Hindawi
Complexity
Volume 2017, Article ID 8581365, 12 pages
https://doi.org/10.1155/2017/8581365

Research Article

Connecting Patterns Inspire Link Prediction in Complex Networks

Ming-Yang Zhou, Hao Liao, Wen-Man Xiong, Xiang-Yang Wu, and Zong-Wen Wei

Guangdong Province Key Laboratory of Popular High Performance Computers, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

Correspondence should be addressed to Hao Liao; jamesliao520@gmail.com

Received 9 August 2017; Revised 27 November 2017; Accepted 6 December 2017; Published 27 December 2017

Academic Editor: Diego Garlaschelli

Link prediction uses observed data to predict future or potential relations in complex networks. An underlying hypothesis is that two nodes have a high likelihood of connecting together if they share many common characteristics. The key issue is to develop different similarity-evaluating approaches. However, in this paper, by characterizing the differences of the similarity scores of existing and nonexisting links, we find an interesting phenomenon that two nodes with some particular low similarity scores also have a high probability to connect together. Thus, we put forward a new framework that utilizes an optimal one-variable function to adjust the similarity scores of two nodes. Theoretical analysis suggests that more links of low similarity scores (long-range links) could be predicted correctly by our method without losing accuracy. Experiments in real networks reveal that our framework not only enhances the precision significantly but also predicts more long-range links than state-of-the-art methods, which deepens our understanding of the structure of complex networks.

1. Introduction

Modern science and engineering techniques increase our availability to various kinds of data including online social networks, scientific collaboration networks, and power grid networks [1–5]. Many interesting phenomena could be uncovered from these networks. For example, analyzing the data of Facebook and Twitter helps find lost friends by only counting their common friends [6, 7] and recommendation systems in online stores [8, 9]. Restricted by instrument accuracy and other obstacles, we only obtain a small fraction or a snapshot of the complete networks [10, 11], promoting us to filter the information in complex networks [12–14]. Link prediction is a straightforward approach to retrieve networks by predicting missing links and distinguishing spurious links [15–17]. Thus great efforts have been devoted to link prediction in recent years [16, 18]. Link prediction is used in different kinds of networks, including unipartite networks and bipartite networks, where unipartite networks consist of nodes with the same type (e.g., social networks and neural networks) and bipartite networks consist of nodes with two types (e.g., user-object purchasing networks and user-movie networks) [19, 20].

In classical link prediction approaches, similarity scores are computed first for two disconnected nodes, and then nonexisting links in the top of the score list are predicted as potential ones [16]. Consequently, the key issue is to search effective score-assigning methods that are mainly divided into three categories [16, 21]: similarity based algorithms, Bayesian algorithms, and maximum likelihood algorithms. First, similarity based algorithms [22–24] suppose that similar nodes have a high probability to link together. Similarities are evaluated by common neighbors, random walk resource allocation, and some other local and global indices. Second, Bayesian algorithms [25–27] abstract the joint probability distribution from the observed networks and then utilize conditional probability to estimate the likelihood of a nonexisting link. Third, maximum likelihood algorithms [28, 29] presuppose that some underlying principles rule the structure of a network, with the detailed rules and specific parameters obtained by maximum likelihood estimation. Scores of nonexisting links are acquired through the details of

these principles. Most of these methods favor predicting links with high similarity scores and perform badly in the detection of long-range links with low similarities.

In the aforementioned methods, the basic hypothesis that two nodes with a high similarity score have a high likelihood of connecting together lacks an in-depth illustration. Recent works have demonstrated that long-range links exist extensively in complex networks and play an important role in routing, epidemical diffusion, and other dynamics [30, 31]. However, in practice, the endpoints of a long-range link usually have weak interaction and low similarity [30], which prevents the detection of long-range links by traditional methods [32, 33]. Hence, the structural patterns underlying the networks are of great importance to study.

Our study takes a different but complementary approach to link prediction problem. By analyzing the score distributions of existing and nonexisting links, respectively, we find an interesting phenomenon that the existing and nonexisting links follow different connecting patterns in respective of their similarity scores. Then, inspired by the precision-recall curves [34–36], we propose a metric, named precision-to-noise ratio (PNR), to characterize the ability to distinguish potential links for different scores. PNR describes the local precision of a given set of links with the same score. Based on PNR, a novel framework, which projects one-variable function to adjust the scores of a given method, is put forward. We argue that the framework finds the optimal transforming function that exploits the full capacities of traditional link prediction methods and improves their performance both on precision and on the detection of long-range links. Experiments in six real-world networks demonstrate the effectiveness of our method.

The rest of the paper is organized as follows. In Section 2, we first brief the link prediction problem and then introduce our proposed method. In Section 3, we compare the performances of our method and the classical methods. Finally, the conclusion is given.

2. Materials and Methods

We give the link prediction formulism in Section 2.1 and the baseline method in Section 2.2. Our proposed framework is introduced in Section 2.3.

2.1. Network Formation and Metrics. Given a network $A = (a_{ij})_{N\times N}$, $E = \{(i, j) \mid a_{ij} \neq 0\}$ with $a_{ij} = 1$ if node i connects to j; otherwise, $a_{ij} = 0$. When evaluating the prediction performance, we usually divide the links randomly into $1-p^H$ training set E^T and p^H probe set E^P ($p^H \in (0, 1)$), with $E^T \bigcap E^P = \emptyset$ and $E^T \bigcup E^P = E$. The goal is to accurately predict the links in probe set only by using the information in training set.

We first assign a score to each nonexisting link and then choose links with the highest top-L scores as potential ones. State-of-the-art similarity evaluation methods could be utilized to carry out link prediction, including common neighbors (CN), Jaccard index (JB), resource allocation index (RA), local path index (LP), and structural perturbation method (SPM) (see the part of *Baseline* and [38]).

There are two popular metrics to characterize the accuracy: area under the receiver operating characteristic curve (AUC) [39] and the precision [40, 41]. AUC can be interpreted as the probability that a randomly chosen missing link (i.e., a link in E^P) has a higher score than a randomly chosen nonexisting link. Then, AUC requires n times of independent comparisons. We randomly choose a real link and a nonexisting link to compare their scores. After n different comparisons, we record n_1 times where real links have higher scores, and n_2 times where the two kinds of links have the same score. The final AUC is calculated as

$$\text{AUC} = \frac{n_1 + 0.5 \times n_2}{n}. \tag{1}$$

If all the scores are given by an independent and identical distribution, then AUC should be around 0.5. A higher AUC is corresponding to a more accurate prediction.

Another metric is precision that characterizes the ratio of correctly predicted links for a given prediction list. That is to say, if the length of prediction list is L, among which L_r links are the right potential links, then the precision is

$$p = \frac{L_r}{L}. \tag{2}$$

Clearly, higher precision means higher prediction accuracy. Intuitively, higher accuracy means higher AUC and higher precision. In the experiments, we will see that precision has little correlation with AUC and that improving the precision may not result in the improvement of AUC.

2.2. Baseline Prediction Methods. There exists a large number of score-assigning approaches in link prediction problem. All these methods could be introduced into our framework. Though we only investigate some state-of-the-art score-assigning approaches, the results and conclusions are also applicable for other score-assigning methods. The five score-assigning approaches [6, 16] are as follows.

(i) Common Neighbor (CN). The metric supposes that if two nodes i and j have more common neighbors, they are more likely to connect together. The neighborhood overlap of the two nodes is as follows:

$$s_{ij}^{\text{CN}} = \left|\Gamma(i) \bigcap \Gamma(j)\right|, \tag{3}$$

where $\Gamma(i)$ is the neighbor set of node i and $|\cdots|$ indicates the size of a set. The drawback of CN is that it favors large-degree nodes. Though the similarity of two large-degree nodes is low, they still have many common neighbors.

