Physica A 557 (2020) 124980

Contents lists available at ScienceDirect

Physica A

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

# Link prediction using node information on local paths

Furqan Aziz [a,b,*], Haji Gul [b], Ishtiaq Muhammad [b], Irfan Uddin [c]

[a] Centre for Computational Biology, University of Birmingham, Birmingham, B15 2TT, UK
[b] Department of Computer Science, IM Sciences, Peshawar, Pakistan
[c] Kohat University of Science and Technology, Kohat, Pakistan

ARTICLE INFO

*Article history:*
Received 17 July 2019
Received in revised form 1 January 2020
Available online 28 July 2020

*Keywords:*
Complex networks
Link prediction
Node information

ABSTRACT

Link prediction is one of the most important and challenging tasks in complex network analysis, which aims to predict missing link based on existing ones in a network. This problem is of both theoretical interest and has applications in diverse scientific disciplines, including social network analysis, recommendation systems, and biological networks. In this paper we propose a novel link prediction method that aims at improving the accuracy of existing path-based methods by incorporating information about the nodes along local paths. We investigate the proposed framework empirically and conduct extensive experiments on real-world datasets obtained from diverse domains. Results show that the proposed method has achieved increased prediction accuracy when compared to existing state-of-the-art link prediction methods.

© 2020 Elsevier B.V. All rights reserved.

## 1. Introduction

Many real-world complex systems can be conveniently represented by means of a complex network whose nodes represent entities and links represent interactions between nodes. The study of complex networks has therefore attracted a large amount of attention in the last few years and has found its applications in many branches of science including social networks [1], transportation networks [2,3], text networks [4], and biological networks [5]. Once a system is modelled as a network, a number of questions related to the underlying system can be answered. One of the most challenging problems associated with complex networks is that of link prediction which aims at estimating the likelihood of the existence of a link between two nodes that are not already connected. This problem has recently attracted the attention of many researchers from diverse scientific disciplines. Some of its applications include friend recommendation in social networks such as Facebook [6], predicting interactions between proteins [7], predict the connectivity of primate cortical networks [8], news recommendation to users [9], and predicting future collaborations between authors of research articles [10].

Over recent years, a number of link prediction algorithms have been proposed and successfully applied. Most of these methods are based on the idea of node similarity. Node similarity can be defined by using the essential attributes of nodes. In other words, two nodes are considered to be similar, if they share many common topological features [11]. Some of the most commonly used similarity based measures include common neighbour [12], Adamic Adar [13], Jaccard coefficient [14], resource allocation [15], and preferential attachment [16]. One of the advantages of local similarity measures is that such measures can be computed very efficiently and perform well in many cases. Another frequently

[*] Corresponding author.
*E-mail addresses:* f.aziz@bham.ac.uk (F. Aziz), ms171606772@imsciences.edu.pk (H. Gul), ms171607089@imsciences.edu.pk (I. Muhammad), irfanuddin@kust.edu.pk (I. Uddin).

https://doi.org/10.1016/j.physa.2020.124980
0378-4371/© 2020 Elsevier B.V. All rights reserved.

used approaches are the quasi local/global metrics. These methods are generally based on local path information. These include, but are not limited to, Katz Index [17], commute time [18], and local path [19]. In a related work Yu et al. [20] have proposed a hybrid method that combines Katz index and resource allocation index to estimate the prediction score. They have empirically demonstrated that their proposed method gives superior performance on five different datasets. A comparison of different local and global link prediction algorithms can be found in several survey papers such as [21] and [22].

Besides the similarity-based prediction algorithms, a number of alternate methods have been proposed over the last two decades. Most of these methods are based on probabilistic and maximum likelihood methods. For example, Gao et al. [23] have recently introduced a linear dynamical response-based similarity measure between nodes and have designed a procedure to efficiently compute it. Zhu et al. [24] have proposed an information theoretic model by investigating the role of network topology in predicting missing links. Clauset et al. [25] proposed a Hierarchical Structure Model that exploits the hierarchical structure of a network to estimate the likelihood of a link. Pech et al. [26] have introduced robust principal component analysis (robust PCA) to estimate the missing entries of the adjacency matrix of a network.

In this paper we propose a new similarity index between two nodes $u$ and $v$, which is computed from the local information of those nodes that lie within a fixed distance from $u$ and $v$. In other words, we do not only include the information of the nodes that are directly connected to $u$ an $v$, but we also incorporate the information about those nodes that lie on set of all possible paths of smaller length from $u$ to $v$. The proposed method combines the advantages of the two most widely used similarity indices, i.e., Adamic Adar [13] and Katz Index [17], and gives superior performance when compared to both of them.

The rest of the paper is organised as follows. In Section 2, we present the link prediction problem and discuss some of the state of the art link prediction algorithms. The proposed algorithm is presented in Section 3. In Section 4, we give experimental evaluations, where we evaluate the accuracy of the proposed method on numerous publicly available benchmark datasets and compare it with existing state-of-the-art methods. Finally Section 5 gives some concluding remarks with possible future directions.

## 2. Overview of link prediction

In this section we present existing state-of-the-art link prediction algorithms. These methods are also used for comparison purposes in Section 4. We commence by providing some basic definitions that will be used throughout the paper. A *network* $G = (V, E)$ consists of a finite nonempty set $V$ of *nodes* and a finite set $E$ of unordered pairs of vertices, called *links*. A network can be directed (where each edge has a direction assigned to it) or undirected (where edges are not assigned any direction). In this paper we will consider undirected simple networks, where multiple links and self loops are not allowed. The *degree* of a vertex $v \in V$ is the number of neighbours linked to $v$. A *walk* $w$ in a network $G = (V, E)$ is a sequence of alternating vertices and edges $v_0, e_1, v_1, e_2, v_2, \ldots, e_k, v_k$ where $v_i \in V$ and $e_i = (v_{i-1}, v_i)$. This walk has length $k$, which is defined as the number of vertices in the walk. The adjacency matrix of a network of an undirected simple network $G = (V, E)$ is a $|V| \times |V|$ matrix, whose $(u, v)^{th}$ entry is 1 if $u$ and $v$ are linked and otherwise 0. The $(u, v)^{th}$ entry of the $k$th power of the adjacency matrix , $(A^k)_{uv}$, represents the number of walks of length $k$ from $u$ to $v$.

