https://doi.org/10.1093/comjnl/bxac157

Link Prediction Based On Local Structure And Node Information Along Local Paths

Tongfeng Li[1,2], Ruisheng Zhang[1,*], Bojuan Niu[1], Yabing Yao[3], Jun Ma[1], Jing Jiang[1] and Zhili Zhao[1]

[1]*School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu 730000, P. R. China*
[2]*Computer College, Qinghai Normal University, Xining, Qinghai 810016, P. R. China*
[3]*School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, P. R. China*
**Corresponding author: zhangrs@lzu.edu.cn*

Link prediction aims at predicting the missing links or new links based on known topological or attribute information of networks, which is one of the most significant and challenging tasks in complex network analysis. Recently, many local similarity-based methods have been proposed and they performed well in most cases. However, most of these methods simultaneously ignore the contributions of the local structure information between endpoints and their common neighbors, as well as transmission abilities of different 3-hop paths. To address these issues, in this paper, we propose a novel link prediction method that aims at improving the prediction accuracy of the existing local similarity-based methods by integrating with local structure information and node degree information along 3-hop paths. Extensive experiments have been performed on nine real-world networks and the results demonstrate that our proposed method is superior to the existing state-of-the-art methods.

Keywords: Link prediction; Node information; Local structure; Complex networks

Received 9 June 2022; Revised 23 September 2022; Editorial Decision 24 September 2022
Handling editor: Yannis Manolopoulos

1. INTRODUCTION

With the in-depth research of complex network theory, network science plays an increasingly significant role nowadays. Benefiting from it, we can further reveal and understand the phenomena in the real world [1]. Due to the large-scale and complexity of real-world networks, it is hard to gather complete and accurate data [2]. To solve these issues, link prediction has attracted researchers from different disciplines in recent years, which aims at predicting existence possibility of a link between two unconnected nodes based on the current topological information of networks [3]. The research of link prediction not only has significant theory value, but also can be seen in numerous application scenarios, including recommending new friends in online social networks, predicting protein interaction in protein networks, discovering the future collaborators in citation networks, as well as recommending favorite products for users on e-commerce websites, resulting in the strengthening of the loyalty of users to the websites [4–6].

In order to address the problem of link prediction, for the past few years, researchers from different disciplines have proposed various methods. Overall speaking, these methods can be split up into two categories: Supervised-based methods and Unsupervised-based methods [7, 8]. The supervised-based methods consider the problem of link prediction as the binary classification and generally have better performance. However, due to process of feature selection and training classification model, these methods are often time-consuming [9–11]. For the unsupervised-based methods, similarity-based methods, which are one of the most popular methods for link prediction, consider a pair of nodes to be formed as a link if they have high similar scores calculated by attributes of nodes or local topological structure information of networks. However, attributes

of nodes are often closely related to specific information which is difficult to obtain because of the privacy protection policy. By contrast, the topological structure information of one network is easier to obtain and can be calculated in less time. Therefore, many researchers have turned their attention to structure-based similarity methods. Structure-based similarity methods assume that the more the local common structure features of two unconnected nodes, the more they tend to be linked in the future. According to the topological structure information used, these kind of methods can be further grouped into three categories [12]: Local methods, Global methods and Quasi-local methods. For the over past decades, many structure-based similarity methods have been implemented for addressing the problem of link prediction, such as Common Neighbors (CN), Adamic–Adar(AA) and Resource Allocation (RA) [13]. CN method takes the number of common neighbors as the similarity score of two endpoints, the higher the similarity score, the more similar they are. The number of common neighbors is equivalent to the number of 2-hop paths. Therefore, from path perspective, it is also viewed as a path-based index. AA method directly measures two endpoints based on their common features. On the other hand, it is also an improved version of CN index and considers not only the shared neighbors, but also the degree of each neighbor node. The smaller degree nodes contribute more than the bigger degree nodes in similarity. Inspired by the resource allocation process, RA method assumes that a node transmits one unit of resources through common neighbor nodes, and the mount of resources obtained by target node is considered as the their similarity score. Moreover, LAS [14] index measures the similarity of a pair of nodes by their common neighbors and their own degrees. Considering the second-level neighborhood carries the essential network topological structure information, CAR [15] index takes into account the nodes that are interlinked with neighbors mostly with filters the noise. Besides, CNDP [16] index calculates the similarity score of a pair of nodes by their common neighbors and the network average clustering coefficient. Compared with other common neighbor-based link prediction methods, CNDP index distinguishes each common neighbor by its degree information.

All the above indices belong to local methods and they all are also viewed as 2-hop path-based similarity indices with low time-consuming complexity. However, they utilize less structure information, which leads to lower prediction accuracy. To address the shortcomings of the local methods and further improve prediction accuracy, many other methods have been exploited. They utilized the whole topological structure information of the networks to calculate similarity score of a node pair, such as Katz index [17], Leicht-Holme-Newman [18] and Random Walk with Restart (RWR) [19]. Although having high precision of prediction, the global indices suffer from the issues of high time consumption. Quasi-local indices, which use more topological structures of networks than local indices and less than global indices, own a balance between performance and complexity, including Local Path index (LP) [20], Local Random Walks (LRW) [21] and Friend Link index (FL) [22]. L3 [23] index can be viewed as a variant of CN index which considers the contributions of 3-hop paths. The basic idea is that if the number of 3-hop paths between two nodes is more, the two nodes are more likely to be connected in the future. To sum up, Table 1 summarizes the characteristics, strengths, weakness of Local methods, Quasi-local methods and Global methods.

Many experiments have demonstrated that quasi-local indices possess better performance than local and global indices, especially in 2-hop paths and 3-hop paths, performing highlight. Recently, Zhou *et al.* [24] made comprehensive experiments on 128 public network datasets and the outcomes indicated that the 3-hop-based indices performed slightly better with a winning rate about 55.88 percent than the 2-hop-based indices. On the other hand, it is worthwhile to consider the contributions of 3-hop paths in local similarity-based indices.

As we all know, the structure-based similarity indices generally rely on the observed topological structure of network. The aim is to get the similarity scores based on the topological structure information. The local similarity-based indices mainly adopt the number of 2-hop paths of two nodes to calculate similarity, such as CN, RA, LAS, IA [25], *et al.* However, they all ignore the local structure information between common neighbors and endpoints. While the quasi-local indices make a trade-off between complexity and accuracy, the prediction accuracy is better than global methods in most cases. For example, LP index both considers the contributions of 3-hop paths and 2-hop paths to measure the similarity of two nodes. Besides, L3 method calculates the similarity scores of node pairs by the number of 3-hop paths. However, each path has a different ability to transmit similarity in real world. Considering the computational complexity and to further improve the prediction accuracy, we need a method that addresses the above issues and integrates with the advantages of the local and quasi-local indices simultaneously.

