Research Article

ECLP: Friend Recommendation Using Ensemble Approach for Detecting Communities Performing Link Prediction

Hasan Saeidinezhad,¹ Elham Parvinnia ,¹ and Reza Boostani ²

¹Department of Computer Engineering, Shiraz Branch, Islamic Azad University, Shiraz, Iran

²CSE and IT Department, Electrical and Computer Engineering Faculty, Shiraz University, Shiraz, Iran

Correspondence should be addressed to Elham Parvinnia; parvinnia@iaushiraz.ac.ir

Received 21 April 2021; Revised 8 May 2021; Accepted 22 May 2021; Published 9 August 2021

Academic Editor: Mohammad Yazdi

Copyright © 2021 Hasan Saeidinezhad et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Social networks provide a variety of online services that play an important role in new connections among members to share their favorite media, document, and opinions. For each member, these networks should precisely recommend (predict) the link of members with the highest common interests. Because of the huge volume of users with different types of information, these networks encounter challenges such as dispersion and accuracy of link prediction. Moreover, networks with numerous users have the problem of computational and time complexity. These problems are caused because all the network nodes contribute to calculations of link prediction and friend suggestions. In order to overcome these drawbacks, this paper presents a new link prediction scheme containing three phases to combine local and global network information. In the proposed manner, dense communities with overlap are first detected based on the ensemble node perception method which leads to more relevant nodes and contributes to the link prediction and speeds up the algorithm. Then, these communities are optimized by applying the binary particle swarm optimization method for merging the close clusters. It maximizes the average clustering coefficient (ACC) of the whole network which results in an accurate and precise prediction. In the last phase, relative links are predicted by Adamic/Adar similarity index for each node. The proposed method is applied to Astro-ph, Blogs, CiteSeer, Cora, and WebKB datasets, and its performance is compared to state-of-the-art schemes in terms of several criteria. The results imply that the proposed scheme has a significant accuracy improvement on these datasets.

1. Introduction

Studies in computer science have rapidly developed in recent decades, and many aspects of computer science are currently being studied [1–7]. These studies include many areas such as image processing and the advanced relevant techniques, machine learning and pattern recognition, data mining, and relative discussions [8–13]. On the other hand, there are many studies dealing with issues such as developing Statistical Process Control (SPC) techniques to take advantage of the advanced measurement technologies and sensors or even assessing the capability of these measurements such as 3D laser scanner that provides high-density data [14–19]. Researchers have also studied social networks and related issues [20–26].

Social networks play an important role in people's daily lives [27–30]. Social media creates a substrate for users to

share their information, in terms of media, documents, and opinions [31–34]. These networks provide online services imitating and simulating real-life interactions and relationships. Social networks also enable making connections among members, in a way that each user can select and add new friends from a long list of candidates, suggested by the recommender system module. Research findings show that users usually connect with their friends/colleagues (who they know and see in real life) as well as new friends who are introduced by the link prediction service of social networks [35, 36].

With the daily growth of information in social networks, the process of introducing proper friends by the link prediction service has been a very challenging task and requires high precision [37]. Recommender systems, which predict the link for users, have been used for more than ten years to

offer products and services to users based on their interests, preferences, and online behavior [38–45]. Another challenge in the LP problem is precise link prediction among members in the large size network [46, 47]. Also in the link suggestion systems, in which actual communities are very scattered, the prediction will be challenging. In other words, the nodes in these communities are associated with only a small fraction of all network nodes. For example, in the case of Facebook, a user typically only connects to about 100 (out of 500 million) nodes. Therefore, this method gives acceptable accuracy of prediction in a way that out of 500 million possible predictions, only 100 errors can occur. One of the most important problems that recommender (link prediction) systems are faced with is the problem of cold start that means there is not enough and sufficient information (ratings) in the system to provide a recommendation [48–52].

In general, link prediction methods are divided into two categories: supervised (with observer) such as decision tree, Bayes, and K-nearest neighbor (KNN) [14] and unsupervised (without observer) which are based on the network's topology information. These methods are divided into three categories: local, global, and semilocal [53].

Local criteria are based on neighborhood topological information. These methods are based on the idea that if there are some similarities between the nodes m_1 and m_2 , they will be more likely to be connected in the near future. The most important local criteria are common neighbors (CN) [48, 54], Jaccard's coefficient [55, 56], and Adamic/Adar index (AA) [57]. These criteria successfully reduce the computational cost; however, they do not have relatively high accuracy in forecasting [58]. Semilocal criteria need more information about the network topology rather than the local criteria but do not require the information of all the network connections. This criterion has been created with the aim of balancing the local and global criteria and does not have the challenges of global criteria such as being time-consuming and difficulties to find the complete network structure. It is also a good trade-off between the complexity and performance, such as local path index (LPI) methods and local random walk (LRW). Global criteria, similar to methods based on all paths in the network [10], need complete information about network topology. In this approach, based on all the paths between the two nodes in the network and all the information extracted from the entire network topology, the similarity of the two nodes is measured. Well-known global criteria are Katz [59], random walk, collision time (Hitting Time), and Sim-Tank, all of which suffer from high computational complexity [60, 61].

