Journal of Intelligent Data Analysis (IDA) 0 (0) 1
IOS Press

Enhancing Link Prediction Efficiency with Shortest Path and Structural Attributes

Muhammad Wasim [a], Feras Al-Obeidat [b], Adnan Amin [c], Haji Gul [c], and Fernando Moreira [d,*]
[a] *Department of Computer Science, City University of Science and Information Technology, Pakistan.*
E-mail: Muhammadwasim443@gmail.com
[b] *College of Technological Innovation, Zayed University, Abu Dhabi.*
E-mail: feras.Al-Obeidat@zu.ac.ae
[c] *Center for Excellence in Information Technology, Institute of Management Sciences, Pakistan.*
E-mails: adnan.amin@imsciences.edu.pk, hajigul1993@gmail.com
[d] *REMIT, IJP, Universidade Portucalense Porto, Portugal IEETA, Universidade de Aveiro, Aveiro, Portugal., Portugal*
E-mail: fmoreira@uportu.pt

Abstract. Link prediction is one of the most essential and crucial tasks in complex network research since it seeks to forecast missing links in a network based on current ones. This problem has applications in a variety of scientific disciplines, including social network research, recommendation systems, and biological networks. In previous work, link prediction has been solved through different methods such as path, social theory, topology, and similarity-based. The main issue is that path-based methods ignore topological features, while structure-based methods also fail to combine the path and structured-based features. As a result, a new technique based on the shortest path and topological features' has been developed. The method uses both local and global similarity indices to measure the similarity. Extensive experiments on real-world datasets from a variety of domains are utilized to empirically test and compare the proposed framework to many state-of-the-art prediction techniques. Over 100 iterations, the collected data showed that the proposed method improved on the other methods in terms of accuracy. SI and AA, among the existing state-of-the-art algorithms, fared best with an AUC value of 82%, while the proposed method has an AUC value of 84%.

Keywords: Link Prediction, Complex Networks, Global Features, Local Features

1. Introduction

Many real-world biological, animal, human, terrorist, and social phenomena can be represented by complex networks [1, 2]. The properties and characteristics of real-world systems can be studied in complex networks whose nodes express the entities while links describe the interactions or relationships between the nodes or entities. Social interaction or relationships can be identified in networks such as Twitter and Facebook, where the node represents a person and the link indicates their relationship [3–5]. similarly, a biological network where each protein describes a node and the interaction between two nodes is as a link [6, 7]. Complex networks are networks with a variety of features and different patterns of connection among their elements that do not follow any particular order. The analysis of complex networks helps to realize and model the interaction between the elements of these networks. Complex

*Corresponding author. E-mail: fmoreira@uportu.pt.

networks are animated objects that develop quickly over time with the increase of nodes and edges, which makes their study and analysis difficult. Examining the evolution and dynamics of the networks is a complicated problem due to the huge number of parameters.
However, when the structure of complex network changes, some links in the graph are lost and others are created. We are trying to conceive of the idea of link prediction, which allows us to predict missing or non-existent links that may exist in the future [8, 9]. The link prediction problem is divided into two categories. One category is missing link prediction, while the other is future link prediction. Both of these problems have the same solution, solved through different methods [10].
A variety of link prediction algorithms have been introduced and applied consistently in recent years. The majority of these solutions are based on the concept of node similarity. The key features of nodes may be used to determine node similarity [11, 12]. Jaccard coefficient [13], Common neighbors [14], Resource allocation [15], Adamic Adar [16], and preferential attachment [17] are some of the most frequently applied similarity-based techniques. The quasi-local/global metric is another common approach. These techniques are often based on information about local paths. Among these are the Katz Index [18], local paths [19], and commuting time [20]. In a related paper, Yu et al. [21] introduced a hybrid technique for estimating node similarity that incorporates the resource allocation index and the Katz index. Several survey publications, such as [9], and [22], compare many local and global link prediction methods.
Over the last two decades, a variety of other techniques have been presented in addition to similarity-based prediction techniques. The majority of these strategies are maximum likelihood or probabilistic in nature. Gao et al. [23], for example, have recently presented a linear dynamical response-based similarity measure between nodes and designed a way to compute it effectively. Zhu et al. [24] investigated the importance of network topology in identifying missing links and provided an information-theoretic model.
The major problem is that path-based approaches neglect topological characteristics, whereas structure-based methods fail to merge path and structured-based data. As a consequence, a new approach based on the shortest path and topological properties is presented in this study. When compared to others, the proposed technique outperforms. The following are our primary contributions.

- We proposed a method for predicting links based on Structural Attributes and the shortest path.
- We conducted research on a wide range of complex networks of varying sizes and structures, evaluating various link prediction techniques.
- We have identified which state-of-the-art link prediction method works better over numerous network data sets belonging to different domains.

The rest of the paper is structured as follows: Section 2 contains information about previous research. Section 3 summarises the findings, experimental setup, and evaluation criteria used. Section 4 contains the results and discussions. The conclusion is in Section 5.
There are a number of applications of link prediction:

1.1. Applications of Link Prediction

Apart from the role of link prediction as a fundamental question, it could be associated with many applications of real-world networks. Link prediction has an outstanding role in many fields such as e-commerce which helps in building recommendation systems [25], spam mail detection, privacy control in different social networks [26], predicting missing references in publications, expert detection, influence detection, network routing, and disease prediction.

Spam Mail Detection : Link prediction indexes can predict spam emails by monitoring traffic on different communication sites [15] based on the graph theory approach.
Social Network Privacy Control : A model uses link prediction to identify trusted users in a weighted network. Link prediction secures users or nodes from unreliable users, it shows the privacy control of users in a social network [27].
Expert Detection : Link prediction has been applied in co-authorship networks for predicting domain experts. A link prediction method was used for ranking candidates in selection for distinguished government posts. [14].
Predicting Missing References in Publication : A model is proposed for the creation of links between referenced and interlinked documents. This model is developed using node-pair graphs for documents, and a merge graphs for the references among documents. The LP model is used in this framework for predicting missing or valuable references in fresh or unused documents.
Recommender Systems : LP has been used widely in recommender systems [28, 29]. Link prediction performed very well in standard collaboration filtering algorithms when applied to provide the recommendation.
Disease Prediction : A link prediction-based method used by [30] to identify the onset of diseases using the existing health condition of a patient.