Let us use Fig. 1 to further illustrate our idea. In the traditional way, when calculating the similarity score between nodes x and y, we usually count the number of common neighbors, paths or integrate with node degree information as the final similarity score. However, most of them ignore the local structure information and the transmission ability of different paths. To solve this issues, in our index, we measure the similarity of two nodes in two steps. Firstly, above the dotted line, in addition to the common neighbor w, which is adopted as the similarity score of nodes x and y, we also utilize the inverse of degree of w, which differentiates the contribution of every common neighbor, and the local structures $\{x, b, w\}$ and $\{x, a, w\}$, which help to improve the precision of prediction. The same reason for nodes w and y. Secondly, we remove the links associated with node w to consider the contribution of path$\{x, f, g, y\}$ under the dotted line. In order to distinguish the transmission ability of path$\{x, f, g, y\}$, we simultaneously consider the degree information of nodes f and g. Inspired

TABLE 1. The characteristics of characteristics, strengths, weakness of Local methods, Quasi-local methods and Global methods.

Method subcategory	Characteristics	Strengths	Weaknesses
Local methods	Only neighborhood information is considered,such as AA, CN, AA, LAS, etc.	Straightforward; low computational complexity	Low accuracy
Global methods	Consider the whole topological information of network, such as Katz, ACT, GLHN,RPR, etc.	Usually with better prediction accuracy	Time-consuming and complexity
Quasi-local methods	Trade-off between local methods and global methods, such as LP, IA, L3, CNDP, etc.	Lower computational complexity than global methods; Higher prediction accuracy than local methods	Dependent on datasets and applications

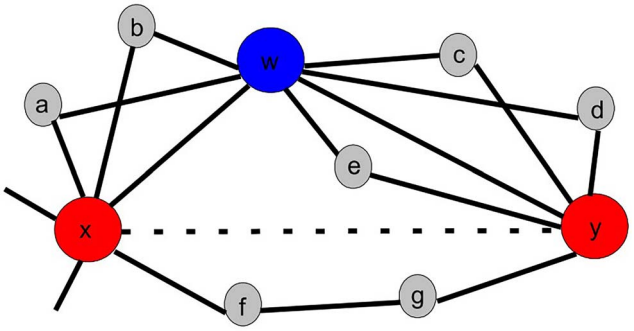


FIGURE 1. A simple network to illustrate the LSNI index.

by above analysis, let us say that the local structure and node degree information have significant contributions toward link prediction. So we come up with a novel link prediction based on Local Structure and Node Information(LSNI) along local paths with taking simultaneously the advantages of local and quasi-local indices.

The rest of the manuscript is organized as follows: In the Methods section, the LSNI index is described and defined in detail. The comparison indices and metrics are given in metrics and baselines section. In the experimental results and analysis section, LSNI index is implemented on different networks and the performances are evaluated. In the last section, we make a conclusion of our work.

2. METHODS

We will utilize four definitions to illustrate our index. Specifically, **Definition 1** and **Definition 2** are used to calculate the local structure information and **Definition 3** is used to distinguish the contribution of every 3-hop path. Combining with **Definitions 1**, **2** and **3**, we give the final definition of the proposed method by **Definition 4**.

DEFINITION 1. *$x, y \in V$ are arbitrary nodes of a network, the neighbor node set of x is represented by $\Gamma(x)$, w is defined as one shared neighbor of x and y, the similarity score of node pair(x, y) can be described as*

$$C^1(x,y) = \sum_{w \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_w} \big(|\Gamma(w) \cap \Gamma(y)| + \alpha \big), \quad (1)$$

where k_w is the degree of w. When $\Gamma(w) \cap \Gamma(y) = \phi$, it only takes the degree of common neighbors into account and deteriorates to the RA index, then α=1. Otherwise, α=0. According to this function, we simultaneously consider the local structures between nodes w and y and distinguish the contribution of every common neighbor. In fact, the local structures between nodes x and w are also helpful to connection between x and y.

DEFINITION 2. *The contributions of the local structures between node x and node w to the similarity of node pair (x, y) can be defined as*

$$C^2(x,y) = \sum_{w \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_w} \big(|\Gamma(x) \cap \Gamma(w)| + \alpha \big). \quad (2)$$

C^1 and C^2 simultaneously consider the contributions of local structures and node degree information to the similarity of x and y. However, the contributions of 3-hop paths are ignored. Therefore, we add the contributions of 3-hop paths and utilize the inverse of degrees of nodes on 3-hop paths to distinguish the contribution of each path.

DEFINITION 3. *P is the collections of 3-hop paths from x to y. p=$\{x, w_1, w_2, y\}$ is one of the P.*

$$C^3(x,y) = \sum_{p \in P} \sum_{w \in \{w_1, w_2\}} \frac{1}{k_w}, \quad (3)$$

where w is the intermediate nodes.

Therefore, overall considering function(1), function(2) and function(3), the ultimate index can be described as follows:

Definition 4. *The final similarity score between nodes x and y is defined as*

$$S_{xy}^{LSNI} = \sum_{w\in\Gamma(x)\cap\Gamma(y)} \frac{1}{k_w}\big(|\Gamma(w)\cap\Gamma(y)|+\alpha\big) + \sum_{w\in\Gamma(x)\cap\Gamma(y)} \frac{1}{k_w}\big(|\Gamma(x)\cap\Gamma(w)|+\alpha\big) + \beta * \sum_{p\in P}\sum_{w\in\{w_1,w_2\}} \frac{1}{k_w}, \quad (4)$$

where β is a free parameter. The implementation framework of LSNI index is showed in Algorithm 1, and the terms and notations are listed in Table 2

Algorithm 1 The implementation of LSNI index

Input: Network $G(V,E)$, node pair(x,y), Parameter: β.

Output: Similarity score S_{xy}.

1: Calculate the shared neighbor set W between nodes x and y.
2: Compute the degree of each $w \in W$.
3: Use Eq.(2) to calculate the similarity C^2 of nodes x and y.
4: Utilize Eq.(1) to calculate the similarity C^1 of nodes x and y.
5: Remove the links associated with neighbor set W.
6: Compute the similarity C^3 between nodes x and y by equation Eq.(3).
7: end for.
8: Similarity score: S_{xy}=C^1+C^2+β*C^3.

3. EXPERIMENT IMPLEMENTATION, COMPLEXITY ANALYSIS AND TIME PERFORMANCE

3.1. Experiment implementation

We implemented our proposed algorithm LSNI using Python3.7 on Ubuntu system. During the implementation of our algorithm, we mainly utilized NetworkX package and numpy package. In detail, we use function NetworkX.common_neighbors () to find the common neighbors of node pairs, NetworkX. degree() to compute the degree of one node and NetworkX.all_simple_paths() to find 3-hop paths. In all intermediate calculations, the common neighbors of node pairs, the 3-hop paths and the similarity scores of node pairs need a certain amount of spatial storage. In terms of partial results, we use List data structure to preserve the common neighbors of any node pairs and 3-hop paths, and use matrix structures from the Numpy package, where every element represents the similarity of two nodes, the larger the value, the more similar they are, to preserve the partial prediction results.

TABLE 2. The notations and definitions.

