2017

Combining Topological Properties and Strong Ties for Link Prediction

Fulan Qian
the School of Computer Science and Technology, Anhui University, Hefei 230601, China.

Yang Gao
the School of Computer Science and Technology, Anhui University, Hefei 230601, China.

Shu Zhao
the School of Computer Science and Technology, Anhui University, Hefei 230601, China.

Jie Tang
de the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.

Yanping Zhang
the School of Computer Science and Technology, Anhui University, Hefei 230601, China.

Follow this and additional works at: https://tsinghuauniversitypress.researchcommons.org/tsinghua-science-and-technology

Part of the Computer Sciences Commons, and the Electrical and Computer Engineering Commons

Recommended Citation
Fulan Qian, Yang Gao, Shu Zhao et al. Combining Topological Properties and Strong Ties for Link Prediction. Tsinghua Science and Technology 2017, 22(6): 595-608.

This Research Article is brought to you for free and open access by Tsinghua University Press: Journals Publishing. It has been accepted for inclusion in Tsinghua Science and Technology by an authorized editor of Tsinghua University Press: Journals Publishing.
Combining Topological Properties and Strong Ties for Link Prediction

Fulan Qian*, Yang Gao, Shu Zhao, Jie Tang, and Yanping Zhang

Abstract: Link prediction is an important task that estimates the probability of there being a link between two disconnected nodes. The similarity-based algorithm is a very popular method that employs the node similarities to find links. Most of these types of algorithms focus only on the contribution of common neighborhoods between two nodes. In sociological theory relationships within three degrees are the strong ties that can trigger social behaviors. Thus, strong ties can provide more connection opportunities for unconnected nodes in the networks. As critical topological properties in networks, nodes degrees and node clustering coefficients are well-suited for describing the tightness of connections between nodes. In this paper, we characterize node similarity by utilizing the strong ties of the ego network (i.e., paths within three degrees) and its close connections (node degrees and node clustering coefficients). We propose a link prediction algorithm that combines topological properties with strong ties, which we called the TPSR algorithm. This algorithm includes TPSR2, TPSR3, and the TPSR4 indices. We evaluate the performance of the proposed algorithm using the metrics of precision and the Area Under the Curve (AUC). Our experimental results show the TPSR algorithm to perform remarkably better than others.

Key words: complex networks; link prediction; strong ties; topological properties

1 Introduction

Different kinds of data in many fields can be represented as networks, with nodes as individuals and edges representing the interaction between them. Examples include friendship and social networks, food webs, protein-protein interaction networks, and the World Wide Web. Research on complex networks has attracted wide attention and become a focus of many branches of science. Link prediction, a fundamental task in link mining and complex network analysis has a wide range of applications, such as recommender systems, information retrieval, and bioinformatics. Two different settings of the link prediction problem are commonly studied. In the first setting, a snapshot of the network at time \( t \) or a sequence of snapshots from time 1 to time \( t \) are used to predict new links that are likely to appear in the near future (at time \( t + 1 \)). In the second setting, the network is treated as static but not fully observed, and the task is to fill in the missing links of the partially observed network. Here we focus on the partially observed setting and do not consider networks that evolve over time. Many algorithms have been proposed from a variety of disciplines. Some are based on Markov chains\(^{1, 2}\) or machine learning and a series of algorithms are based on node similarity\(^{3}\).

In this paper, we mainly focus on similarity-based algorithms. To some extent, these algorithms are based on the intuition that two nodes are similar if they have many common neighbors. According to Lü and Zhou\(^{4}\), similarity-based algorithms can be categorized into three classes: local similarity indices, global similarity indices, and quasi-local similarity indices.
indices. Many methods approach the link prediction question with respect to the local topology structure of the target nodes. These methods have two main aspects. (1) They utilize the influence of the direct neighbors on the target nodes. The descriptions of influence include the node degrees, node clustering coefficients, and the number of common neighbors. For example, the Common Neighbors (CN)\cite{5} and the Jaccard\cite{6} methods directly define the similarity of nodes based on their common neighbors. The Adamic-Adar (AA)\cite{7}, Resource Allocation (RA)\cite{8}, and the Cohesive Common Neighbors (CCN)\cite{9} methods utilize node degrees to describe the similarities between any two nodes. The importance of node clustering coefficients is utilized in some approaches, such as the Clustering Coefficient for Link Prediction (CCLP)\cite{10} and the Clustering Ability (CA)\cite{11} methods. (2) They take advantage of the effect of indirect nodes on the target nodes. In other words, these methods consider local path information. For example, the Katz method\cite{12} counts all paths connecting two nodes and penalizes longer paths. To address problems of low accuracy with respect to local similarity and the high complexity of global similarity indices, the Local Path (LP) method\cite{13} was proposed which is based on Katz method concept but limits the length of paths considered. Other methods based on LPs include the Local Random Walk (LRW)\cite{14}, FriendLink (FL)\cite{15}, and SRank\cite{16} methods. The LRW method is a constrained random walk based method that limits the range of the random walker. In the FL method, long paths are ignored and each path is penalized with a structural coefficient according to its length. The SRank is a typical shortest-paths similarity measure, in which two nodes are considered to be similar if there are multiple small-length paths connecting them.

