrics which utilize users' interests [42] or keywords [43] to measure the similarity between a pair of users. To sum up, non-topology methods have mainly used the users' attributes extracted from non-topological information such as profile, label and content because the information can reflect their personal interests and social behaviors. These methods of attribute extraction and similarity can be directly used for the cold-start link prediction. However, the effectiveness of non-topology metrics depend on the domain and the specific network and information available.

2.2 Learning-Based Link Prediction

Here we deal with the learning-based link prediction methods according to the following three subdivisions.

2.2.1 Classification-Based Methods

In classification-based methods, link prediction is treated as a binary classification problem in which each pair of nodes is an instance, and being positive or being negative is a class label which indicates whether the pair of nodes are connected or not. Many classification models are used in link prediction problems, such as support vector machines (SVM), logistic regression (LR), K-Nearest Neighbors, and Naive Bayes. Like many classification problems, feature selection for classification-based link prediction is the most critical part. These features include topological and non-topological ones. Many link prediction classification models are based on topological features. Lichtenwalter et al. [44] introduced the vertex collocation profile (VCP), one kind of local topological features. Sá et al. [45] used many proximity metrics such as Common Neighbors [17], Jaccard Coefficient [18], Adamic-Adar Coefficient [22], Preferential Attachment [23] and Local Path [25] and deployed them as predictor features in the supervised link prediction model. Leskovec et al. [46] made effort to predict positive and negative links based on the topological features, including degrees of the nodes and triad. Chiang et al. [47] extended Leskovec et al.'s work [46], and made use of features derived from longer cycles in the network. Lichtenwalter et al. [31] presented a high-performance frame work for link prediction, which made great improvement over existing unsupervised methods. These methods are the typically learning-based link prediction on various topological features. To sum up, the learning-based methods construct features by computing the topological similarity based on the neighbors or paths between two users; extensive experiments validate that these topological features are very important and effective in link prediction. However, the topological features are not applicable to a classification model because the loss of topological structure information of cold-start users.

Apart from topological features used in classification model for link prediction, the addition of non-topological features (such as users' location, interests, education) can improve link prediction. Scellato et al. [9] took location features, social features, and global features into consideration on a supervised learning framework. Wohlfarth et al. [11] proposed a semantic and event-based approach to improve the accuracy of the link predictor. These studies indicate that if non-topological features are added to a classification model, the predicting results could be improved. Due to the loss of the topological features in cold-start users, only non-topological features could be used in classification models to solve the cold-start link prediction problem. However, its effectiveness still depends on the domain and the specific network and non-topological features available, etc.

2.2.2 Matrix Factorization-Based Methods

Matrix factorization is a type of technique to get low rank approximation and global information of the adjacency matrices of networks. As is known, many matrix factorization methods (such as singular value decomposition (SVD) [48], non-negative matrix factorization (NMF) [49] and probabilistic matrix factorization (PMF) [50]) have been used in the field of collaborative filtering. Some other matrix factorization methods have been proposed to solve the problem of link prediction. Dong et al. [51] proposed the method of predicting the missing links by using convex nonnegative matrix factorization, and his method was more effective because he combined the concepts of block structure with low rank approximations for matrices. Menon et al. [12] proposed a model which was trained with a ranking loss to address the class imbalance problem that is common in network datasets. However, like topology-based metrics mentioned in Section 2.1, due to the loss of topological structure information of cold-start users, these matrix factorization-based methods are not applicable to cold-start link prediction, either.

2.2.3 Probabilistic Graph Model-Based Methods

Probabilistic Graph Model (PGM) is an important and effective way to model networks. PGM-based methods can capture many complex network structure properties and provide social network analysis with deep insights. Clauset et al. [1] proposed a hierarchical network model, in which a network could be modelled by a hierarchical random graph where leaves correspond to the nodes of network and internal nodes correspond to the linking probability between leaf nodes. By computing the expectation probability in this hierarchical random graph, the linking possibility between network nodes could be gained. A typical PGM-based network model is a stochastic block model [52], [53], which assumes that each node belongs to a group, and the linking probability between two nodes depends on to which groups they belong. Another typical PGM-based network model is a latent-feature network model [14], [15], [54], [55]. It is a probabilistic generative model, in which nodes' latent features are, at first, generated based on some distributions; the edges between nodes are then generated based on the nodes' latent attributes. The latent-feature network model can provide network nodes with a vectorial representation of latent attributes, so it could also be applicable to social network data mining. However, like many methods mentioned before, due to the loss of topological structure information of cold-start users, these PGM-based methods are not applicable to the cold-start link prediction task, either.

2.3 Cold-Start Recommendation

Link prediction is closely related to the problem of collaborative filtering. From the perspective of graph mining, link prediction is to mine the interactions between nodes in unipartite networks, and collaborative filtering is to mine the interactions between two types of nodes (user, item) in bipartite networks. In the field of recommendation, the current studies on cold-start problem mainly focus on incorporating additional attributes or contents from the profiles of entities. Seung et al. [56] proposed a pairwise preference regression (PPR) model to tackle cold-start recommendation, where the model made use of all available information of users and items. Zeno et al. [57] described a method that mapped entity attributes to the latent features of a matrix factorization model. With such mappings, the factors of a matrix factorization model trained by standard techniques could be applied to the new-user and the new-item

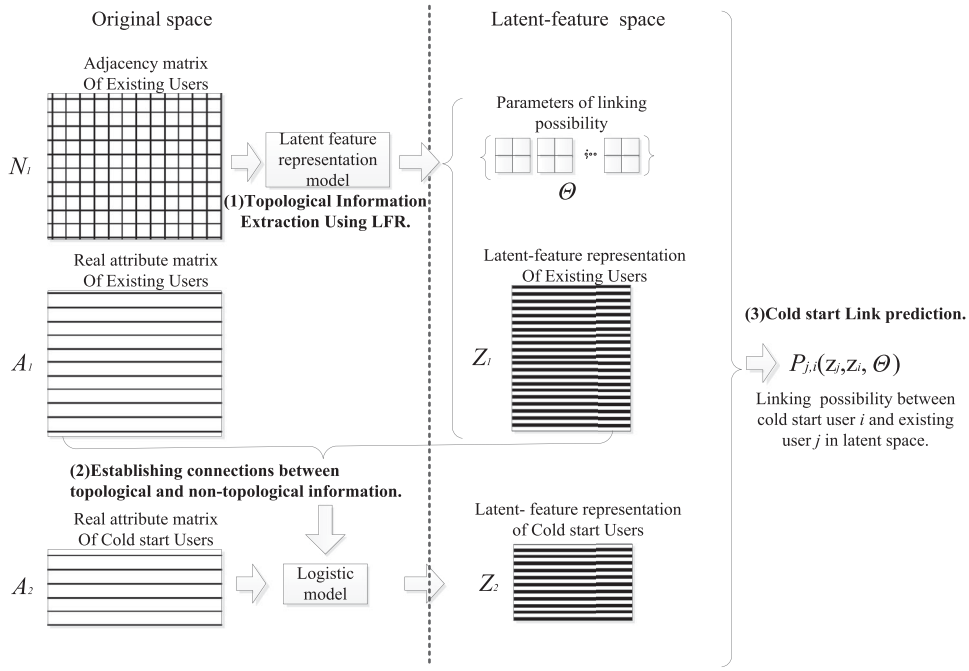


Fig. 3. Cold-start link prediction: Establishing connections between non-topological and topological information.

problem. Leung et al. [58] proposed a novel hybrid recommendation approach which made use of cross-level association rules to integrate content information. Weng et al. [59] combined the implicit relations between users item preferences and the additional taxonomic preferences so as to alleviate the cold-start problem. Martinez et al. [60] presented a hybrid recommendation system which combined a collaborative filtering algorithm with a knowledge-based method. Sedhain et al. [61] proposed a social collaborative filtering (SCF) method for cold-start recommendations, where the method directly utilized user's social network content in a novel extension of item-based collaborative filtering. Besides, many other studies [62], [63], [64] used social network to alleviate cold-start problem. Therefore, how to make use of the auxiliary information is the key to deal with the problems under the cold-start scenarios.