Notation	Definition
$G=(V,E)$	A network G with node set V and edge set E
$N=\|V\|$	The number of nodes in G
$\langle k\rangle=\frac{2\|E\|}{N}$	The average degree of nodes in G
$\beta \geq 0$	A free parameter
$\Gamma(x)$	The neighbor set of node x
k_x	The degree of node x
U	A set of all possible links in network G

3.2. Complexity analysis

Inspired by the complexity analysis approach in [26], we describe the time complexity of LSNI in detail. The time complexity of LSNI algorithm mainly contains three parts. First, we calculate the common neighbors of two nodes, the time complexity is the $O(N\langle k\rangle^2)$, which is equal to the CN index, $\langle k\rangle$ is the average degree of one node and $N=|V|$, where V is the set of nodes. Second, based on the first step, we obtain the local structure information between one endpoint and common neighbors, i.e. C^1 and C^2. The time complexity of C^1 and C^2 is $O(N\langle k\rangle^2)$. Third, the time complexity of finding 3-hop paths is $O(N\langle k\rangle^3)$. Adding the time complexity of all three steps, the whole time complexity is $O(N\langle k\rangle^3+N\langle k\rangle^2+N\langle k\rangle^2)\approx O(N\langle k\rangle^3)$, where $\langle k\rangle \ll N$. In the sparse networks, in the extreme case while maintaining connectivity, the minimum number of edges is $N-1$ and the $\langle k\rangle = 2(N-1)/N \approx 2$, the time complexity of LSNI is $O(N)$. While in the dense networks, in the extreme case, $|E|=N(N-1)/2$, then the $\langle k\rangle = 2|E|/N = N-1$, which approximately equals to N in the case of large N, Therefore, the time complexity of LSNI is $O(N^4)$. In other words, LSNI is complex due to complex 3-hop paths finding and its time complexity is higher than or at least equal with the baseline methods.

3.3. Time performance

In terms of time performance, in order to eliminate randomness, every network is calculated 10 times and we use the average value of 10 times as the final similarity scores of node pairs. In our work, nine networks are utilized and the total time is 1126.708 s of nine networks. Specifically, the time performance of single prediction (all node pairs) in every network is as follows:

Asoiaf: 0.5988s, Blogs: 1.440s, Dnccorecipient: 6.561s, Pajekerdos: 58.159s, Hamsterster: 6.9926s, Adolescent health: 7.2343s, Power grid: 27.813s, Languages: 0.3749s, Virgili: 3.4968s.

4. METRICS AND BASELINES

Given a simple network $G = (V, E)$, $|U| = |V|(|V| - 1)/2$. All observed links in E are randomly divided into two parts: one part is the probe set E^P and the other part is training set E^T. Here, the links in E^P and U-E are named as missing links and non-existent links, respectively. E^T is used to train the proposed link prediction index. Therefore, $E^P \cup E^T = E$ and $E^P \cap E^T = \emptyset$. While the performance of the link prediction index is estimated by E^P. At the same time, we use the AUC (Area Under The Receiver Operating Characteristic Curve), Precision and Recall three metrics [12, 27] to assess the link prediction accuracy of our model and baselines.

The AUC, usually interpreted as the probability that the similarity score of one node pair randomly selected from E^P, is higher than that of one randomly selected from U-E [11]. n times of independent comparisons are run, if there are n' times that the scores of missing links are higher than the non-existent links and n'' times where both of them have equal scores, then the AUC is defined as

$$AUC = \frac{n' + 0.5 * n''}{n}. \tag{5}$$

Generally, the closer the value of AUC is to 1, the more accuracy the link prediction index is.

(2) Precision: Given the scores of non-observed links and ordered in descend, the definition of precision is the ratio of the proportion of related links selected to the number of links chosen. Namely, if top-L links are choose as the predicted links, among which m links are correct, then the accuracy can be mathematically described as

$$precision = \frac{m}{L} \tag{6}$$

(3) The recall is defined as the ratio of m of relevant links in all links R.

$$Recall = \frac{m}{R} \tag{7}$$

4.1. Structure-based similarity baselines

(1) Common Neighbor index (CN) [28], give n an arbitrarily node x in network, the similarity between x and another node y is calculated by the number of their common neighbors, which is defined as follows:

$$S_{xy}^{CN}(x, y) = |\Gamma(x) \cap \Gamma(y)|, \tag{8}$$

where $\Gamma(x)$ is the neighbor collections of node x. CN index is also regarded as the path-based similarity index, which the path length is two.

(2) LAS index [14]. This index can be seen as the improvement of CN index. It not only considers the common neighbors, but also takes degree of the node into account. It is defined as

$$S_{xy}^{LAS} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x)|} + \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(y)|} \tag{9}$$

(3) CAR [15], which filters these noises(the second-level neighborhood) and takes into account nodes that are interlinked with neighbors mostly. The similarity metric is defined as

$$S_{xy}^{CAR} = |\Gamma(x) \cap \Gamma(y)| \sum_{v \in (\Gamma(x) \cap \Gamma(y))} \frac{|\Gamma(v)|}{2} \tag{10}$$

(4) IA index [25]. This index simultaneously considers the common neighbors of nodes and degrees of common neighbors and is defined as

$$S_{xy}^{IA} = \sum_{v \in (\Gamma(x) \cap \Gamma(y))} \frac{|\Gamma(x) \cap \Gamma(y) \cap \Gamma(v)| + 2}{|\Gamma(v)|} \tag{11}$$

(5) L3 [23] is a variation of CN and it only considers the contribution of the 3-hop path, and the connection probability between two nodes is defined as follows:

$$S_{xy}^{L3} = (A^3)_{ij} \tag{12}$$

(6) Adamic-Adar index(AA) index [21]. This index can be viewed as a variant of CN index, which takes the degrees of shared neighbors into account. It is defined as

$$S_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z}, \tag{13}$$

where z is the shared neighbor of x and y. k_z is the degree of z.

(7) Resource Allocation(RA) [6] index. Inspired by the resource allocation mechanism in network, its main goal is to improve prediction precision by punishing shared neighbors with more connection. It is defined as

$$S_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} \tag{14}$$

(8) Local Paths(LP) [20] index. This index only considers the number of 2-hop paths and 3-hop paths and a good trade-off is made between computational complexity and prediction accuracy. It is defined as

$$S_{xy}^{LP} = A^2 + \beta A^3 \tag{15}$$

TABLE 3. Some attributes of the nine real-world networks.

Networks	$\|V\|$	$\|E\|$	$\|C\|$	$\langle k\rangle$	$\langle d\rangle$
Asoiaf	796	32629	0.209	81.982	3.411
Blogs	1224	19025	0.226	31.087	2.772
Dnc corecipient	2029	136602	0.548	134.650	2.713
Pajek erdos	6927	11850	0.036	3.421	3.791
Hamsterster	2426	16631	0.231	13.71	3.669
Adolescent health	2539	12969	0.142	10.215	4.516
Power grid	4941	6594	0.103	2.669	20.094
Languages	868	1255	0.083	2.892	4.076
Virgili	1133	5451	0.166	9.622	3.655

Notes:$|V|$ represents the number of nodes, $|E|$ represents the number of links, $|C|$indicates the clustering coefficient. $\langle k\rangle$ denotes the average degree, $\langle d\rangle$ indicates the average distance.

