International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 6, Issue 3, 2019, PP 31-38
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
DOI: http://dx.doi.org/10.20431/2349-4859.0603004
www.arcjournals.org

Improved Link Prediction in Social Networks using Label Propagation

Shadi sharifpour[1*], Mehdi bateni[2]

[1]*Masters student, Department of Computer Engineering, Sheikhbahaei University, Isfahan, Iran.*

[2]*Assistant Professor, Department of Computer Engineering, Sheikhbahaei University, Isfahan, Iran.*

Corresponding Author:** ***Shadi sharifpour, *Masters student, Department of Computer Engineering, Sheikhbahaei University, Isfahan, Iran.*

Abstract: *Social communications can effectively assist social networks in order to keep their active users and improve services of social networks. Therefore, they should have good prediction of social communications between their users. In this way, link prediction issue in social networks has interested many researchers and increasing prediction accuracy in social networks in under focus. There are three approaches for solving link prediction problem: similarity-based approach, an approach based on maximum likehood and probabilistic model-based approach. In this paper, similarity-based approach is utilized. Also, label propagation concept is used for link prediction and in order to raise accuracy in link prediction algorithm by using label propagation, the distance of the shortest path is applied for label propagation in network. Presented method, with AUC metric is tested on football data set, Email network, science network and power network. Total results of this test show that proposed method is improved basic method about 17.133%.*

Keywords: *social networks, link prediction, label propagation, shortest path distance*

1. INTRODUCTION

Many systems in real world have high complexity. Complexity of these systems is due to high number of entities and their relations that makes them unrealizable. Network modeling of the relation between systems is among methods for reducing the complexity of these systems. These networks are modeled as graph, where entities are illustrated by nodes and relation between entities are showed by edges. Many of social, biological and information systems can be described nicely by networks. So, study on complex networks includes social networks, is interested by many of researchers.

Social networks in online environment are active through social networks' channels through, also, social networks let people to state their relations. Relation with others or link them to other results in potential benefits such as cooperation improvement and data sharing, increase productivity and advanced communications between participants, business participants and customers [1].

Social networks analyzers are following the most optimized communication networks, discover common patterns in such networks and tracking information stream (and other resources) through them. Recognizing these relations and networks, effects on people and organizations. Therefore, social network analyzers, study on two people relation and interpret their performance according to relation of these two people with other members of network [2].

Link prediction is a well-studied and challenging data mining task associated with network theory and social network analysis [3]. Link prediction means likelihood prediction of future communication between two nodes, assuming that currently there is no relation between these two nodes. Link prediction problem can be studied based on three general approach [4]: most likelihood-based approach where by using detailed rules and special parameters that are achieved by maximizing observed structure likelihood, a model for organizing network structure is assumed and probability of making any non-existent link is determined based on these rules and parameters. Common models for network organizing are hierarchical structure model and stochastic block model. Approach based on probabilistic model, where, first a model includes sums of adjustable parameters is created and optimal value of these parameters are discovered by using optimization strategy; such that resulted model can reflect better structures and relations from network characteristics. In the approach, three basic methods

are include probabilistic relational model, probabilistic entity relationship model and stochastic relational model. Similarity-based approach, where all parities of non-existent nodes are ranked based on similarity index, and probability of making link between those pair nodes with maximum similarity, is higher. Similarity-based approach includes two sets of indices: similarity indices based on node properties that use node-related information such as their profile in online social networks for explaining similarity between nodes. Similarity indices based on structure that use network structural (topology) information in order to calculate similarity between two nodes. Structure-based similarity indices are regarded from three viewpoints: local, global and quasi-local indices. Local indices use neighborhood information of nodes.

Global indices need global topologic information, and quasi-local indices do not need global topologic information but use more information than local indices.

In this study, a similarity based approach is used. Also, label propagation concept is used for link prediction and in order to increase accuracy in link prediction algorithms by using label propagation, shortest path distance is used for propagating label in network.

Following, in second section, last researches in link prediction are studied. Then, in section III, link prediction algorithm by using label propagation is studied. In section IV, algorithm optimized method is explained in details. Section V studies on results and finally, discussion and conclusion is presented in section VI.