(ii) Jaccard Coefficient (JB). Jaccard is a conventional similarity metric that aims to suppress the influence of large-degree nodes, which is

$$s_{ij}^{\text{Jaccard}} = \frac{\left|\Gamma(i) \bigcap \Gamma(j)\right|}{\left|\Gamma(i) \bigcup \Gamma(j)\right|}. \tag{4}$$

Since the similarity is normalized by the size of the union set of the two nodes' neighbors, low similarity still exists between two large-degree nodes even though they may have many common neighbors.

(iii) Resource Allocation (RA). This index is inspired by the resource allocation dynamics in complex networks. Given a pair of unconnected nodes i and j, suppose that the node i needs to allocate some resource to j, using common neighbors as transmitters. Each transmitter (common neighbor) starts with a single unit of resource and then distributes it equally among all its neighbors. The similarity between i and j can be calculated as the amount of resource received from their common neighbors:

$$s_{ij}^{\mathrm{RA}} = \sum_{z \in \Gamma(i) \bigcap \Gamma(j)} \frac{1}{k_z}. \tag{5}$$

Comparing with Jaccard method, RA could also suppress the influence of large-degree nodes, but more specifically. Different neighbors contribute to the similarity differently. If two nodes prefer to connect low-degree nodes, it means that they have a higher probability to share common interests or characteristics. However, many pair-nodes have common high-degree neighborhoods, resulting in that high-degree nodes play a weak role when evaluating similarity. Based on the idea, Adamic-Adar (AA) index is obtained by using $\log(k_z)$ instead of k_z in (5).

(iv) Local Path (LP). CN considers the intersection of neighborhoods, which actually utilizes the one-path neighbors to characterize similarity. LP takes a general consideration of paths by considering two-path neighbors:

$$s^{\mathrm{RA}} = A^2 + \epsilon A^3, \tag{6}$$

where A is the adjacent matrix of a network and ϵ is a small positive number. LP supposes that one-path neighbors contribute more to the similarity than two-path neighbors. LP is the low order parts of Katz method ($s^{\mathrm{Katz}} = A^2 + \epsilon A^3 + \epsilon^2 A^4 + \cdots$), but with much lower computing complexity.

(v) Structural Perturbation Method (SPM). Lü et al. [6] suppose that network structure follows consistency after some random perturbation. In SPM, training set A^T is divided into a small fraction of perturbation set ΔA and the remaining set A^R ($A^T = A^R + \Delta A$). A^T has similar eigenvectors with A^R, but different eigenvalues. For the kth largest eigenvalues of A^T and A^R,

$$\Delta\lambda_k = \lambda_k\left(A^T\right) - \lambda_k\left(A^R\right) \approx \frac{x_k^T \Delta A x_k}{x_k^T x_k}, \tag{7}$$

where x_k is the eigenvector of A^R, corresponding to $\lambda_k(A^R)$. The similarity matrix $s = (s_{ij})_{N\times N}$ is

$$s = \sum_{k=1}^{N} \left(\lambda_k + \Delta\lambda_k\right) x_k x_k^T. \tag{8}$$

SPM first divides a network into training set and probe set and further divides the training set into perturbation set and the remaining set. For a given division of training and probe set, we calculate the average of 10 times independent simulations of (8) as the similarity matrix.

Apart from the five similarity metrics introduced above, for more similarity-evaluating methods, please refer to [42, 43].

2.3. The Proposed Method. We start our framework by reinvestigating the definition of precision. Supposing that s_{ij} is the similarity score of nodes i and j obtained by a prediction method f only based on training set E^T, $p_r(s)$ is the similarity distribution that a randomly chosen existing link in training set has score s, and $p_n(s)$ is the similarity distribution that a randomly chosen nonexisting link in the training set has score s. Due to random division of training set and probe set, links in the probe set should have the same similarity distribution with that of the training set at high confidence according to the law of large numbers [44, 45]. Thus we would not differentiate similarity distribution of existing links in the training and probe sets in the following paper. The assumption is reasonable according to the statistical theory if the size of samples goes to infinity [44, 45]. Since classical methods only predict links with high scores, the estimated precision of the method f is written as

$$p_f^0 = \frac{\left|E^P\right| \int_{c_0}^{s_{\max}} p_r(s)\, ds}{\left|U - E^T\right| \int_{c_0}^{s_{\max}} p_n(s)\, ds}, \tag{9}$$

where $|E^P|$ is the size of E^P, c_0 is a constant, and U is the whole set of all possible links ($|U| = (1/2)N(N-1)$). $s_{\max}$ is the maximum score. In real scenarios, the length of the prediction list is usually the size of the probe set [16], which requires c_0 subjecting to $|U - E^T| \int_{c_0}^{s_{\max}} p_n(s)ds = |E^P|$. If $p_r(s) \ll p_n(s)$ at $s > c_0$, the precision $p_f \to 0$. Otherwise, $p_r(s) \gg p_n(s)$ gives rise to a high precision. Since only links with top-L highest scores are predicted as potential links, precision could be calculated by (2) [6, 16]. Equation (2) is a much easier formula to describe precision than (9).

Most previous link prediction methods only predict links with high similarity scores. We generalize (9) by considering links of different similarities. Supposing that links with scores $s_{ij} \in S = (s_1, s_2) \bigcup (s_3, s_4) \cdots \bigcup (s_{2m-1}, s_{2m})$ are predicted as potential links, the precision is as follows:

$$p_f = \frac{\left|E^P\right| \int_{S=(s_1,s_2)\bigcup(s_3,s_4)\cdots\bigcup(s_{2m-1},s_{2m})} p_r(s)\, ds}{\left|U - E^T\right| \int_{S=(s_1,s_2)\bigcup(s_3,s_4)\cdots\bigcup(s_{2m-1},s_{2m})} p_n(s)\, ds}, \tag{10}$$

where $s_1 < s_2 < s_3 < \cdots s_{2m-1} < s_{2m}$. To confine the length of the prediction list, a precondition requires $|U - E^T| \int_S p_n(s)ds = |E^P|$. Note that, in most previous works, $S = (c_0, s_{\max})$, and equation (10) reduces to (9). Our generalized precision equation (10) considers links with both high and low scores.

The main concern is to select appropriate set S in (10) to maximize the precision. We propose precision-to-noise ratio (PNR) to determine S,