2. Literature

In this part, we provide existing state-of-the-art link prediction techniques. These approaches are also utilized for comparison. We begin by presenting some fundamental definitions and an overview of the problem.
The data representation $G(V,E)$ can be described as a graph given a set of nodes V and a set of links E having (x,y) where $(x,y) \in V$. A graph G is called a directed graph if the items in the set of E are ordered pairs. In other terms, when edges carry direction signs such as incoming and outgoing. Regardless of the graph's directionality, if $(x,y) \in E$, node y is a neighbor node to node x. The collection of neighbor nodes of node x and y can be expressed with $\Gamma(x)$ and $\Gamma(y)$.
Link prediction is the challenge of estimating the presence of a link between two nodes in a network in network theory. Identifying friendship links between users in a social network, co-authorship links in a citation network, and interactions between proteins in a biological network are all examples of link prediction. Link prediction can also include a temporal component, with the goal of estimating the connections at time $t+1$ given a snapshot of the collection of links at t. In this paper, undirected complex networks have been used to predict links. Where self-loop, link-weight, and direction are ignored.
Presently in the research field of link prediction problems, several classic and generic indexes have been formed for link prediction problems, some are using nodes, social theory, and topological information to compute the similarities of pairs of nodes, to predict the missing or new links. According to recent literature, besides the similarity-based link prediction methods, a number of different algorithms have been proposed. The majority of these algorithms are dependent on maximum likelihood and probabilistic techniques. For instance, Gao et al. [23] has recently launched a linear dynamical response based on similarity measure and created a method to compute similarity. Haji et al. [31] introduced a novel algorithm based on a double-degree equation with a network feature. Clauset et al. [32] designed a Hierarchical Structure Model that utilizes the network hierarchy to compute the possibility of a link. Zhu et

al. [24] have created an information-theoretic model making use of the topology of the network to predict missing links. To investigate the missing value in a real-world network, [33] proposed robust principal component analysis (robust PCA). In the paper, Haji et al. [34] propose an approach for link prediction in complex networks. The method combines local similarity features and matrix-forest metrics. In order to take into account the similarity of nodes based on their local neighbourhood structure, the authors updated the matrix-forest measure. Another publication [35] on hybridization of feature graphs, in which a novel method for reconstructing protein particle networks employing cutting-edge hybrid features is presented, The method combines topological traits with biochemical characteristics obtained from information about protein sequences.

Obtaining the topology of the node is significantly easier and more convenient than obtaining attribute information, the majority of the input metadata is based on topological properties. The topology-based link prediction method can be used in a variety of networks. Liben-Nowell and Kleinberg [36] investigated graph structural features and suggested a series of topology-based features [37], which are separated into neighborhood-based and path-based features. There are both direct and indirect connections. Neighborhood-based characteristics, according to Wang et al. [38], perform best. To attain a high score with minimal computational cost, the node similarity algorithms simply required information on the nearest neighbors, which is notably efficient and simple for networks with low clustering coefficients. Shang et al.[39], discovered that for a variety of time-varying networks, the direct connection method outperforms the indirect link method, and those topology-based properties are more important. Features that are specific to a neighborhood are also significantly vital. Before proceeding on to the mathematical formulations.

2.1. Local Methods

The goal of these techniques is to find missing links among nodes based on their local connectivity of the topologies. It only employs the node information of their close neighbors. The AA index, for example, gives a common neighbor with a lower degree a higher weight. The assumption is that neighbors with lower degrees are more important in predicting relationships between nodes.

Parameter Dependent : This method was developed by Zhu et al.[40], and used for better accuracy. It can also compute the similarity accurately of unpopular and popular links. Let a free parameter is λ, if the free parameter is equal to one and point five (λ =1, and λ =0.5) it degenerated to LHN and SI index, if the free parameter is equal to zero (λ =0) then it will degenerate CN, respectively as shown in equation 1.

$$PD(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{(|\Gamma(x).\Gamma(y)|)^{\lambda}} \tag{1}$$

Common Neighbor : Common neighbors assign higher similarity to a link between a source and a distinct node if there are a greater number of common nodes. The other name for CN is structural equivalence [14]. It can be calculated by the mathematical formula as shown in equation 2. CN performs efficiently therefore other methods' performance is accessed as its basis. Other mathematical notation for CN is:

$$PA(x,y) = |\Gamma(x) \cdot \Gamma(y)| \tag{2}$$

where A is the adjacency matrix of the network.

Adamic Adar : This index improved the accuracy of common neighbors by assigning more weights to less connected neighbors [16]. Suppose w is a common neighbor of x and y or the weight of a link. This similarity index can be defined in the mathematical form shown in equation 3.

$$AA(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|} \tag{3}$$

Preferential Attachment : PA is a simple concept of social networks that a person with a higher number of friends will tend to add more friends in the future as compared to a person with fewer friends [17]. The mathematical equation is 4. The computational complexity of this method is better when applied to many social network datasets.

$$PA(x,y) = |\Gamma(x) \cdot \Gamma(y)| \tag{4}$$

Sorensen Index : This is also a neighbor-based link prediction index, beside with the size of the common neighbor, it also signalizes lower degrees of vertices have great link likelihood [41]. Mathematically represented in equation 5.

$$SI(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) + \Gamma(y)|} \tag{5}$$

Hub Promoted Index : This will define the topological convergence of two-node x and y [42]. The hub-promoted index values will be computed based on lower-degree nodes. HP can be defined by the following formula in equation 6. This index can also be analyzed by the metabolic network [42].

$$HPI(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{min(|\Gamma(x)|, |\Gamma(y)|)} \tag{6}$$

2.2. Global Methods

Global indices calculate a similarity score based on a graph's global link structure, where nodes have a path distance greater than two. To put it another way, global indices rate each link based on the entire network of topological information. Global indices are superior to local index techniques. Identify all interesting direct and indirect paths to include in the similarity score. Global similarity indices, as opposed to local ones, require topological information. Despite the fact that global indices can produce considerably more accurate predictions than local ones, they have two major drawbacks: (i) Calculating a global index takes a long time and is typically impractical for large-scale networks: (ii) global topological information is not always accessible, especially if the method is to be implemented in a decentralized network. The global methods consider the topology of the whole network only. However, global methods do not consider the degrees of the connecting nodes.

Katz : This approach assembles all routes, counts each pathway between two nodes, and allocates greater weight to shorter path[18]. The mathematical formulation for the Katz index is shown in equation refkatz. In the given formula $Path^{l}_{xy}$ computes all paths between two nodes in the dataset. l and β will be greater than 0 (l,$\beta > 0$).

$$Katz(x,y) = \sum_{l=1}^{\infty} \beta^l . |Path^l_{x,y}| = \beta A + \beta^2 A^2 + ... \tag{7}$$

Leicht Holme Nerman : This method focused on the common neighbors, the LHN index allocates strong similarity among the pairs of the node that have large numbers of common neighbors [43]. This method was proposed for the purpose to measure the nodes' similarity in real word networks. Equation 8 shows the mathematical representation of the index.

$$LHN(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x).\Gamma(y)|} \tag{8}$$

Compared to other works on link prediction [34, 44], this work proposes a novel approach by combining path-based and structure-based approaches into a single algorithm. The proposed technique incorporates both local and global similarity indexes, while previous works had only considered one or the other. This novel technique overcomes the drawbacks of previous approaches by taking into account topological details and combining path and structure-based characteristics.
The research also makes a significant contribution via its comprehensive analysis of link prediction methods. This research performed trials on a wide range of complex networks of varied sizes and structures, in comparison with earlier studies that assessed their approaches on small datasets. Social network analysis, recommendation systems, and even biological networks may all benefit from this method's more thorough and transferable findings. The efficiency of the suggested strategy is evaluated not only in comparison to other methods, but also to several state-of-the-art prediction techniques. The findings demonstrate that the suggested technique achieves greater AUC value than the best-performing algorithm in the prior experiments.

