Strengthening the Common Neighbor Algorithm by the Degree of Importance of Nodes for Predicting Missing Links in Complex Networks

Mourad Charikhi
Department of Computer Science
Mohamed El Bachir El Ibrahimi University
El-Anasser
34030, Bordj Bou Arreridj, Algeria

Social and complex networks analyses are important areas of research. Link prediction is a discipline in this field that attempts to find missing links and predict new links in networks. Node proximity approaches are widely used in link prediction problems; however, they suffer from weak performance. We propose a new method based on node proximity that exploits common neighbors and the degree of importance of nodes using the well-known PageRank algorithm. Our method improves the performance of existing local methods with a low computation time. Experiments conducted on nine datasets show the advantage of our new method compared to the basic methods of common neighbors, Adamic–Adar and resource allocation. We compare our approach and 14 state-of-the-art techniques, including path and local information approaches. The results indicate that our new approach provides a significant improvement in terms of area under the receiver operating characteristic curve (AUC) score with linear computational complexity.

Keywords: link prediction; similarity metrics; PageRank algorithm; network evolution; social network

1. Introduction

Researchers in fields such as computing, applied mathematics and engineering rely on analytical structures known as complex networks to understand system diversity. Complex networks are composed of a large number of nodes (or points) interconnected by links. The points represent the components of the system and the links represent the interactions between these components. Figure 1 is an example of a complex network graph. Link prediction is used to discern network evolution by predicting future links or identifying missing links. Network structure evolution has many parameters that contribute to the link prediction process. Several researchers have investigated this

https://doi.org/10.25088/ComplexSystems.35.2.181

problem, based on the essential components of node information and graph structure. Detecting missing links or predicting new links is used in such applications as recommender systems [1, 2], social networks [3, 4] and academic networks [5, 6]. Today, e-commerce sites use link prediction to provide shoppers with suggestions for new products based on their preferences or purchases. Future friendships can be predicted using social network analysis [4]. In bioinformatics, we cite the prediction of protein–protein interactions [7, 8].

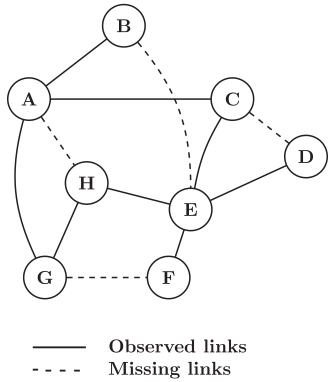


Figure 1. Example of a complex network graph.

We represent a network by an undirected graph G that contains two principal sets of vertices V and edges E. The set of edges E represents the connections between pairs of vertices in V. Link prediction can estimate missing links in the time interval t or predict future links in the interval $t + 1$ [4]. The challenge here is to find a useful method that has minimal complexity and provides good performance. Local methods are a promising solution for this issue. Local methods based on similarity can be used to solve the link prediction problem. These methods use structural information related to the proximity of nodes to calculate the score of unconnected node pairs in a network. For this reason, they are not complex, even in difficult networks. However, according to these techniques, nodes that do not share neighbors or neighbors of neighbors have zero probability to be linked in the future [9]. To overcome this limitation, we combine the PageRank algorithm with existing local similarity measures to determine the degree of importance of the nodes. This degree of node importance represents the attraction force (AF) between the unconnected pair of nodes that do not share any neighbors. Therefore, the naive thought is that, in this case, the more important the AF is for a pair of nodes, the more likely they will be connected in the future. Experiments demonstrate the good performance of our new similarity approach

compared to several state-of-the-art methods, while maintaining the low complexity characteristic of local methods.

The remainder of the paper is organized as follows. In Section 2, we give an overview of the link prediction problem, including different approaches classified in this research field. In Section 3, we present the similarity methods used for link prediction in complex networks and introduce 14 well-known algorithms from the literature. In Section 4, we provide our proposed approach combining the PageRank algorithm with methods based on local information. In Section 5, we describe the properties of the nine datasets used. Then we briefly describe the evaluation methodology and assessment metric. After that, we show the experimental results of the improved similarity methods and compare them with several link prediction methods. Section 6 concludes the paper.

2. Literature Review

Thanks to their well-known paper in 2003, Liben-Nowell and Kleinberg [3] are considered pioneers in work on social networks. According to their research, two nodes could be connected if they have a large number of neighbors in common. They suggested that the similarity between two disconnected nodes is expressed by the number of shared common neighbors (CNs). Mark E. J. Newman [5, 10] proposed link prediction in complex networks to identify the laws that govern network evolution, leading to their structural characteristics. Those authors realized that future links can be estimated from information about the structure of the networks. There are several approaches for finding missing links or predicting new links in complex networks, such as probabilistic approaches [11], supervised learning approaches [12], dimensionality reduction [13] and similarity approaches [14]. Several surveys explore and explain these approaches [15–17].

Similarity-based link prediction methods can be categorized into three main classes: local, global and quasi-local methods. In local methods, the estimated score of getting a new link between two nodes is calculated by the node degree and the number of shared neighbors between pairs of nodes [17]. Several papers were published investigating this approach, such as [5, 18–20]. Adamic–Adar (AA) [18] and resource allocation (RA) [21] are also well-known local methods.

We propose an RA index that leverages information on the next nearest neighbors to improve prediction accuracy. The RA index presents a portion of a resource that one node can send to another through their common neighbors. The similarity between two nodes can be defined as the amount of resource received from one to the

https://doi.org/10.25088/ComplexSystems.35.2.181

other and vice versa. For the AA index, which is a variant measure of RA, we propose for each pair of nodes a score similarity that is determined by the common neighbor measure and the node degree from the neighbors shared by the two nodes. This degree penalizes the similarity score and is equal to the inverse of the logarithmic degree of centrality.

Local methods that are characterized by simplicity are very effective [16]. Global methods use the topological information of the network and perform better than local methods, except that they are less accurate for large networks. Furthermore, they suffer from computational complexity and sensitivity to noise [22]. Several papers have introduced approaches such as the shortest path (SP) [4], the Katz index (KI) [23] and the random walk (RW) [24].

Quasi-local methods are a compromise between local and global approaches. Quasi-local methods are based on the entire topological information of the network and are computationally efficient. Some papers using this approach include the local random walk (LRW) and superposed random walk (SRW) [25] algorithms based on the RW method but with less complexity. Similarly, the local path index (LPI) [14] algorithm limits path lengths to a finite number in the KI.