A is the adjacent matrix of the network and β is a adjustable parameter. According to this index, when β=0, it degrades to A^2, which equals to CN index.

(9) CNDP [16] index, which estimates the connection probability between two nodes x and y in terms of the topological characteristics including common neighbors of each two nodes and the network average clustering coefficient. It is defined as

$$S_{xy}^{CNDP} = |C_z|(|\Gamma(z)|)^{-\beta C} \tag{16}$$

C_z is the number of neighbors of z which consists of the common neighbors of x and y in addition to x and y. C is the network average clustering coefficient, and β is a constant value which needs to be optimized.

(10) Common Neighbor and Distance(CND) [29]. This index regards common neighbor and distance play a key role in calculating the similarity of two nodes. Given any two nodes x and y, a similar score between them is calculated by Eq.(17).

$$S_{xy}^{CND} = \begin{cases} \frac{CN_{xy}+1}{2}, & \Gamma(x)\cap\Gamma(y)\neq\emptyset \\ \frac{1}{d_{xy}}, & otherwise \end{cases}, \tag{17}$$

where Γ(x) represents the neighbor collections of node x, CN_{xy} is the number of common neighbors between node x and y and d_{xy} is defined as the distance of nodes x and y.

(11) CCPA [30] index. Inspired by important properties of nodes, it considers simultaneously common neighbors and centrality to measure the similarity and defined as

$$S_{xy}^{CCPA} = \alpha\,(|\Gamma\,(x)\cap\Gamma\,(y)|) + (1-\alpha) * \frac{N}{d_{xy}} \tag{18}$$

In this index, parameter $\alpha \in [0,1]$ is used to control the weight between common neighbors and centrality. N and d_{xy} are number of nodes and shortest path between nodes x and y, respectively.

5. EXPERIMENTAL RESULTS AND ANALYSIS

5.1. Datasets

In this paper, we utilize nine real-world networks from different domains with different topological structures. The detailed properties of these networks can be seen in Table 3 and are briefly explained as follows: (1)Asoiaf: It is the fictional social network of the series of fantasy novels 'A Song of Ice and Fire' by George R. R. Martin (1996 – *presen*). (2)Blogs: A node represents a blog and an edge represents a hyperlink between two blogs. (3)Dnc corecipient: It is the undirected network of people having received the same email in the 2016 Democratic National Committee email leak. (4)Pajek erdos: It is the co-authorship graph around Paul Erdös. (5)Hamsterster: The network contains friendships and family links between users of the website hamsterster.com. (6)Adolescent health: A node represents a student and an edge between two students shows that the left student chose the right student as a friend. (7)Power grid: The power network of an electrical grid of western US. (8)Languages: This network denotes which languages are spoken in which countries. (9)Virgili: An email network. All above networks can be download from the websites: http://konect.cc/networks/ and http://networkrepository.com/.

5.2. The impact of the variation of β on the prediction accuracy

In our experiments, each network is randomly divided 10 times and the average prediction value of 10 times is used as the final scores of node pairs.

The construction of the LSNI index can be learned from Eq.4; it simultaneously takes into account the common neighbors and different transmission ability of every 3-hop path, as well as degree information of node and local structure information with an adjustable parameter $\beta \in [0,+\infty]$ to make a control. For LSNI index, if we solely take care of common neighbors and local structures, it indicates that the value of β is 0. In this subsection, we focus on the impact of the

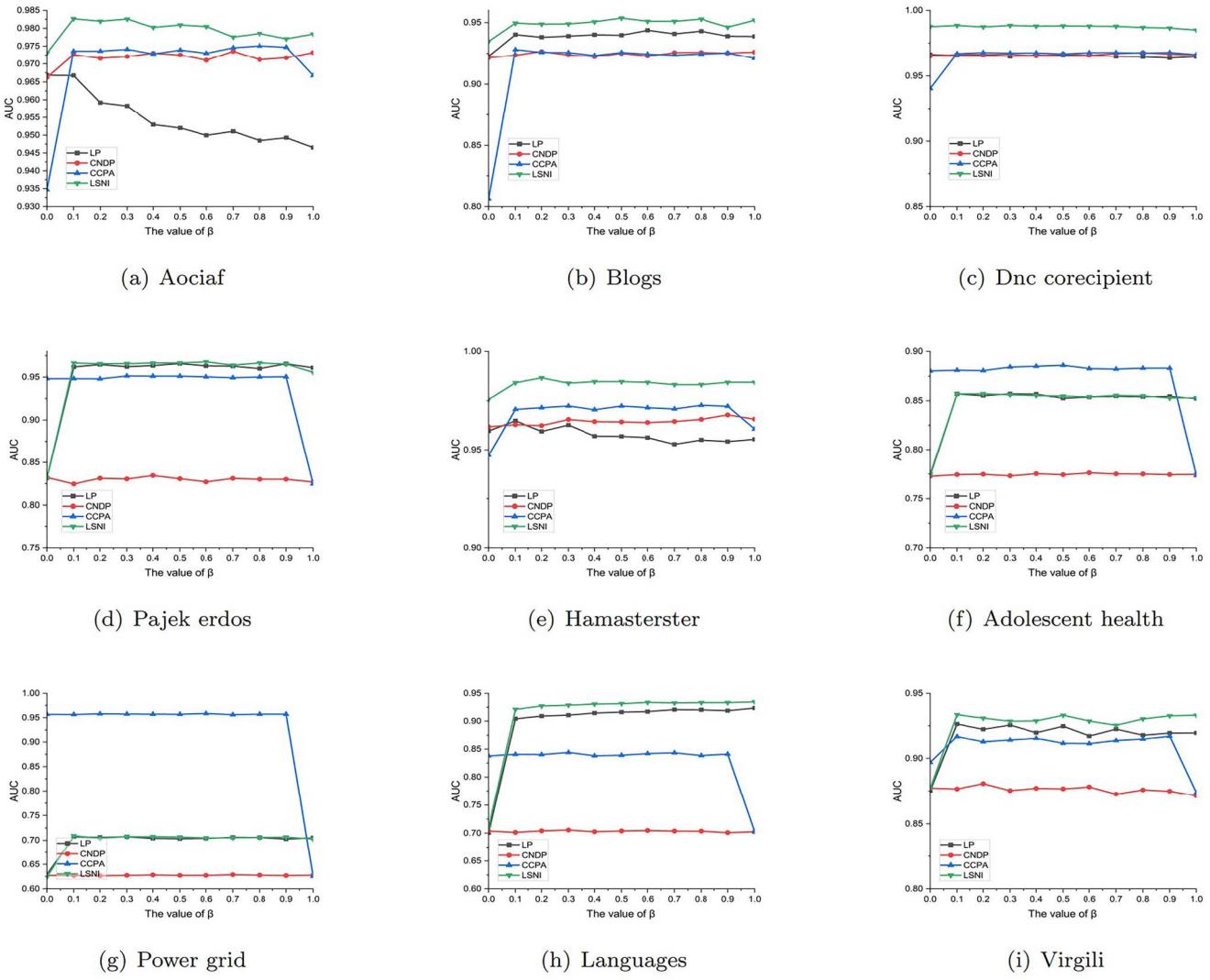


FIGURE 2. The prediction precision of LP, CNDP, CCPA and LSNI indices in nine real networks with different values of $\beta \in [0, 1.0]$. Each value represents the average value of 10 independent operations.