In real life, the more closely two nodes are connected by strong ties, the more new relationships occur between them. In sociological research, relationships within three degrees are considered to be strong ties that can trigger social behaviors\cite{17,18,19}. To shorten the long execution time of the greedy algorithm for the linear threshold model, Lei et al.\cite{20} proposed the heuristic Three Degrees of Influence Algorithm (TDIA). Gong et al.\cite{21} proposed a memetic algorithm for community-based influence maximization in social networks, which optimizes the two-hop influence spread to find the most influential nodes. In social recommendation systems, the Sorec\cite{22}, SoReg\cite{23}, and SocialMF\cite{24} algorithms improve system performance based on the relationships of the direct user. Triggering behavior can provide more connection opportunities for unconnected nodes in the network and has great impact on the generation of new network links. Applying strong ties to link prediction methods can improve prediction performance. In this paper, we construct an ego network with target node-centric paths of length 3. In the ego network, the relationships between nodes and target nodes within three steps are strong ties that have the strong possibility of linking with each other. In addition, we can obtain the node degrees and node clustering coefficients from ego networks. To some extent, the degree and clustering coefficient can reflect the tightness of the connection between the local nodes in the networks as well as the stability of the ego network. These properties also have an important impact on the generation of new links. In this paper, we characterize the similarity between nodes by the strong ties of the ego network and the tightness of the node connections. Node degrees and node clustering coefficients are critical topological properties in networks and are well-suited for describing the tightness of node connections. As such, we propose a new algorithm for link prediction that combines topological properties and strong ties, which we call the TPSR algorithm. This algorithm includes three similarity indices—the TPSR2, TPSR3, and TPSR 4 indices, respectively. To validate our proposed algorithm, we use 14 real-world networks and compare ours with six existing algorithms including CN, Adamic-Adar, RA, LP, Katz, and CCN algorithms. Our experimental results reveal that our TPSR algorithm performs better than the others.

The contributions of this paper can be summarized as follows. Firstly, sociological research has found that relationships within three degrees comprise strong ties that can trigger behaviors in social networks. Our study proves that the consideration of strong ties has great impact on improving the performance of link prediction methods. Secondly, we propose a novel link prediction algorithm, the TPSR. By directly defining the similarity between nodes by combining the topological properties and the strong ties of the ego network, the TPSR algorithm yields higher prediction accuracy. Lastly, our experimental results comparing our TPSR algorithm with six other algorithms with respect to 14 real-world networks reveals the effectiveness of our proposed TPSR algorithm.

The rest of this paper is organized as follows. In
Section 2, we review important link prediction methods. We define the proposed similarity index and describe the TPSR algorithm in Section 3 and discuss our experimental results in Section 4. Finally, we draw our conclusions in Section 5.

2 Similarity-Based Link Prediction Algorithms

Link prediction is a well-known task involving link mining in complex networks, whereby links infer interactions between individuals that are likely to occur in the near future. The majority of existing link prediction methods focus on measuring the similarities between disconnected nodes to predict future connections. Generally speaking, the basis of these methods, known as similarity-based methods, is to define the similarity index between nodes.

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

**CN index**[5]: The basic idea of this index is that two individuals are more likely to interact in the future if they have many common neighbors. To measure the similarity between two different nodes, CN directly counts their common neighbors. The similarity index is defined in Eq. (1) as follows:

$$S_{xy}^{CN} = |\Gamma(x) \cap \Gamma(y)|$$  \hspace{1cm} (1)

**AA index**[7]: This similarity index refines the simple counting of common neighbors by assigning less weight to high-degree neighbors. Compared with CN, AA can differentiate the contributions of neighbors. Its definition is as follows:

$$S_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\lg k_z}$$  \hspace{1cm} (2)

**RA index**[8]: RA can be considered to be a revision of the AA index, which further reduces the contributions of high-degree neighbors. Its definition is as follows:

$$S_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z}$$  \hspace{1cm} (3)

**Katz index**[12]: This index is a global similarity measure. Its basic idea is that the more paths there are connecting two nodes, the greater is their similarity. When computing the similarity of two nodes, Katz directly sums all paths connecting these nodes by exponentially damping the longer paths. The Katz measure is defined as follows:

$$S_{xy}^{Katz} = \sum_{i=1}^{\infty} \beta^i |\text{path}_{x,y}^i|$$  \hspace{1cm} (4)

where $|\text{path}_{x,y}^i|$ is the number of paths connecting $x$ and $y$ with length $i$, and $\beta$ is a free parameter to assign less weight to longer paths. To ensure the convergence of the Katz index, $\beta$ must be lower than the reciprocal of the largest eigenvalue of the adjacency matrix of the network.