$$\mathrm{PNR}(s) = \frac{p_r(s)}{p_n(s)}, \tag{11}$$

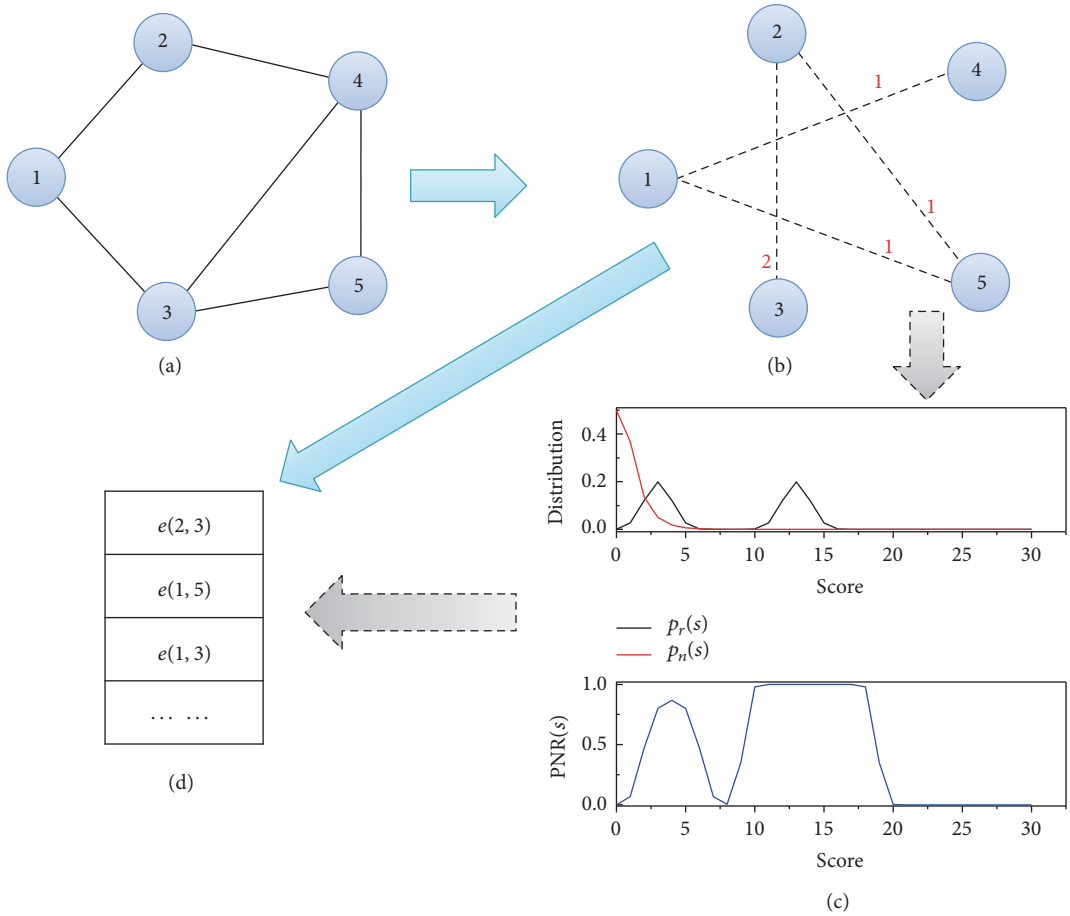

FIGURE 1: Schematic shows the proposed framework based on CN. (a) A snapshot of a large network. (b) Score of nonexisting links calculated by CN method. (c) The top panel is the score distributions of existing and nonexisting links, $p_r(s)$ and $p_n(s)$. The bottom panel is PNR(s) = $p_r(s)/p_n(s)$. (d) Predicted links. State-of-the-art prediction methods follow the path (a)→(b)→(d), while our proposed framework follows the path (a)→(b)→(c)→(d), which has an additional path PNR(s).

where PNR(s) measures the ability to distinguish real links with the same score. Note that a nonexisting link in training set may be an existing link in probe set. Given a nonexisting link in training set with the similarity s_{ij}, the probability that it is an existing link in probe set (i.e., the precision) is $p' = (|E^P| \cdot p_r(s))/(|U - E^T| \cdot p_n(s)) = \alpha \mathrm{PNR}(s)$, where $\alpha = |E^P|/|U - E^T|$ is a constant.

The central issue of our framework is to use PNR(s) to determine the optimal score set S. We first calculate the similarity scores of all links only based on training set by a traditional method. Second, $p_r(s)$, $p_n(s)$, and PNR(s) are computed. Third, we reassign the scores of each link $s'_{ij} = \mathrm{PNR}(s_{ij})$, where s_{ij} is the original similarity score by the first step. Finally, we sort links in the descending order of s' and links with top-L scores are predicted as potential links [16, 18]. The optimal score set S_{opt} corresponds to the original similarity scores whose reassigned scores rank in the top-L score list.

Different kinds of similarity evaluations could be introduced into the framework. Taking CN similarity method as an example, our framework is as follows:

(1) Divide the links of a network into $1 - p^H$ training set and p^H probe set randomly.

(2) Calculate the similarity scores of all existing and nonexisting links by CN method only according to training set.

(3) Calculate PNR(s). Divide the scores into K uniform bins and count how many existing ($n_{e,i}$) and nonexisting ($n_{n,i}$) links locate in each bin i (i.e., calculate discrete $p_r(s)$, $p_n(s)$). Then we obtain $\mathrm{PNR}(s_k) = p_r(s_k)/p_n(s_k)$, $k = 1, 2, \ldots, K$. Note that if $p_n(s_k) = 0$, we define $\mathrm{PNR}(s_k) = 0$.

(4) Obtain the readjusting scores of the nonexisting links in training set by $s' = \mathrm{PNR}(s)$.

(5) Determine the prediction list by choosing links with $L = |E^P|$ highest scores s', and calculate the precision.

Figure 1 depicts the proposed framework based on CN method. After obtaining the similarity scores of links (Figure 1(a)→1(b)), traditional CN method directly predicts potential links according to the scores (Figure 1(b)→1(d)), while the proposed framework calculates PNR(s) (Figures

Table 1: Structural properties of the different real networks. Structural properties include network size (N), link number (E), degree heterogeneity ($H = \langle k^2\rangle/\langle k\rangle^2$), degree assortativity (r), average clustering coefficient ($\langle C\rangle$), average shortest path length ($\langle d\rangle$), and sparsity.

Network	N	E	H	r	$\langle C\rangle$	$\langle d\rangle$	Sparsity
Email	33696	180811	6.070	−0.060	0.170	4.08	3.2×10^{-4}
PDZBase	161	209	2.263	−0.466	0.001	5.11	1.6×10^{-2}
Euroad	1039	1305	1.228	0.090	0.004	18.39	2.4×10^{-3}
Neural	297	2148	1.81	−0.163	0.292	2.46	4.9×10^{-2}
Roundworm	453	2025	4.485	−0.226	0.647	2.66	2.0×10^{-2}
USair	332	2126	3.464	−0.208	0.625	2.74	3.9×10^{-2}

1(b)→1(c)) and later predicts potential links according to the modified scores (Figures 1(c)→1(d)).

An important property of our framework is that if S is determined according to PNR(s), that is, PNR(x) > PNR(y), $\forall x \in S$, $\forall y \in \mathbb{R} - S$, the precision p_L could exploit full capacity of a given similarity-evaluating method. PNR(s) is the optimal transforming function $f_{\text{opt}}(s) = \text{PNR}(s)$. It means that no matter how we transform the similarity by other one-variable function, $s'' = f'(s)$, the precision performance of s'' cannot outperform the proposed method by PNR(s). For the proof of the optimal PNR(s), please see part I in the supplementary materials.

3. Experimental Results

We first describe the six real networks in Section 3.1. The precision comparison between our method and the baseline methods is given in Section 3.2. Finally, the characteristics of the predicted links by different methods are investigated in Section 3.3.

3.1. Datasets. To verify the effectiveness of the proposed method, we measure the performance of our framework in six empirical networks from diverse disciplines and backgrounds: (1) email [46]: Enron email communication network covers all the email communication within a dataset of around half million emails; nodes of the network are email addresses and if an address i sent at least one email to address j, the graph contains an undirected link from i to j; (2) PDZBase [47]: an undirected network of protein-protein interactions from PDZBase; (3) Euroad [48]: international E-road network that locates mostly in Europe; the network is undirected, with nodes representing cities and links denoting e-road between two cities; (4) neural [49]: a directed and weighted neural network of *C. elegans*; (5) USair [6]: an directed network of flights between US airports in 2010; each link represents a connection from one airport to another in 2010; (6) roundworm [49]: a metabolic network of *C. elegans*.

Different real networks contain directed or undirected, weighted or unweighted links. To simplify the problem, we treat all links undirected and unweighted. Besides, only the giant connected components of these networks are taken into account. This is because for a pair of nodes located in two disconnected components, their similarity score will be zero according to most prediction methods. Table 1 shows the basic statistics of those networks.

