A Novel Similarity Measure of Link Prediction in Multi-Layer Social Networks Based on Reliable Paths

Amin Rezaeipanah (amin.Rezaeipanah@gmail.com)
Rahjuyan danesh Borazjan

Research Article

Keywords: social networks, link prediction, multi-layer networks, reliable paths, similarity measure

Posted Date: December 13th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-574927/v1

A Novel Similarity Measure of Link Prediction in Multi-Layer Social Networks Based on Reliable Paths

Amin Rezaeipanah

Department of Computer Engineering, University of Rahjuyan Danesh Borazjan, Bushehr, Iran,
amin.rezaeipanah@gmail.com

Abstract

Online social networks are an integral element of modern societies and significantly influence the formation and consolidation of social relationships. In fact, these networks are multi-layered so that there may be multiple links between a user' on different social networks. In this paper, the link prediction problem for the same user in a two-layer social network is examined, where we consider Twitter and Foursquare networks. Here, information related to the two-layer communication is used to predict links in the Foursquare network. Link prediction aims to discover spurious links or predict the emergence of future links from the current network structure. There are many algorithms for link prediction in unweighted networks, however only a few have been developed for weighted networks. Based on the extraction of topological features from the network structure and the use of reliable paths between users, we developed a novel similarity measure for link prediction. Reliable paths have been proposed to develop unweight local similarity measures to weighted measures. Using these measures, both the existence of links and their weight can be predicted. Empirical analysis shows that the proposed similarity measure achieves superior performance to existing approaches and can more accurately predict future relationships. In addition, the proposed method has better results compared to single-layer networks. Experiments show that the proposed similarity measure has an advantage precision of 1.8% over the Katz and FriendLink measures.

Keywords: social networks, link prediction, multi-layer networks, reliable paths, similarity measure.

1. Introduction

Many real-world systems can be described as networks that have nodes with the role of objects [1]. These networks contain links between nodes that represent relationships or interactions between objects. Therefore, the study of complex networks has become a common focus of many branches of science [2]. As part of recent research on large and complex networks, social network analysis (SNA) has become necessary due to their increasing extension [3]. However, social networks of objects are very dynamic. They grow and change rapidly with the addition of nodes and links. As a result, predicting links in these networks is an interesting and challenging problem that has recently attracted more attention [4]. For example, finding a potential friendship between two users on a social network or a potential collaboration between two scientists may be interesting. This problem is commonly known as the Link Prediction problem.

The link prediction problem assumes the probability of a link between two nodes in a network, so that there is currently no link between them [5]. In this problem, the social network G is assumed to be consecutive times t_0 and t_1. Here, we are looking for a set of links that do not exist in $G[t_0]$, but are likely to appear in $G[t_1]$. The network $G[t_0]$ is used for training and the network $G[t_1]$ for testing. The correctness of the suggestions can be evaluated according to the predicted links and the actual links. The link prediction process is shown in Fig. 1.

The algorithms based on local/global similarity measure (assigning similarity rank to adjacent nodes) are maximum probability approaches and probabilistic models for link prediction [6]. Classical approaches mainly take into account the similarity of the local structure when link prediction [5]. These methods use some similarity measure such as Common Neighbors, FriendLink, Katz, etc. to estimate the probability of adding new links to the network [7]. However, most of them are designed for a specific domain and for this reason they are called "algorithmic small world hypothesis" [5]. Social networks are very big with a large number of users connecting to each other through various types of links. Therefore, predicting these links is still challenging and it is necessary to achieve a predictive method with acceptable precision.

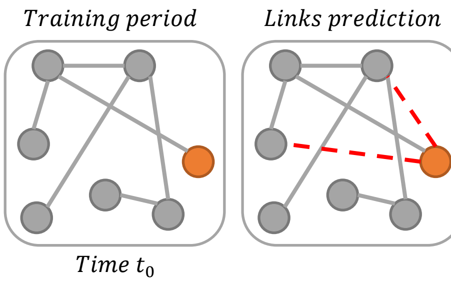

Fig. 1. The link prediction process

Although, the link prediction problem has been extensively studied and various researches have been presented to solve it [7-9]; However, the problem of how to optimally and effectively combine information to describe future communications remains largely unresolved. In [10], to link prediction in social networks, the analysis of user's demographic features has been used. The results show that the cluster coefficient and the shortest path are effective in the link prediction. In [11], a new similarity measure is proposed for the link prediction based on local structures in social networks. This measure is calculated through a supervised learning model with an observer based on estimating the similarity of source and destination nodes on a large database. In [12], a link prediction method based on the Deep Belief Network (DBN) for signed social networks is proposed. Since the DBN distributes the learning on all instances, it can be expected that the proposed links will be distributed along with their model tags. Here, the Bhattacharyya kernel is used to measure the similarity of the k-dimensional Gaussian distributions. Finally, an SVM classifier is trained to predict links based on user similarity information. In [13], link prediction in weighty social networks using learning automata is presented. Here, for each test link, a learning automaton is provided to estimate the actual weight of the link based on the weight of links in the current network. Then, each learning automata will be rewarded or punished according to its influence upon the true weight estimating of the training set.

In recent years, the link prediction problem has become popular on large networks. Researchers have proposed various methods to find missing links [7-9]. Most of these methods are calculated based on a similarity measure on neighboring nodes [11]. These methods also have limitations, because the same value is assumed for all common nodes of a node. In this paper, an efficient solution to the problem of link prediction in multi-layer social networks is presented, where a novel similarity measure is used to calculate similarity. The proposed similarity measure with assigning weight to links considers different values for common nodes of a node.

The remainder of the paper is organized as follows: Section 2 presents the overview of the link prediction problem in multi-layer networks. Section 3 introduces some of the classical similarity measures. Section 4 presents the details of the proposed method and experimental results and discussion are given in Section 5. Finally, the conclusion are described in Section 6.

2. Link prediction in multi-layer networks

Typically, modeling a single-layer social networking platform creates provides [14]. Because, all nodes of a single-layer network are considered to be of the same type, and all communications between the nodes are assumed to be equal. Therefore, this modeling method may lead to incorrect descriptions of some phenomena in the real world. Some real-world platforms have multi-layered structures. Obviously, social networks reflect a multi-layered structure [15]. Meanwhile, users of these networks may be in different groups or even in some cases on different platforms such as Facebook and Twitter. A user probably has different communication structures on Facebook and Twitter networks. Multiplex and heterogeneous networks are two well-known categories of multi-layer networks [16]. Multi-layer networks consist of interconnected nodes of the same type with different types of communications. The nodes are communicated by inter-layer and inside-layer links.

