Physica A 510 (2018) 52–67

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Link prediction in complex networks based on the interactions among paths

Yabing Yao [a,b], Ruisheng Zhang [a,*], Fan Yang [a,c], Jianxin Tang [a,d], Yongna Yuan [a], Rongjing Hu [a]

[a] School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China
[b] Gansu branch of National Prosecutors College, Lanzhou 730010, China
[c] School of Software Engineering, Lanzhou Institute of Technology, Lanzhou 730050, China
[d] School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China

HIGHLIGHTS

- The interactions among paths with different length have been considered and introduced to link prediction.
- The resource-traffic flow mechanism has been adopted to measure the interactions among paths between node pairs.
- A quasi-local path index with better overall performance under AUC and Precision metrics has been proposed.

ARTICLE INFO

Article history:
Received 5 November 2017
Received in revised form 19 May 2018
Available online 13 June 2018

Keywords:
Complex networks
H-index
Link prediction
Resource receiving process

ABSTRACT

Link prediction in incomplete complex networks is an important issue in network science. Recently, various structure-based similarity methods have been proposed. However, most path-dependent methods merely pay attention to the contributions of paths with specific length, which neglects the interactions of paths with different length for performance improvement. Motivated by the resource-traffic flow mechanism on networks, we measure the interaction relationship of paths with a resource receiving process. In this process, each node takes certain initial resources quantified by its H-index, and then the intermediate nodes on paths can receive resources from their neighbours. Based on this process, a local path-based link predictor which emphasizes the effect of the Resources from Short Paths (RSP) is proposed. Experiments on twelve real-world networks demonstrate that the RSP index has better performance than other nine structure-based similarity methods.

1. Introduction

With the booming of network science, complex network theory has become a powerful tool to further reveal mechanisms and understand phenomena in the real world [1,2]. However, due to the difficulty of obtaining data or the mistake in gathering data, the real-world networks is often incomplete or inaccurate, which further affects the results of our investigations and experiments [3,4]. In order to tackle this issue, link prediction in complex networks has drawn much attention from researchers of different fields in recent years. It aims to detect the missing links or forecast the future links based on the existing properties and topological structures of the observed networks [5]. The study of link prediction is of

* Corresponding author.
E-mail addresses: yaoyb14@lzu.edu.cn (Y. Yao), zhangrs@lzu.edu.cn (R. Zhang).

https://doi.org/10.1016/j.physa.2018.06.051
0378-4371/

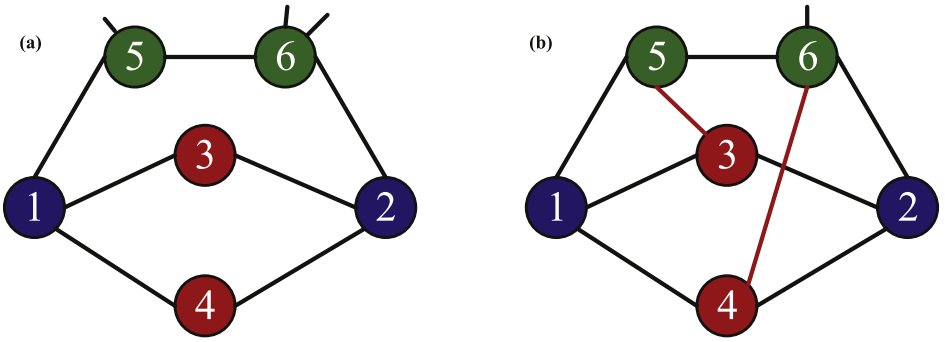


Fig. 1. Two simple networks to illustrate the interactions among paths. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

great theoretical significance. It enables us to extract the implicit information, identify the spurious links in networks and also helps us to model and evaluate the evolution mechanisms of networks [6]. Furthermore, link prediction has been of great practical application value in diverse areas, e.g., uncovering the disease relationships [7] and drug repositioning [8] in biological networks, finding new friends in social networks [9,10], discovering underground criminal groups in terrorist networks [11], predicting the potential collaborators in citation networks [12], as well as recommending the favourite goods for customers in online shopping systems [13].

Various approaches have been proposed for link prediction. Overall, there are mainly two kinds of methods: learning-based methods and similarity metric-based methods [6,11,14]. Although having good performance, the learning-based methods, such as classification-based methods [15,16], probabilistic model-based methods [17–19] and matrix factorization-based methods [20,21], are very time-consuming in constructing training data and learning from training data, with the result that they fail to deal with large scale networks. As a mainstream approach for link prediction, the similarity metric-based methods consider that node pairs will form links with high probability if they are similar in node attributes or network structure features [11]. Nevertheless, the node attribute-based similarity methods are generally oriented towards the specific contexts, which constrains their application scope for diverse kinds of networks. In addition, node attributes are always hidden resulting in the difficulty of obtaining data. By contrast, the topological structures of networks are easier to obtain and have a good universal adaptability with low computational complexity. Therefore, the structure-based similarity methods have been widely explored recently. Different kinds of topological structures have been utilized to improve the accuracy of link prediction, such as neighbour, node degree, clustering coefficient, node centrality, path as well as community, and so forth. For a path between a pair of nodes, most path-based prediction indices generally consider its contribution in complete isolation from the specific length. However, the contribution of interactions of paths with different length is always neglected.

In the real-world networks, the nodes on one path may interact with the nodes on another different path for a pair of nodes. The links between nodes on different paths are called interactions among paths herein. Take Fig. 1(a) for example, there are two paths with length 2 (i.e. $p_1 = \{v_1, v_3, v_2\}$ and $p_2 = \{v_1, v_4, v_2\}$) and one path with length 3 (i.e. $p_3 = \{v_1, v_5, v_6, v_3\}$) between nodes v_1 and v_2. In this network, the intermediate nodes on path p_3 (i.e. v_5 and v_6) do not have links with the common neighbours (i.e. v_3 and v_4) and thus there is no interaction relationship between different paths. However, in Fig. 1(b), the nodes v_5 and v_6 on path with length 3, i.e. $p_3 = \{v_1, v_5, v_6, v_3\}$, connect to the nodes v_3 of path $p_1 = \{v_1, v_3, v_2\}$ and v_4 of path $p_2 = \{v_1, v_4, v_2\}$, respectively. Therefore, the red links in Fig. 1(b) reflect the interactions among paths. In the process of link prediction, it is well known that short paths are better than long paths in improving performance [22,23]. Motivated by this idea, we suppose that the paths have significant contribution towards link prediction if their intermediate nodes prefer to interact with nodes on short paths.

Based on the above discussion, we focus on the structure-based similarity methods and pay special attention to the interactions among paths for link prediction in this paper. Inspired by the resource-traffic flow mechanism on networks [24], we regard the interactions among paths as a receiving resource process through the neighbours of their intermediate nodes. Furthermore, the more the intermediate nodes of a path receive resources from nodes on short paths, the more contribution of this path is. Since the roles of different nodes may be largely different in networks, we treat the knowledge quantity (i.e. H-index) as the virtual resources that every node initially takes according to the knowledge dissemination mechanism [25]. Then, a path-dependent link predictor based on the Resource receiving process from Short Paths (RSP) is proposed. In order to evaluate the performance of RSP index, we conduct experiments on twelve real-world networks drawn from different fields. The experimental results show that the RSP index outperforms four classical predictors and other five path-dependent predictors.

The rest of this paper is organized as follows. The related work is introduced in Section 2. In Section 3, we define and propose our RSP index. The comparison baselines and metrics are given in Section 4. In Section 5, we evaluate the prediction performance of our index with different networks. This work is concluded in Section 6.

2. Related work

In order to solve the problem of link prediction, numerous methods have been developed from different insights, which can be mainly categorized into learning-based methods and similarity metric-based methods. Learning-based methods learn and build models based on the training data that is composed of observed links to predict the likelihood of potential links in networks. Similarity metric-based methods generally calculate the similarity of node pairs according to the node attributes or topological structures. In this section, we will briefly introduce the recent achievements of learning-based and node attribute-based methods and pay more attention to the structure-based similarity methods, since this paper mainly focuses on the link prediction methods based on topological structures of networks.

2.1. Learning-based methods

As one of the typical learning-based methods, the classifier-based methods adopt the idea of binary classification to address link prediction problem. These methods consider that each pair of nodes has a positive or negative class label indicating the existence or non-existence of the corresponding link between two endpoints [15,16]. For instance, Farshad et al. proposed a new similarity measure based network motifs and applied this measure to two supervised learning classifiers: Gradient Boosting Machine (GBM) and Linear Discriminant Analysis (LDA) [26].

Some learning-based methods also consider the statistical and probabilistic models of networks for link prediction. Generally, these methods believe that network formation process always obeys certain known structures and principles, such as Hierarchical Structure Model [17] and Stochastic Block Model [18]. Meanwhile, Liu et al. developed a Fast Blocking probabilistic Model considering the link densities among communities [19]. Considering the clustering mechanism, Pan et al. recently calculated the formation probability of networks via a predefined structural Hamiltonian [27].

Due to its success in recommender systems, the matrix factorization models have attracted much attention and are applied to link prediction problem as well. These models can efficiently extract latent features of nodes and links in networks [20] and are widely applied in different types of networks. By mapping the adjacency matrix of observed network into another space based on kernel functions, Jiao et al. proposed an objective function of the nonnegative matrix factorization according to the mapping space for link prediction [28]. Pech et al. employed the theory of robust principal component analysis and decomposed the network adjacency matrix into two parts: one low-rank matrix representing the network backbone and one sparse matrix indicating the spurious links in network [29]. Ma et al. proposed a graph regularized nonnegative matrix factorization method to solve temporal link problem in dynamic networks [30].