Recent papers combined different metrics in order to obtain better results. The CAR algorithm [26] combines the number of local community links (LCL) method with the CN, AA, RA and PA methods. The CAR algorithm gives good results on large networks [17]. In 2016, the common neighbors and distance (CND) algorithm uses the CN method for nodes that have common neighbors and the SP method for other unconnected nodes [27]. Based on the CND algorithm, the common neighbors and centrality (CNC) algorithm [28] uses the SP and CN metrics by adding a balancing parameter between them to achieve better performance.

Other papers use a probabilistic model to improve the performance of methods based on local information, such as local naive Bayes (LNB) [29] and mutual information (MI) [30]. Both of these methods assume that different common neighbors may play different roles and therefore contribute differently. In practice, the complexity in computation time for both methods is relatively more important compared to the gain in performance.

Beyond techniques that use node proximity and network paths for predicting links, some alternative techniques have been proposed for predicting and reconstructing complex networks. Information-theoretic approaches, such as ARACNE (algorithm for the reconstruction of accurate cellular networks [31]), combine estimating mutual information between variables to infer network structures with data processing inequality principles to eliminate indirect interactions.

These methods have demonstrated considerable success in reconstructing biological networks, a domain in which statistical dependencies are crucial.

In addition to static similarity measures, dynamical modeling frameworks offer another powerful alternative. For example, HiDi [32] introduces a novel method for reconstructing gene regulatory networks from time-series and perturbation data. At its core is a linear differential equation model with adaptive numerical differentiation, a combination that ensures scalability to extremely large regulatory networks.

While these paradigms provide powerful alternatives, they often require strong statistical assumptions or time-series data or have a high computational cost. In contrast, topology-based link prediction methods remain attractive due to their scalability and applicability to large-scale networks. In this paper, we follow this direction and enhance topology-driven prediction by incorporating node importance information through PageRank-based weighting integrated into the proposed similarity metric.

It remains to be seen if the same techniques can be used to predict links that persist in complex networks [33].

3. Similarity Methods for Link Prediction in Complex Networks

We now present the baseline algorithms that are improved by our strategy and introduce 14 well-known algorithms from the literature, including proximity-based methods and shortest paths. All these methods are tested and compared against our proposed similarity measures in Section 5.

3.1 Baseline Algorithms

Our new similarity measure leverages the PageRank algorithm and local methods to estimate new links for all pairs of unconnected nodes. We have selected three well-known local methods based on shared neighbors' information, namely: CN, AA and RA. These methods exhibit a similar level of computational complexity, while some theoretical surveys [16, 17] report an upper bound of $O(e \times k)$, where e denotes the number of edges and k represents the average node degree.

The similarity score between two nodes u and v is determined using an adjacency-list representation and evaluating the intersection of their neighbor sets.

The computational cost for a single pair of nodes is $O(\min(d(u), d(v)))$, where $d(u)$ and $d(v)$ denote the degrees of nodes u

https://doi.org/10.25088/ComplexSystems.35.2.181

and v, respectively. Consequently, computing similarity scores for all candidate node pairs results in an overall complexity of

$$O\left(\sum_{u\in V} d(u)^2\right).$$

For sparse complex networks, where the average node degree is k, this complexity can be approximated by $O(e\times k)$. Therefore, the CN, AA and RA indices remain computationally efficient local similarity methods that are well suited for large-scale networks.

We now introduce the definition and mathematical formulation of the three methods used in our new similarity measure. Γ_x represents the neighborhood of x, which is the set of nodes connected to node x by an edge. $|\Gamma_x|$ represents the degree of a node x and is defined as the number of edges that are connected to the node.

3.1.1 Common Neighbor

The CN method is predicated on the idea that two people can know one another in a social network if they have friends in common. The number of neighbors shared by x and y is the sole component of the score $S(x, y)$, calculated as

$$S^{\mathrm{CN}}(x, y) = |\Gamma_x \cap \Gamma_y|. \tag{1}$$

3.1.2 Adamic–Adar Index

The authors Lada Adamic and Eytan Adar developed the AA index in 2003 to anticipate links in a social network according to the number of shared links between two nodes. The AA index is described as the total of the two nodes' common neighbors' inverse logarithmic degree centralities. This index is based on a simple count of common neighbors, giving more chances for less-connected neighbors, and defines similarity as

$$S^{\mathrm{AA}}(x, y) = \sum_{z\in\Gamma_x\cap\Gamma_y} \frac{1}{\log|\Gamma_z|}. \tag{2}$$

3.1.3 Resource Allocation Index

The RA index approach is similar to the AA method and performs well, especially in networks that are heterogeneous and have a high clustering coefficient. Consider two unconnected nodes x and y, and suppose that node x sends resources to y via nodes common to x and y. The similarity score $S(x, y)$ between these two vertices is then calculated in terms of resources sent from x to y. This is expressed mathematically by

$$S^{\mathrm{RA}}(x, y) = \sum_{z\in\Gamma_x\cap\Gamma_y} \frac{1}{|\Gamma_z|}. \tag{3}$$

3.2 Methods Based on Local Information

We compared our new measures against 10 methods based on local information that simply depend on the neighborhood or local node connections. All these methods are presented in Table 1 using the same notations as Section 3.1 in the functional similarity weight (FSW) score,

$$\gamma = \max(0, \langle \Gamma \rangle - (|\Gamma_x - \Gamma_y| + |\Gamma_x \cap \Gamma_y|)).$$

$\langle \Gamma \rangle$ is the average degree of all nodes from set E.

