International Journal of Modern Physics C
Vol. 28, No. 4 (2017) 1750053 (23 pages)
© World Scientific Publishing Company
DOI: 10.1142/S012918311750053X

Link prediction based on local weighted paths for complex networks

Yabing Yao*, Ruisheng Zhang†, Fan Yang‡, Yongna Yuan§, Rongjing Hu¶
and Zhili Zhao‖

School of Information Science and Engineering
Lanzhou University, Lan zhou, Gansu 730000, P. R. China
**yaoyb14@lzu.edu.cn*
†*zhangrs@lzu.edu.cn*
‡*fanyang2014@lzu.edu.cn*
§*yuanyn@lzu.edu.cn*
¶*hurj@lzu.edu.cn*
‖*zhaozhl@lzu.edu.cn*

Revised 21 January 2017
Accepted 23 January 2017
Published 13 March 2017

As a significant problem in complex networks, link prediction aims to find the missing and future links between two unconnected nodes by estimating the existence likelihood of potential links. It plays an important role in understanding the evolution mechanism of networks and has broad applications in practice. In order to improve prediction performance, a variety of structural similarity-based methods that rely on different topological features have been put forward. As one topological feature, the path information between node pairs is utilized to calculate the node similarity. However, many path-dependent methods neglect the different contributions of paths for a pair of nodes. In this paper, a local weighted path (LWP) index is proposed to differentiate the contributions between paths. The LWP index considers the effect of the link degrees of intermediate links and the connectivity influence of intermediate nodes on paths to quantify the path weight in the prediction procedure. The experimental results on 12 real-world networks show that the LWP index outperforms other seven prediction baselines.

Keywords: Complex networks; link degree; link prediction; node similarity; path weight.

PACS Nos.: 89.20.Ff, 89.75.Fb.

1. Introduction

Various kinds of systems, such as the Internet, the World Wide Web, social networks, food webs and transportation systems, can be considered as complex networks in which nodes are regarded as individuals or units, links represent the interactions between them.[1] With the emergence of small-world[2] and scale-free[3] models, the

†Corresponding author.

complex network theory has become a powerful tool to understand and reveal real-world phenomena over the last decade.[4] Recently, the problem of link prediction in complex networks attracts increasing attention from different disciplinary fields due to its theoretical significance and applicable value. On the one hand, many evolving models have been proposed based on different topological features, which gives rise to the difficulty in judging which model is more advantageous than others to reflect the real evolving process of networks.[5] In principle, each evolving model can be converted to a link prediction algorithm, thus different evolving models can be assessed by the metrics of prediction performance.[6] On the other hand, link prediction can be applied to different domains. In biological networks, it serves to check the connections between proteins,[7] predict the interactions between diseases[8] and reveal the disease mechanisms.[9] It is also helpful to find potential friendships and enhance the loyalties of users in on-line social networks.[10,11] Moreover, in E-commerce networks, it can be used to recommend favorite products for customers and make profits. [12,13]

Link prediction aims to mine the missing and future links, with the aid of the observed links or nodes in networks.[14] In general, link prediction methods can be categorized into three types: (1) node attributes-based methods; (2) maximum-likelihood methods and (3) structural similarity-based methods. Depending on the technology of machine learning, the first methods always utilize the external information of nodes to improve prediction performance, such as sex, ages, incomes and locations of people in social networks. These methods reckon that individuals tend to form links if there are more common features among them.[15] However, the node attributes are unavailable or hidden in most cases, which results in the nonuniversal problem of this type of methods.[6] For example, owing to the privacy protection policy, the information of individuals generally keeps secret in on-line social networks. With the help of the organization principles during the evolution process of networks, some maximum-likelihood methods have been proposed, such as the hierarchical structural model[16] (HSM) and the stochastic block model[17,18] (SBM). Unfortunately, these methods are unable to cope with large scale networks in practice because of high computational cost.[6] By contrast, the structural similarity-based methods solely depend on the network structure and have low time complexity.[5] Therefore, this type of method has drawn a lot of attention. A number of structure features have been utilized by the structural similarity-based methods for link prediction, such as node degree, common neighbor, path, community, and so forth. However, most path-dependent methods simply sum over the total number of paths that have specific length as similarity, which neglect the different contributions of paths even with the same length. This problem is called as the heterogeneous contributions of paths in link prediction.[19]

In this paper, we focus on designing a novel structural similarity-based method, namely local weighted path (LWP) index. Our index concentrates on the characteristic of path heterogeneity in networks, i.e. different paths correspond to different contributions for a pair of nodes, even though they have same length. For a path between two nodes, we leverage the link degree information of its intermediate links and the connectivity influence of its intermediate nodes to measure the weight of this path, which is named as a weighted path. A high-weight path indicates that it has a more important contribution than low-weight paths to node similarity for a pair of nodes. To demonstrate the prediction performance of LWP, we compare it with seven prediction baselines in 12 networks originated from different fields. The experimental results show that the LWP index outperforms other prediction methods, especially in the sparse networks. Moreover, we demonstrate that the performance of the LWP index is inversely proportional to the link degrees of observed links on paths in networks.

This paper is organized as follows. We introduce the related work of link prediction that refers to the structural similarity-based methods in Sec. 2. Then, in Sec. 3, we propose our LWP index. The standard metrics for performance evaluation and the baselines for comparison are given in Sec. 4. The experimental results and analysis are shown in Sec. 5. This work is concluded in Sec. 6.

2. Related Work

The structural similarity-based methods attempt to mine the topological features of networks as many as possible to improve the prediction performance. An unconnected node pair is assigned to a similarity score as the likelihood of forming one link between them. The node pairs with high scores are more likely to connect each other than those with low scores in the future.

Generally, the structural similarity-based methods are divided into three categories: local indices, global indices and quasi-local indices.[6] As one of the simplest measures of local indices, common neighbors (CN) index motivated by a natural assumption that two nodes are more similar if they share more common neighbors.[20] The Adamic–Adar (AA)[21] and resource allocation (RA)[22] indices were developed to distinguish different contributions of common neighbors. Meanwhile, some variants of the CN index were proposed, such as Salton,[23] Sϕrensen,[24] HPI,[25] etc. Compared with local indices, global indices utilize the whole topological knowledges of networks. For instance, the Katz[26] and Leicht–Holme–Newman (LHN2)[27] indices exploit all paths through two nodes. Quasi-local indices are a trade-off between local indices and global indices, i.e. the quasi-local indices exploit more topological information than the former and less than the latter, such as the local path (LP) index,[22,28] local random walk (LRW) index,[29] etc. Overall, among these three types of categories, local indices have worst prediction accuracy with lowest time complexity. The accuracies of quasi-local indices are better than those of local indices, while these methods associate with the high complexity. Global indices own the best accuracies, they certainly have the highest time complexity.

In order to improve the performance for link prediction, different prediction methods concentrate on different structural features in networks, e.g. neighbor,[20–23] clustering coefficient,[30,31] path,[26,22,28] community[7,18,32–34] and so on. However, most path-dependent methods neglect the path heterogeneity of networks that mentioned in Ref. 19, i.e. different paths have different contributions to the prediction accuracy even though they have the same length. In Ref. 19, Zhu *et al.* proposed the SP index that considered the degrees of intermediate nodes on paths to distinguish the different contributions of paths. Although this index has a good prediction accuracy, it only depends on the degrees of intermediate nodes and does not fully exploit the structure information of paths in networks. In addition, the main disadvantage of the SP index is that it is very difficult to find its optimal parameters in different networks because of depending on two free parameters. Tian *et al.* [35] proposed the effective path (EP) index that leveraged the influence of endpoints and the strong connectivity of paths. Li *et al.*[36] proposed the Scop index that considered the different contributions of paths connecting two endpoints and the contributions of endpoints themselves. Zhu *et al.*[5] developed the Neighbor Set Information (NSI) index, which integrates two parts of structural features for a node pair, i.e. the common neighbors of two nodes and the links across two neighbor sets. All these four indices[19,35,36,5] measure the node similarity from the point of view of the intermediate nodes of paths, while the influence of intermediate links on paths is not concerned. By contrast, Xu *et al.*[37] proposed the path entropy (PE) index with penalization to long paths, which measures the entropy of a path as the sum of its intermediate links' entropies.

In this paper, we introduce the topological feature of the link degree of intermediate links to the problem of link prediction and propose a novel LWP index to discriminate the different contributions of paths. For the LWP index, to measure the weight of one path, we simultaneously take into account the link degrees of its intermediate links and the connectivity influence of its intermediate nodes. To evaluate the performance of LWP, we perform experiments on a series of real networks. The results show that the LWP index has a better performance than five mainstream indices CN, RA, AA, LP and Katz. Compared with the path heterogeneity methods, our index also outperforms the PE and NSI indices.

3. Methods

An undirected unweighted simple network is denoted by $G = (V, E)$, V and E are the sets of nodes and links, respectively. The multiple links and self-connections in the network G are not allowed. In order to predict the potential links, all unconnected node pairs are assigned similarity scores based on the structural similarity methods. For the sake of clarity, we denote the similarity score of an unconnected node pair (x, y) by S_{xy}.