2.2. Node attribute-based similarity methods

The node attribute-based methods pay attention to the node information outside the structure of network. Based on user attributes and graph structure in social networks, Yin et al. estimated the link relevance with a random walk algorithm [31]. Wang et al. employed the auxiliary information of users (e.g., profile or microblogs) in Twitter and Facebook, and combined the topological information of users to solve the cold-start link prediction problem [32]. Li et al. proposed a novel feature extraction method to derive node attributes and efficiently solve the feature shortage problem in link prediction [33]. Overall, the node attribute-based methods are a natural way to reflect the similarity of node pairs. However, the effectiveness of this kind of methods mainly rely on the domain and the specific networks.

2.3. Structure-based similarity methods

The structure-based similarity methods always assign similarity scores to unconnected node pairs with the aid of the topology features of networks. Generally, there are mainly four kinds of methods: local approaches, global approaches, quasi-local approaches and community-based approaches [6,11,34].

The local similarity-based approaches calculate the similarity scores of node pairs with consideration of node neighbour-related structural information. Common Neighbour index (CN) [35] is the simplest and less time-consuming local method, except for the Preferential Attachment index (PA) [36]. In order to distinguish the contributions of different common neighbours, several variants of CN have been proposed, such as Adamic–Adar index (AA) [37], Resource Allocation index (RA) [38], Mutual Information index (MI) [39] and CRA index [40]. Furthermore, some researchers investigate the local clustering abilities of nodes and links for calculating similarity. Li et al. developed two novel node-coupling clustering approaches and combined the coupling degrees of common neighbours for predicting links [41]. Wu et al. subsequently explored the influence of node clustering coefficient [42] and link clustering information [43,44] on link prediction. In most cases, local similarity-based approaches can be considered that they use only the information of paths with length 2 for a pair of nodes. Consequently, these approaches own the low computational complexity but also have low prediction accuracy.

On the contrary, the global similarity-based approaches employ the whole network topological information for calculating node similarity, such as Katz index [45], Leicht–Holme–Newman (LHN2) [46], SimRank [47,48] and Random Walk with Restart (RWR) [49]. Although having high prediction accuracy, global approaches suffer the problem of high computational complexity.

The quasi-local similarity-based approaches, which employ more topological information than local approaches and less than global approaches, focus on a trade-off between accuracy and efficiency, such as Local Path index (LP) [22,38], Local Random Walks (LRW) [50], FriendLink (FL) [23]. Considering the structure feature of local networks for a pair of nodes, Zhang et al. measured the node similarity based on relative entropy according to the degree distribution of neighbours of each node in local networks [51].

Recently, community-based link prediction approaches have been also developed based on the fact that nodes in same communities always have similar features in networks. Yan et al. introduced the community structure information to the existing link prediction methods to improve prediction accuracy for the first time [3]. Ding et al. extracted the community structure of networks under different resolutions and applied a frequency statistical model to calculate the probability of missing link [52]. Then, they further considered the relevance between communities and proposed a novel prediction algorithm based on the ruler inference [53]. Wang et al. developed a novel link prediction method with community structure based on hyperbolic mapping technology [54].

There are also some link prediction approaches considering the noise in networks. Lü et al. proposed the structural consistency index to measure the predictability of networks [55]. Wang et al. presented a popularity based structural perturbation method to characterize the existence probability of links [56]. Zhang et al. investigated the robustness of link prediction algorithms under noisy environment [57]. Wang et al. proposed a perturbation-based frame to predict missing links with consideration of random noises and irregular links [21].

Additionally, from the viewpoint of path feature in networks, the path-based structural similarity methods can be categorized into the non-heterogeneity path methods and the heterogeneity path methods [58]. The non-heterogeneity path methods simply sum over the total number of paths as similarity scores and neglect the different contributions of paths even with the same length. For example, CN [35], LP [22,38] and Katz [45] indices can be considered as the typical non-heterogeneity path methods with consideration of paths with length 2, length 2 and 3, and maximum path length, respectively. In order to improve prediction performance, the heterogeneity path methods pay more attention to the different contributions of paths with the same length. The variants of CN-based indices, such as RA [38] and AA [37], actually belong to the heterogeneity path methods which consider paths with length 2. Recently, Zhu et al. proposed the SP [58] and EP [59] indices that distinguish the different contributions of paths with length 2 and 3 by considering the degrees of intermediate nodes on paths. With the help of information theory, Zhu et al. proposed the NSI index that integrates the topological features of common neighbours and the links across neighbour sets for a pair of nodes [60]. Pei et al. extended the NSI index and proposed a more general information-theoretic prediction model with a virtual information allocation process [61]. Liu et al. proposed the ERA index which extended RA index with consideration of the resource transfer process of local paths [62]. Yang et al. considered the significant influence of endpoints and proposed a novel SI index to measure the contributions of strong and weak relations [63]. All these indices [37,38,58–63] consider the contributions of paths from the viewpoint of intermediate nodes on paths. Recently, Xu et al. proposed the PE index that measures the contributions of paths with length 2 and 3 by summing over the entropies of their intermediate links [64]. Based on the knowledge dissemination mechanism, Jia et al. developed the KDLP index, which takes into account the maximum length of paths between two nodes and redefines the weight of each intermediate link on these paths through knowledge quantity [25]. Wu et al. defined the concept of asymmetric link clustering coefficient to measure the contribution of each link and proposed the improved prediction methods based on the models of Local Naïve Bays and mutual information theory [44]. These methods [25,44,64] emphasize the different contributions of paths from the viewpoint of intermediate links on paths.

Generally, most path-based similarity methods only consider the contribution of one path with specific length from the viewpoint of its intermediate nodes or intermediate links, whereas the contributions of interactions of different length paths are always neglected. Carlo et al. proposed a serials of Local Community Paradigm-based (LCP) methods for link prediction, such as CAR and CRA, which account for the LCP structure in networks, i.e., two nodes are more likely to link together if there are more inner links among their common neighbours [40]. However, the LCP-based methods perform poorly when the LCP structure feature of networks is not obvious [42]. In this paper, we consider the interactions among different length paths and propose a RSP index for link prediction. For one path between a pair of nodes, the contribution of its each intermediate node is considered as a process of receiving resources from its neighbours. In comparison to other nine baselines, our index shows better performance in twelve real-world networks, particularly those with low LCP structure.

3. Methods

Consider an undirected unweighted network $G(V, E)$, where V is the set of nodes and E is the set of links. The adjacency matrix of network G is denoted by A whose element a_{ij} is equal to 1 when there is a link between nodes i and j, and 0 otherwise. Given a pair of nodes $(x, y) \in V$, its similarity score is denoted by S_{xy}. All unconnected node pairs are ranked in decreasing order based on their similarity scores. The top ranked node pairs are supposed to have high probabilities of forming links in the future. Since the roles of different nodes are different in networks, in Ref. [25], the knowledge quantity of one node (i.e. H-index) is used to characterize the importance of this node and applied to link prediction. In this paper, we treat the knowledge quantity as the virtual resources that every node initially takes. This virtual resources can distribute to the neighbours of one node on average. Here, the definition of the initial resources of one node is given below.

Definition 1. Let $v_i \in V$ represent an arbitrary node in network G, k_i is the degree of v_i and $\Gamma(i) = \{v_{j1}, v_{j2}, \ldots, v_{jk_i}\}$ is the neighbour set of v_i, for the node v_i, its initial resources are defined as [25,65]:

$$h_i = H(k_{j1}, k_{j2}, \ldots, k_{jk_i}), \tag{1}$$

where k_{jk_i} represents the degree of neighbour v_{jk_i} of node v_i. The operator $H(\cdot)$ returns the maximum number h such that there exist at least h elements in $(k_{j1}, k_{j2}, \ldots, k_{jk_i})$, each of which is no less than h. In this paper, a higher h of one node means that this node takes more initial resources.

According to the mechanism of resource-traffic flow on networks [24], the resources taken by each node can be divided into several pieces and flow to its neighbours, while each node can receive resources from all its neighbours as well. Inspired by this mechanism, we simulate this resource flowing mechanism for link prediction via two steps. First, the initial resources h_i of node v_i can be equally distributed to all its neighbours. Second, the node v_i can receive resources from all its neighbours. Actually, the amount of each piece that flows to the neighbours of node v_i may be different due to the different importance of nodes in networks. Herein, we only consider the simplest case in the first step and assume that the resources of a node are usually divided by its neighbours on average. These two steps are defined as follows.

Definition 2. Given a node $v_i \in V$ with the initial resources h_i, the resources that node v_i distributes to each of its neighbours $v_j \in \Gamma(i)$ are defined as

$$d_{i\to j} = \frac{h_i}{k_i} \tag{2}$$

Meanwhile, the node v_i simultaneously receives resources from all its neighbours, which is defined as

$$R_i = \sum_{v_j \in \Gamma(i)} d_{i\leftarrow j} \qquad = \sum_{v_j \in \Gamma(i)} \frac{h_j}{k_j} \tag{3}$$

where $d_{i\leftarrow j}$ indicates that the resources of node v_j move to node v_i.

In many cases, for a pair of nodes, its short paths are better than long paths in improving performance, which has been confirmed by previous studies [22,23]. Therefore, considering a path with specific length, this path has a significant contribution to performance improvement if its intermediate nodes receive more resources from the nodes on short paths. Owning to the computational complexity, we limit the short paths to those with length 2 in this paper, i.e. the predicted node pair and their common neighbours (intermediate nodes) are only concerned.