3 Method

The aim of the paper is to establish a connection between non-topological and topological information in social networks, and to utilize this connection to predict links for cold-start users. To establish this connection, topological information (existing users' latent-feature representation and linking measure) are at first extracted by a latent-feature representation model, then connections between topological and non-topological information are established through logistic model, finally cold-start users' latent-feature representation is obtained by the logistic model. In this way, the existing users and cold-start users are simultaneously represented in a unified latent-feature space, and the linking possibility between cold-start users and existing users can be calculated in this latent-feature space. Fig. 3 shows the overview of cold-start link prediction method by a latent-feature representation for cold-start users and existing users in a unified space. We summarize our approach in the following three steps:

(1) *Topological Information Extraction.* In this part, a latent-feature representation model is designed to learn a vectorized latent-feature representation for the existing users. The LFR model is a probability generative model, in which the latent features of users in a network are first generated through specific distributions. Each edge in the network between users is then generated through a linking possibility which is defined in the latent-feature space. According to model's generative process, we can finally obtain the existing users' latent-feature representation and some parameters of linking possibility by using a maximum likelihood estimation method. Details of this latent-feature network generative model and latent-feature representation for existing users will be elaborated in Section 3.1.

(2) *Establishing Connections between Topological and Non-Topological Information.* The second part aims to obtain the latent-feature representation of cold-start users who lack topological information. A logistic model is proposed to establish the connections between users' topological and non-topological information. In this way, latent-feature representation of cold-start users is learned. Details of this learning model will be provided in Section 3.2.

(3) *Cold-Start Link Prediction in the Latent-Feature Space.* Based on the latent-feature representation of all the existing users and cold-start users in the unified latent space, we measure their linking probability according to linking measure of the latent space. Thus, the cold-start link prediction could be achieved through this linking possibility in the latent-feature space. Details of the final cold-start link prediction algorithm will be elaborated in Section 3.3.

For convenience, we list out the mainly used notations in this document (see Table 1).

TABLE 1
Notations

Symbol	Explanation
$n_1 \in R$	Number of existing users
$n_2 \in R$	Number of cold-start users
$N_1 \in R^{n_1 \times n_1}$	Adjacency matrix of existing users
$A_1 \in R^{n_1 \times m}$	Real attribute matrix of existing users
$A_2 \in R^{n_2 \times m}$	Real attribute matrix of cold-start users
$Z_1 \in R^{n_1 \times L}$	Latent feature matrix of existing users
$Z_2 \in R^{n_2 \times L}$	Latent feature matrix of cold-start users

3.1 Topological Information Extraction

Many topological characteristics, such as homophily, heavy-tailed degree distributions and small diameter, exist in social network. In this section, we aim to extract the topological information of networks by using one kind of machine learning method, i.e., latent-feature representation.

In LFR, each node of network is modelled as a latent-feature vector and the linking possibility between each pair of nodes is based on one kind of link-affinity matrix which derives from the Multiplicative Attribute Graph model [65]. Intuitively, MAG model is designed to capture many connectively properties in networks such as homophily and heterophily, where homophily means that nodes with a certain feature are more likely to create links among themselves; heterophily means that nodes without a certain feature are more likely to links among themselves. Precisely, in a simplest case, suppose a pair of nodes i and j with the corresponding binary latent-feature vectors $\vec{z_i} = <z_{1,1}, \ldots, z_{i,\ell}, \ldots, z_{i,L}>$ $(z_{i,\ell} \in \{0,1\})$ and $\vec{z_j} = <z_{1,1}, \ldots, z_{j,\ell}, \ldots, z_{j,L}>$ $(z_{j,\ell} \in \{0,1\})$, the probability of an edge p_{ij} (see Equation (1)) is the product over the entries of attribute link-affinity matrices Θ_ℓ in Θ (see Equations (2) and (3)). $\Theta_\ell[z_{i,\ell}, z_{j,\ell}]$ indicates the affinity with which a pair of nodes i and j form a link, given that each ℓth attribute of nodes i and j takes value $z_{i,\ell}$ and value $z_{j,\ell}$ respectively. Intuitively, the higher the value $\Theta_\ell[z_{i,\ell}, z_{j,\ell}]$ is, the stronger the effect of the particular attribute combination $[z_{i,\ell}, z_{j,\ell}]$ is on forming a link

$$p_{ij} = \prod_{\ell=1}^{L} \Theta_\ell[z_{i,\ell}, z_{j,\ell}] \tag{1}$$

$$\Theta_\ell = \begin{bmatrix} a_\ell & b_\ell \\ c_\ell & d_\ell \end{bmatrix} \tag{2}$$

$$\Theta = \{\Theta_1, \ldots, \Theta_\ell, \ldots, \Theta_L\}. \tag{3}$$

As is shown in Fig. 4, nodes i and j possess three-dimension binary attribute vectors [0, 0, 1] and [1, 0, 1] respectively. We then select the corresponding matrixes' elements $\Theta_1[0,1] = b_1$, $\Theta_2[0,0] = a_2$, and $\Theta_3[1,1] = d_3$ as the multiplication factors of the linking probability p_{ij}. The definition of nodes' latent-feature vector allows rich flexibility in modeling the network and the link-affinity matrix can uncover meaningful network topological properties, which will be further explained after introducing the model's generative process.

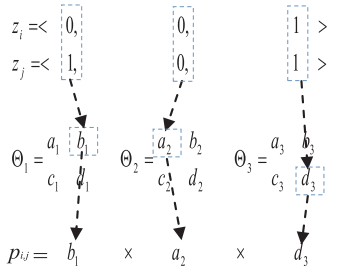


Fig. 4. Linking possibility $p_{i,j}$ between i and j.

The model has a very simple generative process. As is shown in Fig. 5, first, each dimension's latent feature of node i in network N, $z_{i,\ell}(z_{i,\ell} \in \{0,1\})$, is generated from a Bernoulli distribution with parameter α_ℓ, and thus L Bernoulli distributions are used to generate the $\vec{z_i}$. Next, each link in network N, $N_{i,j}$, is also generated with a linking possibility $p_{i,j}$. Finally, network N is generated by such repetitive process

$$\Theta_4 = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}. \tag{4}$$

The meaning of the model construction is that, first of all, modelling the nodes as $\vec{z_i} = <z_{1,1}, \ldots, z_{i,\ell}, \ldots, z_{i,L}>, z_{i,\ell} \in \{0,1\}$ enables nodes to belong to multiple groups at the same time. Secondly, based on the value of ℓth latent feature, nodes in a network can be divided into two groups: the value of the nodes in group one is 1 because they possess the ℓth latent feature, while the value of the nodes in group two is 0 because they don't have the latent feature. In this case, the meaning of link-affinity matrix Θ_ℓ is that it can tell us the possibility of forming links between nodes with or/and without the ℓth latent feature. Because the value of each ℓth latent feature is 0 or 1, the link-affinity matrix corresponding to each ℓth dimension is a 2×2 matrix. The link-affinity matrix can uncover meaningful network topological property. Take the value of link-affinity matric Θ_4 (see Equation (4)) for example, $\Theta_4[0,0] = \Theta_4[1,1] = 0.9$ means that nodes sharing the value 0 or 1 are more likely to link, while $\Theta_4[0,1] = \Theta_4[1,0] = 0.1$ means that the linking possibility between nodes is low if they have different values. This illustrates homophily in a social network

$$\Theta_5 = \begin{bmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{bmatrix}. \tag{5}$$

Equation (5) is an example of heterophily. The value of the link-affinity matrix Θ_5 is just the opposite of

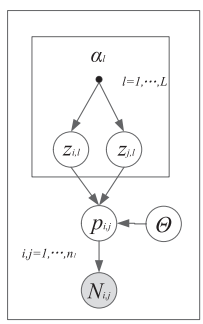


Fig. 5. The generative process in plate notation.

Equation (4), which means that nodes without the same value of the attribute are more likely to link

$$\Theta_6 = \begin{bmatrix} 0.9 & 0.5 \\ 0.5 & 0.1 \end{bmatrix}. \tag{6}$$

Furthermore, the value link-affinity matrix Θ_6 (see Equation (6)) shows the core-periphery characteristic. Specifically, the probability of linking between "0 nodes" is the highest ($\Theta_6[0,0] = 0.9$); the probability of linking between "1 nodes" is the lowest ($\Theta_6[1,1] = 0.1$); the probability of linking between "1 nodes" and "0 nodes" is in the middle. This means that nodes of the core are the most connected and that the nodes of the periphery are more likely to be connected to the core than among themselves [66]. Besides, many other connectivity patterns (such as heavy-tailed degree distributions, small diameter and unique giant connected component) can be captured, which has been proved by Kim and Leskovec [65].