Definition 1. For an observed link e_{ab} of which endpoints are nodes a and b in a network G, its link degree is defined as[38,39]

$$k_{e_{ab}} = k_a k_b, \tag{1}$$

where k_a is the degree of node a. In most cases, the degree is inversely proportional to the prediction accuracy in link prediction. Therefore, for the observed link e_{ab}, we define its contribution to the prediction performance as

$$\omega(e_{ab}) = (k_{e_{ab}})^{-1} = (k_a k_b)^{-1}. \tag{2}$$

For Eq. (2), its general expression form can be written as $(k_a k_b)^{\theta}$, where θ is a free parameter. The influence of θ on the prediction performance will be analyzed in Sec. 5.4.

Definition 2. Given a pair of nodes (x, y) in a network G, let $p^l_{(x,y)} = \{v_0 = x, v_1, v_2, \dots, v_{l-1}, v_l = y\}$ denote a path with length l $(l \geq 2)$ from x to y. The contributions of intermediate links of the path p to the similarity between nodes x and y is defined as

$$\begin{aligned} \xi(p^l_{(x,y)}) &= \sum_{i=0}^{l-1} \omega(e_{v_i v_{i+1}}) \\ &= \sum_{i=0}^{l-1} (k_{v_i} k_{v_{i+1}})^{-1}, \end{aligned} \tag{3}$$

where $\omega(e_{v_i v_{i+1}})$ stands for the contribution of the intermediate link $e_{v_i v_{i+1}}$ on the path p, which is calculated by Eq. (2). $\xi(p)$ is the weight of path p summing over the contributions of its intermediate links. For a node pair (x, y), a high weight value ξ of one path corresponds to an important contribution to node similarity for this path.

Definition 3. Given a pair of nodes (x, y) and a path $p^l_{(x,y)} = \{v_0 = x, v_1, v_2, \dots, v_{l-1}, v_l = y\}$ in a network G, let $I(p) = \{v_1, v_2, \dots, v_{l-1}\}$ represent the sequence set of intermediate nodes on the path p. For the node pair (x, y), the connectivity influence of intermediate nodes $I(p)$ to the similarity score is denoted by $\sigma(p^l_{(x,y)})$, which is defined as the fraction of the cycles of length $l+1$ passing only the intermediate nodes $I(p)$ divided by the maximum possible number of those cycles sharing $I(p)$

$$\sigma(p^l_{(x,y)}) = \begin{cases} \dfrac{N_{\triangle z}}{N_{\wedge z}} & \text{for} \quad l = 2, \\ \dfrac{N_{\square mn}}{N_{\sqcap mn}} & \text{for} \quad l = 3, \\ \dfrac{N_{\text{⌓} v_1 v_2 \dots v_{l-1}}}{N_{\cap v_1 v_2 \dots v_{l-1}}} & \text{for} \quad l \geqslant 4. \end{cases} \tag{4}$$

When $l = 2$, z is the intermediate node on a path $p^2_{(x,y)}$. $N_{\triangle z}$ and $N_{\wedge z}$ are the numbers of triangles and of maximum possible triangles including the node z, respectively. Essentially, $\sigma(p^2_{(x,y)})$ is the clustering coefficient[2] of node z.

When $l = 3$, m and n represent two intermediate nodes on a path $p^3_{(x,y)}$. $N_{\square mn}$ and $N_{\sqcap mn}$ are the numbers of quadrilaterals and of maximum possible quadrilaterals passing nodes m and n, respectively.

When $l \geq 4$, $N_{\text{⌓} v_1 v_2 \dots v_{l-1}}$ and $N_{\cap v_1 v_2 \dots v_{l-1}}$ are the numbers of circles and of potential circles that pass the intermediate nodes $I(p)$, respectively.

According to Definition 3, $\sigma(p)$ represents the connectivity influence of intermediate nodes $I(p)$ of path p and $\sigma(p) \in [0, 1]$.

For one path p, its contribution to node similarity is mainly determined by the weight of path p, i.e. $\xi(p)$. However, it is still difficult to distinguish the contributions of paths that have the same weight values $\xi(p)$. With the aid of Definition 3, we consider that the structure feature of forming cycles that relies on the intermediate nodes of path p can draw the distinction among those paths, i.e. the connectivity influence $\sigma(p)$. For a weighted path having the weight $\xi(p)$, the higher the connectivity of its intermediate nodes is, the greater the contribution of this path to node similarity is. Here, we regard $\sigma(p)$ as the enhancement factor of the weight of path p, i.e. $\xi(p)$. The weight of path $\xi(p)$ is enhanced significantly if the path p has a high value $\sigma(p)$. Therefore, the paths that have great contributions to node similarity can be further distinguished. With respect to paths that have same weight values or paths that have low weight values, one of them can have a highly ranked order according to the connectivity influence.

Definition 4. Given a pair of nodes (x, y) in a network G and a path $p^l_{(x,y)}$ between nodes x and y, a weighted path $M(p^l_{(x,y)})$ simultaneously considers the contributions of its intermediate links and the connectivity influence of its intermediate nodes. It is defined as

$$\begin{aligned} M(p^l_{(x,y)}) &= \xi(p^l_{(x,y)}) + \sigma(p^l_{(x,y)})\xi(p^l_{(x,y)}) \\ &= (1+\sigma(p^l_{(x,y)}))\xi(p^l_{(x,y)}). \end{aligned} \tag{5}$$

For Eq. (5), when $\sigma(p) = 0$, $M(p)$ is equal to the path wight $\xi(p)$. The more the value of $\sigma(p)$ is close to 1, the higher the weight of path p is. Generally, for a specified path length l, a path with higher value M has a more important contribution to the node similarity for node pair (x, y).

Definition 5. Given a pair of nodes (x, y), a weighted path index can be defined as

$$\begin{aligned} S_{xy} = &\sum_{p^2_{(x,y)} \in P^2_{(x,y)}} M(p^2_{(x,y)}) + \lambda \sum_{p^3_{(x,y)} \in P^3_{(x,y)}} M(p^3_{(x,y)}) + \cdots \\ &+ \lambda^{l-1} \sum_{p^l_{(x,y)} \in P^l_{(x,y)}} M(p^l_{(x,y)}), \end{aligned} \tag{6}$$

where $\lambda \in [0,+\infty)$ is a tunable parameter and $P^l_{(x,y)}$ is the set of paths with length l that have the endpoints of nodes x and y. This index sums over the contributions of all paths of which length is not greater than l by weighted form.

Considering the high computation complexity and little contributions of long paths,[28] we only focus on the paths with length 2 and 3 for a pair of nodes (x, y), named LWP index. It is defined as

$$S^{\mathrm{LWP}}_{xy} = \sum_{p \in P^2_{(x,y)}} M(p) + \lambda \sum_{q \in P^3_{(x,y)}} M(q), \tag{7}$$

where $M(p)$ is defined in Eq. (5).

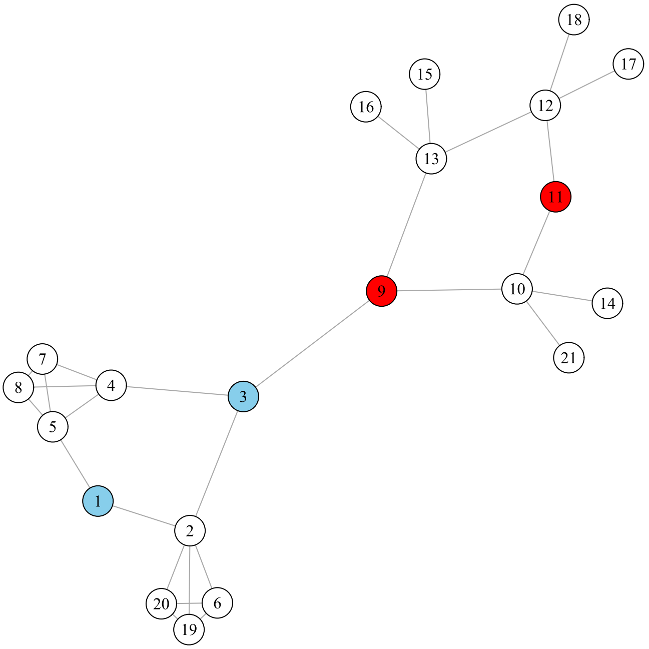

Fig. 1. (Color online) An illustration about the computational process of the LWP index.