2. RELATED WORK

Label Propagation [5] is a network community detection algorithm first proposed by U. N.Raghavan et al. An effective and efficient method that uses network structure for process evaluation and does not need any prerequisite such as community number and size. Assume that X node is in network. X has neighbors X_1, X_2,…X_K that each neighbor includes a label that indicates its community type. For example, X_1 belongs to societ1 and X_2 belongs to community 2. In this algorithm, each nodes in graph has a dedicated label. Then, in each step, each node take its neighbor's label and replace more frequented label with its one. If label frequency is even, then randomly selects a label. At last, all nodes in community, get a label.

Research [6] are introduced link prediction based on node-connected clustering approach for predicting link. In this connection method, nodes of common neighbor mix a predicted node with cluster geometry of node. Approaches in this study not only consider coordination between connection degree of nodes of common neighborhood and cluster information of network, but also introduces different roles of nodes for link prediction. According to tests of this study, approaches of clustering information of a network by using clustering each node, need less time complexity. Also, approaches in this study are more proper for very large scale networks.

Ref. [3] use graph parallelism for link prediction. Bulk synchronous parallel model is an invented framework for parallel graph algorithms. This paper uses parallel label propagation algorithm to detect community and a Adamic-Adar-based metric based on parallelism information in order to link prediction. These algorithms are coordinated by using infrared synchronous programming model and are tested through large network with various domains.

Cui et al. [7] stated that link prediction method based on common neighbor-based indices have tremendous progress. This research proposes that if limit is considered under similarity degree of node-pair, then it would be possible to considerably improve this method.

Sun et al. [8] have introduced concept of node degree and the idea of community structure, proposed a new similarity index called "local affinity structure", and described local affinity structure of a node and their common neighbors.

In [9] have proposed balanced factorization machine to facing challenges of link prediction, where link predictions in very scattered networks are implemented through learning interactions between network nodes and edges in a supervised learning environment.

Ding et al. [10] in order to link prediction, have extracted community structure by using network local information. Then, relation of each pair of societies are calculated by using new indices of community relationship and finally use a simple prediction model based on linear deduction to evaluate probability of non-existed links.

In [11] have proposed a new idea for link prediction by merging two methods of link prediction based on similarity score and stochastic link prediction that is placed in a new category of link prediction methods. This idea uses probabilistic techniques to obtain similarity scores between nodes which has better results than other methods in standard data sets.

Ahmed et al. [12] have introduced a fast similarity-based method for predicting probabilistic links in time networks. In this method, first snapshots of network are connected to a weight diagram. Then, through stochastic walking, an under axis graph is created in each node in weigh diagram. This under-graph area includes a set of start paths of considered node. Because similarity score in such small axis under-graphs are calculated in each node, then algorithm can reduce considerably the calculation time.

3. MATERIALS AND METHODS

3.1. Link Prediction Algorithm by using Label Propagation

In [13] have proposed a type of improved label propagation process related with link prediction that uses dynamic interaction processes in order to obtain likely links. Also, uses network structure that is very effective in predicting non-existent links and future links. Proposed algorithm in this paper that is introduced by link prediction by using improved label propagation, in first iteration assign a distinct memory and special label to each node in network (Fig1(a)). In next iteration processes. Each node is randomly selected to get a label from its neighbors. Then, neighbors randomly select a label from their memory and send for applicant node. For example, assume that in Fig1 graph, nodes {3,5,1,4,2}are selected randomly as receiver list. Therefore, node 3 is the first receiver of information from its neighbors. It is clear that node3 receives labels 1,2 and 5 because its neighbors just have one label in their memory.

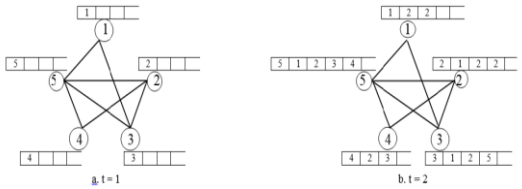

Fig1. *The trend of ILP label propagation*

Then, node5 is the next receiver. When node 3 sends its label, randomly selects one of the elements {3,1,2,5} from memory because in this time, node 3 is finished second iteration. Assume node 3 randomly selects and sends label 3, so node5 includes label {5,1,2,3,4} in its memory.