Some of the link prediction methods are designed based on data mining methods such as clustering that can categorize each group of members with mutual interests into one cluster. This technique can highly limit the search space and suggests a list of candidate members who are selected from the corresponding cluster for each member in a cluster. This process can be used to improve the matching process and provide suggestions on social networks and also predict future events. Previous researchers clustered the nodes of a social network for their link prediction algorithm and

deduced that there was a significant relation between communities related to the network structure and the precision of their algorithm. Moreover, using the similarity-based link prediction algorithm, which used the clustering information, improved the accuracy of link prediction [62]. A gravity-based link prediction method was proposed which included community information of networks. The researchers used the Louvain algorithm for community detection as well as parallelism in the community processing to speed up the prediction globally [59].

None of the previous work considered overlapping between the recognized clusters; also, they did not check merging clusters to find an optimum combination of clusters for this particular problem. Therefore, a three-phase method is proposed in this paper which considers all the pre-discussed subjects. In the first phase of the algorithm (community detection), ensemble of clustering is applied. In this phase, dense communities are searched by deploying node perception algorithm, which, despite the previous works, is a clustering method that considers overlapping between the detected clusters by using three hyperparameters. To find proper values for the parameters, a greedy search is applied. In the second phase, the clustering process is continued by applying binary particle swarm optimization (BPSO) algorithm to optimize clusters by merging near-enough neighbor clusters. The fitness function of BPSO is defined to maximize average clustering coefficient (ACC) of the network. Then, those communities are merged where their merge enhances the fitness value. Executing this phase leads to more suitable communities because they improve network ACC. Eventually, in the third phase of our work, a similarity-based link prediction algorithm (Adamic/Adar) is applied with the hope of achieving high performance in link prediction in social networks.

The rest of this paper is organized as follows. Section 2 presents previous related works. Our proposed algorithm is described in Section 3 in detail. Section 4 describes the dataset and evaluation criteria. The results and analysis of our work are reported in Section 5, and the last section concludes our paper.

2. Related Work

Due to the importance of friend suggestions in social networks, this issue has been investigated in several studies. Bastami et al. [59] introduced a new similarity measure called Triadic, which used some units called Motifs. They defined Motifs as the different forms of a Triadic network including three nodes. Their main proposed method consisted of two phases; in the first phase, a data table was created including the information of each node in the graph. Then, in the second phase, they used a classifier to be trained on the data table entries to predict the links. They applied their work on directed datasets to identify the Motifs, which means that the undirected networks cannot be analyzed by their algorithm. They also did not consider the problem of an imbalance dataset in their model. Datasets in link prediction are very imbalanced because the positive class (existing links) is in minority compared with nonexisting links.

Rafiee et al. [60] proposed link prediction based on common neighbors degree penalization (CNDP) and used quasilocal techniques instead of using local or global techniques, by considering neighbors of neighbors instead of only direct neighbors. They defined a new similarity score which is determined according to the topological structure of the network including common neighbors of each two nodes and the average clustering coefficient of the network. They performed their experiments on BUP, CEG, SMG, and HMT datasets and showed that their proposed method was superior to other methods such as adaptive degree penalization (ADP), the node-coupling clustering coefficient of node (NCCCN), and Triadic measure. They showed that contributing the common neighbors of three Triadic nodes including x , y , and z as the two desired nodes for calculating their similarity (x, y) and common neighbor (z) can affect the prediction performance. They also calculated the degree penalization scores in terms of area under the curve (AUC) and precision metrics using regression over different datasets which improved their work compared with the ADP algorithm that used a fixed value for degree penalization.

Most of the studies consider a static network to assess their method while this is a rough approximation because all social networks are updated over time. In this regard, Yao et al. [61] introduced a hybrid method based on link prediction that is assessed over dynamic networks with three metrics: the time-varied weight (revealing topological structure changes over time), the degree of common neighbors, and the intimacy between common neighbors. The last metric checks whether two common neighbor nodes have mutual relationships and checks the probability of a link appearing between the increases. They redefined the common neighbors by considering the nodes within two hops to achieve better performance. In their proposed method, they only focused on the topological features of common neighbors that can change over time, but the other metrics related to other features such as the node's attributes or graph topology on a larger scale (communities) or the change of path between two nodes were ignored. Moreover, one of their metrics, which check if intimacy between the common neighbors exists or not, is independent of time and can be used in static networks. They performed their experiments on DBLP datasets and used the Jaccard similarity criterion. They also used TPR and execution time to evaluate the performance of their method.