3. Proposed Work

The link prediction problem has been solved through different methods, such as local similarities, global similarities, and quasi-based link prediction. The problem in previous work was that local similarities were ignored in order to utilize global, complex network features. For example, common neighbors assign higher similarity scores to nodes that have a large number of common neighbours. Similarly, global link prediction methods that ignore using local complex network features, such as the local path link prediction method, assign higher accuracy to nodes that are close to each other, with a level of 2 and a maximum of 3. As a result, a link prediction approach that can utilize both local and global similarity indexes to increase link prediction accuracy is required.
In this paper, the author introduces a novel framework for link prediction. The proposed algorithm is based on two advanced features of a complex network graph. In the first part, it computes the similarity matrix based on the graph's shortest path, and secondly, the other similarity matrix is computed based on the more dense nodes of node x source and node y destination. The shortest-way issue is an extra-contemplative subject in the field of computer science, explicitly in the complex network graph. From the source to the destination node, an ideal shortest path is unified with the base length models. Most of the shortest path methods fall into two general classifications. The primary class is the single derivation

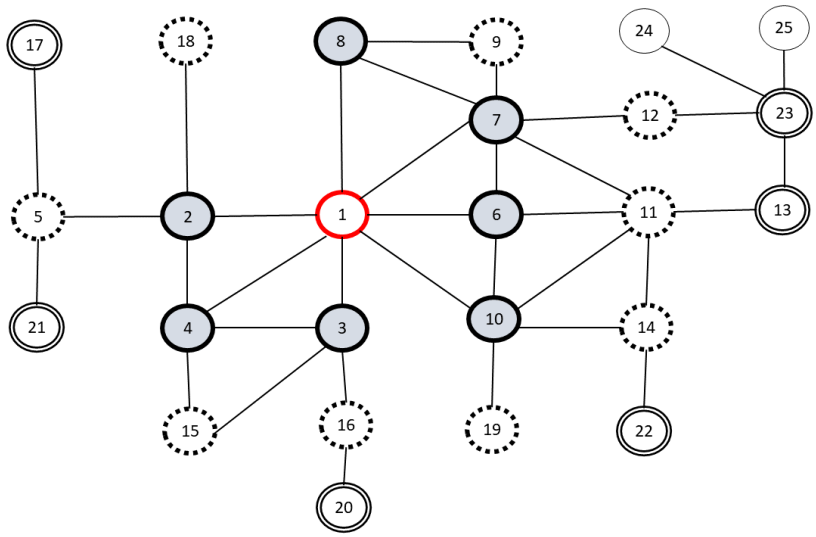

Fig. 1. Network Depiction for Proposed Algorithm

shortest path, where the goal is to track down the most limited ways from a solitary source vertex to any remaining vertices. While the second type of shortest path method is, where the goal is to track down the most limited ways between all sets of vertices in a complex network. The calculation of the briefest way can produce either precise or rough arrangements. The decision of which calculation to utilize depends upon the attributes of the complex network and the necessary application.

The proposed method computes the similarity matrix based on the shortest path and with more dense nodes between the source node and destination node. Starting from the x_i source/initial node to every destination y_j node. First of all, it computed destinations between x_i and all other destination nodes $y2, 3, 4.....n$. Then all the distances are compared to each other, and the best or shortest distance is suggested between node x_i and destination node y_j. Next, it finds the shortest distance between node x_{i+1} and y_j, x_{i+1} work as a source node and all others as destination nodes. Further, it computed the distance between $X_{i+2}, X_{i+3}, X_{i+4}.....X_n$.The shortest path can be computed in a complex network or graph G that comprises a bunch of edges E and vertices V. The network graph is characterized as $G = (V, E)$. The edges can be coordinated or undirected. If the edge exists between a pair of nodes x and y can be represented by $e = (x, y)$, Where $e \in E$ or unweighted Suppose V is a set of nodes, where the starting node can be represented by Xi and the finishing node expressed by y_j, wheres X_i and $y_j \in V$.

The second part of the proposed algorithm computes the similarity of the node based on more dense nodes. Suppose X_i is the starting node and Y_j is the ending or destination node. The shortest destine nodes with X_i are $Y_{i+2}, Y_{i+2a}, Y_{i+2b}.....Y_{i+2n}$.The more dense the nodes, the higher the similarity, and the more accurate the destination node for link prediction. Combining both of these parameters-shortest distance and higher dense node similarities—provided a higher predicted similarity matrix.

The AUC (area under the curve) has been used as an evaluation matrix in order to check which link prediction algorithm is best in terms of accuracy. Broadly speaking, AUC takes the predicted matrix as an input and suggests the best link predictor. Each dataset has 10% of its connections deleted, and the remaining 90% of the dataset is used to create a connected graph. This is an important element to remember when splitting the datasets: the remaining 90% of the dataset must remain connected. After the dataset has been segmented into training and testing, the prediction methods for link prediction are

used. It's also worth noting that modified training and testing sets must be in the form of a connected, complex network. Assume U represents the set of all possible links, or the network density, which can be calculated using a simple formula where V represents a positive value. Suppose E represents the set of non-existent edges. The mathematical form is,$\acute{E} = \frac{U}{E}$. The group of known edges E is irregularly partitioned into two disconnected groups, i.e., the set of training E^T which is being used in experiments for training purposes while the second test set E^P is used for the purpose of testing. Also, the information provided by E^T is utilized to estimate lost edges, while the information acquired by E^P is utilized to measure the functioning of the link estimation method. Clearly, two groups of sets model the separation of the E set, i.e., $E = E^T \cup E^P$, and $E^T \cap E^P = \phi$. A model metric to compute the accuracy of the prediction algorithm is AUC [10]. The evaluation criteria have been graphically viewed in Figure 2. The higher dense node in an undirected complex network can be analyzed as,

$$Adjacency\ matrix\ (adj) = A_{xy}(Graphrepresentation) \tag{9}$$

In the literature A_{xy} expressed in different forms,

$$A_{x,y}, if\ there\ an\ edge\ e = (x,y)\ , otherwise\ e \neq (x,y). \tag{10}$$

$$Density\ of\ whole\ matrix = \sum_{i}^{n} A_{x,y} \tag{11}$$

$$Specific\ node\ Density = \sum_{i}^{n} a_{xy} \tag{12}$$

Finally, in Formula (14) the highest dense node has been evaluated. While in the first part, the shortest distance is evaluated between source node x and destination node y.