We take Fig. 1 as an example to gain a deeper understanding of the LWP index. For node pairs (1,3) and (9,11), they have same similarity scores if we only count the number of paths between them. There are one path with length 2 (i.e. $e_{1,2} \leftrightarrow e_{2,3}$) and one path with length 3 (i.e. $e_{1,5} \leftrightarrow e_{5,4} \leftrightarrow e_{4,3}$) in terms of the node pair (1,3). Similarly, for the node pair (9,10), there are one path with length 2 (i.e. $e_{9,10} \leftrightarrow e_{10,11}$) and one path with length 3 (i.e. $e_{9,13} \leftrightarrow e_{13,12} \leftrightarrow e_{12,11}$). For the CN and LP indices, these two node pairs have same similarity scores, i.e. $S^{\rm CN}_{(1,3)} = S^{\rm CN}_{(9,11)} = 1$; $S^{\rm LP}_{(1,3)} = S^{\rm LP}_{(9,11)} = 1 + \lambda * 1 = 1 + \lambda$. For the RA and AA indices, the similarity of node pair $(1,3)$ is less than that of node pair $(9,11)$, i.e. $S^{RA}_{(1,3)} = \frac{1}{5} < S^{\rm RA}_{(9,11)} = \frac{1}{4}$; $S^{\rm AA}_{(1,3)} = \frac{1}{\log(5)} < S^{\rm AA}_{(9,11)} = \frac{1}{\log(4)}$. However, intuitively, the node pair $(1,3)$ is more likely to be linked than node pair $(9,11)$. Then, we calculate their similarity scores with the LWP index as follows (λ is fixed at 0.2 based on the experiment results analysis in Sec. 5): $S^{\rm LWP}_{(1,3)} = M(p^2_{(1,3)}) + 0.2 * M(p^3_{(1,3)}) = (1 + \frac{3}{10}) * (\frac{1}{5*2} + \frac{1}{5*3}) + 0.2 * (1 + \frac{1}{6}) * (\frac{1}{2*4} + \frac{1}{4*4} + \frac{1}{4*3}) = 0.2797$; $S^{\rm LWP}_{(9,11)} = M(p^2_{(9,11)}) + 0.2 * M(p^3_{(9,11)}) = (1 + \frac{0}{6}) * (\frac{1}{4*3} + \frac{1}{4*2}) + 0.2 * (1 + \frac{0}{15}) * (\frac{1}{3*4} + \frac{1}{4*4} + \frac{1}{4*2}) = 0.2625$. Obviously, the similarity score of node pair $(1,3)$ is greater than that of node pair $(9,11)$. It means that the connection probability of node pair $(1,3)$ is higher than that of node pair $(9,11)$. This result confirms the effect of heterogeneous paths, i.e. the LWP index distinguishes the different contributions of paths by the information of intermediate links and nodes on paths.

4. Metrics and Baselines

Given a simple network $G(V, E)$, the total number of possible node pairs in this network is denoted by $U = |V|(|V| - 1)/2$, where $|V|$ is the node size of the network G. In order to evaluate the accuracy of one prediction index, the observed links E are divided into two parts: the training set E^T and the probe set E^P. E^T that represents the known information of network G is used for calculating the similarity scores, while E^P is used to test the performance of the prediction index. In our experiments, the 90% observed links E are randomly picked as the training set E^T and the remaining 10% links are regarded as the probe set E^P. Note that, we guarantee that the network of training set is also a connected graph. Obviously, $E^T \cup E^P = E$ and $E^T \cap E^P = \emptyset$. Herein, we call a link in E^P is the missing link and a link in $U - E$ is the nonexistent link.

4.1. *Evaluation metrics*

In this paper, two widely used metrics are applied to evaluate the accuracy of one prediction index, i.e. AUC and Precision.

(1) AUC, *the area under the receiver operating characteristic* (*ROC*) *curve*,[40] can be viewed as the probability that the similarity score of one randomly chosen missing link is higher than that of one randomly chosen nonexistent link.[41] In practice, when n times of interdependent comparisons are performed, if there are n' times that the missing links have higher scores and n'' times that the scores of missing links equal to those of nonexistent links, then AUC can be defined as:

$$\mathrm{AUC} = \frac{n' + 0.5 \times n''}{n}. \tag{8}$$

Clearly, AUC should be roughly equal to 0.5 if the scores are generated from independent and identical distribution. The extent to which the AUC value of one predictor is higher than 0.5 reflects its accuracy than pure chance. Therefore, a high AUC value always corresponds to a high performance for one prediction index.

(2) Precision only concentrates on the accuracy of top ranked links.[42] Given the descending order of similarity scores of all nonobserved links, if there are L_r links belong to the probe set E^P, when we consider *top-L* links, then

$$\text{Precision} = \frac{L_r}{L}. \tag{9}$$

Obviously, the higher precision always corresponds to the higher accuracy.

4.2. *Baselines*

We choose seven representative indices as baselines for performance comparison.

(1) Common Neighbor index (CN).[20] As a simplest prediction algorithm, this index directly counts the number of all common neighbors between a node pair as the

similarity score. It is defined as

$$S_{xy}^{\mathrm{CN}} = |\Gamma(x) \cap \Gamma(y)|, \tag{10}$$

where $\Gamma(x)$ is the set of neighbors of node x.

(2) Adamic–Adar index (AA).[21] As a variant of CN, AA index not only considers the common neighbors of a pair of nodes, but also concerns the different contributions of common neighbors via giving low degree neighbors more weights. It is defined as

$$S_{xy}^{\mathrm{AA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k(z)}, \tag{11}$$

where $k(z)$ is the node degree of one common neighbor z for a node pair (x, y).

(3) Resource Allocation index (RA).[22] Motivated by the resource allocation process taking place on networks, this index also considers the different contributions of common neighbors and punishes the large degree nodes more severely than AA index. It is defined as

$$S_{xy}^{\mathrm{RA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k(z)}. \tag{12}$$

(4) Local Path index (LP).[28,22] To have a trade-off between the accuracy and computational complexity, this index takes into account local paths, i.e. the paths with length 2 and 3 between a pair of nodes, and assigns the longer paths to lower weight by a control parameter. It is defined as

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

where A is the adjacency matrix of network and λ is a free parameter which is always set to 0.01 for obtaining the approximately optimal performance.

(5) Katz index (Katz).[26] This index sums over all length paths that connect two nodes and punishes the longer paths less weights. It is defined as

$$S_{xy}^{\mathrm{Katz}} = \sum_{l=1}^{\infty} \beta^l \cdot |\mathrm{paths}_{xy}^{\langle l \rangle}| = \beta A_{xy} + \beta^2 (A^2)_{xy} + \beta^3 (A^3)_{xy} + \cdots, \tag{14}$$

where $|\mathrm{paths}_{xy}^{\langle l \rangle}|$ is the number of paths with length l connecting nodes x and y, β is a free parameter that is set to 0.01 in our experiments. This index can be redefined as $S = (I - \beta A)^{-1} - I$ when β is lower than the reciprocal of the maximum of the eigenvalues of adjacent matrix A.

(6) Path Entropy index (PE).[37] This index considers the information entropies of the shortest paths between two nodes by penalizing long paths. It is defined as

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

where l is the max length of paths and L_{xy}^1 denotes the probability that node pair $(x$,$y)$ has a link, and D_{xy}^i is a simple path with length i between two nodes x and y. In our experiments, l is set to three for PE index.

(7) Neighbor Set Information index (NSI).[5] From the perspective of information theory, this index measures the contributions of different structural features as information entropies. It integrates two parts of structural features: common neighbors of two nodes and links across two neighbor sets for a node pair. It is defined as

$$S_{xy}^{\mathrm{NSI}} = -I(L_{xy}^1|O_{xy}) - \lambda I(L_{xy}^1|P_{xy}), \tag{16}$$

where λ is a free parameter that can be fixed at 0.1 to get a reasonable performance, L_{xy}^1 is connection probability between node pair (x, y), O_{xy} is common neighbors between nodes x and y, P_{xy} denotes the set of links that across neighbors set of nodes x and y. Essentially, this index can be treated as one local path-based index considering the paths with length 2 and 3, although the authors do not mention in Ref. 5.

5. Experiments and Analysis

In this section, we demonstrate the prediction performance of the LWP index from different aspects in our experiments.

5.1. *Datasets*

We take into account 12 real-world networks drawn from disparate fields. All these networks are treated as the undirected and unweighted networks. Besides, the giant component of these networks is only considered. The reason is that, for most structural similarity-based indices, the similarity scores of node pairs which locate in different components are always given to zero.[28] A brief description of these networks is given as follows: (1) Karate[43]: a friendship network of a karate club. (2) Lesmis[44]: a co-appearance network of characters in the novel "Les Miserables". (3) Physicians[45]: an innovation spread network among physicians. (4) FoodWeb[46]: a food ecosystem of Florida bay. (5) Celegans[2]: a neural network of the nematode worm Caenorhabditis elegans. (6) Metabolic[47]: a metabolic network of C.elegans. (7) WebGoogle:[48] an information network of web pages from Google. (8) Names[49–51]: a co-appearance network of nouns in the King James Version of the Bible. (9) Watt[48,52,53]: a miscellaneous network. (10) Hamster[54]: a friendship and familylink network of users on the website hamsterster.com. (11) GrQc[48,53,55]: a scientific collaboration network of individuals. (12) WebSpam[48,53,56]: an information network of web pages.

Table 1 lists the basic topological features of these experimental networks.

5.2. *The performance accuracy of the LWP index with changing of λ*

In this subsection, we investigate the prediction performance of the LWP index as a function of λ. According to Eq. (7), the value of the parameter λ ranges from 0 to $+\infty$. When $\lambda = 0$, the performance of the LWP index only depends on the paths with length 2. Conversely, its performance is completely driven by the paths with length 3

Table 1. The basic topological features of experimental networks. N and M are the total number of nodes and links, respectively. $\langle k \rangle$ is the average degree, $S = \frac{2M}{N(N-1)}$ denotes the network density, r stands for the assortative coefficient, d is the average shortest path distance, C represents the average clustering coefficient, $H = \frac{\langle k^2 \rangle}{\langle k \rangle^2}$ is the degree heterogeneity, e is the network efficiency.

