Article

JIS

Journal of Information Science
2015, Vol. 41(2) 167–177

DOI: 10.1177/0165551514560121
jis.sagepub.com
SAGE

Accurate similarity index based on the contributions of paths and end nodes for link prediction

Longjie Li
School of Information Science and Engineering, Lanzhou University, China

Lvjian Qian
School of Information Science and Engineering, Lanzhou University, China

Jianjun Cheng
School of Information Science and Engineering, Lanzhou University, China

Min Ma
School of Information Science and Engineering, Lanzhou University, China

Xiaoyun Chen
School of Information Science and Engineering, Lanzhou University, China

Abstract

Link prediction whose intent is to discover the likelihood of the existence of a link between two disconnected nodes is an important task in complex network analysis. To perform this task, a similarity-based algorithm that employs the similarities of nodes to find links is a very popular solution. However, when calculating the similarity between two nodes, most of the similarity-based algorithms only focus on the contributions of paths connecting these two nodes but ignore the influences of these two nodes themselves. Therefore, their results are not accurate enough. In this paper, a novel similarity index, called Scop, is proposed for link prediction. By directly defining the contributions of paths to their end nodes and the contributions of end nodes themselves, Scop not only distinguishes the contributions of different paths but also integrates the contributions of end nodes. Hence, Scop can obtain better performance on accuracy. Experiments on 10 networks compared with six baselines indicate that Scop is remarkably better than others.

Keywords

Complex network; end node contribution; link prediction; path contribution; similarity index

1. Introduction

The purpose of link prediction is to find the missing links or predict the emergence of new links that do not present currently in a complex network. For some networks, in particular, biological networks, such as protein–protein interaction networks and metabolic networks, the current knowledge about those networks is substantially incomplete, and discovering missing links is expensive in laboratories. For others, for example a social network, very likely but not yet existent links can be recommended to users as a promising friendship proposal [1]. Basically, link prediction has the following three functions. First of all, it is applied to settle a missing data problem. Take protein–protein interaction networks as

Corresponding author:
Xiaoyun Chen, School of Information Science and Engineering, Lanzhou University, China.
Email: chenxy@lzu.edu.cn

an example. Owing to the limitation of our knowledge and the very large scale of those networks, it is too difficult to find missing data within our capacity. Thus, link prediction methods indeed play an inportant role. Second, the data in constructing biological and social networks may contain inaccurate information and hence result in spurious links [2, 3]. Link prediction, as an inexpensive and competent method, can be applied to identifying these spurious links [4]. Finally, link prediction algorithms can be utilized to predict the links that may appear in the future along with the evolution of networks. Instead of blindly checking all possible interactions, predicting based on the observed structural contexts and focusing on most likely existent links can sharply reduce the experimental costs if the prediction method is accurate enough [5]. Therefore, developing accurate link prediction methods based on the known structural contexts plays a vital role.

Link prediction, which is a fundamental task in link mining and complex network analysis, has a wide range of applications, such as recommend systems, information retrieval and bioinformatics. Many algorithms have been proposed from different disciplines. Some are based on Markov chains [6–8] or machine learning [9, 10], and a series of algorithms are based on node similarity [11, 12].

In this paper, we mainly focus on the similarity-based algorithms. To some extent, these algorithms are intuitively based on the idea that *two nodes are considered to be similar if they share many common neighbours*. According to Lü and Zhou [11], similarity-based algorithms can be categorized into three classes: local similarity index, global similarity index and quasi-local similarity index. The *Common Neighbours index* [13], which simply counts the number of the common neighbours between two disconnected nodes, is a typical local index. The *Jaccard index* [14], which was proposed over 100 years ago, can be considered as a normalized version of Common Neighbours index. It is defined as the number of common neighbours divided by the number of all distinct neighbours of these two nodes. Both Common Neighbours and Jaccard indices do not distinguish the different contributions of their different common neighbours, and hence treat each common neighbour equally. To alleviate this weakness, some indices with a penalization of large-degree common neighbours have been proposed, such as the *Adamic-Adar index* [15] and the *Resource Allocation index* [16]. The main advantage of local similarity indices is the best performance in efficiency. However, only taking the immediate neighbours into consideration makes this index suffer from the relatively poor performance in prediction. In contrast, the global similarity index achieves a better prediction performance but bears a higher computational complexity. The *Katz index* [17] and the *SimRank index* [18] are two famous global similarity indices. The Katz index makes use of all paths connecting two nodes with a penalization of longer paths. The SimRank index is a self-consistent method, based on the intuition that *two nodes are similar if they are linked by similar nodes*. When calculating the similarity score between two different nodes, SimRank adopts the average similarity score of all their neighbour pairs. The quasi-local index provides a good tradeoff of performance on accuracy and computational complexity between local and global indices. The *Local Path index* [1], the *Local Random Walk index* [19], the *FriendLink index* [20] and the *SRank index* [21] are several quasi-local methods. The Local Path index takes the idea of the Katz index but limits the length of paths under consideration. When computing similarities in FriendLink, long paths are ignored and each path is penalized with a structural coefficient according to its length. The Local Random Walk index is a constrained random walk-based method by limiting the range that a random walker can surf. SRank is a typical shortest paths similarity measure. In SRank, two nodes are considered as similar if there are multiple small-length paths connecting them. Most recently, the Significant Path index was proposed by Zhu *et al.*[22]. The basic idea of this index comes from the intuition that short paths including small-degree intermediate nodes make stronger evidence of a missing link connecting its two ends. In practice, the Significant Path index only employs the paths with lengths 2 and 3. Experimental results in Zhu *et al.*[22] showed that the index outperforms many existing methods.