$$Distance(x,y) = Shortestlength(length(x,y)) \tag{13}$$

It must also be noted that, length≠connected nodes. In the given equation 14 notations A expresses adjacency matrix, a is node density and e indicates the link between two nodes if exists. Finally, the complete proposed algorithm mathematical formula can be written as,

$$Propose\ (x,y) = \sum_{x=1}^{n}\sum_{y=1}^{n} \left[Sl(D)'_{x,y}\ ,\ max(a)_{xy}\right] \tag{14}$$

Algorithm 1 : Link Prediction

Require: $Modified\ Matrix\ E^U - E = E^P$
Ensure: $Predicted\ matrix\ \Gamma(E)$
  $Max \leftarrow max(max(data(:,1))$
  $SM \leftarrow zeros(Max)$
  while $i <= length(data(:,2))$ **do**
    $SM \leftarrow (data(:,1), data(:,2)) = 1$
    $Modify\ Creation\ SM - \Gamma(E) = E \leftarrow (E^T U E^P)$
    $Note : E^T \cap E^P = \phi$
    $Compute\ shortest\ path\ similarity\ S_1 \leftarrow length(x,y)$
    $Higher\ dense\ node\ S_2 \leftarrow \alpha_{i,j}$
  end while
$Predicted\ matrix \leftarrow (S_1\ and\ S_2)$

3.1. Experimental Setup

For experiments, nine different data sets are being used, which belong to different domains and are associated with biomedical, animal, social, human relations, and transport. Data sets are concisely introduced in Section 3.1. To begin the initial interaction of LP, the adjacency square matrix has been made equivalent, as far as possible, to the size of a complex network. The adjacency matrix contains only 0 and 1, where 0 indicates that there is no connection between the pair of hubs, while 1 infers that the connection exists between the pair of hubs in the arrangement and can be numerically addressed by $e = (x, y)$.The code of this paper is available at [1].

Datasts: All the network data sets utilized in the tests are real-world complex networks and were downloaded from [2],[3]. In the experimentation section of this work, directions and weights are ignored. All datasets are briefly explained.

karate: Karate is a well-known and freely available dataset; it is also known as the Zachary Karate Club, which was compiled in 1977 from university karate club members. Members of the karate club serve as the dataset's "vertices," or "nodes," while "edges," or "links," represent an association of ties between two club members. The karate club dataset contains 34 vertices and 78 edges, which can be denoted by V and E, respectively [45].

Dolphin: This is the animal social network of dolphins, where the links represent their frequent interactions with one another and the nodes are dolphin expressions. This network contains 62 nodes and 159 links or associations [46].

Sampson: This undirected network holds ratings between monks related to a crisis in a cloister (or monastery) in New England (USA), causing some monks to leave. This dataset combines multiple accessible ratings (liking, positive/negative influence, praise or blame) into a single rating that is positive if all original ratings were positive and negative if all original ratings were negative. If there was a tie, the rating is 0. A monk is represented by a vertex that has 18 nodes, whereas an edge between two monks is represented by an edge of 48 nodes.

Contiguous USA: The contiguous USA dataset is made up of 48 contiguous states in the United States

[1] https://github.com/mw364/IDA-Code
[2] https://snap.stanford.edu/data/
[3] https://graphchallenge.mit.edu/data-sets

of America. The connection represents the shared border of two states, and the node represents the states of America. Its undirected graph has 107 links and does not contain any loops.

Kinder-Garten: This dataset is an undirected network with 15 nodes, which are students, and 57 edges, which represent the interactions between them.The dataset contains the status of immunisations for students.

Terrorist: Train bombing is an unweighted and undirected real-world system [47], that consists of 63 suspected terrorists who were accepted to be involved in the March 11, 2004 train bombing in Madrid. In this real-world network, nodes represent terrorists Thus, the links between two nodes are created when a couple of terrorists join training together, which is 243 in number. This can also be explained by saying that in this network, the nodes are represented by terrorists, and the relationship/association between them is represented by links.

Zebra: This is a complex real-world network that has 26 nodes and 113 links. The direction and weight of the network are also ignored here. A vertex expresses a single animal, while an edge indicates the interaction between two animals.

Kangaroo: Dolphin and kangaroo networks show the interaction of animals. This undirected network consists of relationships among free-ranging grey kangaroos. There are 14 nodes and 91 links, where the vertex expresses kangaroos, and a link between two kangaroos shows that there was an interaction between them. The edge values denote the total count of interactions.

Human Contact: It is an undirected real-world complex network with 41 nodes, each of which can represent a person, and 336 edges, each of which can represent communication between two people.

3.2. Evaluation Procedure

The AUC (area under the curve) has been used as an evaluation matrix in order to check which link prediction algorithm is best in terms of accuracy. Broadly speaking, AUC takes the predicted matrix as an input and suggests the best link predictor. Each dataset has 10% of its connections deleted, and the remaining 90% of the dataset is used to create a connected graph. This is an important element to remember when splitting the datasets: the remaining 90% of the dataset must remain connected. After the dataset has been segmented into training and testing, the prediction methods for link prediction are used. It's also worth noting that modified training and testing sets must be in the form of a connected, complex network. Assume U represents the set of all possible links, or the network density, which can be calculated using a simple formula where V represents a positive value. Suppose E represents the set of non-existent edges. The mathematical form is,$\acute{E} = \frac{U}{E}$. The group of known edges E is irregularly partitioned into two disconnected groups, i.e., the set of training E^T which is being used in experiments for training purposes while the second test set E^P is used for the purpose of testing. Also, the information provided by E^T is utilized to estimate lost edges while the information acquired by E^P is utilized to measure the functioning of the link estimation method. Clearly, two groups of sets model a separation of E set i.e., $E = E^T \cup E^P$, and $E^T \cap E^P = \phi$. A model metric to compute the accuracy of the prediction algorithm is AUC [10]. The execution of each algorithm is repeated 100 times across every dataset to produce a singular numerical result. This criterion was applied to all datasets and methods. The evaluation criteria have been graphically viewed in Figure 2.

