Statistics and Its Interface Volume 15 (2022) 323–334

Local neighborhood-based approach of link prediction in networks[*]

Chunning Wang[†] and Bingyi Jing

Network structure has been widely studied in recent decades. One particular usage of the network is to represent the relationship among nodes. Therefore, link prediction plays a crucial role in network analysis. A key issue of link prediction is to estimate the likelihood of potential links between nodes in the network. However, the complex network structure makes such estimation very challenging. In this paper, we propose a link prediction method based on nodes' local neighborhood (LN), which constructs a local neighborhood for each node and calculates the likelihood of connection between nodes based on their neighbors. Further, we extend the LN method to solve the link prediction problems in a network with node covariates and community structure. Experimental studies on synthetic and real networks demonstrate that the performance of our methods is competitive.

AMS 2000 subject classifications: Primary 62-07; secondary 91D30.

Keywords and phrases: Network analysis, Link prediction, Local neighborhood, Assortative mixing, Disassortative mixing.

1. INTRODUCTION

Network has been widely employed to describe the interactions and relations between individual units in many fields, such as social networks, protein–protein interactions, gene regulatory networks, food webs, and computer networks. A network consists of nodes and edges. For example, in social networks, nodes correspond to people and edges represent friendships; in biological networks, a node may indicate a gene or a protein while edges stand for regulatory relationships. The study of networks has attracted much attention in recent years, and a large body of literature demonstrates the development of this topic [3, 8, 35, 37]. Among others, a vital problem in network research is link prediction. The aim of link prediction is to estimate the likelihood of a link between two unconnected nodes on the basis of the data of observed links (network structure) and attributes of nodes (node covariates). We refer to [32, 31, 44, 20] for a detailed description of link prediction. Generally, two types of link prediction exist: predicting missing links and predicting links that may appear in the future [42]. This paper considers the former.

Link prediction has been extensively applied to various fields. For example, in online social networks, the system recommends new friends to users to preserve their loyalties; in biological networks, checking every potential link between two nodes in protein–protein interaction and gene regulatory networks is time-consuming and costly. Instead of blindly checking all possible interactions, link prediction is able to provide specific targets for future experiments, hence sharply reducing the experiment cost while achieving accurate link prediction [32, 52]. In terrorist networks, link prediction helps determine if the particular individuals are in a same group even though their interactions are not directly observed [31].

In recent years, many novel methods have been proposed within their own application scenarios. Among them, the similarity-based method provides the simplest framework, which is determined by similarity score. A similarity score is assigned for each pair of nodes that are unconnected in the current network. All non-observed links are sorted in descending order according to their similarity scores. The higher the similarity scores, the more possible that the node pairs are connected. The similarity-based method enjoys several advantages over other approaches. For example, it is easy to implement with low computation complexity. Clearly, the similarity score plays a critical role in the method. It is an infeasible measure and unobservable, which is why the design of similarity score is crucial. A well-designed similarity score can significantly improve the effectiveness of the approach [33]. Some methods combine the attributes of nodes and structure of network to define the similarity score. However, the data of node attributes are often hard to obtain. Most existing works design the similarity score index from the network topology. The similarity indices can be classified into three categories: local similarity indices, global similarity indices, and quasi-local indices. To learn more details, readers can refer to the review literature [31, 32, 44].

Other statistical models have also been proposed. Claus et al. [11] shows that a hierarchical structure model per-

[*] The project was supported by National Natural Science Foundation of China (No. 11971214), Natural Science Foundation of Gansu Province (No. 20JR5RA204), Innovation Ability Promotion Project of Colleges and Universities in Gansu Province (No. 2020A-059) and Natural Science Foundation of the Anhui Higher Education Institutions of China (No. KJ2020A0023).

[†] Corresponding author.

forms well in predicting the missing links. However, it suffers from computational complexity. The stochastic block model (SBM) [24] assumes that nodes are in blocks and the probability of an edge between two nodes is determined by the blocks to which they belong. The main challenge in fitting the SBM model is the estimation of the blocks themselves, see [2, 4, 7, 41]. The latent space model (LSM) is introduced by Hoff et al. [23] for social networks. It models the probability of a link between nodes depending on the positions of individuals in a latent space and on their observed covariates. LSM has been extended in many directions, including treatment of transitivity, clustering, and homophily on observed attributes [18, 28]. Recently, a competing method called the popularity-scaled latent space model (PSLSM) was proposed on the basis of LSM by Chang et al. [10]. To address the degree heterogeneity in a large-scale directed social network, nodes' popularity index (PI) is considered for calculating the link probability between nodes. Unlike other latent space models, PSLSM is simple to implement and has a competitive result for directed networks. The exponential random graph model [25, 13] incorporates network statistics such as triangles, k-stars, and degrees in an exponential family. Another popular model for network analysis is the exchangeable random graph model, which is characterized by a nonnegative function f named graphon; see [51, 46, 12, 39, 16]. The link probability of two nodes is defined as $P_{ij} = f(u_i, u_j)$, where u_i and u_j are latent variables. These methods always need to make some assumptions about graphon. Probabilistic relational models [15], probabilistic entity relationship model [21], and stochastic relational model [48] are some mainstream methods of probabilistic models, which aim to extract the structure from the observed network and then predict the missing links via the learned model [32]. In addition to the network structure, the probabilistic models also require information about node attributes, thereby limiting their applications to a certain extent [40]. Despite these achievements, how to design effective and efficient algorithms is still a big challenge.

As mentioned above, most link prediction methods rely on the assumption that those similar nodes are more likely to be connected with each other. The assumption may be appropriate for networks of assortative mixing. For example, in social networks, people tend to affiliate with those who have similar background and interests with them, such as age and income level. The assumption, however, is challenged by disassortative networks, in which no evidence shows the difference between the numbers of connections among the analogous nodes and among the non-analogous nodes. For example, predators do not typically feed on each other in a food web. Recently, Zhao et al. [52] proposed a link prediction model for partially observed networks (LPPON). They assume that similar node pairs have similar link probabilities [52]. This method does not need the assumption that similar nodes are more likely to be connected with each other. Newman and Leicht [38] found that nodes can be clustered according to the similarity of the connection pattern between them, where they also provided some reasons for the importance of connection patterns in network analysis. More recently, Zhang et al. [51] proposed a novel neighborhood smoothing (NBS) to estimate the underlying probability matrix whose elements represent the possibility of links among nodes. Inspired by the works of [38, 51, 52], in this paper, we propose a new approach named local neighborhood-based score index (LN) to solve the aforementioned problems. Different from the existing approaches, this approach involves forming a group for each node that consists of its local neighbors[1], and then calculates the score between nodes on the basis of their local neighbors. The proposed method has the following advantages. First, it constructs a new mechanism of link prediction on the basis of network topology. The information of the connection pattern between nodes is fully maximized by constructing the neighborhood set. Second, it is obviously applicable to both assortative and disassortative networks. Third, the LN method improves the accuracy for link prediction nearly without increased complexity. The methodology is illustrated and the performance of the proposed prediction methods is proven by a numerical experiment based on simulation network and real network datasets in Section 4.

The rest of this paper is organized as follows: The link prediction problem and some similarity indices that will be used in the paper are stated in Section 2. The proposed prediction method (LN) is provided in Section 3. Numerical studies are presented in Section 4. The paper is concluded in Section 5.

