Contents lists available at ScienceDirect

Physica A

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

CNDP: Link prediction based on common neighbors degree penalization

Samira Rafiee [*], Chiman Salavati, Alireza Abdollahpouri

Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran

ARTICLE INFO

Article history:
Received 10 January 2019
Received in revised form 3 June 2019
Available online 26 September 2019

Keywords:
Link prediction
Network topology
Similarity metric
Social networks

ABSTRACT

In social network analysis, link prediction is a fundamental tool to determine new relationships among users which are most likely to occur in the future. Link prediction by means of a similarity metric is common in which a pair of similar nodes is likely to be connected. In this paper, we propose a similarity-based link prediction algorithm, referred to as CNDP, which similarity score is determined according to the structure and specific characteristics of the network, as well as the topological characteristics. In the proposed method, a new metric for link prediction is introduced, considering clustering coefficient as a structural property of the network. Moreover, the presented method considers the neighbors of shared neighbors in addition to only shared neighbors of each pair of nodes, which leads to achieve better performance than other similar link prediction methods. The empirical results of evaluation on synthetic and real-world networks demonstrate that the proposed algorithm achieves higher accuracy prediction results with lower complexity, and performs superior compared to other algorithms.

1. Introduction

With the rapid spread of the amount of information in social networks, accurately predict a potential link become an important and challenging problem in many domains, such as recommender systems, decision making and criminal investigations [1]. Link prediction involves computing the likelihood of the missing or future links among nodes in a network [2,3]. In order to accurately define the link prediction problem, assume that G is an undirected graph $G = (V, E)$, where V is a set of labeled nodes and E is a set of edges between pairs of entities from set V. Considering a snapshot of the network G at time t, the link prediction problem involves defining the missing subsets in the current snapshot, which is formed at time $t + \Delta$ [4]. Link prediction problem is confronted with two main challenges. The first one is the massive data, which requires the prediction approaches with low complexity; and the second challenge is that the prediction approaches involving high prediction accuracy. Traditional data mining approaches, however, ignore the relationships between entities and are incapable to solve the link prediction problem efficiently.

There are several studies that follow different approaches to deal with the link prediction problem [2,5], where most of them are based on the similarity between nodes [2,6]. In similarity-based techniques links are more likely to form between similar nodes. Moreover, there are some link prediction methods that considered nodes that share a higher number of neighbors in the link prediction processes [7]. In order to assigns a similarity score to every pair of nodes in the network in these methods, first a function $s(x, y)$ is defined. Different features such as topological features and particular features of

[*] Corresponding author.
E-mail addresses: samira.rafiee@eng.uok.ac.ir (S. Rafiee), chiman.salavati@eng.uok.ac.ir (C. Salavati), abdollahpouri@uok.ac.ir (A. Abdollahpouri).

https://doi.org/10.1016/j.physa.2019.122950
0378-4371/© 2019 Published by Elsevier B.V.

the network are taken into account in the similarity score. Then, all pairs of nodes in the network are ranked in decreasing order of their scores, and links with the highest rank are suggested as the foreseen links in the list of lost links.

Similarity-based methods can be divided into three categories: local, global and quasi-local, according to the amount of information that they take into account in the process of computing the similarity function [2,6,8]. In local techniques, more attention is paid to the direct neighbor information. These techniques are appropriate for huge networks owing to have high precision versus a linear time complexity, but suffer from low accuracy. Global techniques by using the entire network topology are capable to compute the similarity between each pair of nodes and are not limited to the shared neighbors of nodes. However, global approaches have lower precision in comparison with local approaches due to sensitivity to the noise and involving high computational complexity. These two categories, regardless of the network they are applied, always work in the same way. Quasi-local techniques seek to take the advantages of both local and global techniques, by considering neighbors of neighbors instead of only direct neighbors, and limiting the distance between each pair of nodes. These methods can find a balance between the amount of considered information and the computational complexity. There are two major issues in the link prediction scope. The first issue is introducing a link prediction approach with low computational complexity when it is particularly confronted with huge datasets. The second one is the prediction accuracy so that high-accuracy link prediction approaches are desirable. Thus, two main challenges in link prediction are time complexity and accuracy. Since traditional link prediction approaches are not able to solve the problem efficiently, new approaches have been proposed in this area. However, an approach which can achieve low complexity and also high accuracy at the same time has not yet been suggested. In this paper, a similarity-based link prediction method is proposed based on common neighbors degree penalization, namely *CNDP*, in which the similarity score is determined according to the topological characteristics including common neighbors of each two nodes and the average clustering coefficient of the network. The main difference between presented method and other similar algorithms is that it distinguishes between the common neighbors of the nodes. The presented method is a quasi-local similarity method that benefits both the local and global features, simultaneously. The empirical results in the experiments reveal the superiority of the proposed method to the similar methods in terms of accuracy and computational complexity.

The rest of the paper is organized as follows. Section 2 provides related works and then some required background including adaptive degree penalization (ADP) and clustering coefficient are explained briefly in Section 3. Our proposed CNDP method is introduced and explained in detail in Section 4. The performance of our method is discussed and compared with some other similar methods in Section 5. Finally, we conclude the paper in Section 6.

2. Related works

The most well-known similarity-based approaches in the link prediction area are Common Neighbor (CN) [9], Adamaic Adar (AA) [10], and Resource Allocation (RA) [11]. The similarity score given by CN to each pair of nodes involves the number of shared neighbors between these nodes. CN assume that, if two nodes have many shared neighbors, the probability of formation edge among them will be increased. There are several works that have confirmed the strong correlation between the number of shared neighbors of two nodes and the likelihood of forming a link between them [12]. If Γ_x is the set of neighbors of node x and $|\Gamma_x|$ is the number of neighbors of node x, CN is formulated according to Eq. (1).

$$S_{CN}(x, y) = |\Gamma_x \cap \Gamma_y| \tag{1}$$

AA method panelizes each shared neighbor by its degree and assigns a similarity score to nodes by investigating the common neighbors between each two nodes as Eq. (2).

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

RA considers the distribution of resources between two non-connected nodes x and y through their neighbors, so that each neighbor node receives some resources and distributes them equally among its neighbors. Similarity criterion between two nodes x and y can be the amount of resources received from node x from node y through their shared neighbors. It can be defined as Eq. (3).