Definition 3. Given a pair of nodes (x, y), the node set on paths with length 2 between the pair (x, y), which includes the common neighbours and the predicted node pair, is denoted by $Q_{xy} = \{\Gamma(x) \bigcap \Gamma(y) \bigcup \{x, y\}\}$. For a path $p^t_{(x,y)} = \{v_0 = x, v_1, \ldots, v_{t-1}, v_t = y\}$ with length t ($t \geqslant 2$) between (x, y), its intermediate node set is denoted by $I(p_t) = \{v_1, \ldots, v_{t-1}\}$. The contribution of one of intermediate nodes $v_j \in I(p_t)$ is defined as follows:

$$\xi(v_j) = \frac{R_{j|Q_{xy}}}{R_{j|\Gamma(j)}} = \frac{\sum_{v_{i'} \in Q_{xy}} d_{j\leftarrow i'} a_{i'j}}{\sum_{v_i \in \Gamma(j)} d_{j\leftarrow i}} \tag{4}$$

where $R_{j|Q_{xy}}$ is the resources of node v_j receiving from Q_{xy} and $R_{j|\Gamma(j)}$ is the resources receiving from all neighbours of node v_j. $a_{i'j}$ is equal to 1 when there is a link between nodes i' and j, and 0 otherwise. According to Eq. (4), $\xi(v_j) \in [0, 1]$, 1 corresponds to $Q_{xy} = \Gamma(j)$ and 0 indicates $Q_{xy} \bigcap \Gamma(j) = \varnothing$. Moreover, the higher the value of $\xi(v_j)$ is, the more the contribution of node v_j is.

For the node pair (x, y), the contribution of path $p^t_{(x,y)}$ sums over all the contributions of intermediate nodes on this path, which is defined as

$$\omega(p^t_{(x,y)}) = \sum_{v_j \in I(p_t)} \xi(v_j) \tag{5}$$

Accordingly, $\omega(p^t_{(x,y)})$ considers all intermediate nodes on the path $p^t_{(x,y)}$. The higher value of $\omega(p^t_{(x,y)})$ corresponds to the more significant contribution of this path on performance improvement.

Definition 4. Given a pair of nodes (x, y), l is the maximum length of paths between them, the contribution of all paths between (x, y) is defined as

$$S_{xy} = \sum_{p^2_{(x,y)} \in P_2} \omega(p^2_{(x,y)}) + \lambda \sum_{p^3_{(x,y)} \in P_3} \omega(p^3_{(x,y)}) + \cdots + \lambda^{l-2} \sum_{p^l_{(x,y)} \in P_l} \omega(p^l_{(x,y)}) \tag{6}$$

where $\lambda \in [0, +\infty)$ is a tunable parameter, P_l is the set of paths with length l between node pair (x, y) and $p^l_{(x,y)}$ is one of the paths in P_l.

Considering the high computational complexity, we only pay attention to the local paths with length 2 and 3 between x and y, i.e. $l \leqslant 3$. Therefore, the prediction index is defined as

$$S_{xy}^{RSP} = \sum_{p_2 \in P_2} \omega(p_2) + \lambda \sum_{p_3 \in P_3} \omega(p_3) \tag{7}$$

where p_2 represents a path with length 2, p_3 represents a path with length 3 and $\omega(p)$ is defined in Eq. (5). Due to the dependence of the Resources from Short Paths, this index is named as RSP in this paper. Algorithm 1 shows the implementation framework of RSP index.

Algorithm 1: The implementation procedure of RSP index.

Input: Network $G(V, E)$, parameter λ, node pair(x, y).
Output: Similarity score S_{xy}.

for $v_i \in V$ **do**
  compute the initial resources h_i of node v_i by Eq. (1)
  compute the resources that v_i distributes to each of its neighbours by Eq. (2)
end for
construct the node set $Q_{xy} = \{\Gamma(x) \bigcap \Gamma(y) \bigcup \{x, y\}\}$
initialize the similarity score $s = s' = 0$
find the set of paths with length 2 between (x, y) P_2
for $p_2 \in P_2$ **do**
  compute the contribution $\omega(p_2)$ by Eqs. (4)(5).
  $s = s + \omega(p_2)$
end for
find the set of paths with length 3 between (x, y) P_3
for $p_3 \in P_3$ **do**
  compute the contribution $\omega(p_3)$ by Eqs. (4)(5).
  $s' = s' + \omega(p_3)$
end for
$S_{xy} = s + \lambda * s'$
return S_{xy}

4. Baselines and metrics

In this paper, we choose nine baselines for performance comparison and utilize two widely used metrics (AUC and Precision) to evaluate prediction accuracy.

4.1. Comparison baselines

(1) Common Neighbour index (CN) [35,66]. This index counts the number of all common neighbours as similarity score and is defined as

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

where $\Gamma(x)$ represents the neighbour set of node x.

(2) Adamic–Adar index (AA) [37]. This index is a variant of CN, which draws a distinction among common neighbours. It is defined as

$$S_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z}, \tag{9}$$

where k_z is the degree of node z that is one of common neighbours between x and y.

(3) Resource Allocation index (RA) [38]. Motivated by the resource allocation mechanism on networks, this index punishes the common neighbours having large degree more heavily than AA. It is defined as

$$S_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z}. \tag{10}$$

(4) CRA index [40]. This index takes into account the total number of common neighbours and the inner links between common neighbours. It is defined as

$$S_{xy}^{CRA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{|\gamma(z)|}{|\Gamma(z)|}, \tag{11}$$

where $\gamma(z)$ is the local community degree of node z, i.e. the sub-set of neighbours of z that are also common neighbours of nodes x and y.

(5) Katz index [45]. This index considers all paths between two nodes and assigns less weights to longer paths. It is defined as

$$S_{xy}^{Katz} = \sum_{l=1}^{\infty} \beta^l \cdot |paths_{xy}^{\langle l \rangle}| = \beta A_{xy} + \beta^2 (A^2)_{xy} + \beta^3 (A^3)_{xy} + \cdots, \tag{12}$$

where A is the adjacency matrix of network and $|paths_{xy}^{\langle l \rangle}|$ is the number of paths with length l between x and y. β is a tunable parameter that is always fixed at a very small value. If β is lower than the reciprocal of the maximum of the eigenvalues of adjacent matrix A, this index can be redefined as $S = (I - \beta A)^{-1} - I$.

(6) Local Path index (LP) [22,38]. This index only counts the number of paths with length 2 and 3 between two nodes and is defined as

$$S_{xy}^{LP} = A^2 + \lambda A^3, \tag{13}$$

where λ is a free parameter.

(7) Neighbour Set Information index (NSI) [60]. Based on the information theory, this index employs two structure features to compute similarity: the common neighbours of two nodes and the links across two neighbour sets. It is defined as

$$S_{xy}^{NSI} = -I(L_{xy}^1 | O_{xy}) - \lambda I(L_{xy}^1 | P_{xy}), \tag{14}$$

where L_{xy}^1 is connection probability between node pair (x, y), O_{xy} represents the common neighbours and P_{xy} denotes the set of observed links across two neighbour sets.

(8) Path Entropy index (PE) [64]. This index measures the different contributions of paths between two nodes by considering the entropies of intermediate links on paths. It is defined as

$$S_{xy}^{PE} = -I(L_{xy}^1 | \bigcup_{i=2}^{l} \{D_{xy}^i\}), \tag{15}$$

where l is the max length of paths (it is set to 3 in our experiments) and L_{xy}^1 is the connectivity probability of node pair (x, y) and D_{xy}^i is a path with length i.

(9) Knowledge Dissemination Link Predictor index (KDLP) [25]. Considering the knowledge quantity of every node in networks, each observed link is assigned to a weight value based on the knowledge dissemination mechanism. It is defined as

$$S_{xy}^{KDLP} = \sum_{l=1}^{\infty} \beta^l \cdot |paths_{xy}^{\langle l \rangle}| = \beta W_{xy} + \beta^2 (W^2)_{xy} + \beta^3 (W^3)_{xy} + \cdots, \tag{16}$$

where W_{xy} is the weighted adjacency matrix of networks, β is a free parameter and l is the length of paths between x and y.

4.2. Evaluation metrics

Consider a simple network $G(V, E)$, U denotes all possible links in this network. In order to test the accuracy of predictors, all existing links E are randomly divided into two parts: the training set E^T and the probe set E^P. The links in E^P are named as missing links and the links in $U - E$ are called as non-existent links. E^T is regarded as the known information of networks to calculate node similarity, while E^P is used to validate the accuracy of predictors and does not participate in the prediction procedure. Obviously, $E^T \cup E^P = E$ and $E^T \cap E^P = \varnothing$. In our experiments, ten percent of existing links E are chosen as the probe set and the rest percent existing links are considered as the training set. We use two metrics to measure the prediction accuracy and describe them as follows.

(1) AUC (area under the receiver operating characteristic curve [67]) can be considered as the probability that the similarity score of one randomly chosen missing link is higher than that of one randomly chosen non-existent link [68]. When we perform n times of independent comparisons, if there are n' times where the missing links have higher scores than the non-existent links and n'' times where both of them have equal scores, then AUC is defined as

$$AUC = \frac{n' + 0.5 \times n''}{n}. \tag{17}$$

If all the similarity scores are generated from an independent and identical distribution, AUC ≈ 0.5. For a prediction index, the extent to which its AUC exceeds 0.5 reflects how much better its prediction accuracy than pure chance.

Table 1
The topology properties of twelve networks.

Networks	N	M	$\langle k \rangle$	S	r	d	C	H	e	LCP_{corr}
Karate	34	78	4.588	0.139	−0.476	2.408	0.588	1.693	0.492	0.756
Dolphins	62	159	5.129	0.084	−0.044	3.357	0.303	1.327	0.379	0.907
Football	115	613	10.661	0.094	0.162	2.508	0.403	1.007	0.450	0.893
Physicians	117	465	7.949	0.068	−0.084	2.587	0.219	1.253	0.432	0.791
Celegans	297	2148	14.465	0.049	−0.163	2.455	0.308	1.801	0.445	0.906
WikiVote	889	2914	6.556	0.007	−0.029	4.096	0.195	2.766	0.274	0.893
SmaGri	1024	4916	9.602	0.009	−0.192	2.981	0.349	3.947	0.358	0.946
Blogs	1222	16714	27.355	0.022	−0.221	2.737	0.360	2.971	0.398	0.929
Yeast	2224	6609	5.943	0.003	−0.105	4.376	0.201	2.803	0.246	0.886
Kohonen	3704	12673	6.843	0.002	−0.121	3.670	0.304	9.317	0.296	0.857
Erdos	4680	7030	3.004	0.001	−0.460	5.491	0.273	5.454	0.194	0.792
Power	4941	6594	2.669	5.4e−4	0.004	18.989	0.106	1.450	0.063	0.846

Note: N and M are the node size and link size. $\langle k \rangle$, S and r represent the average degree, network density and assortative coefficient, respectively. d and C are the average distance and average clustering coefficient. H is the degree heterogeneity. e is the network efficiency. $LCP_{corr} = \frac{cov(CN,LCL)}{\sigma_{CN} \cdot \sigma_{LCL}}$ denotes the Pearson correlation coefficient of common neighbours (CN) and inner links between common neighbours (LCL) [40].