Nets	N	M	$\langle k \rangle$	S	r	d	C	H	e
Karate	34	78	4.588	0.139	−0.476	2.408	0.588	1.693	0.492
Lesmis	77	254	6.597	0.087	−0.165	2.641	0.736	1.827	0.435
Physicians	117	465	7.949	0.068	−0.084	2.587	0.219	1.253	0.432
FoodWeb	128	2106	32.906	0.259	−0.104	1.772	0.335	1.231	0.624
Celegans	297	2148	14.465	0.049	−0.163	2.455	0.308	1.801	0.445
Metabolic	453	2025	8.94	0.02	−0.226	2.664	0.655	4.485	0.407
WebGoogle	1299	2773	4.269	0.003	−0.055	6.48	0.61	2.648	0.173
Names	1707	9059	10.614	0.006	−0.052	3.376	0.71	3.922	0.32
Watt	1856	4942	5.325	0.003	−0.033	14.231	0.044	1.332	0.1
Hamster	2000	16098	16.098	0.008	0.023	3.589	0.573	2.719	0.306
GrQc	4158	13422	6.456	0.002	0.639	6.049	0.665	2.785	0.179
WebSpam	4767	37375	15.681	0.003	0	3.793	0.359	4.612	0.291

*All these networks can be found at http://networkrepository.com/networks.php and http://konect.uni-koblenz.de/networks/.

when $\lambda \to +\infty$. We find that the optimal parameter of λ always falls from 0 to 1 for all networks (see the detailed results in Appendix A). Therefore, in order to observe the performance variation of the LWP index, we plot its AUC and Precision results in Figs. 2 and 3, respectively, when λ ranges from 0 to 1 in steps of size 0.01. Owning to the parameter dependency of the LP and NSI indices, we also show their performances in these figures.

As shown in the figures, the performance of the LWP index is always improved when we simultaneously consider the paths with length 2 and 3. It indicates that, compared with the performance that only based on the paths with length 2, it is effective to improve the prediction performance with the information of paths with length 2 and 3 for the LWP index. In some cases, the performance that only depends on the paths with length 3 is better than that solely relies on the paths with length 2 no matter for LWP, NSI or LP, such as in the networks of Karate, FoodWeb and Celegans. The performance of the LWP index with paths of length 3 is even close to the optimal prediction performance, such as in the FoodWeb network. This may be due to the special topological features of these networks. Additionally, when λ changes, we find that the LWP index performs remarkably better than the LP and NSI indices in most cases, and the LP index always has the worst performance. With the increment of λ, the performances of the LP and NSI indices are hardly improved and even decrease from the beginning, especially in the networks of Lesmis, Metabolic and Names. This indicates that the information of paths with length 3 becomes noise for these two predictors. Conversely, for all networks, the performance of the LWP index can be further improved. This result further exhibits that the LWP index has an advantage in making use of the information of long paths.

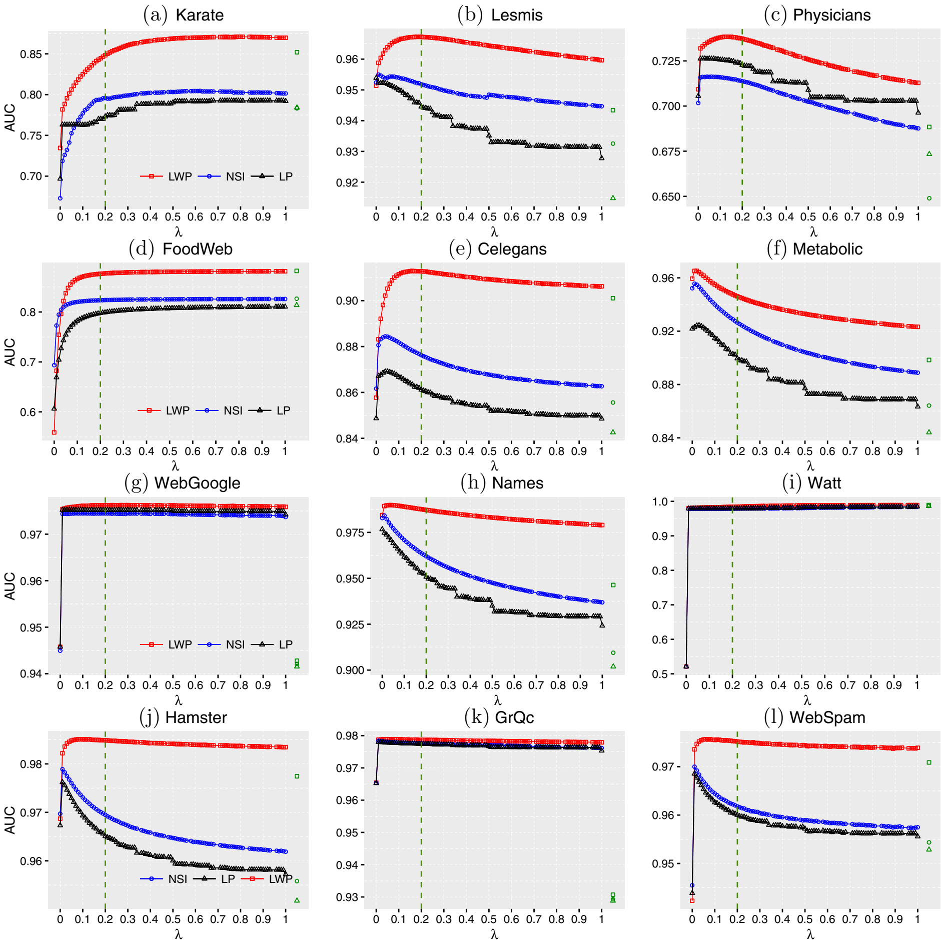

Fig. 2. (Color online) The accuracies of the LWP, NSI and LP indices measured by the metric of AUC in different networks. Each value is averaged over 50 independent implementations. The green points of square, circle and triangle on the far right in figures represent the accuracies of the LWP, NSI and LP indices that only exploit paths with length 3, respectively. Each green vertical line presents the AUC performance of LWP when $\lambda = 0.2$.

Although there are different trends in the performance of the LWP index with changing of λ, we find that $\lambda = 0.2$ always corresponds to a reasonable performance in terms of different networks for the LWP index (see Figs. 2 and 3). Since the metric of Precision depends on the number of L, we present the performance of the LWP index with $\lambda = 0.2$ when L changes from 10 to 100 in steps of size 10 in Fig. 4 as well. As shown in the figure, compared with other baselines, the Precision performance of the LWP index has a great advantage especially for the sparse networks (i.e. the network density is less than 0.05 shown in Table 1), such as Celegans, Metabolic, Watt, Hamster and GrQc. Tables 2 and 3 show the performance of the LWP index

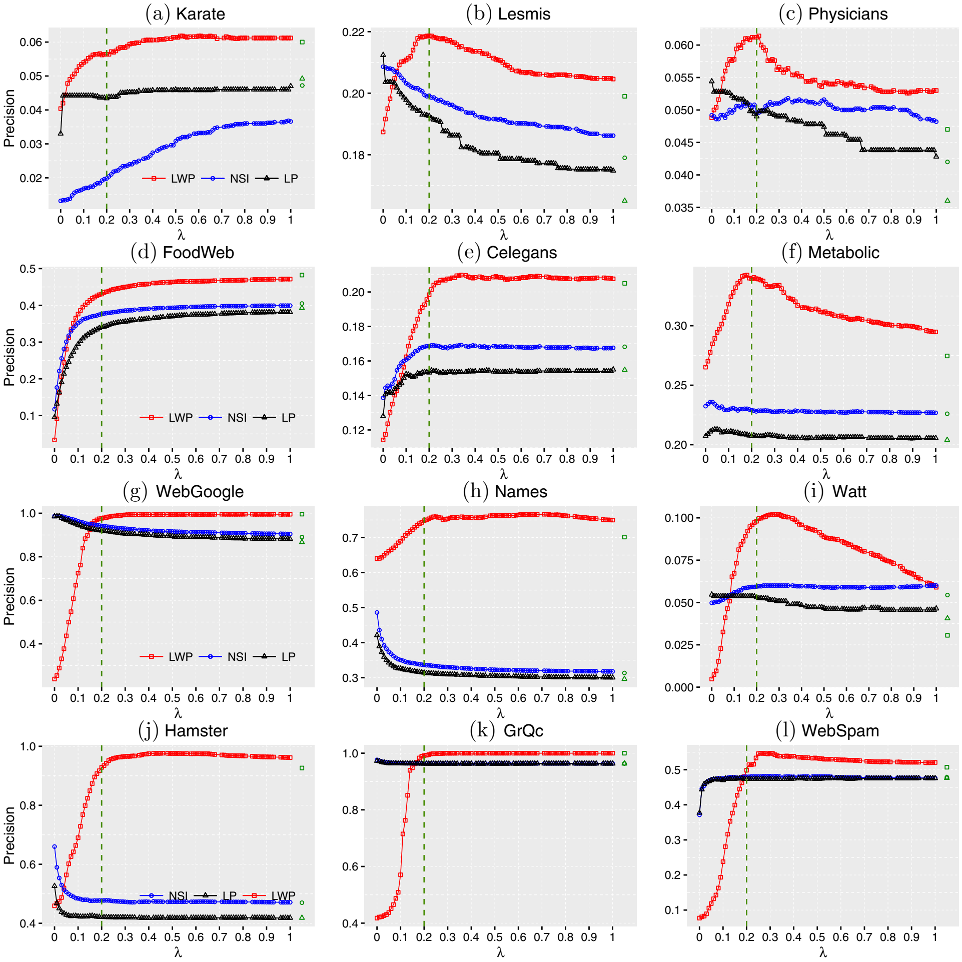

Fig. 3. (Color online) The accuracies of the LWP, NSI and LP indices measured by the metric of Precision with top-100 in different networks. Each value is averaged over 50 independent implementations. The green points of square, circle and triangle on the far right in figures represent the accuracies of the LWP, NSI and LP indices that only exploit paths with length 3, respectively. Each green vertical line presents the Precision performance of LWP when $\lambda = 0.2$.

with a fixed value 0.2 of λ (i.e. the column of $\mathrm{LWP}_{0.2}$). As shown in the tables, the LWP index with $\lambda = 0.2$ provides a competitively accurate performance compared with its optimal performance (i.e. the column of $\mathrm{LWP}_{\mathrm{opt}}$). In most cases, it has a better performance than other baselines. This result is highly significant for enhancing the application value of the LWP index. One can keep λ as a fixed value 0.2 and save the searching time of the optimal parameter λ in the prediction procedure for the LWP index.