2. PRELIMINARIES

2.1 Problem description

Considering a network $G(V, E)$ with n nodes, where V is the set of nodes and E is the set of edges. Multiple links and self loops are not considered in this paper. Let U denote the universal possible link set, which contains all $|V|(|V|-1)/2$ possible links, where $|V|$ is the number of elements in set V. Hence, U/E denotes the missing links. The purpose of link prediction is to find the possible links in the set of U/E. For each pair of nodes $(i, j) \in U/E$, a score S_{ij} is assigned to estimate the connection likelihood between nodes i and j. All unconnected node pairs are sorted in descending order according to their scores S_{ij}, and the node pairs at the top of order list are more likely to be connected than those at the bottom of the order list. In other words, a higher score indicates a higher probability of connection between the nodes, and vice versa.

2.2 Similarity-based indices

In this section, we will briefly introduce the definitions of similarity indices that will be used later.

[1] Here, we say the node A is a local neighbor of node B if they have a similar pattern of connections with other nodes in the whole network.

- **Common Neighbors Index (CN)** [14]. The CN index assumes that two nodes sharing more common neighbors are more likely to have a link. Let $\Gamma(i)$ denote the set of neighbors of node i. The CN index is defined as

$$S_{ij}^{CN} = |\Gamma(i) \cap \Gamma(j)|,$$

where $|\cdot|$ is the cardinality of a set.
- **Adamic-Adar Index (AA)** [1]. The AA index refines the CN index by assigning more weights to lower-degree nodes in the common neighbors set. It is defined as

$$S_{ij}^{AA} = \sum_{l \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log d_l},$$

where d_l denotes the degree of the node l.
- **Resource Allocation Index (RA)** [53]. The RA index is similar to the AA index. It is motivated by the resource allocation process, which takes place in the networks with complex distribution [32]. Considering a pair of nodes (i, j) that are unconnected, node i can send some resource to node j via their common neighbors, which play the role of transmitters. In the simplest case, every transmitter is assumed to be a unit of resource and will be distributed equally to all its neighbors. Consequently, the similarity between i and j may be defined as the amount of resource that j has received from i. The RA index is defined as

$$S_{ij}^{RA} = \sum_{l \in \Gamma(i) \cap \Gamma(j)} \frac{1}{d_l},$$

where d_l denotes the degree of the node l.
- **Preferential Attachment Index (PA)**. The mechanism of PA can be used to generate evolving scale-free networks, where the probability that a new link is connected to the node i is proportional to d_i [5]. Motivated by this mechanism, the PA index is defined as

$$S_{ij}^{PA} = d_i \times d_j.$$

The PA index indicates that new links are more likely to be connected with the higher-degree nodes than the lower-degree ones.
- **Jaccard Index (Jac)** [26]. The Jaccard index can be used to evaluate the similarity of the neighbor sets of the two nodes, and is defined as

$$S_{ij}^{Jaccard} = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|}.$$

- **Salton Index (Sal)** [32]. This index is also called the cosine similarity. It is defined as

$$S_{ij}^{Sal} = \frac{|\Gamma(i) \cap \Gamma(j)|}{\sqrt{d_i \times d_j}}.$$

- **Sørensen Index (Sør)** [32]. The index is used mainly for comparing the similarity between ecological community data and is defined as

$$S_{ij}^{Sørensen} = \frac{2|\Gamma(i) \cap \Gamma(j)|}{d_i + d_j}.$$

- **Hub Promoted Index (HPI)** [32]. The HPI index promotes link formation between low-degree nodes and hubs [34], defined as

$$S_{ij}^{HPI} = \frac{|\Gamma(i) \cap \Gamma(j)|}{min(d_i, d_j)}.$$

- **Hub Depressed Index (HDI)** [32]. The HDI index is analogous to the above index but has an opposite goal. It is defined as

$$S_{ij}^{HDI} = \frac{|\Gamma(i) \cap \Gamma(j)|}{max(d_i, d_j)}.$$

- **Leicht-Holme-Newman Index (LHNI)** [30]. It is defined as

$$S_{ij}^{LHN1} = \frac{|\Gamma(i) \cap \Gamma(j)|}{d_i \times d_j}$$

- **Katz Index (Katz)** [27]. The Katz index is based on the path ensemble method. It sums over all paths between nodes i and j. Its expression is

$$S_{ij}^{Katz} = \sum_{k=1}^{\infty} \beta^k \cdot |paths_{ij}^k|,$$

where β is a free parameter and $paths_{ij}^k$ is the set of all paths with length k from i to j.
- **Local Path Index (LP)** [32]. The LP index considers the information of all paths with lengths 2 and 3 between nodes i and j. It can be defined as

$$S_{ij}^{LP} = (A^2)_{ij} + \epsilon (A^3)_{ij},$$

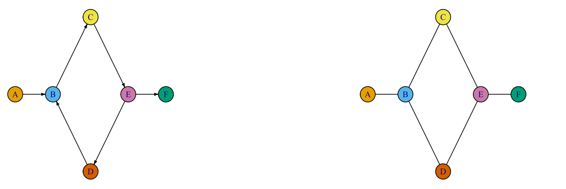
where ϵ is a free parameter and A is the adjacency matrix of the network.

3. METHODOLOGY

Consider a network with n vertices. Let A be the adjacency matrix of the network. The relationship between two nodes in the network can be represented by a $n \times n$ adjacency matrix $A = (A_{ij})_{n \times n}$, where

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge from } i \text{ to } j, \\ 0 & \text{otherwise.} \end{cases}$$

Adjacency matrix representation provides a concise mathematical structure of the topology of networks. Each row $A_{i \cdot}$ of the adjacency matrix represents the link relationship



(a) directed network

(b) undirected network

Figure 1. Directed network (a) and undirected network (b).

from node i to the other nodes. In this paper, we shall consider the link prediction problem for directed and undirected networks, as displayed in Figure 1. Therefore, the adjacency matrix A can be either symmetric (for undirected networks) or asymmetric (for directed networks).

Our approach has two key steps. The first step is to construct a local neighborhood N_i for the node $i, i = 1, ..., n$. Second, each pair of nodes, i and j, is assigned a score S_{ij} based on their local neighborhoods. The details of the method are given in the following sections.

3.1 Construction of local neighborhood

We define a matrix $C = (C_{ij})_{n\times n}$ to measure the level of distinction of connection patterns, where C_{ij} denotes the level of the distinction of the connection pattern between node i and j. Here, we demonstrate the level of distinction of connection patterns via the inverse proportion of the similarity of the connection patterns between the same nodes. Precisely, we define the C_{ij} as

$$C_{ij} = 1/W_{ij},$$

where W_{ij} denotes the similarity of the connection pattern between nodes i and j. That is, W_{ij} would be large if nodes i and j have a similar pattern of connections with other nodes.

Various ways can be used to define the W_{ij}. A simple choice is

$$W_{ij} := |\{m : A_{im} = A_{jm}\}|, \tag{1}$$

where $|\cdot|$ denotes the cardinality of a set. However, this measure is not very efficient because many networks are sparse, and most elements of $A_{i\cdot}$ may be 0, which means that most of W_{ij} in Equation (1) would be large. To avoid the effect of network sparsity, a choice of W_{ij} is the Jaccard index [26]. For an undirected network, the Jaccard index is defined as

$$W_{ij} = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|}, \tag{2}$$

where $\Gamma(i) = \{l : A_{il} = 1\}$ denotes the set of nodes that are connected with the node i. Therefore, a larger Jaccard index implies that two nodes tend to have more common connections, which means they have a similar patterns of connections with other nodes. For directed networks, we consider the "in" links and "out" links separately in the definition of (2)

$$W_{ij} = \frac{|\Gamma_r(i) \cap \Gamma_r(j)|}{|\Gamma_r(i) \cup \Gamma_r(j)|} + \frac{|\Gamma_c(i) \cap \Gamma_c(j)|}{|\Gamma_c(i) \cup \Gamma_c(j)|}, \tag{3}$$

where $\Gamma_r(i) = \{l : A_{il} = 1\}$ and $\Gamma_c(i) = \{l : A_{li} = 1\}$.