**CCN[9]**: This index considers the contribution of two- and three-hop common neighbors. In this index, each two- or three-hop common neighbor plays different roles in the node connection probability according to their degrees. Its definition is as follows:

$$S_{xy}^{CCN} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} + \sum_{m \in \Gamma(x), n \in \Gamma(y), m \in \Gamma(z)} \frac{1}{k_m \cdot k_n}$$  \hspace{1cm} (5)

**LP index**[13]: This index is a variant of the Katz index in which the length of paths is limited to the range of 2–3. As a result, it provides a good tradeoff between prediction accuracy and computational complexity. Its definition is as follows:

$$S_{xy}^{LP} = |\text{path}_{x,y}^2| + \varepsilon |\text{path}_{x,y}^3|$$  \hspace{1cm} (6)

where $\varepsilon$ is similar to the $\beta$ parameter in Katz. When $\varepsilon = 0$, this index degenerates to the CN index.

3 Proposed Method

The core of our similarity-based link prediction method is the computation of the similarities between nodes. The higher is the similarity score of two individuals, the higher is the probability that they will link with each other. In this section, we propose three novel similarity indices for link prediction, the TPSR2, TPSR3, and TPSR4 indices. Collectively, we refer to the above three methods as the TPSR algorithm. Before defining these new similarity indices in Section 3.2, we first explain our motivation for their development in Section 3.1.

3.1 Motivation

The basic idea of the proposed algorithm has two origins.

Firstly, using the CN-based indices gives a similarity score of zero when two nodes have no common neighbors. These indices only consider nodes with path lengths of 2, i.e., nodes with strong ties within two degrees. This criteria does not conform to objective facts. For example, there are no paths with lengths of 2 between nodes $x$ and $y$ in Fig. 1. So, the CN-based indices give a zero score for their similarity. In fact, they have a probability of linking with each other by the trigger behavior of the strong ties because there are three paths of length 3 between nodes $x$ and $y$, i.e.,
<x, i, j, y>, <x, f, j, y>, and <x, f, i, y>. As such, the similarity of nodes \(a\) and \(b\) should not be zero because these paths may contribute to the similarity of the nodes. This fact inspired us to introduce LP information into link prediction.

Secondly, we may want to calculate the similarity of nodes \(a\) and \(b\) and nodes \(b\) and \(y\), as shown in Fig. 1, with the following: \(S_{ab}^{CA} = S_{by}^{CA}\), \(S_{ab}^{AA} = S_{by}^{AA}\), \(S_{ab}^{RA} = S_{by}^{RA}\), \(S_{ab}^{LP2} = S_{by}^{LP2}\), \(S_{ab}^{CCN2} = S_{by}^{CCN2}\). Obviously, this is a bad result because these indices cannot efficiently distinguish between the contributions of common neighbors. In our opinion, although the degree of nodes \(c\) and \(p\) equals to the degree of nodes \(o\) and \(k\), the contribution of nodes \(c\) and \(p\) to \((a, b)\) is greater than that of nodes \(o\) and \(k\) to \((b, y)\), because the clustering coefficient and the degree of the common neighbor nodes yield different contributions to point nodes. Clustering coefficients and degree are very important topological properties in the networks. To some extent, they reflect the degree of connection closeness between local nodes in networks and the degree of stability of the ego network. In graph theory, the clustering coefficient is a well-studied attribute that measures the degree to which nodes in graph tend to cluster together. This coefficient has a great effect on the generation of new links. As such, making full use of the topological properties of nodes will benefit link prediction.

### 3.2 TPSR algorithm

In this section, we formally define the similarity indices of the TPSR algorithm whereby the more the path lengths differ, the more different is the similarity index. In this paper, we consider path lengths of 2, 3, and 4 only because the longer is the path between two nodes, the lower is the contribution of the path to the similarity of the nodes. Calculating the contribution of longer paths is also very time-consuming. The “small-world phenomenon”

\[ S_{xy}^{TPSR2} = \sum_{z \in L_{x,y}^{3}, z \neq x, z \neq y} \frac{1}{k_z} + c_z \]

where \(k_z\) is the degree of node \(z\) and \(c_z\) is the clustering coefficient of node \(z\). This is the definition of the TPSR2 index.

**Definition 1:** Given two nodes \(x\) and \(y\) in the graph \(G(V, E)\), \(L_{x,y}^{2}\) denotes the set of paths connecting \(x\) and \(y\) with length 2. The similarity between \(x\) and \(y\) is defined as follows:

\[ S_{xy}^{TPSR3} = S_{xy}^{TPSR2} + \theta |L_{x,y}^{3}| \]

where \(|L_{x,y}^{3}|\) denotes the number of paths connecting \(x\) and \(y\) with length 3 and \(\theta\) is similar to the parameter in the LP index. This is the definition of the TPSR3 index.