From the above mentioned latent-feature representation model's generative process, we know that each link is independently generated in this model. Therefore, the likelihood $P(N/\Theta, \vec{\alpha})$ of a given network N can be expressed as the product of the linking probabilities over the edges and non-edges of the network, as follows:

$$\begin{aligned} P(N/\Theta, \vec{\alpha}) &= \sum_{Z} P(N, Z/\Theta, \vec{\alpha}) \\ &= \sum_{Z} P(N/\Theta, Z) P(Z/\vec{\alpha}) \\ &= \prod_{N_{i,j}=1} p_{i,j} \prod_{N_{i,j}=0} (1 - p_{i,j}) \\ &\quad \prod_{z_{i,\ell}=1} \alpha_\ell \prod_{z_{i,\ell}=0} (1 - \alpha_\ell). \end{aligned} \tag{7}$$

In this section, our aim is to extract topological information by learning nodes' latent-feature representation Z and the link-affinity matrix set Θ. Therefore, we need to estimate these parameters by maximizing the network log-likelihood $\ell(Z, \Theta)$

$$\ell(Z, \Theta) = log P(N/\Theta, \vec{\alpha}) = log \sum_{Z} P(N, Z/\Theta, \vec{\alpha}). \tag{8}$$

Variational expectation-maximization (EM) method, which is the most commonly-used optimization method is utilized to solve the network representation learning problem. Because the log-likelihood $\ell(Z, \Theta)$ is non-convex, a variational distribution $Q(Z)$ is introduced to approximate the posterior distribution $P(Z/N, \Theta, \vec{\alpha})$

$$Q(Z) = \prod_{i,\ell} Q(z_{i,\ell}), \tag{9}$$

where $Q(z_{i,\ell}) = \phi_{i,\ell}^{z_{i,\ell}} (1 - \phi_{i,\ell})^{1 - z_{i,\ell}}$ and $\phi = \{\phi_{i,\ell}\}$ are variational parameters according to [55]. Then, we deduce the lower bound $\ell_Q(Z, \Theta)$ of $\ell(Z, \Theta)$ (see Equation (10)), and maximize $\ell(Z, \Theta)$ through maximizing $\ell_Q(Z, \Theta)$

$$\begin{aligned} \ell(Z, \Theta) &= log \sum_{Z} Q(Z) \frac{P(N, Z/\Theta, \vec{\alpha})}{Q(Z)} \\ &\geq \sum_{Z} Q(Z) log \frac{P(N, Z/\Theta, \vec{\alpha})}{Q(Z)} = \ell_Q(Z, \Theta). \end{aligned} \tag{10}$$

In the E-step of the variational EM algorithm, we estimate the parameters ϕ by fixing parameters Z and Θ, and we aim to estimate the parameters Z and Θ by fixing parameters ϕ in the M-step. After the algorithm converges, related parameters in this model are calculated. Finally, the model's solution can provide us with the existing users' latent-feature representation Z_1 and the link-affinity matrix set Θ:

$$Z_1 = \begin{pmatrix} z_{1,1} & \cdots & z_{1,L} \\ z_{2,1} & \cdots & z_{2,L} \\ \cdots & \cdots & \cdots \\ z_{n_1,1} & \cdots & z_{n_1 L} \end{pmatrix} \tag{11}$$

$$\Theta = \{\Theta_1, \ldots, \Theta_\ell, \ldots, \Theta_L\}. \tag{12}$$

The final learned parameters (see in Equations (11) and (12)) are the representation of topological information of the network which is the basis of establishing the connections between topological information and non-topology information in the next section.

3.2 Establishing Connections between Topological and Non-Topological Information

Apart from topological information, a social network is often associated with rich auxiliary information, such as users' profile and rich text information. The main focus of this section is how to establish the relation between topological information of network structure and non-topological information, which is also the key to cold-start link prediction.

In Section 3.1, we extract the topological information of the network structure, i.e., existing users' latent-feature representation Z_1 and the link-affinity matrix set Θ, which is the representation of the network structure in L dimension latent space. To establish connections between the topological and non-topological information, we need to map users into this L dimension latent space by using their non-topological information (the auxiliary information). Suppose that we denote each user i's vectorized non-topological information as $\vec{a_i} = <a_{i,1}, \ldots, a_{i,m}>$, all the existing users' real feature matrix is denoted as A_1

$$A_1 = \begin{pmatrix} a_{1,1} & \cdots & a_{1,m} \\ a_{2,1} & \cdots & a_{2,m} \\ \cdots & \cdots & \cdots \\ a_{n_1,1} & \cdots & a_{n_1,m} \end{pmatrix}, \tag{13}$$

where each line of A_1 corresponds to the real features of a user. Next, the problem in this section can be formally denoted as follows:

$$\begin{pmatrix} a_{1,1} & \cdots & a_{1,m} \\ a_{2,1} & \cdots & a_{2,m} \\ \cdots & \cdots & \cdots \\ a_{n_1,1} & \cdots & a_{n_1,m} \end{pmatrix} \Rightarrow \begin{pmatrix} z_{1,1} & \cdots & z_{1,L} \\ z_{2,1} & \cdots & z_{2,L} \\ \cdots & \cdots & \cdots \\ z_{n_1,1} & \cdots & z_{n_1,L} \end{pmatrix}. \tag{14}$$

If the users' 0-1 latent-feature representation in each dimension ℓ of Z_1 could be seen as a binary partition in the latent space, then all the representation of L latent dimensions are L latent binary partitions. In order to establish this connections between A_1 and Z_1, a logistic model is proposed, where users could be mapped into the L dimensions'

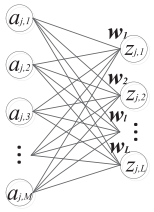


Fig. 6. Relationship between real and latent features in the L logistic models.

latent partitions by using L logistic models. Each logistic model is used to establish the relation between each ℓ dimension's partition and A_1

$$\begin{pmatrix} a_{1,1} & \dots & a_{1,m} \\ a_{2,1} & \dots & a_{2,m} \\ \dots & \dots & \dots \\ a_{n_1,1} & \dots & a_{n_1,m} \end{pmatrix} \Rightarrow \begin{pmatrix} z_{1,\ell} \\ z_{2,\ell} \\ \dots \\ z_{n_1,\ell} \end{pmatrix}. \tag{15}$$

For each node i's real feature $\vec{a_i}$ of A_1, we have

$$\begin{aligned} y_{i,\ell} &= \frac{1}{1+exp(-\vec{w_\ell}^T \vec{a_i})} \\ z_{i,\ell} &\sim Bernoulli(y_{i,\ell}), \end{aligned} \tag{16}$$

where, $\vec{w_\ell}^T$ is the logistic model parameter for the ℓth latent dimension. Fig. 6 shows the relationship between a user i's real features and his/her latent features in the L logistic models.

To establish the relationship between real and latent features for every user, we need to train the L logistic models (parameter estimation for $\{\vec{w_\ell}^T\}_{l=1,\dots,L}$), and predict the latent-feature representation of a user's real-attribute vector input. Algorithm 1 gives the pseudo code of parameter estimation and the latent-feature prediction for cold-start users.

Algorithm 1. Cold-Start Users' Latent-Feature Learning Based on Logistic Model

Input: Existing users' real attributes $\vec{a_1}, \dots, \vec{a_{n_1}}$, latent-feature representation $\vec{z_1}, \dots, \vec{z_{n_1}}$ and cold-start users' real attributes $\vec{a_{n_1+1}}, \dots, \vec{a_n}$.
Output: cold-start users' latent-feature representation $\vec{z_{n_1+1}}, \dots, \vec{z_n}$.