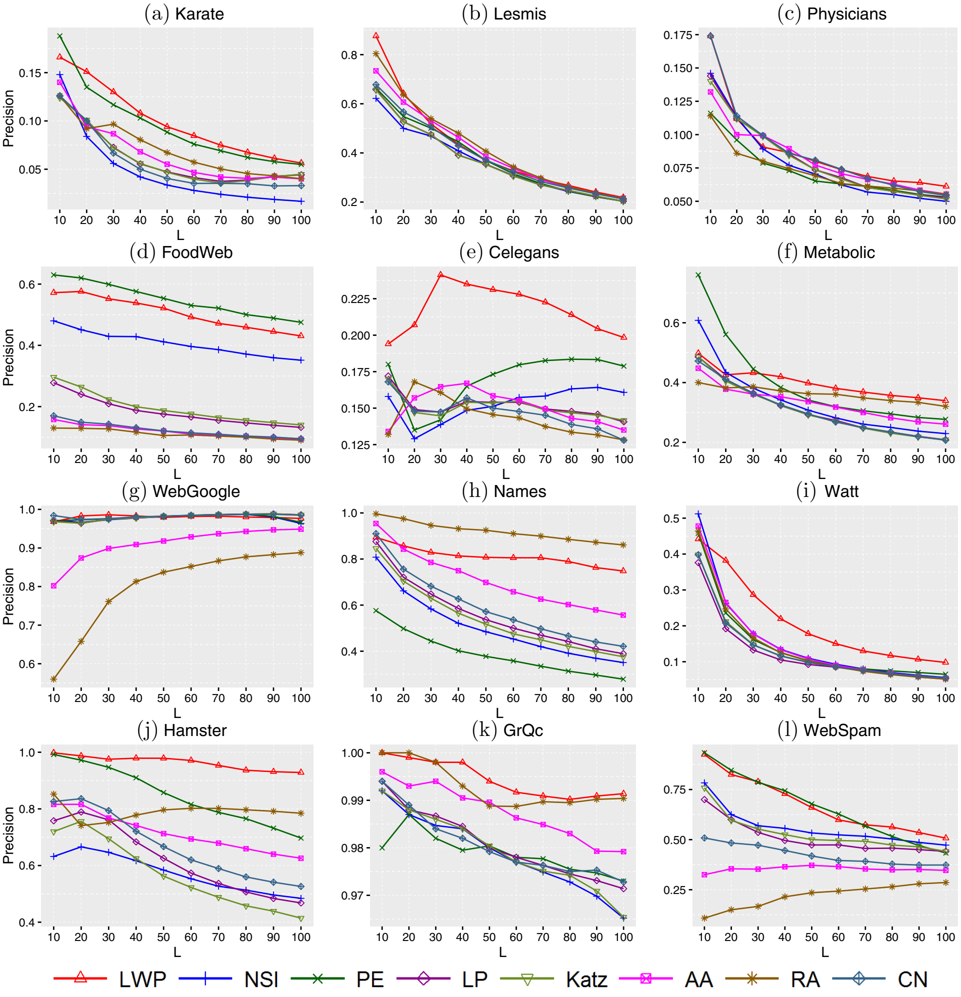

Fig. 4. (Color online) Comparison of the accuracies of Precision with the variation of L in different networks. Each value is averaged over 50 independent implementations. The free parameters of the LWP, NSI, LP and Katz indices are typically fixed at 0.2, 0.1, 0.01 and 0.01.

5.3. *The prediction accuracy of the LWP index compared with baselines*

Table 2 presents the prediction accuracy of eight indices measured by the metric of AUC. According to the AUC results, our LWP index has the best performance on ten out of 12 networks, except for the Katz index on the networks of WebGoogle and GrQc. The reason is that, the Katz index exploits the information of all path length and these two networks have the higher average shortest path length than other networks, as shown in Table 1. Furthermore, the LWP index still has the second-best performance on these two networks. Meanwhile, the AUC values are significantly

Table 2. Comparison of the accuracies measured by AUC in different networks. Each value is obtained by averaging over 50 implementations with independent random divisions of training set (90%) and probe set (10%).

Nets	CN	RA	AA	LP_{opt}	Katz	PE	NSI_{opt}	$\text{LWP}_{0.2}$	LWP_{opt}
Karate	0.6968	0.7355	0.7273	0.7926	0.7596	0.8414	0.8045	**0.8483**	**0.8709***
Lesmis	0.9539	0.9628	0.9624	0.9539	0.952	0.9484	0.9555	**0.9672**	**0.9673***
Physicians	0.7056	0.7108	0.7111	0.7263	0.7244	0.7142	0.7163	**0.7373**	**0.7384***
FoodWeb	0.6063	0.6115	0.6078	0.8135	0.6752	0.8595	0.8266	**0.8764**	**0.8824***
Celegans	0.8487	0.8704	0.8659	0.8691	0.8646	0.8848	0.8844	**0.9128**	**0.913***
Metabolic	0.9218	**0.9596**	0.9548	0.9244	0.9216	0.9107	0.9556	0.9462	**0.9652***
WebGoogle	0.9458	0.9464	0.9464	0.9752	**0.9896***	0.9704	0.9745	0.9762	**0.9763**
Names	0.9767	0.9865	0.9851	0.9767	0.9735	0.6148	0.9859	**0.9872**	**0.9899***
Watt	0.521	0.521	0.5209	0.986	0.9821	0.9852	**0.9869**	0.9841	**0.9893***
Hamster	0.9673	0.973	0.9714	0.9764	0.9746	0.5928	0.9791	**0.9849**	**0.9851***
GrQc	0.9653	0.9658	0.9658	0.978	**0.9864***	0.9764	0.9785	0.9787	**0.9789**
WebSpam	0.944	0.9474	0.9466	0.9687	0.9616	0.5279	0.9705	**0.9752**	**0.9757***

Note: The top-2 best accuracies are emphasized by bold and * denotes the highest accuracies in each network. LP_{opt}, NSI_{opt} and LWP_{opt} are the best performance of the LP, NSI and LWP indices in each network in Fig. 2. $\text{LWP}_{0.2}$ represents the performance of the LWP index when $\lambda = 0.2$.

improved by the LWP index in the networks of Karate, FoodWeb and Celegans compared with other baseline indices.

Table 3 presents the performance accuracy measured by the metric of Precision. According to the Precision results, our LWP index obtains the best performance on all networks, except for the RA index on the network of Names. It also has the second-best performance on the Names network for the LWP index. Furthermore, in contrast to the relatively dense networks, the Precision accuracy of the LWP index is

Table 3. Comparison of the accuracies measured by the metric of Precision with top-100 in different networks. Each value is obtained by averaging over 50 implementations with independent random divisions of training set (90%) and probe set (10%).

Nets	CN	RA	AA	LP_{opt}	Katz	PE	NSI_{opt}	$\text{LWP}_{0.2}$	LWP_{opt}
Karate	0.033	0.0402	0.04	0.0496	0.0448	0.055	0.0476	**0.0564**	**0.0618***
Lesmis	0.2124	0.2148	**0.2154**	0.2124	0.2032	0.2038	0.2086	**0.2186***	**0.2186***
Physicians	0.0544	0.0552	0.0554	0.0544	0.0518	0.0532	0.0518	**0.0612**	**0.0614***
FoodWeb	0.0952	0.0902	0.0936	0.3922	0.1402	**0.4752**	0.4048	0.4308	**0.4824***
Celegans	0.128	0.1282	0.135	0.155	0.1416	0.1788	0.1692	**0.1984**	**0.2098***
Metabolic	0.2072	0.3204	0.261	0.2128	0.2088	0.277	0.2358	**0.3394**	**0.3426***
WebGoogle	0.9848	0.888	0.9488	0.986	0.9858	0.9664	**0.9872**	0.9766	**0.9968***
Names	0.421	**0.8612***	0.5568	0.421	0.3758	0.2786	0.486	0.748	**0.7668**
Watt	0.0546	0.0514	0.0542	0.0546	0.054	0.065	0.06	**0.098**	**0.1022***
Hamster	0.5262	0.7846	0.6258	0.5262	0.4144	0.697	0.66	**0.9284**	**0.9766***
GrQc	0.9728	0.9904	0.9792	0.9734	0.9654	0.973	0.9762	**0.9914**	**1***
WebSpam	0.3771	0.2864	0.3507	0.4764	0.4479	0.4321	0.4814	**0.4986**	**0.5471***

Note: The top-2 best accuracies are emphasized by bold and * denotes the highest accuracies in each network. LP_{opt}, NSI_{opt} and LWP_{opt} are the best performance of the LP, NSI and LWP indices in each network in Fig. 3. $\text{LWP}_{0.2}$ represents the performance of the LWP index when $\lambda = 0.2$.

more advantageous than other baselines in the sparse networks (i.e. the network density is less than 0.05 shown in Table 1), including the networks of Celegans, Metabolic, Watt, Hamster and GrQc. Taking the Hamster network as an example, the Precision accuracy of LWP index achieves 0.9766, while the best performance of other baselines is only 0.7846 (i.e. RA index).