In this paper, a novel similarity measure for link prediction in a two-layer platform is presented. Here, the link prediction problem for same users on two social networks including Twitter and Foursquare is performed. In general, a multi-layered social network has different types of links. A social network architecture with two-layer (i.e., G_α and G_β) and three types of links (i.e., R_1, R_2 and R_3) is shown in Fig. 2.

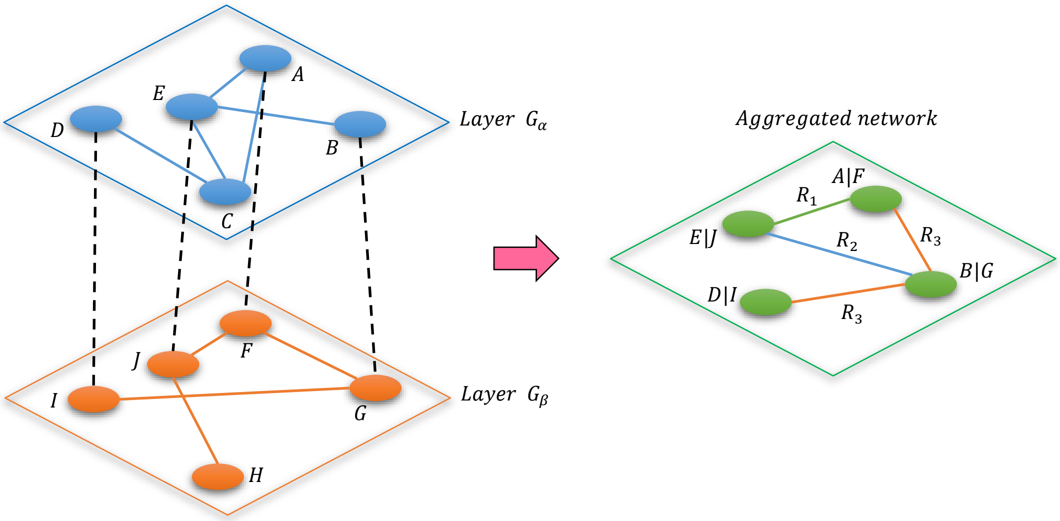

Fig. 2. Two-layer social network architecture

Here, there is a social network with two-layers G_α and G_β, where both networks are considered undirection. There are different types of links including two-layer links (R_1), single-layer links on G_α (R_2) and single-layer links on G_β (R_3). In this paper, the link prediction problem on two-layer networks is considered. Generally, link prediction is more useful in multi-layer networks than in single-layer networks, because multiple layers may provide more information about a node than a single-layer network.

It is important to study link prediction in multi-layer networks. Multi-layer networks consist of several layers with the number of same nodes in each layer [6]. The information from these layers may be used to predict missing links in a layer. The use of inter-layer information for solving the link prediction problem in multi-layer networks has already been considered in a number of researches [17]. In [18], an iterative degree penalty (IDP) algorithm for link prediction in multi-layer social networks is presented. IDP performs better than current methods based on network structure when the average network degree and nodes overlapping rate are low. In [19], hyperbolic geometry was used to predict links in multiplex networks. Here, link prediction is performed using two new similarity measures based on hyperbolic distance. In [20], a decision tree classification model is proposed to link prediction in a multiplex collaboration network with three layers. In [6], the supervised classification model is used to link prediction in a two-layer network including Twitter and Foursquare. Weight networks may provide more information than unweight networks in which each link has a specific weight [21]. These weights are effective in more accurately predicting links [21]. Currently, multi-layer networks with weighted links are more popular for solving link prediction problems [22].

3. Classical link prediction measures

Most link prediction methods assign a weight value to the link of each node pair (u, v) based on a similarity measure [11-13]. This value is a score for predicting missing links between two nodes. The two nodes with the highest similarity scores are more likely to be linked in the future. In this section, some of the most popular classical similarity measures for the link prediction problem are introduced.

Common Neighbors (CN) measure: In CN measure, the score for link prediction is computed by finding the number of common neighbors that are directly connected to the two nodes under evaluation [23]. The CN measure can be represented by Eq. (1).

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

Where, u and v are nodes, and $\Gamma(u)$ and $\Gamma(v)$ show the neighbors of nodes u and v, respectively.

Jaccard (JA) measure: This measure was developed in 1901 based on a statistic to compare similarity and diversity of sample sets [24]. The JA similarity measure refers to the ratio of common neighbors of nodes u and v to the all neighbor's nodes of u and v and prevents higher degree nodes to have high similarity measure with other nodes. The JA measure can be represented by Eq. (2).

$$S_{JA}(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|} \tag{2}$$

Adamic-Adar (AA) measure: This measure is related to the JA measure and used to compute the closeness of nodes based on their common neighbors [25]. AA measure gives more importance to common neighbors who have fewer neighbors. The AA measure can be represented by Eq. (3).

$$S_{AA}(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{1}{\log k_z} \tag{3}$$

Where, z is the common neighbor of u and v nodes and k_z is the degree of z node.

Katz (KT) measure: KT measure is a global structure based similarity index and considers all paths between two nodes in calculating the similarity score [26]. This measure introduces the concept of node centrality. The KT measure can be represented by Eq. (4).

$$S_{KT}(u,v) = \sum_{l=2}^{\infty} \beta^{l} . \left|paths_{u,v}^{<l>}\right| \tag{4}$$

Where, $paths_{u,v}^{<l>}$ the number of length paths l is between nodes u and v, and β is a damping factor used to control path weights, where $0 < \beta < 1$. In fact, β is a factor to reduce the effect of long paths in calculating similarity scores.

FriendLink (FL) measure: FL measure is a quasi-local structure based similarity index and uses paths longer than 2 to calculate similarity [5]. The FL measure can be represented by Eq. (5).

$$S_{FL}(u,v) = \sum_{l=2}^{L} \frac{1}{l-1} . \frac{\left|paths_{u,v}^{<l>}\right|}{\prod_{j=2}^{l}(n-j)} \tag{5}$$

Where, n is the number of nodes in network, L is the maximum path length considered and $1/(l-1)$ is the attenuation factor that weights path according to length l. In addition, $\prod_{j=2}^{l}(n-j)$ is the number of possible length l-paths from u to v.

4. The proposed similarity measure

In this section, an efficient solution to the link prediction problem in the two-layer social network is presented. The proposed method performs link prediction based on same users on Twitter and Foursquare networks. First, same users are identified based on the maximum similarity in their profiles. Then, inter-layer and inside-layer links of users are configured. Then, users are assigned to two sets of training (E^{tr}) and testing (E^{te}). The purpose is to apply link prediction to Furasquare based on topological information from both Twitter and Furasquare layers. The flowchart of the proposed method is shown in Fig. 3.