(Training process)
Initialize $\vec{w_1}, \dots, \vec{w_L}$ of the L logistic models
repeat
  for $\ell = 1$ to L **do**
    $\vec{w_l}^{new} := \vec{w_\ell}^{old} + \lambda \nabla_{\vec{w_\ell}} \ell_w$
  end for
until Convergence

(Predicting process)
for $n = n_1 + 1$ to $n_1 + n_2$ **do**
  for $\ell = 1$ to L **do**
    $z_{n,\ell} \Leftarrow \frac{1}{exp(-\vec{w_\ell}^T \vec{a_n})}$
  end for
end for
Return: $\vec{z_{n_1+1}}, \dots, \vec{z_n}$

As shown in Algorithm 1, a stochastic gradient method is used to update the logistic model's parameters in the training process, where $\nabla_{\vec{w_\ell}} \ell_w$ is the stochastic gradient to update each $\vec{w_\ell}$ in a random dot $\vec{a_i}$

$$\nabla_{\vec{w_\ell}} \ell_w = (z_{i,\ell} - y_{i,\ell}) \vec{a_i}. \tag{17}$$

After the parameters $\vec{w_1}, \dots, \vec{w_L}$ are learned, they, together with cold-start users' real attributes $\vec{a_{n_1+1}}, \dots, \vec{a_n}$, could be seen as the input of the algorithm predicting process. Finally, we obtain cold-start users' latent-feature representation $\vec{z_{n_1+1}}, \dots, \vec{z_n}$.

3.3 Cold-Start Link Prediction in Latent Space

In Sections 3.1 and 3.2, we have described the extraction of topological information and the establishment of connections between non-topological information and topological information respectively, and in this section we will focus on cold-start link prediction in the latent space.

Suppose we have vectorized users' features $\vec{a_1}, \dots, \vec{a_n}$, each cold-start user's latent-feature representation (see Algorithm 1) could be obtained by using the learned logistic model mentioned in Section 3.2. Then the existing users and cold-start users are represented in a unified latent space. Thus, predictions could be made about whether there will be links between cold-start users and the existing users by measuring the linking possibility. Assume that one latent attribute vector of a cold-start user j is $\vec{z_j} = < z_{j,1}, \dots, z_{j,\ell}, \dots, z_{j,L} >$, and one latent-feature vector of an existing user i is $\vec{z_i} = < z_{i,1}, \dots, z_{i,\ell}, \dots, z_{i,L} >$, then the linking possibility between cold-start user j and existing user i could be defined as follows:

$$p_{j,i} = \prod_{\ell=1}^{L} \Theta_\ell[z_{j,\ell}, z_{i,\ell}]. \tag{18}$$

The pseudo code of the cold-start link prediction mechanism will be shown in Algorithm 2.

Algorithm 2. Cold-Start Link Prediction

Input: Existing users' real attributes $\vec{a_1}, \dots, \vec{a_{n_1}}$, their network N_1, and cold-start user j's real attributes $\vec{a_{n_j}}$.
Output: Link list L_j of cold-start user j.
$L_j \leftarrow \varnothing$
Extract the topological information (existing users' latent representation Z_1 and link affinity matrix set Θ) by using the LFR
Learn cold-start user j's latent representation $\vec{z_j} = < z_{j,1}, \dots, z_{j,\ell}, \dots, z_{j,L} >$ by using the logistic model
for each existing user i **do**
  $p_{j,i} = \prod_{\ell=1}^{L} \Theta_\ell[z_{j,\ell}, z_{i,\ell}]$
  Put the link (j, i) in L_j
end for
Sort all links (j, i) in decreasing order according to $p_{j,i}$
Put the top k links (j, i) in the sorted list to L_j

4 Experiments

In this section, we conduct several experiments with the following purposes: (1) to find out the effective non-topological features in our cold-start link prediction method, (2) to find out whether our method is better in establishing connections

TABLE 2
Statistics of the Three Types of Networks

Datasets	Edges	Nodes	Density
Weibo1	6,371	340	5.53%
Weibo2	3,124	213	6.92%
Facebook1	2,519	333	2.28%
Facebook2	3,192	224	6.39%
Twitter1	2,861	231	5.38%
Twitter2	1,099	159	4.37%

between topological and non-topological information, (3) to find out whether our method is superior to other methods in cold-start link prediction, (4) to find out the impact of latent space dimension L on the performance of cold-start link prediction.

4.1 Data Sets

According to the experimental requirements of cold-start link prediction, each social network data set should contain a network structure and the auxiliary information of users. Finally, we adopt three types of datasets to conduct our experiments: Weibo, Facebook and Twitter.

Weibo,[1] which enables users to send and read microblogs, is a well-known online social network service in China. In Weibo, the two-way relationship between users form directed social networks, and its users have rich auxiliary information. The Weibo datasets[2] we will use in our study is derived from [67], which is crawled from Sina platform. Based on the original data, we extract two datasets for our experiments by selecting the users who have more than 50 followees and 300 microblogs. We use two types of features for each Weibo user, i.e., profile and microblogs. Among them, profile like gender and location is the basic feature for the Weibo users. We divide them into numerical features and non-numerical features. For the numerical features, such as age and the number of microblogs, we normalize them into [0, 1]. For the non-numerical features, such as location and verification status, we treat them as 0-1 category feature, that is to say, when a user has a non-numerical feature in his/her profile, the value of the feature is '1', and otherwise, it is '0'. For the microblogs, we extract users' keywords and weights to vectorize users' text features. Specifically, because each microblog is very short, we treat all the microblogs that one user have posted as one text document, then calculate TF-IDF value of each keyword in users' level, and extract keywords feature for each user according to the TF-IDF value. After that, by normalizing the TF-IDF value from 0 to 1, we obtain the vectorized users' keywords feature. Apart from the auxiliary information, details of the two Weibo networks Weibo1 and Weibo2 are shown in Table 2.

Apart from the Weibo datasets, we also use two types of open datasets: Facebook[3] and Twitter,[4] which are also well-known online social network services for exchanging and sharing information. Among them, Twitter is similar to Weibo, which is also a user-user directed network, while Facebook is different from them, which is a user-user undirected network. Both datasets are derived from Stanford Network Analysis Platform.[5] For each type of social network, we also select two datasets to evaluate our experiments and these datasets have been anonymized by replacing the internal IDs for each user with a new value, and users' attributes have been 0-1 vectorized by a unified treatment. In addition to these attributes, these networks Facebook1, Facebook2, Twitter1 and Twitter2 are respectively shown in Table 2.

4.2 Experimental Design and Metrics

According to the general experimental protocol for link prediction [2], [6], [9], [46], we first split the observed links into training data and testing data. In carrying out the cold-start link prediction experiments, we should note that the testing links should be obtained from all links of the selected users in a social network. Specifically, the testing data is the links of cold-start users which refer to some randomly selected nodes in each social network, and the training data is the links of existing users which refer to the remaining network (which is also called observed network in this paper) nodes. In this data partition process, the remaining network should be a connected graph. Thereupon, we use different amounts of training data (90, 92, 94, 96, 98 percent) to test the algorithms. For example, training data 90 percent means we randomly select 10 percent of the users from a social network as cold-start users, and all the links of these cold-start users as testing data, and the links of the remaining 90 percent existing users as the training data. We carry out the random selection five times independently.

In the data partition process, we need to know the following details: (1) For the topological information, when we divide the original network links into training data and testing data, it will make a difference between original network and the remaining network in topological information. We will analyze the correlation between this topological differences and results of cold-start link prediction in Section 4.4.2. (2) For the non-topological information, the partition of a network dataset do not change the non-topological feature spaces of nodes because the dataset partition will only remove the links of a part of nodes in a network, but the profile of nodes do not change.

The link prediction problem is intrinsically a very difficult binary prediction problem because of the sparseness of the social network. Given n nodes, we have a space of $n^2 - n$ possible links. Among them, only a very small fraction of links actually exist. As illustrated in Table 2, the existing links constitute a very small percentage of all possible links. Similar to many existing link prediction studies [6], we use the most frequently used metrics "Area under the receiver operating characteristic" (AUC) to measure the cold-start link prediction. This metric is viewed as a robust measure in the presence of imbalance [7]. Given the rank of all unobserved links for cold-start users, the AUC value can be interpreted as the probability of that a randomly chosen unobserved link is given a higher score than a randomly chosen nonexistent link [68].