Despite its good performance on accuracy, the Significant Path index did not fully exploit the structure information of a network. One important part which can further improve its performance on accuracy is neglected. It is the contribution that end nodes themselves make to their similarity scores. Take the network shown in Figure 1 as an example. The respective paths connecting node pairs (a, b) and (x, y) are equivalent; hence the Significant Path index will assign the same similarity scores to (a, b) and (x, y). However, this result is visually incorrect. In our viewpoint, the similarity of (a, b) should be greater than that of (x, y) since the contribution that (a, b) makes to their similarity is greater than the counterpart that (x, y) makes. In fact, this situation is not considered by not only the Significant Path index but also many other similarity indices, such as the Common Neighbours index and the Local Path index. Additionally, the Significant Path index has a very serious problem, which is that its parameters are very difficult to set. In Zhu *et al.*[22], a large number of combinations of parameters α and β were tried to find the optimal one for each network.

To solve the aforementioned issues, in this paper, we propose a new node similarity index, called *Scop* (short for Similarity based on the Contributions of end nOdes and Paths), for link prediction. When computing the similarity between two different nodes, the proposed index takes into account not only the contributions of paths connecting these two nodes but also the contributions of these two end nodes themselves. In Scop, the contributions of different paths and

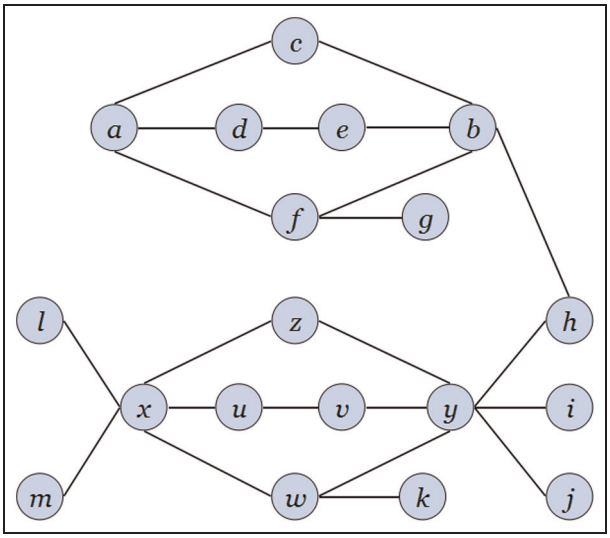

Figure 1. An example network to describe the Scop index.

end nodes are distinguished. Therefore, this index can assign a more accurate similarity score to any node pair. Although two parameters, that is, damping factor d and path length L, are also employed in Scop, their values can be set conveniently. We use 10 real-world networks to validate the proposed index compared with six baselines: the Common Neighbours index, the Adamic-Adar index, the Resource Allocation index, the Local Path index, the FriendLink index and the Significant Path index. Experimental results show that our index performs far better than the others.

The contributions of this paper are summarized below. First, a common issue that exists in many previous similarity indices is figured out. That is that the contributions made by end nodes themselves to their similarity scores were not taken into account. Second, a novel similarity index, Scop, is proposed, which not only considers the contributions of both end nodes and paths connecting these two nodes, but also differentiates the contributions of different end nodes and paths. Therefore, Scop can assign more accurate similarity to any node pair than many previous methods. Finally, we experimentally study the effectiveness of Scop compared with six other similarity indices on 10 real-world networks.

The rest of this paper is organized as follows. In Section 2, we review some important link prediction methods. The definition of the proposed similarity index is described in Section 3. The experimental results are demonstrated in Section 4. Finally, we conclude our work in Section 5.

2. Similarity-based link prediction methods

A complex network is modelled as an undirected graph $G(V, E)$ with V as the node set and E as the edge set. For a node $u \in V$, notations $\Gamma(u)$ and k_u denote the neighbour set and degree of node u, respectively. Given two nodes u and v in G, notations s_{uv} and s_{uv}^M represent the real similarity and the similarity measured by method M between u and v, respectively.

Link prediction is a well-known task of link mining in complex networks, which tries to infer new interactions among individuals that are likely to occur in the near future. The majority of previous link prediction methods pay attention to measuring the similarities between disconnected nodes at present to predict new connections in the future. Generally speaking, these methods are called similarity-based methods. The basis of these methods is to define the similarity index between nodes.

In this section, we briefly introduce some important similarity indices.