Now we are ready to construct a local neighborhood N_i of the ith node from the matrix C. By the definition of C, the local neighborhood N_i should be a collection of the nodes such that $C_{i\cdot}$s are small (under some specified threshold of tolerance). Here, we adopt the method proposed by [51] to construct the local neighborhood N_i, which is defined as:

$$N_i = \{k : C_{ik} < Q_i(\alpha), k \neq i\},$$

where $Q_i(\alpha)$ is the αth quantile of the set $\{C_{ik}, k \neq i\}$, with α serving as a tuning parameter. Quantile is chosen as the threshold because it is a data-driving choice (soft threshold) and shows significant advantage in the stability of performance compared with an absolute (hard) threshold, and we choose $\alpha = (\log n/n)^{\frac{1}{2}}$ as in [51], where n is the number of nodes in a network.

3.2 Calculating the score between nodes

Now, we calculate the scores between nodes. On the basis of the set of the local neighborhood N_i, we define the score of the node pair (i, j), S_{ij}^{LN}, as the proportion of nodes in N_i that are connected to the node j. Specifically, for an undirected network, we define S_{ij}^{LN} by

$$S_{ij}^{LN} = \frac{1}{|N_i|}\sum_{k\in N_i} A_{kj} + \frac{1}{|N_j|}\sum_{k\in N_j} A_{ki}, \tag{4}$$

and for a directed network, it is defined by

$$S_{ij}^{LN} = \frac{1}{|N_i|}\sum_{k\in N_i} A_{kj}. \tag{5}$$

After the scores of all unconnected node pairs are calculated using the LN method, these node pairs are sorted in descending order of the scores. The higher the ranking, the more likely the node pairs are to be connected.

A detail needs to be pointed out that although the LN method still follows the principle that a larger S_{ij}^{LN} corresponds to a higher likelihood that the nodes i, j are to be connected, the calculation principle of S_{ij}^{LN} is completely different from the calculation based on similar score indexes. As mentioned before, in assortative networks, the assumption that similar nodes are more likely form links is valid. One just needs to define an appropriate similarity score that effectively measures the degree of similarity between nodes. Classical indexes, such as AA, RA, and Katz, perform well in link prediction. However, the above assumption is no longer valid in disassortative networks. At this time, the similarity indexes, especially defined by CN and its variations, are no longer valid because they assume that the more common neighbors between nodes, the more likely nodes are to be connected. However, in disassortative networks such as food webs, although predators have many common prey objects, they obviously do not feed on each other. To address this issue, we propose the LN method. In this method, we collect those nodes that have a similar pattern of connections with node i to construct its neighborhood N_i. After the neighborhood N_i is constructed, a natural idea to predict the possibility that node i is connected to j is to compute the proportion of nodes in N_i that connects to j. After all, nodes in N_i and node i have a similar pattern of connections with other nodes. If more nodes in N_i are connected to j, then the probability that node i and j are connected is also greater, which indicates a larger S_{ij}^{LN}. From the data analysis results in Section 4, especially in the food webs, the performance of the LN method is relatively better.

Remark 3.1. *We may also consider other designs for W. For instance, one may use*

$$W_{ij} = \frac{|\Gamma_r(i) \cap \Gamma_r(j)| + |\Gamma_c(i) \cap \Gamma_c(j)|}{|\Gamma_r(i) \cup \Gamma_r(j) \cup \Gamma_c(i) \cup \Gamma_c(j)|}.$$

We can also consider a two-dimension similarity index (W_{ij}, W'_{ij}) as follows:

$$W_{ij} = \frac{|\Gamma_r(i) \cap \Gamma_r(j)|}{|\Gamma_r(i) \cup \Gamma_r(j)|},\ W'_{ij} = \frac{|\Gamma_c(i) \cap \Gamma_c(j)|}{|\Gamma_c(i) \cup \Gamma_c(j)|}.$$

We design W using Equation (2) for undirected networks and Equation (3) for directed networks in this paper.

Remark 3.2. *For the definition of W_{ij} in Equations (2) and (3), W_{ij} may be 0, especially in a sparse network. If W_{ij} equals 0, then C_{ij} is defined as ∞. According to the definition of neighborhood of node i, if $C_{ij} = \infty$, the node j will be excluded from the neighborhood of node i. In other words, we collect only those nodes with a small value of C_{ij} to construct the neighborhood of node i.*

Remark 3.3. *If the node covariates are observed and related to the structure of network, then we can define C_{ij} as $\frac{1}{W_{ij}+\lambda F_{ij}}$, where F_{ij} measures the similarity between nodes covariates and λ is a parameter to tune the weights of the network structure information and node covariates information. In this case, C_{ij} represents the distinction between nodes i, j, which is measured by the connection pattern and the node covariates. We use LNC to denote the extension method of the LN after adding node covariates. The effectiveness of the extension method is verified by simulation data and a real dataset in Section 4.*

Remark 3.4. *Many empirical networks display an inherent tendency to cluster. In some cases, compared with the members of the inter-community, the members of the intra-community should have more similar patterns of connection to other nodes [38]. If the network has community structures, then our method can also be extended by adding appropriate community information to W. Modularity proposed by Newman [36] is used to evaluate the qualities of detected communities. They defined a modularity matrix M with elements $M_{ij} = A_{ij} - \frac{d_i d_j}{2M}$, where M denotes the total number of edges in the network. If node i and j are in the same community, then the contribution of the node pair to the modularity is M_{ij}. According to [9], we call M_{ij} the modularity contribution of nodes i and j. The value of modularity contribution M_{ij} measures the difference between the number of real edges and the number of expected edges between nods i and j. A larger M_{ij} means the nodes have a higher chance of being in the same community [9]. To avoid the accuracy of link prediction results being affected by the quality of community detection, we choose M_{ij} as the information of the network community and add it to W_{ij} after normalizing. In a simple case, C_{ij} can be defined as $\frac{1}{W_{ij}+M_{ij}}$. Let LNM denotes the extension method of the LN after adding the modularity contribution. We will test the performance of the LNM on simulation networks and apply it for link prediction on four real datasets in Section 4.*

4. NUMERICAL EVALUATION

In this section, we first introduce the datasets and evaluation metrics. Then we experimentally evaluate the performance of our proposed method on synthetic networks and real networks. In each network, the parameter of the LN method is set as $\alpha = (\log n/n)^{\frac{1}{2}}$, where n denotes the total number of nodes in the network.