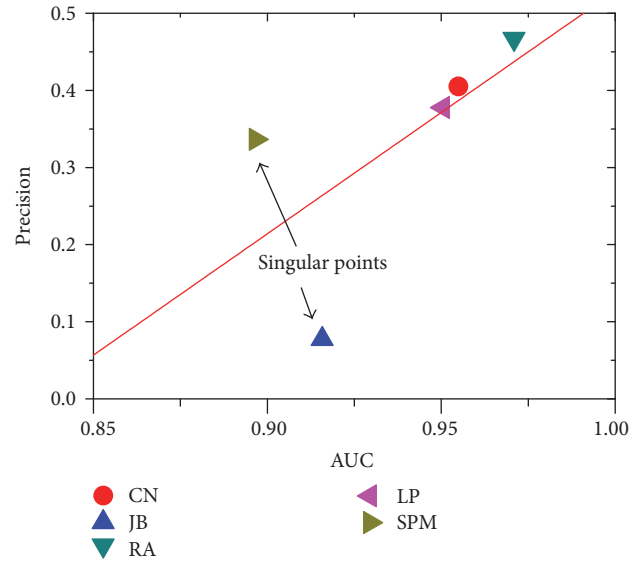

Figure 2: AUC and precision of the USair network obtained by five different approaches: common neighbors (CN), Jaccard index (JB), resource allocation index (RA), local path index (LP), and structural perturbation method (SPM). The results are obtained by 50 independent simulations. SPM method achieves high precision, yet low AUC, and JB has low precision, but high AUC (>0.9).

3.2. Precision Evaluation. In the experiments, we set $p^H = 10\%$ that means the networks are randomly divided into 90% training set and 10% probe set. All the experiments are the average of 50 independent simulations.

Figure 2 shows AUC and precision of five different methods in USair network. In Figure 2, CN method achieves low AUC, yet high precision, whereas RA method achieves similar AUC with methods of CN, JB, and SPM, but much lower precision. Apart from USair network, the deviation between AUC and precision also exists in other real-world networks (see FIG. S1 in the supplementary materials). The main reason is that AUC characterizes the score difference between existing and nonexisting links in the whole networks, whereas precision only counts the links with top-L high scores. Specifically, from the perspective of score distributions, $\text{AUC} = \int_{-\infty}^{+\infty} p_r(x) \int_{-\infty}^{x} p_n(y)dx\,dy$. Comparing with (10), the definitions of the two metrics are completely different, resulting in little correlation between them.

Figure 3 shows PNR and the score distributions of existing and nonexisting links for USair network by CN method.

Table 2: Maximal precision comparison of the proposed methods and traditional high-similarity methods for six real-world networks. Traditional precision is obtained by the maximum of traditional methods, that is, max{CN, Jaccard, RA, LP, SPM}. Proposed precision is obtained by our framework, that is, max{PNR_{CN}, $PNR_{Jaccard}$, PNR_{RA}, PNR_{LP}, PNR_{SPM}}.

	Email	PDZBase	Euroad	Neural	Roundworm	USair
Traditional p_{max}	0.0171	0.0032	0.0052	0.0107	0.2651	0.4670
Proposed p_{max}	**0.0313**	**0.3159**	**0.0674**	**0.0392**	**0.3475**	**0.5309**

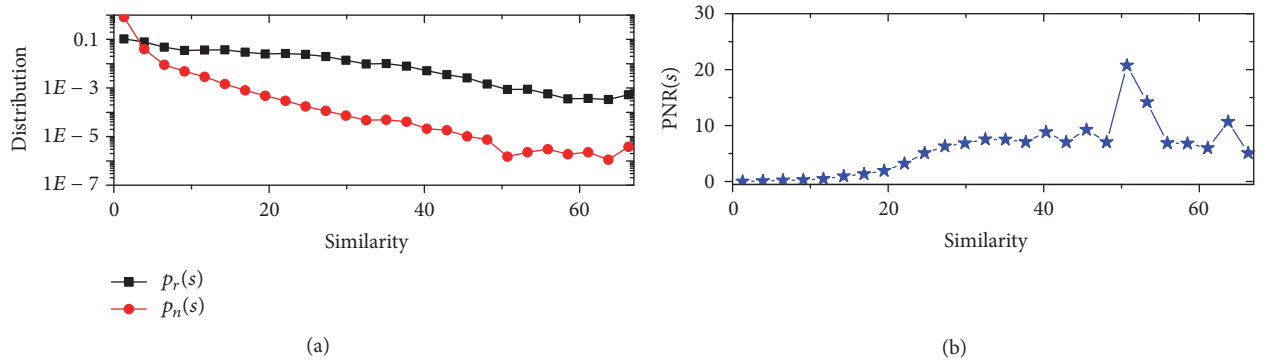

Figure 3: Similarity distributions and the corresponding PNR(s) of USair network, where the similarity is obtained by CN method. (a) Similarity distributions of the existing and nonexisting links, $p_r(s)$ and $p_n(s)$, respectively. (b) PNR(s) as a function of similarity in USair network.

In Figure 3(a), the scores of existing and nonexisting links follow power law distribution largely. High scores sometimes correspond to low PNR, especially at Similarity $\approx$ 60 (see Figure 3(b)). Nevertheless, some low scores achieve high PNR, indicating that for a nonexisting link in training set with this particular score, the link is likely to be an existing link in probe set. For a nonexisting link in training set with high score, yet with low PNR, it has a high probability not to be an existing link in probe set. The similar phenomenon also exists in other networks (see FIG. S2 in the supplementary materials). In consequence, the foundation of traditional methods, which suppose that similar nodes have a high likelihood to form links, is confronted with great challenges in precisely predicting links of low similarities.

Figure 6 shows the precision difference between the proposed PNR methods and the baseline methods. Our proposed method enhances precision remarkably compared with the original methods in most cases. Some fluctuation exists in these methods, due to the limited size of networks. Table 2 gives the maximal precision increasement in the six networks. In Table 2, precision is obtained by the maximum of traditional methods and PNR methods, respectively, that is, max{CN, Jaccard, RA, LP, SPM} and max{PNR_{CN}, $PNR_{Jaccard}$, PNR_{RA}, PNR_{LP}, PNR_{SPM}}. Our method outperforms state-of-the-art methods in the six networks. Besides, Figure 4 shows the influence of the probe set size on the precision performance. We find that our method outperforms classical methods when $p^H > 0.85$, except for JB method when $p^H > 0.6$. Other networks have similar results (see FIG. S3 in the supplementary materials). However, according to the theoretical analysis (see the first part in the supplementary materials), our method should perform better than, or at least equally to, the classical

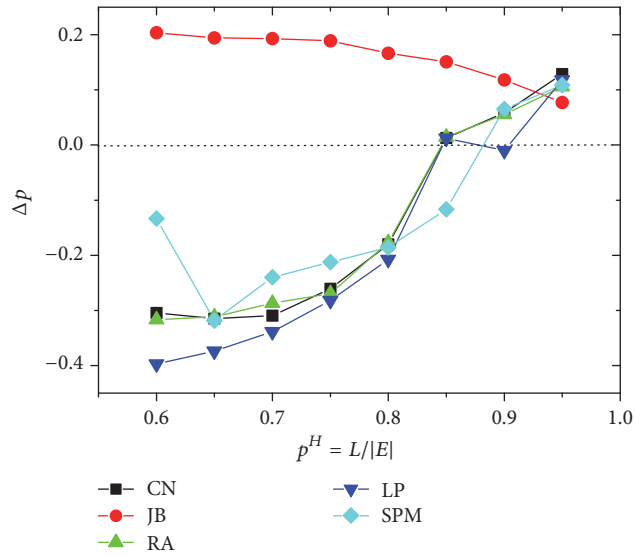

Figure 4: The precision difference Δp as a function of probe set size $p^H = L/|E|$ in Usair network. $\Delta p = p_{PNR} - p_{original}$. $\Delta p > 0$ means that our method outperforms the original methods. In the panel, when $p^H > 0.85$, $\Delta p > 0$.

methods. The reason is that we suppose the network structure is not influenced by the random division of training and probe set. Thus, the training subnetwork should have similar structure with the original networks. The assumption is rational when p^H is small. If the size p^H of the probe set is large, the training sets have many differences with the entire networks, which violates the assumption of our method. Therefore, our method performs well when the fraction of the probe set is small.