1. http://www.weibo.com
2. https://aminer.org/Influencelocality
3. http://www.facebook.com
4. http://www.twitter.com
5. http://snap.stanford.edu

4.3 Comparison Methods

We will compare our method with three metric-based and two learning-based methods in link prediction field, and three typical cold-start recommendation methods.

4.3.1 Metric-Based Methods

Suppose a cold-start user i and an existing user j, let $\vec{a_i} = <a_{i,1}, a_{i,2}, \ldots, a_{i,m}>$ and $\vec{a_j} = <a_{j,1}, a_{j,2} \ldots, a_{j,m}>$ be the vectorized feature of user i and user j respectively.

(1) Leroy et al.'s [41] two-phase bootstrap probabilistic method can be directly adopted in the cold-start link prediction. In our comparison experiments, we will test various features in [41], and select the best one for comparison. The specific feature used in the comparison method will be made clear in the parts of experimental analysis (see Sections 4.4.2 and 4.4.3). For convenience, this method is denoted as Leroy.

(2) Pearson's Correlation (PC)-based method. This PC index is a measure of the linear correlation between i and j

$$PC = \frac{\sum_{k=1}^{m}(a_{i,k} - \overline{a_i})(a_{j,k} - \overline{a_j})}{\sqrt{\sum_{k=1}^{m}(a_{i,k} - \overline{a_i})^2} \cdot \sqrt{\sum_{k=1}^{m}(a_{j,k} - \overline{a_j})^2}}. \quad (19)$$

In the PC-based link prediction method, the PC value is positively correlated with linking possibility.

(3) Cosine Similarity-based method. It is a commonly-used measure for measuring similarity between users [42]

$$COS = \frac{\sum_{k=1}^{m} a_{i,k} a_{j,k}}{\sqrt{\sum_{k=1}^{m}(a_{i,k})^2} \cdot \sqrt{\sum_{k=1}^{m}(a_{j,k})^2}}. \quad (20)$$

For link prediction, the COS-based method indicates that two users are more likely to have a link if they have a larger COS value.

4.3.2 Learning-Based Methods

(4) Logistic regression-based link prediction [16], [46], [47], [69]. In our data sets, we use the common profile between two users to define the features of each pair of users, that is, if two users have a common profile such as common school, common location, and common profession, the feature value in this dimension is denoted by 1, otherwise it is denoted by 0. In this way, each pair of users can be vectorized, and used for the classification model.

(5) Support Vector Machines-based link prediction [10], [11], [16], [70], [71]. We use the same features as LR model, and use linear and rbf kernel respectively.

For the LR and SVM models, we employ Scikit-learn toolkit [72] to train and predict the formation of links.

4.3.3 Cold-Start Recommendation Methods

Cold-start recommendation methods are employed in recommendation field for making the cold-start link prediction.

(6) Social Collaborative Filtering. It is a social collaborative filtering method [61] that generalizes standard item-based collaborative filtering in the cold-start recommendation setting.

(7) Map-knn. It is a learning attribute-to-feature mappings method for cold-start recommendations [57].

(8) Pairwise Preference Regression. It is a regression approach for cold-start recommendation [56].

It is very essential to note that for the sake of convenience, our cold-start link prediction method based on latent-feature representation is denoted as CSLP-LFR.

TABLE 3
Effectiveness of Different Types of Non-Topological Information in CSLP-LFR

Users' feature	Weibo1	Weibo2
Profile	0.640 ± 0.042	0.653 ± 0.022
Keywords	0.635 ± 0.036	0.650 ± 0.020
Profile+Keywords	$\mathbf{0.659 \pm 0.044}$	$\mathbf{0.672 \pm 0.027}$

4.4 Experimental Results

All the experiments are conducted on an Intel i7-2006 2 Core 3.4 GHz with 8 GB memory. In our CSLP-LFR method, we select the latent dimension $L = 7$ for datasets Weibo1, Facebook1, and Twitter1, and $L = 6$ for datasets Weibo2, Facebook2, and Twitter2. The selection of L will be discussed in Section 4.4.4.

4.4.1 Find Out the Effective Non-Topological Feature in Our Method

Features are critical to machine learning techniques, and they are also critical to our cold-start link prediction method. Therefore, we compare different features to find out the effective non-topological features in our cold-start link prediction method. The results of the cold-start link prediction by using different types of non-topological features are shown in Table 3 (This experiments use 90 percent training data setting which has been introduced in Section 4.2).

Table 3 shows the results of two types of features and their combined features in our cold-start link prediction model with datasets Weibo1 and Weibo2. The results show little difference by using each type feature alone. When the two types of features are combined, we get the best experimental results. Therefore, in the next experiments with datasets from Weibo, we employ the combined features in our method and other methods.

4.4.2 Correlation Analysis

The purpose of CSLP-LFR is to solve problem by establishing connections between non-topological and topological information in social networks. In this section, we will analyze the correlation between results of cold-start link prediction and differences of topology between existing network and target network, and then verify whether our method is better than the other methods in establishing connections between topological and non-topological information.

To conveniently analyze the correlation, we will define a kind of topological similarity between existing

TABLE 4
Correlation between Topological Similarities and Results of Different Cold-Start Methods

Datasets	CSLP-LFR	Leroy(prod)	PC	COS	LR	SVM(linear)	SVM(rbf)	SCF	Map-knn	PPR
Weibo1	0.995	0.369	0.937	0.660	0.961	0.013	0.013	0.506	0.420	-0.630
Weibo2	0.704	0.006	0.701	0.632	0.148	0.036	0.036	−0.508	0.263	0.007

network and target network based on LFR. As is introduced in Section 3, given a network N, we can extract the representation of topological information Z_N and Θ_N based on LFR, and the topological similarity measure could be defined as follows:

Definition 1. *Let Z_{N_1} and Θ_{N_1} be the latent-feature representation of network N_1, and let Z_{N_2} and Θ_{N_2} be the latent-feature representation of network N_2. The topological similarity between network N_1 and N_2 is defined as*

$$S(N_1, N_2) = \frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{2}\left(\frac{1}{exp(|\mu_{1,\ell}-\mu_{2,\ell}|)}+\frac{1}{exp(|\Theta_{1,\ell}-\Theta_{2,\ell}|)}\right), \tag{21}$$

where $\mu_{1,\ell}=\frac{1}{n_1}\sum_{i=1}^{n_1} z_{i,\ell}$ and $\mu_{2,\ell}=\frac{1}{n_2}\sum_{j=1}^{n_2} z_{j,\ell}$ correspond to 0-1 value distribution of Z_{N_1} and Z_{N_2} in ℓth dimension respectively. The definition is easy to understand, for it just combines two parts (the value of latent-feature distribution $\frac{1}{exp(|\mu_{1,\ell}-\mu_{2,\ell}|)}$ and connectivity pattern $\frac{1}{exp(|\Theta_{1,\ell}-\Theta_{2,\ell}|)}$ in each ℓ dimension) of the similarities.

To quantitatively analyze the correlation between topological similarities and results of different cold-start link prediction methods, we use Pearson Correlation Coefficient to measure the correlation. Suppose that we have a series of observed networks $N_1, N_2, \ldots, N_D$, and a target network N, the topological similarities between each existing network and the target network are $S(N, N_1)$, $S(N, N_2)$, $\ldots$, $S(N, N_D)$, and the results of one cold-start link prediction method f for each observed network are $AUC_1, AUC_2, \ldots, AUC_D$. The Pearson Correlation Coefficient between the cold-start results and topological similarities is defined as

$$\frac{\sum_{d=1}^{D}(S(N,N_d)-\overline{S(N,N_d)})(AUC_d-\overline{AUC_d})}{\sqrt{\sum_{d=1}^{D}(S(N,N_d)-\overline{S(N,N_d)})^2}\cdot\sqrt{\sum_{d=1}^{D}(AUC_d-\overline{AUC_d})^2}}.$$

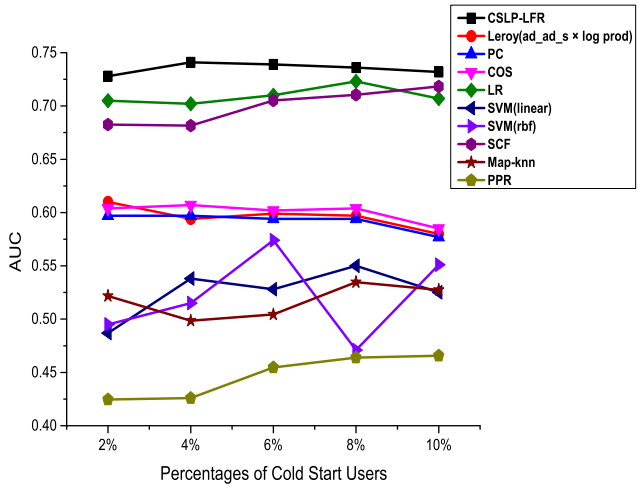
The results of the correlation among different cold-start link prediction methods are shown in Table 4. CSLP-LFR and PC show high positive correlation between results of the cold-start link prediction and topological similarities, and our method shows the highest positive correlation, which reveals that our method performs better in establishing connections between topological and non-topological information in social network.