We now present some commonly used local and quasi local/global link prediction methods. The same methods are used as a baseline for comparison in experimental evaluation section. These are defined as follows:

**Common Neighbour (CN)** [12]: According to this method, two nodes are most likely to have a link if they share many common neighbours. In other words, this simple measure counts the number of common neighbours of two nodes. This index is also called structural equivalence [12]. Given two nodes it can be computed as:

$$CN(u, v) = |\Gamma(u) \cap \Gamma(v)| \tag{1}$$

CN can be computed very efficiently and it generally performs very well in many cases [27]. Therefore, it is considered as a baseline to assess the performance of other methods. Note that CN can also be expressed as $CN(u, v) = \left(A^2\right)_{uv}$, where $A$ is the adjacency matrix of the network.

**Adamic Adar (AA)** [13]: This index aims at improving the accuracy of common neighbours by assigning more weights to less connected neighbours. This index is defined as:

$$AA(u, v) = \sum_{w \in \{\Gamma(u) \cap \Gamma(v)\}} \frac{1}{\log |\Gamma(w)|}, \tag{2}$$

where $w$ is a common neighbour of $u$ and $v$. In social networks this index can be interpreted as follows: An unpopular person (someone with less number of friends) may be more likely to introduce a particular pair of his friends to each other.

**Jaccard Coefficient (JC)** [14]: The Jaccard coefficient, also know as Jaccard index, is another similarity measure that is computed from the common neighbours of two nodes. It is defined as:

$$JC(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|} \tag{3}$$

The value of JC can be interpreted as the probability that a neighbour of nodes $u$ and $v$ is a neighbour of both $u$ and $v$.

**Katz Index** ($Katz_\beta$) [17]: This method considers set of all paths between pair of nodes. The paths are exponentially damped by the length of the path to give more weightage to shorter paths. Since this method is based on the topology of the whole network, therefore it is generally computationally more expensive than local indices. Mathematically, it can be computed as

$$Katz(u, v) = \sum_{l=1}^{\infty} \beta^l \cdot \left[\text{path}_{uv}^{\langle l \rangle}\right] \tag{4}$$

where $0 < \beta < 1$ is the parameter that controls the weight of the paths of different lengths. A very small value of $\beta$ will result in penalising the paths of longer lengths and measure reduces to CN in this case. Note that the similarity matrix $S$ can also be computed as $(I - \beta A)^{-1} - I$, where $I$ represent identity matrix of size $V$.

**Local Paths (LP)** [19]: To provide a good trade-off between accuracy and Complexity Lü et al. [19] have introduced an index that considers local paths of shorter lengths. It is defined as

$$LP(u, v) = \sum_{i=2}^{l} \beta^{i-2} A^i, \tag{5}$$

where $A$ is the adjacency matrix of the network. As with Katz index, $\beta$ is set to a small value so that shorter paths get more weights. Note that this index degenerates to common neighbour, if $\beta = 0$. In practice, due to its computational complexity, this measure is usually used with $l = 3$, where it reduces to $LP(u, v) = A^2 + \beta A^3$.

**Preferential Attachment (PA)** [16]: This index is based on the well known concept in social network that users with many friends tend to get more connections in future. It is defined as:

$$PA(u, v) = |\Gamma(u)| \cdot |\Gamma(v)| \tag{6}$$

This index can be computed very efficiently, and it performs very well for many social network datasets.

## 3. Proposed framework

### 3.1. Motivation

The similarity-based link prediction methods generally depend upon the topological structure of a network. The objective of such methods is to predict missing links between a pair of nodes based on their local connectivity structure. The local link prediction methods use the node information of immediate neighbours of two nodes to predict link between them. For instance, the AA index assigns higher weight to a common neighbour with smaller degree. The intuition here is that common neighbours with lesser degrees are more significant in predicting links between two nodes. While the local similarity indices consider only those nodes that are directly connected to the query nodes, the global methods (such as Katz index and LP index) consider the topology of the whole network. However, the global methods do not consider the degrees of the connecting nodes. Therefore, there is a need of a link prediction method that can simultaneously take advantage of local and global similarity indices to measure the similarity.

### 3.2. Model definition

Consider an unweighted network $G = (V, E)$. Suppose we want to predict a link between two nodes $u$ and $v$ with $(u, v) \in E'$. Inspired by the above discussion, to estimate the likelihood of a potential connection between node $u$ and $v$ we consider all those nodes of the network lying within a fixed distance from the two nodes $u$ and $v$. Our proposed similarity index (PSI) is defined as follows:

**Definition 1.** Let $\Gamma^k(u)$ represent the set of all nodes whose shortest distance from the node $u$ is $k$. Then the similarity score, $PSI(u, v)$, between two nodes $u$ and $v$ is defined as:

$$PSI(u, v) = \sum_{\substack{i+j=2 \\ i,j>0}}^{l} \beta^{i+j-2} \left[ \sum_{w \in \{\Gamma^i(u) \cap \Gamma^j(v)\}} \frac{1}{\log|\Gamma(w)|} \right] \tag{7}$$

where $\beta$ is the damping factor that assigns the weights in such a way that the nodes at a distance greater than 1 get lower weights, and $l$ is the length of the longest path that is considered in computation of $PSI(u, v)$. □

The proposed method combines the advantages of Adamic Adar and LP index. It considers local paths with wider horizon than AA (AA considers only immediate neighbours of the two nodes). However the degree of each node is also considered (LP index only counts local paths and ignores degree). In Section 4, we have empirically demonstrated that PSI has outperformed both the AA and LP on numerous datasets. Note that, since the shortest distance between a pair of reachable nodes in a network is always unique, therefore, each node is considered exactly once in computation of the proposed similarity index. This led us to the following equivalent definition of $PSI$:

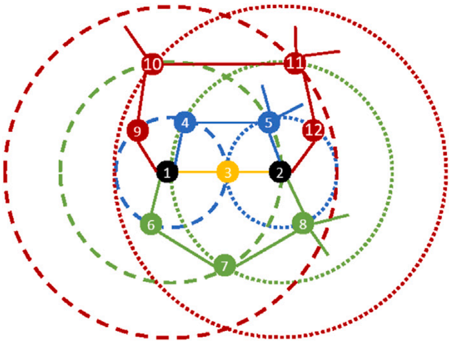


**Fig. 1.** Proposed Similarity Index.

**Definition 2.** The similarity score, $PSI(u, v)$, between two nodes $u$ and $v$ is defined as:

$$PSI(u, v) = \sum_{\substack{w \in V \setminus \{u,v\} \\ 1<d=\delta(u,w)+\delta(v,w)<l}} \frac{\beta^{d-2}}{\log |\Gamma(w)|} \tag{8}$$

where $\delta(u, v)$ represents the shortest distance between nodes $u$ and $v$. □