**Definition 2:** Given two nodes \(x\) and \(y\) in graph \(G(V, E)\), \(L_{x,y}^{4}\) denotes the set of paths connecting \(x\) and \(y\) with length 3. The similarity between \(x\) and \(y\) is defined as follows:

\[ S_{xy}^{TPSR4} = S_{xy}^{TPSR3} + \theta |L_{x,y}^{4}| + \theta^2 |L_{x,y}^{4}| \]

where \(|L_{x,y}^{4}|\) denotes the number of paths connecting \(x\) and \(y\) with length 4. This is the definition of the TPSR4 index. Next, we explore the effect of path length on the prediction accuracy by comparing the TPSR2, TPSR3, and TPSR4 indices with the other start-of-the-art algorithms. Our experimental results show that the indices within three degrees can achieve better results in the majority of datasets. We provide specific details in Section 4.

In this paper, with the TPSR3 and TPSR4 indices, we accept the suggestion of the LP index that sets the parameter \(\theta=0\). Our method differs somewhat from the others. Firstly, the TPSR2 index uses the degrees and clustering coefficients of common nodes to calculate the similarity of the end nodes. It can apply more topological features information of nodes in the network and then embody differences between common nodes, and it also improves the performance of the link prediction algorithm by exploiting these differences. Secondly, the TPSR3 index considers the length-3 paths based on the TPSR2 index and makes full use of the effect of the strong ties in the ego network. The more
paths there are between nodes, the greater is the chance that they will connect. This not only has the advantage of the TPSR2 index, but also performs better with respect to nodes that have no common neighbors. This means that the TPSR3 index expands the prediction range of the TPSR2 index. Lastly, if we compare the performance of the TPSR4 with that of the TPSR2 and TPSR3, we can conclude that the indices within length 3 achieved better results. So, there is no need to take time to calculate longer TPSR path indices. This comparison also shows that strong ties can play a role in improving the accuracy of link prediction.

As illustrated in Algorithm 1, the main operations of the TPSR algorithm consist of lines 3–6 and lines 7–9. In the worst case, the time complexity of lines 3–6 is $O(N^2)$ and the time complexity of lines 7–9 is $O(N^3)$. Therefore, the overall time complexity of the TPSR algorithm is $O(N^3)$. Thus, it has the same time complexity as the LP algorithm and less than that of the Katz algorithm.

4 Evaluation and Results

In this section, we experimentally evaluate the efficient of the TPSR compared with those of CN, AA, RA, LP, Katz, and CCN algorithms.

4.1 Evaluation metrics

We used two standard metrics to quantify the accuracy of the prediction algorithms: the area under the receiver-operator curve (AUC) and precision.

AUC: By providing the rank of all non-observed links, the AUC value can be interpreted to be the probability that a randomly chosen missing link will be given a higher score than a randomly chosen nonexistent link. In algorithmic implementation, we usually calculate the score of each non-observed link rather than the list order since the latter task is more time-consuming. Then, each time, we randomly select a missing link and a nonexistent link to compare their scores. If among $n$ independent comparisons there are $n'$ times for which the missing link had a higher score and $n''$ times for which they had the same score, the AUC value is as follows:

$$\text{AUC} = \frac{n' + 0.5n''}{n} \quad (10)$$

Precision: Given the ranking of the non-observed links, we define precision as the ratio of relevant items to the number of items selected. That is to say, if we take the top-$L$ links as the predicted links, among which $m$ links are correct, then the precision value is as follows:

$$\text{Precision} = \frac{m}{L} \quad (11)$$

Generally $L = 100$. Clearly, higher precision indicates higher prediction accuracy.

4.2 Networks

In this study, we performed experiments on 14 real-world networks. We briefly introduce each benchmark network below:

(1) C.elegan\textsuperscript{[28]}. This network represents the connections of the frontal ganglia of the nematode worm C.elegans.

(2) NetScience (NS)\textsuperscript{[29]}. In this network, nodes and links represent scientists and the co-authorships between them, respectively.

(3) Jazz\textsuperscript{[30]}. This is a network of jazz bands, in which a link between two bands is established if they have a common musician.

(4) USAir97\textsuperscript{[31]}. This is the USAir transportation network.

(5) Food Web. This comprises four food chain networks, including FWEW\textsuperscript{[32]}, FEMW\textsuperscript{[33]}, FWFW, and FWFD\textsuperscript{[32]}.

(6) Facebook (FB)\textsuperscript{[34]}. This is the well-known friendship network.

(7) Political Book (PB)\textsuperscript{[35]}. This pol-blog network is extracted from a set of weblogs about US politics.
(8) Router\[36\]. The router network has 5022 nodes and 6258 links.

(9) Power\[37\]. This network comprises an electrical power grid in the western US, wherein nodes representing generators, substations, transformers, and edges representing the high tension lines between them.

(10) WCGScience (WCGS)\[38\]. This is a scientific co-authorships network.

(11) Yeast\[39\]. This is a protein-protein interaction network.