Common Neighbour index (CN) [13]. This index is based on the basic idea that two individuals are more likely to develop an interaction in the future if they have many common neighbours. To measure the similarity between two different nodes, CN directly counts their common neighbours. The definition of this similarity index is defined as follows:

$$s_{uv}^{\mathrm{CN}} = |\Gamma(u) \cap \Gamma(v)| \tag{1}$$

Adamic-Adar index (AA) [15]. This similarity index refines the simple counting of common neighbours by assigning the high-degree neighbours less weight. Compared with CN, AA can differentiate the contributions of different neighbours. Its definition is as follows:

$$s_{uv}^{\mathrm{AA}} = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log(k_z)} \tag{2}$$

Resource Allocation index (RA) [16]. RA can be considered as a revision of AA with the contributions of large-degree neighbours further reduced. RA is defined as follows:

$$s_{uv}^{\mathrm{RA}} = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{k_z} \tag{3}$$

Katz index [17]. This index is a global similarity measure. The idea of the Katz index is that the more paths connect two nodes, the greater the similarity between them is. When computing similarity between two nodes, Katz directly sums all paths connecting these two nodes by exponentially damping the longer paths. The Katz measure is defined as follows:

$$s_{uv}^{\mathrm{Katz}} = \sum_{i=1}^{\infty} \beta^i \left| path_{u,v}^i \right| \tag{4}$$

where $\left| path_{u,v}^i \right|$ is the number of paths connecting u and v with length i, and β is a free parameter to give less weights to longer paths. To ensure that the Katz index is converged, β must be lower than the reciprocal of the largest eigenvalue of the adjacency matrix of the network.

SimRank (SR) [18]. SimRank is a well-known global similarity index based on the structural context. The idea of SimRank is that two nodes are similar if they are linked by similar nodes. Following this idea, when computing the similarity between two different nodes, SimRank averages the similarities of their neighbour pairs. The definition of SimRank is as follows:

$$s_{uv}^{\mathrm{SR}} = \frac{C \times \sum_{x \in \Gamma(u)} \sum_{y \in \Gamma(v)} s_{xy}^{\mathrm{SR}}}{k_u \times k_v} \tag{5}$$

where C is a damping factor ($0 < C < 1$).

Local Path index (LP) [1]. This index is a variant of the Katz index by limiting the length of paths in the range of 2–3. As a result, it provides a good tradeoff between prediction accuracy and computational complexity. Its definition is as follows:

$$s_{uv}^{\mathrm{LP}} = \left| path_{u,v}^2 \right| + \epsilon \left| path_{u,v}^3 \right| \tag{6}$$

where ϵ is similar to β in Katz. When $\epsilon = 0$, this index degenerates to the Common Neighbours index.

FriendLink index (FL) [20]. The similarity in FriendLink method is another variant of Katz index by traversing all paths connecting two nodes within a small length. It is defined as the count of paths of varying length, as:

$$s_{uv}^{\mathrm{FL}} = \sum_{i=2}^{L} \frac{1}{i-1} \frac{\left| P_{uv}^i \right|}{\prod_{j=2}^{i} (|V| - j)} \tag{7}$$

where L represents the maximum length of a path under consideration, P_{uv}^i is the set of paths connecting u and v with length i, and $1/(i-1)$. is a penalty factor that penalizes longer paths.

Significant Path index (SP) [22]. This recently proposed index takes into account not only the count of paths between two different nodes but also the degrees of intermediate nodes in those paths, as:

$$s_{uv}^{\mathrm{SP}} = \sum_{p \in P_{uv}^2} \sum_{z \in M(p)} k_z^{\beta} + \alpha \sum_{q \in P_{uv}^3} \sum_{w \in M(q)} k_w^{\beta} \tag{8}$$

where $M(p)$ indicates the set of intermediate nodes of path p, and α and β are two penalty factors that penalize paths and neighbours, respectively.

3. The proposed method

The core of the similarity-based link prediction method is the computation of similarities between nodes. The higher the similarity score two individuals have, the higher possibly they will become friends. In this section, we propose a new

similarity index, called *Scop*, for link prediction. Before giving the definition of the Scop index in Section 3.2, we first show the motivations inspiring our similarity index in Section 3.1.

3.1. Motivation

The basic idea of the proposed index is motivated by the facts that (a) different paths connecting two nodes usually make different contributions to the similarity score between these two nodes and (b) the contributions of these two nodes themselves should be taken into account. Let us review the network shown in Figure 1. There are three paths connecting nodes a and b, $p_1 = \langle a, c, b\rangle$, $p_2 = \langle a, d, e, b\rangle$ and $p_3 = \langle a, f, b\rangle$. Intuitively, the contributions made by these paths to the similarity between a and b are different. p_2 makes the least contribution since it is the longest path. Although, p_1 and p_3 have the same length, the contribution of p_3 is smaller than that of p_1. The reason for this is that the degree of node f is bigger than that of node c. That is to say, different paths may make different contributions to the similarity of their end nodes. This is the first thing we aim to settle in our similarity measure.

Then, we consider the second thing that affects the similarity between two nodes. It is the status of participation of these two nodes into their common neighbourhood. Usually, this information is ignored by many of the existing indices. As shown in Figure 1, the paths connecting nodes x and y are equivalent to the paths connecting nodes a and b. Thus, many indices assign the same similarity score between x and y with the similarity score between a and b, for example, $s_{xy}^{\mathrm{CN}} = s_{ab}^{\mathrm{CN}} = 2$, $s_{xy}^{\mathrm{RA}} = s_{ab}^{\mathrm{RA}} = 1/2 + 1/3 = 6/5$ and $s_{xy}^{\mathrm{FL}} = s_{ab}^{\mathrm{FL}} = 2/17 + 1/2 \times 1/(17 \times 16) = 0.12$.