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

Preferential Attachment Index [13], Jaccard Index [14], Sørensen Index [15], Hub-Promoted and Hub-Depressed Indices [16], and Leicht–Holme–Newman Index [8] are the other similarity-based measures employed in the link prediction. There are several similarity based methods in the link prediction scope in which the probability of being links between two nodes is determined based on their shared neighbors. Yao, Lin, et al. [17] proposed a link prediction algorithm based on common-neighbors for dynamic social network in which a weight is assigned to all the edges between two nodes, using three particular metrics, and then sum of these weights determines the probability of being a link between the mentioned nodes. These metrics are defined according to the number of edges, Euclidean distance, and the shared neighbors between two nodes. Li et al. in [1] proposed node-coupling clustering coefficient of node n, referred to as NCCCN. In this work, the portion of common neighbors between two nodes is combined with the information achieved from clustering. In this method, clustering coefficient of common neighbors that is the same for every node is applied. Authors in [18] consider the lower bound of node pairs' similarity scores. They utilized a parallel computing scheme to obtain all node pairs with CN values larger than the lower bound which declines the computational time. Dong et al. [19], established a bi-scale link prediction on networks which combines the information of micro-scale (neighbors) and meso-scale (communities). Triadic similarity measure [20] is a recently published neighborhood-based similarity index, which considers the structural units of the network called motifs. Motifs are 13 different forms of small networks consist of three nodes in directed networks. In this metric, for determining the similarity score between two nodes x and y for each common neighbor z, the number of motifs which are constructed by x, y and z are divided into 13, and then, the result obtained for all common neighbors are gathered together. Afterwards, the outcome is divided into the number of these common neighbors, making a similarity score to the pair of nodes x and y.

Chuan et al. [21] proposed a metric, referred to as LDAcosin, for link prediction in the co-authorship network based on the content similarity. This method presents mathematical notions of the link prediction in the co-authorship network and a link prediction algorithm based on topic modeling. Authors in [22] prove the range of the interval of the classical transferring similarity. They present a specific state to denote the degree of similarity using fuzzy system theory. This method is able to combine the transferring similarity of different intermediates using information fusion approach. Two-Phase Selection Link Prediction for Vertex in Graph Streams (TPLP) [23], aims to predict the top-k vertices, i.e., the top-k links that are most likely to connect to the target vertex in graph streams. Bütün et al. [24] introduced a temporal link prediction metric, taking advantage of both local and global topological structures in complex networks. They formulated the predict citation count of scientists (PCCS) problem as a link prediction problem in citation networks in which link prediction approach predicts links and the weight of links by employing the temporal link prediction metric. Influential links prediction (ILP) [25] predicts links to users that could apply social influence on them. To this end, an influence maximization algorithm is used to determine a set of possible influential users from the set of current influential users of the target [26–29]. Bastami et al. [30] proposed an unsupervised gravitation-based link prediction approach to accuracy improvement of local and global predictions using combination of node features, community information and graph characteristics. Moreover, authors in [31] stress on network features or integrating place features by taking a different perspective. This work focuses on user, hemophilic and locational features and integrating them in order to evaluate their impact and efficacy.

3. Adaptive degree penalization (ADP)

This section provides a background of the two core concepts of this article: adaptive degree penalization (ADP) [4] and clustering coefficient [32].

3.1. Adaptive degree penalization (ADP)

It will be more acceptable if a link prediction method has the accurate results on the networks by different structural characteristics on various domains. In adaptive degree penalization (ADP) [4], the degree of the common neighbors is penalized according to the clustering coefficient of the network. In general, CN, AA, and RA can be defined as Eq. (4).

$$S(x,y)=\sum_{z\in\Gamma_x\cap\Gamma_y}|\Gamma_z|^{-\alpha} \tag{4}$$

where, α is a constant value, z is a common neighbor between x and y, and $|\Gamma_z|$ is the degree of z. The only difference between CN, AA and RA is the value of α. For example, this parameter in CN is set to 0, in RA is 1 and in AA, α is in range [0,1]. In ADP, a general value is considered for parameter α. For this purpose, the relevance of α by the shortest path between nodes and average clustering coefficient are considered, and it is resulted that there is a strong correlation between α and clustering coefficient, and as a consequence, the probability of being a link between two nodes x and y is formulated as Eq. (5).

$$S_{ADP}(x,y)=\sum_{z\in\Gamma_x\cap\Gamma_y}|\Gamma_z|^{-\beta C} \tag{5}$$

where C is the average clustering coefficient (see Section 3.2), and β is a constant value. In the ADP, average value 2.5 is set for β based on a set of heterogeneous networks. The number of shared neighbors between each pair of nodes is panelized, regarding the clustering coefficient as a structural property of the network. ADP is performed on a variety of networks and achieved good performance.

3.2. Clustering coefficient

There are several structural properties in a network. Some of these features are the shortest path between nodes, information entropy of the paths between nodes, diameter of the network which is the longest shortest path in the network, and clustering coefficient of the nodes. Clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. In a social network, the clustering coefficient would measure the tendency of the friends of a given user to be also friends among themselves. Average clustering coefficient is a constant value for the whole network. In order to obtain average clustering coefficient, the clustering coefficient is computed for each node in the network. Therefore, clustering coefficient for node x is computed according to Eq. (6).

$$C_x = \frac{\sum_{y,z\in \Gamma_x \text{ and } E_{y,z}\in E} E_{y,z}}{|\Gamma_x|(|\Gamma_x|-1)} \tag{6}$$

Therefore, average clustering coefficient is defined as Eq. (7).

$$\bar{C} = \frac{\sum_{x\in V} C_x}{|V|} \tag{7}$$

4. Proposed method

Similarity based link prediction methods have the same framework, and the similarity index among nodes is the only difference between different methods. The main purpose of them is to present a more accurate index to estimate the probability of link existence between nodes in the network. This index is a similarity score among every pair of nodes. There is no doubt that the probability of existence a link between every two nodes is dependent on the number of common neighbors between them. Other methods, such as CN, AA, and RA penalize common neighbors with a constant value. They are unaware of the fact that degree of penalization must be different according to the characteristics and structure of the network. On the other hand, ADP method utilizes the average clustering coefficient in the similarity index formula properly to pay attention to the characteristics and network structure. However, the major weakness of this method is the lack of attention to the form of common neighbors. In order to overcome to this challenge, the proposed method looks at the neighbors with a new perspective. The proposed method has improved the ADP method by distinguishing between the common neighbors. For example, assume that Fig. 1 is a small part of a real network and we want to compute the similarity score between node x and node y. According to the figure, there are 3 common neighbors z_1, z_2, z_4 for two nodes x and y. However, the main point is that there is a major difference between z_4, which is not a neighbor for any of the other common neighbors, and z_1, z_2. In other words, there is a link between z_1 and z_2 but z_4 involves no relationships with the other common neighbors. This difference is considered in our metric in order to achieve a higher efficiency in link prediction. This point is also evident in friendship relationship when we want to predict the likelihood of friendship between x and y. If x and y have friends who are already friend, the probability of friendship between x and y in the future will be greater than the case when x and y have common friends who are not friend. It should be noted that when the number of common neighbors z, which are themselves a common neighbor, increases, the accuracy and efficiency of the link prediction method improved. Therefore, a similarity based link prediction method is proposed which considers the shared neighbors in different way. Moreover, this metric is capable to adapt by the structure of network by considering $-\beta$C. The presented method is formulated as Eq. (8).

$$S(x,y) = \sum_{z\in \Gamma_x\cap\Gamma_y} |C_z|\,(|\Gamma_z|)^{-\beta C} \tag{8}$$

where z is a common neighbor for two nodes x and y, $|C_z|$ is the number of neighbors of z which consist of the common neighbors of x and y in addition to x and y. $|\Gamma_z|$ is the number of neighbors of z, C is the average clustering coefficient, and β is a constant value which is tuned in the experimental section.