In this paper, the similarity between users is calculated based on four topological features including the number of common neighbors (f_1), the number of common posts (f_2), the number of multi-layer paths (f_3) and the number of common multi-layer paths (f_4). Features are extracted for each user pair u and v, where $u \in E^{te}$ and $v \in E^{tr}$. The f_1 feature expresses the number of users that link to both u and v users. The f_2 feature represents the number of common keywords used in u and v user's posts. The f_3 feature expresses the number of multi-layer paths between users u and v. A multi-layer path of length 2 between users u and v is defined as G_T, where there is a link between users u and w in layer G_T (Twitter network) and user w is linked to user v in layer G_F (Foursquare Network). Here, the feature of the number of multi-layer paths between two users u and v is calculated based on the number of similar users w, where the path length is assumed to be 2. For example, in Fig. 4, there are two multi-layer paths of length 2 between users u_1 and u_2, including $u_1 \xrightarrow{G_T} u_4 \xrightarrow{G_F} u_2$ and $u_1 \xrightarrow{G_T} u_3 \xrightarrow{G_F} u_2$. Hence, the value of this feature is equal to 2. The f_4 feature is similar to the number of multi-layer paths, except that one of the links must appear in both layers. Thus, a common multi-layer path of length 2 between users u and v is defined as $u \xrightarrow{G_T} w \xrightarrow{G_T,G_F} v$ or $u \xrightarrow{G_T,G_F,} w \xrightarrow{G_F} v$. Here, the feature of the number of common multi-layer paths between two users u and v is calculated based on the number of similar users w, where the path length is assumed to be 2. For example, in Fig. 4, there is a common multi-layer path of length 2 between users u_1 and u_2 as $u_1 \xrightarrow{G_T,G_F} u_4 \xrightarrow{G_F} u_2$. Hence, the value of this feature is equal to 1.

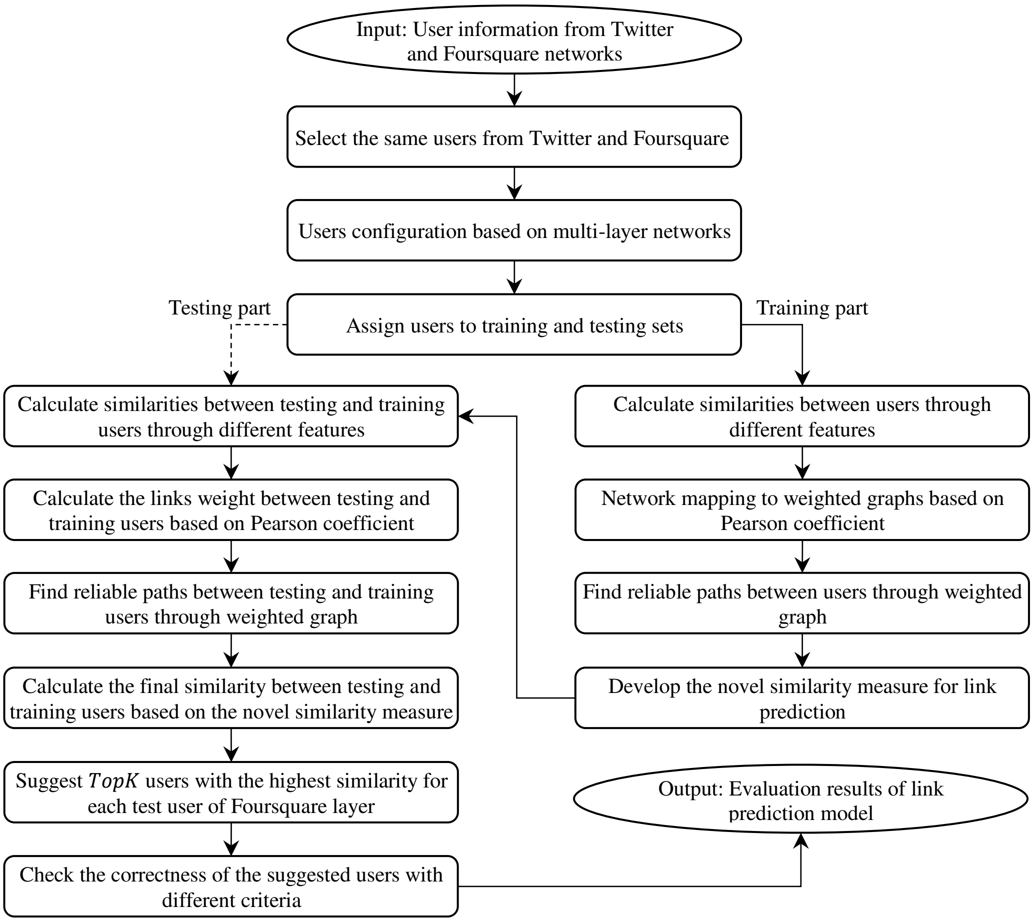
Input: User information from Twitter and Foursquare networks
Select the same users from Twitter and Foursquare
Users configuration based on multi-layer networks
Testing part
Assign users to training and testing sets
Training part
Calculate similarities between testing and training users through different features
Calculate the links weight between testing and training users based on Pearson coefficient
Find reliable paths between testing and training users through weighted graph
Calculate the final similarity between testing and training users based on the novel similarity measure
Suggest $TopK$ users with the highest similarity for each test user of Foursquare layer
Check the correctness of the suggested users with different criteria
Calculate similarities between users through different features
Network mapping to weighted graphs based on Pearson coefficient
Find reliable paths between users through weighted graph
Develop the novel similarity measure for link prediction
Output: Evaluation results of link prediction model

Fig. 3. Flowchart of the proposed method

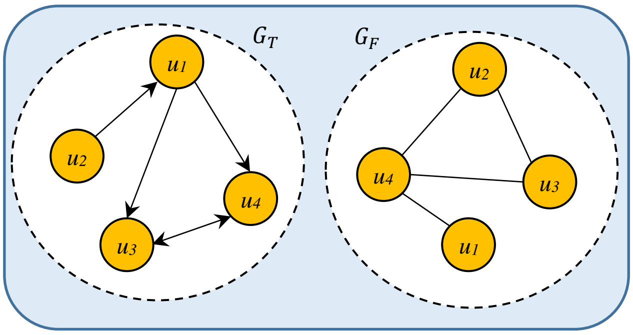

Fig. 4. An example of features the multi-layer paths and the number of common multi-layer paths

In the following method, the similarities between users are calculated based on the extracted features. Hence, the network is mapped to a weighted graph, where similarity is calculated based on the Pearson coefficient between each pair of users. Pearson correlation coefficient is calculated as Eq. (6).

$$p(u,v) = \frac{\sum_{i=1}^{m}(u_{fi} - \overline{u_f})(v_{fi} - \overline{v_f})}{\sum_{i=1}^{m}(u_{fi} - \overline{u_f})^2 \sum_{i=1}^{m}(v_{fi} - \overline{v_f})^2} \tag{6}$$

Where, u_{fi} and v_{fi} are the i-th features for users u and v, respectively, $\overline{u_f}$ and $\overline{v_f}$ are the average of all the features for users u and v, respectively, and m represents the total number of features extracted.