Name and Reference	Score (x, y)
Jaccard coefficient index (JC) [19]	$S^{\mathrm{JC}}(x, y) = \frac{\lvert\Gamma_x \cap \Gamma_y\rvert}{\lvert\Gamma_x \cup \Gamma_y\rvert}$
hub promoted index (HPI) [34]	$S^{\mathrm{HPI}}(x, y) = \frac{\lvert\Gamma_x \cap \Gamma_y\rvert}{\min(\lvert\Gamma_x\rvert, \lvert\Gamma_y\rvert)}$
hub depressed index (HDI) [34]	$S^{\mathrm{HDI}}(x, y) = \frac{\lvert\Gamma_x \cap \Gamma_y\rvert}{\max(\lvert\Gamma_x\rvert, \lvert\Gamma_y\rvert)}$
CAR-based resource allocation index (CARRA) [26]	$S^{\mathrm{CARRA}}(x, y) = 1 + \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{\lvert\Gamma_x \cap \Gamma_y \cap \Gamma_z\rvert}{\lvert\Gamma_z\rvert}$
CAR-based common neighbors index (CARCN) [26]	$S^{\mathrm{CARCN}}(x, y) = 1 + \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{\lvert\Gamma_x \cap \Gamma_y \cap \Gamma_z\rvert}{2}$
functional similarity weight (FSW) [35]	$S^{\mathrm{FSW}}(x, y) = \left(\left(\frac{2\lvert\Gamma_x \cap \Gamma_y\rvert}{\lvert\Gamma_x - \Gamma_y\rvert + 2\lvert\Gamma_x \cap \Gamma_y\rvert + \gamma}\right)^2\right)$
preferential attachment (PA) [36]	$S^{\mathrm{PA}}(x, y) = \lvert\Gamma_x\rvert \times \lvert\Gamma_y\rvert$
SAM index (SAM) [37]	$S^{\mathrm{SAM}}(x, y) = \frac{\lvert\Gamma_x \cap \Gamma_y\rvert/\lvert\Gamma_x\rvert + \lvert\Gamma_x \cap \Gamma_y\rvert/\lvert\Gamma_y\rvert}{2}$
common neighbor to degree (CN2D) [38]	$S^{\mathrm{CN2D}}(x, y) = \lvert\Gamma_x \cap \Gamma_y\rvert + \beta\left(\frac{1}{\max(\lvert\Gamma_x\rvert, \lvert\Gamma_y\rvert)} \sum_{z \in \{\Gamma_x \cap \Gamma_y\}} \lvert\Gamma_z\rvert\right)$
PageRank and resource allocation (PR-RA) [9]	$S^{(\mathrm{PR\text{-}RA})}(x, y) = \begin{cases} S^{\mathrm{RA}}(x, y), & \text{if } \lvert\Gamma_x \cap \Gamma_y\rvert \neq \emptyset \\ \max(\mathrm{PR}(x), \mathrm{PR}(y)), & \text{otherwise} \end{cases}$

Table 1. Compared methods based on local information.

https://doi.org/10.25088/ComplexSystems.35.2.181

3.3 Methods Based on Shortest Path and Local Information

Table 2 shows a summary of the five measures based on shortest path and local information. For the measures CNC, CND and SP, dist(x, y) represents the shortest path or the distance between two nodes x and y. In the CNC method, the parameter $\alpha \in [0, 1]$ is a user-defined value. In our experiment, the best result was obtained with $\alpha = 0.8$.

Name and Reference	Score (x, y)
common neighbor and centrality (CNC) [28]	$S^{\mathrm{CNC}}(x, y) = \alpha(\lvert\Gamma_x \cap \Gamma_y\rvert) + \frac{N(1-\alpha)}{\mathrm{dist}(x,y)}$
shortest path (SP) [4]	$S^{\mathrm{SP}}(x, y) = -\lvert\mathrm{dist}(x, y)\rvert$
Katz index (KI) [23]	$S^{\mathrm{KI}}(x, y) = \sum_{l=1}^{\infty} \alpha^l \lvert\mathrm{dist}(x, y)^l\rvert$
random walk with restart (RWR) [39]	$S^{\mathrm{RWR}}(x, y) = \vec{P}_y^x + \vec{P}_x^y$

Table 2. Compared methods based on shortest path and local information.

We now provide descriptions and definitions of the Katz and RWR indices.

The Katz index (KI) compiles the set of possible paths between two pairs of unconnected nodes by progressively penalizing the path lengths. The KI is defined by

$$S^{\mathrm{KI}}(x, y) = \sum_{l=1}^{\infty} \alpha^l \lvert\mathrm{dist}_{(x,y)}^l\rvert, \tag{4}$$

where $\mathrm{dist}_{(x,y)}^l$ is the set of path lengths l between nodes x and y. The parameter α is a damping factor with $0 < \alpha < 1$. The similarity between all pairs of nodes can be directly calculated using

$$S = (I - \alpha A)^{-1} - I, \tag{5}$$

where A is the adjacency matrix of the network, I is the identity matrix, and S is the similarity for each pair of nodes x and y ($S_{(x,y)}$ is the (x, y)-element of the matrix S). The time complexity of this method is $O(v^3)$. For our experiment, the best result was obtained with $\alpha = 0.001$.

The random walk with restart (RWR) model is defined by Tong [39] as a random walk optimization problem whose solution is given by

$$\vec{P}^x = (1 - \alpha)\left(I - \alpha M^T\right)^{-1}\vec{S}^x, \tag{6}$$

where α is the probability of moving from one chosen node to another following a random walk. M is the transition probability matrix defined by the adjacency matrix. $\vec{P}^x$ is a vector containing the

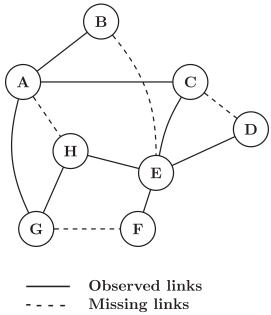
probability of reaching any node using x as a starting point, and $\vec{S}^x$ is the seed vector of length $|V|$ with all its elements set to 0 except for $S_x^x = 1$. It can be iteratively approximated by

$$\vec{P}^x(t) = \alpha M^T \vec{P}^x(t-1) + (1-\alpha)\vec{S}^x. \tag{7}$$

Equation (7) is applied iteratively, where the value zero is initially assigned to all elements of $\vec{P}^x(0)$ and the final score RWR(x, y) is equal to $P_y^x + P_x^y$. The computational time complexity for this method is $O(mv^2k)$, where m is the number of iterations until convergence. The best result for our test was obtained with $\alpha = 0.9$.

4. Proposed Method

Local information methods based on node proximity such as the CN method and its variants are frequently used to solve the link prediction problem. They offer good performance relative to the computational time consumed. However, these local methods have two weaknesses. First, they assume that nodes that have no common neighbors (or neighbors of neighbors) have no chance of being connected in the future. Second, these methods assign the same score to all pairs of nodes that have the same number of common neighbors, regardless of their different importance in the network. Figure 2 shows the drawback of local methods for link prediction. For example, the pair of unconnected nodes (A,H) and (C,D) will score 1 using the CN measure, even though these two pairs do not have the same importance in the graph. On the other hand, the score of pairs (G,F) and (B,E) that do not share any neighbors will be zero according to local methods.



PR Score ($d = 0.85$)	CN Score	AF-CN Score
A: 0.2559,	CN(A,H)=1	AF-CN(A,H)=2.0236
B: 0.1696,	CN(B,E)=0	AF-CN(B,E)=0.7380
C: 0.1030,	CN(C,D)=1	AF-CN(C,D)=1.4120
D: 0.0579,	CN(G,F)=0	AF-CN(G,F)=0.3166
E: 0.1845,		
F: 0.0579,		
G: 0.0792,		
H: 0.0916		

Figure 2. An illustrative example for calculating CN and AF-CN scores.

https://doi.org/10.25088/ComplexSystems.35.2.181

We propose a new method to overcome these limitations using the PageRank algorithm to measure the importance of a node and reinforce local methods by leveraging node importance in the network. This allows for the possibility of connecting a pair of nodes that do not share any neighbors by using their node importance value. Furthermore, pairs of unconnected nodes with the same number of neighbors will be scored differently, depending on the importance of their nodes in the network. We first present the basic notions and theoretical foundations concerning the PageRank algorithm before going into the details and mathematical formulation of our new similarity measure.