The proposed method consists of six steps, which the pseudo-code is illustrated in Algorithm 1. In step 1, the clustering coefficient is computed for each node, and then the average clustering coefficient is obtained. Step 2 involves of dividing the network into two sets: train and test. Dividing the network is done in two ways. In the first way, only 10 percent of the total edges from the general network are randomly assigned to the test set, and the remained 90 percent are assigned to the train set. In the second way, 5-fold cross validation is employed so that the total edges of the original network is divided into five equal sections and each section is considered as test and train set, respectively. In this step, test set is consisted of nonexistent edges plus non-observation edges, which the similarity score for each of them is computed in step 3. In the next step, the similarity scores are arranged in descending order and then the edges from arranged list is added to the train list. In step 5, exactly the same number of edges as the test set and subtracted arbitrarily from the main network should be added to the edges of the train network (G_{train}). These added edges are the predicted edges. Finally, in step 6, the number of true positives (correctly predicted) and false positives (incorrectly predicted) are determined. The number of correctly predicted edges is the number of edges that exist in both test set and predicted sets, and the number of incorrectly predicted edges is the difference of the number of edges in test set and the number of correctly predicted edges. Consequently, precision can easily be calculated. For more comprehension, an example of a toy network is shown

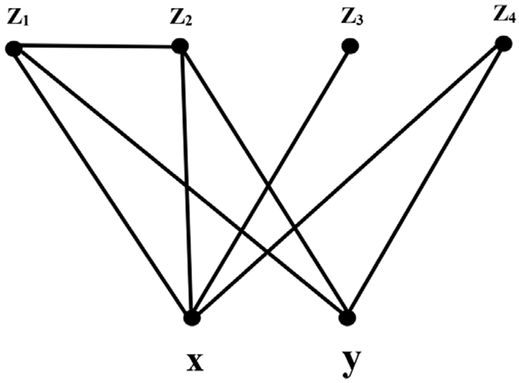


Fig. 1. The representation of similarity between x and y.

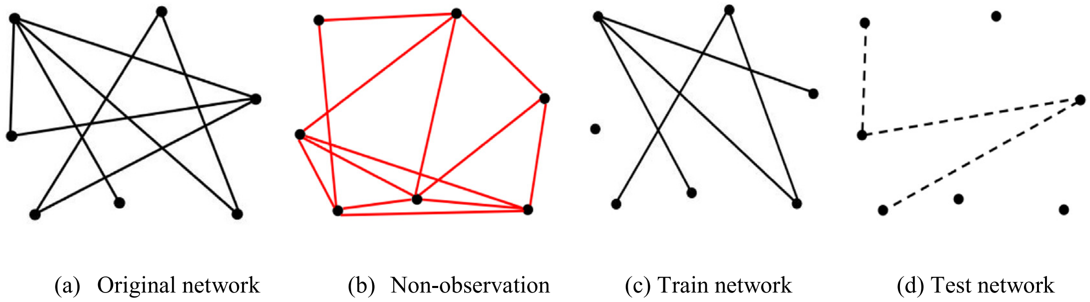
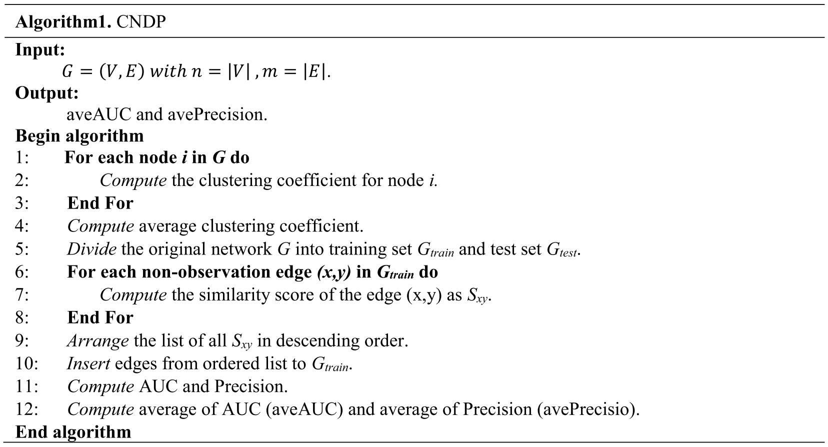

(a) Original network (b) Non-observation (c) Train network (d) Test network

Fig. 2. An example of a toy network.

in Fig. 2. In this sample, Fig. 2(a) is the original network, Fig. 2(b) is non-observation edges in red lines, Fig. 2(c) is G_{train} that depicts by black lines, and Fig. 2(d) is G_{test} that is shown with dashed lines. Therefore, the total disappeared edges include dashed lines and red lines.


Algorithm1. CNDP

```
Input:
        $G = (V, E)\ with\ n = |V|, m = |E|$.
Output:
        aveAUC and avePrecision.
Begin algorithm
    For each node i in G do
        Compute the clustering coefficient for node i.
    End For
    Compute average clustering coefficient.
    Divide the original network G into training set $G_{train}$ and test set $G_{test}$.
    For each non-observation edge (x,y) in $G_{train}$ do
        Compute the similarity score of the edge (x,y) as $S_{xy}$.
    End For
    Arrange the list of all $S_{xy}$ in descending order.
    Insert edges from ordered list to $G_{train}$.
    Compute AUC and Precision.
    Compute average of AUC (aveAUC) and average of Precision (avePrecisio).
End algorithm
```

5. Experimental analysis

In this section, some experiments are performed on the proposed link prediction method (CNDP), and the results of the algorithm are reported and analyzed. For evaluating the performance of the proposed method, it is compared to state-of-the-art link prediction methods, including Triadic measure [20], adaptive degree penalization (ADP) [4] and node-coupling clustering coefficient of node (NCCCN) [1]. Details on the experiments, such as the selected datasets, the evaluation criteria, the determination of parameter β, and the numerical results are explained below.