Table 2
Prediction accuracy measured by AUC in twelve networks.

Networks	CN	RA	AA	CRA	Katz	KDLP	LP_{opt}	PE	NSI_{opt}	$RSP_{0.1}$	RSP_{opt}
Karate	0.7008	0.7420	0.7342	0.5973	0.7573	0.8032	0.7869	**0.8401***	0.8115	0.8129	**0.8248**
Dolphins	0.7973	0.7976	0.7985	0.6358	**0.8460***	0.8351	0.8339	0.8223	0.8320	0.8370	**0.8371**
Football	0.8541	0.8537	0.8538	0.8244	0.8641	**0.8692***	0.8641	0.8646	0.8617	0.8687	**0.8690**
Physicians	0.7101	0.7141	0.7141	0.5721	0.7293	**0.7389***	0.7310	0.7165	0.7232	0.7358	**0.7360**
Celegans	0.8482	0.8697	0.8649	0.7697	0.8644	0.8760	0.8692	0.8858	0.8806	**0.8966**	**0.8969***
WikiVote	0.8098	0.8115	0.8121	0.6430	**0.9107**	**0.9191***	0.9001	0.9042	0.9006	0.9067	0.9071
SmaGri	0.8488	0.8591	0.8591	0.7168	0.8977	0.9099	0.9027	0.7119	0.9079	**0.9249**	**0.9250***
Blogs	0.9242	0.9288	0.9276	0.8985	0.9332	0.9196	0.9404	0.4857	0.9410	**0.9473**	**0.9475***
Yeast	0.7313	0.7316	0.7318	0.6108	**0.9119***	**0.9104**	0.8888	0.8905	0.8890	0.8898	0.8916
Kohonen	0.8281	0.8356	0.8356	0.6501	0.9088	**0.9294***	0.9142	0.7050	0.9188	0.9281	**0.9284**
Erdos	0.7351	0.7354	0.7353	0.5464	**0.9344**	**0.9388***	0.8592	0.8590	0.8593	0.8595	0.8598
Power	0.6250	0.6251	0.6250	0.5176	**0.9640**	**0.9676***	0.6973	0.6972	0.6973	0.6971	0.6973

Note: The mean AUC is obtained by the mean of 50 independent implementations with a random 90%–10% division of training set and probe set. The top two best performance in each network is emphasized in bold and $*$ denotes the highest performance. LP_{opt}, NSI_{opt}, RSP_{opt} represents the best performance of LP, NSI and RSP indices when $\lambda \in [0, +\infty)$, respectively. $RSP_{0.1}$ denotes the performance of RSP with $\lambda = 0.1$.

(2) Precision only pays attention to the top-ranked links [69]. In practice, all non-observed links, including the missing links and the non-existent links, are ranked in descending order according to their similarity scores. Considering *top-L* non-observed links, if there are L_r links belong to missing links, then Precision is defined as

$$Precision = \frac{L_r}{L}. \tag{18}$$

Therefore, a high Precision value indicates a high prediction accuracy.

5. Experimental results and analysis

5.1. Datasets

In this paper, twelve real-world networks from different fields with different structures are considered. The basic topology properties of these networks are given in Table 1. In our experiments, the giant component of each network is only considered. These networks are briefly described as follows: (1) Karate [70]: a friendship network of a karate club at a US university in the 1970s; (2) Dolphins [71]: a social network between dolphins in a community living off Doubtful Sound, New Zealand; (3) Football [72]: an American football game network; (4) Physicians [73]: an innovation spread network among physicians; (5) Celegans [74]: a neural network of the nematode Caenorhabditis elegans; (6) WikiVote [75,76]: an election and vote history network from Wikipedia; (7) SmaGri [77]: a network that composed of citations to Small & Griffith and Descendants; (8) Blogs [78]: an US political blog network; (9) Yeast [79]: a protein–protein interaction network of yeast; (10) Kohonen [75]: a network of articles with topic self-organizing maps or references to Kohonen; (11) Erdos [75]: a collaboration network; (12) Power [80]: a power grid network of the western US. All these networks can be downloaded from the websites of http://konect.uni-koblenz.de/ and http://networkrepository.com/networks.php.

Table 3
Prediction accuracy measured by top-100 Precision in twelve networks.

Networks	CN	RA	AA	CRA	Katz	KDLP	LP_{opt}	PE	NSI_{opt}	$RSP_{0.1}$	RSP_{opt}
Karate	0.0344	0.0434	0.0433	0.0216	0.0441	0.0518	0.0479	**0.0552***	0.0506	0.0535	**0.0545**
Dolphins	**0.0688**	0.0648	0.0676	0.0480	0.0648	0.0506	**0.0688**	0.0594	0.0620	0.0686	**0.0694***
Football	0.2988	0.2956	0.2950	**0.3258**	0.2894	0.3150	0.2988	0.2766	0.2834	0.3180	**0.3284***
Physicians	0.0558	0.0590	0.0556	0.0522	0.0544	0.0578	0.0558	0.0532	0.0576	**0.0594**	**0.0602***
Celegans	0.1272	0.1335	0.1375	0.1577	0.1379	0.1032	0.1602	**0.1798***	0.1591	0.1646	**0.1656**
WikiVote	0.1294	0.1228	0.1402	0.1259	0.1300	0.0112	0.1298	0.1363	0.1382	**0.1447**	**0.1463***
SmaGri	0.1953	0.1900	0.2003	0.2210	0.1950	0.1027	0.2017	0.1063	0.2020	**0.2383**	**0.2397***
Blogs	0.4227	0.2380	0.3760	0.4920	0.4573	0.0080	0.5113	0.0047	0.5073	**0.5140**	**0.5173***
Yeast	0.2233	0.1957	0.2323	0.2943	0.2217	0.0057	0.2487	0.2970	0.2693	**0.3173**	**0.3207***
Kohonen	0.1516	0.1400	0.1471	0.2452	0.1526	0.0577	0.1597	0.2139	0.1655	**0.2558**	**0.2932***
Erdos	0.2500	0.2150	0.2433	0.2117	0.2490	0.0157	0.2543	0.2483	0.2667	**0.2730**	**0.2990***
Power	0.1287	0.0783	0.0947	**0.1893***	0.1317	0.0113	0.1347	0.1163	0.1380	0.1670	**0.1770**

Note: The mean Precision is obtained by the mean of 50 independent implementations with a random 90%–10% division of training set and probe set. The top two best performance in each network is emphasized in bold and $*$ denotes the highest performance. LP_{opt}, NSI_{opt}, RSP_{opt} represents the best performance of LP, NSI and RSP indices when $\lambda \in [0, +\infty)$, respectively. $RSP_{0.1}$ denotes the performance of RSP with $\lambda = 0.1$.

5.2. The prediction accuracy with variation of parameter λ

According to the definition of RSP index in Eq. (7), it simultaneously considers the paths with length 2 and 3 by penalizing long paths with a tunable parameter $\lambda \in [0, +\infty)$. For RSP index, if we solely take paths with length 2 or paths with length 3 into account, it indicates that λ corresponds to 0 or $+\infty$, respectively. In this subsection, we focus on the changing of prediction accuracy with variation of parameter λ for RSP index. The AUC and Precision results are shown in Figs. 2 and 3. Although λ can be increased from 0 to $+\infty$, yet its optimal parameter always lies within 0 and 1 (see detailed reason in Appendix). Meanwhile, due to the parameter dependence of NSI and LP indices that consider the paths with 2 and 3, their accuracies are also plotted in these two figures for comparison.