However, in our opinion, s_{xy} should be smaller than s_{ab}. Since the respective degrees of end nodes x and y are bigger than those of end nodes a and b, the contributions of x, y to s_{xy} are less than the counterparts of a, b to s_{ab}, respectively. So, in our similarity index, when computing the similarity between two nodes, we also take their own contributions into account.

3.2. Scop index

In this section, we formally give the definition of the Scop index. As stated in Section 3.1, Scop is composed of two parts: path contribution and end node contribution.

Definition 1: Given two nodes x and y in graph $G(V, E)$, P_{xy}^{i} denotes the set of paths connecting x and y with length i. Let $p \in P_{xy}^{i}$ be a path connecting x and y; the contribution of p to the similarity between x and y is defined as:

$$\sigma(p) = \sum\nolimits_{z \in M(p)} min(\pi_{zx}(p), \pi_{zy}(p)) \tag{9}$$

where $M(p)$ is the set of intermediate nodes in path p and $\pi_{zx}(p)$ is the probability from z to x via path p. Suppose $p = \langle x, v_1, \cdots, z = v_m, \cdots, v_{i-2}, y\rangle$, then $M(p) = \{v_1, \cdots, v_{i-2}\}$ and $\pi_{zx}(p) = \prod_{j=m}^{1} d/k_{v_j}$; here $d \in (0, 1]$ is a damping factor.

To some extent, $\pi_{zx}(p)$ denotes the intimacy between nodes z and x via path p. Intuitively, the intimacy between x and y through z is determined by the smaller one between $\pi_{zx}(p)$ and $\pi_{zy}(p)$. Therefore, in formula (9), we use $min(\pi_{zx}(p), \pi_{zy}(p))$. Suppose p and q are two paths connecting x and y with the sane length. According to formula (9), if some corresponding nodes in p and q have different degrees, the contributions of p and q to the similarity between x and y may be different.

Definition 2: The score of Scop between two different nodes x and y, denoted as s_{xy}^{Scop}, is proportional to the sum of contributions of paths connecting x and y, as

$$s_{xy}^{\mathrm{Scop}} \propto \sum\nolimits_{i=2}^{L} \sum\nolimits_{p \in P_{xy}^{i}} \sigma(p) \tag{10}$$

where L denotes the maximum length of a path under consideration.

In formula (10), we do not penalize path length as other similarity indices do, such as the FriendLink index and the Significant Path index. The reason for this is that the damping factor d in Definition 1 will penalize the longer paths. For

the three paths connecting nodes a and b in Figure 1, $p_1 = \langle a, c, b\rangle$, $p_2 = \langle a, d, e, b\rangle$ and $p_3 = \langle a, f, b\rangle$, the respective contributions of these three paths are $\sigma(p_1) = d/2 = 0.3$, $\sigma(p_2) = d^2/4 + d^2/4 = 0.18$ and $\sigma(p_3) = d/3 = 0.2$ with $d = 0.6$. Therefore, $\sigma(p_1) > \sigma(p_3) > \sigma(p_2)$, that is, p_1 makes the most contribution and p_2 makes the least contribution. Next, we consider the second part in the Scop index, that is, end node contribution.

Definition 3: Given two different nodes x and y in graph G, the set of common (intermediate and immediate) neighbours between x and y is defined as:

$$\Gamma^L(x, y) = \left\{ z | z \in M(p), p \in P^i_{xy}, i = 2, \cdots, L \right\} \tag{11}$$

The contribution of node x to the similarity between x and y is defined as:

$$\sigma_{xy}(x) = \frac{\left|\Gamma(x) \cap \Gamma^L(x, y)\right|}{k_x} \tag{12}$$

In formula (12), nodes in $\Gamma(x) \cap \Gamma^L(x, y)$ are the neighbours of x which are simultaneously present in paths connecting x and y. In other words, $\Gamma(x) \cap \Gamma^L(x, y)$ only contain the neighbours of x that directly contribute to the similarity between x and y. Using the example network shown in Figure 1, we get $\sigma_{ab}(a) = 3/3 = 1$, $\sigma_{ab}(b) = 3/4 = 0.75$, $\sigma_{xy}(x) = 3/5 = 0.6$ and $\sigma_{xy}(y) = 3/6 = 0.5$.

Combining the aforementioned two parts, we show the formal computation of the Scop index as follows:

$$s^{\text{Scop}}_{xy} = \frac{\sigma_{xy}(x) + \sigma_{xy}(y)}{2} \times \sum\nolimits_{i=2}^{L} \sum\nolimits_{p \in P^i_{xy}} \sigma(p) \tag{13}$$

Consider the example network in Figure 1 again; we compute the similarity of (a, b) and that of (x, y), respectively, in the light of the Scop index. We obtain $s^{\text{Scop}}_{ab} = (1 + 0.75)/2 \times (0.3 + 0.18 + 0.2) = 0.595$ and $s^{\text{Scop}}_{xy} = (0.6 + 0.5)/2 \times (0.3 + 0.18 + 0.2) = 0.374$. Therefore, $s^{\text{Scop}}_{xy} < s^{\text{Scop}}_{ab}$. This result is reasonable.