The formula we have provided is an approximation of the AUC (Area Under the Curve) metric, which is commonly used to evaluate the performance of a binary classifier in link prediction tasks. In this formula, n represents the total number of positive and negative examples in the dataset, n' represents

the number of concordant pairs (i.e., pairs where the predicted score for the positive example is higher than the predicted score for the negative example), and n'' represents the number of tied pairs (i.e., pairs where the predicted scores for both examples are equal). The formula is derived based on the Mann-Whitney U statistic, which is a non-parametric test used to compare two independent samples. In the context of link prediction, the Mann-Whitney U statistic is used to compare the distribution of scores assigned to positive and negative examples by a given classifier. The AUC can be calculated from the Mann-Whitney U statistic as follows:

$$AUC = \frac{U}{n_{pos} \times n_{neg}} \tag{15}$$

where U is the Mann-Whitney U statistic, n_{pos} is the number of positive examples, and n_{neg} is the number of negative examples. The formula you provided is an approximation of the AUC that is often used when dealing with large datasets or when calculating the AUC in an online setting. It is based on the fact that the number of concordant pairs (n') and tied pairs (n'') can be estimated from the total number of pairs (n) and the sum of ranks assigned to the positive examples (R_{pos}):

$$n' = R_{pos} - \frac{n_{pos} \times (n_{pos} + 1)}{2} \; n'' \qquad = n - n' - (n_{pos} \times n_{neg}) \tag{16}$$

Where R_{pos} is the sum of ranks assigned to the positive examples, which can be calculated as follows:

$$R_{pos} = \sum_{i=1}^{n_{pos}} rank_i \tag{17}$$

where $rank_i$ is the rank assigned to the i^{th} positive example. Substituting these estimates into the AUC formula, we get:

$$AUC = \frac{(n' + 0.5n'')}{n} \tag{18}$$

This formula provides an approximation of the AUC that is computationally efficient and does not require storing the scores for all pairs. However, it may not be as accurate as the exact AUC calculation, especially when dealing with imbalanced datasets or when the number of tied pairs is large.
Due to its capacity to evaluate a model's capability to rank pairs of nodes in order of the probability of being related, Area Under the Curve (AUC) is a popular performance metric for link prediction challenges. Predicting whether or not a link exists between two nodes in a network is known as a "link prediction problem," and it requires knowledge of both the nodes and their relationships. In link prediction tasks, AUC is often used due to its ability to give a global measure of performance that is independent of the threshold employed to generate predictions. That is, it evaluates the model's ability to classify positive and negative associations over an infinite range of threshold values. This is relevant because various models may have different optimum thresholds based on the nature of the data and the issue being addressed, and the threshold used to generate link predictions may have a substantial influence on a model's performance.

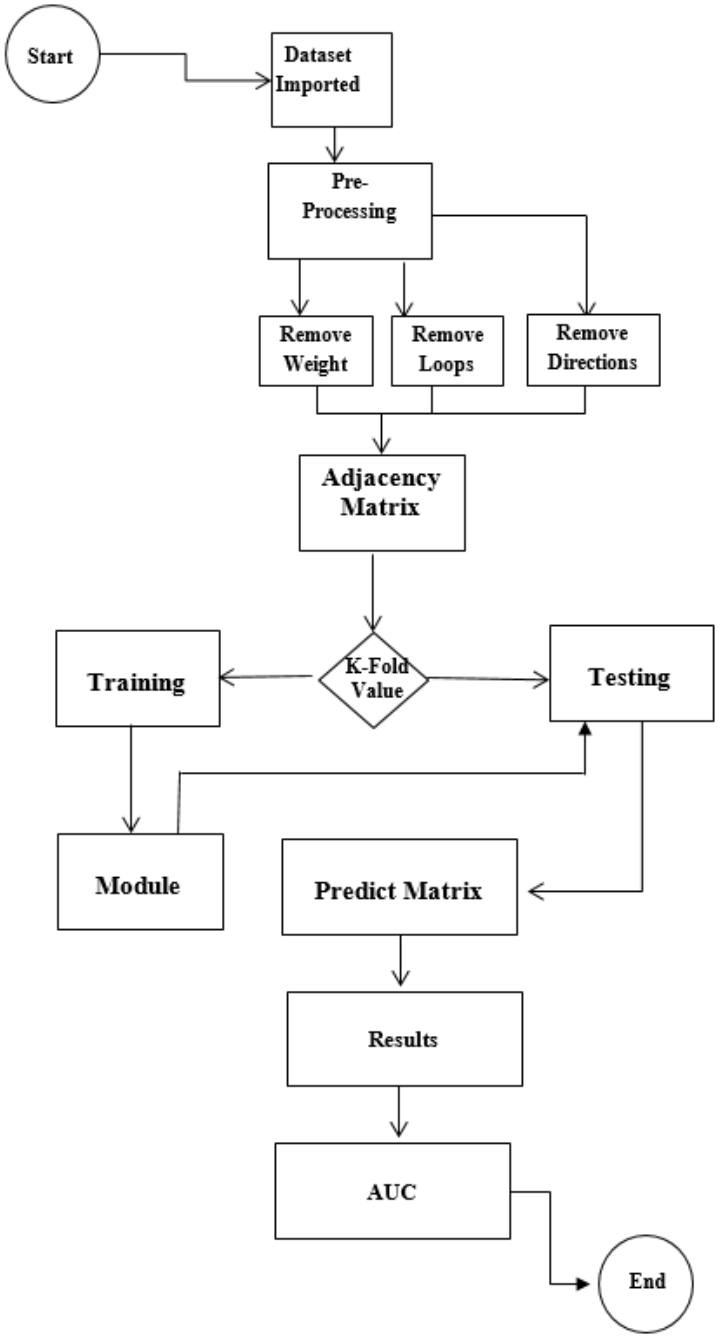

Fig. 2. Evaluation Criteria

4. Results and Discussion

The execution of each algorithm is repeated 100 times across every dataset to produce a singular numerical result. This criterion was applied across nine different datasets are given in Table 1 also graphi-

Table 1
Numerical Comparison Results of Proposed Work

Data-sets	LHN	PD	AA	CN	PA	KATZ	HP	SI	PROPOSED
Karate	0.58988	0.37902	0.72181	0.69104	0.73089	0.71936	0.67104	0.65859	**0.77899**
Dolphin	0.77443	0.61702	0.79779	0.79529	0.66928	0.8054	0.77529	0.80243	**0.84694**
Sampson	0.53682	0.44791	0.62451	0.68899	0.71063	0.718	0.66859	**0.80243**	0.64355
Conti USA	0.91982	0.82333	0.90468	0.89822	0.44507	0.80612	0.79822	0.80888	**0.92674**
Kinder-Garten	0.80311	0.68414	0.80778	**0.81051**	0.61988	0.73101	0.80041	0.80437	0.79203
Terrorist	0.8063	0.75739	0.87959	0.85658	0.67877	0.71293	0.85658	0.85785	**0.8819**
Zebra	0.93681	0.64041	0.96246	0.95098	0.81566	0.88513	0.91813	0.94837	**0.96725**
Kangaroo	0.53729	0.6244	0.91175	0.90271	0.88862	0.88834	0.87271	0.9065	**0.91143**
Human Contact	0.73602	0.61208	0.85781	**0.84944**	0.71809	0.72795	0.84944	0.84913	0.85674

* Note: All the short terms are extended here as, LHN stands for Leicht Holme Newman, PD Path Dependent, AA Adami-Adar, CN Common Neighbor, PA Preferential Attachment, Katz, HP Hub Promoted, Sorensen index, and Proposed algorithm.

cally represented in Figure 3.