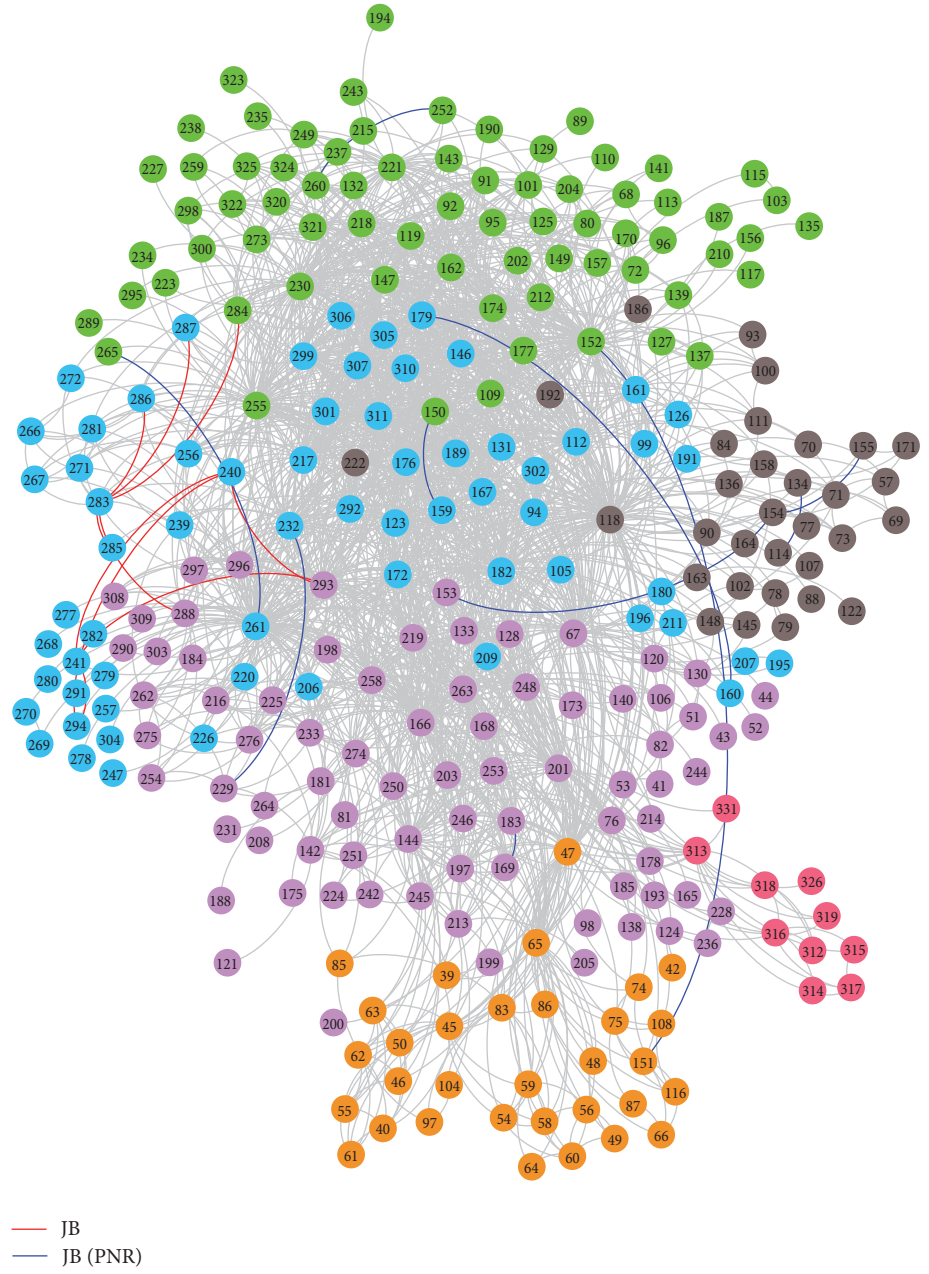

FIGURE 5: The comparison of the predicted edges between JB and the corresponding PNR methods in the Usair network. In the panel, we predict 10 edges for both JB and PNR_{JB} methods. The Usair network is divided into different communities by the method in [37]. Nodes in the same community have the same color and short geographical distances. Our method (blue lines) predicts more edges between faraway nodes in different communities, while the original JB method (red lines) only predicts edges between close nodes.

3.3. Characteristics of the Predicted Links. Long-range links play an important role in the dynamics of networks and it is of much significance to predict long-range links [32, 50]. Figure 5 gives a comparison of the predicted links between JB and the corresponding PNR methods in the Usair network. In Figure 5, our method predicts more links between faraway nodes in different communities, while the original JB method only predicts links between close nodes. Community detection method in [37] is utilized in Figure 5. However, it is difficult to evaluate long-range links solely based on community divisions. Since long-range links usually have long distances and low similarities, we would investigate the average distance and average similarity of the predicted links by our proposed framework.

The distance d_{ij} of a link e_{ij} is the shortest distance between nodes i and j only based on training set. Since the endpoints of the predicted links do not connect directly, $d_{ij} \geq 2$. The average distance of the predicted links is

$$\overline{d}_{\text{predict}} = \frac{1}{L} \sum_{e_{ij} \in \text{predcited links}} d_{ij}. \tag{12}$$