In the following, a novel measure is developed to calculate the similarity between users based on reliable paths. The proposed similarity measure is based on weighted networks and is developed based on the KT measure. Local similarity measures, such as CN and JA, apply to the link prediction problem based on all paths between two users of length 2. Meanwhile, quasi-local and global similarity measure such as FL and KT solve the link prediction problem using all paths between two users with a length of more than 2. These measures have been shown to provide more accurate link prediction than local similarity measure, because in them the diversity of communication paths is considered [5, 26]. However, quasi-local and global similarity measure only consider the number and length of different paths and do not consider the importance of any of the path links. For example, in the KT similarity measure, $\left|paths_{u,v}^{l}\right|$ considers only the number of paths that exist between users u and v with length l.

A reliable path is provided by generalizing the quasi-local or global similarity measure of unweight networks to weight networks, where the importance of each link in the path based on its weight is considered to calculate the final similarity [7]. Thus, a reliable path between two users includes the path with the highest weight, which shows the similarity between them. In general, the weight of a link indicates that it is probability to be safe on the path, which can be considered as the reliability of that path. Hence, a reliable path is a combination of the probabilities of all links in that path, which can help the link prediction in social networks.

In various studies, it has been shown that a reliable path between two users can be calculated based on the "multiply the path links weight" [7]. This is because the sum of the path weights cannot express the importance of the path with respect to the path length. However, this technique has so far been used on local similarity measures and in this paper is the first time that it is applied on quasi-local and global similarity measure. Here, a novel similarity measure based on KT global similarity measure is presented. In the KT, only the number and length of paths between users are considered, while in the proposed similarity measure, the effect of each link in the path is also applied through the path weight. Therefore, the proposed measure maps the similarity of the number of paths in KT to the number of weighted paths. Eq. (7) shows the proposed similarity measure for calculating the similarity between users u and v.

$$Sim(u,v) = \sum_{l=2}^{L} \beta^{l}. \left[\sum_{p \in paths_{u,v}^{<l>}} \prod_{(x,y) \in p} w(x,y) \right] \tag{7}$$

Where, $paths_{u,v}^{<l>}$ is the set of paths between users u and v of length l, and p represents a path of $paths_{u,v}^{<l>}$. (x,y) are two consecutive nodes of the path p that provide a link. $w(x,y)$ is the weight associated with the link (x,y). The damping factor β is defined similar to the KT measure to reduce the impact of paths with longer lengths. In addition, L is considered the maximum path length.

For a better understanding, consider the graph in Fig 5. In this example, assuming $\beta = 0.05$ and $L = 3$, the similarity between the two users u_2 and u_5 is calculated based on the KT measure. Based on 2 paths with length 2 (i.e., $\langle u_2 \rightarrow u_3 \rightarrow u_5 \rangle$ and $\langle u_2 \rightarrow u_6 \rightarrow u_5 \rangle$) and 1 path with length 3 (i.e., $\langle u_2 \rightarrow u_3 \rightarrow u_4 \rightarrow u_5 \rangle$) the final similarity score as Eq. (8).

$$S_{KT}(u_2,u_5) = [0.05^2 \times 2] + [0.05^3 \times 1] = 0.0051 \tag{8}$$

The results show that the KT measure does not consider the difference between links in the path, while there may be a strong link between two users in the path that has a high impact on future communication.

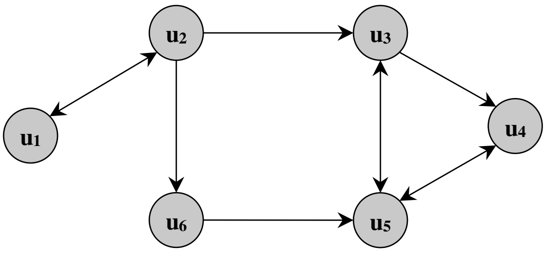

Fig. 5. An example of the similarity calculation in the KT measure

In order to observe the effect of the linked weight in calculating the final similarity, consider Fig. 6, where the link strength can be seen in the graph through the link weight. In this figure, there are 2 paths of length 2 between users u_2 and u_5 (i.e., $\langle u_2 \rightarrow u_3 \rightarrow u_5 \rangle$ and $\langle u_2 \rightarrow u_6 \rightarrow u_5 \rangle$). Given the links weight, the paths are shown as $\langle u_2 \xrightarrow{0.4} u_3 \xrightarrow{0.9} u_5 \rangle$ and $\langle u_2 \xrightarrow{0.3} u_6 \xrightarrow{0.1} u_5 \rangle$. Here, there is a stronger link in the first path, i.e., $\langle u_2 \rightarrow u_3 \rightarrow u_5 \rangle$. Because the total weight for the first path is 1.3, but this score is 0.4 for the second path. In addition, there is 1 path with length 3 as $\langle u_2 \xrightarrow{0.4} u_3 \xrightarrow{0.9} u_4 \xrightarrow{0.6} u_5 \rangle$ between users u_2 and u_5. Considering the proposed similarity measure, the final similarity score is according to Eq. (9).

$$Sim(u_2,u_5) = \left[0.05^2 \times \left((0.4 \times 0.9) + (0.3 \times 0.1)\right)\right] + [0.05^3 \times (0.4 \times 0.9 \times 0.6)] = 0.0010 \tag{9}$$

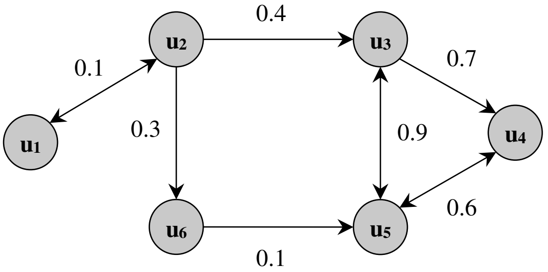

Fig. 6. An example of the similarity calculation in the proposed measure

We will now increase the link weight between users u_2 and u_5 to 0.8 to make this path more important for communication between these users in the future. Based on this, the similarity is calculated according to Eq. (10).

$$Sim(u_2, u_5) = \left[0.05^2 \times \left((0.4 \times 0.9) + (0.3 \times 0.8)\right)\right] + [0.05^3 \times (0.4 \times 0.9 \times 0.6)] = 0.0015 \quad (10)$$

It is clear that the proposed similarity measure increases the likelihood of linking between users u_2 and u_5 in the future due to increased link weights. Therefore, this technique considers safe and strong communication between links in calculating similarity between users, which can be effective for link prediction.