TABLE 4. The prediction accuracy measured by AUC.

Networks	*CN*	*IA*	*CAR*	*CND*	*L3*	*AA*	*RA*	*LP*	*LAS*	*CNDP*	*CCPA*	*LSNI*
Asoiaf	0.9660	0.9741	0.9505	0.9748	0.9339	0.9744	0.9729	0.9668	0.9466	0.9726	0.9751	**0.9827**
Blogs	0.9199	0.9260	0.8828	0.9246	0.9387	0.9246	0.9268	0.9430	0.8687	0.9257	0.9258	**0.9537**
Dnc corecipient	0.9659	0.9667	0.9645	0.9681	0.9626	0.9663	0.9674	0.9657	0.9543	0.9663	0.9671	**0.9885**
Pajek erdos	0.8289	0.8273	0.8191	0.9514	0.9604	0.8254	0.8295	0.9558	0.8224	0.8347	0.9515	**0.9679**
Hamasterster	0.9637	0.9658	0.9559	0.9716	0.9479	0.9638	0.9678	0.9648	0.9603	0.9678	0.9723	**0.9867**
Adolescent health	0.7731	0.7746	0.7725	0.8830	0.8247	0.7776	0.7728	0.8569	0.7725	0.7745	**0.8831**	0.8571
Power grid	0.6285	0.6269	0.6286	0.9573	0.6369	0.6259	0.6285	0.7073	0.6295	0.6289	**0.9581**	0.7091
Languages	0.7014	0.7042	0.6882	0.8408	0.9211	0.7019	0.7011	0.9209	0.6840	0.7050	0.8440	**0.9347**
Virgili	0.8619	0.8622	0.8579	0.9077	0.8998	0.8652	0.8633	0.9195	0.8572	0.8645	0.9118	**0.9272**

Notes: The AUC is measured by the mean of 10 independent implementations with a random 90%–10% division of training set and probe set. The best performance in each network is highlighted in bold.

TABLE 5. The prediction accuracy measured by Precision.

Networks	*CN*	*IA*	*CAR*	*CND*	*L3*	*AA*	*RA*	*LP*	*LAS*	*CNDP*	*CCPA*	*LSNI*
Asoiaf	0.108	0.107	0.118	0.025	0.052	0.097	0.102	0.061	0.258	0.113	0.10	**0.273**
Blogs	0.16	0.155	0.128	0.027	0.09	0.157	0.153	0.092	0.177	0.153	0.005	**0.181**
Dnc corecipient	0.172	0.182	0.138	0.01	0.145	0.18	0.181	0.145	0.293	0.17	0.172	**0.316**
Pajek erdos	0.155	0.155	0.128	0.027	0.09	0.157	0.153	0.098	0.177	0.157	0.005	**0.188**
Hamasterster	0.102	0.105	0.087	0.022	0.047	0.103	0.106	0.06	0.482	0.103	0.1	**0.518**
Adolescent health	0.113	0.143	0.12	0.007	0.017	0.11	0.135	0.04	**0.288**	0.108	0.113	0.147
Power grid	0.133	0.148	0.127	0.025	0.005	0.14	0.153	0.038	**0.423**	0.133	0.122	0.140
Languages	0.126	0.125	0.148	0.052	0.003	0.122	0.113	0.02	0.145	0.123	0.001	**0.127**
Virgili	0.119	0.105	0.115	0.052	0.003	0.121	0.113	0.005	0.115	0.123	0.11	**0.126**

Notes: The Precision is measured by the mean of 10 independent implementations with a random 90%–10% division of training set and probe set. The best performance in each network is highlighted in bold.

TABLE 6. The prediction accuracy measured by Recall.

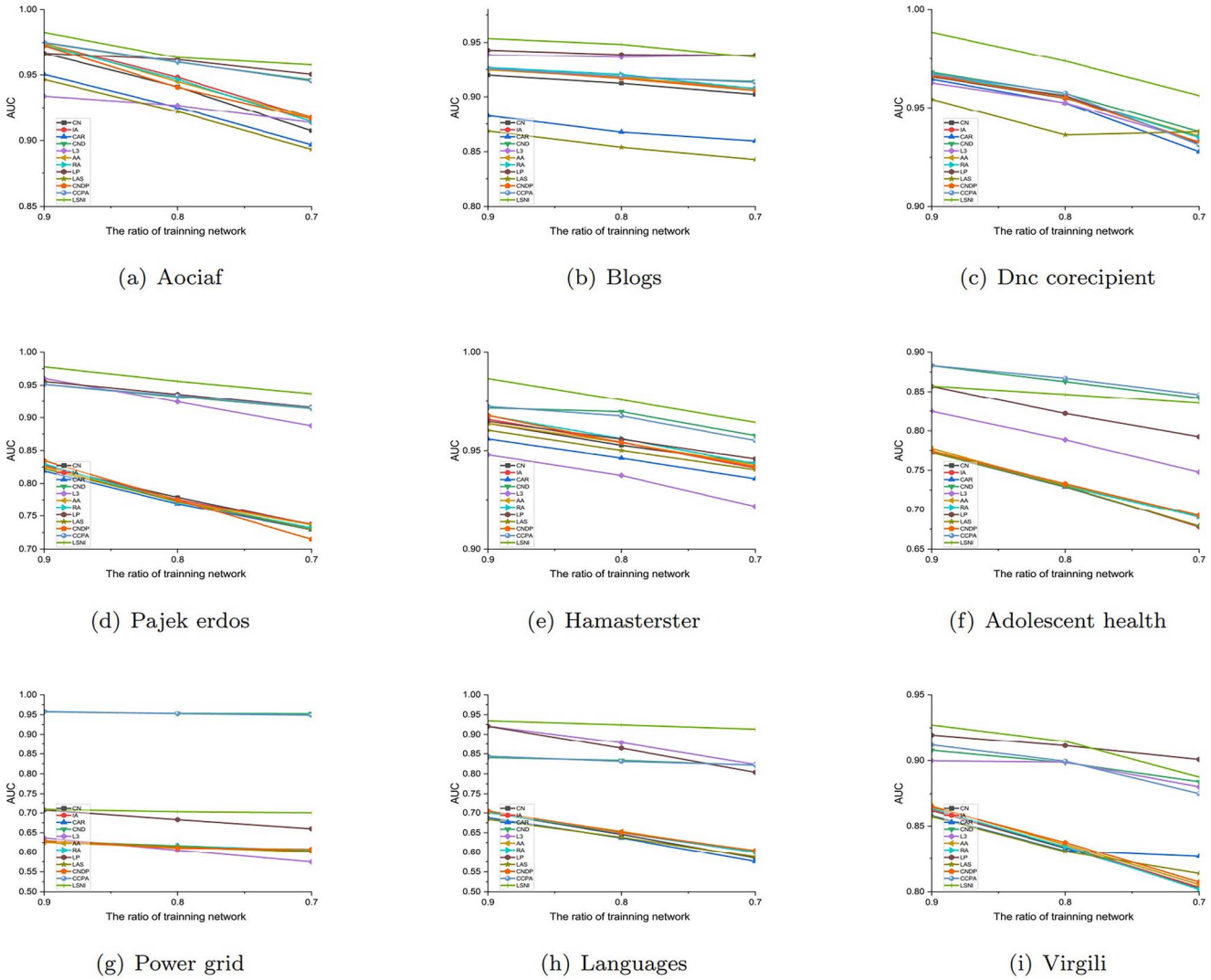
Networks	*CN*	*IA*	*CAR*	*CND*	*L3*	*AA*	*RA*	*LP*	*LAS*	*CNDP*	*CCPA*	*LSNI*
Asoiaf	0.110	0.109	0.115	0.056	0.071	0.109	0.108	0.083	0.196	0.109	0.110	**0.211**
Blogs	0.127	0.124	0.123	0.061	0.1	0.126	0.123	0.101	0.141	0.127	0.001	**0.153**
Dnc corecipient	0.131	0.134	0.126	0.016	0.113	0.131	0.134	0.114	0.130	0.132	0.131	**0.273**
Pajek erdos	0.127	0.124	0.123	0.061	0.1	0.126	0.123	0.103	0.141	0.127	0.001	**0.161**
Hamasterster	0.104	0.101	0.103	0.043	0.063	0.103	0.100	0.07	0.124	0.101	0.104	**0.163**
Adolescent health	0.136	0.135	0.125	0.068	0.012	0.137	0.136	0.05	**0.158**	0.137	0.133	0.152
Power grid	0.144	0.144	0.144	0.128	0.003	0.144	0.144	0.060	0.144	0.134	**0.155**	0.149
Languages	0.123	0.128	0.158	0.055	0.005	0.119	0.120	0.02	0.157	0.125	0.002	**0.165**
Virgili	0.119	0.105	0.115	0.052	0.003	0.121	0.113	0.005	0.115	0.123	0.11	**0.126**