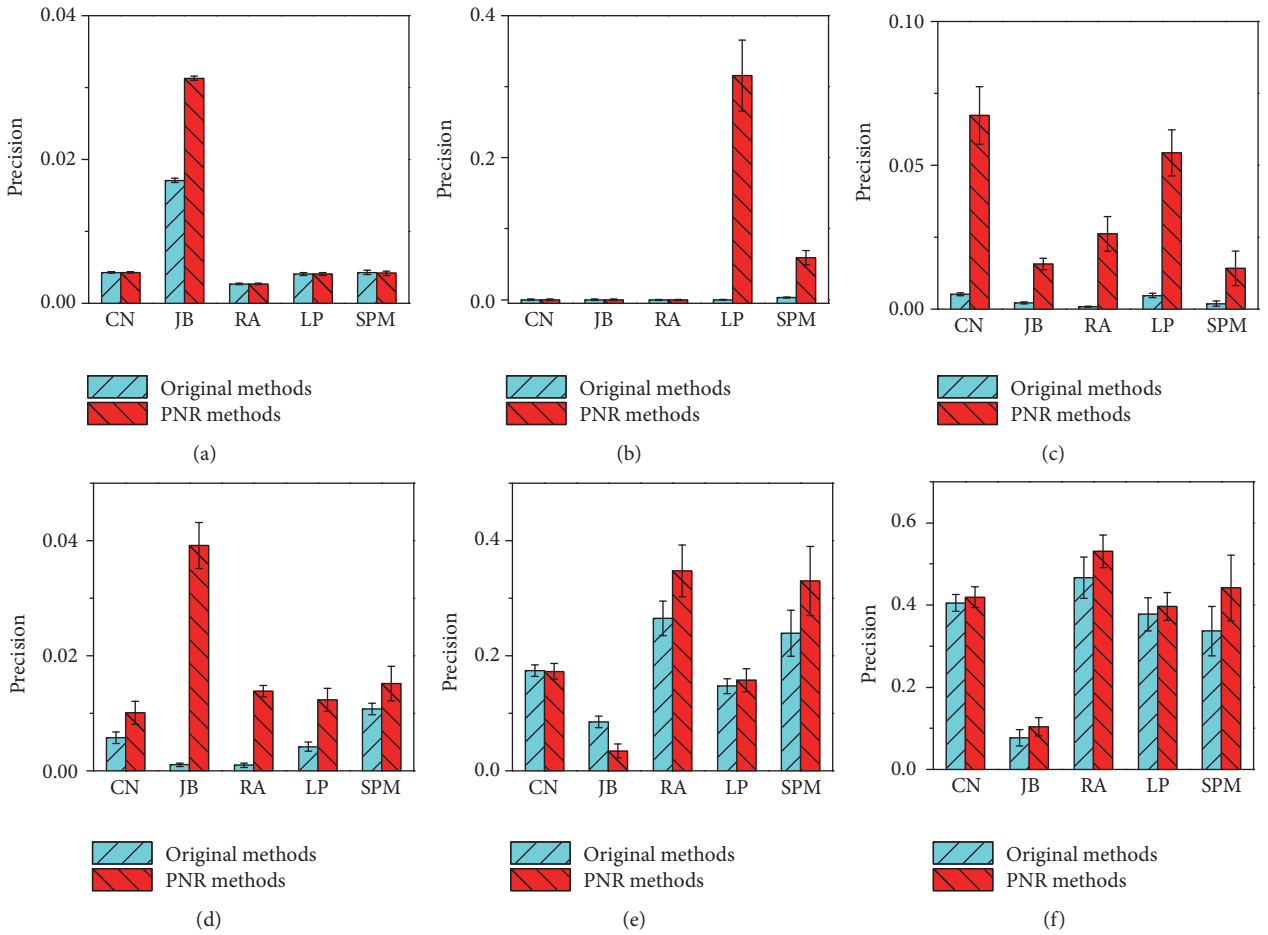

FIGURE 6: Precision comparison of the proposed methods (red) and traditional high similarity based methods (cyan) for six real-world networks. (a) Email network. (b) PDZBase network. (c) Euroad network. (d) Neural network. (e) Roundworm network. (f) USair network. Results are the average of 50 independent simulations. Our proposed framework increases precision in most cases.

Analogously, the average similarity of the predicted links is

$$\bar{s}_{\text{predict}} = \frac{1}{L} \sum_{e_{ij} \in \text{predcited links}} s_{ij}, \tag{13}$$

where s_{ij} is the similarity of nodes i and j in training set.

Figure 7 shows the difference of the average distances obtained by PNR method and the corresponding original methods. Generally, PNR method achieves a higher average distance than the corresponding original methods in the six networks, especially for SPM in Email network and LP in USair network, whereas for many cases, PNR and the original methods have the same average distance $\bar{d} = 2$. It is because that the distance of most unconnected nodes are 2, revealing that most commonly used methods incline to predict triangle edges. Therefore, our method has little influence on the average distance. However for some sparser networks, such as neural and USair networks, the average distance is improved by our framework, especially for LP in USair network. Previous works show that the two endpoints of a long-range link usually have a high distance or low similarity. Since PNR framework could increase the average distance of the predicted links, it can be conjectured that more long-range links are predicted. Besides, integrating Figures 6 and 7, we can find that our framework predict more long-range links correctly.

Furthermore, Figure 8 shows the difference of average similarity obtained by PNR method and the corresponding original methods. In Figure 8, PNR method achieves a lower average similarity than the corresponding original methods in the six networks, except RA method in roundworm network. The reason is that PNR has much fluctuations due to the limited size of networks, bringing about the unusual phenomenon of RA in roundworm network. Similar to the analysis of average distance, we show that PNR methods are beneficial to the prediction of long-range links, which agrees with the conclusion from Figure 7.

4. Conclusion

In summary, we systematically study the drawbacks of similarity-based link prediction methods and show that some link prediction methods achieve high AUC, yet low precision. Based on the differences of the similarity distributions of existing and nonexisting links, we propose a metric (PNR) to explain the problem of high AUC and low precision.

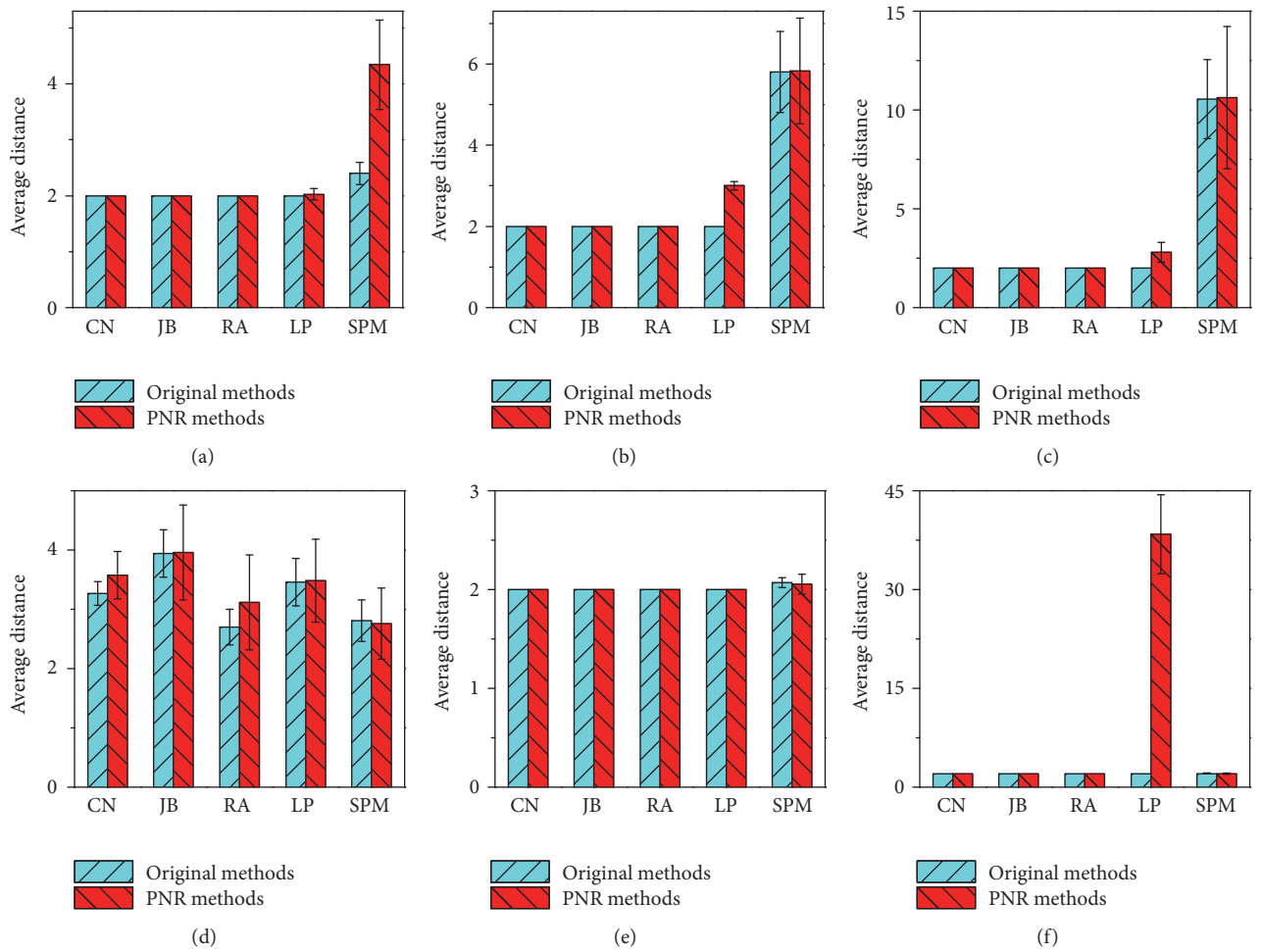

Figure 7: Comparison of average distance of the PNR predicted links with that of the corresponding original methods for different networks. (a) Email network. (b) PDZBase network. (c) Euroad network. (d) Neural network. (e) Roundworm network. (f) USair network. Results are the average of 50 independent simulations. Our proposed framework increases the average distance on the whole, which indicates that more long-range links are predicted correctly.