In the experiments, we treated all of the networks listed above as undirected and unweighted networks, whether or not they are actually weighted and/or directed. We removed all loops and allowed no multilinks. Table 1 lists the basic statistics of these networks. To ensure a fair comparison, we computed all the evaluation metrics values as the average values of 100 iterations. For each iteration, we used a set of 90% randomly selected network interactions as a training set for the algorithms and the remaining 10% interactions were used for the test set.

### 4.3 Results and analysis

In this section, we compare our method with six other similarity indices—CN, AA, RA, LP, Katz, and CCN indices. First, we compare the performance of the TPSR algorithm with those of the six other similarity-based algorithms. Next, we analyze the performance of TPSR2, TPSR3, and TPSR4 indices, respectively. Third, we make some suggestions to the different networks for selecting the best link prediction method, based on our use of the Newman-Watt (NW) small-world network model\[40\] to generate ten networks with different features. Lastly, we discuss the time complexity of the proposed method.

#### 4.3.1 Performance in real-world networks of TPSR compared with other algorithms

Table 2 shows the prediction accuracies as measured by the AUC and Table 3 presents the results for another widely used metric—precision. The highest AUC/precision for each network (in each column) is shown in boldface. For the AUC metric, the TPSR algorithm performs best in most networks. Compared with the Katz algorithm on the Food Networks, in particular, the AUC of the TPSR is more than 10% better. For the precision metric, the TPSR algorithm performs best in 13 out of 14 networks and is 11.3% better on the Yeast network. In addition, the TPSR3 index performs better than the LP index on the most networks, including Yeast (by 11.8%), FWFW (by 21.2%), FWFD (by 22%), NS (by 10%), and Router (by 9.3%). Based on these results, we can see that the proposed algorithm is either the best or very close to the best.

We can also verify the robustness of proposed algorithm by Figs. 2 and 3 since, in most networks, the accuracy of the TPSR algorithm is either the best or very close to the best, even when we varied the size of the training sets (ratio of known edges). These results indicate that, on sparse networks, the TPSR algorithm also has better prediction accuracy than other algorithms (i.e., CN, RA, LP, and Katz). The main reason for this is that the proposed algorithm utilizes the effect of the strong ties on the ego network and the contribution of node degrees and clustering coefficients.

Table 4 shows a comparison of the prediction accuracy of the TPSR and CCN algorithms on all tested networks. The difference between the TPSR and CCN is that the CCN algorithm does not consider the clustering coefficient of common neighbor nodes. From Table 4, we find that in most cases the TPSR algorithm has better prediction accuracy than the CCN algorithm. Specifically, on the FWFW network, the TPSR precision was greater by 18.2%, on the FWFD network it was greater by 18.3%, on the PB network it was greater by 20.5%, on the Router network it was greater by 11.3%, and on the Yeast network precision was greater by 22%.

---

Table 1 Basic network statistics. $N$ is the number of nodes, $M$ is the number of edges, $<c>$ is the average clustering coefficient of a network, $<d>$ is the average degree of the network, and $<m>$ is the network density.

| Network | $N$ | $M$ | $<c>$ | $<d>$ | $<m>$ |
|---------|-----|-----|-------|-------|-------|
| FWEW    | 69  | 880 | 0.5521| 25.0702| 0.3751 |
| FWMW    | 97  | 1446| 0.4683| 29.8144| 0.3106 |
| FWFW    | 128 | 2075| 0.3346| 32.4219| 0.2553 |
| FWFD    | 128 | 2105| 0.3347| 32.8906| 0.2590 |
| Jazz    | 198 | 2742| 0.6175| 27.6970| 0.1406 |
| Celegan | 297 | 2148| 0.2924| 14.4646| 0.0489 |
| USAir   | 332 | 2126| 0.6252| 12.8072| 0.0387 |
| PB      | 1222| 16714|0.3203| 27.3552| 0.0224 |
| NS      | 1589| 2742| 0.6378| 3.4512 | 0.0022 |
| Yeast   | 2375| 11693|0.3057| 9.8467 | 0.0041 |
| FB      | 4039| 88234|0.6055| 43.6910| 0.0108 |
| Power   | 4941| 6594 |0.0801| 2.6691 | 0.0005 |
| Router  | 5022| 6258 |0.0116| 2.4922 | 0.0004 |
| WCGS    | 7343| 11898|0.4075| 3.2406 | 0.0004 |
The above results confirm the advantages of the TPSR algorithm with respect to prediction accuracy and the fact that strong ties (paths within three degrees) and close connections (node degrees and clustering coefficients) positively impact the performance of the link prediction algorithm.