Bao et al. [61] proposed an incremental dynamic link prediction method that predicts relative links in dynamic social networks. Their algorithm called DLP-IRA is based on the resource allocation (RA) algorithm which is an algorithm based on the nodes' degree and considers the fact that the smaller the degree two nodes have, the higher the chance of establishing a link. Therefore, they defined a weight on each nonexisting link which is inverse of the maximum degree of source and destination nodes. They also stated that common neighbors of two nodes have a huge impact on link appearance, so they contributed the common neighbors' degree in their calculation. To adapt their algorithm with the dynamic network, they update the changing node set S with all the nodes with changed connected links. When no new node was added, subgraphs including the nodes in S and

their neighbors are considered as the input of their improved resource allocation algorithm. They claimed that their algorithm can avoid recomputing when the graph structure changes and also speeds up the process. The basis of their algorithm is nodes' degree, and no other extra information of the network had contributed. Even though the statement of the reverse correlation between the nodes' degree and their chance of link creation is generally true, in many situations, it is totally reverse, like when a TV star becomes famous, and as his popularity increases, more fans follow him, or when a person is extrovert, he tries to connect with more and more people. Furthermore, they did not consider the graph change evolution over time and just relied on the last change.

3. Proposed Method

In this research, a combined method based on three phases has been proposed for link prediction (suggesting friends) in social networks. In this regard, one can easily combine local, global, and community information in very large spaces with millions of users and billions of links, where it can increase the prediction accuracy and cover all possible modes that solve the problem of data scattering. It also improves the problem of the cold start of users and the accuracy of the offer. The three stages of the proposed method are (i) community detection, which detects dense communities with the node perception algorithm, (ii) optimization task with binary particle swarm optimization (BPSO) algorithm, and (iii) applying link prediction algorithm to the output of the former phase. The roadmap of the proposed method is illustrated in Figure 1 and Tables 1 and 2.

Since the simulations and algorithms are applied to the graph, we shortly discuss the role of graphs in link prediction systems.

3.1. Graph Theory. Interactions on social networks and their structure can be simulated and implemented on a graph. A graph $G = (V, E)$ consists of a set of nodes V and a set of edges E , where the edges connect the nodes. V demonstrates the role of users in the social network and E is the relationship between users. Adjacent matrix A of graph G is a two-dimensional matrix; if two users are connected, the corresponding edge is set to 1; otherwise, it sets to zero as shown in Figure 2.

3.2. Community Detection. As shown in Figure 1, the first phase is community detection. To detect the communities, nodes should be clustered, and each cluster can be considered as a community. For each node, we determined the clustering coefficient (CC) as follows:

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

where Γ_x demonstrates the neighbors of node x and $|\Gamma_x|$ is the degree of node x . $E_{y,z}$ represents the edge that exists between two nodes y and z . Therefore, to calculate the clustering coefficient of node x , all the edges between the

FIGURE 1: Diagram of the proposed method.

TABLE 1: Nomenclature table.

Symbol	Meaning
$ V $	Number of nodes
Γ_x	Neighbors of node x
$ \Gamma_x $	Degree of node x
$E_{y,z}$	Edge between two nodes y and z
C_x	Clustering coefficient of node x
CC	Clustering coefficient
ACC	Average clustering coefficient
t	Time snapshot
X_{ij}	A binary array for indicating the communities i and j are merged together or not
V_{ij}	Velocity function as a probability threshold value that determines whether the position vector is zero or one
$s(z)$	Sigmoid function that takes the velocity function of next time as input

TABLE 2: The statistical information of experimental datasets [51, 52, 63–65].

Dataset	NN	NE	NC	MO	DA	AD	ACC	NT
Astro-ph	18771	198050	321	0.619	014	21.09	0.677	1351441
Blogs	1224	19025	010	0.426	008	31.38	0.213	101043
CiteSeer	3312	4732	467	0.887	008	1.421	0.080	1223
Cora	2708	5429	098	0.805	015	3.891	0.293	1630
WebKB	0877	1608	043	0.757	010	3.372	0.332	389

neighbors of node x are counted and divided into the total number of possible mutual relationships between the neighbors of node x . In other words, we determine how many people (nodes) are neighbors of our desired node, and among whom, how many of the nodes are friends with each

other. More friendships between the neighbors of the desired node lead to its higher CC value. After determining the CC value of all nodes in each community, we determined the value average clustering coefficient (ACC) as the aggregation of that community according to the following equation:

FIGURE 2: As shown above, a graph consists of three users, and their interaction is expressed in the corresponding matrix.

$$ACC = \frac{\sum_{x \in V} C_x}{|V|}, \quad (2)$$

where C_x represents the clustering coefficient of node x and $|V|$ is the number of nodes in the graph. After determining the ACC of each community, all of them are summed together, and the result is divided into the number of the communities. To detect each community by the node perception algorithm, first, we identify small communities [23]. This algorithm has an acceptable performance even when encountering millions of users and billions of links. This method has some properties compared to other competitive methods in which one user can be assigned to more than one cluster. This algorithm uses three hyperparameters to perform clustering. To properly set the parameters, the ensemble of clustering is used, which performs a greedy search on parameters in the search space. This algorithm consists of two phases that run alternately. Suppose that we are dealing with a network with n vertices. First, we take each vertex equal to a community. Hence, first, we have several communities equal to the number of vertices; then for each vertex i , we find the neighboring community j in a way that for removing i from its community and joining it to community j , modularity is maximized, and then, we add vertex i to community j . This is done only if the modularity is increased; otherwise, vertex i remains in its community. This is repeated frequently for all vertices until no change happens. At this stage, the first phase is finished. It should be noted that in the first phase, a vertex may be moved several times between different communities. The first phase stops at a locally optimal point. Then, in the second phase, clustering continues by merging small groups that have the ability to form larger groups. We continue this process until we reach the desired number of communities or exceed a certain threshold. This process should satisfy some constraints. The size of each community should be lower than a threshold.