4.4.3 Comparisons with Different Methods in Open Datasets

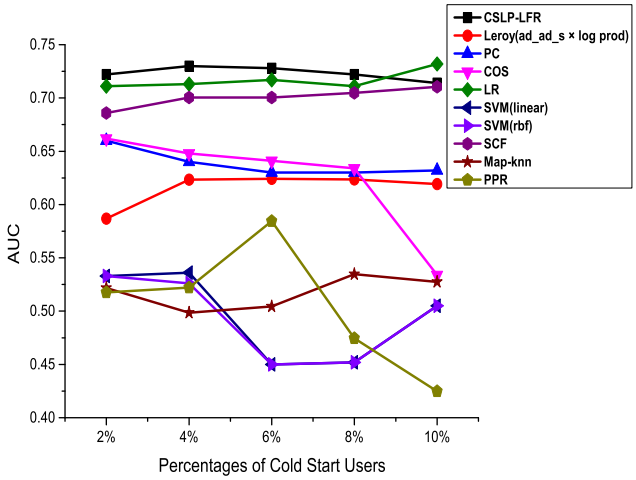
For the open datasets, Facebook and Twitter, we use different amounts of training data (90, 92, 94, 96, 98 percent) setting to compare different methods.

As shown in Figs. 7a and 7b, our method almost consistently outperforms other approaches in all the settings of the two undirected Facebook datasets.

The metric-based methods COS and PC directly define similarity and correlation between users by using users' attributes, but their low performances indicate that they are not highly effective for predicting links. For the Leroy ($ad\_ad\_s \times logprod$) method [41], we used the best performing results with the feature combination $ad\_ad\_s \times logprod$. However, Leroy ($ad\_ad\_s \times logprod$) results is still not good enough in Facebook1 (see Fig. 7a) or in Facebook2 (see Fig. 7b). Instead of directly defining the linking measure by using users' real features like COS and PC, our method first extracts the topological information (representation of users' latent features and link affinity matrix) of the network, and then establishes

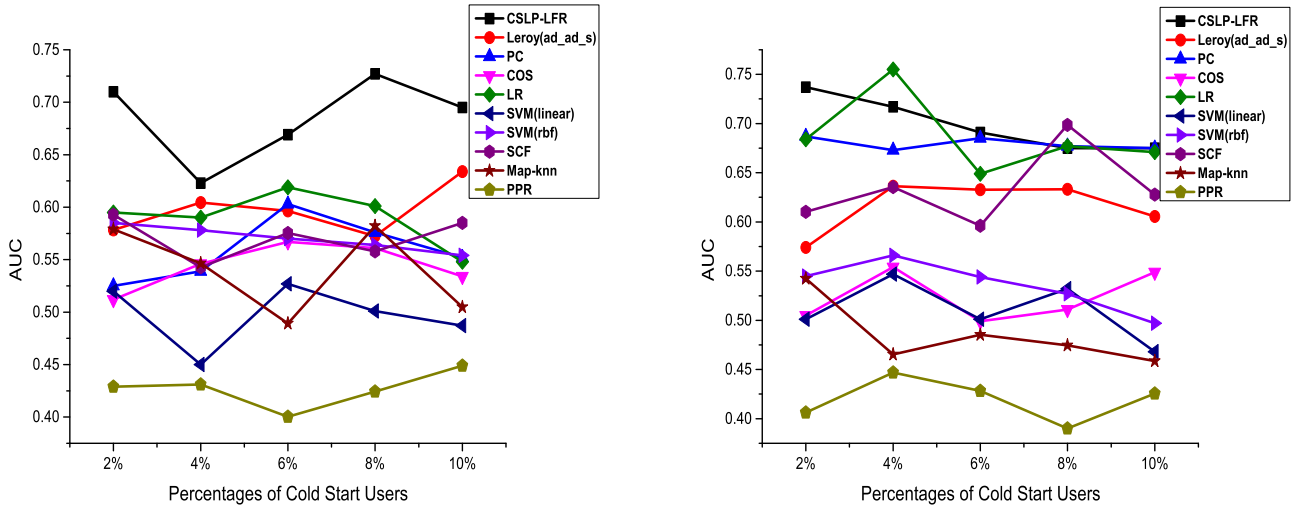


(a) Results on Facebook1



(b) Results on Facebook2

Fig. 7. Performance of cold-start link prediction on Facebook datasets.



(a) Results on Twitter1 (b) Results on Twitter2

Fig. 8. Performance of cold-start link prediction on Twitter datasets.

the connections between topological information and non-topology information (users' auxiliary information). Therefore, our method is superior to the metric-based methods.

As far as the classification-based methods are concerned, the performance of LR-based link prediction is superior to SVM-based link prediction, and the LR model performs better than all the other compared metric-based methods. However, it is less effective than our method. Classification-based link prediction methods can establish a connection between users' auxiliary information and network structure by training a model in our data sets, while topological features of cold-start users in the classification model are absent. Therefore, no topological features could be used to train a classification model and the performances of classification-based methods are seriously affected. As mentioned in the related work in Section 2.2 "Classification-based methods", many works have verified that the combination of non-topological features and topological features in a classification model can improve the performance of link prediction, while the non-topological features alone is not sufficient to get good performance.

As for the three cold-start recommendation methods SCF, Map-knn and PPR, SCF is the best among them. However, it is inferior to our method. The two other methods Map-knn and PPR performed poorly in the compared methods. Therefore, it is not proper for us to directly adopt the cold-start recommendation methods to deal with the cold-start link prediction problem.

Another finding in Figs. 7a and 7b is that as the percentages of cold-start users increases, the performance of our method decreases to some extent, but in most cases its performance is better than that of other methods.

The different methods on two directed networks, Twitter1 and Twitter2, are also compared, and the comparison of their performances is shown in Figs. 8a and 8b. Similarly, our method outperforms other approaches. The difference is that the results show fluctuation of the compared methods in the two datasets.

4.4.4 *Impact of Latent Feature Dimension L*

In our cold-start link prediction method, the model's parameter L controls the dimensions of the latent-feature space of network users. If L is small, the model will not be sufficient to capture the structural property of this network, and this is described as underfitting. By contrast, if L is large, overfitting will occur in the learning process. In this case, our model will produce lower results in the link prediction phase. The value of L also affects the time in the learning process. In this section, we will discuss how the model parameter L affects the performance of the final cold-start link prediction and the time of model learning.

Figs. 9a, 9b, 9c, 9d, and 9e show the performance of cold-start link prediction in Facebook and Twitter along the dimensions L. As L increases, the improvement of AUC becomes increasingly marginal, and converge to a stable accurate value.

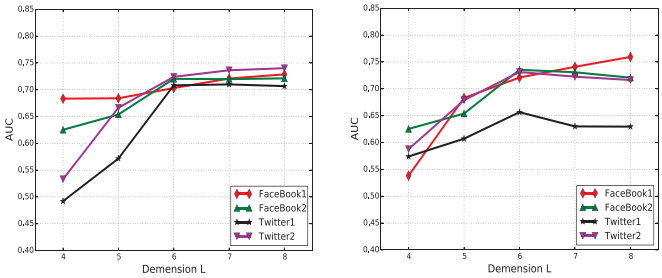
Figs. 10a and 10b show that the time of model learning in both directed and undirected networks is substantially linear.

Hence, we should limit the value of L so that an acceptable compromise could be reached between the performances of cold-start link prediction and time consumption.