The PageRank algorithm was developed by Google cofounders Larry Page and Sergey Brin at Stanford University [40]. According to the paper, PageRank is defined as follows: for a given page P pointed to by pages $S_1, \ldots, S_n$, the PageRank of P is given by the recursive formula:

$$\mathrm{PR}(P) = \frac{(1-d)}{N} + d\left(\frac{\mathrm{PR}(S_1)}{C(S_1)} + \cdots + \frac{\mathrm{PR}(S_n)}{C(S_n)}\right). \tag{8}$$

From equation (8), N is the total number of pages in the web graph, $C(P)$ is the number of outgoing links from the page P, and d is the damping factor that can take real values from 0 to 1. Therefore, a straightforward iterative technique with linear complexity equal to $O(m \times e)$ can be used to calculate the PageRank index, where e is the number of edges and m is the iteration number until convergence.

The PageRank algorithm for a particular page in the web graph can be thought of as a sort of vote that all the other pages conduct to decide the importance of this page. We adapt this principle to complex network analysis, utilizing the PageRank score as a metric for the structural significance of nodes. Local similarity methods such as CN and RA assume that link formation depends strictly on the shared neighborhood structure. However, many real-world networks exhibit hub-attraction behavior, where connections are generally determined by the most influential node involved in the interaction. Examples of this phenomenon include:

- Transportation systems: Regional airports preferentially connect to major international hubs.
- Coauthorship networks: Researchers often seek highly influential authors as collaborators to increase visibility.

We introduce an attraction component based on $\max(\mathrm{PR}(x), \mathrm{PR}(y))$ to model this phenomenon that relies on the dominant structural influence of highly central nodes. This formulation complements local similarity information by integrating global importance into the prediction process. In our method, the PageRank score indicates the

AF of the nodes. Thus, we define the AF between unconnected pairs of x and y nodes for the V vertex set, which is the first measure of similarity, using

$$S^{\mathrm{AF}}(x, y) = \frac{N}{2} \times \max(\mathrm{PR}(x), \mathrm{PR}(y)), \tag{9}$$

where N is the number of nodes in the graph, and $\mathrm{PR}(x)$ and $\mathrm{PR}(y)$ are the two values returned by the PageRank algorithm for nodes x and y, respectively. The normalization factor $N/2$ is applied to align the numerical range of PageRank scores with that of local similarity indices.

The second similarity measure we adopt is a neighborhood-based method, which can be, for instance, any variant of the CN method.

Finally, the resulting similarity measure is the sum of the two previously mentioned measures:

$$S^{\mathrm{MD}}_{\mathrm{AF}}(x, y) = S^{\mathrm{AF}}(x, y) + S^{\mathrm{MD}}(x, y), \tag{10}$$

in which $S^{\mathrm{AF}}(x, y)$ is the AF score, and $S^{\mathrm{MD}}(x, y)$ presents a variant of the common neighbor metric, which can be the CN, AA or RA index.

Figure 2 shows a calculation of similarity scores. Here a sample graph is represented with four missing links along with a weights vector. Column one shows the PageRank score of the nodes. Column two contains the scores given by using only the basic CN method. Column three shows the scores obtained by applying the enhanced CN measure (AF-CN). Notice that the links (A,H) and (C,D) have the same number of shared neighbors, one, and the score returned by CN for the two previous links is one. The scores are different with our new AF-CN measure: 2.0236 and 1.412 for the links (A,H) and (C,D), respectively. This difference is obtained because the pair of nodes A and H have a more important PageRank value than the pair of nodes C and D. Furthermore, we can see that even the links (B,E) and (G,F) do not share any neighbors; the scores generated by the new similarity measure are different from zero, as opposed to the basic CN measure, where we see that the resulting scores are null.

5. Experimental Results

We now introduce the metric and evaluation methodology used in this study and present the dataset properties of nine real-world networks. The performance of our enhanced methods AF-CN, AF-AA and AF-RA are assessed on the nine datasets. A comparison of the area under the receiver operating characteristic curve (AUC) score is made between our new measure and other methods, including baseline, local, shortest path and global approaches.

https://doi.org/10.25088/ComplexSystems.35.2.181

5.1 Metric and Evaluation Methodology

We use the rotation estimation validation model with a five-fold cross-validation approach to evaluate the performance of our similarity method. The network is divided into five subgraphs of equivalent size. Of these subnetworks, only one is regarded as a test graph G_{test}. The remaining components are assembled into a training graph G_{train}. In another way, the training graph is constructed by removing the test edges from the original network. The procedure is repeated for each subgraph, using it once as the test graph and using the other subgraphs as the training graph. The final evaluation score is determined by averaging the outcomes from each round of applying the performance metrics. All structural features, including PageRank and methods based on shortest path, are computed exclusively on the training graph G_{train} within each cross-validation fold. This prevents any information leakage from test edges during the learning and evaluation processes. This evaluation protocol follows common practice in link prediction studies [16].

The predicted graph G_{pred} is produced by applying a link prediction algorithm to the training graph. Given E_{train}, E_{test} and E_{pred} are the set of edges, respectively, included in G_{train}, G_{test} and G_{pred} where $E_{\text{train}} \bigcap E_{\text{test}} = \emptyset$. An edge e is a true positive (TP) link if $e \in E_{\text{test}}$ and $e \in E_{\text{pred}}$, that is, the link is present in both the predicted and test graphs. If $e \notin E_{\text{test}}$ and $e \in E_{\text{pred}}$, e is considered as a false positive (FP) link. The findings may be assessed using several metrics by calculating the number of TP and FP links. We utilize the AUC metric that is frequently used to assess how well link prediction algorithms function in complex networks [41]. The AUC value measures the likelihood that a randomly selected edge from the test set E_{test} obtains a higher similarity score $S(x, y)$ than a randomly selected pair of unconnected nodes used as a negative sample:

$$\text{AUC} = \frac{n_1 + 0.5n_2}{n}, \tag{11}$$

where n is the number of comparisons of independent observations, n_1 is the number of links where the missing link was better ranked than the nonexistent link, and n_2 is the number of pairs where the two links gave an equal score. The AUC score should be close to 0.5 if each score has a distinct and identical distribution. The performance of a particular strategy in relation to random chance is therefore shown by the value at which the AUC exceeds 0.5. Good link prediction algorithms should achieve an AUC value that is close to 1.