Table 1
The basic statistical properties of the real and artificial networks.

Network	$\|V\|$	$\|E\|$	$\langle K\rangle$	$\langle C\rangle$	ASPL	D	H
Data1	100	1 002	10	0.11	2.23	4	1.090
Data2	500	5 032	10	0.02	2.93	5	1.094
Data3	1 500	12 088	8	0.005	3.74	7	1.124
BUP	105	441	8.4	0.49	3.08	7	1.42
CEG	297	2 148	14.46	0.29	2.46	5	1.80
UAL	332	2 126	12.81	0.63	2.74	6	3.46
INF	410	2 765	13.49	0.46	3.63	9	1.39
SMG	1 024	4 916	9.6	0.31	2.98	6	3.95
EML	1 133	5 451	9.62	0.22	3.61	8	1.94
NSC	1 461	2 742	3.75	0.69	2.59	17	1.85
HMT	2 426	16 630	13.71	0.54	3.15	10	3.10
ADV	5 155	39 285	15.24	0.25	3.22	9	5.41
LDG	8 324	41 532	9.98	0.31	4.37	16	6.189
PGP	10 680	24 316	4.55	0.27	7.49	24	4.147

5.1. Datasets

The proposed method is evaluated on both real-world and synthetic networks. The synthetic networks generated by Girvan Newman [33] include four networks. One of these networks are aimed to evaluate the correctness of the presented method, and the other ones, namely Data1, Data2, and Data3, are utilized for evaluating the efficiently of the proposed method on networks with different sizes and various clustering coefficients. Furthermore, the presented method is investigated on eleven real networks which are different in terms of topological properties such as the number of nodes, edges and average clustering coefficient. The properties of the real and synthetic networks are shown in Table 1. Not that the synthetic network which are used in order to approve the correctness of the proposed method is too small and thus its properties are not provided in this table. Columns from left to right of Table 1 include: network name, number of nodes ($|V|$), number of edges ($|E|$), average degree ($\langle K\rangle$), average clustering coefficient ($\langle C\rangle$), average shortest path length (ASPL), diameter (D), and heterogeneity (H).

Real networks are collected from different sources and domains. BUP [12] is a network of political blogs. CEG [32] is a biological network. UAL [34] is an airport traffic network. INF [7] is a network of face-to-face contacts in an exhibition. SMG [34], NSC [35], and LDG [34] are co-authorship networks for different fields of study. EML [5] is a network of users who shared emails. HMT [36] and ADV [37] are social networks. Finally, PGP [38] is an interaction network of users of the Pretty Good Privacy algorithm. These networks are available at http://noesis.ikor.org/datasets/link-prediction.

5.2. The evaluation criteria

In this section, two evaluation criteria are employed for evaluating the performance of the proposed link prediction method in comparison with other methods. The first criterion is area under the receiver operating characteristic curve or area under the curve (AUC). The AUC is the probability of being higher similarity score of an edge selected from the test set in relation to the similarity score of an edge selected from the nonexistent set. In order to calculate AUC, a link from the test set and a link from the nonexistent set are selected, and this process is repeated for several times. In each iteration, the score of two links is compared. The formula for calculating the AUC is according to (9).

$$AUC = \frac{n_1 + 0.5n_2}{n} \tag{9}$$

where, n is the number of comparisons, n_1 is the number of times that score of the link chosen from the test set is more than the other link, and n_2 is the number of times that both links have the same score. Link selection is performed in two ways. In the first way, selection can be randomly so that the number of times that a link is selected (n) is large. In the second way, every pair of links of two sets must be compared. The number of selected edges in this case is equal to the multiplication of the number of test set members and the number of members of the nonexistent set. In this paper, the second way is applied. At random case, the AUC is approximately 0.5; therefore, if AUC is larger than 0.5, the performance of the algorithm will be significantly better than other randomly cases.

Precision is another effective means for validation of link prediction, which is used in this article to evaluate the accuracy achieved by various methods. Precision is the percentage of correctly predicted links, and it is computed as Eq. (10).

$$precision = \frac{A}{T} \tag{10}$$

where, A is the number of correctly predicted links, and T is the total number of predicted links. A correctly predicted link is a link which belongs to both predicted set and test set. The total predicted links are a collection of correctly predicted links and incorrectly predicted links.