We have an overlap threshold that controls the overlap tolerance of communities. A threshold is determined for the tolerance required to merge communities. The threshold parameter is set between 0.1 and 1, and we change the value of this parameter by increasing step 0.1. We set the smallest community size that can be detected. To properly set the parameters, we used an ensemble of clustering. The ensemble is a set of parameters set in node preservation with different values, which is considered to increase the value of aggregation indicating the ACC. To obtain the ACC value in a community, we must first calculate the value of the clustering coefficient for all nodes in the community based on equation (1). Then, based on equation (2), we determine the ACC of each community, and exceeding this value from a threshold implies that it is a valid community.

3.3. Clusters' Optimization by BPSO. Particle swarm optimization (PSO) is one of the known evolutionary algorithms introduced by Kennedy and Supermart (1995). Similar to other metaheuristic algorithms, PSO starts with a population of particles, each having the potential to be a solution to the problem. These particles are first randomly initialized all over the search space. After determining the fitness value of all particles, one of them has the highest fitness score, and it attracts all particles toward itself. After the second iteration, particles are moved according to two directions: the direction to the position of the best particle (gbest) and the direction to the position of best fitness that the corresponding particle has experienced (pbest). Unlike other evolutionary computational techniques, each particle (particle) in PSO depends on the position, velocity, and acceleration. The particles continuously move in the search space at an acceleration that is dynamically calculated according to their previous behaviors. Meanwhile, through the movement of particles, two random parameters try to insert a degree of randomness to enhance the exploration of this algorithm while PSO preserves a good exploitation property. If through the movement of particles, the fitness of one particle exceeds gbest, this particle is considered as the new gbest [24, 25].

Due to the discrete nature of our problem, the continuous operators of PSO should be discretized. In binary PSO (BPSO), introduced by Kennedy and Eberhart [24, 25], velocity is used as a probability threshold value that determines whether the position vector is zero or one. Suppose that X_{ij} is the value of the j^{th} bit of the binary vector which represents the position of the i^{th} particle. In this case, the following equation describes how the binary PSO algorithm works:

$$X_{ij}(t+1) = \begin{cases} 0 & \text{if } \text{rand}(\cdot) \geq S(v_{ij}(t+1)) \\ 1 & \text{if } \text{rand}(\cdot) < S(v_{ij}(t+1)) \end{cases}, \quad (3)$$

$$S_x = \frac{1}{1 + e^{-x}},$$

$$v_{ij}(t+1) = m * v_{ij}(t) + c_1 r_{1j}(t) [y_{ij}(t) - z_{ij}(t)] + c_2 r_{2j}(t) [y_j(t) - z_{ij}(t)],$$

where X_{ij} is an array with 1 and 0 elements, which, respectively, indicate if the corresponding community should merge to other communities or not. t indicates the iteration, v_{ij} is a velocity function, $s(z)$ is a sigmoid function that takes the velocity function of the next time as an input, and then, if the output of the sigmoid function is greater than a random number, the value of 1 is placed in X_{ij} array for the corresponding community; otherwise, 0 is placed.

In binary PSO, the particle velocity is changed similar to the standard PSO algorithm. In addition to introduced parameters, the number of particles (initial population), which can increase the exploration, and the maximum number of iterations are important parameters of the algorithm. BPSO optimizes the communities that were detected in the first stage to achieve greater ACC. The fitness function of each particle (network) is the summation of ACC of its corresponding elements (community). We come to an array for each particle whose size is equal to the number of the communities. Then, we specify the position of each community, which can be zero and one. It should be noted that for multiple communities, integrated by BPSO, ACC plays the main role in merging communities. A high value indicates that the integration has been done properly; otherwise, we ignore merging.

For example, if we have a network with 5 communities, we consider a 5×1 array whose elements are 0 or 1, which is initialized at random. For example, if the array elements are $[0, 0, 0, 1, 1]$, this particle means that the first and fourth communities should be merged. Then, for all the resulting community nodes, we calculate the clustering coefficient and consequently calculate the ACC as the fitness of the integrated community. These operations are repeated until all possible particles of integration are checked. This is where the work of the second step ends, and the best possible communities were selected.