### 3.3. Algorithm

We now explain how the proposed similarity index is computed. For $l = 2$, the PSI degenerates to AA (LP index degenerates to CN when used with $l = 2$). So we assume $l > 2$, in which case the value of $PSI(u, v)$ is computed as follows. Assume $u$ and $v$ are not connected, first the common neighbours of the nodes $u$ and $v$ are determined and their score is computed by taking the inverse of log of their degrees. This score is added to the similarity index $PSI(u, v)$. Next we consider those nodes that are directly connected to one of the two nodes $u$ and $v$, but whose shortest path from the other node is exactly two. Their score is computed in a similar way, i.e., by taking the inverse of log of its degree. However, they are assigned less weights in computation of prediction score. This is done by multiplying their score with $\beta < 1$ and the value is then added to $PSI(u, v)$. This process continuous until the score of all nodes whose distance from nodes $u$ and $v$ is less than or equal to $l$ is computed. Note that the maximum value of $l$ could be $2d$, where $d$ is the diameter of the network which is defined as the longest of the shortest paths between all pairs of nodes in the network. To get a better understanding of the process, consider the graph of Fig. 1.

Suppose we have to estimate the likelihood of existence of a link between node 1 and node 2 (black nodes). The only node that is directly connected to both node 1 and node 2 is node 3 which has no direct connections with any other nodes. Therefore, node 3 will have highest contribution in similarity score. Next the node 4 and node 5 will have the second highest contribution to similarity score, since both nodes lie on a path of length 3 (blue path) from node 1 to node 2. This is done by multiplying their score with $\beta < 1$. Similarly, nodes lying on path of length 4 (green path) will have less contribution to nodes with same degree lying on path of length 3 but will have higher contribution compared to the nodes with the same degree lying on path of length 5 (red path). Finally, the contribution of a node in $s_{uv}$ also depends upon the number of its neighbours. So, for example, nodes 4, 7 and 10 will have higher contribution to similarity score compared to nodes 5, 8 and 11 respectively. The procedure to compute similarity score using proposed framework is outlined in Algorithm 1.

**Algorithm 1:** Link Prediction.

```
Input : Network $G^T = (V, E^T)$, with adjacency matrix $A$, the value of $l$, and a damping parameter $\beta$.
Output: A matrix of similarity scores S.
$s \leftarrow 0$                                        ▷ Initialise similarity matrix to 0
$\delta \leftarrow$ AllPairShortestPaths($G^T$)          ▷ Compute shortest paths
for $(u, v) \in E \setminus E^T$ do
    for $w \in V \setminus \{u, v\}$ do
        $d = \delta(u, w) + \delta(v, w)$
        if $d \leq l$ then
            $PSI(u, v) \leftarrow PSI(u, v) + \beta^{d-2}(\log |w|)^{-1}$     ▷ Update similarity score
        end if
    end for
end for
```

We briefly describe all the steps performed in Algorithm 1.[1] This algorithm accepts as input the adjacency matrix of the network $G^T$ which is obtained by removing the links in $E^P$. It also accepts the damping parameter $\beta$ and the $l$ (which is the length of the largest path to be considered). In the experiments section we have discussed how these values are chosen. In the procedure, we first compute the shortest paths between all pairs of nodes in the network. The for loop (Line 5 to 12) iterates through all pairs of nodes that are not connected and can possibly be linked in future. The similarity score is estimated by considering all nodes in the network, whose sum of shortest path from the two node is less than $l$. The similarity score is updated in line 11 according to Eq. (8).

### 3.4. Time analysis

We next analyse the running time of Algorithm 1. Line 3 in the algorithm requires $O(|V|^2)$ steps. Floyd–Warshall algorithm can be used to compute the shortest paths between all pairs of nodes of the network. The running time of Floyd–Warshall algorithm is bounded by $\Theta(|V|^3)$. The for loop (line 5 to 12) can be computed in $O(|V \parallel E|)$ in worst case. Hence the total running time of the algorithm is bounded by $\Theta(|V|^3)$. The running time can be accelerated for the case when $|E| \ll |V|^2$. This can be done by repeatedly running Dijkstra's algorithm for each $v \in |V|$. Dijkstra's algorithm can achieve a worst case running time of $O\,(|E| + |V| \log |V|)$ when used with a min-priority queue implemented by a Fibonacci heap. Therefore the worst case running time of the proposed algorithm for this case is given by $O\left(|V \parallel E| + |V|^2 \log |V|\right)$. The execution can be accelerated, if we choose a smaller value of $l$, but this can result in reduced performance. Note that the worst case running time of Katz index is also $O(|V|^3)$. However, in practice, Katz index runs faster than Algorithm 1 because it requires one matrix inversion operation which can be efficiently computed and requires less time compared to finding all pairs of shortest paths in the network.

## 4. Experiments and discussion

In this section we present the experiment evaluation results of the proposed method on real-world datasets. We also compare our methods with state-of-the-art prediction algorithms presented in Section 2. Note that we consider two different versions of the proposed similarity index, i.e., PSI (global measure) and PLSI (local measure). PSI considers all reachable nodes in the graph and is computationally expensive. For PLSI, we have set $l = 3$. PLSI can be computed efficiently but may have reduced performance when compared to PSI.

### 4.1. Datasets

We have used twelve different publicly available datasets, most of which can be downloaded from KONECT [28]. A brief introduction of each of these datasets is given below, while their topological properties are summarised in Table 1.

**Karate** [29]: This is a well known dataset, also known as Zachary karate club, which was collected in 1977 from members of a university karate club. Nodes represents members of the club while edges represent a tie between two members.

**US Roads** [3]: This road network consists of 49 nodes where 48 nodes represent 48 contiguous states of United States and one node represents District Colombia. A link shows there exists at least one drivable road between the two nodes.

**Dolphins** [30]: A social networks of dolphins, where nodes represent dolphins and links represent frequent association between dolphins.

**Train bombing** [31]: This is an undirected network of 64 suspected terrorists who were believed to be involved in the train bombing of Madrid on March 11, 2004. Nodes represent terrorist and a link between two nodes is established, if the two terrorists are friends or have co-participated in training camps.

**Caenorhabditis elegans (neurons)** [5]: This dataset consists of a total of 279 neurons. This dataset has originally 2990 connections, which included 1584 unidirectional and 1406 bidirectional connections. The unidirectional links are replaced with bidirectional links which resulted in 2287 bidirectional links.

**E.Coli** [32]: This is the protein–protein interaction network of Escherichia coli. The edges are directed but the direction was ignored in our experiments. The original network has 424 nodes and 519 connections. In our experiment we have used the largest connected component with 329 nodes and 456 links.