In addition, the length value of list $L$ in link prediction is 100 in the conventional calculation of the precision metric. However, in this case only, the precision cannot fully explain the performance of the prediction algorithm. In view of this, we set $L$ to different values. As shown in Fig. 4, the comparative algorithms included the CN, LP3 (LP algorithm with path length 3), and LP4 algorithms (LP algorithm with path length 4). The $\varepsilon$ value is 0.01, which is the parameter value proposed by the LP algorithm and that proposed in Eq. (6). From Fig. 4, we can see clearly that the TPSR performs better than the other algorithms (i.e., CN, LP3, and LP4) in 12 of the 14 networks. In the Jazz, NS, and WCGS networks, when compared with the CN, LP3, and LP4 algorithms, the TPSR algorithm also obtained the best precision with the various $L$ values. The main reason for this is that the topological properties of these three networks greatly influence the prediction accuracy. Compared to the CN algorithm, the TPSR always yielded the best precision in the networks. This indicates the importance of strong ties in the ego network. In other words, these results reveal the TPSR algorithm to still yield higher prediction accuracy for the various $L$ values, thus demonstrating that the TPSR algorithm has a wide range of applicability.
Fig. 2 Comparison of prediction accuracy with respect to the AUC metric. The ratio of known edges varied from 0.5 to 0.9.
Fig. 3 Precision curves for CN, LP3, LP4, TPSR2, TPSR3, and TRSR4 indices on 14 real-world networks.
4.3.2 Analysis of the performance of the three TPSR indices

Figures 5 and 6 show the AUC and precision results, respectively, for the TPSR2, TPSR3, and TPSR4 indices. From Fig. 5, we can see that the TPSR2 index achieves the best AUC performance in three out of 14 networks, whereas the TPSR3 index obtains the best performance in the other seven. In the Power network, the TPSR4 obtained the best performance. From Fig. 6, we see that the TPSR2 index obtained the best performance in four networks, including the Jazz, USAir, NS, and WCGS. The TPSR3 index obtained the best performance in seven networks, including Celegan, FWFW, FWEW, FWFD, FWMW, PB, Router, and Yeast. In the FB network, the indices obtained the same performance. These results indicate that LPs with lengths of 2 and/or 3 are feasible and suitable for the TPSR algorithm, which reflects the fact that strong ties play a role in the generating new links on networks. In addition, the TPSR3 index obtained a higher prediction accuracy than the TPSR4 index in most networks. This result illustrates that the strong ties of the ego network have an important impact on the performance of the prediction algorithm. From Fig. 4, we see that the precision curves of the TPSR3 and the TPSR4 indices also demonstrate the influence of the strong ties.

4.3.3 Performance of TPSR in NW small-world model

Because real-networks may have some characteristics that are difficult to detect, it is difficult to establish a prediction method that is appropriate for application to all real-world networks. Here, to explore the general applicability of the TPSR method, we use the NW small-world model to construct 10 networks with different attributes (these networks are similar to the real networks generated by the NW model, so the artificial networks are similar to the real networks). The NW model obtained networks with different average clustering coefficients and average degrees by adjusting the \( m \) and \( p \) parameters. Parameter \( m \) regulates the average degree of the network and parameter \( p \) adjusts the average clustering coefficient of the network. We also use these two node attributes in the method we propose in this paper. Table 5 shows the feature information of the artificial networks and we set \( N = 1000 \). The experimental results show that the prediction performance of the TPSR method is related to the network topology. Figures 7 and 8 show the AUC and precision, respectively. We can see that the TPSR algorithm has better prediction performance in networks with large \( <c> \) and \( <d> \) values, whereby the smaller are the \( <c> \) and \( <d> \) values, the worse is the performance of TPSR algorithm. As such, there is a positive correlation between them.

4.3.4 Discussion of the time complexity of TPSR indices

In our discussion of the time complexity of the proposed algorithm, we consider only the TPSR2 and TPSR3 indices, the results of which are shown in Table 6. In general, some indices require topological properties, including the AA, RA, CCN, and TPSR indices, which can be incorporated in the data processing stage. The
Fig. 4 Comparison of prediction accuracies with respect to the precision metric. The ratio of known edges is varied from 0.5 to 0.9.
Fig. 2 AUC of TPSR2, TPSR3, and TPSR4 indices.

Fig. 3 Precision of TPSR2, TPSR3, and TPSR4 indices.

Fig. 4 Relation between the AUC of the TPSR indices and the average of degree and clustering coefficient values of the networks.

Fig. 5 Relation between the precision of the TPSR indices and the average degree and clustering coefficient values of the networks.

Table 5 Parameters $m$, $p$, and the features of the 10 networks.

| $m$ | $p$ | Network | $N$ | $M$ | $<c>$ | $<d>$ | $<m>$ |
|-----|-----|---------|-----|-----|-------|-------|-------|
| 10  | 0.001| net1    | 1000| 10009| 0.7093| 20.02 | 0.020 |
| 10  | 0.05 | net2    | 1000| 10522| 0.6446| 21.04 | 0.021 |
| 10  | 0.1  | net3    | 1000| 10963| 0.5975| 21.93 | 0.022 |
| 8   | 0.1  | net4    | 1000| 8825 | 0.5813| 17.65 | 0.018 |
| 6   | 0.15 | net5    | 1000| 6859 | 0.5317| 13.72 | 0.014 |
| 4   | 0.15 | net6    | 1000| 4622 | 0.4959| 9.24  | 0.009 |
| 4   | 0.3  | net7    | 1000| 5234 | 0.3896| 10.47 | 0.010 |
| 4   | 0.5  | net8    | 1000| 5997 | 0.295 | 11.99 | 0.010 |
| 3   | 0.7  | net9    | 1000| 5113 | 0.2128| 10.23 | 0.010 |
| 2   | 1.0  | net10   | 1000| 3996 | 0.1223| 7.99  | 0.008 |