3.4. Link Prediction. In the first phase, we had an initial graph, and we came to divide this network into smaller communities that had more density. Then, in the second phase, we optimized the communities of the first part by BPSO. We created subgraphs that merged with each other to give better fitness results, leading to the production of more valuable communities. In this regard, we considered each of these subgraphs as input and using the Adamic/Adar index method (AA) to find the suggested link according to equation (4). Then, for each edge of the graph based on the obtained score in equation (4), a high value indicated that the edge can appear.

$$\text{Score}(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log|\Gamma_z|} \quad (4)$$

where x and y are two nodes, which we tend to calculate their similarity and z is the set of their common neighbors. Γ_z demonstrates the neighbors of node z and $|\Gamma_z|$ is the degree of node z . In this regard, we find common neighbors for two nodes (x, y) , then for each common neighbor obtained, we divide 1 by the log (logarithm) of the desired number of

nodes, and finally, the values obtained from the common nodes are added together, which results in a score for each edge connected to the two nodes. This score indicates the amount of similarity between the two nodes x and y , creating a link between them. Assuming that if x and y are nodes that have a large number of common neighbors, then according to the AA index, a higher score is determined. In other words, it can be said that there is a high probability that a connection is established between x and y in the future (link). At this stage, a threshold is considered, and based on it, if the value of that edge exceeds this threshold, it means that there is a very high probability that there will be a link between them; otherwise, no link is suggested. For example, in Figure 3, to calculate AA (2, 5), we have

$$\text{AA}(2, 5) = \sum_{z \in \Gamma_2 \cap \Gamma_5} \frac{1}{\log|\Gamma(z)|} \quad (5)$$

4. Datasets and Evaluation Criteria

To evaluate the proposed method, we divided the edges in the graph into two categories: E_train and E_test. E_train are the edges that are used for calculating the similarity score between each pair of nodes in the graph and E_test are the edges that are used to validate the proposed method. Based on this, we use the following data to show the effectiveness of the proposed method. The employed datasets are Astro-ph [26], Blogs [27], CiteSeer [28], Cora [26], and WebKB [28], and the features that make up the dataset can be expressed as the number of nodes (NN), number of edges (NE), number of communities (NC), modularity (MO), diameter (DA), average degree (AD), average clustering coefficient (ACC), and number of triangles (NT).

The criteria used to assess the proposed method, compared to the counterparts, are expressed as follows. The confusion matrix includes true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Based on these four parameters, it is possible to determine sensitivity, specificity, precision, recall, TP rate (TPR), FP rate (FPR), receiver operative curve (ROC), and (AUC). Precision means the number of correctly identified links divided by the total number of links classified as positive, and accuracy means the number of correctly identified links divided by the total number of links. Recall means the number of classified positive links divided by the total number of real positive links. ROC curve sketches TPR versus FPR, and AUC is the area under ROC. Important indicators that have shown their performance in link prediction are Adamic/Adar (AA) index, common neighbor (CN), Jaccard index (JC), and KI.

5. Experimental Results

The purpose of the experiments was to evaluate the proposed method (ECLP) versus state-of-the-art methods over the introduced datasets in terms of the described criteria. The tests were performed on a system with Intel Core i7 GHz CPU specifications and 16 GB RAM. The test results of the proposed method (ECLP) were tested and compared with

FIGURE 3: Comparison of the algorithms for different datasets in terms of AUC. (a) Comparison of the algorithms together for each dataset separately. (b) The accuracies obtained for different datasets for each algorithm.

KI, CN, JC, AA, and GLP methods. It should be noted that in the deployed graphs, the edges are considered weightless.

In order to evaluate our new method, we use fivefold cross-validation which divides the data into five folds; each time picks a fold as test data and keeps the rest as train data. Then, based on the proposed method, the similarity scores are calculated for the nodes in four folds of train data, and these scores are evaluated on the test fold and thus accuracy is determined. This process is repeated for all five folds. Finally, AUC is calculated based on the five accuracy values obtained in fivefold cross-validation and this average accuracy is recorded in Table 3.

As shown in Table 3, the proposed method works well in real-world datasets. Evaluation of experimental results is expressed in terms of the AUC metric. Table 3 indicates that our method is significantly improved compared to state-of-the-art methods. A comparison of columns two and three in Table 3 shows that ECLP outperforms the GLP algorithm in all datasets except one of them, and by comparing ECLP to the methods of columns four to seven, it is obvious that ECLP significantly improved (more than 4%) the prediction accuracy in terms of AUC.

Figure 3 visualizes the quantitative results of Table 3, to better understand the difference between the aforementioned algorithm results. Figure 3(a) represents the accuracy (AUC) values for a unique dataset when different algorithms are applied. As shown in Figure 3, ECLP outperforms other algorithms for the four datasets (Blogs, CiteSeer, Cora, and WebKB). It is noteworthy to say that ECLP performs better than even GLP for these four datasets, especially for the CiteSeer dataset, where there was a big difference between ECLP and GLP accuracy. Figure 4 shows the results in a different way, facilitating comparison among the algorithms' performance. It is so obvious that ECLP performs better than state-of-the-art algorithms such as GLP, AA, CN, JC, and KI in terms of AUC criterion. Moreover, it can be interpreted that ECLP performs better on two datasets (Cora and CiteSeer) compared with the other two.

TABLE 3: Accuracy of the proposed and compared methods (in terms of AUC).