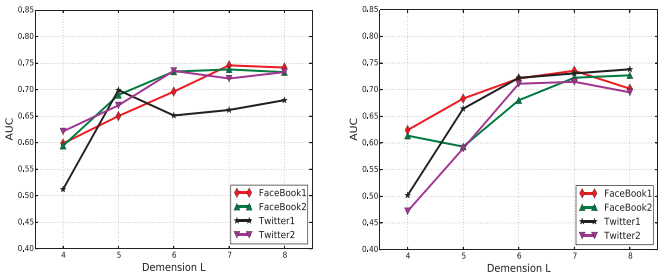
5 Conclusion and Future Work

How to accurately infer links for cold-start users in a network by using non-topological information is a difficult problem in link prediction. Many traditional link prediction methods have been proposed and they have indeed contributed a lot to the solution of link prediction problems. Due to the fact that topological structure information of cold-start users could not be found, most of the traditional methods could not satisfactorily address the cold-start link prediction problem.

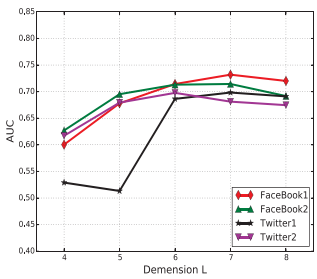
To solve this problem, we propose the latent-feature representation in a unified latent-feature space for cold-start link prediction, in which the connections between topological and non-topological information are established, and the latent features of existing users and cold-start users are represented by a LFR model and a logistic mapping model respectively. In the unified latent-feature space, the linking possibility between cold-start users and existing users are calculated. To assess the performance of the latent-feature



(a) 2% as cold-start users (b) 4% as cold-start users



(c) 6% as cold-start users (d) 8% as cold-start users



(e) 10% as cold-start users

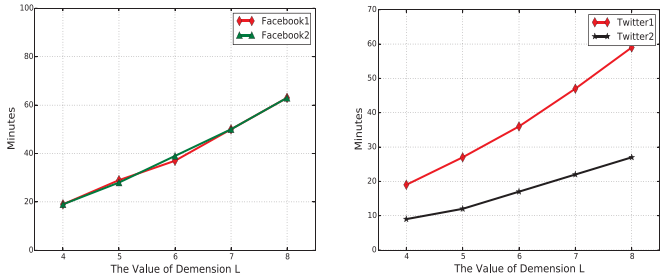
Fig. 9. Performance of cold-start link prediction in Facebook and Twitter along the dimensions L.

representation method for the cold-start link prediction, we compare our approach with three metric-based methods and two classification-based methods in link prediction field, and three cold-start recommendation methods. Experiments with three types of real-world social networks Weibo, Facebook and Twitter show that the proposed approach is very effective in cold-start link prediction.

This work has many potential directions in the future. For example, we could study how to conduct incremental learning on latent-feature representation model so that the model could be adapted to a dynamic circumstance when cold-start nodes consecutively join a network. Besides, it also has many real applications based on the cold-start link prediction model. For example, our method can contribute to speed up the initial growth for a new online App based on social networking services (SNS).

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 61432011, U1435212, 61373082, 61322211), National Key Basic Research and Development Program of China (973) (No. 2013CB329404), National High Technology Research and Development Program of China (863) (No. 2015AA015407), and Open Project Foundation of Information Security Evaluation Center of Civil Aviation, Civil Aviation University of China (No. CAAC-ISECCA-201402). Jiye Liang is the corresponding author.



(a) Results in Facebook (b) Results in Twitter

Fig. 10. Impact of dimension L on time of latent-feature representation in Facebook and Twitter.

References

[1] A. Clauset, C. Moore, and M. E. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, 2008.

[2] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," *Eur. Physical J. B-Condensed Matter Complex Syst.*, vol. 71, no. 4, pp. 623–630, 2009.

[3] K. Jahanbakhsh, V. King, and G. C. Shoja, "Predicting missing contacts in mobile social networks," *Pervasive Mobile Comput.*, vol. 8, no. 5, pp. 698–716, 2012.

[4] K. Juszczyszyn, K. Musial, and M. Budka, "Link prediction based on subgraph evolution in dynamic social networks," in *Proc. 2011 IEEE Int. Conf. Privacy Secur. Risk Trust IEEE Int. Conf. Social Comput.*, 2011, pp. 27–34.

[5] B. Bringmann, M. Berlingerio, F. Bonchi, and A. Gionis, "Learning and predicting the evolution of social networks," *Intell. Syst.*, vol. 25, no. 4, pp. 26–35, 2010.

[6] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: Statistical Mechanics Appl.*, vol. 390, no. 6, pp. 1150–1170, 2011.

[7] P. Wang, B. Xu, Y. Wu, and X. Zhou, "Link prediction in social networks: The state-of-the-art," *Sci. China Inf. Sci.*, vol. 58, no. 1, pp. 1–38, 2014.

[8] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019–1031, 2007.

[9] S. Scellato, A. Noulas, and C. Mascolo, "Exploiting place features in link prediction on location-based social networks," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1046–1054.

[10] M. Pavlov and R. Ichise, "Finding experts by link prediction in co-authorship networks," in *Proc. 2nd Int. Conf. Finding Experts Web Semantics*, 2007, vol. 290, pp. 42–55.

[11] T. Wohlfarth and R. Ichise, "Semantic and event-based approach for link prediction," in *Proc. 7th Int. Conf. Practical Aspects Knowl. Manage.*, 2008, pp. 50–61.

[12] A. K. Menon and C. Elkan, "Link prediction via matrix factorization," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2011, pp. 437–452.

[13] D. J. Marchette and C. E. Priebe, "Predicting unobserved links in incompletely observed networks," *Comput. Statistics Data Anal.*, vol. 52, no. 3, pp. 1373–1386, 2008.

[14] K. Palla, D. Knowles, and Z. Ghahramani, "An infinite latent attribute model for network data," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 1607–1614.

[15] K. Miller, M. I. Jordan, and T. L. Griffiths, "Nonparametric latent feature models for link prediction," in *Proc. Advances Neural Inf. Process. Syst.*, 2009, pp. 1276–1284.

[16] J. Zhang, Z. Fang, W. Chen, and J. Tang, "Diffusion of "following" links in microblogging networks," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 8, pp. 2093–2106, Aug. 2015.

[17] F. Lorrain and H. C. White, "Structural equivalence of individuals in social networks," *J. Math. Sociology*, vol. MS-1, no. 1, pp. 49–80, 1971.

[18] P. Jaccard, *Etude Comparative De La Distribution Florale Dans Une Portion Des Alpes Et Du Jura*. Montreux, Switzerland: Impr. Corbaz, 1901.

[19] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[20] E. Leicht, P. Holme, and M. E. Newman, "Vertex similarity in networks," *Physical Rev. E*, vol. 73, no. 2, 2006, Art. no. 026120.

[21] Y.-X. Zhu, L. Lü, Q.-M. Zhang, and T. Zhou, "Uncovering missing links with cold ends," *Physica A: Statistical Mechanics Appl.*, vol. 391, no. 22, pp. 5769–5778, 2012.

[22] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social Netw.*, vol. 25, no. 3, pp. 211–230, 2003.

[23] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[24] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. P-18, no. 1, pp. 39–43, 1953.

[25] L. Lü, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Physical Rev. E*, vol. 80, no. 4, 2009, Art. no. 046122.

[26] H.-H. Chen, L. Gou, X. L. Zhang, and C. L. Giles, "Discovering missing links in networks using vertex similarity measures," in *Proc. 27th Annu. ACM Symp. Appl. Comput.*, 2012, pp. 138–143.

[27] A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos, "Fast and accurate link prediction in social networking systems," *J. Syst. Softw.*, vol. 85, no. 9, pp. 2119–2132, 2012.

[28] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 355–369, Mar. 2007.

[29] G. Jeh and J. Widom, "SimRank: A measure of structural-context similarity," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 538–543.

[30] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Comput. Netw. ISDN Syst.*, vol. 30, no. 1, pp. 107–117, 1998.

[31] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New perspectives and methods in link prediction," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 243–252.

[32] X. Xie, "Potential friend recommendation in online social network," in *Proc. IEEE/ACM Int. Conf. Green Comput. Commun. Int. Conf. Cyber Physical Social Comput.*, 2010, pp. 831–835.