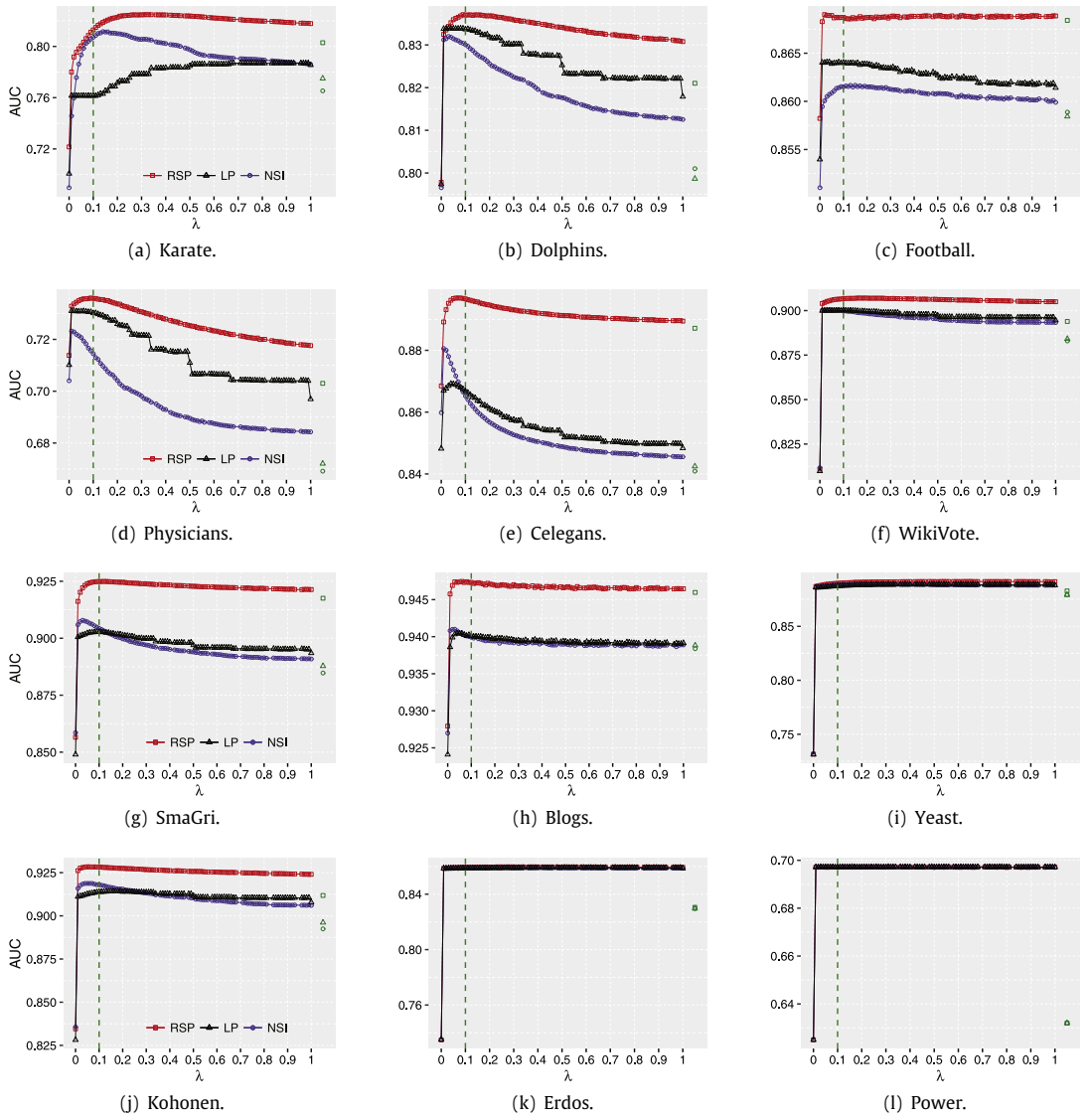
As shown in Figs. 2 and 3, the prediction accuracies of RSP, NSI and LP can be improved with the increase of parameter λ in most cases. This result means that the paths with length 3 pay a significant role in performance improvement and should not be neglected. In comparison with NSI and LP indices, our RSP index performs best in exploiting path information, in particular for the Precision metric. For example, in the networks of Physicians, WikiVote, SmaGri, Kohonen and Erdos in Fig. 3, with the changing of λ, the accuracies of NSI and LP hardly increase or even decrease from the beginning. However, the accuracy of RSP index can be improved in all networks. In most cases, when $\lambda = 0$, our RSP index has better performance than NSI and LP. This advantage becomes more significant when $\lambda = +\infty$, i.e. the paths with length 3 are only considered. This result demonstrates that our RSP index has the effect of employing the interactions among paths for link prediction. According to the results shown in Fig. 2 and Fig. 3 (i.e. vertical green lines at $\lambda = 0.1$), regardless of the metrics of AUC or Precision, we also find that $\lambda = 0.1$ always achieves the approximately optimal performance for RSP index in all networks. The performance of RSP with $\lambda = 0.1$ will be further analysed in the following section.

5.3. The prediction accuracy of RSP index compared with other baselines

To validate the prediction accuracy of RSP index, we compare it with other nine baseline indices in twelve networks, including four local similarity indices: CN, RA, AA and CRA, two global similarity indices: Katz and KDLP, as well as three quasi-local similarity indices: LP, PE and NSI. The prediction results of these indices are shown in Tables 2 and 3, respectively. The top-2 highest accuracies in each network are emphasized in black.

Table 2 shows the accuracies measured by AUC. Compared with other nine baselines, RSP index has the top two best performance on eight out of twelve networks. With the exception of Katz and KDLP, our RSP index performs best or nearly best in all networks in most cases and also provides competitive prediction accuracy compared with two global indices (Katz and KDLP). Since the metric of AUC concentrates on the macroscopic accuracy of all non-observed links, as the global prediction approaches, Katz and KDLP indices are of great benefit to resolve the problem of "degeneracy of the states" [38] and make node similarity scores more distinguishable, therefore these two methods have good performance. Compared to Katz index, KDLP index has the better performance in most cases. This is because KDLP index not only considers the number of all paths of any length between two nodes, but also emphasizes the different contributions of paths with same length. Among four quasi-local methods, RSP index outperforms other three indices. Following RSP, NSI outperforms PE and LP methods. Since AA and RA are variants of CN index, therefore, they have better performance than CN. Although CRA considers the interactions between paths with length 2, yet its performance is worst. This is because this index neglects the contributions of intermediate nodes that do not interact with other common neighbours according to its definition in Eq. (11).

Table 3 presents the accuracies measured by Precision. Our RSP index has an obvious advantage over other nine indices in all networks, particularly in the networks of SmaGri, Blogs, Yeast, Kohonen and Erdos. It has the best performance in nine out



(a) Karate. (b) Dolphins. (c) Football.

(d) Physicians. (e) Celegans. (f) WikiVote.

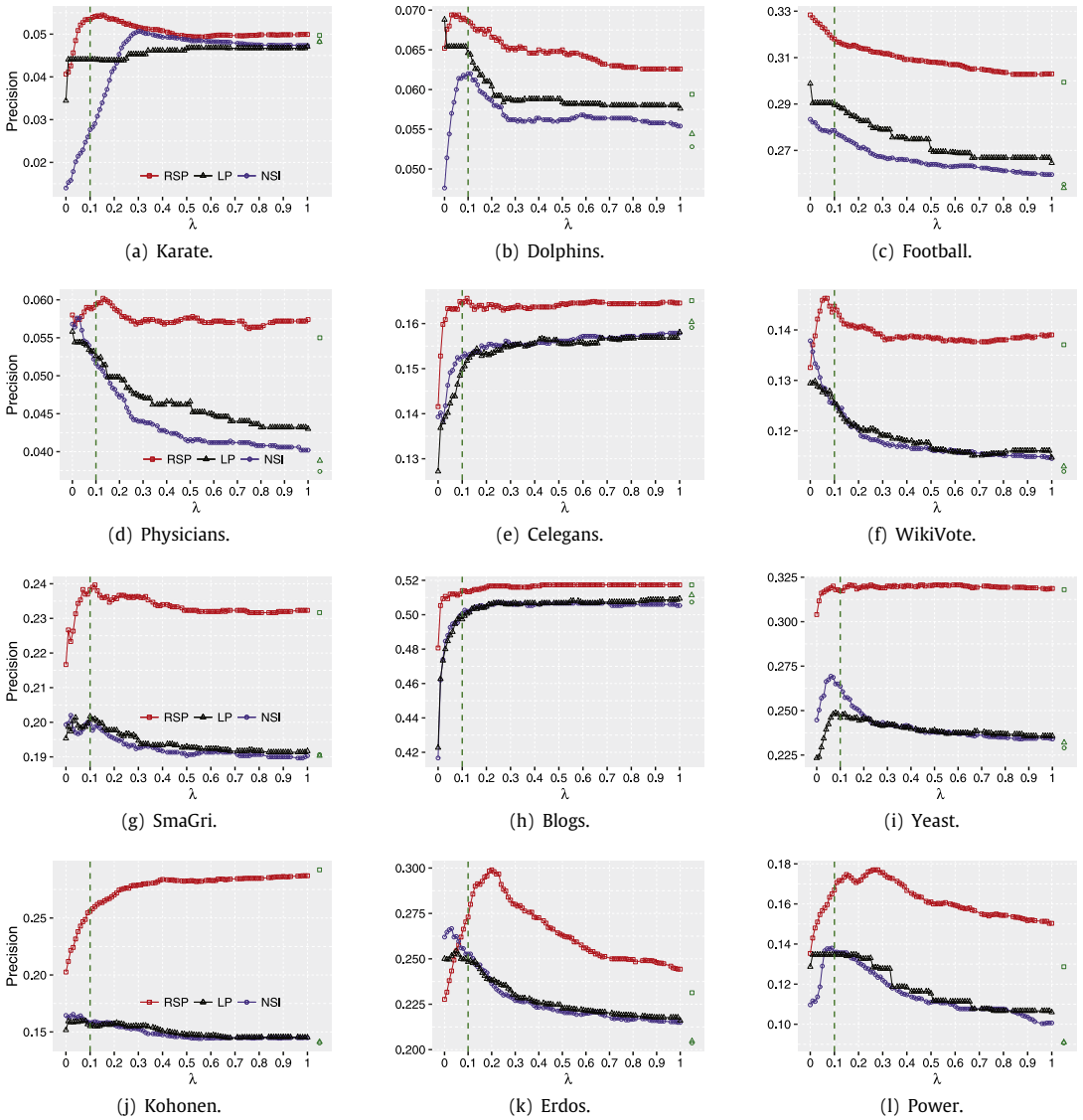
(g) SmaGri. (h) Blogs. (i) Yeast.

(j) Kohonen. (k) Erdos. (l) Power.

Fig. 2. The prediction performance of RSP, NSI and LP indices measured by AUC in twelve networks with different values of λ. Each point represents the average value over 50 independent implementations. The green points of square, circle and triangle on the far right in each figure are the performance of RSP, NSI and LP indices that depend only on the paths with length 3, respectively.

of twelve networks and also has the second-best performance in rest three networks. In comparison with the AUC results, two global indices have poor performance especially for KDLP index which always has the worst performance in all networks in most cases. The reason is that the metric of Precision only focuses on the microscopic accuracy of top ranked links, Katz and KDLP take into account all paths between two nodes, which results in much noise information introduced by long paths in prediction procedure and affects the rankings of missing links. Overall, the performance of quasi-local methods is better than that of local methods, since the quasi-local methods exploit more structure information than local methods. Note that CRA index outperforms CN, RA and AA in most cases and also has the best or nearly best performance in Football, SmaGri, Yeast and Power networks except for RSP. This is because the LCP structure features of these networks are significant, e.g., the LCP_{corr} of SmaGri and Blogs correspond to 0.946 and 0.929 (see Table 1). For the networks with low LCP_{corr}, such as Karate (0.756), Physicians (0.791) and Erdos (0.792), CRA index generally has the worst performance, whereas our RSP index always performs best. This result indicates that it is beneficial to improve prediction accuracy via considering the interactions among paths.