Two nodes with some particular low scores also have a high likelihood of forming links between them. Furthermore, we prove that PNR is the optimal one-variable function to adjust the likelihood scores of links. Experiments in real networks demonstrate the effectiveness of PNR, and the precision is greatly enhanced. Additionally, the proposed framework could also reduce the average similarity and increase the average distance of the predicted links, which indicates that more missing long-range links can be detected correctly.

Though the proposed approach investigates link prediction in unipartite networks, it could also be generalized to bipartite and other kinds of networks. What is more, our method provides a novel way to explore the connecting patterns of real networks that may inspire other better score-assigning methods in the future.

Conflicts of Interest

The authors declare no competing financial interests.

Acknowledgments

The authors thank Dr. Alexandre Vidmer for his fruitful discussion and comments. This work is jointly supported by the National Natural Science Foundation of China (61703281, 11547040), the Ph.D. Start-Up Fund of Natural Science Foundation of Guangdong Province, China (2017A030310374 and 2016A030313036), the Science and Technology Innovation Commission of Shenzhen (JCYJ20160520162743717, JCYJ20150625101524056, JCYJ20140418095735561, JCYJ20150731160834611, JCYJ20150324140036842, and SGLH20131010163759789), Shenzhen Science and Technology Foundation (JCYJ20150529164656096, JCYJ20170302153955969), the Young Teachers Start-Up Fund of Natural Science Foundation of Shenzhen University, and Tencent Open Research Fund.

Supplementary Materials

In the supplementary materials, we prove that PNR(x) is the optimal transferring function in Section 1. The deviation of AUC and precision in different networks is shown in

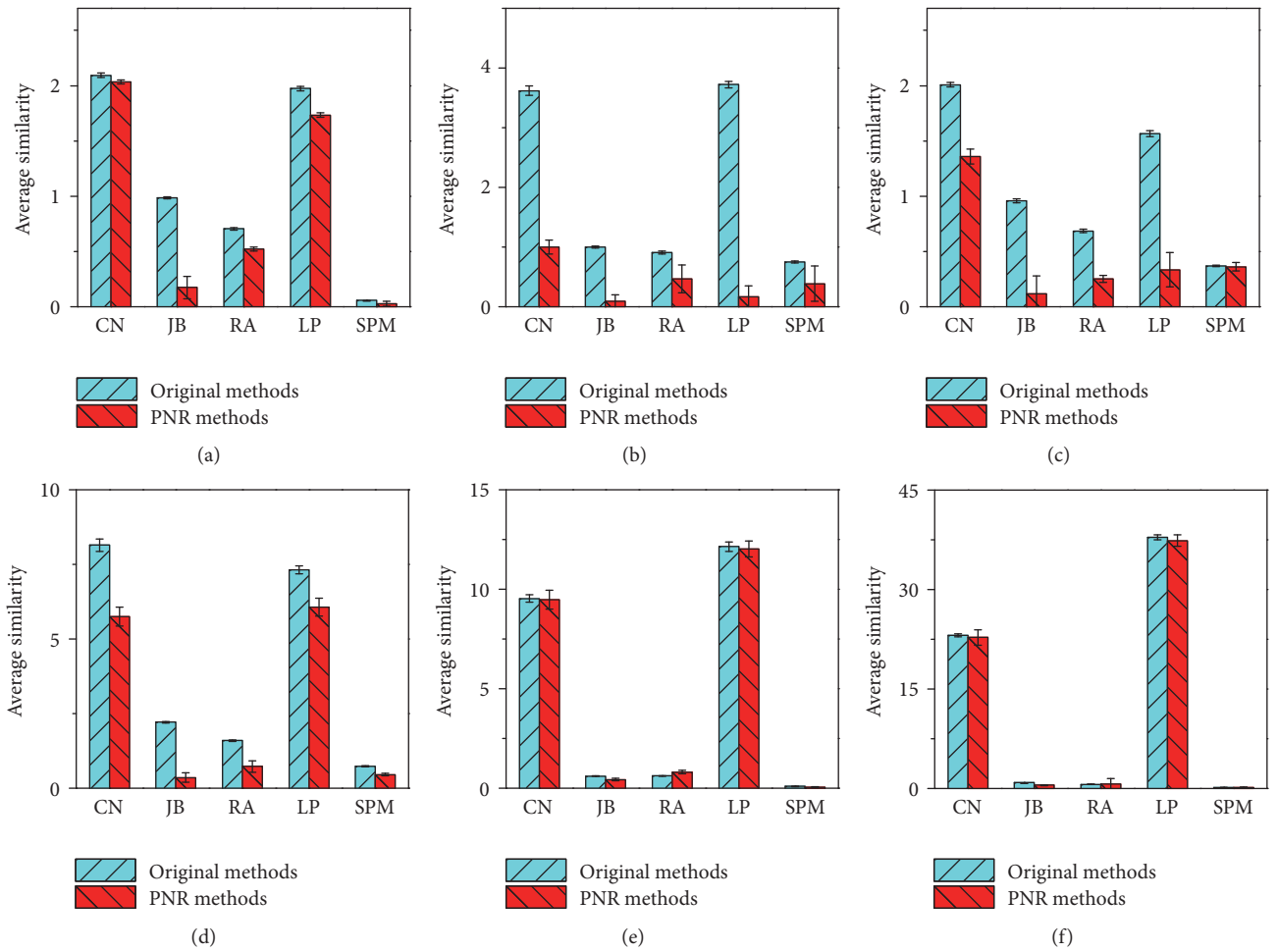

Figure 8: Comparison of average similarity of the PNR predicted is linked with that of the corresponding original methods for different networks. (a) Email network. (b) PDZBase network. (c) Euroad network. (d) Neural network. (e) Roundworm network. (f) USair network. Results are the average of 50 independent simulations. Our proposed framework reduces the average similarity on the whole, which indicates that more long-range links are predicted correctly.

Section 2. The PNR performances of different methods in different networks are shown in Section 3. In Section 3, we first plot the PNR(x) by different methods in FIG. S2 and then show the influence of the probe set size on the precision in Fig. S3. FIG. S1 (color online), AUC and precision of six real-world networks (see Table 2) by five different popular approaches. Results are average of 50 independent simulations. In the experiments, $p^H = 0.1$ means that we utilize 90% existing edges as training set to predict the other 10% edges (probe set). FIG. S2 (color online), PNR for six networks by five different methods. (a) Email network. (b) PDZBase network. (c) Euroad network. (d) Neural network. (e) Roundworm network. (f) USair network. Results are the average of 50 independent simulations and are obtained only according to training set. For different methods and different networks, scores are normalized to 0~1 with $s_{new} = (s - s_{min})/(s_{max} - s_{min})$. FIG. S3 (Color online), the precision difference Δp as a function of probe set size $p^H = L/|E|$ in the four networks, where Δp is the difference between the five classical and the corresponding PNR methods, $\Delta p = p_{PNR} - p_{original}$. $\Delta p > 0$ means that our method outperforms the original methods. In the panels, when $p^H > 0.85$, $\Delta p > 0$. *(Supplementary Materials)*

References

[1] J. Esquivel-Gómez, R. E. Balderas-Navarro, P. D. Arjona-Villicaña, P. Castillo-Castillo, O. Rico-Trejo, and J. Acosta-Elias, "On the emergence of islands in complex networks," *Complexity*, 2017.

[2] W. X. Wang, Y. C. Lai, and C. Grebogi, "Data based identification and prediction of nonlinear and complex dynamical systems," *Physics Reports*, vol. 644, pp. 1–76, 2016.

[3] A.-L. Barabási, "Network science," *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, vol. 371, no. 1987, article no. 0375, 2013.

[4] T. G. Lewis, *Network science: Theory and applications*, vol. 8, John Wiley and Sons, 2011.