4.1 Datasets

The following well-known network datasets are used in our analysis:

- **Zachary karate club network dataset (Karate)**: The dataset contains social ties among the members of a university karate club collected by Wayne Zachary in 1977, which is available on http://networkrepository.com/soc-karate.php;
- **Jazz musicians network dataset (Jazz)**: It is a collaboration network between jazz musicians, available on http://konect.cc/networks/arenas-jazz/;
- **US Air Transportation Network (USAir)**: This dataset contains 332 US airports with the largest amount of traffic from publicly available data and is available on http://vlado.fmf.uni-lj.si/pub/networks/data/;
- **Yeast**: This is a biological dataset of a protein–protein interaction network, available on http://vlado.fmf.uni-lj.si/pub/networks/data/;
- **C. elegans**: This is a neural network of the nematode worm C. elegans, available on http://www-personal.umich.edu/~mejn/netdata/;

- **FWFW**: This is an Ecology Network dataset that describes the food web of the Florida ecosystem and is available on http://vlado.fmf.uni-lj.si/pub/networks/data/;
- **StMarks**: This is an Ecology Network dataset that describes the food-webs and is available on http://vlado.fmf.uni-lj.si/pub/networks/data/;
- **Everglades**: This is a network of food web in Everglades Graminoids during the wet season, available on http://vlado.fmf.uni-lj.si/pub/networks/data/.
- **Lawyer**: This is a network of corporate law partnership that was created in a Northeastern US corporate law firm and is available on http://moreno.ss.uci.edu/data.html.

The nine real datasets we consider are from different fields, including social networks, transportation networks, biological networks, and ecology networks. Furthermore, among the datasets, four datasets are undirected networks (Karate, Jazz, USAir, and Yeast) and the others are directed networks. For the undirected networks, Karate [49] is a social network of friendship collected by Zachary at a US university in the 1970s; Jazz [17] is a collaboration network that consists of jazz bands, in which each node is a jazz musician and an edge denotes that two musicians have played together in a band; USAir [6] is a US air transportation system network; Yeast [43] is a protein–protein interaction network in budding yeast that has large components, containing 2,375 nodes and 11,693 interactions. For the directed networks, C. elegans [45] is a neural network of the nematode worm C. elegans; FWFW, StMarks, and Everglades are all the food web networks, where FWFW is the food web of the Florida ecosystem, and StMarks and Everglades are the food webs in St. Marks River Flow and Everglades Graminoids, respectively. The Lawyer [29] dataset represents the friendship among 69 lawyers after 2 isolated nodes are removed. Each node has seven attributes.

The graphic statistics of these network datasets are summarized in Table 1. In the table, V and E refer to the total numbers of nodes and edges in the network, respectively. AD denotes the value of average degree of the network, and C is the clustering coefficient.

Table 1. Summarized statistics of the nine networks

Network Type	Datasets	V	E	AD	C
undirected	Karate	34	78	4.588	0.588
	Jazz	198	2742	27.697	0.633
	USAir	332	2126	12.807	0.749
	Yeast	2375	11693	9.847	0.388
directed	C. elegans	297	2345	7.896	0.174
	FWFW	128	2106	16.453	0.177
	StMarks	54	353	5.116	0.210
	Everglades	69	911	13.203	0.303
with covariates	Lawyer	69	575	8.33	0.407

4.2 Evaluation metrics

Recall that U denotes the set of all possible links and E the set of all observed links. We randomly divide the observed links E into two sets: the training set E_t, and the probe set E_p. Following the common practice, the training set E_t contains 90% elements of E, and the remaining 10% elements are left to form the probe set. That is, we randomly delete 10% links, from the whole link set E. This yields a new network, namely, the training set E_t, from which the scores of those deleted links (or node pairs) and the other possible links are computed according to formulas (4) and (5).

To quantify the accuracy of prediction methods, we consider two standard metrics that are commonly used for evaluating the performance of link prediction methods: precision [22] and the area under the receiver operating characteristic curve (AUC) [19]. To implement, the basic preparation for the two metrics is the same [47]. Both metrics are based on the information of the same training and probe sets. We state the precision first. According to our prediction rule, we predict that two nodes are connected if the score is in the top of the score lists. Therefore, on the basis of the scores calculated from the training set, suppose that the top L links are selected and thus they are predicted to be connected. Furthermore, assume that L_r links in the top list are also the elements of the probe set, that is, they are predicted correctly. The precision is defined as the ratio of the true positive and total positive, that is,

$$precision = \frac{L_r}{L}.$$

In the numerical experiment, we set $L = 20$ for networks with fewer than 1,000 links and $L = 100$ for networks with more than 1,000 links.

Secondly, we consider the AUC. AUC evaluates the performance of link prediction methods according to the rank list. Given the rank of all non-observed links, it can be interpreted as the probability that a randomly chosen missing link in E_p is given a higher score than a randomly chosen nonexistent link in $U - E$. If among n independent comparisons, there are n' times the missing link having a high similarity score and n'' times are the same, then the calculation of AUC can be written as follows:

$$AUC = \frac{n' + 0.5n''}{n}.$$

A high AUC value corresponds to a better the prediction result.

4.3 Simulation studies

In this section, we compare the performance of LN and other methods on simulated networks, including the classic indices introduced in Section 2 and recently proposed methods, such as PI [10], LPPON [52], and NBS [51] in simulation 1. Further, we compare the performance of LN with that of its two extension methods LNM and LNC in simulation 2.

4.3.1 Simulation 1

The first simulation investigates the performance of all methods under different types of network structure. We generate a stochastic block model with three clusters, where each cluster has 100 nodes. We first consider the assortative network. In this case, we define

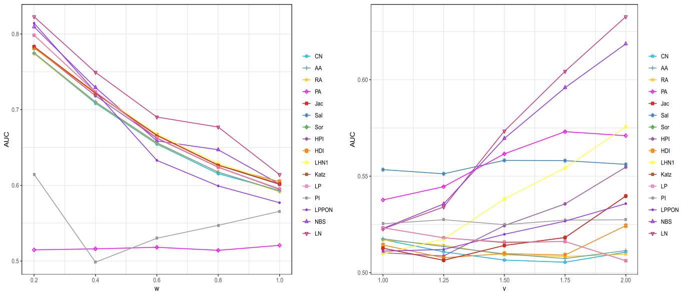
$$B_1 = \begin{pmatrix} 0.3 & 0.12w & 0.08w \\ 0.12w & 0.3 & 0.08w \\ 0.08w & 0.08w & 0.25 \end{pmatrix}.$$

The within-cluster link probabilities between nodes i and j are 0.3, 0.3, and 0.25. The link probability is $0.12w$ if nodes i and j come from clusters 1 and 2 and $0.08w$ for other cases, in which $w \in [0.2, 1]$. A small w corresponds the stronger assortativity of the network. In the disassortative network, we generate a directed network. In this case, we set

$$B_2 = \begin{pmatrix} 0.1 & 0.2v & 0.18v \\ 0.12v & 0.1 & 0.2v \\ 0.12v & 0.12v & 0.09 \end{pmatrix}.$$

If nodes i and j in the same cluster, then the link probability is 0.1, 0.1, and 0.09. The link probability from cluster 1 to 2 or 2 to 3 is $0.2v$, the link probability from cluster 1 to 3 is $0.18v$, and $0.12v$ for other cases, in which $v \in [1, 2]$. The larger v means that the disassortative of the network is stronger.

Figure 2 shows the mean of AUC for various methods as the degree of network assortativity (disassortativity) changes. Figure 2(a) shows that, with the increase in w, the assortativity of the network becomes weaker, and the AUC values of most methods decrease. However, the AUC value of the LN method is always higher than the others. Figure 2(b) shows a similar conclusion. As v increases, the AUC value of LN also increases, while the AUC values of other methods do not changed much. Overall, the simulation results show that the LN method can deal with both assortative and disassortative networks effectively.