**Network Science** [33]: This dataset contains a co-authorship network of scientists working on network theory and experiments. The dataset originally consists of 1461 nodes. Here we have considered the largest connected component that consists fo 379 nodes.

**Infectious** [34]: This network describes the face-to-face behaviour of individuals during the exhibition, infectious: stay away, in 2009 at the Science Gallery in Dublin. Here nodes represent people and links represent face-to-face contacts that were active for at least 20 s.

**Caenorhabditis elegans (metabolic)** [35]: This is the metabolic network of the roundworm Caenorhabditis elegans. Nodes are metabolites (e.g., proteins), and links are interactions between them. All the links are undirected. Multiple links and loops may exist, which are ignored in this work.

[1] Code to compute PSI is available at https://github.com/azizfurqan/PSI.

**Table 1**
Topological properties of the networks used in experiments. $|V|$ and $|E|$ are the number of nodes and links respectively. $CC$ is the clustering coefficient. $\langle k\rangle$ and $\langle d\rangle$ are average degree and average path length. Finally $\rho$ denotes the density of the network while $H$ is the heterogeneity defined as $H = \frac{\langle k^2\rangle}{\langle k\rangle^2}$.

| Datasets | $\lvert V\rvert$ | $\lvert E\rvert$ | $CC$ | $\langle k\rangle$ | $\langle d\rangle$ | $\rho$ | $H$ |
|---|---|---|---|---|---|---|---|
| Karate | 34 | 78 | 0.588 | 4.588 | 1.204 | 0.139 | 7.769 |
| US Roads | 49 | 107 | 0.507 | 4.367 | 2.082 | 0.091 | 4.935 |
| Dolphin | 62 | 159 | 0.303 | 5.129 | 1.678 | 0.084 | 6.805 |
| Train Bombing | 64 | 243 | 0.711 | 7.594 | 1.345 | 0.121 | 12.597 |
| Neurons | 279 | 2287 | 0.337 | 16.394 | 1.218 | 0.059 | 25.916 |
| E.Coli | 329 | 456 | 0.222 | 2.772 | 2.421 | 0.008 | 12.314 |
| Netscience | 379 | 914 | 0.798 | 4.823 | 3.021 | 0.013 | 8.021 |
| Infectious | 410 | 17 298 | 0.467 | 84.38 | 1.815 | 0.206 | 2.992 |
| Metabolic | 453 | 4596 | 0.782 | 20.291 | 1.332 | 0.045 | 17.903 |
| US Air | 500 | 2980 | 0.726 | 11.92 | 1.496 | 0.024 | 53.785 |
| Email | 1133 | 5451 | 0.254 | 9.622 | 1.803 | 0.009 | 18.688 |
| Yeast | 2375 | 11 693 | 0.388 | 9.847 | 2.548 | 0.004 | 34.223 |

**US Air** [36,37]: This is the network of travel connections among the 500 airports in US. Here nodes represent airports and two nodes are connected if there is a direct flight between the corresponding airports.
**Email** [38]: This is the email communication network between individuals at the University Rovira i Virgili in Tarragona in the south of Catalonia in Spain. Nodes represent users while link represent that at least one of the two connected users have sent email to the other user. The direction of the email and the number of emails are ignored.
**Yeast** [39]: This is a protein–protein interaction network whose nodes are the proteins of yeast and links represent interactions between proteins.

### *4.2. Evaluation metric*

In order to evaluate the performance of the proposed method and compare it with alternate methods, we use the performance metric *AUC* [22] which is commonly used in literature. We consider an undirected network $G = (V, E)$, where multiple edges and self connections are not allowed. Let $U$ be the set of all possible $\frac{|V|(|V|-1)}{2}$ links and let $E'$ be the set of nonexistent links. Clearly $E' = U \setminus E$. The set of observed links, $E$, is randomly divided into two disjoint sets, i.e., training set $E^T$ used for training purpose and the probe set $E^P$ used for test purpose. In other words, the information in $E^T$ is used to predict missing links while the information in $E^P$ is used to evaluate the performance of the prediction algorithm. Clearly, the two sets form a partition of the set $E$, i.e., $E = E^T \cup E^P$, and $E^T \cap E^P = \Phi$. To estimate the accuracy of the prediction algorithm, we use a standard metric, AUC, area under the receiver operating characteristic (ROC). In our case, this property can be interpreted as the probability that a randomly chosen missing link in $E^P$ gets higher score than a randomly chosen nonexistent link in $E'$. If among $n$ independent comparisons, $n'$ is the number of times the missing link has higher score and $n''$ is the number of times a missing link and a nonexistent link having the same score, then the accuracy is defined as

$$AUC = \frac{n' + 0.5n''}{n}$$

Note that If all the link scores are randomly generated according to an independent identical distribution, then the accuracy should be about 0.5. Therefore, a value greater than 0.5 indicates how well the prediction algorithm performs when compared to pure chance.

### *4.3. Experimental results*

As discussed earlier, in our experiments, we have randomly divided the set of links $E$ of a network into two sets, i.e., a training set $E^T$ and a probe set $E^P$. The training set contains 90% of the links while the remaining 10% of the links were used for testing purpose. The same sets $E^T$ and $E^P$ were used to evaluate the performances of both the proposed and the alternate methods. The value of the parameter, $\beta$, was set to 0.005 for PSI, PLSI, Katz Index, and LP. To provide a fair comparison between PSI and LP, both measures were used with $l = 3$. The accuracies of all the 12 datasets are reported in Table 2. The experiments were executed 100 times with independent random splitting of $E$ into $E^P$ and $E^T$. The table reports the average accuracies of all the 100 experiments.