5.2 Datasets

Nine real-world complex network datasets with various topological characteristics were chosen from a variety of application sectors. The

datasets were obtained from noesis.ikor.org/datasets/link-prediction. Table 3 summarizes the fundamental topological characteristics and references of the chosen network datasets.

Name	Type	Reference	$\|V\|$	$\|E\|$	$\langle k \rangle$	C	ASPL
BUP	social network	[30]	105	441	8.4	0.49	3.09
CEG	biological network	[42]	297	2148	14.46	0.29	2.46
UAL	transportation network	[43]	332	2126	12.81	0.63	2.74
EML	social network	[44]	1133	5451	9.62	0.22	3.61
INF	social network	[45]	410	2765	13.49	0.46	3.63
SMG	co-authorship network	[46]	1024	4916	9.6	0.31	2.98
YST	biological network	[7]	2284	6646	5.82	0.13	4.29
HMT	social network	[47]	2426	16630	13.71	0.54	3.15
KHN	social network	[47]	3772	12718	6.74	0.25	3.63

Table 3. Properties of the nine real-world networks.

5.3 Results and Discussion

We ran experiments on nine datasets from nine real-world networks to demonstrate the effectiveness of our new methods using five cross-validations. Section 5.3.1 compares the results of our new similarity methods and the three baseline methods to show the improvements obtained. Section 5.3.2 shows the competitiveness of our new AF similarity methods against 14 well-known methods, including methods based on the shortest path and global approaches.

5.3.1 Comparison with Baseline Methods

This section presents the results of comparing our new similarity methods against three baseline methods to show the enhancements achieved by our AF approach. Table 4 presents the AUC evaluation of AF similarity and basic methods performed under nine real-world networks, namely: our three enhanced measures AF-CN, AF-AA and AF-RA and the three baseline methods CN, AA and RA, respectively. We assigned the PageRank algorithm damping factor of $\alpha = 0.01$ for the experiments, yielding remarkable results with faster convergence (fewer than 100 iterations). We calculated the network average node degree for each dataset used. Then, we classified the datasets and reported the AUC scores according to this average node degree in Table 4.

The AUC score results show that the AF value used in our new similarity measures is definitely advantageous for all techniques in all databases used (Table 4). It should be noted that all baseline methods performed well on datasets with a high average node degree (greater than 12). For example, the HMT dataset with an average node degree

of 13.71 obtained a score of about 0.95 by all methods. On the other hand, for the YST biological network, whose average node degree is 5.82, the score is close to 0.68. It is clear that the proposed approach improves all basic methods with low average node degree datasets. For example, for the lowest average node degree YST dataset, the performance gain achieved by all new methods is about 18%. Figure 3 shows the average AUC scores obtained under the nine datasets. According to the histogram, the achieved improvement is around 5% compared to the other three baseline methods in all nine tested datasets. As can be seen here, the approach associated with the RA algorithm was ranked first eight times and consequently produced the best performance.

	YST	KHN	BUP	SMG	EML	UAL	INF	HMT	CEG
$k = \frac{2\|E\|}{\|V\|}$	5.82	6.74	8.40	9.60	9.62	12.80	13.49	13.71	14.46
AF-CN	0.8631	0.892	0.8829	0.8898	0.8759	0.9427	0.9289	0.9595	0.8467
CN	0.685	0.7782	0.8691	0.8138	0.8234	0.9278	0.9264	0.9523	0.8276
Gain	0.18	0.11	0.01	0.08	0.05	0.01	0.00	0.01	0.02
AF-AA	0.8631	0.8973	0.8868	0.8936	0.8774	0.9512	0.9327	0.9624	0.8547
AA	0.6855	0.7839	0.8781	0.8224	0.8251	0.9391	0.93	0.9553	0.845
Gain	0.18	0.11	0.01	0.07	0.05	0.01	0.00	0.01	0.01
AF-RA	0.8634	0.8996	0.8891	0.9022	0.877	0.9614	0.9332	0.9628	0.8624
RA	0.6854	0.7839	0.8801	0.8224	0.8248	0.9439	0.9305	0.9561	0.8485
Gain	0.18	0.12	0.01	0.08	0.05	0.02	0.00	0.01	0.01

Table 4. Evaluation of the prediction accuracy measured by AUC under the nine real-world networks.

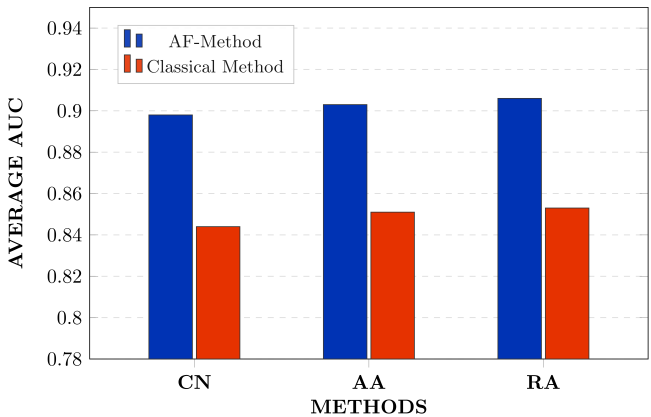


Figure 3. Histogram of the average AUC of the evaluated algorithms. Our measure of AF similarities reaches around 0.90 of the average AUC score, while the baseline measures reach 0.85.

To analyze the accuracy of the proposed approach, we gradually decrease the similarity scores obtained by the two methods (CN and AF-CN) and calculate the resulting number of correctly predicted links. Note that this test was conducted on the datasets CEG, KHN and YST. Figure 4 shows the curves corresponding to the number of TP scores compared to the number of predicted links. Obviously, from the curve, we can see a significant improvement when the number of links equals 3000 in the CGE and KHN datasets, while the advantage of our approach can be seen when the number of links equals 5000 in the YST dataset. Therefore, the results of our approach AF-CN show a clear improvement in accuracy in all three selected datasets. In Section 5.3.2, we present the evaluation results and a second comparison of our new method against other methods using local and path information.