Overall, the performances of the path-based indices, i.e. LWP, NSI, PE, LP and Katz, are better than those of the local indices that are based on the information of nodes, i.e. CN, AA and RA, regardless of the AUC or Precision. This is because the path-based indices exploit more topological information than the node-based indices. As variants of CN, the AA and RA indices have better performance than CN, because they consider the different contributions of neighbors. In most cases, due to the negligence of heterogeneous contributions of paths, the performances of the LP and Katz indices are worse than those of the path heterogeneity indices, i.e. LWP, NSI and PE. Among the path heterogeneity indices, the LWP index always performs better than the NSI and PE indices. The reason is that the LWP index makes use of more topological information than NSI and PE. It not only considers the different contributions of intermediate links on paths, but also adopts the different contributions of intermediate nodes on the paths. Therefore, we can conclude that the LWP index has an excellent performance with respect to different metrics and disparate networks.

5.4. *The influence of θ on the prediction accuracy for the LWP index*

In the above experiments, we adopt $\omega(e_{ab}) = (k_a k_b)^{-1}$ as the contribution of an existent link e_{ab} for the LWP index in Eq. (2). Actually, the general form of the contribution of e_{ab} can be redefined as a function of θ for $\omega(e_{ab})$, i.e. $\omega(e_{ab}^{\theta}) = (k_a k_b)^{\theta}$. The definition of Eq. (2) can be considered as a special case of $\omega(e_{ab}^{\theta})$ when $\theta = -1$. In this subsection, we pay attention to observe the influence of θ on the performance of the LWP index. Here, we fix λ at the reasonable value 0.2 and change θ from -3 to 3 by increment of 0.1 each time. The performance of AUC and Precision (top-100) of the LWP index are presented in Figs. 5 and 6, respectively.

On the whole, the optimal value of θ is less than 0, regardless of the AUC or Precision. It suggests that the contribution of an intermediate link on paths is inversely proportional to the product of the degrees of its two endpoints. This result is consistent with the conclusions of AA and RA, i.e. we can obtain a good prediction performance by punishing the large degree nodes heavily.[22] Moreover, the optimal θ is always very close or equal to -1 in most networks, such as in the networks of Lesmis, Physicians, FoodWeb, Celegans, Metabolic, WebGoogle, Names, Hammster and GrQc. This experimental results verify the reasonableness of the definition of Eq. (2) to measure the contribution of an intermediate link on paths.

5.5. *Comparison of the computational complexity*

In this subsection, we discuss the computational complexities of the eight similarities indices, i.e. CN, AA, RA, LP, Katz, PE, NSI and LWP. Given a simple undirected

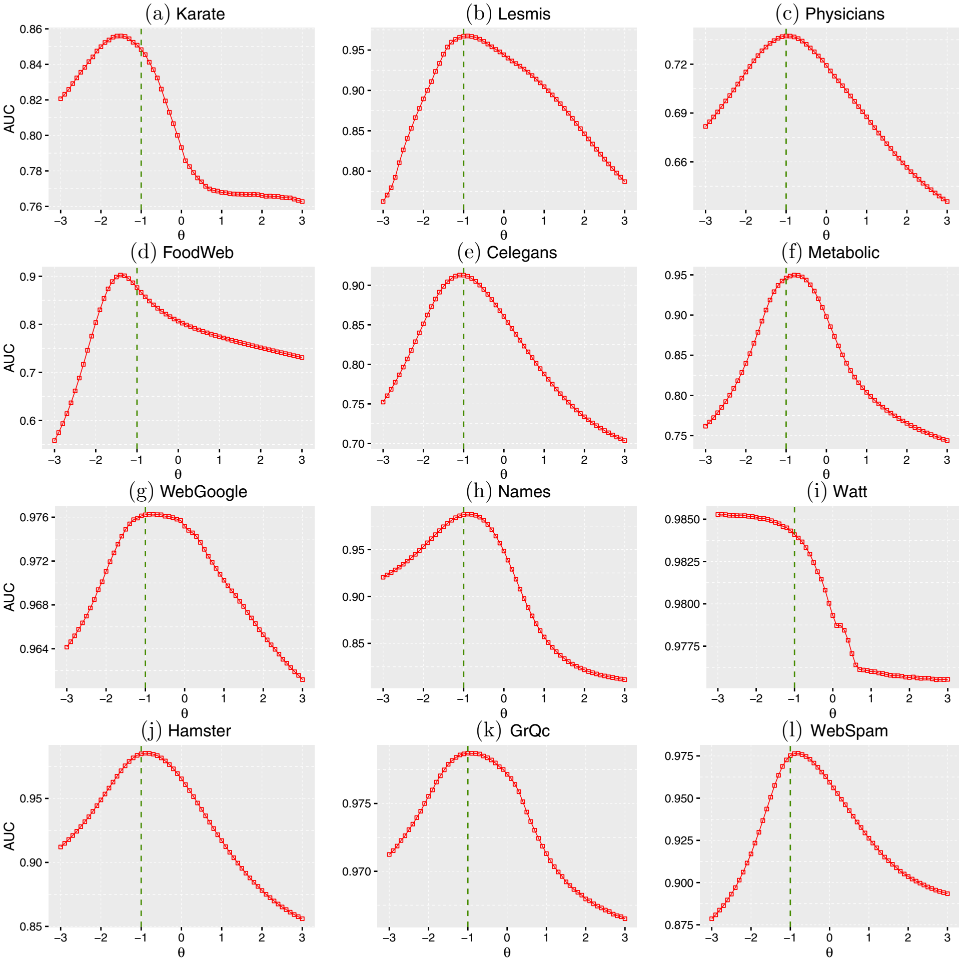

Fig. 5. (Color online) The prediction accuracy of the LWP index that measured by AUC varies according to the θ parameter in different networks. Each value is averaged over 50 independent implementations. Each green vertical line corresponds to the AUC performance of LWP when $\theta = -1$.

network $G(V, E)$, $N = |V|$ and $M = |E|$ are the node size and the link number of the network G, respectively. Let $\langle k \rangle$ denote the average degree of a node in the network G, obviously, $M = N\langle k \rangle/2$. For a node x, it takes time $O(\langle k \rangle^2)$ and $O(\langle k \rangle^3)$ to search the paths with length 2 (i.e. the common neighbors) and length 3 from x to another node y, respectively.[28] Since the CN, AA and RA indices can be considered as the path-based indices, which make use of paths with length 2, the time complexities in calculating these three indices are $O(N\langle k \rangle^2)$. The LP and PE indices employ the information of paths with length 2 and 3, thus the time complexities of these two predictors are $O(N\langle k \rangle^3)$. Due to the requirement of matrix inversion operator, the time complexity of Katz index is $O(N^3)$.[57] For the NSI and LWP indices, in addition to consider the paths with lengths 2 and 3, they also take time to