As mentioned in the above Section 5.2, the parameter $\lambda = 0.1$ always corresponds to the approximately optimal performance for RSP index. Therefore, in Tables 2 and 3, we also show the accuracies of RSP index when $\lambda = 0.1$ and



(a) Karate. (b) Dolphins. (c) Football.

(d) Physicians. (e) Celegans. (f) WikiVote.

(g) SmaGri. (h) Blogs. (i) Yeast.

(j) Kohonen. (k) Erdos. (l) Power.

Fig. 3. The prediction performance of RSP, NSI and LP indices measured by Precision in twelve networks with different values of λ. Each point represents the average value over 50 independent implementations. The green points of square, circle and triangle on the far right in each figure are the performance of RSP, NSI and LP indices that depend only on the paths with length 3, respectively.

denote $RSP_{0.1}$ for the sake of convenience. As shown in these two tables, the performance of $RSP_{0.1}$ is close to the optimal performance and has a competitive prediction performance compared with other baselines. In addition, due to the parameter dependence of Precision metric, i.e. L in Eq. (18), we plot the performance of $RSP_{0.1}$ in Fig. 4 when L ranges from 10 to 100 at 10 intervals. As shown in this figure, $RSP_{0.1}$ has a better performance than other methods in most case, in particular for the networks of Physicians, WikiVote, SmaGri, Blogs, Yeast, Kohonen and Erdos. Although KDLP index also measures the different roles of nodes through H-index, its accuracies always perform worst in most cases. According to the results of AUC and Precision, compared with other nine methods, $RSP_{0.1}$ has a good performance. This find can enhance the application value of our RSP index, since it is time-consuming to search the optimal parameter λ for different networks.

In summary, our RSP index has better performance than other nine baselines due to three reasons. First, RSP index considers more topological structure information of networks than local similarity indices, i.e. paths with length 2 and 3. Whereas the local similarity indices (CN, RA, AA and CRA) only use the information of common neighbours, i.e. paths with length 2. Second, considering the interactions among paths with different length, RSP index distinguishes the different contributions of paths even with the same length. Although LP also considers the paths with 2 and 3, it neglects the different contributions of paths with same length. As two heterogeneity path methods, PE and NSI focus on the different contributions of paths with specific length but neglect the interaction relationship among paths with different length. Third, compared to

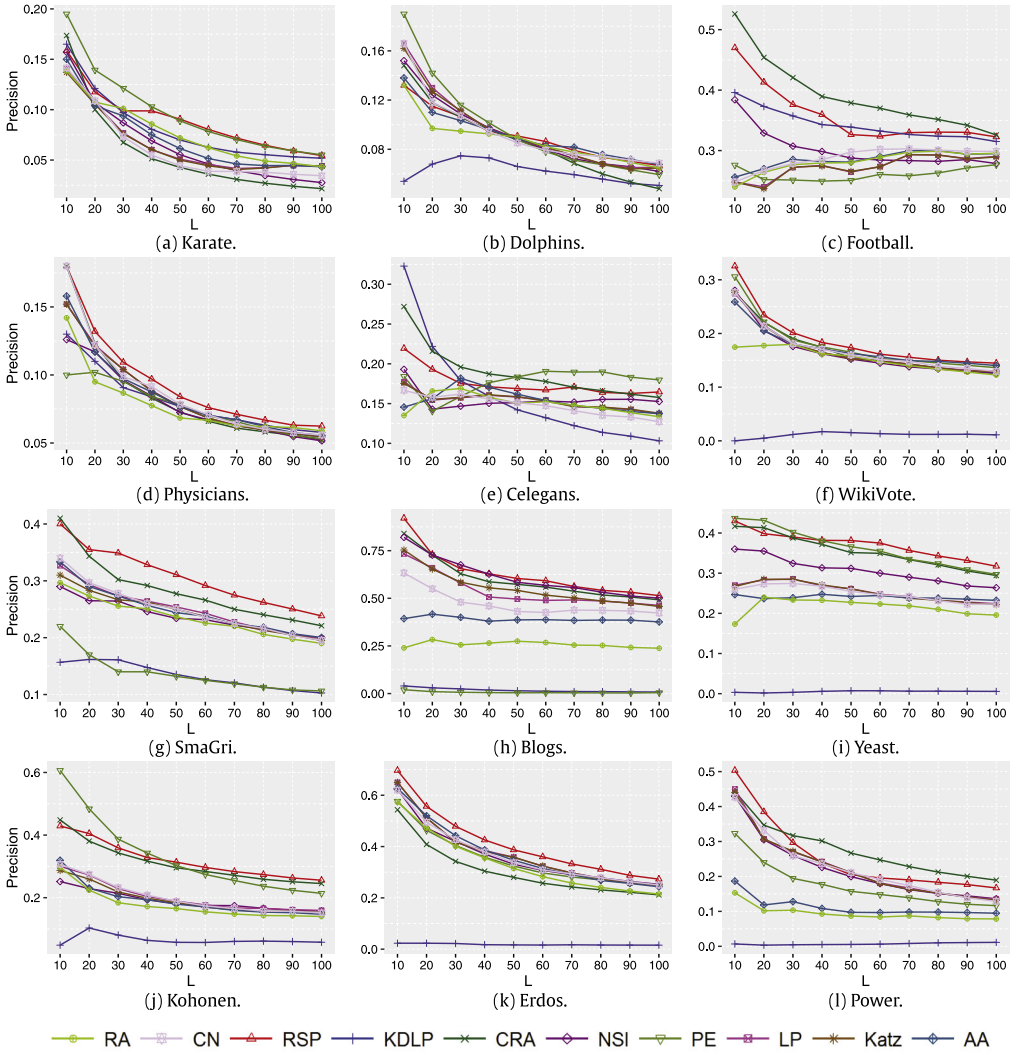


Fig. 4. Prediction performance of ten indices measured by Precision with the changing of L in twelve networks. Each point is averaged over 50 independent implementations. The parameters of Katz, KDLP, LP, NSI and RSP are set to 0.01, 0.01, 0.01, 0.1 and 0.1, respectively.

global methods (Katz and KDLP), RSP index only takes into account local paths with length 2 and 3, which avoids introducing much noise information in prediction procedure and reduces the computational complexity at the same time.

6. Conclusion

In this paper, we develop a path-dependent RSP index, which considers the interactions of paths with different length based on the resource-traffic flow mechanism on networks. The contribution of one path is treated as a process of receiving resources from short paths, which are limited to paths with length 2 due to computational complexity. To measure the different roles of nodes, with the aid of knowledge dissemination mechanism, the knowledge quantity is regarded as the initial resources that each node takes. For a path between a pair of nodes, if its intermediate nodes receive more resources from nodes on paths with length 2, this path is considered having significant contribution to node similarity. Experimental results on twelve real-world networks demonstrate that our RSP index has better performance than other nine prediction indices measured by AUC and Precision. With respect to the networks with low LCP_{corr}, our index also has a significant advantage compared to the LCP-based index, i.e. CRA. Furthermore, although our index depends on a tunable parameter to weight long paths, yet this parameter can be approximately fixed at a constant value, which enhances the application value of our index. For RSP index, the H-index of one node is employed as its initial resources, it is worth exploring other possible ways to measure the initial resources of one node for performance improvement in the future.