According to the proposed similarity measure, the similarity between each pair of users u and v is calculated, where $u \in E^{te}$ and $v \in E^{tr}$. Then, for each user such as u, number of $TopK$ users such as v is suggested with the highest similarity score. In the link prediction process for a user such as u versus a user such as v, a direct link between them (If there is a directed link or even an undirected link) should not be considered. In fact, the purpose is to predict the existence of direct link between these two users based on other links.

5. Simulation analysis

In this section, we perform extensive experiments on real data sets to evaluate the effectiveness of the proposed method. The proposed method is simulated in MATLAB R2019a software. All experiments were performed on a PC with 3.2 GHz Intel Core i7 CPU, 32 GB of RAM and Windows 10 operating system. In order to more accurate evaluation and fair comparisons, the all results are presented by the 10-fold cross validation technique. According to this technique, users are divided into two sets of training (E^{tr}) and testing (E^{te}) so that $E = E^{tr} \cup E^{te}$ and $E^{tr} \cap E^{te} = \emptyset$. Meanwhile, E is the total number of links between users on both layers. The purpose is to recommend the number of $TopK$ users with the highest rank of the E^{tr} collection to users in the E^{te} collection, where this process is performed for all users of the E^{te} collection who have links to at least one user from the E^{te} collection. This section consists of 4 subsections: (1) dataset description, (2) evaluation criteria, (3) parameter analysis, and (4) results and comparisons.

5.1 Dataset description

The dataset used in this paper is a two-layer social network including Twitter (as a microblogging social network) and Foursquare (as a location-based social network). The datasets used from both social networks were surveyed in November 2012 and are available from https://data.world/datasets [27, 28]. The Twitter network allows users to share tweets (messages) with a maximum of 140 characters. In this network, users can follow these tweets, where the link of the follower users to following users forms a directional network. Foursquare is the undirected network that allows users to share their location with friends by "checking-in" at a given place using their smartphone. In this paper, the social communications of same users in Twitter and Forasquare social networks are considered, where based on the communications in the Twitter network, link prediction is done for Forasquare network users. The same users were searched based on a similarity score greater than the threshold value for all profile pairs. Meanwhile, there are about 45,000 potentially identical users identified on the profile alone. All paths between users on both networks can be found by performing the DFS search. Here is a collection of 1517 users that can be used to simulate the proposed method. Statistical data on these two datasets are available in Table 1.

Table 1. Statistical data from the Twitter and Foursquare datasets

Platform	No. same users	No. links	Average degree of nodes	No. common links
Twitter	1517	15172	in = 10.05, out = 10	6551

Foursquare		18481	24.4	

5.2 Evaluation criteria

Different evaluation criteria such as Precision, Recall and F-measure are used to confirm the performance of the proposed method [29]. Precision is defined as the ratio of the number of correct users suggested ($TopK$) to the total number of users suggested. If k_R are only links from $TopK$ in the E^{te} collection; then the precision criterion can be represented by Eq. (11). Recall is defined as the ratio of the number of correct users suggested to the total number of actual related users. If k_s contains users from the E^{te} collection that link to the target user (actual related users), then the recall criterion can be represented by Eq. (12). Finally, the f-measure can be interpreted as a weighted harmonic meaning of precision and recall, where it considers both false positives and false negatives. This criterion is defined according to Eq. (13).

$$Precision = \frac{k_R}{TopK} \tag{11}$$

$$Recall = \frac{k_R}{k_s} \tag{12}$$

$$F\_measure = \frac{2 \times Precision \times Recall}{recision + Recall} \tag{13}$$

In order to calculate the evaluation criteria, first precision, recall and f-measure are computed for each user of E^{te} collection (target users) and then the final results are reported on average for all users.

5.3 Parameter analysis

In this section, the parameters of the proposed method to improve performance in link prediction are analyzed. These parameters include path length (L), path length impact factor (β), similarity calculation coefficient, two-layer platform, extracted features, and attenuation factors. The analysis is performed in order to find the optimal value of these parameters in the proposed method. When simulation is applied for a parameter with different values, the other parameters are set to $L = 3$, $\beta = 0.05$ and $TopK = 5$ on the default values.

First, the proposed similarity measure analysis with different path lengths is reported in Table 2. The results show that the proposed similarity measure with a maximum path length of 3 has the best performance. However, the results for $L = 2$ and $L = 4$ are also promising. Optimal value are shown in bold type.

Table 2. Proposed similarity measure analysis with different path lengths

Criteria	1	2	3	4	5
Precision	0.637	0.735	0.765	0.758	0.703
Recall	0.504	0.530	0.558	0.551	0.498
F-measure	0.563	0.616	**0.645**	0.638	0.583

In the following, the value of the path length impact factor in the proposed similarity measure according to Table 3 is investigated. The results clearly show the higher efficiency of the path length impact factor with a value of 0.05. However, the $\beta = 0.01$ also reports suitable results. The value shown in bold in the table indicates the comparatively better optimum value.

Table 3. Proposed similarity measure analysis with different values of path length impact factor

Criteria	0.01	0.05	0.1	0.15	0.2
Precision	0.751	0.765	0.708	0.660	0.611
Recall	0.598	0.558	0.458	0.511	0.472
F-measure	0.666	**0.645**	0.556	0.577	0.533

In the proposed method, Pearson correlation coefficient is used to calculate the similarity of the two users based on the extracted features. However, there are other coefficients such as Cosine, Jaccard, etc. to calculate similarity. Here, the Pearson coefficient is compared with the Cosine and Jaccard coefficients for better performance in the link prediction problem. The results of this comparison in Table 4 show the superiority of the Pearson coefficient, although this superiority is negligible.

Table 4. Comparison of the efficiency of different similarity coefficients in the proposed method

Criteria	Cosine	Jaccard	Pearson
Precision	0.748	0.744	0.765
Recall	0.556	0.530	0.558
F-measure	0.638	0.620	**0.645**

In the proposed method, a two-layer platform including Twitter and Foursquare is used to undirected predict links in the Foursquare layer, so that each network is considered in a separate layer. Therefore, to predict links in the Foursquare network, it is necessary to have communication information in both Twitter and FuraSquare networks. In other words, the communication topology of Twitter and Forasquare networks is used to predict links in Forasquare. Here, we proved that using both layers of information improves link prediction performance compared to using single-layer information alone. In the single-layer platform, only the Foursquare network topology information and the f_1 and f_2 features are used for link prediction work. The results of this comparison in Table 5 clearly show the superiority of link prediction in the two-layer platform, where this superiority in the precision criterion is more than 19%. Optimal value are shown in bold type.