Notes: The Precision is measured by the mean of 10 independent implementations with a random 90%–10% division of training set and probe set. The best performance in each network is highlighted in bold.

variation of β on the prediction accuracy. We will employ AUC, Precision and Recall as the evaluation metrics to evaluate the performance of the LSNI index. Due to the AUC measures the performance of a link prediction method as a whole, we only plot the AUC results of different β values of LSNI index in Fig. 2. Theoretically, the value of β can be ranged from 0 to $+\infty$. However, through practical experiments, it is found that its optimal parameter always located at 0 and 1. Meanwhile, LP, CCPA and CNDP indices consider common neighbors and all rely on parameters; their AUC results are also plotted in this figure for comparison. In the following text, we will make a description of experimental results in detail. As shown in Fig. 2 the LSNI index gives higher prediction accuracy than LP, CCPA and CNDP index in nine real networks when β at an optimal value. In general, the prediction accuracy increases with the increases of the value of β and decreases when it reaches a certain value. In most networks, an optimal AUC value can be obtained at or close to 0.1. This clearly shows that by incorporating 3-hop paths, the AUC values of LP, CNDP, CCPA and LSNI algorithm significantly increase. In detail, when β equals to 0, LSNI, CNDP and LP index all degenerate to common neighbor index. However, the prediction accuracy of LSNI is better than LP and CNDP in most networks, which is a benefit from the degree of nodes and local structures between common neighbors and endpoints. It also clearly shows that considering the local structure and node degree information does help to improve similarities. When the value of β is greater than 0, the AUC of LSNI is significantly higher than other indicators, the reason being that it also takes into account the number of 3-hop paths and the contribution of each path is distinguished. From the path aspect, CCPA index only takes 2-hop paths and one fixed length path between source node and target node into account. When $\beta = 1$, it degenerates to CN index, which is why there is a big drop in most networks. Through the above analysis, the LSNI index makes effective employing the local structure information and degrees of nodes, as well as the different contribution of every 3-hop path for link prediction.

5.3. The AUC, Precision and Recall values of LSNI index compared with baselines

In this paper, our work focuses on the local similarity-based methods, Therefore, to assess the performance of the LSNI index, we compared it with 11 similarity indices on nine networks, including local similarity indices and quasi-local



(a) Aociaf (b) Blogs (c) Dnc corecipient

(d) Pajek erdos (e) Hamasterster (f) Adolescent health

(g) Power grid (h) Languages (i) Virgili

FIGURE 3. The prediction precision measured by AUC of all indices on nine real networks with different ratios of training sets. Each point represents the average value of 10 independent operations.

similarity indices. The prediction outcomes of these similarity indices are displayed in Table 4, Table 5 and Table 6. The peak of the prediction values in each network are shown in black. Compared with other 11 baselines, the proposed LSNI index has the best performance on most real networks. Compared with CN index, CCPA and CND indices are cut a fine figure. This is because they both take simultaneously the number of common neighbors and centrality into account. However, CCPA performs better than CND. Next, from Table 4, Table 5 and Table 6, we also note that RA and IA indices perform better than CN index in most cases. It demonstrates that incorporating the degree information of node can enhance the prediction accuracy of the algorithm. According to the results measured by AUC, LP index gives better prediction results than the other baselines except for LSNI index in most networks. Because it not only considers the number of 2-hop paths but also pays attention to the important role of 3-hop paths. The results show that paths with 3-hop actually enhance the prediction accuracy and should be considered. Inspired by the above analysis, our proposed LSNI index takes both the degree of nodes, local structure information between common neighbors and endpoints, the number of 3-hop paths and different transmitting similarity ability of every 3-hop path into account. This is why the LSNI index has better prediction accuracy than other baselines on most real networks.

5.4. The robustness of LSNI index

LSNI index is a good prediction algorithm, which can also perform well in sparse networks. Therefore, in order to further investigate the robustness of LSNI index, different size of training set is chosen, such as 90%, 80%, 70%, and compared