(a) assortative network (b) disassortative network

Figure 2. Mean of AUC vs. the degree of network assortativity (a) and disassortativity (b). Each result is obtained by averaging 50 realizations with a probe set containing 10% random links.

4.3.2 Simulation 2

In the second simulation, we compare the performances of LN, LNM, and LNC. We generate a network with node covariates. Each network contains 300 nodes, and node i's covariates X_i are generated from a Gaussian mixture distribution with three components. Specifically, we first generate independent three centers $\mu_l \sim N(0, \eta^2 I_p), l = 1, 2, 3, p = 5$. Then, for each μ_l, we generate 100 nodes, with $X_i \sim N(\mu_l, I_p)$. Finally, each A_{ij} is generated from Bernoulli distribution, that is,

$$A_{ij} \sim Bernoulli(p_{ij}),$$

where p_{ij} is obtained from the distance model in [23],

$$logit(p_{ij}) = -||X_i - X_j||,$$

where $|| \cdot ||$ denotes the Euclidean norm. The covariates X_i of node i can be regarded as the position of the node in the "social space" [52]. The above model can generate both directed and undirected networks. For undirected networks, we just set $A_{ij} = A_{ji}$.

The node covariates are continuous variables and affect the probabilities of links. Thus, we define the similarity of node covariates by Gaussian similarity function $F_{ij} = exp\{-||X_i - X_j||^2/(2\sigma^2)\}$ with $\sigma = 1$. To evaluate the performance of the LN method and its extension methods, we randomly delete 10% edges in each simulated network as the test set and report the mean of AUC based on 50 repetitions. The results are shown in Table 2. As we know, μ_l will become more separated as η become larger, which means the nodes become more "clustered". Table 2 shows that the performance of each method improved as the nodes became more "clustered". According to the simulation, the connection probabilities between nodes are directly affected by the node covariates. As expected, from the second and third columns of Table 2, we can see that the LNC method, which uses both network topology and node covariates, performs better than the LN method, which uses only network topology. From the second and fourth columns of Table 2, we can derive the same conclusion. In summary, the performance of the LN can be slightly improved as node covariates or community information is included.

Table 2. Mean of AUC for simulated network based on 50 replications. For LNC, the parameter $\lambda = 1$

η^2	C_{ij}		
	only using W_{ij} (LN)	using W_{ij} and F_{ij} (LNC)	using W_{ij} and M_{ij} (LNM)
1	0.6525	0.7506	0.6579
4	0.7133	0.7958	0.7203
9	0.7515	0.8171	0.7607
16	0.8165	0.8816	0.8275

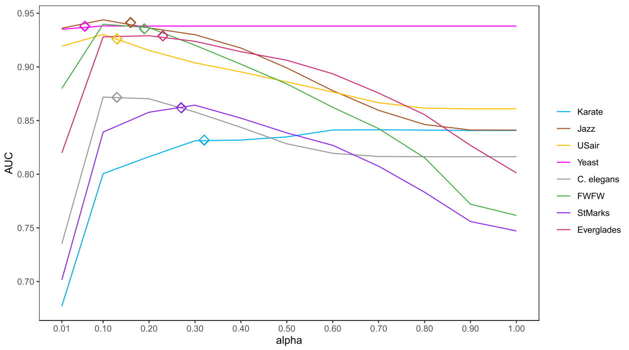


Figure 3. Prediction accuracy measured by AUC as a function of α. The eight graph lines show the average AUC values for different networks with α changing from 0.01 to 1. The values of $\diamond$ correspond to the α values obtained using $\sqrt{\log n/n}$.

Table 3. Similarity indices for directed networks

Index	S_{ij} for directed netwoerk
CN	$\lvert\Gamma_r(i) \cap \Gamma_c(j)\rvert$
AA	$\sum\limits_{k\in\Gamma_r(i)\cap\Gamma_c(j)} \frac{1}{\log d_k^r}$
RA	$\sum\limits_{k\in\Gamma_r(i)\cap\Gamma_c(j)} \frac{1}{d_k^r}$
PA	$d_i^r \times d_j^c$
Jaccard	$\frac{\lvert\Gamma_r(i)\cap\Gamma_c(j)\rvert}{\lvert\Gamma_r(i)\cup\Gamma_c(j)\rvert}$
Salton	$\frac{\lvert\Gamma_r(i)\cap\Gamma_c(j)\rvert}{\sqrt{d_i^r \times d^{c_j}}}$
Sørensen	$\frac{2\lvert\Gamma_r(i)\cap\Gamma_c(j)\rvert}{d_i^r + d_j^c}$
HPI	$\frac{\lvert\Gamma_r(i)\cap\Gamma_c(j)\rvert}{min(d_i^r, d_j^c)}$
HDI	$\frac{\lvert\Gamma_r(i)\cap\Gamma_c(j)\rvert}{max(d_i^r, d_j^c)}$
LNHI	$\frac{\lvert\Gamma_r(i)\cap\Gamma_c(j)\rvert}{d_i^r \times d_j^c}$

4.4 Real-world dataset

4.4.1 Selection of the parameter α

The size of node i's local neighborhood N_i is controlled by adjusting the parameter α. An appropriate α must be chosen. In real dataset analysis, we still use $\alpha = \sqrt{\log n/n}$ proposed by [51], where n is the number of nodes in a network. Further, by discussing the influence of the change of parameter α on the empirical results, we demonstrate the effectiveness of this parameter selection method proposed by [51].

In eight real networks, we measure the performance of our method by AUC as α varies. For each dataset, the results are averaged over 50 replications, as presented in Figure 3. Let α^* denote the corresponding α value obtained by $\sqrt{\log n/n}$. From Figure 3, we can find that for most datasets, the AUC values returned by LN method at α^* ($\diamond$ in Figure 3) are the highest. In the Karate, Jazz, and FWFW datasets, the highest AUC values of LN corresponding to α values are near α^*, not far from α^*. In general, LN performs well when α is chosen as $\sqrt{\log n/n}$.

4.4.2 Comparison of LN with other methods

We first examine the performances of the 16 methods, which are the same as the one in simulation 1.

For directed networks, the local similarity indices are defined as in [50]; see Table 3. Here, $\Gamma_r(i) = \{l : A_{il} = 1\}$, $\Gamma_c(i) = \{l : A_{li} = 1\}$, $d_i^r = |\Gamma_r(i)|$ and $d_i^c = |\Gamma_c(i)|$.

Table 4 compares the prediction accuracy quantified by precision. The highest precision among the 16 methods in each dataset is highlighted in boldface. From Table 4, we see that our proposed method performs the best in four network datasets: Karate, C. elegans, FWFW, and StMarks, while the AA method works the best in the Jazz dataset, the RA shows the best performance in the USAir dataset, LPPON shows the best result in the Yeast dataset, and NBS performs the best for the Everglades dataset. Moreover, the LN method shows a strong ability to deal with the directed networks, performing the best on three of the four directed networks: C. elegans, FWFW, and StMarks. The PI index does not perform well in undirected networks, probably because the index itself is designed for directed networks. For the networks in which the LN method does not perform the best, it still shows satisfactory performance in most of cases. We derive a similar conclusion from Table 5; clearly the LN method is more preferred. Under the AUC metric, the LN method again performs the best in the Karate, C. elegans, FWFW, and StMarks datasets, as well as in the Everglades dataset. For the other three networks, the RA method shows the best performance on the Jazz and USAir datasets, and LPPON still shows the best performance on the Yeast dataset. Therefore, our proposed LN method appears to be the best one among the 16 prediction methods.