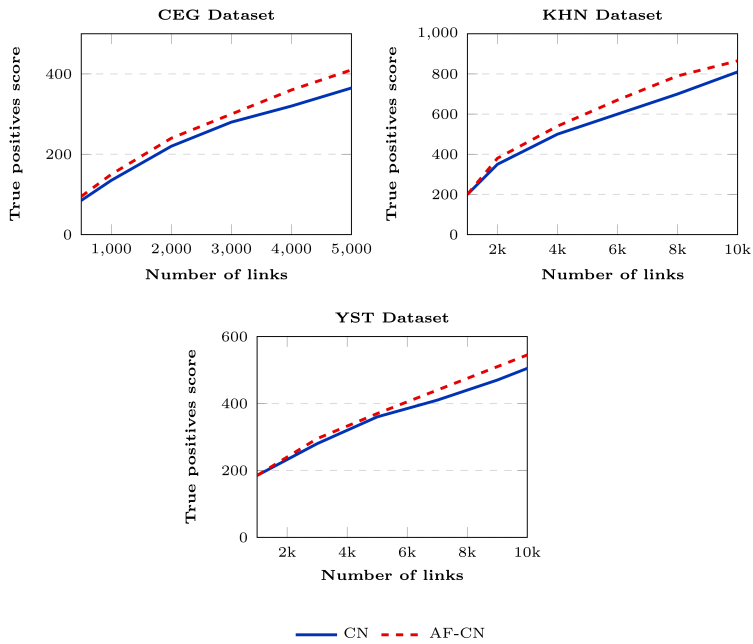


Figure 4. Number of correctly predicted links for the CEG, KHN and YST datasets. The curves correspond to the number of TP scores. Three curves present the obtained results of the three datasets CGE, KHN and YST. The curves show that when there are 3000 links in the CGE and KHN datasets, our proposed measure starts outperforming the basic CN method. In the YST dataset, the advantage of our approach can be seen at 5000 links.

5.3.2 Comparison with Methods Using Local and Path Information

We compare our obtained evaluation results to the 14 methods mentioned in Section 3 to test the competitiveness of our approach. The

AUC score obtained for each algorithm is presented in Table 5. Figure 5 shows a comparison of the average AUC of our AF similarity measure and the compared algorithms.

–	BUP	CEG	UAL	EML	INF	SMG	KHN	YST	HMT
JC	0.8559	0.7767	0.8927	0.8215	0.9286	0.7751	0.7618	0.6837	0.9492
HDI	0.8476	0.7680	0.8879	0.8215	0.9277	0.7747	0.6837	0.9483	0.7618
HPI	0.8682	0.7933	0.8667	0.8186	0.9255	0.7866	0.7677	0.6835	0.9484
CARRA	0.8691	0.8276	0.9278	0.8234	0.9264	0.8138	0.7782	0.6850	0.9523
CARCN	0.8675	0.8240	0.9283	0.8233	0.9265	0.8096	0.6850	0.9521	0.7782
FSW	0.8646	0.8046	0.9112	0.8222	0.9270	0.7952	0.6842	0.9493	0.7700
PA	0.6655	0.7489	0.8840	0.7775	0.7023	0.8346	0.7734	0.8472	0.8107
SP	0.8455	0.7303	0.7641	0.8347	0.9048	0.7827	0.7704	0.7297	0.6135
KATZ	**0.8946**	0.8491	0.9172	**0.8858**	**0.9526**	0.8622	0.8432	0.8044	0.9621
RWR	0.8450	0.7474	0.8559	0.8465	0.9305	0.8042	0.8379	0.7896	0.9337
SAM	0.8665	0.7905	0.8863	0.8197	0.9276	0.7839	0.7659	0.6835	0.9496
CNC	0.8895	0.8316	0.9191	0.8707	0.9492	0.8334	0.7904	**0.9612**	0.8268
PR-RA	0.8882	0.8571	0.9554	0.8769	0.9333	0.8903	0.8928	0.8629	**0.9632**
AF-RA	0.8891	**0.8624**	**0.9614**	0.8770	0.9332	**0.9022**	**0.8996**	0.8634	0.9628

Table 5. Comparison of the prediction accuracy measured by AUC score on nine real-world networks.

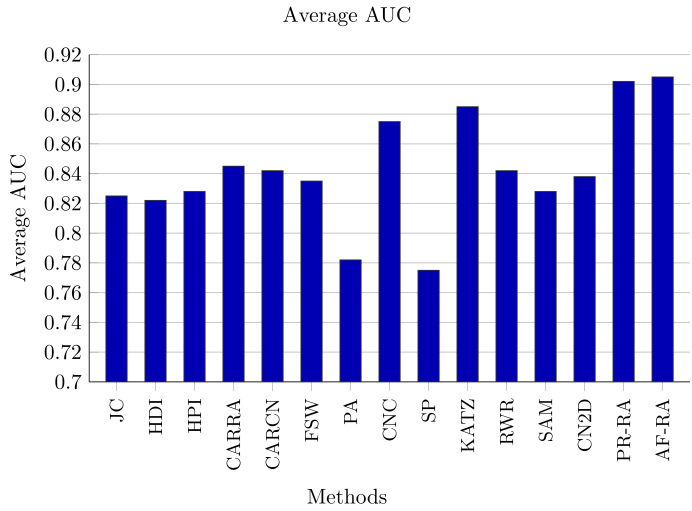


Figure 5. Average AUC score of our AF similarity measure and the compared algorithms.

According to Table 5, there is no single method that is best for all datasets. However, some methods seem to perform better than others on multiple datasets.

We use AF-RA, the best new resource allocation method, to present the competitiveness of our approach. The histogram of the average AUC using the different methods applied to the nine datasets presented in Figure 3 clearly shows that our new AF-RA method obtained the highest average AUC.

Table 5 shows the average AUC scores of the 14 link prediction algorithms applied to nine different datasets; we calculated the standard deviation, minimum value and maximum value. The top 1 and 2 are listed for the four most successful methods, as presented in Table 6.

Algorithms	CNC	KATZ	PR-RA	AF-RA
average AUC	0.8747	0.8857	0.9022	0.9057
standard deviation	0.0594	0.0521	0.0390	0.0386
min value	0.7904	0.8044	0.8629	0.8624
max value	0.9612	0.9621	0.9632	0.9628
number of times ranked first	1	3	1	4
number of times ranked second	1	1	5	3

Table 6. Performance summary of four algorithms CNC, KATZ, PR-RA and AF-RA.