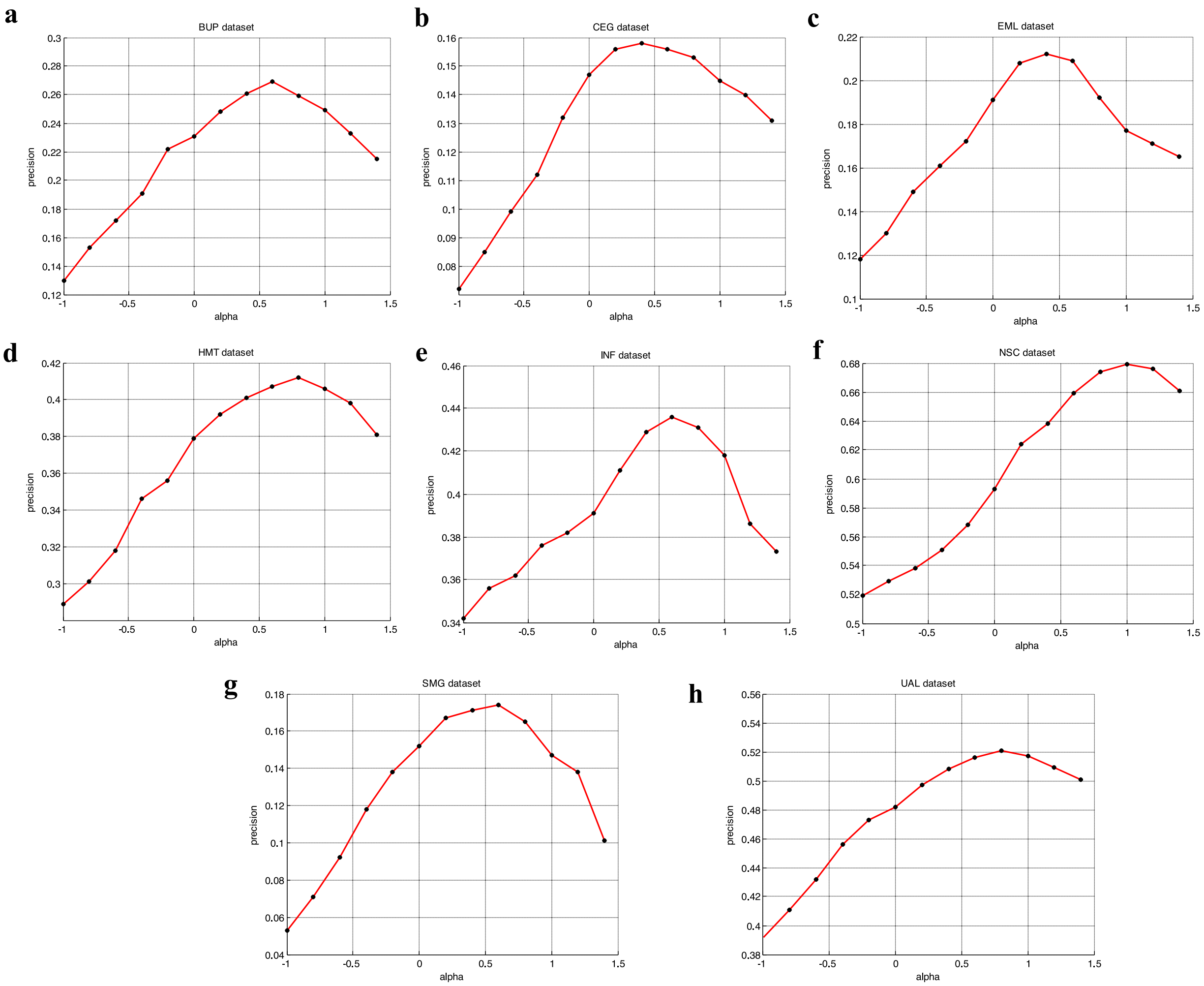


Fig. 3. Precision obtained by our link predictor varying the alpha parameter for different networks.

5.3. Determination of parameter β

As mentioned previously, β is a constant parameter that its value affects the probabilistic of existing link between two nodes and the performance of the proposed method depends on this parameter in some extent. Therefore, considering the proposed index for computing the likelihood of link existence between two nodes, different values is set for β. In this paper, trial and error method is utilized. First, the evaluation criteria are applied based on α for each network, and then the best-performing α value is obtained for different networks. Multiple values of α in range $[-1, 1.5]$ are considered, and for each network a best-performing α value is achieved (Figs. 3 and 4). Then, a linear regression between the clustering coefficients of the networks and the best-performing alpha values is performed to determine β value (Figs. 5 and 6). Therefore, for precision measurement, the value of β is obtained equal to $\beta = 1.84$ and for AUC, the value of $\beta = 1.76$.

5.4. Experimental results

In the experimental results section, at first, 10 percent of links is selected randomly and removed from the network. This selection is performed five times for more precise results and avoiding random behavior of the algorithm. Next, 5-fold cross-validation method is applied so that the network is divided into five equal sections, each time a section is considered as a test or credit set and the remainder as the train set. To ensure the correctness of the algorithm, it is conducted on a small synthetic network which is illustrated in Fig. 7.

The network shown In Fig. 7 has 14 edges. Through the application of the first criterion, selecting 10 percent of the edges, 1 edge is considered as test set. Therefore, the test set contains only 1 edge. This edge is removed from the network. After applying the proposed link prediction method on the network, the most similarity score is given to the mentioned

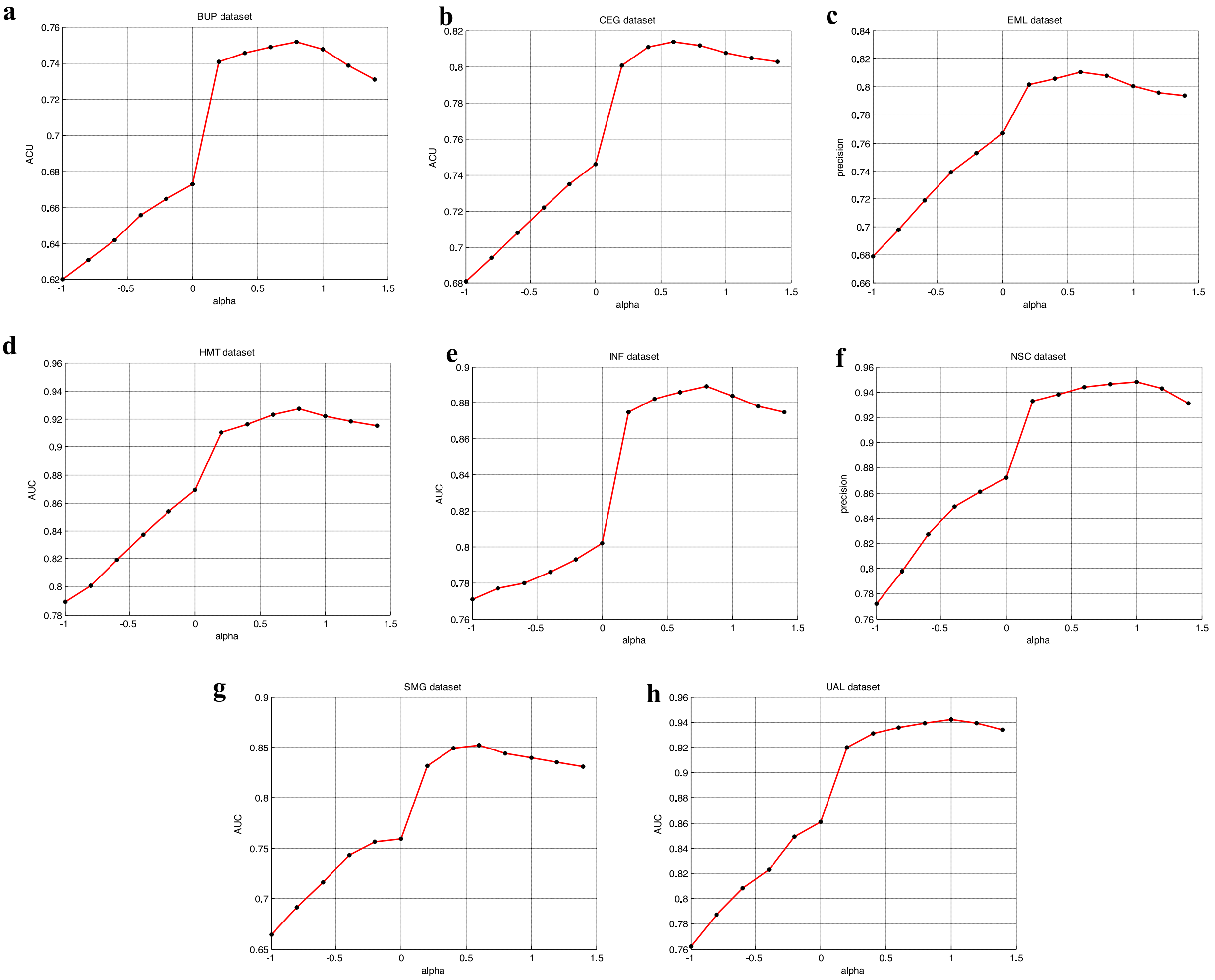


Fig. 4. AUC obtained by our link predictor varying the alpha parameter for different networks.