The cells highlighted with grey colour in Table 2 show best performance (or close to best performance) while the cells highlighted with light grey colour show the second best performance (or close to second best performance). These results demonstrate that the proposed algorithm can outperform the state-of-the-art algorithms. It is clear from the table that PSI always gives superior performance when compared to Katz index, and the difference is significant for some of the datasets like Karate network, US Roads network and metabolic network. Note that the proposed method gives better performance when used with $l = 3$. It is also worth noting that proposed index always gives superior performance when

**Table 2**
The prediction accuracy of the proposed (1st column) and the alternate methods, measured by AUC. Each value is obtained by averaging over 100 executions of experiments with independently random divisions of training set and probe set.

| Datasets | PSI | PLSI | Katz | LP | CN | AA | PA | JC |
|---|---|---|---|---|---|---|---|---|
| **Karate** | 0.7843 | 0.7711 | 0.7628 | 0.7662 | 0.7028 | 0.7366 | 0.7318 | 0.6138 |
| **US Roads** | 0.9250 | 0.9184 | 0.8944 | 0.8955 | 0.8975 | 0.9042 | 0.4406 | 0.9187 |
| **Dolphins** | 0.8425 | 0.8213 | 0.8384 | 0.8269 | 0.7863 | 0.7902 | 0.6674 | 0.7850 |
| **Train Bombing** | 0.9485 | 0.9444 | 0.9309 | 0.9312 | 0.9322 | 0.9438 | 0.7985 | 0.9298 |
| **Neurons** | 0.8828 | 0.8781 | 0.8660 | 0.8670 | 0.8593 | 0.8747 | 0.7238 | 0.8305 |
| **E.Coli** | 0.8949 | 0.8625 | 0.8846 | 0.8663 | 0.6213 | 0.6281 | 0.8788 | 0.6120 |
| **Netscience** | 0.9929 | 0.9911 | 0.9861 | 0.9860 | 0.9811 | 0.9849 | 0.6613 | 0.9782 |
| **Infectious** | 0.9554 | 0.9481 | 0.9483 | 0.9477 | 0.9123 | 0.9151 | 0.6969 | 0.9149 |
| **Metabolic** | 0.9277 | 0.9260 | 0.8995 | 0.9003 | 0.8671 | 0.9055 | 0.8480 | 0.7517 |
| **US Air** | 0.9665 | 0.9672 | 0.9513 | 0.9522 | 0.9522 | 0.9621 | 0.9196 | 0.9111 |
| **Email** | 0.9366 | 0.9260 | 0.9336 | 0.9254 | 0.8654 | 0.8678 | 0.8194 | 0.8624 |
| **Yeast** | 0.9731 | 0.9700 | 0.9706 | 0.9695 | 0.9142 | 0.9149 | 0.8648 | 0.9132 |

compared to Adamic-Adar, which only depends upon the information of the immediate neighbours of the nodes. This clearly shows that by incorporating information about the reachable nodes, the classification accuracy of the prediction algorithm significantly increases.

In order to further investigate the performance of the proposed framework, we evaluate its classification accuracy with different partitioning sizes of training and probe sets. For this purpose we choose different size of the probe sets as 20%, 30%, 40%, 50% respectively, and evaluate the performance of the proposed method and other methods on all the datasets used in this paper. For each split, we have computed the accuracies of all the methods on all the datasets. To visualise and compare those results, we have plotted the average accuracies of 100 independent runs of each experiment in Fig. 2.

It is evident from the visualisation results of Fig. 2 that PSI gives higher accuracy and outperforms other local and global link prediction methods on most of the datasets, when evaluated with different sizes of training sets and probe sets. Another important observation that can be made from the results of the figure is that the performances of local similarity measures generally decrease with decrease in the size of training set. This decrease is high for local indices such as AA and CN, intermediate for quasi local indices such as LP an PLSI and low for global indices such as Katz and PSI. In particular, the difference between the global similarity indices and local similarity indices becomes very high when about 50% of the links are deleted.

Note that the value of parameter $\beta$ can play an important role in the prediction algorithms PSI, Katz index, PLSI and LP. In practice, $\beta$ is set to a very small value so that the nearest neighbours get higher weights in prediction. For example, in our experiments we have chosen $\beta = 0.005$, which is common in literature [19]. In our last experiment, we have studied the effect of choosing a different value of $\beta$. This is demonstrated in Fig. 3, which shows that a value of $\beta$, greater than 0.005 will generally result in reduced performance. We have compared the results of all the four methods that require us to adjust the value of $\beta$ before prediction. Note that for a larger value of $\beta$, Katz index may not converge. For comparison purposes, the index was approximated for larger values of $\beta$ by considering first six terms. Finally, it is also important to note that as long as the value of $\beta$ is less than 0.01, PLSI generally performs well when compared to alternate path based methods.

### 4.4. Synthetic data

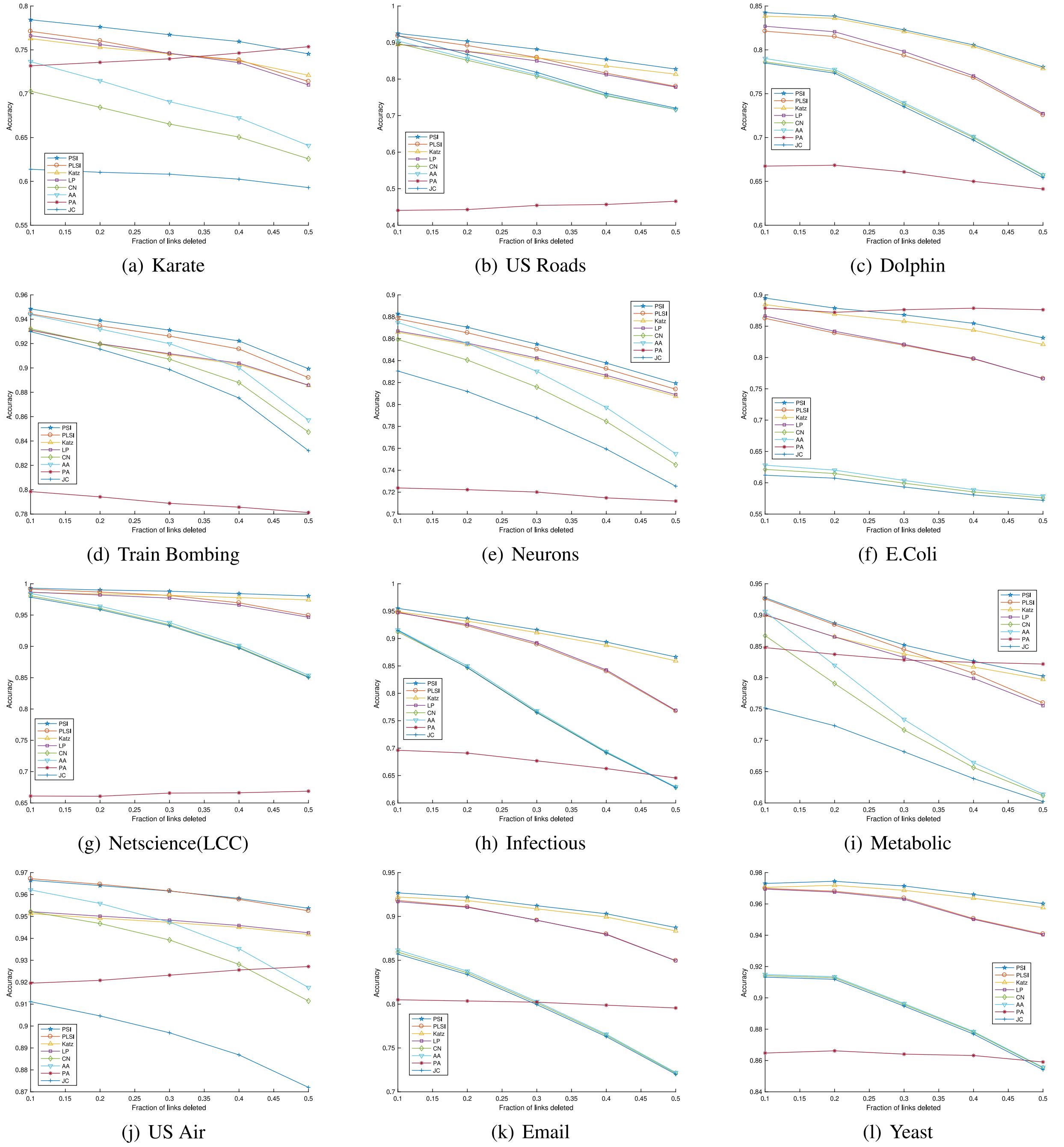
In this section, we compare the performance of the proposed similarity index on networks that are generated according to different network models. In our experiments, we have considered the following three network models.