4. Evaluation and results

In this section, we experimentally evaluate the performance of Scop compared with CN, AA, RA, LP, FL and SP in terms of efficiency. In Section 4.1, we first introduce the evaluation metrics used in our experiments. Then, the benchmark networks are described in Section 4.2. Finally, Section 4.3 presents the experimental results.

4.1. Evaluation metrics

Given a network $G(V, E)$, multiple links and self-connections are not allowed in G. Let U be the universal set containing all $|V|(|V| - 1)/2$ possible links, where $|V|$ is the number of nodes in V. Then, the set of non-existent links is $U \backslash E$. For each node pair (x, y), a similarity-based link prediction method assigns a similarity score to it. To test the accuracy of link prediction methods, we randomly divide the link set E into two parts: the training set, E^{T}, and the probe set, E^{P}, such that $E = E^{\text{T}} \cup E^{\text{P}}$ and $E^{\text{T}} \cap E^{\text{P}} = \varnothing$. For fair comparisons, we adopt the *K-fold cross-validation* in our experiments. That is, the link set E is randomly partitioned into K subsets. And each time one subset is chosen as the probe set, the rest of $K - 1$ subsets make up the training set. In this paper, K is 10 as suggested in Lü and Zhou [11].

We employ two standard metrics to quantify the accuracy of prediction methods: *AUC* [23] and *Precision* [24]. The value of AUC can be thought of as the probability that a randomly selected link in E^{P} is given a higher score than a randomly selected link in $U \backslash E$. In the implementation, at each time, we select a link in E^{P} and a link in $U \backslash E$ to compare their similarity scores. If among n independent comparisons, there are n' times the missing link having higher score and n'' times the missing link and non-existent link having the same score, the AUC value is defined as follows:

$$\text{AUC} = \frac{n' + 0.5n''}{n} \tag{14}$$

Precision is computed as the ratio of the number of correctly predicted links to the number of missing links. In order to calculate this metric, first, all of the missing and non-existent links are ranked in decreasing order according to their

similarity scores. Then, we consider the top-L links (in this paper, L is 100 [19, 25]). If among these top-L links, there are l links which are correctly predicted, that is, these l links are in E^P, then Precision can be defined as:

$$\text{Precision} = \frac{l}{L} \tag{15}$$

4.2. Networks

In this paper, we perform experiments on 10 real networks. The brief information of each benchmark network is as follows:

(1) *C.elegans* (CE) [26] – the neural network of the nematode worm *Caenorhabditis elegans*;
(2) Football network (FB) [27] – the network of American football games between Division IA colleges during regular season Fall 2000;
(3) SmallWorld (SW) [28] – the network of papers that cited S. Milgram's Psychology Today paper [29] or using Small World in title;
(4) Political Book (Book) [30] – nodes are books about US politics published around the time of the 2004 US Election and sold by Amazon.com; edges exist in frequently co-purchased books by the same customers;
(5) Political blogs (Blog) [31] – a network of weblogs on US politics;
(6) Jazz [32] – the network of Jazz musicians;
(7) US Air97 (USAir) [28] – the network of the US air transportation system;
(8) Food Web of Florida ecosystem (Food) [33] – this network contains the carbon exchanges in the cypress wetlands of South Florida during the wet season;
(9) Slavko [34] – a Facebook friendships network of Slavko Zitnik;
(10) Infectious (IN) [35] – the network of people's face-to-face contact during the exhibition 'Infectious: Stay Away' in 2009 at the Science Gallery in Dublin.

In this paper, we treat all of the networks listed above as undirected and unweighted networks whether they are weighted and/or directed networks or not. Meanwhile, loops are removed and multilinks are not allowed. The basic statistics of these networks are listed in Table 1.

4.3. Results and analysis

In this section, first, we describe the performance of Scop with different path length L. Then, we show the experimental results of each similarity index on all benchmarks.

Generally, longer paths need more computation but contribute little to predicting missing links [19]. The 'small-world phenomenon' [29] also suggests selecting short paths. Therefore, we only try lengths of 2, 3 and 4 on each network for the AUC and Precision metrics. The parameter d introduced in Definition 1 has a similar effect to the damping factor in PageRank [36]. In our experiments, we set the value of d to be 0.6. The results are shown in Figures 2 and 3, respectively. From Figure 2, we can see that Scop achieves the best performance measured by AUC on seven out of 10 networks when $L = 3$ while obtaining the best performance on the three other networks when $L = 2$. For the Precision metric shown in Figure 3, Scop obtains the best performance on two networks when $L = 3$ and gets the best performance on seven networks when $L = 2$. On the network Book, Scop gets very close performance when L = 2 and 3. Only on the network

Table 1. Statistics of the 10 networks. $|V|$ and $|E|$ are the total numbers of nodes and edges, respectively. $\langle K \rangle$ represents the average degree and $\langle C \rangle$ indicates the clustering coefficient [28]