Table 5. Comparison of link prediction in Foursquare network based on one-layer and two-layer platforms

Criteria	One-layer platform	Two-layer platform
Precision	0.639	0.765
Recall	0.441	0.558
F-measure	0.521	**0.645**

In the following, the effect of 4 extracted features on the performance of link prediction is investigated. Here, the effect of each feature on link prediction is calculated by deactivating that feature from the proposed method. The results of this experiment are shown in Fig. 7 based on various criteria. The results show that the accuracy of link prediction is reduced by deactivating the f_3 feature (i.e., the number of multi-layer paths) compared to other features. Meanwhile, this feature has the greatest impact on the performance of the proposed similarity measure. However, the proposed method offers the best performance considering all the features.

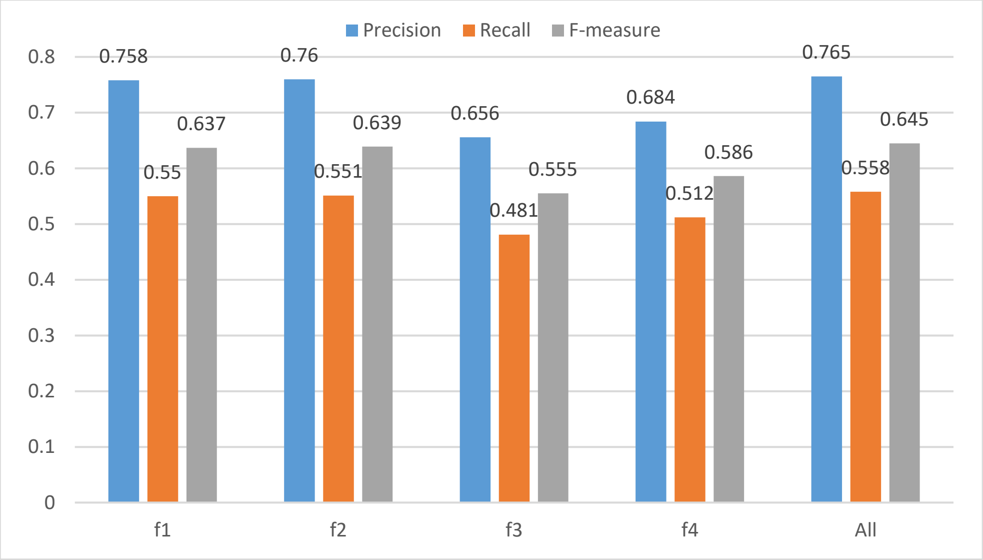

Fig. 7. The effect of extracted features on the efficiency of the proposed method

The proposed similarity measure is designed with attenuation factors β^l. The purpose of this factor is to assign a score less similarity to paths with more length. In the following, the efficiency of this factor is compared against the number of different attenuation factors. The results of this comparison in Table 6 prove the best performance for the attenuation factor β^l.

Table 6. Proposed similarity measure analysis with different attenuation factors

Criteria	$\boldsymbol{\beta^l}$	$\boldsymbol{\frac{1}{l-1}}$	$\boldsymbol{\frac{1}{2l}}$	$\boldsymbol{\frac{1-\beta^l}{l-1}}$	$\boldsymbol{\frac{1}{l^2}}$
Precision	0.765	0.734	0.685	0.745	0.692
Recall	0.558	0.544	0.529	0.562	0.525
F-measure	0.645	0.625	0.599	0.641	0.597

5.4 Results and comparisons

The proposed similarity measure uses only the topographic information of the network, so its results should be compared with other classical similarity measures in the link prediction problem. Here, five classical similarity measures including Common Neighbors (CN), Jaccard (JA), Adamic-Adar (AA), Katz (KT) and FriendLink (FL) are used for comparison. Meanwhile, for each different number of suggested users ($TopK$), the evaluation criteria including precision, recall, and f-measure are calculated as average for all users of the E^{te} collection.

Figure 8 shows the results of the comparison of the proposed similarity measure with other classical similarity measures based on precision criterion. Comparisons are presented based on different $TopK$ from 1 to 30. Here, PM refers to the results of the proposed similarity measure. The results of this comparison show that the proposed similarity measure in most cases has better precision than other classical similarity measures. At best, when $TopK = 1$, the precision results for the proposed method are 0.877. However, as the number of suggested users increases, the results of the precision criterion decrease. This is clearly visible for all similarity measures. The reason for this decrease is the precision criterion calculation process, where at the denominator is the $TopK$ value.

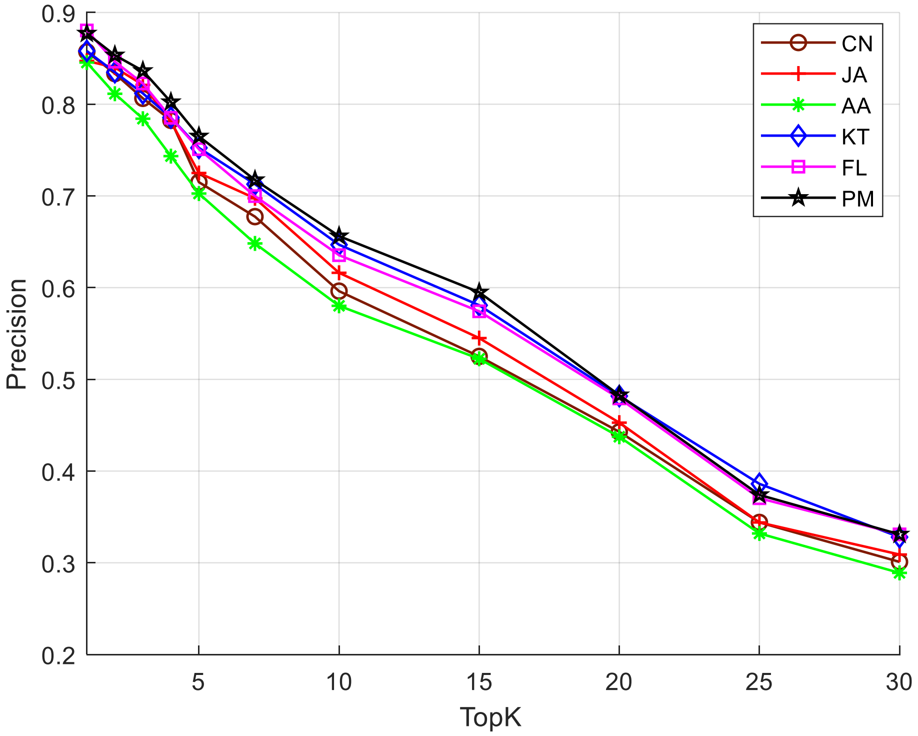

Fig. 8. Comparison results based on precision criterion with different number of suggested users

In another experiment, the proposed similarity measure (PM) was compared to the classical similarity measures based on recall criterion. The results of this comparison with different $TopK$ from 1 to 30 are shown in Fig. 9. The results of this experiment also show the superiority of the proposed similarity measure in most comparisons. At best, when $TopK =$ 30, the recall criterion results for the proposed method are 0.635. However, the recall value decreased as the number of suggested users decreased. This is clearly visible for all similarity measures. The reason for this decrease is the recall criterion calculation process, where at the denominator is the number of actual related users.