In addition, we delete 10% of edges at random and calculate the receiver operating characteristic (ROC) curve for predicting the missing links on four networks, as shown in Figure 4. Again, the LN method performs better than the other methods.

In Figure 5, we display the complete graphs of the values of precision against the threshold number L, which investigates the dependence pattern between the precision and threshold number L, for all eight network datasets. In the graphs, the precision curves are plotted against 10 different values of L, where L are for 2 (10) to 20 (100) with step 2 (10), for the network dataset with links fewer (more) than 1,000, respectively. From Figure 5, we see that for most datasets, the precision curve decreases when the threshold number L increases, with the Yeast network as the only exception, in which the precision curve changes to be relatively flat. This finding demonstrates that the LN method achieves the highest precision in most networks, uniformly

Table 4. Mean of precision for real networks. Each result is obtained by averaging 100 realizations with a probe set containing 10% random links. Parameters in Katz and LP are tuned to their optimal values subject to maximal precision. The highest values are emphasized in bold

method	Karate	Jazz	USAir	Yeast	C. elegans	FWFW	StMarks	Everglades
	precision							
CN	0.0965	0.8136	0.5810	0.6811	0.0710	0.0883	0.1325	0.2665
AA	0.0930	**0.8318**	0.6034	0.7013	0.0613	0.1166	0.1125	0.4080
RA	0.0980	0.8142	**0.6246**	0.4910	0.0707	0.1252	0.1345	0.4812
PA	0.0615	0.1936	0.4577	0.5021	0.0354	0.1791	0.2115	0.3705
Jac	0.0105	0.7526	0.0206	0.0050	0.0090	0.0162	0.0310	0.0721
Sal	0.0065	0.7650	0.0195	0.0050	0.0121	0.0299	0.0605	0.1250
Sør	0.0105	0.7526	0.0206	0.0052	0.0097	0.0162	0.0311	0.0720
HPI	0.0435	0.0456	0.0021	0.0170	0.0715	0.1212	0.0800	0.0050
HDI	0.0190	0.7058	0.0256	0.0050	0.0090	0.0119	0.0020	0.0605
LHNI	0.0055	0.0670	0.0062	0.0012	0.0119	0.0052	0.0053	0.0005
Katz	0.0933	0.8046	0.5904	0.6554	0.0661	0.0882	0.1027	0.0205
LP	0.0931	0.7786	0.5860	0.6670	0.0662	0.0881	0.1330	0.2575
PI	0.0195	0.0000	0.0055	0.0000	0.0344	0.0892	0.1915	0.4093
LPPON	0.1005	0.5482	0.2951	**0.8380**	0.0055	0.2684	0.1120	0.5310
NBS	0.1125	0.7692	0.5797	0.6033	0.1371	0.3559	0.3683	**0.6275**
LN	**0.1325**	0.8038	0.6157	0.5316	**0.1603**	**0.4461**	**0.3982**	0.5585

Table 5. Mean of AUC for real networks. Each result is obtained by averaging 100 realizations with a probe set containing 10% random links. Parameters in Katz and LP are tuned to their optimal values subject to maximal AUC. The highest values are emphasized in bold

method	Karate	Jazz	USAir	Yeast	C. elegans	FWFW	StMarks	Everglades
	AUC							
CN	0.7080	0.9549	0.9338	0.8967	0.7847	0.7306	0.7024	0.7555
AA	0.7403	0.9619	0.9449	0.8972	0.7899	0.7346	0.6990	0.7637
RA	0.7487	**0.9707**	**0.9505**	0.8974	0.7907	0.7381	0.6985	0.7622
PA	0.7238	0.7682	0.8869	0.8276	0.7836	0.8366	0.7973	0.8798
Jaccard	0.6049	0.9611	0.8990	0.8905	0.7571	0.6072	0.6353	0.7032
Salton	0.6406	0.9671	0.9250	0.9102	0.6048	0.7169	0.6759	0.7645
Sørense	0.6049	0.9611	0.8990	0.8905	0.7742	0.7013	0.6885	0.6808
HPI	0.7165	0.9493	0.8840	0.9091	0.7845	0.6998	0.7059	0.7119
HDI	0.5907	0.9519	0.8926	0.8904	0.7726	0.7031	0.6863	0.6721
LHNI	0.5996	0.9030	0.7744	0.9059	0.7697	0.6692	0.6833	0.5984
Katz	0.7403	0.9512	0.9245	0.9215	0.8139	0.6477	0.6592	0.7852
LP	0.7481	0.9475	0.9272	0.9415	0.8136	0.6429	0.6527	0.7827
PI	0.5703	0.5475	0.6693	0.8495	0.7682	0.7385	0.7030	0.7502
LPPON	0.8212	0.8888	0.8979	**0.9645**	0.7641	0.8101	0.7111	0.7396
NBS	0.8316	0.9337	0.9371	0.9095	0.8147	0.9093	0.8446	0.9226
LN	**0.8318**	0.9413	0.9247	0.9386	**0.8777**	**0.9355**	**0.8619**	**0.9292**

in the threshold number L. We also observe that the methods based on classical similarity indices, such as CN and AA, do not perform well, particularly in the directed networks, whereas NBS shows better performances on some datasets.

We also discuss the robustness of LN against other methods by varying the ration of probe sets from 10% to 50%. The results are shown in Figure 6. As the information of known edges decreases, LN achieves higher accuracy compared with other methods, which suggests that LN has a reasonable robustness.

To summarize, in view of the results based on two metrics of precision and AUC, the LN method shows the overall best performance, and it works well particularly in the food web datasets. The main reason is that the LN approach employs the nodes' neighbors to calculate the scores (those neighbors always have similar connection patterns with the node), and provide useful information about network structure in the prediction. In addition, unlike the classical similarity score indices, the LN captures useful connection structure information on networks. Finally, the LN method can handle both assortative and disassortative networks.

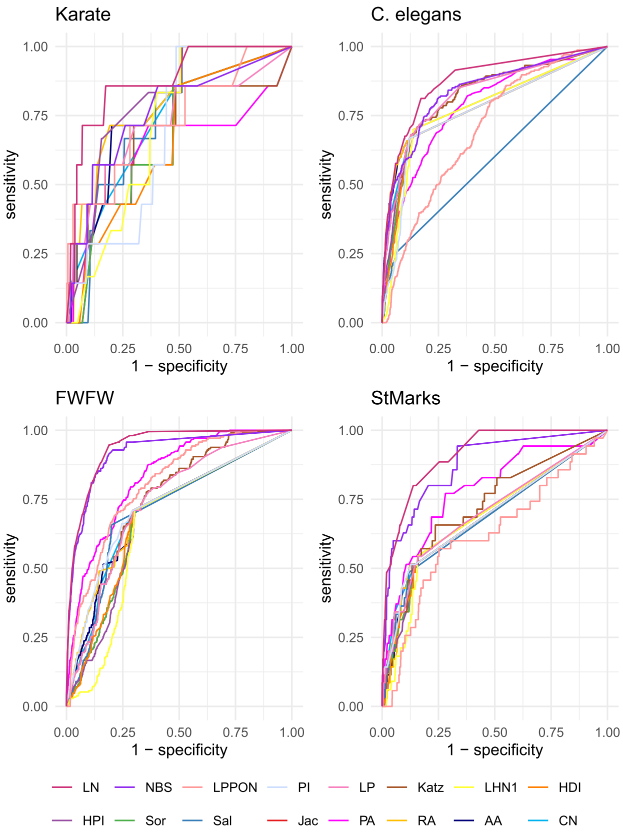


Figure 4. ROC curves for link prediction on four networks; 10% of edges are missing at random.