Performance scores for the new AF-RA method range from 0.8624 to 0.9628 across the nine different datasets, with the highest average AUC of 0.9057 and the lowest standard deviation of 0.0386. Despite the low complexity of AF-RA, it is the most efficient and stable in this experiment. Additionally, it obtained the highest score in four datasets (BUP, CEG, UAL and HMT). It was ranked second three times. PR-RA performance scores range from 0.8629 to 0.9632 across the nine datasets, with the second-highest average AUC of 0.9022 and the second-lowest standard deviation of 0.0386. This method was ranked first once in the HMT dataset and second four times behind the AF-AR method, with very close scores. The KI has great predictive power. Its performance scores range from 0.8044 to 0.9621 across the nine datasets, with the third highest average AUC of 0.8857 and the third lowest standard deviation of 0.0386. It was ranked first three times and second one time out of the nine datasets. Unfortunately, it is only applicable to small networks because of the high algorithmic complexity required to calculate the inverse of a matrix [16]. The CNC method scores are competitive, but the results are unstable on datasets with similar topological characteristics. Indeed, it ranked first on one occasion and second on another occasion across the nine datasets, with AUC scores ranging from 0.7904 to 0.9621 and an average AUC of 0.8747. The inclusion of node

attraction in this measure has shown promising results. This will be the subject of future work.

6. Conclusion

Link prediction is an area of research in network analysis and applications such as predicting disease outbreak, controlling network privacy and detecting spam emails. Methods based on local information are very successful in this domain, and many researchers have focused on this approach to solve link prediction problems. One disadvantage of this method is that it predicts links only for nodes that have common neighbors. Numerous variations of local information-based methods have been proposed to improve performance. We propose a new solution in this paper using the PageRank algorithm to represent the attraction force (AF) of nodes. Therefore, all nodes that do not share neighbors might be linked in the future according to their PageRank properties. We performed a series of experiments under nine network datasets to study our new method. We compared our new method against three basic methods, with results that show a significant improvement in the area under the receiver operating characteristic curve (AUC) score. Furthermore, for more credibility and to assess the competitiveness of our proposed similarity method, we compared our approach with 14 other methods that are based on local and path information. The results show that our new similarity methods clearly outperform the existing methods in terms of average AUC measurement. There are many link prediction problems that need to be solved, especially in weighted and directed networks. We propose extending the use of our new method by taking into account the particular characteristics of weighted and directed networks.

References

[1] J. B. Schafer, D. Frankowski, J. Herlocker and S. Sen, "Collaborative Filtering Recommender Systems," *The Adaptive Web: Methods and Strategies of Web Personalization* (P. Brusilovsky, A. Kobsa and W. Nejdl, eds.), Berlin, Heidelberg: Springer, 2007 pp. 291–324. doi:10.1007/978-3-540-72079-9_9.

[2] Z. Kurt, Ö. N. Gerek, A. Bilge and K. Özkan, "A Graph-Based Recommendation Algorithm on Quaternion Algebra," *SN Computer Science*, **3**(4), 2022 299. doi:10.1007/s42979-022-01171-4.

[3] D. Liben-Nowell and J. Kleinberg, "The Link Prediction Problem for Social Networks," in *Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM '03)*, New Orleans, New York: Association for Computing Machinery, 2003 pp. 556–559. doi:10.1145/956863.956972.

[4] 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 pp. 1019–1031. doi:10.1002/asi.20591.

[5] M. E. J. Newman, "The Structure of Scientific Collaboration Networks," *Proceedings of the National Academy of Sciences*, **98**(2), 2001 pp. 404–409. doi:10.1073/pnas.98.2.404.

[6] D. Mohdeb, A. Boubetra and M. Charikhi, "Strength-Based Link Prediction in Scientific Bibliographic Networks," *Journal of Information Technology Research*, **10**(3), 2017 pp. 84–106. doi:10.4018/JITR.2017070106.

[7] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang et al., "Topological Structure Analysis of the Protein–Protein Interaction Network in Budding Yeast," *Nucleic Acids Research*, **31**(9), 2003 pp. 2443–2450. doi:10.1093/nar/gkg340.

[8] B. Mewara and S. Lalwani, "A Survey on Deep Networks Approaches in Prediction of Sequence-Based Protein–Protein Interactions," *SN Computer Science*, **3**(4), 2022 298. doi:10.1007/s42979-022-01197-8.

[9] M. Charikhi, "Association of the PageRank Algorithm with Similarity-Based Methods for Link Prediction in Complex Networks," *Physica A: Statistical Mechanics and Its Applications*, **637**, 2024 129552, doi:10.1016/j.physa.2024.129552.

[10] M. E. J. Newman, "Finding Community Structure in Networks Using the Eigenvectors of Matrices," *Physical Review E*, **74**(3), 2006 036104. doi:10.1103/PhysRevE.74.036104.

[11] R. R. Sarukkai, "Link Prediction and Path Analysis Using Markov Chains," *Computer Networks*, **33**(1), 2000 pp. 377–386. doi:10.1016/S1389-1286(00)00044-X.

[12] M. Al Hasan, V. Chaoji, S. Salem and M. Zaki, "Link Prediction Using Supervised Learning," in *SDM06: Workshop on Link Analysis, Counter-Terrorism and Security*, Bethesda, MD, volume 30, 2006 pp. 798–805.

[13] A. Pecli, B. Giovanini, C. C. Pacheco, C. Moreira, F. Ferreira, F. Tosta, J. Tesolin et al., "Dimensionality Reduction for Supervised Learning in Link Prediction Problems," in *Proceedings of the 17th International Conference on Enterprise Information Systems - Volume 2 (ICEIS 2015)*, Barcelona, Spain (S. Hammoudi, L. Maciaszek, E. Teniente, O. Camp and J. Cordeiro, eds.), Cham: Springer, 2015 pp. 295–302. doi:10.5220/0005371802950302.

[14] 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. doi:10.1103/PhysRevE.80.046122.

[15] M. A. Hasan and M. J. Zaki, "A Survey of Link Prediction in Social Networks," *Social Network Data Analytics* (C. C. Aggarwal, ed.), Boston, MA: Springer, 2011 pp. 243–275. doi:10.1007/978-1-4419-8462-3_9.

[16] V. Martínez, F. Berzal and J.-C. Cubero, "A Survey of Link Prediction in Complex Networks," *ACM Computing Surveys*, **49**(4), 2016 pp. 1–33. doi:10.1145/3012704.

[17] 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, doi:10.1016/j.physa.2020.124289.

[18] L. A. Adamic and E. Adar, "Friends and Neighbors on the Web," *Social Networks*, **25**(3), 2003 pp. 211–230. doi:10.1016/S0378-8733(03)00009-1.

[19] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et du Jura," *Bulletin de la Société Vaudoise des Sciences Naturelles*, **37**, 1901 pp. 547–579. doi:10.5169/seals-266450.

[20] S. Li, J. Huang, Z. Zhang, J. Liu, T. Huang and H. Chen, "Similarity-Based Future Common Neighbors Model for Link Prediction in Complex Networks," *Scientific Reports*, **8**(1), 2018 17014. doi:10.1038/s41598-018-35423-2.