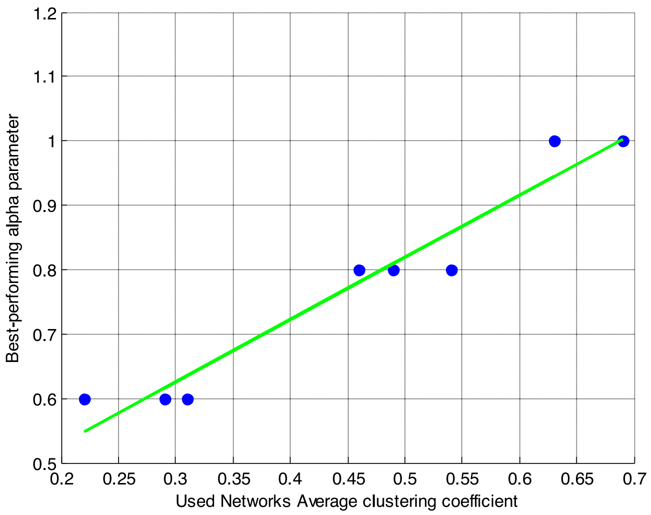


Fig. 5. Average clustering coefficient shown against the best-performing alpha value according to precision for the networks used in our experiments.

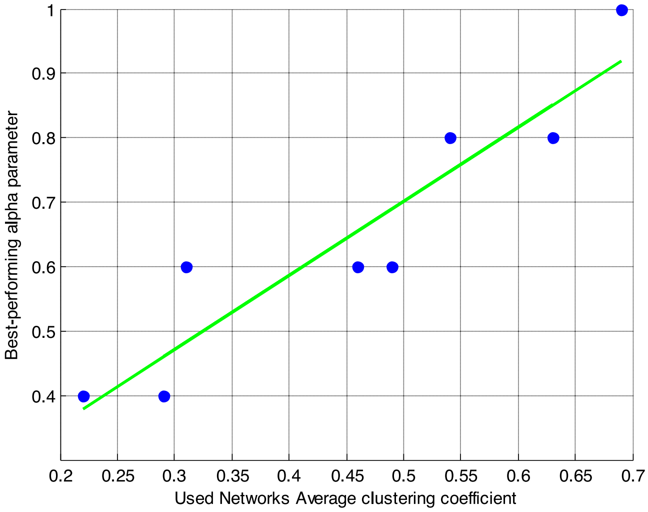


Fig. 6. Average clustering coefficient shown against the best-performing alpha value according to AUC for the networks used in our experiments.

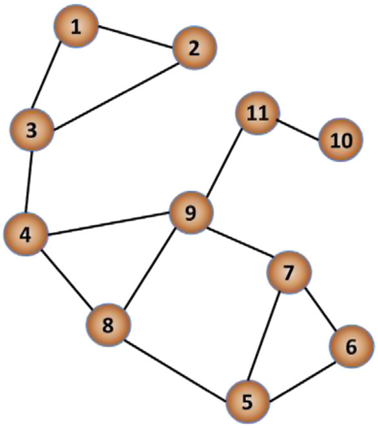


Fig. 7. A synthetic network.

edge by the algorithm, and in fact it can be considered as a predicted edge. In this case, the precision value is 1 as the ratio of the correct predicted edges respect to the total number of predicted edges, which is precisely the same with the correct prediction of the edge. Furthermore, the achieved value for AUC is 0.939 which is an acceptable result. According to the precision and AUC values and regarding the edge that the proposed method provides as outcome, it can be concluded that the proposed method is able to predict the missed link correctly.

The performance of the proposed method is compared with ADP and NCCCN, and Triadic measure. These methods are based on similarity and compared to prior methods have had better performance. ADP in comparison with CN, AA and RA has significant results. NCCCN, similar to the proposed method, has focused on the neighbors of z and the clustering coefficient of the network. Triadic measure is a recently published work based on neighborhood and similarity for directed networks. Since the datasets employed in this paper are undirected, Triadic measure is exerted by considering both directions between each pair of nodes that are already connected with an edge, and its results are then compared with the proposed method. The empirical experiments reveal the efficiency of the presented metric to the others. In the following, the results of the proposed algorithm and the other three algorithms on real and synthetic networks are shown in Tables 2–5.

Table 2 illustrates the performance of different algorithms by the selection of 10 percent of edges of the eleven real-world networks. It is observed that CNDP has the best performance in all states, compared with the other methods except for ADP on the NSC network and NCCCN on the HMT network. Table 3 illustrates the results of the proposed method and other algorithms on synthetic networks with various structures. As can be seen from the numerical results, our metric has also a good performance encountering with networks with different characteristics. Furthermore, Tables 4 and 5 demonstrate the performance of different algorithms by 5-fold cross-validation on different real and synthetic networks.

Table 2
The performance of different algorithms by the selection of 10 percent of edges on eleven real-world networks.

Algorithm	Evaluation metric	BUP	CEG	UAL	INF	SMG	EML	NSC	HMT	ADV	LDG	PGP
ADP	Precision	0.367	0.15	0.523	0.429	0.152	0.187	0.672	0.409	0.18	0.169	0.367
	AUC	0.925	0.799	0.944	0.858	0.839	0.790	**0.959**	0.987	**0.88**	0.914	0.925
NCCCN	Precision	0.354	0.152	0.52	0.419	0.149	0.192	0.665	0.419	0.170	0.161	0.354
	AUC	0.927	0.785	0.936	0.85	0.828	0.811	0.951	**0.992**	0.857	0.894	0.927
Triadic measure	Precision	0.332	0.129	0.499	0.409	0.130	0.169	0.639	0.382	0.159	0.137	0.332
	AUC	0.900	0.769	0.927	0.829	0.798	0.759	0.873	0.928	0.844	0.864	0.900
CNDP	Precision	0.375	0.158	0.525	0.433	0.168	0.207	0.679	0.422	0.186	0.174	0.375
	AUC	**0.936**	**0.811**	**0.946**	**0.878**	**0.842**	**0.832**	0.954	0.989	**0.88**	**0.918**	**0.936**

Table 3
The performance of different algorithms by the selection of 10 percent of edges on synthetic networks.