Comparing the proposed method to the other eight link prediction strategies, accuracy is generally greater for the proposed approach. The LHN and PD findings, however, are less favorable than those of the others. In the given Figure 3, the majority of the algorithms performed better and attained accuracy levels of around 90% across three datasets (the contiguous United States, Zebra, and Kangaroo). The performance of all techniques was then assessed using dolphin and kindergarten datasets, with some variation in findings shown with terrorist and human contact networks. Because the results across these two datasets are around 86%, 75%, 71%, and 60%. Furthermore, across the Sampson and Karate networks, all of the techniques produce significantly different results from one another. Three of the state-of-the-art strategies performed well over the majority of the datasets when compared to the others, except for the proposed one.

The proposed algorithm was experimentally demonstrated over nine datasets and compared with nine other state-of-the-art link prediction algorithms. From the experimental results, it has been clear that the proposed algorithm has higher accuracy over seven datasets. For more detail, each dataset is further divided and the test ratio is increased to 20%, 30%, 40%, and 50%. All these results are expressed in Figure 4.

Based on the features and structure of complex networks, various algorithms perform differently on different datasets. Suppose the karate dataset is given in Figure 4, where the performance of LHN is very low as compared to PA. Next, over the dolphin dataset, LHN has higher accuracy as compared to PA. Similarly, there is variation in the results of some of the algorithms, such as the PD, LHN, PA, and SI over Karate, Dolphin, Zebra, Human, Contiguous USA, Kindergarten, Kangaroo, and Sampson.

4.1. Comparative Analysis

Table 2 contrasts the performance of the proposed technique with that of other well-known methods for link prediction problems. The proposed technique increased accuracy by 0.17%, 0.75%, 8.69%, 1.74%, and 0.06%, for the various complex network datasets shown in Table 2. Global topological techniques only act on global features, whereas local topological methods only function on local features. This is

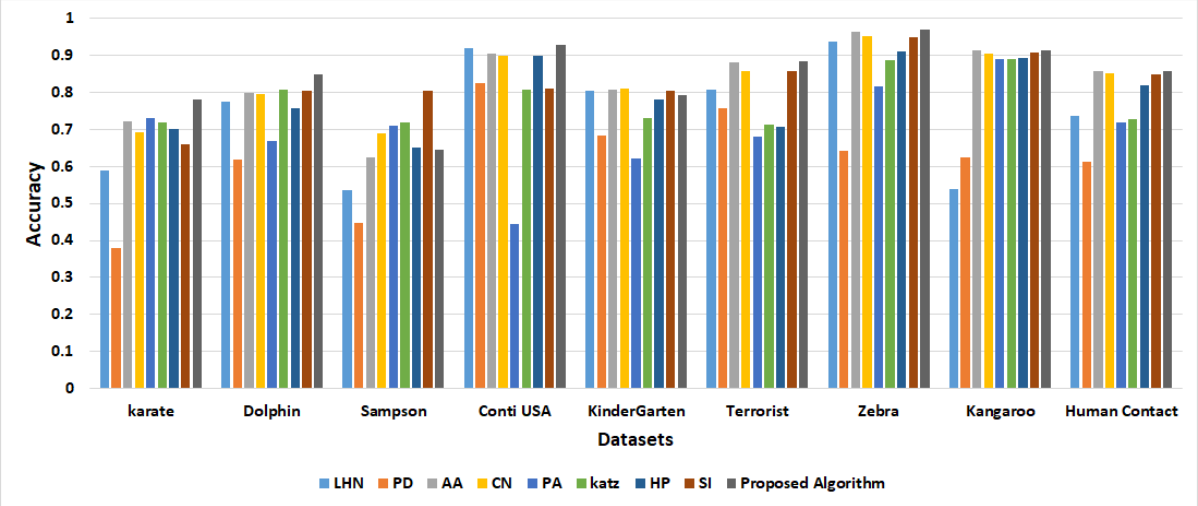

Fig. 3. Graphical Representation of Overall Results

due to the irregular structure and behaviours of complex networks, which affect how various algorithms operate and provide different AUC values. Some complex networks have significant local or global features, which clearly affect the AUC value. The AUC results are often closely associated with the complex network topology. Table 2 contains columns where the letter "X" may be found, which shows that the experiments in the paper did not use this dataset.

Table 2
Comparative Analysis of Proposed Technique with Other Well-known approaches

Datasets	Karate	Dolphin	Conti-USA	Zebra	Terrorist
Double-Degree [44]	72.21%	81.10%	83.99%	94.99%	*X*
Iteration Based LP [48]	*X*	83.71%	*X*	*X*	*X*
Preference RW LP [49]	*X*	80.01%	*X*	*X*	*X*
LPXGB [50]	77.72%	*X*	*X*	*X*	*X*
MFI with LF [50]	73.84%	*X*	*X*	*X*	88.13%
LP via Evolutionary Perturbations[51]	*X*	69.8%	*X*	*X*	87.8%
CND [52]	72.5%	79%%	*X*	80.16%	*X*
Efficient LP [34]	73.84%	*X*	*X*	*X*	*X*
Proposed	**77.89%**	**84.46%**	**92.67%**	**96.73%**	**88.19%**