[21] T. Zhou, L. Lü and Y.-C. Zhang, "Predicting Missing Links via Local Information," *The European Physical Journal B*, **71**(4), 2009 pp. 623–630. doi:10.1140/EPJB/E2009-00335-8.

[22] S. Rafiee, C. Salavati and A. Abdollahpouri, "CNDP: Link Prediction Based on Common Neighbors Degree Penalization," *Physica A: Statistical Mechanics and Its Applications*, **539**, 2020 122950. doi:10.1016/j.physa.2019.122950.

[23] L. Katz, "A New Status Index Derived from Sociometric Analysis," *Psychometrika*, **18**(1), 1953 pp. 39–43. doi:10.1007/BF02289026.

[24] W. Liu and L. Lü, "Link Prediction Based on Local Random Walk," *Europhysics Letters*, **89**(5), 2010 58007. doi:10.1209/0295-5075/89/58007.

[25] F. Xia, J. Liu, H. Nie, Y. Fu, L. Wan and X. Kong, "Random Walks: A Review of Algorithms and Applications," *IEEE Transactions on Emerging Topics in Computational Intelligence*, **4**(2), 2019 pp. 95–107. doi:10.1109/TETCI.2019.2952908.

[26] C. V. Cannistraci, G. Alanis-Lobato and T. Ravasi, "From Link-Prediction in Brain Connectomes and Protein Interactomes to the Local-Community-Paradigm in Complex Networks," *Scientific Reports*, **3**(1), 2013 1613. doi:10/1038/srep01613.

[27] J. Yang and X.-D. Zhang, "Predicting Missing Links in Complex Networks Based on Common Neighbors and Distance," *Scientific Reports*, **6**(1), 2016 38208. doi:10.1038/srep38208.

[28] I. Ahmad, M. U. Akhtar, S. Noor and A. Shahnaz, "Missing Link Prediction Using Common Neighbor and Centrality Based Parameterized Algorithm," *Scientific Reports*, **10**(1), 2020 364. doi:10.1038/s41598-019-57304-y.

[29] Z. Liu, Q.-M. Zhang, L. Lü and T. Zhou, "Link Prediction in Complex Networks: A Local Naïve Bayes Model," *Europhysics Letters*, **96**(4), 2011 48007. doi:10.1209/0295-5075/96/48007.

[30] F. Tan, Y. Xia and B. Zhu, "Link Prediction in Complex Networks: A Mutual Information Perspective," *PloS One*, **9**(9), 2014 e107056. doi:10.1371/journal.pone.0107056.

[31] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera and A. Califano, "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context," *BMC Bioinformatics*, 7(1), 2006 S7. doi:10.1186/1471-2105-7-S1-S7.

[32] Y. Deng, H. Zenil, J. Tegnér and N. A. Kiani, "HiDi: An Efficient Reverse Engineering Schema for Large-Scale Dynamic Regulatory Network Reconstruction Using Adaptive Differentiation," *Bioinformatics*, **33**(24), 2017 pp. 3964–3972. doi:10.1093/bioinformatics/btx501.

[33] D. Mohdeb, A. Boubetra and M. Charikhi, "Tie Persistence in Academic Social Networks," *Informatica*, **40**(3), 2016 1225. www.informatica.si/index.php/informatica/article/view/1225.

[34] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai and A.-L. Barabasi, "Hierarchical Organization of Modularity in Metabolic Networks," *Science*, **297**(5586), 2002 pp. 1551–1555. doi:10.1126/science.1073374.

[35] H. N. Chua, W.-K. Sung and L. Wong, "Exploiting Indirect Neighbours and Topological Weight to Predict Protein Function from Protein–Protein Interactions," *Bioinformatics*, **22**(13), 2006 pp. 1623–1630. doi:10.1093/bioinformatics/btl145.

[36] M. E. J. Newman, "The Structure of Scientific Collaboration Networks," *Proceedings of the National Academy of Sciences*, **98**(2), 2001 pp. 404–409. doi:10.1073/pnas.98.2.404.

[37] A. Samad, M. Qadir and I. Nawaz, "SAM: A Similarity Measure for Link Prediction in Social Network," in *13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS '2019)*, Karachi, Pakistan, IEEE, 2019 pp. 1–9. doi:10.1109/MACS48846.2019.9024762.

[38] D. Mumin, L.-L. Shi and L. Liu, "An Efficient Algorithm for Link Prediction Based on Local Information: Considering the Effect of Node Degree," *Concurrency and Computation: Practice and Experience*, **34**(7), 2022 e6289. doi:10.1002/cpe.6289.

[39] H. Tong, C. Faloutsos and J.-Y. Pan, "Random Walk with Restart: Fast Solutions and Applications," *Knowledge and Information Systems*, **14**(3), 2008 pp. 327–346. doi:10.1007/s10115-007-0094-2.

[40] L. Page, S. Brin, R. Motwani and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web." Technical Report, Stanford InfoLab, 1999.

[41] J. A. Hanley and B. J. McNeil, "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, **143**(1), 1982 pp. 29–36. doi:10.1148/radiology.143.1.7063747.

[42] Y.-N. Ye, Z.-G. Hua, J. Huang, N. Rao and F.-B. Guo, "CEG: A Database of Essential Gene Clusters," *BMC Genomics*, **14**(1), 2013 769. doi:10.1186/1471-2164-14-769.

[43] V. Batagelj and A. Mrvar, "Analysis of Large Networks with Pajek," in *Proceedings of Pajek Workshop at XXVI Sunbelt Conference*, 2006. api.semanticscholar.org/CorpusID:61725297.

[44] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt and A. Arenas, "Self-Similar Community Structure in a Network of Human Interactions," *Physical Review E*, **68**(6), 2003 065103(R). doi:10.1103/PhysRevE.68.065103.

[45] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton and W. V. den Broeck, "What's in a Crowd? Analysis of Face-to-Face Behavioral Networks," *Journal of Theoretical Biology*, **271**(1), 2011 pp. 166–180. doi:10.1016/j.jtbi.2010.11.033.

[46] M. Heidari, M. Asadpour and H. Faili, "SMG: Fast Scalable Greedy Algorithm for Influence Maximization in Social Networks," *Physica A: Statistical Mechanics and Its Applications*, **420**, 2015 pp. 124–133. doi:10.1016/j.physa.2014.10.088.

[47] J. Kunegis, "KONECT: The Koblenz Network Collection," in *Proceedings of the 22nd International Conference on World Wide Web*, Rio de Janeiro, Brazil, New York: Association for Computing Machinery, 2013 pp. 1343–1350. doi:10.1145/2487788.2488173.