Algorithm	Evaluation metric	Data1	Data2	Data3
ADP	Precision	0.1753	0.4126	0.2306
	AUC	0.4635	0.7832	0.6911
NCCCN	Precision	0.1487	0.3878	0.2017
	AUC	0.4089	0.7529	0.6492
Triadic measure	Precision	0.1168	0.3541	0.1783
	AUC	0.3982	0.7106	0.6054
CNDP	Precision	0.2419	0.4365	0.2530
	AUC	**0.4821**	**0.8031**	**0.7122**

Table 4
The performance of different algorithms by 5-fold cross-validation on eleven real-world networks.

Algorithm	Evaluation metric	BUP	CEG	UAL	INF	SMG	EML	NSC	HMT	ADV	LDG	PGP
ADP	Precision	0.254	0.153	0.515	0.419	0.159	0.197	0.667	0.396	0.186	0.173	0.376
	AUC	0.749	0.800	0.931	0.869	**0.833**	0.797	0.937	0.911	0.873	0.909	0.932
NCCCN	Precision	0.242	0.160	0.502	0.393	0.148	0.211	0.653	0.409	0.178	0.164	0.368
	AUC	0.753	0.781	0.919	0.879	0.812	0.802	0.932	0.899	0.864	0.889	0.921
Triadic measure	Precision	0.229	0.147	0.483	0.364	0.136	0.173	0.635		0.156	0.143	0.329
	AUC	0.733	0.734	0.893	0.483	0.782	0.785	0.876	0.865	0.838	0.864	0.899
CNDP	Precision	0.265	0.162	0.522	0.431	0.17	0.213	0.672	0.412	0.197	0.194	0.382
	AUC	**0.755**	**0.812**	**0.941**	**0.886**	0.832	**0.810**	**0.947**	**0.925**	**0.885**	**0.921**	**0.942**

Table 5
The performance of different algorithms by 5-fold cross-validation on synthetic networks.

Algorithm	Evaluation metric	Data1	Data2	Data3
ADP	Precision	0.1974	0.4071	0.2479
	AUC	0.4765	0.7916	0.6802
NCCCN	Precision	0.122	0.3958	0.2137
	AUC	0.42	0.7771	0.6659
Triadic measure	Precision	0.104	0.3693	0.1982
	AUC	0.3671	0.7259	0.6147
CNDP	Precision	0.2669	0.4267	0.2634
	AUC	**0.4992**	**0.8248**	**0.7036**

According to the results, CNDP achieved the first rank over all the datasets in terms of precision value in comparison with the other link prediction methods. On the SMG network, ADP has better performance with a slight difference from the proposed method when AUC is considered as metric. According to the findings, only in two comparisons out of 64 comparisons, our method cannot perform better than other methods. These two cases are related to the time when method is executed randomly; however, if the proposed method is repeated more frequently and the results are averaged, it may perform better even in those two cases.

5.5. Computational complexity and execution time

As it is previously mentioned, computational complexity and execution time are crucial challenges in link prediction methods. Thus, a method is superior when it is capable to make a good balance between execution time and efficiency based on evaluation criteria. Overall, all neighborhood-based similarity measures have the same process. The merely

Table 6
Execution time (seconds).

Datasets	ADP	NCCCN	Triadic measure	CNDP
Data1	26	29	22	25
Data2	87	94	82	88
Data3	273	287	256	261
BUP	29	34	23	29
CEG	58	65	51	59
UAL	61	66	57	60
INF	72	76	68	73
SMG	201	219	187	203
EML	231	246	217	230
NSC	281	293	272	282
HMT	472	494	469	471
ADV	858	872	839	859
LDG	1481	1497	1476	1473
PGP	1679	1691	1667	1668
Average	414.92	425.928	406.142	412.928

Table 7
Ranking of methods.

	ADP	NCCCN	Triadic measure	CNDP
Precision	2.73	2.27	1.00	**4.00**
AUC	2.77	2.45	1.00	**3.77**

Table 8
The result of Friedman test for comparison between the algorithms.

	N	Chi-Square	df	*P*-value
Precision	11	30.382	3	0.000
AUC	11	26.284	3	0.000

difference between all of these methods is the procedure of computing similarity. In calculating the CNDP, for each node x, we first search all x's neighbors. The time complexity to traverse the neighborhood of a node is simply $\langle K\rangle$. In the next step, we need to check all neighbors of each of x's second-order neighbors [39,40]. Therefore, the time complexity in calculating the CNDP index is $O\left(n\left\langle K\right\rangle^{2}\right)$, which n is the number of nodes and $\langle K\rangle$ is the average degree. Besides the time complexity, memory space is another limitation for algorithmic implementation confronting with huge networks. In calculating CNDP index, the memory required is of the order $O(n\left\langle K\right\rangle)$, thus it requires relatively less memory and CPU time.

This section of experiments concerns assessment of the runtime of CNDP as compared to other similarity based algorithms over all real and synthetic datasets when 5-fold cross validation is employed. A comparison of the runtimes (in Seconds) of the CNDP method with the other methods is provided in Table 6. Based on these results, the average execution time of the proposed method is approximately equal to ADP, and the execution time of the CNDP and ADP is lower than NCCCN. Meanwhile, it is evident from the achieved results in previous sections that the performance of the proposed CNDP along with the ADP, is better than that of NCCCN. Moreover, although the execution time of the Triadic measure is lower than that of the proposed method and ADP, it cannot be an acceptable achievement due to its inefficiency. Overall, it can be concluded that the main advantage of our approach includes achieving high performance, better than that of similarity-based methods, with very low computational complexity, which significantly reduces run-time in the benchmark datasets.

5.6. Statistical test result

In this section, using the Friedman test [41], the obtained results from different similarity-based link prediction methods will be analyzed. Friedman test is a nonparametric statistical test for finding differences in behavior across multiple approaches. Nonparametric (distribution-free) means the test doesn't assume that the data comes from a particular distribution. Friedman test can be used to evaluate the results of N different methods on K datasets. In this test, the methods are ranked based on their performance criterion. In this paper, the evaluation criteria are precision and AUC and thus the method by higher rank has the best performance. Table 7 shows the ranking of mentioned link prediction methods. As it can be seen, the ranking of the proposed method is higher than other similarity based methods, accounting for 4.00 and 3.77 for precision and AUC, respectively. According to Table 8, the *P*-value is less than 0.05; thus, it has been found that these results are statistically significant.