**Watts and Strogatz model (WS)** [40]: A *WS* graph $G(n, k, p)$ is constructed in the following way. First construct a regular ring lattice on the node. Then for a graph with $n$ vertices, each vertex is connected to $k$ nearest vertices, with $k/2$ connecting on each side. Then for every vertex we take each edge and rewire it with probability $p$.

**Barabási–Albert model (BA)** [41]: A *BA* graph $G(n, n_0, m)$ is constructed from an initial fully connected graph with $n_0$ vertices. New vertices are added to the graph one at a time. Each new vertex is connected to $m$ previous vertices with a probability that is proportional to the number of pre-existing links for each node.

**Delaunay triangulations (DT)** [42]: A *DT* for a set of points in a Euclidean space is a triangulation, DT(P), such that no point in P is inside the circumcircle of any triangle in DT(P). This is a spatial graph where nodes represent the points in the space and the edges represent adjacencies of the Voronoi regions containing the points.

For WS and BA we generate graphs with 200 nodes. The other parameters were chosen in such a way that the two graphs have approximately the same number of edges. The DT was constructed by generating 166 random points in a Euclidean space. Table 3 show the other topological properties of the networks.



(a) Karate

(b) US Roads

(c) Dolphin

(d) Train Bombing

(e) Neurons

(f) E.Coli

(g) Netscience(LCC)

(h) Infectious

(i) Metabolic

(j) US Air

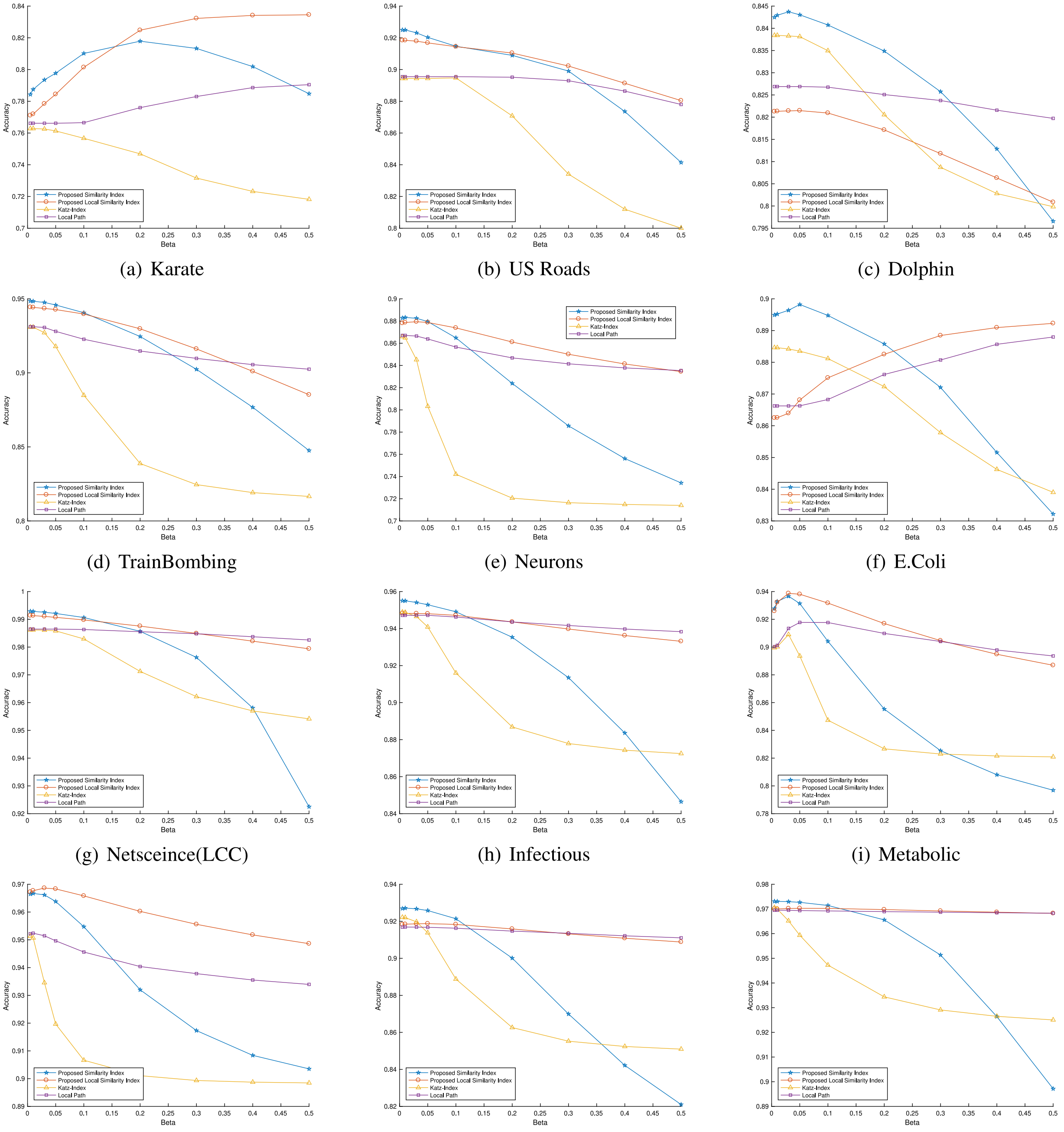
(k) Email

(l) Yeast

**Fig. 2.** The prediction accuracy of the proposed and the alternate methods, measured by AUC, with different split of training and probe set. As with the previous experiments, each value is obtained by averaging over 100 executions of experiments with independently random divisions of training set and probe set.