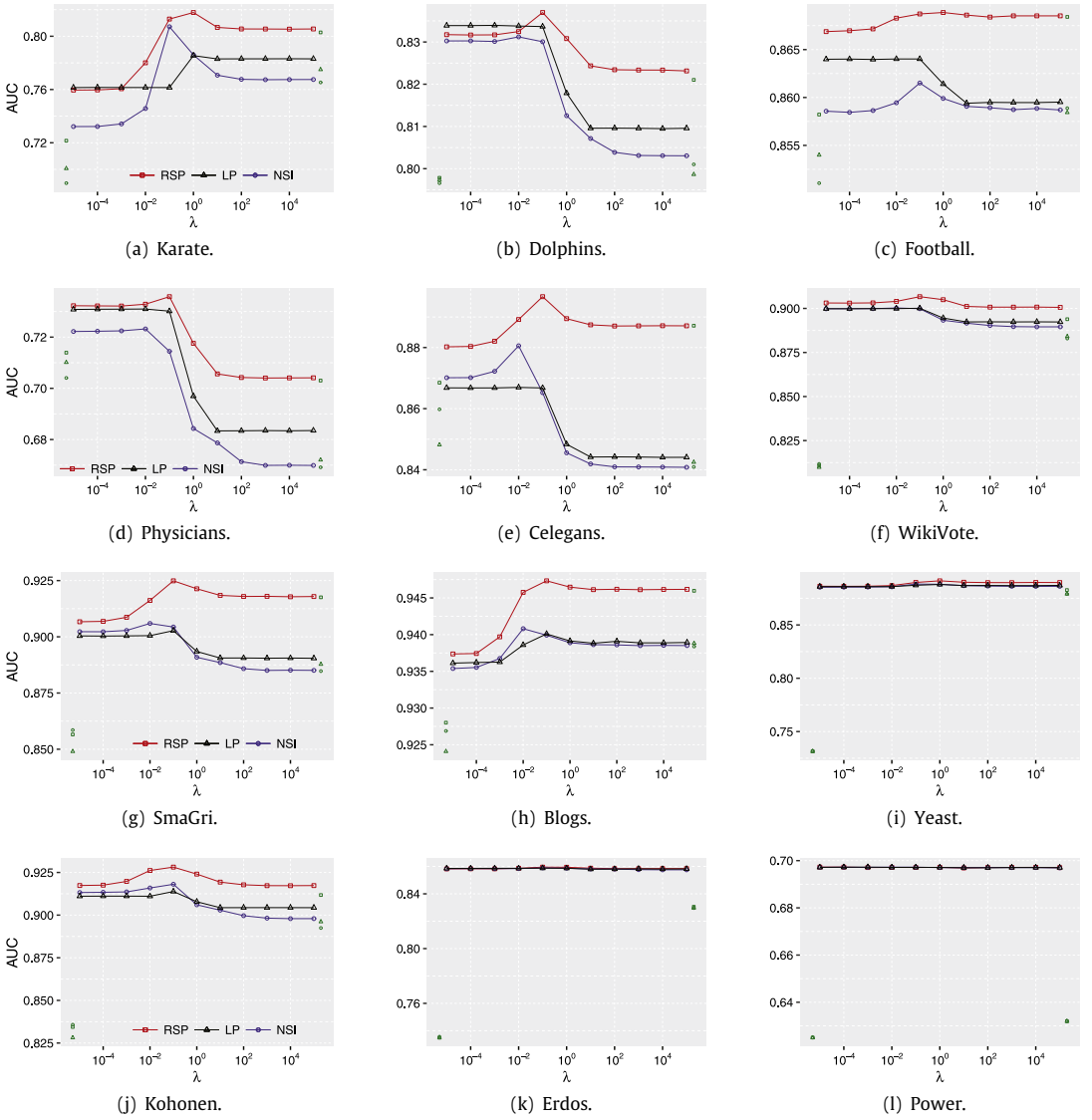


Fig. A.1. The prediction performance of RSP, NSI and LP indices measured by AUC in twelve networks with different values of λ. Each point represents the average value over 50 independent implementations. The green points on the far left and the far right in each figure are the performance that depends only on the paths with length 2 and 3, respectively. The points of square, circle and triangle correspond to the performance of RSP, NSI and LP indices, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (No. 21503101), the Natural Science Foundation of Gansu Province, China (No. 1506RJZA223), the Project-sponsored by SRF for ROCS, SEM, China (No. SEM[2015]311) and the Scientific Research Projects of Gansu Colleges and Universities, China (No. 2017A-106).

Appendix. The optimal scope of parameter λ for RSP index

In order to get the optimal value of parameter λ, we explore the prediction accuracies of RSP index when λ ranges from 10^{-5} to 10^{5} with a ratio of 10. Figs. A.1 and A.2 show the results according to AUC and Precision, respectively. As shown in the figures, with the changing of λ, we can always get the optimal accuracies of RSP index when λ changes from 0 to 1. This results indicate that the parameter λ can be fixed at an interval of [0, 1] to reduce the searching time of optimal parameter λ for RSP index.

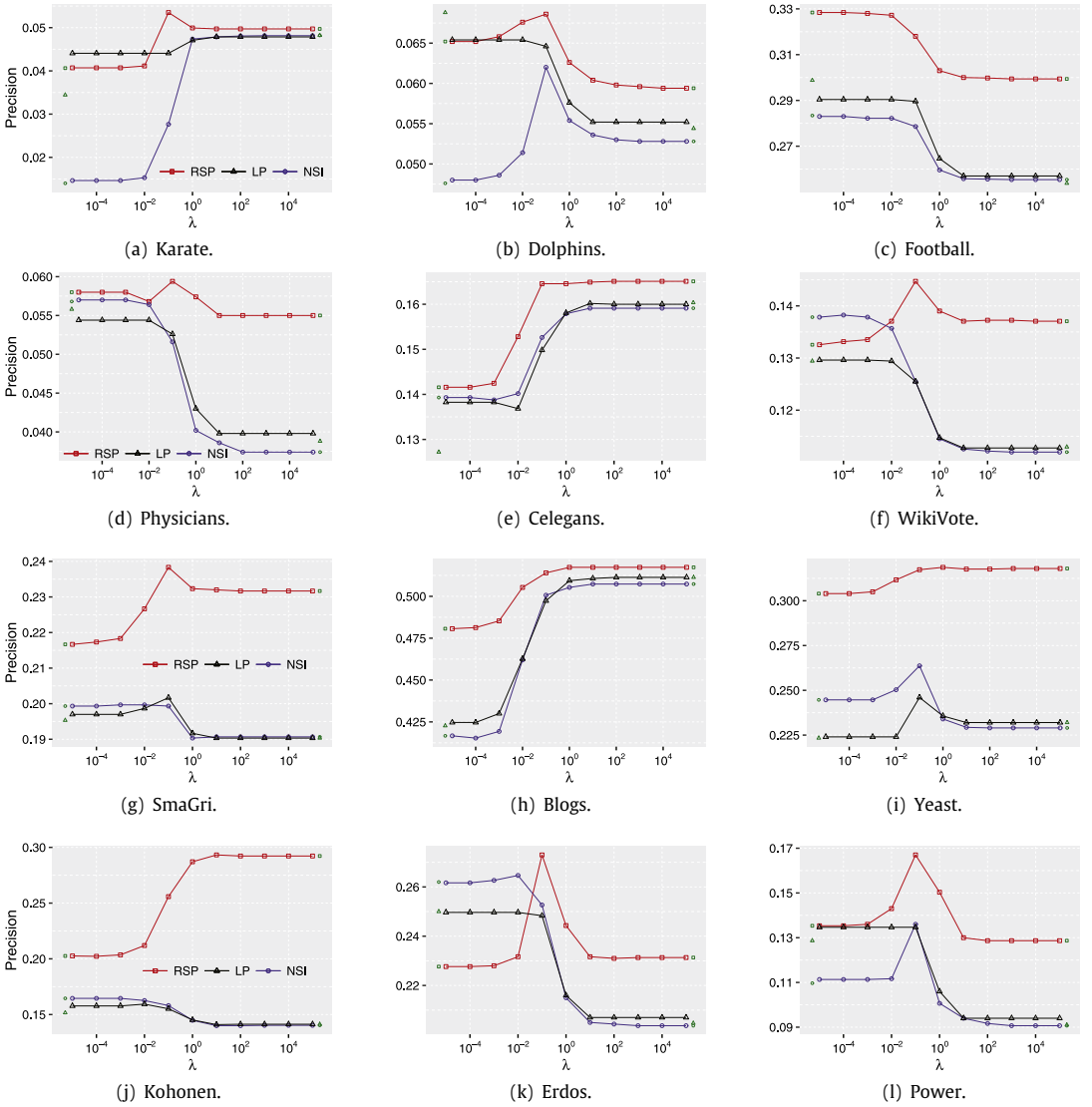


Fig. A.2. The prediction performance of RSP, NSI and LP indices measured by Precision (top-100) in twelve networks with different values of λ. Each point represents the average value over 50 independent implementations. The green points on the far left and the far right in each figure are the performance that depends only on the paths with length 2 and 3, respectively. The points of square, circle and triangle correspond to the performance of RSP, NSI and LP indices, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

References

[1] A.-L. Barabási, The network takeover, Nat. Phys. 8 (1) (2012) 14–16.
[2] A.-L. Barabási, Network science, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 371 (1987) (2013) 20120375.
[3] B. Yan, S. Gregory, Finding missing edges in networks based on their community structure, Phys. Rev. E 85 (5) (2012) 056112.
[4] P. Zhang, X. Wang, F. Wang, A. Zeng, J. Xiao, Measuring the robustness of link prediction algorithms under noisy environment, Sci. Rep. 6 (2016) 18881.
[5] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, J. Assoc. Inform. Sci. Technol. 58 (7) (2007) 1019–1031.
[6] V. Martínez, F. Berzal, J.-C. Cubero, A survey of link prediction in complex networks, ACM Comput. Surv. 49 (4) (2016) 69.
[7] J. Menche, A. Sharma, M. Kitsak, S.D. Ghiassian, M. Vidal, J. Loscalzo, A.-L. Barabási, Uncovering disease-disease relationships through the incomplete interactome, Science 347 (6224) (2015) 1257601.
[8] Y. Lu, Y. Guo, A. Korhonen, Link prediction in drug-target interactions network using similarity indices, BMC Bioinformatics 18 (1) (2017) 39.
[9] L.M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, F. Menczer, Friendship prediction and homophily in social media, ACM Trans. Web (TWEB) 6 (2) (2012) 9.
[10] D. Li, Y. Zhang, Z. Xu, D. Chu, S. Li, Exploiting information diffusion feature for link prediction in sina weibo, Sci. Rep. 6 (2016) 20058.
[11] L. Lü, T. Zhou, Link prediction in complex networks: a survey, Physica A 390 (6) (2011) 1150–1170.
[12] V. Ciotti, M. Bonaventura, V. Nicosia, P. Panzarasa, V. Latora, Homophily and missing links in citation networks, EPJ Data Sci. 5 (1) (2016) 7.