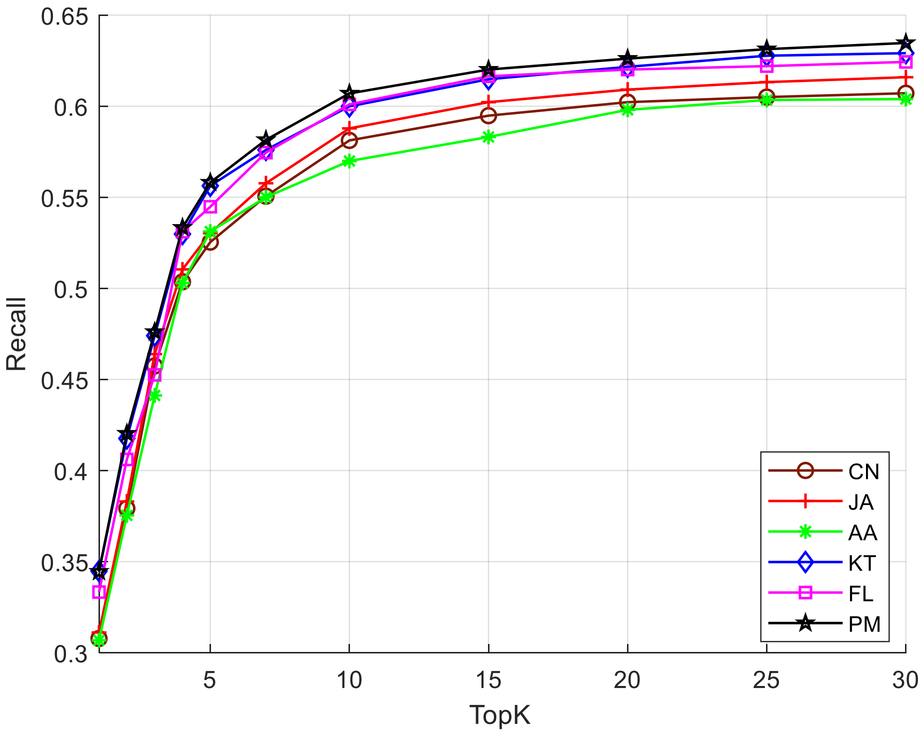

Fig. 9. Comparison results based on recall criterion with different number of suggested users

Figure 10 shows the f-measure criterion results for the proposed similarity measure (PM) and other classical similarity measures including CN, JA, AA, KT, and FL. F-measure criterion makes it easier to visualize and compare the performance of link prediction methods in different operating conditions than ROC [30]. Here, the results are reviewed for different number of suggested users from 1 to 30. The results of this comparison show that when $TopK = 5$, the proposed method achieves the best f-measure result of 0.645. After the proposed method, the similarity criteria of KT, FL, JA, CN and AA are in the next ranks in terms of efficiency, respectively. These results prove the superiority of the proposed similarity measure over other classical similarity measures.

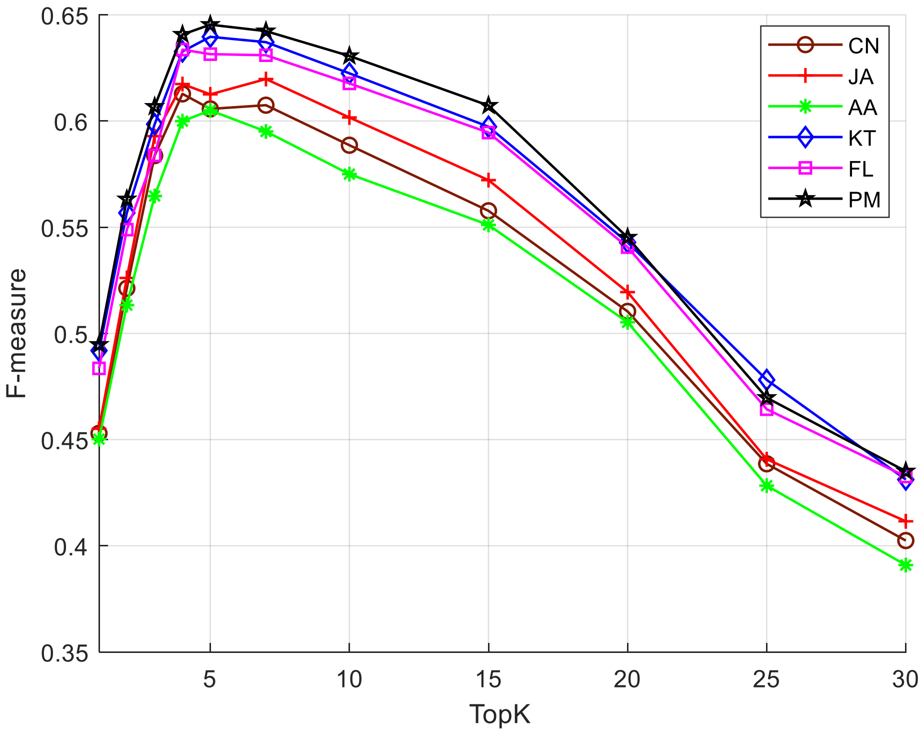

Fig. 10. Comparison results based on f-measure criterion with different number of suggested users

Finally, Table 1 shows the numerical results of the comparison of the proposed similarity measure with other classical similarity measures. Here, $TopK = 5$ is considered for comparison. The results clearly show the superiority of the proposed method over other measures. After the proposed method, the Katz and FriendLink measures show better performance, respectively. However, in precision criterion, the proposed similarity measure is about 1.8% superior to the Katz and FriendLink measures.