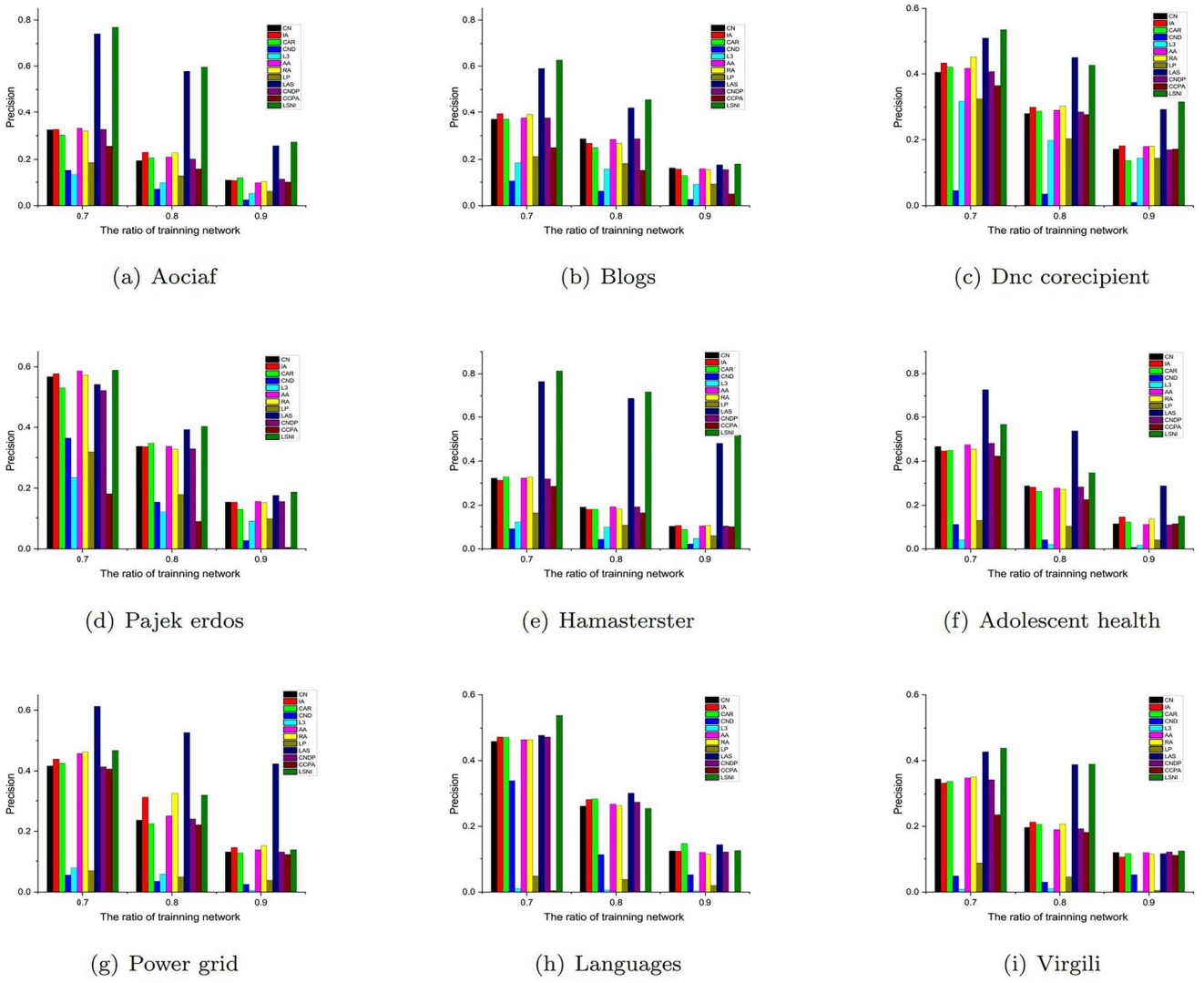


FIGURE 4. The prediction precision measure by Precision of all indices on nine real networks with different ratios of training sets. Each point represents the average value of 10 independent operations.

with 11 baselines on nine real networks. For each splitting set, we run the LSNI index and the other 11 baselines 10 times independently, and plot the average results measured by AUC, Precision and Recall in Fig. 3, Fig. 4 and Fig. 5 to make a comparison, respectively.

From the results of Fig. 3, we can see that the LSNI index owns a higher prediction precision and exceeds the other baselines including local and quasi-local indices on most networks. We also note that the prediction accuracy of the index generally decreases with the decrease in the size of training sets, because less known topological structure information is used. As can be seen in Fig. 4 and Fig. 5, the Precision and Recall results decrease as the proportion of the training set increases on all networks. This is because the test set increases and a link is chosen with a higher probability. But the LSNI index performs better than the other baselines on most networks. Therefore, from the results measured by AUC, Precision and Recall it can be seen that the LSNI index also performs better on the sparse networks.

6. CONCLUSION

In this paper, we develop a novel index called LSNI, which significantly improves the prediction accuracy compared with the state-of-the-art methods. In detail, it incorporates the degrees of nodes, common neighbors and the number of 3-hop paths, as well as local structures between common neighbors and endpoints. Besides, the 3-hop paths are distinguished by the inverse of degrees of nodes. A another advantage of the LNSI index is that it can adapt to different sizes of training sets with