Table 6 Time complexity of these prediction methods. $l$ is the length of considering the longest paths in the Katz index. $N$ is the number of nodes in the networks.

| Prediction Method | $O(N^2)$ | $O(N^3)$ | $O(N^3)$ | $O(N^3)$ | $O(N^3)$ | $O(N^3)$ |
|-------------------|-----------|-----------|-----------|-----------|-----------|-----------|
| CN                | $O(N^2)$  | $O(N^2)$  | $O(N^2)$  | $O(N^2)$  | $O(N^2)$  | $O(N^2)$  |
| AA                | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  |
| CCN               | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  |
| LP                | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  |
| Katz              | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  |
| TPSR2             | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  |
| TPSR3             | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  | $O(N^3)$  |

5 Conclusion

In this paper, we proposed a new link prediction algorithm, the TPSR, which is based on the strong ties of the ego network and its topological properties. According to the length of the path, we considered three new similarity indices, TPSR2, TPSR3, and TPSR4. We then compared the performance of six existing similarity-based algorithms with that of the proposed algorithm. We presented the experimental result for 14 real-world networks, which reveal the good prediction performance of the proposed algorithm. The TPSR algorithm can provide higher prediction accuracy than the CN, AA, LP, CCN, RA, and Katz algorithms. In addition, the TPSR has higher precision and less time complexity than the Katz algorithm. Our proposed algorithm captures more local network information via the strong ties (paths within three degree) and
topological properties (node degrees and clustering coefficients). In addition, we explored the relationship between the prediction accuracy of the TPSR algorithm and the average degrees and clustering coefficients of the networks, which exhibit positive correlations.

In future research work, we plan to evaluate our algorithm in directed and weighted networks.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China (Nos. 61673020, 61402006, and 61702003), the National High-Tech Research and Development (863) Program of China (No. 2015AA124102), Humanities and Social Science Research on Youth Fund Project, Ministry of Education (No. 14YJC860020), and Anhui Provincial Natural Science Foundation (No. 1708085MF160).

References

[1] F. Fouss, L. Yen, A. Pirotte, and M. Saerens, An experimental investigation of graph kernels on a collaborative recommendation task, in International Conference on Data Mining, 2006.

[2] Y. Yang, G. Hao, T. Tian, and H. Li, Link prediction in brain networks based on a hierarchical random graph model, Tsinghua Science and Technology, vol. 20, no. 3, pp. 306–315, 2015.

[3] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas, Characterization of complex networks: A survey of measurements, Advances in Physics, vol. 56, no. 1, pp. 167–242, 2010.

[4] L. Lü and T. Zhou, Link prediction in complex networks: A survey, Physica A Statistical Mechanics & Its Applications, vol. 390, no. 6, pp. 1150–1170, 2011.

[5] F. Lorrain and H. C. White, Structural equivalence of individuals in social networks, The Journal of Mathematical Sociology, vol. 1, no. 1, pp. 49–80, 1971.

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

[7] L. A. Adamic and E. Adar, Friends and neighbors on the web, Social Networks, vol. 25, no. 3, pp. 211–230, 2003.

[8] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King, Recommender systems with social regularization, in Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, 2011.

[9] M. Jamali and M. Ester, A matrix factorization technique with trust propagation for recommendation in social networks, in ACM Conference on Recommender Systems, Recsys 2010, Barcelona, Spain, 2010.

[10] L. Katz, A new status index derived from sociometric analysis, Psychometrika, vol. 18, no. 1, pp. 39–43, 1953.

[11] L. Lü, C. H. Jin, and T. Zhou,Similarity index based on local paths for link prediction of complex networks, Physical Review E, vol. 80, no. 4, p. 046122, 2009.

[12] A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos, Fast and accurate link prediction in social networking systems, Journal of Systems and Software, vol. 85, no. 9, pp. 2119–2132, 2012.

[13] H. Khosravi-Farsani, M. Nematbakhsh, and G. Lausen, SRank: Shortest paths as distance between nodes of a graph with application to RDF clustering, Journal of Information Science, vol. 39, no. 2, pp. 198–210, 2013.

[14] S. K. Walker, Connected: The surprising power of our friends’ friends’ friends affect everything you feel, think, and do, Mathematics and Computer Education, vol. 48, no. 2, p. 201, 2014.

[15] W. H. Lei, Q. Yang, and H. Wang, Positive influence maximization algorithm based on three degrees of influence, in International Conference on Intelligent Data Engineering and Automated Learning, 2016.