Dataset	Accuracy					
	ECLP	GLP	CN	AA	JC	KI
Astro-ph	0.9340	0.9313	0.8912	0.8653	0.8732	0.8695
Blogs	0.9559	0.9014	0.8786	0.8679	0.8274	0.8483
CiteSeer	0.9985	0.8612	0.8534	0.8658	0.8677	0.8565
Cora	0.9972	0.9455	0.9123	0.8975	0.8865	0.9078
WebKB	0.9768	0.9468	0.9233	0.8897	0.9011	0.9236
Average (%)	97.2	91.7	89.1	87.7	87.1	88.1

FIGURE 4: Comparison of the algorithms in terms of standard deviation.

As shown in Figure 4, our proposed method had a lower standard deviation compared to the other algorithms except CN. This indicates that our proposed method is more robust.

6. Conclusion

This paper proposes a three-stage scheme by optimizing the community detection to improve link prediction precision in social networks. In the first stage of the proposed method, we are faced with an initial graph, which should be divided into smaller dense subgraphs (communities). This is performed by applying node ensemble perception, which

incorporates global information. Then, in the second phase, we applied BPSO to optimize the communities of the former phase and, therefore, merge those communities that alleviate the fitness function. The second phase preserves semilocal information of the network. The third phase computes the similarity between each pair of nodes by applying Adamic/Adar similarity-based algorithm and then the edges with a high similarity score considered as the predicted edges of the algorithm. This phase enables ECLP to capture local information of the network. Moreover, our method considers the overlap between communities which means a node may belong to multiple communities while GLP uses the Louvain algorithm for community detection that does not consider the overlapping between communities. Our proposed algorithm was evaluated and analyzed by applying it to different datasets, and the results indicated that it performs better than all rival algorithms.

Data Availability

The data used to support the findings of this study are available from the first author, Hasan Saeidinezhad, upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] J. Zhao, J. Liu, J. Jiang, and F. Gao, "Efficient deployment with geometric analysis for mmWave UAV communications," *IEEE Wireless Communications Letters*, vol. 9, 2020.
- [2] Z. Xiong, N. Xiao, F. Xu et al., "An equivalent exchange based data forwarding incentive scheme for socially aware networks," *Journal of Signal Processing Systems*, vol. 93, 2021.
- [3] K. Shi, Y. Tang, X. Liu, and S. Zhong, "Secondary delay-partition approach on robust performance analysis for uncertain time-varying Lurie nonlinear control system," *Optimal Control Applications & Methods*, vol. 38, 2017.
- [4] M. Mortazavi, F. T. Hocanin, and A. Davarpanah, "Application of quantitative computer-based analysis for student's learning tendency on the efficient utilization of mobile phones during lecture hours," *Sustainability*, 2020.
- [5] B. Daryayehsalameh, M. Nabavi, and B. Vaferi, "Modeling of CO₂ capture ability of [Bmim][BF₄] ionic liquid using connectionist smart paradigms," *Environmental Technology & Innovation*, vol. 22, 2021.
- [6] A. Davarpanah, R. Shirmohammadi, B. Mirshekari, and A. Aslani, "Analysis of hydraulic fracturing techniques: hybrid fuzzy approaches," *Arabian Journal of Geosciences*, vol. 12, 2019.
- [7] X. Ma, K. Zhang, L. Zhang et al., "Data-driven niching differential evolution with adaptive parameters control for history matching and uncertainty quantification," *SPE Journal*, vol. 26, 2021.
- [8] R. Dastoorian and L. Wells, "Gauge capability studies for high-density data: SPC phase 0," *Procedia Manufacturing*, vol. 48, 2020.
- [9] M. Abasi, A. Saffarian, M. Joorabian, and S. G. Seifossadat, "Fault location in double-circuit transmission lines compensated by generalized unified power flow controller (GUPFC) based on synchronous current and voltage phasors," *IEEE Systems Journal*, 2020.
- [10] M. Abasi, A. Saffarian, M. Joorabian, and S. Ghodrattollah Seifossadat, "Fault classification and fault area detection in GUPFC-compensated double-circuit transmission lines based on the analysis of active and reactive powers measured by PMUs," *Journal of the International Measurement Confederation*, vol. 169, 2021.
- [11] S. Qu, W. Xu, J. Zhao, and H. Zhang, "Design and implementation of a fast sliding-mode speed controller with disturbance compensation for SPMSM system," *IEEE Transactions on Transportation Electrification*, 2021.
- [12] X. Xue, K. Zhang, K. C. Tan et al., "Affine transformation-enhanced multifactorial optimization for heterogeneous problems," *IEEE Transactions on Cybernetics*, vol. 8, 2020.
- [13] A. Davarpanah, "Parametric study of polymer-nanoparticles-assisted injectivity performance for axisymmetric two-phase flow in EOR processes," *Nanomaterials*, vol. 10, 2020.
- [14] Q. Jiang, F. Shao, W. Lin, K. Gu, G. Jiang, and H. Sun, "Optimizing multistage discriminative dictionaries for blind image quality assessment," *IEEE Trans Multimed*, vol. 20, 2018.
- [15] Q. Jiang, F. Shao, W. Gao, Z. Chen, G. Jiang, and Y. S. Ho, "Unified No-reference quality assessment of singly and multiply distorted stereoscopic images," *IEEE Transactions on Image Processing*, vol. 28, 2019.
- [16] S. He, F. Guo, Q. Zou, and H. Ding, "MRMD2.0: a Python tool for machine learning with feature ranking and reduction," *Current Bioinformatics*, vol. 15, 2020.
- [17] S. Lv and Y. Liu, "PLVA: privacy-preserving and lightweight V2I authentication protocol," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, 2021.
- [18] B. H. Li, Y. Liu, A. M. Zhang, W. H. Wang, and S. Wan, "A survey on blocking technology of entity resolution," *Journal of Computer Science and Technology*, vol. 25, 2020.
- [19] B. Li, R. Liang, W. Zhou, H. Yin, H. Gao, and K. Cai, "LBS meets blockchain: an efficient method with security preserving trust in SAGIN," *IEEE Internet of Things Journal*, vol. 1, 2021.
- [20] L. J. Wells, R. Dastoorian, and J. A. Camello, "A novel NURBS surface approach to statistically monitor manufacturing processes with point cloud data," *Journal of Intelligent Manufacturing*, vol. 32, 2021.
- [21] C. Zuo, Q. Chen, L. Tian, L. Waller, and A. Asundi, "Transport of intensity phase retrieval and computational imaging for partially coherent fields: the phase space perspective," *Optics and Lasers in Engineering*, vol. 71, 2015.
- [22] C. Zuo, J. Sun, J. Li, J. Zhang, A. Asundi, and Q. Chen, "High-resolution transport-of-intensity quantitative phase microscopy with annular illumination," *Scientific Reports*, vol. 7654, 2017.
- [23] Y. Zhou, L. Tian, C. Zhu, X. Jin, and Y. Sun, "Video coding optimization for virtual reality 360-degree source," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, 2020.
- [24] H. Sheng, S. Wang, Y. Zhang et al., "Near-online tracking with Co-occurrence constraints in blockchain-based edge computing," *IEEE Internet of Things Journal*, vol. 8, 2021.
- [25] K. Zhang, J. Zhang, X. Ma et al., "History matching of naturally fractured reservoirs using a deep sparse autoencoder," *SPE Journal*, 2021.
- [26] A. Armin Razmjoo, A. Sumper, and A. Davarpanah, "Energy sustainability analysis based on SDGs for developing