Table 7. Comparison results of the proposed similarity measure against other classical similarity measures

Criteria	Common Neighbors	Jaccard	Adamic-Adar	Katz	FriendLink	Proposed Method
Precision	0.715	0.725	0.703	0.752	0.751	0.765
Recall	0.525	0.530	0.531	0.556	0.545	0.558
F-measure	0.606	0.613	0.605	0.640	0.631	0.645

6. Conclusions and future work

Link prediction approaches in social networks predict the possibility of a link between the two nodes in the future. These approaches are essential for inferring social interactions or friend recommendations to users. Link prediction is a fundamental social network problem analysis. Using similarity measure to predict the probability of future interactions is one of the common methods in link prediction problem. In general, there are various link prediction techniques that use similarity measures to estimate the proximity of nodes in the network. In this paper, a novel similarity measure based on reliable paths and topological features for two-layer network was developed. This study was performed on Twitter (directed platform) and Foursquare (undirected platform) networks, where the structural data of the layers is used to link prediction in the Foursquare network. Here, reliable paths have been developed by generalizing the global similarity measure from unweight to weight networks based on Pearson correlation coefficient. The proposed similarity measure uses only the topographic information of the network, so its results should be compared with other classical similarity measures in the link prediction problem. Experiments show that the proposed method based on the information extraction between layers can significantly improve the performance of link prediction. These observations strongly prove that the topological information of network provides suitable knowledge for estimating the similarity between users. Mostly our future work will focus on the link prediction problem by combining the proposed method with the content and semantic information of network nodes. In addition, we will examine more real datasets based on networks with more than two-layers.

Compliance with ethical standards

Conflict of interest

The authors declare that there is no conflict of interest.

Informed consent

For this type of study formal consent is not required.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Reference

[1] Wang, H., & Le, Z. (2020). Seven-Layer Model in Complex Networks Link Prediction: A Survey. *Sensors*, *20*(22), 6560.

[2] Aziz, F., Gul, H., Muhammad, I., & Uddin, I. (2020). Link prediction using node information on local paths. *Physica A: Statistical Mechanics and its Applications*, *557*, 124980.

[3] Wang, Z., Wang, Y., Ma, J., Li, W., Chen, N., & Zhu, X. (2019). Link prediction based on weighted synthetical influence of degree and H-index on complex networks. *Physica A: Statistical Mechanics and its Applications*, *527*, 121184.

[4] Rezaeipanah, A., Mokhtari, M. J., & Boshkani, M. (2020). Providing a new method for link prediction in social networks based on the meta-heuristic algorithm. *International Journal of Cloud Computing and Database Management*, 1(1), 28-36.

[5] Papadimitriou, A., Symeonidis, P., & Manolopoulos, Y. (2012). Fast and accurate link prediction in social networking systems. *Journal of Systems and Software*, 85(9), 2119-2132.

[6] Rezaeipanah, A., Ahmadi, G., & Matoori, S. S. (2020). A classification approach to link prediction in multiplex online ego-social networks. *Social Netw. Analys. Mining*, *10*(1), 27.

[7] Zhao, J., Miao, L., Yang, J., Fang, H., Zhang, Q. M., Nie, M., ... & Zhou, T. (2015). Prediction of links and weights in networks by reliable routes. *Scientific reports*, *5*, 12261.

[8] Ayoub, J., Lotfi, D., El Marraki, M., & Hammouch, A. (2020). Accurate link prediction method based on path length between a pair of unlinked nodes and their degree. *Social Network Analysis and Mining*, *10*(1), 9.

[9] Yan, R., Li, Y., Li, D., Wu, W., & Wang, Y. (2020). SSDBA: the stretch shrink distance based algorithm for link prediction in social networks. *Frontiers of Computer Science*, *15*(1), 1-8.

[10] Madahali, L., Najjar, L., & Hall, M. (2019, March). Exploratory factor analysis of graphical features for link prediction in social networks. In *International Workshop on Complex Networks* (pp. 17-31). Springer, Cham.

[11] Aghabozorgi, F., & Khayyambashi, M. R. (2018). A new similarity measure for link prediction based on local structures in social networks. *Physica A: Statistical Mechanics and its Applications*, *501*, 12-23.

[12] Yuan, W., He, K., Guan, D., Zhou, L., & Li, C. (2019). Graph kernel based link prediction for signed social networks. *Information Fusion*, *46*, 1-10.

[13] Moradabadi, B., & Meybodi, M. R. (2018). Link prediction in weighted social networks using learning automata. *Engineering Applications of Artificial Intelligence*, *70*, 16-24.

[14] Najari, S., Salehi, M., Ranjbar, V., & Jalili, M. (2019). Link prediction in multiplex networks based on interlayer similarity. *Physica A: Statistical Mechanics and its Applications*, *536*, 120978.

[15] Yasami, Y., & Safaei, F. (2018). A novel multilayer model for missing link prediction and future link forecasting in dynamic complex networks. *Physica A: Statistical Mechanics and its Applications*, *492*, 2166-2197.

[16] De Bacco, C., Power, E. A., Larremore, D. B., & Moore, C. (2017). Community detection, link prediction, and layer interdependence in multilayer networks. *Physical Review E*, *95*(4), 042317.

[17] Pan, L., Zhou, T., Lü, L., & Hu, C. K. (2016). Predicting missing links and identifying spurious links via likelihood analysis. *Scientific reports*, *6*(1), 1-10.

[18] Tang, J., Cui, Y., Li, Q., Ren, K., Liu, J., & Buyya, R. (2016). Ensuring security and privacy preservation for cloud data services. *ACM Computing Surveys (CSUR)*, *49*(1), 1-39.

[19] Samei, Z., & Jalili, M. (2019). Application of hyperbolic geometry in link prediction of multiplex networks. *Scientific reports*, *9*(1), 1-11.

[20] Maruyama, W. T., & Digiampietri, L. A. (2016, July). Co-authorship prediction in academic social network. In *Anais do V Brazilian workshop on social network analysis and mining* (pp. 61-72). SBC.

[21] Gupta, N., & Singh, A. (2014, December). A novel strategy for link prediction in social networks. In *Proceedings of the 2014 CoNEXT on student workshop* (pp. 12-14).

[22] Malik, D., & Singh, A. (2020). Link prediction in multilayer networks. *International Journal of Business Intelligence and Data Mining*, *16*(4), 490-505.

[23] Lorrain, F., & White, H. C. (1971). Structural equivalence of individuals in social networks. *The Journal of mathematical sociology*, *1*(1), 49-80.

[24] Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013, March). Using of Jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists* (Vol. 1, No. 6, pp. 380-384).

[25] Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. *Social networks*, 25(3), 211-230.

[26] Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, *18*(1), 39-43.

[27] Kong, X., Zhang, J., & Yu, P. S. (2013, October). Inferring anchor links across multiple heterogeneous social networks. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 179-188).

[28] Zhang, J., Yu, P. S., & Zhou, Z. H. (2014, August). Meta-path based multi-network collective link prediction. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1286-1295).

[29] Tang, R., Jiang, S., Chen, X., Wang, H., Wang, W., & Wang, W. (2020). Interlayer link prediction in multiplex social networks: an iterative degree penalty algorithm. *Knowledge-Based Systems*, 105598.

[30] Lichtnwalter, R., & Chawla, N. V. (2012, August). Link prediction: fair and effective evaluation. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 376-383). IEEE.