[5] G. Chen, R. Mao, and K. Lu, "A parallel computing framework for big data," *Frontiers of Computer Science*, vol. 11, no. 4, pp. 608–621, 2017.

[6] L. Lü, L. Pan, T. Zhou, Y. C. Zhang, and H. E. Stanley, "Toward link predictability of complex networks," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 112, no. 8, pp. 2325–2330, 2015.

[7] T. Wang, M. Y. Zhou, and Z. Q. Fu, "Link prediction in evolving networks based on the popularity of nodes".

[8] L. Sharma and A. Gera, "A survey of recommendation system: Research challenges," *International Journal of Engineering Trends and Technology (IJETT)*, vol. 4, no. 5, pp. 1989–1992, 2013.

[9] J. Lu, D. S. Wu, M. S. Mao, W. Wang, and G. Zhang, "Recommender system application developments: a survey," *Decision Support Systems*, vol. 74, pp. 12–32, 2015.

[10] W.-X. Wang, Y.-C. Lai, C. Grebogi, and J. Ye, "Network Reconstruction Based on Evolutionary-Game Data via Compressive Sensing," *Physical Review X*, vol. 1, no. 2, Article ID 021021, pp. 1–7, 2011.

[11] Z. Shen, W.-X. Wang, Y. Fan, Z. Di, and Y.-C. Lai, "Reconstructing propagation networks with natural diversity and identifying hidden sources," *Nature Communications*, vol. 5, article no. 5323, 2014.

[12] H. Liao, A. Zeng, M. Zhou, R. Mao, and B. Wang, "Information mining in weighted complex networks with nonlinear rating projection," *Communications in Nonlinear Science and Numerical Simulation*, vol. 51, pp. 115–123, 2017.

[13] Y. Wang, J. Wang, H. Liao, and H. Chen, "An efficient semi-supervised representatives feature selection algorithm based on information theory," *Pattern Recognition*, vol. 61, pp. 511–523, 2017.

[14] A. L. Barabási, *Network Science*, vol. 4, Cambridge University Press, 2016.

[15] R. Mao, P. Zhang, X. Li, X. Liu, and M. Lu, "Pivot selection for metric-space indexing," *International Journal of Machine Learning and Cybernetics*, vol. 7, no. 2, pp. 311–323, 2016.

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

[17] A. Zeng and G. Cimini, "Removing spurious interactions in complex networks," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 85, no. 3, Article ID 036101, 2012.

[18] V. Martínez, F. Berzal, and J.-C. Cubero, "A survey of link prediction in complex networks," *ACM Computing Surveys*, vol. 49, no. 4, article no. 69, 2016.

[19] A. Javari and M. Jalili, "A probabilistic model to resolve diversity–accuracy challenge of recommendation systems," *Knowledge and Information Systems*, vol. 44, no. 3, pp. 609–627, 2015.

[20] A. Javari and M. Jalili, "Accurate and novel recommendations: An algorithm based on popularity forecasting," *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 4, 2015.

[21] P. Zhang, X. Wang, F. Wang, A. Zeng, and J. Xiao, "Measuring the robustness of link prediction algorithms under noisy environment," *Scientific Reports*, vol. 6, Article ID 18881, 2016.

[22] Z. Huang, X. Li, and H. Chen, "Link prediction approach to collaborative filtering," in *Proceedings of the the 5th ACM/IEEE-CS joint conference*, p. 141, Denver, CO, USA, June 2005.

[23] D. K. Lin, "An information-theoretic definition of similarity," in *Proceedings of the In 15th International Conference on Machine Learning (ICML '15)*, vol. 98, pp. 296–304, 1998.

[24] E. A. Leicht, P. Holme, and M. E. J. Newman, "Vertex similarity in networks," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 73, no. 2, Article ID 026120, 2006.

[25] Z. Liu, Q.-M. Zhang, L. Lü, and T. Zhou, "Link prediction in complex networks: a local naïve Bayes model," *EPL (Europhysics Letters)*, vol. 96, no. 4, Article ID 48007, 2011.

[26] D. Heckerman, C. Meek, and D. Koller, "Probabilistic entity-relationship models, prms, and plate models," in *Introduction to Statistical Relational Learning*, pp. 201–238, 2007.

[27] K. Yu, W. Chu, S. Yu, V. Tresp, and Z. Xu, "Stochastic relational models for discriminative link prediction," in *Proceedings of the 20th Annual Conference on Neural Information Processing Systems(NIPS '06)*, pp. 1553–1560, December 2006.

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

[29] C. Zhou, L. Zemanová, G. Zamora, C. C. Hilgetag, and J. Kurths, "Hierarchical organization unveiled by functional connectivity in complex brain networks," *Physical Review Letters*, vol. 97, no. 23, Article ID 238103, 2006.

[30] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, "On facebook, most ties are weak," *Communications of the ACM*, vol. 57, no. 11, pp. 78–84, 2014.

[31] S. Aral, "The future of weak ties," *American Journal of Sociology*, vol. 121, no. 6, pp. 1931–1939, 2016.

[32] X. Wang, W. Lu, M. Ester, C. Wang, and C. Chen, "Social recommendation with strong and weak ties," in *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016*, pp. 5–14, USA, October 2016.

[33] A. L. Kavanaugh, D. D. Reese, J. M. Carroll, and M. B. Rosson, "Weak ties in networked communities," *Information Society*, vol. 21, no. 2, pp. 119–131, 2005.

[34] B. Ozenne, F. Subtil, and D. Maucort-Boulch, "The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases," *Journal of Clinical Epidemiology*, vol. 68, no. 8, pp. 855–859, 2015.

[35] J. M. Lobo, A. Jiménez-valverde, and R. Real, "AUC: A misleading measure of the performance of predictive distribution models," *Global Ecology and Biogeography*, vol. 17, no. 2, pp. 145–151, 2008.

[36] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, pp. 243–275, ACM, NY, USA, June 2006.

[37] R. Lambiotte, J. C. Delvenne, and M. Barahona, *Laplacian Dynamics and Multiscale Modular Structure in Networks*, Physics, 2008.

[38] H. Liao, M. S. Mariani, M. s. Medo, Y.-C. Zhang, and M.-Y. Zhou, "Ranking in evolving complex networks," *Physics Reports*, vol. 689, pp. 1–54, 2017.

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

[40] I. R. Dunsmore and S. Geisser, "Predictive Inference: An Introduction.," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, vol. 158, no. 1, p. 191, 1995.

[41] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information and System Security*, vol. 22, no. 1, pp. 5–53, 2004.

[42] L. Getoor and C. P. Diehl, "Link mining: a survey," *SIGKDD Explorations Newsletter*, vol. 7, no. 2, pp. 3–12, 2005.

[43] M. A. Hasan and M. J. Zaki, "A survey of link prediction in social networks," in *Social Network Data Analytics*, pp. 243–275, Springer, NY, USA, 2011.

[44] C. C. Heyde, "A supplement to the strong law of large numbers," *Journal of Applied Probability*, vol. 12, pp. 173–175, 1975.

[45] K. L. Chung, "The strong law of large numbers," *Selected Works of Kai Lai Chung*, pp. 145–156, 2008.

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

[47] T. Beuming, L. Skrabanek, M. Y. Niv, P. Mukherjee, and H. Weinstein, "PDZBase: A protein-protein interaction database for PDZ-domains," *Bioinformatics*, vol. 21, no. 6, pp. 827-828, 2005.

[48] L. Subelj and M. Bajec, "Robust network community detection using balanced propagation," *The European Physical Journal B*, vol. 81, no. 3, pp. 353–362, 2011.

[49] University of koblenz-Landau, "the koblenz network collection," http://konect.uni-koblenz.de/networks/.

[50] X. F. Wang and G. Chen, "Complex networks: small-world, scale-free and beyond," *IEEE Circuits and Systems Magazine*, vol. 3, no. 1, pp. 6–20, 2003.