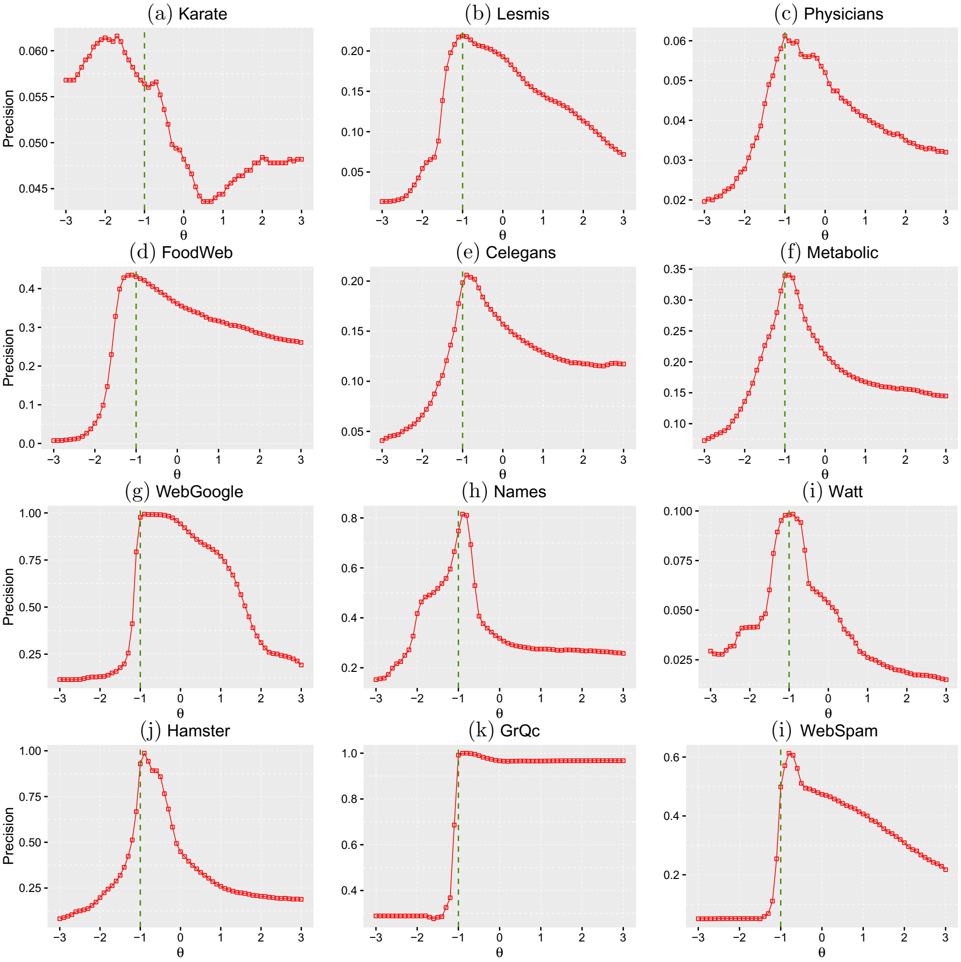

Fig. 6. (Color online) The prediction accuracy of the LWP index that measured by top-100 Precision varies according to the θ parameter in different networks. Each value is averaged over 50 independent implementations. Each green vertical line corresponds to the Precision performance of LWP when $\theta = -1$.

calculate the probabilities of forming triangles (i.e. the clustering coefficient) and quadrilaterals for the intermediate nodes on these paths. Generally, it takes time $O(\langle k\rangle^2)$ to calculate the clustering coefficient of an intermediate node for paths with length 2, while the same time is consumed to obtain the probability of forming quadrilaterals for two intermediate nodes of paths with length 3. In the worst case, for the LWP and NSI indices, the time complexity of calculating the clustering coefficient for all intermediate nodes is $O(N\langle k\rangle^2)$ and the time complexity of calculating the quadrilateral probability for all two intermediate nodes is $O(M\langle k\rangle^2) = O(N\langle k\rangle^3/2)$. Therefore, the time complexities in calculating the LWP and NSI indices are $O(N\langle k\rangle^3 + N\langle k\rangle^2 + N\langle k\rangle^3/2) \propto O(N\langle k\rangle^3)$.

From all experimental results mentioned above, we can observe that the LWP index outperforms five traditional methods (i.e. CN, RA, AA, LP, Katz) measured by the metrics of AUC and Precision. Compared with two representative path heterogeneity indices, i.e. PE and NSI, the LWP index also has a better performance. Especially, the LWP index has a great advantage in the sparse networks. This is because the LWP index exploits more information to discriminate the different contributions of paths, which is neglected by other traditional prediction methods. The free parameter of λ can be approximatively fixed at 0.2 to get a reasonable performance for the LWP index. This results further enhance the application value of our index. Meanwhile, we confirm that the contribution of a link is inversely proportional to its link degree. By the analysis of complexity, the LWP index has the same time complexity $O(N\langle k\rangle^3)$ as other local path-dependent indices. As a whole, the LWP index has a great performance, whether the prediction accuracy or efficiency.

6. Conclusion

Link prediction is a significant problem in complex network, which aims to estimate the existence likelihood of the missing links and future links. In this paper, we pay attention to the structural similarity-based method and address the problem of path heterogeneity in link prediction. First, we define the weight of a path based on the topological feature of the link degrees of all intermediate links on this path. Second, considering the influence of intermediate nodes on the path, we enhance its path weight with the aid of the connectivity influence of its intermediate nodes. Finally, we propose a novel LWP index that exploits the information of local paths with lengths 2 and 3. For the LWP index, it takes account of the different contributions of different paths between a pair of nodes. In order to validate the effectiveness of LWP index, we perform experiments on 12 networks from different application domains. The results demonstrate that our LWP index outperforms other seven similarity indices that measured by AUC and Precision, especially in the sparse networks. Although the LWP index depends on a free parameter to control the weight of long paths, yet this parameter can be fixed at a constant number, which enhances its application value in practice. Besides, we demonstrate that the performance of LWP index is inversely proportional to the link degrees of observed links. In the future, we will examine the prediction performance by other topological features in networks that measure the importance of observed links or paths. Another interesting attempt is how to extend our approach in the bipartite networks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 21503101), the Natural Science Foundation of Gansu Province, China (Grant No. 1506RJZA223) and the Project-sponsored by SRF for ROCS, SEM (Grant No. SEM[2015]311).

Appendix A. The Prediction Accuracy of LWP Index When λ Ranges from 10^{-5} to 10^5

According to Eq. (7), the parameter λ ranges from 0 to $+\infty$ to control the weight of paths with length 3 for the LWP index. We observe the performance of LWP index when its parameter λ ranges from 10^{-5} to 10^5 with a ratio of 10. The AUC and Precision results in different networks are shown in Figs. A.1 and A.2, respectively. As shown in the figures, the optimal parameter λ that corresponds to the best

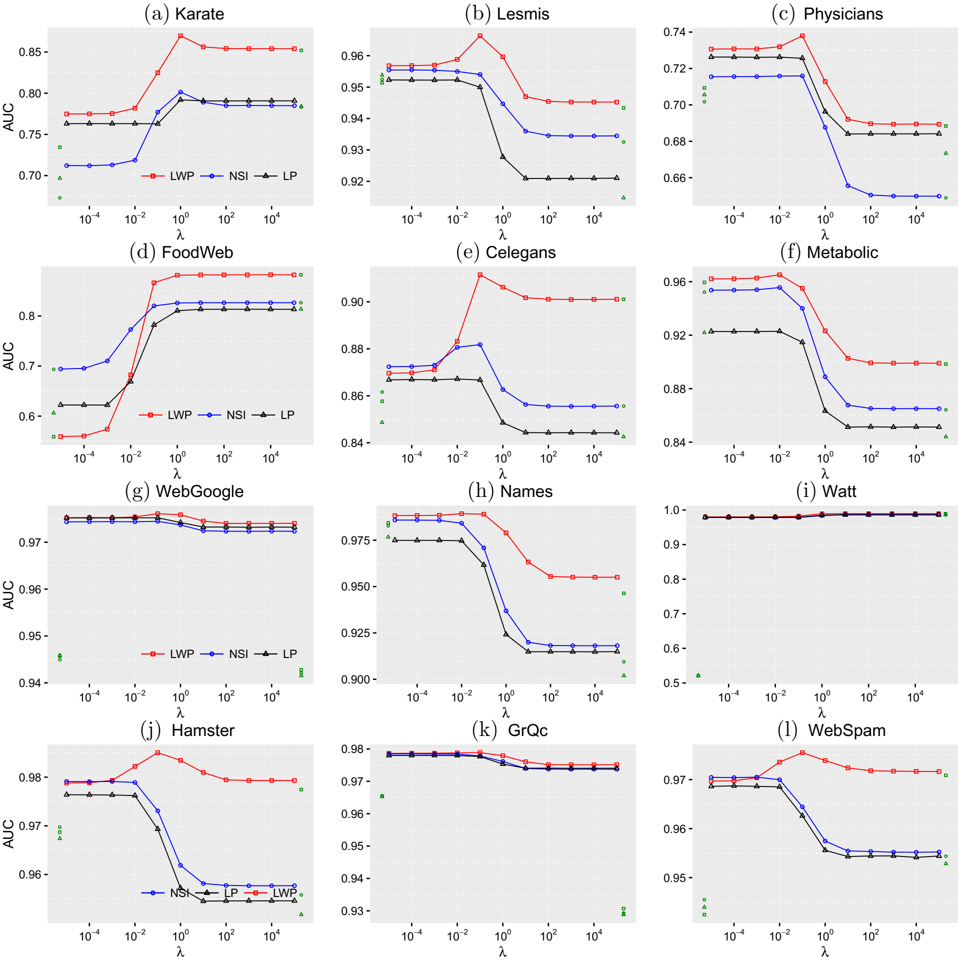

Fig. A.1. (Color online) The accuracies of LWP, NSI and LP indices measured by AUC in different networks. Each value is averaged over 50 independent implementations. The green points of square, circle and triangle on the far left in figures represent the accuracies of LWP, NSI and LP indices that only exploit paths with length 2, respectively. The green points of square, circle and triangle on the far right in figures represent the accuracies of LWP, NSI and LP indices that only exploit paths with length 3, respectively.