4.4.3 Comparison of LN with its extension methods

On the basis of the evaluation criteria of AUC, we first compare the predictive performance of LNC and LN on the Lawyer network. The Lawyer friendship network contains 69 nodes after 2 isolated nodes are removed. Each node has seven attributes describing their personal information, namely, formal status (fs), office location (ol), practice (p), gender (g), law school (ls), age, and years with the firm (yf). The first five attributes are categorical variables, and the remaining attributes are continuous variables. The continuous attribute variables are transformed into categorical attribute variables to facilitate the calculation of attribute similarity. Specifically, if the age is less than 40, then reset the attribute of age as 0; otherwise, it is set as 1. A similar reset is made for the year of firm, with the threshold selected as 7. Then, the similarity between nodes covariates F_{ij} is defined as

$$F_{ij} = (I_{\{fs_i=fs_j\}} + I_{\{ol_i=ol_j\}} + I_{\{p_i=p_j\}} + I_{\{g_i=g_j\}} + I_{\{ls_i=ls_j\}} + I_{\{age_i=age_j\}} + I_{\{yf_i=yf_j\}})/7,$$

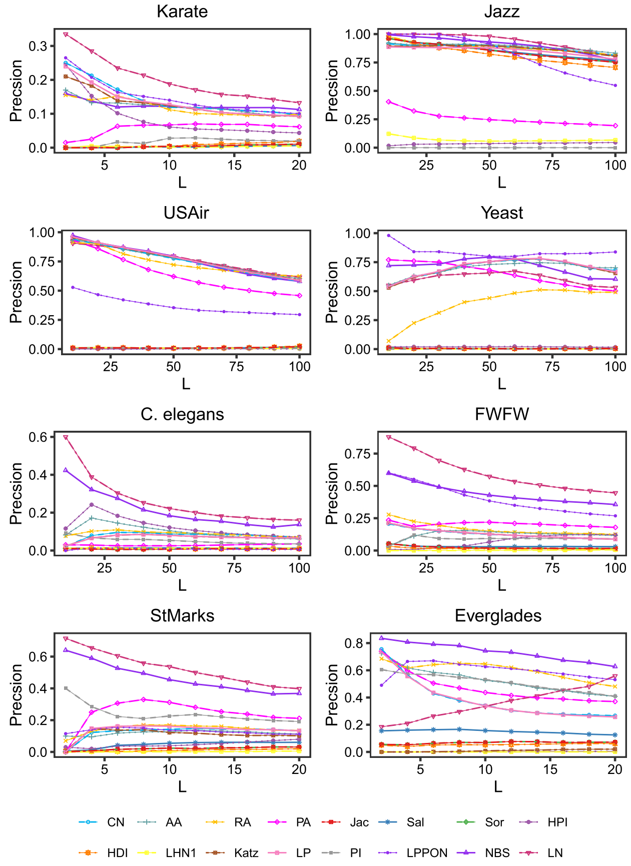


Figure 5. Precision as a function of L. For each L, the result is obtained by averaging over 50 independent realizations with a probe set containing 10% random links.

Table 6. Mean of AUC for the Lawyer friend network based on 50 replications. Parameters in LNC are tuned to their optimal values subject to maximal AUC

data set	C_{ij}	
	using W_{ij} (LN)	using W_{ij} and F_{ij} (LNC)
Lawyer	0.8273	0.8349

and the numerical results are shown in Table 6. In this experiment, we choose the parameter $\lambda = 0.5$ subject to maximal AUC. As we can see, the performance of the LNC method is slightly better than that of the LN method.

We also explore the performances of LN and LNM on four undirected networks; the results are shown in Table 7. After the information of modularity is added, the performance of LNM is slightly better than that of LN on three networks. However, the opposite results were found on the Karate network, possibly because simply replacing community information with modularity information is not appropriate for Karate data.

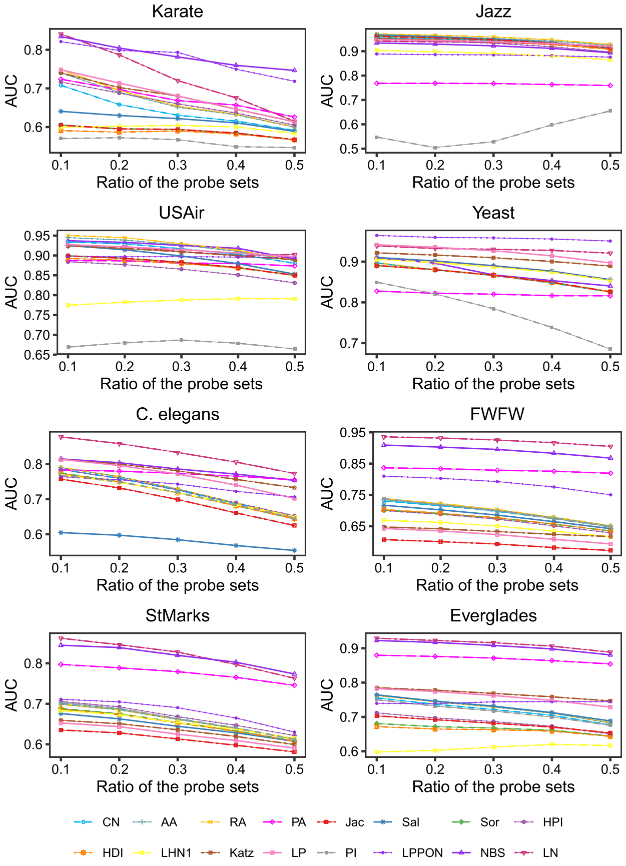


Figure 6. Value of AUC vs. the ratio of the probe sets on all datasets. For each ratio, the result is obtained by averaging over 50 independent realizations with probe set containing 10% random links.

Table 7. Mean of AUC for real network based on 50 replications

data set	C_{ij}	
	using W_{ij} (LN)	using W_{ij} and M_{ij} (LNM)
Karate	0.8318	0.8240
Jazz	0.9413	0.9546
UsAir	0.9247	0.9519
Yeast	0.9386	0.9427

5. CONCLUSIONS

In this paper, we propose the LN method for link prediction by constructing nodes' local neighborhoods. First, we construct a local neighborhood N_i for each node i, in which node i and its neighbors have a similar pattern of connections with other nodes in the network. Second, the score between nodes i, j is defined as the proportion of nodes in N_i that are connected to the node j. By calculating the scores in this way, our method captures useful connection structure information and is not limited to the assumption of classical methods that similar nodes are more likely to be connected with each other. As a result, it can effectively handle both assortative mixing and disassortative mixing networks. The performance of the proposed method is demonstrated with eight real network datasets, including four undirected networks and four directed networks. Experiment results show that our method either outperforms or works comparatively with the other existing methods. In addition, we extend the LN method to solve the link prediction problems in a network with node covariates and community structure. Experimental studies on synthetic and real networks show that including additional useful information can improve the performance of the LN method.