As with the real-world dataset, we have randomly divided the set of links $E$ of a network into a training set $E^T$ (90%) and a prob set $E^P$ (10%). All the other parameters were assigned the same values as in the previous experiments. The experiments were executed 100 times with independent random splitting of $E$ into $E^P$ and $E^T$. Table 4 reports the average accuracies of all the 100 experiments.

We have also investigated the performances of the proposed methods and the alternate methods with different sizes of training and probe sets. The resultant accuracies with different sizes of probe sets are shown in Fig. 4.



(a) Karate (b) US Roads (c) Dolphin

(d) TrainBombing (e) Neurons (f) E.Coli

(g) Netsceince(LCC) (h) Infectious (i) Metabolic

(j) US Air (k) Email (l) Yeast

**Fig. 3.** The prediction accuracies of PSI, LP, and Katz, measured by AUC, with different values of the parameter $\beta$. Again, the experiment was performed 100 times and the average values are reported.

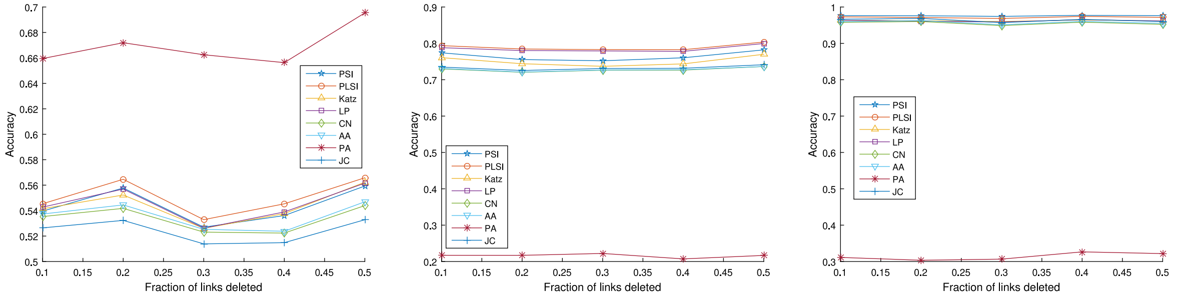
There are a number of important observations that can be made from the results of Table 4 and Fig. 4. Firstly, as expected, the PA index outperforms all the other indices on BA network. This is obvious as one of the key assumption of the BA model is preferential attachment. However, for the WS and DT, the PA index gives poor performance. Secondly, the quasi local indices (LP and PLSI) give superior performance on the WS networks when compared to both local and global indices. In particular, the proposed PLSI outperforms all the other indices on WS networks. This is probably due to the small-world property of WS network. Finally, the proposed similarity indices outperforms all the other indices on DT.

**Table 3**
Topological properties of the networks generated according to different network models.

| Datasets | $\|V\|$ | $\|E\|$ | $CC$ | $\langle k\rangle$ | $\langle d\rangle$ | $\rho$ | $H$ |
|---|---|---|---|---|---|---|---|
| WS | 200 | 398 | 0.254 | 3.98 | 2.446 | 0.02 | 4.158 |
| BA | 200 | 396 | 0.102 | 3.96 | 1.618 | 0.02 | 9.856 |
| DT | 166 | 468 | 0.446 | 5.639 | 2.998 | 0.034 | 5.902 |

**Table 4**
The prediction accuracy of the proposed (1st column) and the alternate methods, measured by AUC. Each value is obtained by averaging over 100 executions of experiments with independently random divisions of training set and probe set.

| Datasets | PSI | PLSI | Katz | LP | CN | AA | PA | JC |
|---|---|---|---|---|---|---|---|---|
| **BA** | 0.5395 | 0.5451 | 0.5419 | 0.5431 | 0.5357 | 0.5375 | 0.6599 | 0.5262 |
| **WS** | 0.7701 | 0.7912 | 0.7614 | 0.7871 | 0.7281 | 0.7285 | 0.2 | 0.7311 |
| **DT** | 0.9772 | 0.9731 | 0.9634 | 0.9635 | 0.9595 | 0.9602 | 0.3151 | 0.9644 |



(a) BA (b) WS (c) DT

**Fig. 4.** The prediction accuracies of PSI, LP, and Katz, measured by AUC, with different values of the $\beta$. The experiments were performed 100 times and the average values are reported.

These results show that the proposed similarity indices can predict links with higher accuracy. Note that we have also applied our method on graphs that are generated according to **Erdős–R(́e)nyi (ER)** [43] network model. An ER network is a graph network, where every link has an equal probability of being present. Since such networks are generated randomly, it is not possible to predict which links are missing in the network as every non-existent and missing links are equally probable to be present. We have therefore not reported the accuracies of such methods in this paper.