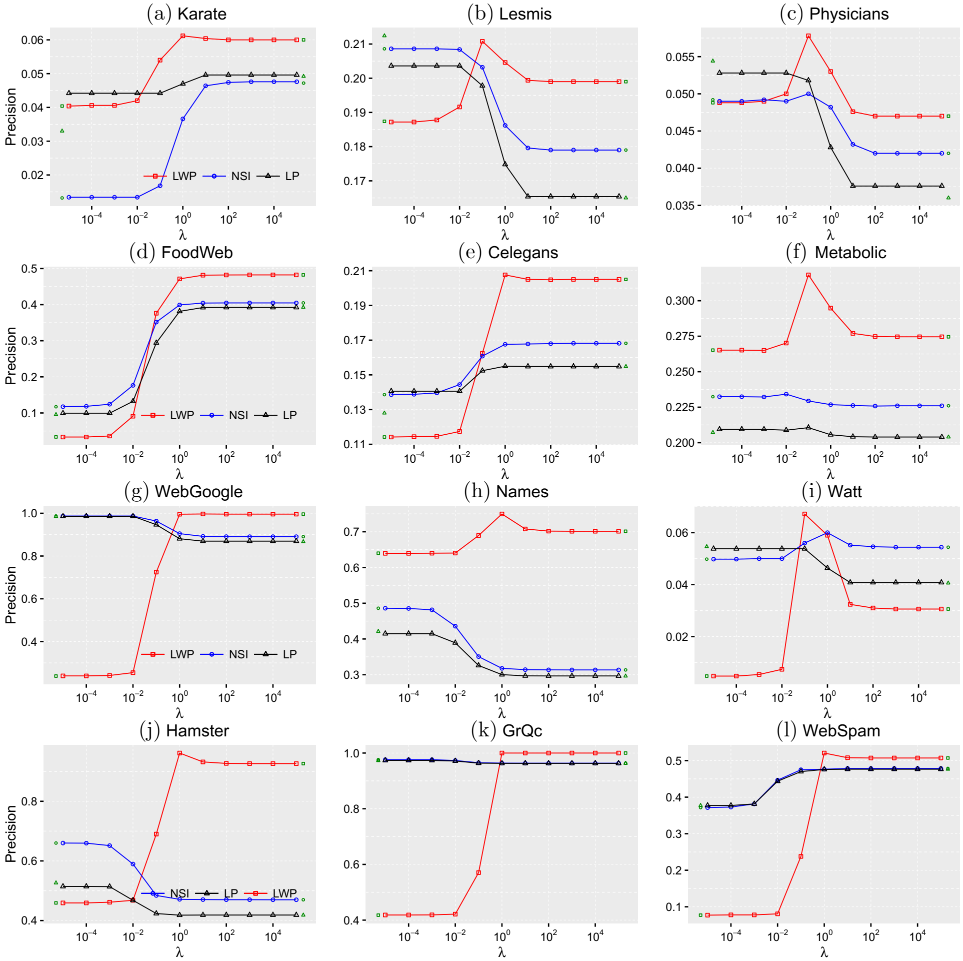

Fig. A.2. (Color online) The accuracies of LWP, NSI and LP indices measured by Precision with top-100 in different networks. Each value is averaged over 50 independent implementations. The green points of square, circle and triangle on the far left in figures represent the accuracies of LWP, NSI and LP indices that only exploit paths with length 2, respectively. The green points of square, circle and triangle on the far right in figures represent the accuracies of LWP, NSI and LP indices that only exploit paths with length 3, respectively.

prediction performance of LWP index always falls between 0 and 1. In other words, we can find the optimal parameter of λ from 0 to 1 for the LWP index.

References

1. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D.-U. Hwang, *Phys. Rep.* **424**, 175 (2006).
2. D. J. Watts and S. H. Strogatz, *Nature* **393**, 440 (1998).
3. A.-L. Barabási and R. Albert, *Science* **286**, 509 (1999).

4. A.-L. Barabási, *Nat. Phys.* **8**, 14 (2011).
5. B. Zhu and Y. Xia, *Sci. Rep.* **5** (2015) 13707 pp. Nature Publishing Group.
6. L. Lü and T. Zhou, *Phys. A, Stat. Mech. Appl.* **390**, 1150 (2011).
7. C. V. Cannistraci, G. Alanis-Lobato and T. Ravasi, *Sci. Rep.* **3** (2013) 1613 pp. Nature Publishing Group.
8. B. Kaya and M. Poyraz, *Comput. Biol. Med.* **63**, 1 (2015).
9. B. Barzel and A.-L. Barabási, *Nat. Biotech.* **31**, 720 (2013).
10. G. Kossinets, *Soc. Netw.* **28**, 247 (2006).
11. V. StröEle, G. ZimbrãO and J. M. Souza, *J. Syst. Softw.* **86**, 1819 (2013).
12. Q.-M. Zhang, A. Zeng and M.-S. Shang, *PloS one* **8**, e62624 (2013).
13. H. Liao, A. Zeng, R. Xiao, Z.-M. Ren, D.-B. Chen and Y.-C. Zhang, *PloS one* **9**, e97146 (2014).
14. D. Liben-Nowell and J. Kleinberg, *J. Am. Soc. Inf. Sci. Technol.* **58**, 1019 (2007).
15. D. Lin, An information-theoretic definition of similarity, in *ICML*, 1998.
16. A. Clauset, C. Moore and M. E. Newman, *Nature* **453**, 98 (2008).
17. R. Guimerà and M. Sales-Pardo, *Proc. Nat. Acad. Sci.* **106**, 22073 (2009).
18. Z. Liu, J.-L. He, K. Kapoor and J. Srivastava, *PloS one* **8**, e72908 (2013).
19. X. Zhu, H. Tian, S. Cai, J. Huang and T. Zhou, *Europhys. Lett.* **106**, 18008 (2014).
20. M. E. Newman, *Phys. Rev. E* **64**, 025102 (2001).
21. L. A. Adamic and E. Adar, *Soc. Netw.* **25**, 211 (2003).
22. T. Zhou, L. Lü and Y.-C. Zhang, *Eur. Phys. J. B* **71**, 623 (2009).
23. G. Salton and M. J. McGill *Introduction to Modern Information Retrieval* (McGraw-Hill Inc., 1983), pp. 305–30X.
24. T. Sorenson, *K. Dan. Vidensk. Selsk.* **5**, 4 (1948).
25. E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai and A.-L. Barabási, *Science* **297**, 1551 (2002).
26. L. Katz, *Psychometrika* **18**, 39 (1953).
27. E. A. Leicht, P. Holme and M. E. Newman, *Phys. Rev. E* **73**, 026120 (2006).
28. L. Lü, C.-H. Jin and T. Zhou, *Phys. Rev. E* **80**, 046122 (2009).
29. W. Liu and L. Lü, *Europhys. Lett.* **89**, 58007 (2010).
30. Z. Wu, Y. Lin and Y. Zhao, arXiv:1504.01018.
31. Z. Wu, Y. Lin, J. Wang and S. Gregory, *Phys. A, Stat. Mech. Appl.* **452**, 1 (2016).
32. B. Yan and S. Gregory, *Phys. Rev. E* **85**, 056112 (2012).
33. J. Ding, L. Jiao, J. Wu, Y. Hou and Y. Qi, *Phys. A, Stat. Mech. Appl.* **417**, 76 (2015).
34. J. Ding, L. Jiao, J. Wu and F. Liu, *Knowl.-Based Syst.* **98**, 200 (2016).
35. X. Zhu, H. Tian and S. Cai, *Phys. A, Stat. Mech. Appl.* **413**, 515 (2014).
36. L. Li, L. Qian, J. Cheng, M. Ma and X. Chen, *J. Inf. Sci.* **41**, 167 (2014).
37. Z. Xu, C. Pu and J. Yang, *Phys. A, Stat. Mech. Appl.* **456**, 294 (2016).
38. P. Holme, B. J. Kim, C. N. Yoon and S. K. Han, *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **65**, 634 (2002).
39. G. Georgiadis and L. Kirousis, *Complexus* **3**, 147 (2006).
40. J. A. Hanley and B. J. McNeil, *Radiology* **143**, 29 (1982).
41. P. Pudil, J. Novovičová and J. Kittler, *Pattern Recognit. Lett.* **15**, 1119 (1994).
42. J. L. Herlocker, J. A. Konstan, L. G. Terveen and J. T. Riedl, *ACM Trans. Inf. Syst.* **22**, 5 (2004).
43. W. W. Zachary, An information flow model for conflict and fission in small groups, *J. Anthropol. Res.* (1977), pp. 452–473, JSTOR.
44. D. E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing* (Addison-Wesley Reading, 1993).
45. J. Coleman, E. Katz and H. Menzel, *Sociometry* **20**, 253 (1957).

46. R. Ulanowicz, C. Bondavalli and M. Egnotovich. Network analysis of trophic dynamics in South Florida ecosystem, FY97: The South Florida Ecosystem, Annual Report to the United States Geological Service Biological Resources Division Ref. No. [UMCES] CBL, pp. 98–123 (1998).
47. J. Duch and A. Arenas, *Phys. Rev. E* **72**, 027104 (2005).
48. R. A. Rossi and 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.
49. King james network dataset — KONECT (January 2016).
50. http://chrisharrison.net/projects/bibleviz/index.html, Accessed: 2014-08-22.
51. J. Kunegis, KONECT — The Koblenz Network Collection, in *Proc. Int. Conf. on World Wide Web Companion*, 2013.
52. R. A. Rossi and N. K. Ahmed, watt-1 — Miscellaneous Networks (2013). http://networkrepository.com/watt-1.php
53. R. A. Rossi and N. K. Ahmed, *SIGKDD Explor.* **17**, 37 (2016).
54. Hamsterster network dataset — KONECT (2016). http://konect.uni-koldenz.de/networks/petsler-hamster
55. R. A. Rossi and N. K. Ahmed, ca-grqc — miscellaneous networks (2013).
56. R. A. Rossi and N. K. Ahmed, web-spam — web graphs (2013).
57. G. H. Golub and C. F. van Loan, Matrix Computations, 1996, John Hopkins University Press, Baltimore MD, USA, pp. 374–426.