Networks	CE	FB	SW	Book	Blog	Jazz	USAir	Food	Slavko	IN
$\|V\|$	453	115	395	105	1224	198	332	128	334	410
$\|E\|$	2025	613	994	441	16715	2742	2126	2075	2218	2765
$\langle K \rangle$	8.94	10.66	5.03	8.40	27.31	27.70	12.81	32.42	13.28	13.49
$\langle C \rangle$	0.65	0.40	0.33	0.49	0.32	0.62	0.63	0.33	0.45	0.46

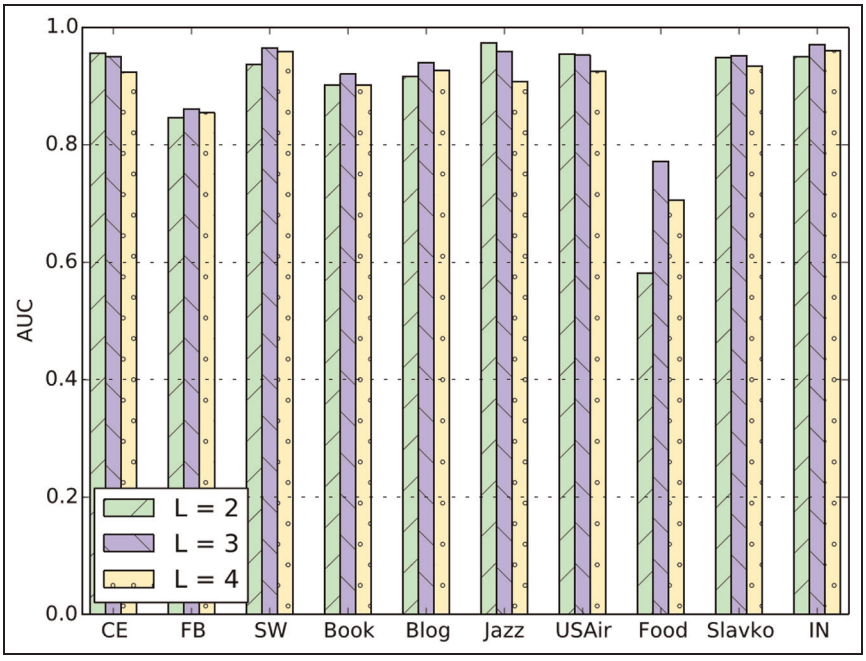

Figure 2. The AUC results of Scop index on 10 benchmark networks under different values of *L*.

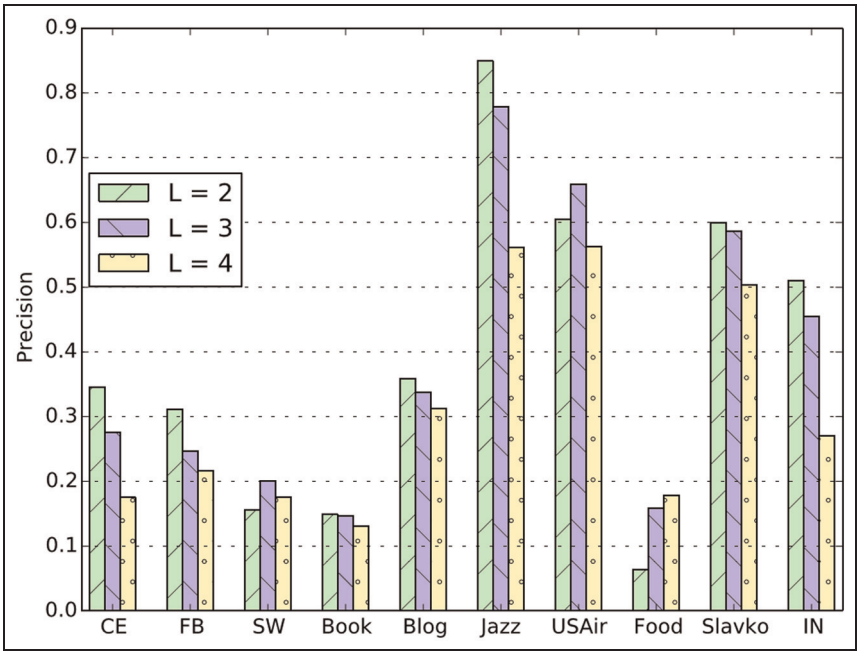

Figure 3. The Precision results of Scop index on 10 benchmark networks under different values of *L*.

Food, Scop achieves the best performance when $L = 4$. These results indicate that choosing the paths with length 2 and/or 3 is feasible and suffices for the Scop index.

In order to demonstrate the ability to predict missing links, we use the values of the best performance measure by both AUC and Precision metrics on 10 networks. For impartial comparison with the baselines, we hope to use their optimal parameter configurations. Lü et al. [1] suggested that, setting a small positive number to ϵ, the LP index could obtain a near-optimal accuracy. Thus in our experiments, we set $\epsilon = 0.001$ for LP index. In Papadimitriou et al. [20], the

Table 2. Performance of each similarity index measured by AUC metric on the 10 networks