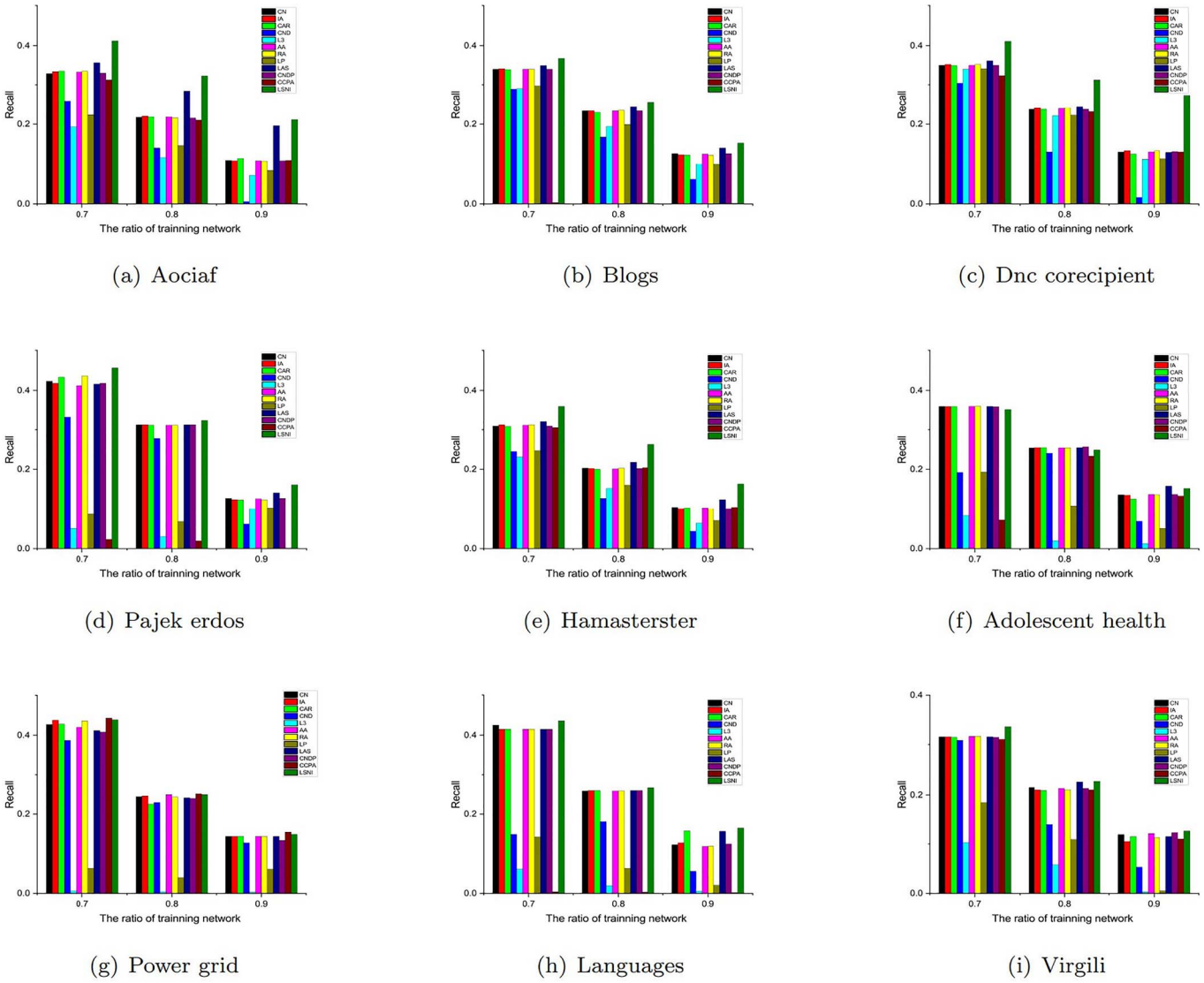


FIGURE 5. The prediction precision measure by Recall of all indices on nine real networks with different ratios of training sets. Each point represents the average value of ten independent operations.

good performance and compensates for the shortcomings of local and quasi-local indices, such as RA, CCPA, CND, LP, LAS, etc.

DATA AVAILABILITY STATEMENTS

The data underlying this article will be shared on reasonable request to the corresponding author.

ACKNOWLEDGEMENTS

This work was supported by the Longyuan Youth Innovation and Entrepreneurship Talents Team Project of Gansu(No.2021 LQTD24),the Higher Education Innovation Fund project of Gansu (No.2022A-022).

REFERENCES

[1] Yao, Y., Zhang, R., Yang, F., Tang, J., Yuan, Y. and Hu, R. (2018) Link prediction in complex networks based on the interactions among paths. *Physica A: Statistical Mechanics and its Applications*, 510, 52–67.

[2] Zeng, S. (2016) Link prediction based on local information considering preferential attachment. *Physica A: Statistical Mechanics and its Applications*, 443, 537–542.

[3] Yi, T., Zhang, S., Bu, Z., Du, J. and Fang, C. (2022) Link prediction based on higher-order structure extraction and autoencoder

learning in directed networks. *Knowledge-Based Systems*, 241, 108241.
[4] Liu, Y., Liu, S., Yu, F. and Yang, X. (2022) Link prediction algorithm based on the initial information contribution of nodes. *Inform. Sci.*, 608, 1591–1616.
[5] Zhao, Z., Gou, Z., Du, Y., Ma, J., Li, T. and Zhang, R. (2022) A novel link prediction algorithm based on inductive matrix completion. *Expert Systems with Applications*, 188, 116033.
[6] Daud, N.N., Ab Hamid, S.H., Saadoon, M., Sahran, F. and Anuar, N.B. (2020) Applications of link prediction in social networks: A review. *Journal of Network and Computer Applications*, 166, 102716.
[7] Tan, F., Xia, Y. and Zhu, B. (2014) Link prediction in complex networks: a mutual information perspective. *PloS one*, 9, e107056.
[8] Cui, Y., Liu, Y., Hu, J. and Li, H. (eds). (2018) A survey of link prediction in information networks. In *2018 IEEE International Conference on Smart Internet of Things (SmartIoT)*Xi'an, China, 17-19 August 2018, pp. 29–33. IEEE, New York.
[9] Li, J.C., Zhao, D.L., Ge, B.F., Yang, K.W. and Chen, Y.W. (2017) A link prediction method for heterogeneous networks based on bp neural network. *Physica A: Statistical Mechanics and its Applications*, 495, 1–17.
[10] Wang, L., Ren, J., Xu, B., Li, J. and Xia, F. (2020) Model: Motif-based deep feature learning for link prediction. *IEEE Transactions on Computational Social Systems*, 7, 1–14.
[11] Pudil, P., Novovičová, J. and Kittler, J. (1994) Floating search methods in feature selection. *Pattern recognition letters*, 15, 1119–1125.
[12] Mutlu, E.C., Oghaz, T., Rajabi, A. and Garibay, I. (2020) Review on learning and extracting graph features for link prediction. *Machine Learning and Knowledge Extraction*, 2, 672–704.
[13] Kumar, A., Singh, S.S., Singh, K. and Biswas, B. (2020) Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and its Applications*, 553, 124289.
[14] Sun, Q., Hu, R., Yang, Z., Yao, Y. and Yang, F. (eds). (2017) An improved link prediction algorithm based on degrees and similarities of nodes. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*Wuhan, China, 24-26 May 2017, pp. 13–18. IEEE, New York.
[15] Cannistraci, C.V., Alanis-Lobato, G. and Ravasi, T. (2013) From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Sci. Rep.*, 3, 1–14.
[16] Rafiee, S., Salavati, C. and Abdollahpouri, A. (2020) Cndp: Link prediction based on common neighbors degree penalization. *Physica A: Statistical Mechanics and its Applications*, 539, 122950.
[17] Katz, L. (1953) A new status index derived from sociometric analysis. *Psychometrika*, 18, 39–43.
[18] Leicht, E.A., Holme, P. and Newman, M.E.J. (2006) Vertex similarity in networks. *Phys. Rev. E*, 73, 026120.
[19] Brin, S. and Page, L. (1998) The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30, 107–117.
[20] Lü, L., Jin, C.H. and Zhou, T. (2009) Similarity index based on local paths for link prediction of complex networks. *Physical Review E*, 80, 046122.
[21] Lü, L. and Zhou, T. (2011) Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390, 1150–1170.
[22] Papadimitriou, A., Symeonidis, P. and Manolopoulos, Y. (2012) Fast and accurate link prediction in social networking systems. *Journal of Systems and Software*, 85, 2119–2132.
[23] Kovács, I.A. *et al.* (2019) Network-based prediction of protein interactions. *Nat. Commun.*, 10, 1–8.
[24] Zhou, T., Lee, Y.-L. and Wang, G. (2021) Experimental analyses on 2-hop-based and 3-hop-based link prediction algorithms. *Physica A: Statistical Mechanics and its Applications*, 564, 125532.
[25] Dong, Y., Ke, Q., Wang, B. and Wu, B. (2011) Link prediction based on local information. In *2011 International Conference on Advances in Social Networks Analysis and Mining*Kaohsiung, Taiwan, 25-27 July 2011, pp. 382–386. IEEE, New York.
[26] Lü, L., Jin, C.-H. and Zhou, T. (2009) Similarity index based on local paths for link prediction of complex networks. *Physical Review E*, 80, 046122.
[27] 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.
[28] Newman, M.E. (2001) Clustering and preferential attachment in growing networks. *Physical review E*, 64, 025102.
[29] Yang, J. and Zhang, X.D. (2016) Predicting missing links in complex networks based on common neighbors and distance. *Sci. Rep.*, 6, 38208.
[30] Ahmad, I., Akhtar, M.U., Noor, S. and Shahnaz, A. (2020) Missing link prediction using common neighbor and centrality based parameterized algorithm. *Sci. Rep.*, 10, 364.