6. Conclusion

In this paper, a new metric for link prediction is proposed, considering clustering coefficient which is a structural property of the network. On the other hand, the presented method considers the neighbors of shared neighbors in addition to only shared neighbors of each pair of nodes, which leads to achieve better performance than the other similar link prediction method (NCCCN). To verify the efficiency of our method, we conducted a comparative experiment on eleven real-world networks. The advantages of our method can be seen clearly in the experiments. The experimental results, on several networks with different sizes, demonstrated that our scheme obtained considerable results compared to other algorithms.

For future works, we will try to present a new systematic way to improve the proposed method on big data by presenting a method that can be implemented in parallel and can increase efficiency significantly. Additionally, the proposed method can be employed on weighted, directed and also bipartite networks. Moreover, proposing a way to determine an appropriate value for β in the proposed metric in this paper can also be a possible future research direction.

References

[1] F. Li, J. He, G. Huang, Y. Zhang, Y. Shi, R. Zhou, Node-coupling clustering approaches for link prediction, Knowl.-Based Syst. 89 (2015) 669–680.
[2] L. Lü, T. Zhou, Link prediction in complex networks: A survey, Physica A 390 (2011) 1150–1170.
[3] B. Taskar, M.-F. Wong, P. Abbeel, D. Koller, Link prediction in relational data, in: Advances in Neural Information Processing Systems, 2004, pp. 659-666.
[4] V. Martínez, F. Berzal, J.-C. Cubero, Adaptive degree penalization for link prediction, J. Comput. Sci. 13 (2016) 1–9.
[5] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, Topological structure analysis of the protein–protein interaction network in budding yeast, Nucleic Acids Res. 31 (2003) 2443–2450.
[6] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, J. Am. Soc. Inf. Sci. Technol. 58 (2007) 1019–1031.
[7] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, W. Van den Broeck, What's in a crowd? Analysis of face-to-face behavioral networks, J. Theoret. Biol. 271 (2011) 166–180.
[8] E.A. Leicht, P. Holme, M.E. Newman, Vertex similarity in networks, Phys. Rev. E 73 (2006) 026120.
[9] M.E. Newman, Clustering and preferential attachment in growing networks, Phys. Rev. E 64 (2001) 025102.
[10] L.A. Adamic, E. Adar, Friends and neighbors on the web, Social Networks 25 (2003) 211–230.
[11] T. Zhou, L. Lü, Y.-C. Zhang, Predicting missing links via local information, Eur. Phys. J. B 71 (2009) 623–630.
[12] V. Krebs, A network of books about recent US politics sold by the online bookseller amazon. com, Unpublished http://www.orgnet.com, (2008).
[13] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.
[14] P. Jaccard, Étude comparative de la distribution florale dans une portion des Alpes et des Jura, Bull. Soc. Vaudoise Sci. Nat. 37 (1901) 547–579.
[15] T. Sørensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons, Biol. Skr. 5 (1948) 1–34.
[16] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.-L. Barabási, Hierarchical organization of modularity in metabolic networks, Science 297 (2002) 1551–1555.
[17] L. Yao, L. Wang, L. Pan, K. Yao, Link prediction based on common-neighbors for dynamic social network, Procedia Comput. Sci. 83 (2016) 82–89.
[18] W. Cui, C. Pu, Z. Xu, S. Cai, J. Yang, A. Michaelson, Bounded link prediction in very large networks, Physica A 457 (2016) 202–214.
[19] E. Dong, J. Li, Z. Xie, N. Wu, Bi-scale link prediction on networks, Chaos Solitons Fractals 78 (2015) 140–147.
[20] F. Aghabozorgi, M.R. Khayyambashi, A new similarity measure for link prediction based on local structures in social networks, Physica A 501 (2018) 12–23.
[21] P.M. Chuan, M. Ali, T.D. Khang, N. Dey, Link prediction in co-authorship networks based on hybrid content similarity metric, Appl. Intell. 48 (2018) 2470–2486.
[22] L. Yin, Y. Deng, Measuring transferring similarity via local information, Physica A 498 (2018) 102–115.
[23] Y. Xiao, H. Huang, F. Zhao, H. Jin, TPLP: Two-phase selection link prediction for vertex in graph streams, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2019, pp. 514–525.
[24] E. Bütün, M. Kaya, Predicting citation count of scientists as a link prediction problem, IEEE Trans. Cybern. (2019).
[25] A. Monteserin, M.G. Armentano, Influence me! Predicting links to influential users, Inf. Retr. J. 22 (2019) 32–54.
[26] C. Salavati, A. Abdollahpouri, Z. Manbari, BridgeRank: A novel fast centrality measure based on local structure of the network, Physica A 496 (2018) 635–653.
[27] C. Salavati, A. Abdollahpouri, Z. Manbari, Ranking nodes in complex networks based on local structure and improving closeness centrality, Neurocomputing 336 (2019) 36–45.
[28] A. Sheikhahmadi, M.A. Nematbakhsh, A. Zareie, Identification of influential users by neighbors in online social networks, Physica A 486 (2017) 517–534.
[29] A. Zareie, A. Sheikhahmadi, M. Jalili, Influential node ranking in social networks based on neighborhood diversity, Future Gener. Comput. Syst. 94 (2019) 120–129.
[30] E. Bastami, A. Mahabadi, E. Taghizadeh, A gravitation-based link prediction approach in social networks, Swarm Evol. Comput. 44 (2019) 176–186.
[31] L. Eberhard, C. Trattner, M. Atzmueller, Predicting trading interactions in an online marketplace through location-based and online social networks, Inf. Retr. J. 22 (2019) 55–92.
[32] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (1998) 440.
[33] M.E. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2004) 026113.
[34] V. Batagelj, A. Mrvar, Pajek datasets: Reuters terror news network, Online: http://vlado.fmf.uni-lj.si/pub/networks/data/CRA/terror.htm, (2006).
[35] M.E. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (2006) 036104.
[36] J. Kunegis, Hamsterster full network dataset–KONECT, in, 2014.
[37] P. Massa, M. Salvetti, D. Tomasoni, Bowling alone and trust decline in social network sites, in: 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, IEEE, 2009, pp. 658–663.
[38] G. Caldarelli, Large Scale Structure and Dynamics of Complex Networks: From Information Technology to Finance and Natural Science, World Scientific, 2007.
[39] L. Lü, C.-H. Jin, T. Zhou, Similarity index based on local paths for link prediction of complex networks, Phys. Rev. E 80 (2009) 046122.
[40] F. Gao, K. Musial, C. Cooper, S. Tsoka, Link prediction methods and their accuracy for different social networks and network metrics, Sci. Program. 2015 (2015) 1.
[41] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Ann. Math. Stat. 11 (1940) 86–92.