AUC	CE	FB	SW	Book	Blog	Jazz	USAir	Food	Slavko	IN
AA	0.951	0.840	0.940	0.898	0.920	0.962	0.947	0.608	0.946	0.948
CN	0.917	0.838	0.931	0.888	0.917	0.955	0.935	0.606	0.942	0.945
LP	0.912	0.854	0.952	0.901	0.926	0.951	0.928	0.622	0.942	0.963
RA	0.955	0.840	0.941	0.900	0.921	0.970	0.953	0.612	0.947	0.949
FL	0.912	0.854	0.952	0.901	0.926	0.951	0.928	0.636	0.942	0.963
SP	0.953	0.854	0.957	0.915	0.932	0.971	0.953	**0.845**	**0.951**	0.970
Scop	**0.956**	**0.861**	**0.964**	**0.920**	**0.940**	**0.973**	**0.954**	0.771	**0.951**	**0.971**

Table 3. Performance of each similarity index measured by Precision metric on the 10 networks

Precision	CE	FB	SW	Book	Blog	Jazz	USAir	Food	Slavko	IN
AA	0.252	0.287	0.160	0.144	0.382	0.833	0.605	0.087	0.583	0.426
CN	0.198	0.295	0.140	0.115	0.420	0.804	0.583	0.088	0.552	0.386
LP	0.193	0.279	0.144	0.120	**0.422**	0.796	0.587	0.091	0.545	0.365
RA	0.308	0.286	0.173	0.143	0.247	0.826	0.624	0.084	0.612	0.493
FL	0.193	0.279	0.144	0.120	0.419	0.796	0.587	0.099	0.545	0.365
SP	0.313	0.286	0.176	0.140	0.241	0.719	0.554	**0.349**	**0.621**	0.440
Scop	**0.346**	**0.311**	**0.201**	**0.149**	0.359	**0.849**	**0.659**	0.178	0.600	**0.510**

authors of the FL index tested different lengths of paths and found that the best performance was attained with length 3. Therefore we also set length $L = 3$ for the FL index in our test. Concerning the SP index, it is difficult to look for the optimal values of α and β for different datasets. Among the 10 benchmarks, CE, Jazz, USAir, Food, Slavko and IN were also used in the work in which the SP index was proposed [22]. Therefore we use the optimal combinations of α and β reported in Zhu et al. [22] for SP, but for four other datasets, we set $\alpha = 0.001$ and $\beta = -1.1$. For the three other indices, CN, AA and RA, no parameters are needed.

The algorithmic AUC values of Scop and baselines on these 10 real networks are presented in Table 2. Clearly, Scop obtains the best performance on nine out of 10 datasets and achieves the second-best on the remaining dataset (Food). In addition, the SP index achieves the second-best performance. Although penalizing large degree nodes and considering long paths make SP better than other baseline indices, it is still worse than Scop, since it does not consider the contributions of end nodes. In Scop, the contributions of both paths and end nodes are taken into consideration, and hence it achieves the best performance.

In Table 3, the Precision values of each similarity index on the 10 networks are listed. Unsurprisingly, Scop obtains the best performance on seven out of 10 datasets. For the remaining three datasets (Blog, Food and Slavko), Scop is second on Food and third on Slavko. Only on Blog does Scop achieve the worse performance (fifth). From Table 3, we find that all of the methods which differentiate the contributions of paths or neighbours, such as RA and SP, give a poor performance on Blog. Conversely, CN and LP show their advantages on this network. This result shows that Blog is a particular network and has its special property. In Narang et al. [37], the link prediction task from different network flows has been studied. On the whole, these results indicate that Scop is a very accurate similarity index and can be applied to the link prediction problem.

From Table 2 and 3, we can see an interesting phenomenon, that the results of LP and FL indices are very similar. The reason for this phenomenon is that both of the two indices utilize the paths with lengths 2 and 3 connecting two nodes and do not distinguish the contributions of paths of the same length. Consequently, we can draw the conclusion that, when predicting missing links, incorporating the contributions of paths as well as the contributions of end nodes can further improve the performance of accuracy.

5. Conclusion and discussion

Link prediction, which is to find the missing links and predict future links in a network, is an important task in complex network analysis. To better solve this problem, a wide range of achievements, particularly similarity-based link prediction methods, have been proposed. We carefully investigate many similarity-based methods and find that even the most

powerful quasi-local similarity index at present, the Significant Path index, did not fully utilize the structure information of a network. One thing that can further improve its performance on accuracy is the contributions that end nodes themselves make to their similarity scores.

In this paper, we proposed a new quasi-local similarity index, Scop, to predict links in complex networks. In the viewpoint of Scop, not only do the paths connecting two end nodes make contributions to their similarity but also the status of participation of these two nodes in their common neighbourhood determines their similarity. By incorporating the contributions of both paths and end nodes, Scop achieves good accuracy.

To estimate the proposed index, we perform experiments on 10 real networks from disparate fields compared with CN, AA, RA, LP, FL and SP. To quantify the accuracy of prediction, two standard metrics, AUC and Precision, are employed. The experimental results show that a great improvement in accuracy is achieved by the proposed index. In summary, taking into account both path contribution and end node contribution in similarity index is a good and solid idea for investigating link prediction and will facilitate applications in link prediction.

Funding

This work was supported by the Fundamental Research Funds for the Central Universities (no. lzujbky-2014-47, lzujbky-2014-54).