[13] X. Li, H. Chen, Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach, Decis. Support Syst. 54 (2) (2013) 880–890.
[14] Z.L. Li, X. Fang, O.R.L. Sheng, A survey of link recommendation for social networks: methods, theoretical foundations, and future research directions, ACM Trans. Manage. Inf. Syst. 9 (1) (2017) 1:1–1:26.
[15] W. Cukierski, B. Hamner, B. Yang, Graph-based features for supervised link prediction, in: Neural Networks (IJCNN), The 2011 International Joint Conference on, IEEE, 2011, pp. 1237–1244.
[16] R.N. Lichtenwalter, J.T. Lussier, N.V. Chawla, New perspectives and methods in link prediction, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2010, pp. 243–252.
[17] A. Clauset, C. Moore, M.E. Newman, Hierarchical structure and the prediction of missing links in networks, Nature 453 (7191) (2008) 98–101.
[18] R. Guimerà, M. Sales-Pardo, Missing and spurious interactions and the reconstruction of complex networks, Proc. Natl. Acad. Sci. 106 (52) (2009) 22073–22078.
[19] Z. Liu, J.L. He, K. Kapoor, J. Srivastava, Correlations between community structure and link formation in complex networks, Plos One 8 (9) (2013) e72908.
[20] A.K. Menon, C. Elkan, Link prediction via matrix factorization, in: Machine Learning and Knowledge Discovery in Databases - European Conference, Ecml Pkdd 2011, Athens, Greece, September 5–9, 2011, Proceedings, 2011, pp. 437–452.
[21] W. Wang, C. Fei, P. Jiao, P. Lin, A perturbation-based framework for link prediction via non-negative matrix factorization, Sci. Rep. 6 (2016) 38938.
[22] L. Lü, C.-H. Jin, T. Zhou, Similarity index based on local paths for link prediction of complex networks, Phys. Rev. E 80 (4) (2009) 046122.
[23] A. Papadimitriou, P. Symeonidis, Y. Manolopoulos, Fast and accurate link prediction in social networking systems, J. Syst. Softw. 85 (9) (2012) 2119–2132.
[24] Q. Ou, Y.-D. Jin, T. Zhou, B.-H. Wang, B.-Q. Yin, Power-law strength-degree correlation from resource-allocation dynamics on weighted networks, Phys. Rev. E 75 (2) (2007) 021102.
[25] W. Zhou, Y. Jia, Predicting links based on knowledge dissemination in complex network, Physica A 471 (2017) 561–568.
[26] F. Aghabozorgi, M.R. Khayyambashi, A new similarity measure for link prediction based on local structures in social networks, Physica A 501 (2018) 12–23.
[27] L. Pan, Z. Tao, L. LÃ, C.K. Hu, Predicting missing links and identifying spurious links via likelihood analysis, Sci. Rep. 6 (2016) 22955.
[28] P. Jiao, F. Cai, Y. Feng, W. Wang, Link predication based on matrix factorization by fusion of multi class organizations of the network, Sci. Rep. 7 (1) (2017) 8937.
[29] R. Pech, D. Hao, L. Pan, H. Cheng, T. Zhou, Link prediction via matrix completion, Eur. Phys. Lett. 117 (3) (2017) 38002.
[30] X. Ma, P. Sun, Y. Wang, Graph regularized nonnegative matrix factorization for temporal link prediction in dynamic networks, Physica A 496 (2018) 121–136.
[31] Z. Yin, M. Gupta, T. Weninger, J. Han, Linkrec:a unified framework for link recommendation with user attributes and graph structure, in: International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, Usa, April, 2010, pp. 1211–1212.
[32] Z. Wang, J. Liang, R. Li, Y. Qian, An approach to cold-start link prediction: establishing connections between non-topological and topological information, IEEE Trans. Knowl. Data Eng. 28 (11) (2016) 2857–2870.
[33] T. Li, J. Wang, M. Tu, Y. Zhang, Y. Yan, Enhancing link prediction using gradient boosting features, in: Intelligent Computing Theories and Application, Springer International Publishing, 2016, pp. 81–92.
[34] E. Bastami, A. Mahabadi, E. Taghizadeh, A gravitation-based link prediction approach in social networks, Swarm Evol. Comput. (2018) Available online.
[35] F. Lorrain, H.C. White, Structural equivalence of individuals in social networks, J. Math. Sociol. 1 (1) (1971) 49–80.
[36] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.
[37] L.A. Adamic, E. Adar, Friends and neighbors on the web, Soc. Netw. 25 (3) (2003) 211–230.
[38] T. Zhou, L. Lü, Y.-C. Zhang, Predicting missing links via local information, Eur. Phys. J. B 71 (4) (2009) 623–630.
[39] F. Tan, Y. Xia, B. Zhu, Link prediction in complex networks: a mutual information perspective, PLoS One 9 (9) (2014) e107056.
[40] C.V. Cannistraci, G. Alanis-Lobato, T. Ravasi, From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks, Sci. Rep. 3 (2013) 1613.
[41] F. Li, J. He, G. Huang, Y. Zhang, Y. Shi, R. Zhou, Node-coupling clustering approaches for link prediction, Knowl.-Based Syst. 89 (C) (2015) 669–680.
[42] Z. Wu, Y. Lin, J. Wang, S. Gregory, Link prediction with node clustering coefficient, Physica A 452 (2016) 1–8.
[43] Z. Wu, Y. Lin, H. Wan, W. Jamil, Predicting top-l missing links with node and link clustering information in large-scale networks, J. Stat. Mech. Theory Exp. (8) (2016) 083202.
[44] Z. Wu, Y. Lin, Y. Zhao, H. Yan, Improving local clustering based top-l link prediction methods via asymmetric link clustering information, Physica A 492 (2018) 1859–1874.
[45] L. Katz, A new status index derived from sociometric analysis, Psychometrika 18 (1) (1953) 39–43.
[46] E.A. Leicht, P. Holme, M.E. Newman, Vertex similarity in networks, Phys. Rev. E 73 (2) (2006) 026120.
[47] G. Jeh, J. Widom, SimRank: a measure of structural-context similarity, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2002, pp. 538–543.
[48] W. Zheng, L. Zou, L. Chen, D. Zhao, Efficient simrank-based similarity join, ACM Trans. Database Syst. 42 (3) (2017) 16.
[49] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Comput. Netw. ISDN Syst. 30 (1) (1998) 107–117.
[50] W. Liu, L. Lü, Link prediction based on local random walk, Europhys. Lett. 89 (5) (2010) 58007.
[51] Q. Zhang, M. Li, Y. Deng, Measure the structure similarity of nodes in complex networks based on relative entropy, Physica A 491 (2017) 749–763.
[52] J. Ding, L. Jiao, J. Wu, Y. Hou, Y. Qi, Prediction of missing links based on multi-resolution community division, Physica A 417 (2015) 76–85.
[53] J. Ding, L. Jiao, J. Wu, F. Liu, Prediction of missing links based on community relevance and ruler inference, Knowl.-Based Syst. 98 (2016) 200–215.
[54] Z. Wang, Y. Wu, Q. Li, F. Jin, W. Xiong, Link prediction based on hyperbolic mapping with community structure for complex networks, Physica A 450 (2016) 609–623.
[55] L. Lü, L. Pan, T. Zhou, Y.C. Zhang, H.E. Stanley, Toward link predictability of complex networks, Proc. Natl. Acad. Sci. USA 112 (8) (2015) 2325–2330.
[56] T. Wang, X.S. He, M.Y. Zhou, Z.Q. Fu, Link prediction in evolving networks based on popularity of nodes, Sci. Rep. 7 (1) (2017) 7147.
[57] Z. Peng, W. Xiang, F. Wang, Z. An, J. Xiao, Measuring the robustness of link prediction algorithms under noisy environment, Sci. Rep. 6 (2016) 18881.
[58] X. Zhu, H. Tian, S. Cai, J. Huang, T. Zhou, Predicting missing links via significant paths, Europhys. Lett. 106 (1) (2014) 18008.
[59] X. Zhu, H. Tian, S. Cai, Predicting missing links via effective paths, Physica A 413 (2014) 515–522.
[60] B. Zhu, Y. Xia, An information-theoretic model for link prediction in complex networks, Sci. Rep. 5 (2015) 13707.
[61] P. Pei, B. Liu, L. Jiao, Link prediction in complex networks based on an information allocation index, Physica A 470 (2017) 1–11.
[62] S. Liu, X. Ji, C. Liu, Y. Bai, Extended resource allocation index for link prediction of complex network, Physica A 479 (2017) 174–183.
[63] Y. Yang, J. Zhang, X. Zhu, L. Tian, Link prediction via significant influence, Physica A 492 (2018) 1523–1530.
[64] Z. Xu, C. Pu, J. Yang, Link prediction based on path entropy, Physica A 456 (2016) 294–301.
[65] L. Lü, T. Zhou, Q.-M. Zhang, H.E. Stanley, The h-index of a network node and its relation to degree and coreness, Nature Commun. 7 (2016) 10168.
[66] M.E. Newman, Clustering and preferential attachment in growing networks, Phys. Rev. E 64 (2) (2001) 025102.
[67] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve., Radiology 143 (1) (1982) 29–36.

[68] P. Pudil, J. Novovičová, J. Kittler, Floating search methods in feature selection, Pattern Recognit. Lett. 15 (11) (1994) 1119–1125.
[69] J.L. Herlocker, J.A. Konstan, L.G. Terveen, J.T. Riedl, Evaluating collaborative filtering recommender systems, ACM Trans. Inf. Syst. (TOIS) 22 (1) (2004) 5–53.
[70] W.W. Zachary, An information flow model for conflict and fission in small groups, J. Anthropol. Res. 33 (4) (1977) 452–473.
[71] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, S.M. Dawson, The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations, Behav. Ecol. Sociobiol. 54 (4) (2003) 396–405.
[72] M. Girvan, M.E. Newman, Community structure in social and biological networks, Proc. Nat. Acad. Sci. 99 (12) (2002) 7821–7826.
[73] J. Coleman, E. Katz, H. Menzel, The diffusion of an innovation among physicians, Sociometry 20 (4) (1957) 253–270.
[74] J.G. White, E. Southgate, J.N. Thomson, S. Brenner, The structure of the nervous system of the nematode caenorhabditis elegans, Philos. Trans. R. Soc. Lond. B Biol. Sci. 314 (1165) (1986) 1–340.
[75] R.A. Rossi, N.K. Ahmed, The network data repository with interactive graph analytics and visualization, in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015, http://networkrepository.com.
[76] J. Leskovec, D. Huttenlocher, J. Kleinberg, Signed networks in social media, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2010, pp. 1361–1370.
[77] N.P. Hummon, P. Dereian, Connectivity in a citation network: The development of DNA theory, Soc. Netw. 11 (1) (1989) 39–63.
[78] L.A. Adamic, N. Glance, The political blogosphere and the 2004 US election: divided they blog, in: Proceedings of the 3rd International Workshop on Link Discovery, ACM, 2005, pp. 36–43.
[79] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, et al., Topological structure analysis of the protein–protein interaction network in budding yeast, Nucl. Acids Res. 31 (9) (2003) 2443–2450.
[80] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (6684) (1998) 440–442.