[16] M. Gong, C. Song, C. Duan, and L. Ma, An efficient memetic algorithm for influence maximization in social networks, IEEE Computational Intelligence Magazine, vol. 11, no. 3, pp. 22–33, 2016.

[17] H. Ma, H. Yang, M. R. Lyu, and I. King, SoRec: Social recommendation using probabilistic matrix factorization, in ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, CA, USA, 2008.

[18] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King, Recommender systems with social regularization, in Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, 2011.

[19] M. Jamali and M. Ester, A matrix factorization technique with trust propagation for recommendation in social networks, in ACM Conference on Recommender Systems, Recsys 2010, Barcelona, Spain, 2010.

[20] S. Milgram, The small world problem, Psychology Today, vol. 2, no. 1, pp. 60–67, 1967.

[21] J. A. Hanley and B. J. McNeil, A method of comparing the areas under receiver operating characteristic curves derived from the same cases, Radiology, vol. 148, no. 3, pp. 839–843, 1983.

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

[23] D. J. Watts and S. H. Strogatz, Collective dynamics of “smallworld” networks, Nature, vol. 393, no. 6684, pp. 440–442, 1998.
[29] M. E. J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E, vol. 74, p. 036104, 2006.

[30] B. P. Gleiser and L. Danon, Community structure in jazz, Advances in Complex Systems, vol. 6, no. 4, p. 565, 2003.

[31] L. Li, L. Qian, J. Cheng, M. Ma, X. Chen, Accurate similarity index based on the contributions of paths and end nodes for link prediction, Journal of Information Science, vol. 41, no. 2, pp. 167–177, 2014.

[32] R. E. Ulanowicz, J. J. Heymans, and M. S. Egnotovich, Network analysis of trophic dynamics in South Florida ecosystems, in Proceedings of South Florida Restoration Science Forum, 1999.

[33] D. Baird, J. Luczkovich, and R. R. Christian, Assessment of spatial and temporal variability in ecosystem attributes of the St Marks National Wildlife Refuge, Apalachee Bay, Florida, Estuarine Coastal & Shelf Science, vol. 47, no. 3, pp. 329–349, 1998.

[34] F. Li, J. He, G. Huang, Y. Zhang, Y. Shi, and R. Zhou, Nodecoupling clustering approaches for link prediction, Knowledge-Based Systems, vol. 89, pp. 669–680, 2015.

[35] L. A. Adamic and N. Glance, The political blogosphere and the 2004 U.S. election: Divided they blog, in Proceedings of the 3rd International Workshop on Link Discovery, 2005.

[36] L. Lü, L. Pan, T. Zhou, Y. C. Zhang, and H. E. Stanley, Toward link predictability of complex networks, Proceedings of the National Academy of Sciences, vol. 112, no. 8, pp. 2325–2330, 2015.

[37] A. K. Menon and C. Elkan, Link prediction via matrix factorization, in European Conference on Machine Learning and Knowledge Discovery in Databases, 2011.

[38] L. Lü and T. Zhou, Role of weak ties in link prediction of complex networks, in ACM International Workshop on Complex Networks Meet Information & Knowledge Management, 2009.

[39] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, Comparative assessment of largescale data sets of protein–protein interactions, Nature, vol. 417, no. 6887, pp. 399–403, 2002.

[40] M. E. Newman and D. J. Watts, Renormalization group analysis of the small-world network model, Physics Letters A, vol. 263, no. 4, pp. 341–346, 1999.

[41] P. Wang, B. Xu, Y. Wu, and X. Zhou, Link prediction in social networks: The state-of-the-art, Science China Information Sciences, vol. 58, no. 1, pp. 1–38, 2014.

Fulan Qian received the MS degree in computer application in 2002 and PhD degree in computer application in 2016 from Anhui University. She is currently an assistant professor in the School of Computer Science and Technology at Anhui University. Her main research interests include quotient space theory, social network, and recommender system.

Yang Gao received the bachelor degree from Anhui Jianzhu University in 2014. He is currently a master student in the School of Computer Science and Technology at Anhui University. His main research interests include link prediction, complex networks, recommendation system, and so on.

Shu Zhao received the PhD degree in computer science from Anhui University in 2007. She is now a professor in the School of Computer Science and Technology, Anhui University. Her current research interests include quotient space theory, granular computing, social network, and machine learning.

Jie Tang obtained the PhD degree from Tsinghua University in 2006. He is now an associate professor in Tsinghua University. His research interests include social network theories, data mining methodologies, machine learning algorithms, and semantic web technologies.

Yanping Zhang is currently a professor in Anhui University. She received the PhD degree from Anhui University in 2005. Her main research interests include computational intelligence, quotient space theory, artificial neural networks and intelligent information processing, machine learning, and so on. She is the Principal Investigator of some 973 projects and the leader of National Natural Science Foundations of China and Anhui Province.