- countries," *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, vol. 42, 2019.
- [27] H. Safari-Katesari, S. Y. Samadi, and S. Zaroudi, "Modelling count data via copulas," *Statistics*, vol. 54, 2020.
 - [28] H. Safari Katesari and B. Fathi Vajargah, "Testing adverse selection using frank copula approach in Iran insurance markets," *Journal of Mathematical and Computational Science*, vol. 15, 2015.
 - [29] S. S. Kalantari and A. A. Taleizadeh, "Mathematical modelling for determining the replenishment policy for deteriorating items in an EPQ model with multiple shipments," *Journal of Systems Science: Operations & Logistics*, vol. 7, 2020.
 - [30] A. Roghani, S. H. Nyarko, and C. Sparks, "The first family formation among young Americans: the role of family process," *SN Social Science*, vol. 50, 2021.
 - [31] A. Chapnevis, I. Güvenç, and E. Bulut, "Traffic shifting based resource optimization in aggregated IoT communication," in *Proceedings of the 45th IEEE Conference on Local Computer Networks (LCN)*, Sydney, Australia, October 2020.
 - [32] M. Masoumnezhad, M. Rajabi, A. Chapnevis et al., "An approach for the global stability of mathematical model of an infectious disease," *Symmetry (Basel)*, vol. 12, 2020.
 - [33] S. Arasteh, M. Mahdavi, P. N. Bideh, S. Hosseini, and A. Chapnevis, "Security analysis of two key based watermarking schemes based on QR decomposition," in *26th Iranian Conference on Electrical Engineering (ICEE)*, Mashhad, Iran, May 2018.
 - [34] A. Chapnevis and B. Sadeghiyan, "A secure two-party computation protocol for intersection detection between two convex hulls," 2020, <https://arxiv.org/abs/2011.00319v3>.
 - [35] D. M. Boyd and N. B. Ellison, "Social network sites: definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, 2007.
 - [36] A. Azma, E. Narreie, A. Shojaaddini et al., "Statistical modeling for spatial groundwater potential map based on gis technique," *Sustainability*, vol. 13, 2021.
 - [37] F. Li, J. He, G. Huang, Y. Zhang, Y. Shi, and R. Zhou, "Node-coupling clustering approaches for link prediction," *Knowledge-Based System*, vol. 38, 2015.
 - [38] D. M. Boyd and N. B. Ellison, "Social network sites: definition, history, and scholarship," *IEEE Engineering Management Review*, vol. 38, 2010.
 - [39] M. Jalali Sarvestani and P. Charehjou, "Fullerene (C20) as a potential adsorbent and sensor for the removal and detection of picric acid contaminant: DFT Studies," *Central Asian Journal of Environmental Science and Technology Innovation*, vol. 2, no. 1, 2021.
 - [40] Y. Maina, B. Kyari, and M. Jimme, "Impact of household fuel expenditure on the environment: the quest for sustainable energy in Nigeria," *Central Asian Journal of Environmental Science and Technology Innovation*, vol. 1, no. 2, pp. 109–118, 2020.
 - [41] M. Mohammadi, "An in-depth study of specific pathway linked to abnormal reproductive system: a meta-analysis in a specific Caucasian ethnicity," *Central Asian Journal of Medical and Pharmaceutical Sciences Innovation*, vol. 1, no. 1, pp. 1–7, 2021.
 - [42] M. Sabernezhad, "Quantitative analysis of the p53 substitution mutation and breast cancer: an informative study in Iranian population," *Central Asian Journal of Medical and Pharmaceutical Sciences Innovation*, vol. 1, no. 1, pp. 8–14, 2021.
 - [43] R. Asgari, "The role of ESR1 PvuII T/C variant in the female reproductive process: a review," *Central Asian Journal of Medical and Pharmaceutical Sciences Innovation*, vol. 1, no. 1, pp. 22–27, 2021.