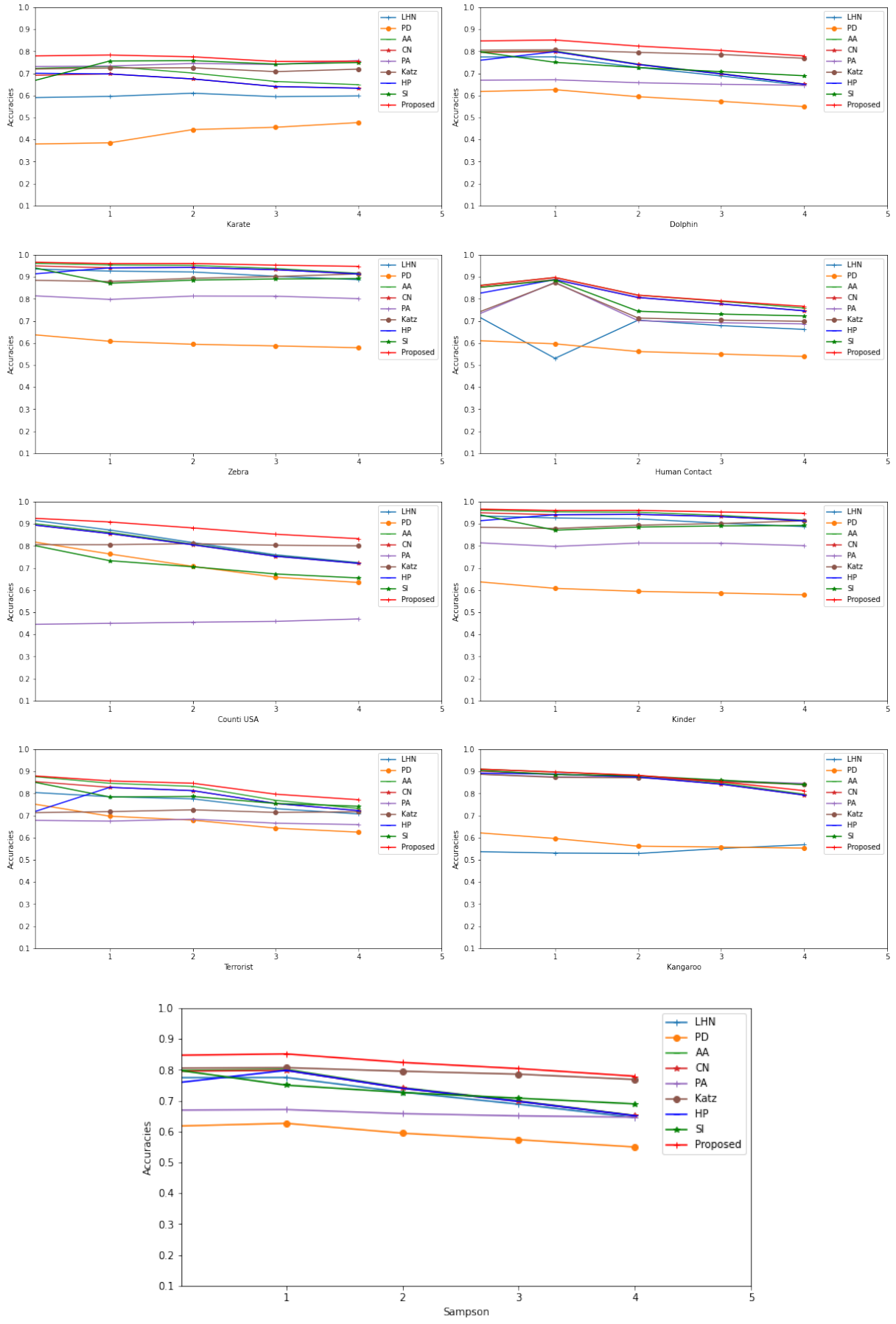

Fig. 4. The accuracy of the proposed and other techniques, as measured by AUC, using various training and probing set division ratio.

5. Conclusion

Link prediction is perhaps the most significant and testing region in a complex network investigation. The objective of link prediction is to appraise the probability of missing connections based on existing connections, highlights, and data in a network. It can help identify potential collaborations or connections. In a business or social network, link prediction can help identify potential partners or connections that might be valuable for a company or individual. In this paper, a novel link prediction technique is proposed that can be utilized in many ways for the treatment of different kinds of diseases, recommender systems, coauthorship prediction, spam mail detection, etc. The proposed framework depends on the shortest path and mutual nodes between two vertices in a real-world complex network. The experimental results over nine complex network data-sets show that the proposed technique gives higher accuracy, as estimated by AUC, compared with other state-of-the-art methods. The proposed work can be extended to directed and weighted real-world complex networks in the future.

6. Acknowledgements

This work was supported by the FCT – Fundação para a Ciência e a Tecnologia, I.P. [Project UIDB/05105/2020]

References

[1] M. Sales-Pardo, R. Guimera, A.A. Moreira and L.A.N. Amaral, Extracting the hierarchical organization of complex systems, *Proceedings of the National Academy of Sciences* **104**(39) (2007), 15224–15229.
[2] W. Li, T. Li and K. Berahmand, An effective link prediction method in multiplex social networks using local random walk towards dependable pathways, *Journal of Combinatorial Optimization* **45**(1) (2023), 31.
[3] A. Dellnitz and W. Rödder, An entropy-based framework to analyze structural power and power alliances in social networks, *Scientific reports* **10**(1) (2020), 1–12.
[4] S. Martinčić-Ipšić, E. Močibob and M. Perc, Link prediction on Twitter, *PloS one* **12**(7) (2017), e0181079.
[5] M. Fire, L. Tenenboim-Chekina, R. Puzis, O. Lesser, L. Rokach and Y. Elovici, Computationally Efficient Link Prediction in a Variety of Social Networks, *ACM Trans. Intell. Syst. Technol.* **5**(1) (2014). doi:10.1145/2542182.2542192.
[6] M. Sumathipala and S.T. Weiss, Predicting mirna-based disease-disease relationships through network diffusion on multi-omics biological data, *Scientific reports* **10**(1) (2020), 1–12.
[7] H. Gul, F. Al-Obeidat, A. Amin, F. Moreira and K. Huang, Hill Climbing-Based Efficient Model for Link Prediction in Undirected Graphs, *Mathematics* **10**(22) (2022), 4265.
[8] A. Kumar, S.S. Singh, K. Singh and B. Biswas, Link prediction techniques, applications, and performance: A survey, *Physica A: Statistical Mechanics and its Applications* **553** (2020), 124289.
[9] L. Lü and T. Zhou, Link prediction in complex networks: A survey, *Physica A: statistical mechanics and its applications* **390**(6) (2011), 1150–1170.
[10] H. Wang and Z. Le, Seven-Layer Model in Complex Networks Link Prediction: A Survey, *Sensors* **20**(22) (2020), 6560.
[11] D. Lin et al., An information-theoretic definition of similarity., in: *Icml*, Vol. 98, Citeseer, 1998, pp. 296–304.
[12] H. Gul, F. Al-Obeidat, A. Amin, M. Tahir and F. Moreira, A systematic analysis of community detection in complex networks, *Procedia Computer Science* **201** (2022), 343–350.
[13] 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.
[14] F. Lorrain and H.C. White, Structural equivalence of individuals in social networks, *The Journal of mathematical sociology* **1**(1) (1971), 49–80.
[15] T. Zhou, L. Lü and Y.-C. Zhang, Predicting missing links via local information, *The European Physical Journal B* **71**(4) (2009), 623–630.
[16] L.A. Adamic and E. Adar, Friends and neighbors on the web, *Social networks* **25**(3) (2003), 211–230.
[17] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, *science* **286**(5439) (1999), 509–512.