References

[1] Lü L, Jin C H, and Zhou T. Similarity index based on local paths for link prediction of complex networks. *Physical Review E* 2009; 80(4): 046122.
[2] Von Mering C, Krause R, Snel B, et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 2002; 417(6887): 399–403.
[3] Butts CT. Network inference, error, and informant (in) accuracy: A Bayesian approach. *Social Networks* 2003; 25(2): 103–140.
[4] Guimerà R and Sales-Pardo M. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Science* 2009; 106(52): 22073–22078.
[5] Clauset A, Moore C and Newman MEJ. Hierarchical structure and the prediction of missing links in networks. *Nature* 2008; 453(7191): 98–101.
[6] Sarukkai RR. Link prediction and path analysis using markov chains. *Computer Networks* 2000; 33(1): 377–386.
[7] Zhu J, Hong J, and Hughes JG. Using Markov chains for link prediction in adaptive web sites. In: *Soft-Ware 2002: Computing in an imperfect world*. Heidelberg: Springer, 2002, pp.60–73.
[8] Bilgic M, Namata GM, and Getoor L. Combining collective classification and link prediction. In: *Proceedings of the 7th IEEE international conference on data mining workshops*. Washington, DC: IEEE Computer Society, 2007, pp. 381–386.
[9] Popescul A, and Ungar LH. Statistical relational learning for link prediction. In: *Proceedings of the workshop on learning statistical models from relational data at IJCAI 2003*. New York: ACM, 2003, pp. 81–90.
[10] Yu K, Chu W, Yu S, et al. Stochastic relational models for discriminative link prediction. In: *Proceedings of the 2006 neural information processing systems*. Vancouver: MIT Press, 2006, pp. 1553–1560.
[11] Lü L, and Zhou T. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications* 2011; 390(6): 1150–1170.
[12] Liben-Nowell D and Kleinberg J. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology* 2007; 58(7): 1019–1031.
[13] Lorrain F and White HC. Structural equivalence of individuals in social networks. *The Journal of Mathematical Sociology* 1971; 1(1): 49–80.
[14] Jaccard P. Etude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin del la Societe Vaudoise des Sciences Naturelles* 1901; 37: 547–579.
[15] Adamic LA and Adar E. Friends and neighbors on the web. *Social Networks* 2003; 25(3): 211–230.
[16] Zhou T, Lü L, and Zhang YC. Predicting missing links via local information. *The European Physical Journal B – Condensed Matter and Complex Systems* 2009; 71(4): 623–630.
[17] Katz L. A new status index derived from sociometric analysis. *Psychometrika* 1953; 18(1): 39–43.
[18] Jeh G and Widom J. SimRank: A measure of structural-context similarity. In: *Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining*. New York: ACM, 2002, pp. 538–543.
[19] Liu W and Lü L. Link prediction based on local random walk. *EPL (Europhysics Papers)* 2010; 89(5): 58007.
[20] Papadimitriou A, Symeonidis P, and Manolopoulos Y. Fast and accurate link prediction in social networking systems. *Journal of Systems and Software* 2012; 85(9): 2119–2132.
[21] Khosravi-Farsani H, Nematbakhsh M and Lausen G. SRank: Shortest paths as distance between nodes of a graph with application to RDF clustering. *Journal of Information Science* 2013; 39(2): 198–210.
[22] Zhu X, Tian H, Cai S, et al. Predicting missing links via significant paths. *EPL (Europhysics Papers)* 2014; 106(1): 18008.

[23] Hanley JA and McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983; 148(3): 839–843.
[24] Herlocker JL, Konstan JA, Terveen LG, et al. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems* 2004; 22(1): 5–53.
[25] Lü L and Zhou T. Link prediction in weighted networks: The role of weak ties. *EPL (Europhysics Papers)* 2010; 89(1): 18001.
[26] Watts DJ and Strogatz SH. Collective dynamics of 'small-world' networks. *Nature* 1998; 393(6684): 440–442.
[27] Girvan M and Newman MEJ. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 2002; 99(12): 7821–7826.
[28] Vladimir B and Andrej M. http://vlado.fmf.uni-lj.si/pub/networks/data/
[29] Milgram S. The small world problem. *Psychology Today* 1967; 2(1): 60–67.
[30] Valdis K. http://www.orgnet.com/.
[31] Adamic LA and Glance N. The political blogosphere and the 2004 US election: Divided they blog. In: *Proceedings of the 3rd international workshop on link discovery*. New York: ACM, 2005, pp. 36–43.
[32] Gleiser PM and Danon L. Community structure in jazz. *Advances in Complex Systems* 2003; 6(04): 565–573.
[33] Melián CJ and Bascompte J. Food web cohesion. *Ecology* 2004; 85(2): 352–358.
[34] Blagus N, Šubelj L, and Bajec M. Self-similar scaling of density in complex real-world networks. *Physica A: Statistical Mechanics and its Applications* 2012; 391(8): 2794–2802.
[35] Isella L, Stehlé J, Barrat A, et al. What's in a crowd? Analysis of face-to-face behavioral networks. *Journal of Theoretical Biology* 2011; 271(1): 166–180.
[36] Page L, Brin S, et al. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999.
[37] Narang K, Lerman K, and Kumaraguru P. Network flows and the link prediction problem. In: *Proceedings of the 7th workshop on social network mining and analysis*. New York: ACM, 2013, p. 3.