ACKNOWLEDGEMENTS

The authors would like to thank an associate editor and anonymous referees for their careful reading and helpful comments.

Received 18 August 2020

REFERENCES

[1] Adamic, L. A. and Adar, E. (2003). Friends and Neighbors on the Web. *Social Networks* **25** 211–230.

[2] Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. C. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* **9** 1981–2014.

[3] Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Review of Modern Physics* **74** 47–97. MR1895096

[4] Amini, A. A., Chen, A., Bickel, P. J., and Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics* **41** 2097–2122. MR3127859

[5] Barabasi, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512. MR2091634

[6] Batageli, V. and Mrvar, A. Pajek datasets. http://vlado.fmf.uni-lj.si/pub/networks/data.

[7] Bickel, P. J. and Chen, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences* **106** 21068–21073.

[8] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics Reports* **424** 175–308. MR2193621

[9] Caiyan, D., Chen, L., and Li, B. (2017). Link prediction in complex network based on modularity. *Soft Computing* **21** 4197–4214.

[10] Chang, X., Huang, D., and Wang, H. (2019). A Popularity Scaled Latent Space Model for Large-Scale Directed Social Network. *Statistica Sinica* **29** 1277–1299. MR3932518

[11] Clauset, A., Moore, Cristopher, and Newman, M. E. J. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature* **453** 98–101.

[12] David, C. (2017). Co-clustering of nonsmooth graphons. *The Annals of Statstics* **45** 1488–1515. MR3670186

[13] Frank, O. and Strauss, D. (1986). Markov Graphs. *Journal of the American Statistical Association* **81** 832–842. MR0860518

[14] Franois, L. and White, H. C. (1971). Structural equivalence of individuals in networks. *The Journal of Mathematical Sociology* **1** 49–80.

[15] Friedman, N., Getoor, L., Koller, D., and Pfeffer, A. (2000). *Learning probabilistic relational models*. Springer, Berlin.

[16] Gao, C., Lu, Y., and Zhou, H. H. (2015). Rate-optimal graphon estimation. *The Annals of Statistics* **43** 2624–2652. MR3405606

[17] Gleiser, P. M. and Danon, L. (2003). COMMUNITY STRUCTURE IN JAZZ. *Advances in Complex Systems* **6** 565–573.

[18] Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society* **170** 301–354. MR2364300

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

[20] Hasan, M. A. and Zaki, M. J. (2011). *A Survey of Link Prediction in Social Networks*. Springer, US. MR3014051

[21] Heckerman, D., Meek, C., and Koller, D. (2004). Probabilistic Entity-Relationship Models, PRMs, and Plate Models. In *:Proceedings of the 21st International Conference on Machine Learning, Banff, Canada,* 55–60.

[22] Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004). Evaluating Collaborative Filtering Recommender Systems. *ACM Trans. Inf. Syst.* **22** 5–53.

[23] Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association* **97** 1090–1098. MR1951262

[24] Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5** 109–137. MR0718088

[25] Holland, P. W. and Leinhardt, S. (1981). An Exponential Family of Probability Distributions for Directed Graphs. *Journal of the American Statistical Association* **76** 33–50. MR0608176

[26] Jaccard, P. (1901). étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles* **37** 547–579.

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

[28] Krivitsky, P. N., Handcock, M. S., Raftery, A. E., and Hoff, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks* **31** 204–213.

[29] Lazega, E. (2001). *The Collegial Phenomenon: The Social Mechanisms of Cooperation among Peers in a Corporate Law Partnership*. Oxford University Press.

[30] Leicht, E. A., Holme, P., and Newman, M. E. J. (2006). Vertex similarity in networks. *Physical Review E Statal Nonlinear & Soft Matter Physics* **73** 026120.

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

[32] Lü, L. and Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A* **390** 1150–1170.

[33] Ma, C., Bao, Z., and Zhang, H. (2017). Improving link prediction in complex networks by adaptively exploiting multiple structural features of networks. *Physics Letters A* **381** 3369–3376.

[34] Martínez, V., Berzal, F., and Cubero, J. C. (2016). A Survey of Link Prediction in Complex Networks. *Acm Computing Surveys* **49** 69.

[35] Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *Computer Physics Communications* **147** 40–45. MR1913364

[36] Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America* **103** 8577–8582.

[37] Newman, M. E. J. (2010). *Networks: An introduction*. Oxford University Press, Inc. MR2676073

[38] Newman, M. E. J. and Leicht, E. A. (2007). Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences of the United States of America* **104** 9564–9569.

[39] Olhede, S. C. and Wolfe, P. J. (2014). Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences of the United States of America* **111** 14722–14727.

[40] Pech, R., Hao, D., Lee, Y.-L., Yuan, Y., and Zhou, T. (2019). Link prediction via linear optimization. *Physica A* **528** 121319. MR3952211

[41] Rohe, K., Chatterjee, S., and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics* **39** 1878–1915. MR2893856

[42] Song, A., Liu, Y., Wu, Z., Zhai, M., and Luo, J. (2019). A Local Random Walk Model for Complex Networks Based on Discriminative Feature Combinations. *Expert Systems with Applications* **118** 329–339.

[43] Sun, S., Ling, L., Zhang, N., Li, G., and Chen, R. (2003). Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Research* **31** 2443–2450.

[44] Wang, P., Xu, B., Wu, Y., and Zhou, X. (2015). Link prediction in social networks:the state-of-the-art. *Science China: Information Sciences* **58** 1–38.

[45] Watts, D. J. and Strogatz, S. H. (1998). Collective Dynamics of Small World Networks. *Nature* **393** 440–442.

[46] Wolfe, P. J. and Olhede, S. C. (2013). Nonparametric graphon estimation. *Eprint Arxiv.*

[47] Wu, Z., Lin, Y., Wang, j., and Gregory, S. (2016). Link prediction with node clustering coefficient. *Physica A* **452** 1–8.

[48] Yu, K., Chu, W., Yu, S., Tresp, V., and Xu, Z. (2006). Stochastic Relational Models for Discriminative Link Prediction. *Advances in Neural Information Processing Systems* **19** 1553–1560.

[49] Zachary, W. W. (1977). An Information Flow Model for Conflict and Fission in Small Groups1. *Journal of Anthropological Research* **33** 452–473.

[50] Zhang, X., Zhao, C., Wang, X., and Yi, D. (2015). Identifying Missing and Spurious Interactions in Directed Networks. *International Journal of Distributed Sensor Networks* **11** 507386–10.

[51] Zhang, Y., Levina, E., and Zhu, J. (2017). Estimating network edge probabilities by neighborhood smoothing. *Biometrika* **104** 771–783. MR3737303

[52] Zhao, Y., Wu, Y.-J., Levina, E., and Zhu, J. (2017). Link Prediction for Partially Observed Networks. *Journal of Computational and Graphical Statistics* **26** 725–733. MR3698680

[53] Zhou, T., Lü, L., and Zhang, Y. (2009). Predicting missing links via local information. *European Physical Journal B* **71** 623–630.

Chunning Wang
School of Mathematics and Statistics
Lanzhou University
Lanzhou
China
School of Statistics
Lanzhou University of Finance and Economics
Lanzhou
China
E-mail address: leven_wong@163.com

Bingyi Jing
Department of Mathematics
Hong Kong University of Science and Technology
Hong Kong
China
E-mail address: majing@ust.hk