## 5. Conclusion

Link prediction is one of the most important and challenging areas in complex network analysis. The goal of a link prediction algorithm is to estimate the likelihood of the missing links based on the existing links in a network. In this paper, we have designed a novel link prediction method that significantly improves the accuracy compared to existing state-of-the-art methods. The proposed method is based on path-based index that incorporates the information about the nodes on the paths to predict the likelihood of existence of a link between two nodes in a network. Unlike the local path index that is based on the frequency of local paths to predict missing links, the proposed approach also incorporates the information about the nodes on these paths. The experiments presented in this paper on both real-world and synthetic datasets demonstrate that the proposed method can give higher accuracy measured by AUC, compared to other state-of-the-art methods. There are several directions in which the work reported here can be extended. From a theoretical perspective, it would be interesting to combine other local and path-based indices that can be used to predict links with higher accuracy. From an application perspective, it would be interesting to explore the applications of the proposed link prediction method in other domains.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] L. Guo, B. Zhang, Mining structural influence to analyze relationships in social network, Physica A 523 (2019) 301–309.
[2] Y. Hadas, G. Gnecco, M. Sanguineti, An approach to transportation network analysis via transferable utility games, Transp. Res. B 105 (2017) 120–143.
[3] D.E. Knuth, The Art of Computer Programming, Volume 4, Fascicle 0: Introduction to Combinatorial Algorithms and Boolean Functions (Art of Computer Programming), first ed., Addison-Wesley Professional, 2008.
[4] L. Antiqueira, O.N. Oliveira, L. da Fontoura Costa, M. das Graças Volpe Nunes, A complex network approach to text summarization, Inform. Sci. 179 (5) (2009) 584–599.
[5] K. Jinseop, K. Marcus, From caenorhabditis elegans to the human connectome: A specific modular organization increases metabolic, functional and developmental efficiency, Philos. Trans. R. Soc. B 369 (2014).
[6] Z. Wang, J. Liao, Q. Cao, H. Qi, Z. Wang, Friendbook: A semantic-based friend recommendation system for social networks, IEEE Trans. Mob. Comput. 14 (3) (2015) 538–551.
[7] I.A. Kovács, K. Luck, K. Spirohn, Y. Wang, C. Pollis, S. Schlabach, W. Bian, D.-K. Kim, N. Kishore, T. Hao, M.A. Calderwood, M. Vidal, A.-L. Barabási, Network-based prediction of protein interactions, Nature Commun. (2019).
[8] L.F. Costa, M. Kaiser, C.C. Hilgetag, Predicting the connectivity of primate cortical networks from topological and spatial node properties, BMC Syst. Biol. (2007) 16.
[9] D. Wei, T. Zhou, G. Cimini, P. Wu, W. Liu, Y.-C. Zhang, Effective mechanism for social recommendation of news, Physica A 390 (11) (2011) 2117–2126.
[10] M. Pavlov, R. Ichise, Finding experts by link prediction in co-authorship networks, in: Proceedings of the 2Nd International Conference on Finding Experts on the Web with Semantics - Volume 290, in: FEWS'07, 2007, pp. 42–55.
[11] D. Lin, An information-theoretic definition of similarity, in: Proceedings of the 15th International Conference on Machine Learning, Morgan Kaufmann, 1998, pp. 296–304.
[12] F. cois Lorrain, H.C. White, Structural equivalence of individuals in social networks, J. Math. Sociol. 1 (1) (1971) 49–80.
[13] L. A.A.damic, E. Adar, Friends and neighbors on the web, Social Networks 25 (3) (2003) 211–230.
[14] P. Jaccard, Étude comparative de la distribution florale dans une portion des Alpes et des Jura, Bull. Soc. Vaudoise Sci. Nat. 37 (1901) 547–579.
[15] T. Zhou, L. Lü, Y.-C. Zhang, Predicting missing links via local information, Eur. Phys. J. B 71 (4) (2009) 623–630.
[16] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.
[17] L. Katz, A new status index derived from sociometric analysis, Psychometrika 18 (1) (1953) 39–43.
[18] F. Fouss, A. Pirotte, J. Renders, M. Saerens, Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation, IEEE Trans. Knowl. Data Eng. 19 (3) (2007) 355–369.
[19] L. Lü, C.-H. Jin, T. Zhou, Similarity index based on local paths for link prediction of complex networks, Phys. Rev. E 80 (2009) 046122.
[20] C. Yu, X. Zhao, L. An, X. Lin, Similarity-based link prediction in social networks: A path and node combined approach, J. Inf. Sci. 43 (5) (2017) 683–695.
[21] P. Zhang, D. Qiu, A. Zeng, J. Xiao, A comprehensive comparison of network similarities for link prediction and spurious link elimination, Physica A 500 (2018) 97–105.
[22] L. Lü, T. Zhou, Link prediction in complex networks: A survey, Physica A 390 (6) (2011) 1150–1170.
[23] H. Gao, J. Huang, Q. Cheng, H. Sun, B. Wang, H. Li, Link prediction based on linear dynamical response, Physica A 527 (2019) 121397.
[24] B. Zhu, Y. Xia, An information-theoretic model for link prediction in complex networks, Sci. Rep. 5 (2015) 13707.
[25] A. Clauset, C. Moore, M.E.J. Newman, Hierarchical structure and the prediction of missing links in networks, Nature 453 (2008).
[26] R. Pech, D. Hao, L. Pan, H. Cheng, T. Zhou, Link prediction via matrix completion, Europhys. Lett. 117 (3) (2017) 38002.
[27] M.E.J. Newman, Clustering and preferential attachment in growing networks, Phys. Rev. E 64 (2001) 025102.
[28] J. Kunegis, KONECT: The koblenz network collection, in: Proceedings of the 22nd International Conference on World Wide Web, in: WWW '13 Companion, 2013, pp. 1343–1350.
[29] W.W. Zachary, An information flow model for conflict and fission in small groups, J. Anthropol. Res. 33 (4) (1977) 452–473.
[30] R.A. Rossi, N.K. Ahmed, The Network Data Repository with Interactive Graph Analytics and Visualization, in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, 2015, 4292–4293.
[31] B. Hayes, Connecting the dots. can the tools of graph theory and social-network studies unravel the next big plot? Am. Sci. 94 (5) (2006) 400–404.
[32] S. Shen-Orr, R. Milo, S. Mangan, U. Alon, Network motifs in the transcriptional regulation network of escherichia coli, Nature Genet. 31 (2002) 64–68.
[33] M.E.J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (2006) 036104.
[34] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, W.V. den Broeck, What's in a crowd? Analysis of face-to-face behavioral networks, J. Theoret. Biol. 271 (1) (2011) 166–180.
[35] J. Duch, A. Arenas, Community detection in complex networks using extremal optimization, Phys. Rev. E 72 (2005) 027104.
[36] V. Colizza, R. Pastor-Satorras, A. Vespignani, Reaction - diffusion processes and metapopulation models in heterogeneous networks, Nat. Phys. 3 (2007) 027104.
[37] G. Caldarelli, Complex networks: Principles, methods and applications by vito latora, vincenzo nicosia and giovanni russo, J. Complex Netw. 6 (5) (2018) 830.
[38] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, A. Arenas, Self-similar community structure in a network of human interactions, Phys. Rev. E 68 (2003) 065103.
[39] C. von Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, P. Bork, Comparative assessment of large-scale data sets of protein–protein interactions, Nature 417 (2002) 399–403.
[40] D.J. Watts, S. Strogatz, Collective dynamics of 'small-world' networks, Nature (1998) 440–442.
[41] A. Barabási, R. Albert, Emergence of scaling in random networks, Science (1999) 509–512.
[42] B. Delaunay, Sur la sphére vide, Izvestia Akademii Nauk SSSR, Otdelenie Matematicheskikh i Estestvennykh Nauk (1934) 793–800.
[43] R. Erdõs, A.: On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci. (1960) 17–61.