[33] Y. Dong, et al., "Link prediction and recommendation across heterogeneous social networks," in *Proc. IEEE 12th Int. Conf. Data Mining*, 2012, pp. 181–190.

[34] S. Wan, Y. Lan, J. Guo, C. Fan, and X. Cheng, "Informational friend recommendation in social media," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2013, pp. 1045–1048.

[35] J. Tang, et al., "Topic level expertise search over heterogeneous networks," *Mach. Learn.*, vol. 82, no. 2, pp. 211–237, 2011.

[36] W. X. Zhao, J. Wang, Y. He, J.-Y. Nie, J.-R. Wen, and X. Li, "Incorporating social role theory into topic models for social media content analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 4, pp. 1032–1044, Apr. 2015.

[37] J. Tang, T. Lou, J. Kleinberg, and S. Wu, "Transfer link prediction across heterogeneous social networks," *ACM Trans. Inf. Syst.*, vol. 9, no. 4, 2010, Art. no. 43.

[38] Q. Cao, Z. Wang, H. Qi, and J. Liao, "Friendbook: A semantic-based friend recommendation system for social networks," *IEEE Trans. Mobile Comput.*, vol. 14, no. 3, pp. 538–551, Mar. 2015.

[39] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[40] L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer, "Friendship prediction and homophily in social media," *ACM Trans. Web*, vol. 6, no. 2, pp. 1–33, 2012.

[41] V. Leroy, B. B. Cambazoglu, and F. Bonchi, "Cold start link prediction," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 393–402.

[42] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Effects of user similarity in social media," in *Proc. 5th ACM Int. Conf. Web Search Data Mining*, 2012, pp. 703–712.

[43] P. Bhattacharyya, A. Garg, and S. F. Wu, "Analysis of user keyword similarity in online social networks," *Social Netw. Anal. Mining*, vol. 1, no. 3, pp. 143–158, 2011.

[44] R. N. Lichtenwalter and N. V. Chawla, "Vertex collocation profiles: Subgraph counting for link analysis and prediction," in *Proc. 21st Int. Conf. World Wide Web*, 2012, pp. 1019–1028.

[45] H. R. De Sá and R. B. Prudêncio, "Supervised link prediction in weighted networks," in *Proc. Int. Joint Conf. Neural Netw.*, 2011, pp. 2281–2288.

[46] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 641–650.

[47] K.-Y. Chiang, N. Natarajan, A. Tewari, and I. S. Dhillon, "Exploiting longer cycles for link prediction in signed networks," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage.*, 2011, pp. 1157–1162.

[48] G. Golub and W. Kahan, "Calculating the singular values and pseudo-inverse of a matrix," *J. Soc. Ind. Appl. Mathematics Series B: Numerical Anal.*, vol. SIAM-2, no. 2, pp. 205–224, 1965.

[49] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Advances Neural Inf. Process. Syst.*, 2001, pp. 556–562.

[50] A. Mnih and R. Salakhutdinov, "Probabilistic matrix factorization," in *Proc. Advances Neural Inf. Process. Syst.*, 2007, pp. 1257–1264.

[51] E. Dong, J. Li, and Z. Xie, "Link prediction via convex nonnegative matrix factorization on multiscale blocks," *J. Appl. Mathematics*, vol. 2014, pp. 1–9, 2014.

[52] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Netw.*, vol. SN-5, no. 2, pp. 109–137, 1983.

[53] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic blockmodels," *J. Mach. Learn. Res.*, vol. 9, pp. 1981–2014, 2008.

[54] J. Zhu, "Max-margin nonparametric latent feature models for link prediction," presented at the 29th Int. Conf. Mach. Learn., Edinburgh, Scotland, 2012.

[55] M. Kim and J. Leskovec, "Modeling social networks with node attributes using the multiplicative attribute graph model," presented at the 27th Conf. Uncertainty Artificial Intell., Barcelona, Spain, 2011.

[56] S.-T. Park and W. Chu, "Pairwise preference regression for cold-start recommendation," in *Proc. 3rd ACM Conf. Recommender Syst.*, 2009, pp. 21–28.

[57] Z. Gantner, L. Drumond, C. Freudenthaler, S. Rendle, and L. Schmidt-Thieme, "Learning attribute-to-feature mappings for cold-start recommendations," in *Proc. IEEE 10th Int. Conf. Data Mining*, 2010, pp. 176–185.

[58] W. K. Leung, C. F. Chan, and F. L. Chung, "An empirical study of a cross-level association rule mining approach to cold-start recommendations," *Knowl.-Based Syst.*, vol. 21, no. 7, pp. 515–529, 2008.

[59] L. T. Weng, Y. Xu, Y. Li, and R. Nayak, "Exploiting item taxonomy for solving cold-start problem in recommendation making," in *Proc. 20th IEEE Int. Conf. Tools Artificial Intell.*, 2008, pp. 113–120.

[60] L. Martinez, L. G. Perez, and M. J. Barranco, "Incomplete preference relations to smooth out the cold-start in collaborative recommender systems," in *Proc. Annu. Meeting North Amer. Fuzzy Inf. Process. Soc.*, 2009, pp. 1–6.

[61] S. Sedhain, S. Sanner, D. Braziunas, L. Xie, and J. Christensen, "Social collaborative filtering for cold-start recommendations," in *Proc. 8th ACM Conf. Recommender Syst.*, 2014, pp. 345–348.

[62] H. Ma, H. Yang, M. R. Lyu, and I. King, "SoRec: Social recommendation using probabilistic matrix factorization," in *Proc. 17th ACM Conf. Inf. Knowl. Manage.*, 2010, pp. 931–940.

[63] M. Jamali and M. Ester, "A matrix factorization technique with trust propagation for recommendation in social networks," in *Proc. ACM Conf. Recommender Syst.*, 2010, pp. 1055–1066.

[64] P. Massa and P. Avesani, "Trust-aware recommender systems," in *Proc. ACM Conf. Recommender Syst.*, 2007, pp. 17–24.

[65] M. Kim and J. Leskovec, "Multiplicative attribute graph model of real-world networks," *Internet Mathematics*, vol. 8, no. 1/2, pp. 113–160, 2012.

[66] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," *Internet Mathematics*, vol. 6, no. 1, pp. 29–123, 2009.

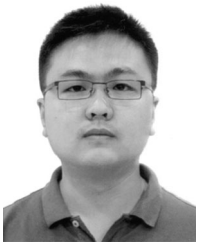
[67] J. Zhang, B. Liu, J. Tang, T. Chen, and J. Li, "Social influence locality for modeling retweeting behaviors," in *Proc. 23rd Int. Joint Conf. Artificial Intell.*, 2013, pp. 2761–2767.

[68] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. R-143, no. 1, pp. 29–36, 1982.

[69] Z. Lu, B. Savas, W. Tang, and I. S. Dhillon, "Supervised link prediction using multiple sources," in *Proc. IEEE 10th Int. Conf. Data Mining*, 2010, pp. 923–928.

[70] X. Li and H. Chen, "Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach," *Decision Support Syst.*, vol. 54, no. 2, pp. 880–890, 2013.

[71] M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki, "Link prediction using supervised learning," in *Proc. SDM Workshop Link Anal. Counter-Terrorism Secur.*, 2006.
[72] F. Pedregosa, et al., "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.



Zhiqiang Wang received the MS degree from the School of Computer and Information Technology, Shanxi University, in 2012. He is working toward the PhD degree in the School of Computer and Information Technology, Shanxi University. His research interests include data mining and social network analysis.

Jiye Liang received the PhD degree from Xi'an Jiaotong University. He is a professor in the School of Computer and Information Technology, Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University. His research interests include artificial intelligence, granular computing, data mining, and machine learning. He has published more than 80 articles in international journals.

Ru Li received the PhD degree from Shanxi University. She is a professor in the School of Computer and Information Technology, Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University. Her research interests include social media data mining, natural language processing, and information retrieval.

Yuhua Qian received the PhD degree from Shanxi University. He is a professor with the Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, China. His research interests include pattern recognition, feature selection, rough set theory, granular computing, and artificial intelligence. He has published more than 60 articles on these topics in international journals. He is a member of the IEEE.