[18] L. Katz, A new status index derived from sociometric analysis, *Psychometrika* **18**(1) (1953), 39–43.
[19] L. Lü, C.-H. Jin and T. Zhou, Similarity index based on local paths for link prediction of complex networks, *Physical Review E* **80**(4) (2009), 046122.
[20] F. Fouss, A. Pirotte, J.-M. Renders and M. Saerens, Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation, *IEEE Transactions on knowledge and data engineering* **19**(3) (2007), 355–369.
[21] C. Yu, X. Zhao, L. An and X. Lin, Similarity-based link prediction in social networks: A path and node combined approach, *Journal of Information Science* **43**(5) (2017), 683–695.
[22] P. Zhang, D. Qiu, A. Zeng and J. Xiao, A comprehensive comparison of network similarities for link prediction and spurious link elimination, *Physica A: Statistical Mechanics and its Applications* **500** (2018), 97–105.
[23] H. Gao, J. Huang, Q. Cheng, H. Sun, B. Wang and H. Li, Link prediction based on linear dynamical response, *Physica A: Statistical Mechanics and its Applications* **527** (2019), 121397.
[24] B. Zhu and Y. Xia, An information-theoretic model for link prediction in complex networks, *Scientific reports* **5** (2015), 13707.
[25] H. Wang and Z. Le, Expert recommendations based on link prediction during the COVID-19 outbreak, *Scientometrics* **126**(6) (2021), 4639–4658.
[26] S.S. Singh, S. Mishra, A. Kumar and B. Biswas, Link Prediction on Social Networks Based on Centrality Measures, in: *Principles of Social Networking*, Springer, 2022, pp. 71–89.
[27] E. Estrada, *The structure of complex networks: theory and applications*, Oxford University Press, 2012.
[28] I. Esslimani, A. Brun and A. Boyer, Densifying a behavioral recommender system by social networks link prediction methods, *Social Network Analysis and Mining* **1**(3) (2011), 159–172.
[29] J. Xu, A. Liu, N. Xiong, T. Wang and Z. Zuo, Integrated collaborative filtering recommendation in social cyber-physical systems, *International Journal of Distributed Sensor Networks* **13**(12) (2017), 1550147717749745.
[30] F. Folino and C. Pizzuti, Link prediction approaches for disease networks, in: *International Conference on Information Technology in Bio-and Medical Informatics*, Springer, 2012, pp. 99–108.
[31] M. Wasim, Link Prediction Using Double Degree Equation with Mutual and Popular Nodes, *Trends and Applications in Information Systems and Technologies: Volume 4*, 328.
[32] A. Clauset, C. Moore and M.E. Newman, Hierarchical structure and the prediction of missing links in networks, *Nature* **453**(7191) (2008), 98.
[33] R. Pech, D. Hao, L. Pan, H. Cheng and T. Zhou, Link prediction via matrix completion, *EPL (Europhysics Letters)* **117**(3) (2017), 38002.
[34] H. Gul, F. Al-Obeidat, A. Amin, M. Tahir and K. Huang, Efficient link prediction model for real-world complex networks using matrix-forest metric with local similarity features, *Journal of Complex Networks* **10**(5) (2022), cnac039.
[35] H. Gul, F. Al-Obeidat, F. Moreira, M. Tahir and A. Amin, Real-world protein particle network reconstruction based on advanced hybrid features, in: *Proceedings of International Conference on Information Technology and Applications: ICITA 2021*, Springer, 2022, pp. 15–22.
[36] D. Liben-Nowell and J. Kleinberg, The link-prediction problem for social networks, *Journal of the American society for information science and technology* **58**(7) (2007), 1019–1031.
[37] M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach and Y. Elovici, Link prediction in social networks using computationally efficient topological features, in: *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, IEEE, 2011, pp. 73–80.
[38] J. Wang and L. Rong, Similarity index based on the information of neighbor nodes for link prediction of complex network, *Modern Physics Letters B* **27**(06) (2013), 1350039.
[39] K.-k. Shang, M. Small, X.-k. Xu and W.-s. Yan, The role of direct links for link prediction in evolving networks, *EPL (Europhysics Letters)* **117**(2) (2017), 28002.
[40] Y.-X. Zhu, L. Lü, Q.-M. Zhang and T. Zhou, Uncovering missing links with cold ends, *Physica A: Statistical Mechanics and its Applications* **391**(22) (2012), 5769–5778.
[41] T.A. Sorensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons, *Biol. Skar.* **5** (1948), 1–34.
[42] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai and A.-L. Barabási, Hierarchical organization of modularity in metabolic networks, *science* **297**(5586) (2002), 1551–1555.
[43] E.A. Leicht, P. Holme and M.E. Newman, Vertex similarity in networks, *Physical Review E* **73**(2) (2006), 026120.
[44] H. Gul, A. Amin, F. Nasir, S.J. Ahmad and M. Wasim, Link prediction using double degree equation with mutual and popular nodes, in: *Trends and Applications in Information Systems and Technologies: Volume 4 9*, Springer, 2021, pp. 328–337.
[45] W.W. Zachary, An information flow model for conflict and fission in small groups, *Journal of anthropological research* **33**(4) (1977), 452–473.

[46] R. Rossi and N. Ahmed, The network data repository with interactive graph analytics and visualization, in: *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
[47] B. Hayes, Connecting the dots, *American Scientist* **94**(5) (2006), 400–404.
[48] Y. Yao, R. Zhang, F. Yang, J. Tang, Y. Yuan and R. Hu, Link prediction in complex networks based on the interactions among paths, *Physica A: Statistical Mechanics and its Applications* **510** (2018), 52–67.
[49] K. Berahmand, E. Nasiri, S. Forouzandeh and Y. Li, A preference random walk algorithm for link prediction through mutual influence nodes in complex networks, *Journal of King Saud University - Computer and Information Sciences* **34**(8, Part A) (2022), 5375–5387. doi:https://doi.org/10.1016/j.jksuci.2021.05.006. https://www.sciencedirect.com/science/article/pii/S1319157821001099.
[50] D.K. Behera, M. Das, S. Swetanisha, J. Nayak, S. Vimal and B. Naik, Follower link prediction using the XGBoost classification model with multiple graph features, *Wireless Personal Communications* (2021), 1–20.
[51] S. Yu, M. Zhao, C. Fu, J. Zheng, H. Huang, X. Shu, Q. Xuan and G. Chen, Target defense against link-prediction-based attacks via evolutionary perturbations, *IEEE Transactions on Knowledge and Data Engineering* **33**(2) (2019), 754–767.
[52] J. Yang and X.-D. Zhang, Predicting missing links in complex networks based on common neighbors and distance, *Scientific Reports* **6** (2016), 38208.