 - [44] S. Qayyum, I. Khan, K. Meng, Y. Zhao, and C. Peng, "A review on remediation technologies for heavy metals contaminated soil," *Central Asian Journal of Environmental Science and Technology Innovation*, vol. 1, no. 1, pp. 21–29, 2020.
 - [45] A. Rivera-Diaz, C. Patricia Ortiz, and D. Ricardo Delgado, "The crucial role of estrogen/androgen hormones and their receptors in male infertility risk," *Central Asian Journal of Medical and Pharmaceutical Sciences Innovation*, vol. 1, no. 1, pp. 35–43, 2021.
 - [46] T. Keim, "Extending the applicability of recommender systems: a multilayer framework for matching human resources," in *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, Big Island, Hawaii, January 2007.
 - [47] L. Jin, J. B. D. Joshi, and M. Anwar, "Mutual-friend based attacks in social network systems," *Computers and Security*, vol. 37, 2013.
 - [48] Y. Chen, V. M. Patel, P. J. Phillips et al., "An optimizing and differentially private clustering algorithm for mixed data in SDN-based smart grid," *IEEE Access*, vol. 7, 2018.
 - [49] X. Chen, D. Y. Wang, J. B. Tang, W. C. Ma, and Y. Liu, "Geotechnical stability analysis considering strain softening using micro-polar continuum finite element method," *Journal of Central South University*, vol. 28, no. 1, pp. 297–310, 2021.
 - [50] M. Abasi, M. Razaz, G. Seifossadat, and S. Moosapour, "Presenting a new formulation to analyze and determine unbalance voltage produced at the place of load resulting from network and loads unbalance and asymmetry of transmission lines in radial power systems," *Majlesi Journal of Energy Management*, vol. 6, no. 3, p. 4, 2015.
 - [51] J. Leskovec, "Astro physics collaboration network," <https://snap.stanford.edu/data/ca-AstroPh.html>.
 - [52] Lise's INQuisitive Students, "Link based classification," <http://www.cs.umd.edu/projects/linqs/projects/lbc>.
 - [53] L. L. Linyuan and T. Zhou, "Link prediction in complex networks: a survey," *Physica A: Statistical Mechanics and its Applications*, vol. 39, 2011.
 - [54] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, 2007.
 - [55] R. H. Li, J. X. Yu, and J. Liu, "Link prediction: the power of maximal entropy random walk," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, Glasgow, Scotland, October 2011.
 - [56] C. A. Bliss, M. R. Frank, C. M. Danforth, and P. S. Dodds, "An evolutionary algorithm approach to link prediction in dynamic social networks," *Journal of Computational Science*, vol. 5, 2014.
 - [57] A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos, "Fast and accurate link prediction in social networking systems," *Journal of Systems and Software*, vol. 85, 2012.
 - [58] L. Adamic and E. Adar, "How to search a social network," *Social Networks*, vol. 27, 2005.
 - [59] E. Bastami, A. Mahabadi, and E. Taghizadeh, "A gravitation-based link prediction approach in social networks," *Swarm and Evolutionary Computation*, vol. 44, 2019.
 - [60] S. Rafiee, C. Salavati, and A. Abdollahpour, "CNDP: link prediction based on common neighbors degree penalization,"

Physica A: Statistical Mechanics and its Applications, vol. 539, 2020.

- [61] L. Yao, L. Wang, L. Pan, and K. Yao, "Link prediction based on common-neighbors for dynamic social network," *Procedia Computer Science*, vol. 83, 2016.
- [62] Z. Liu, Q. M. Zhang, L. Lü, and T. Zhou, "Link prediction in complex networks: a local naïve Bayes model," *EPL*, vol. 96, 2011.
- [63] <https://paperswithcode.com/dataset/citeseer>.
- [64] <https://paperswithcode.com/dataset/cora>.
- [65] <https://paperswithcode.com/dataset/webkb>.