During the receiving process, assume that node 1 receives two label 2 from nodes 3 and 5, node 4 receives label 2 from node 2 and label 3 from node 5, node 2 receives labels {1,2,2} from nodes 3,4,5. Fig1(b) shows final result after second iteration.

In order to clear the trend of calculating the possibility of likely links, trend of label reproduction is stopped after second iteration. Fig2 shows that three links (1,2), (1,4) and (3,4) are non-existent in graph. So, in order to get possibility of link (1,2), memory of nodes 1 and 2 are investigated. There are two label 2 in memory of node1. Instead, there is a label 1 in memory of node 2. Therefore, possibility of making a link between node 1 and node 2 is MAX {1,2}=2. Also the possibility of making a link between nodes (1,4) is zero and possibility of making link between nodes (3,4) is 1.

3.2. Improved Method

Input of problem is improved method, a undirected and unweight network, and output of algorithm are links that are predicted for network. In first step, a memory is assigned to each node of network. Main step of this algorithm is filling memory of nodes by using the shortest path algorithm. In next step, non-existent links are scored by using similarity of labels located in memory, and finally link prediction is implemented according to these scores. Steps of implementing proposed method is illustrated in block diagram of Fig2.

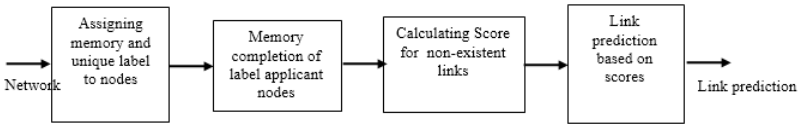

Fig2. *Block diagram of main steps of research*

3.2.1. Assigning Memory and Unique Label to Nodes

A simple, undirected and un-weighted network is the input of this step. In this step, a dedicated memory is assigned to each node of network, where a unique label that is considered for each node, is stored in first cell of memory of each node (Fig.3). Then, the nodes centrality is calculated based on degree centrality, i.e. nodes with higher degree are placed in center. Therefore, nodes are selected based on their degree centrality order and receive the label of their neighbors. In Fig.3 nodes selection based on their degree centrality are 5,3,2,4,1 respectively.

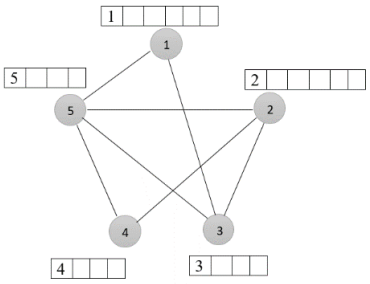

Fig3. *Memory assignment for each node and storing the unique label of each node*

3.2.2. Memory Filling of Label Requested Nodes

In this step, the first node is selected based on higher centrality, to receive label from adjacent nodes. Then, adjacent nodes use shortest path algorithm to transfer in-memory labels such that, adjacent node selects a label from its memory with minimum distance to itself. In must be noted that in this step, adjacent nodes beware of calculating the shortest path between their node and node with label of applicant node. Also, do not consider the shortest path to their node. If there are a number of label with the same distance then, adjacent node randomly selects one of the labels and send for nodes requested for label. In the following, to better understand the trend, steps of proposed method are discussed with an example.

In Fig.4, according to maximum degree centrality, node5 is selected as the first node that receives label from its neighbors. Because adjacent nodes of node5 only have one label in their memory, they send the same labels for node5 and node5 receives labels 1,2,3,4 from its neighbors.

Due to degree centrality of nodes, node3 is selected as next node to complete its memory with labels that are received from its neighbors. Node3 is the node that receives label from nodes 1,2,5. Because node5 complete its memory in this step, transfer a label from its memory with the minimum distance to itself. As the shortest path from node5 to itself is zero, it is not calculated. Also, calculating the shortest path to label of receiver node is ignored (label3) because, selection and transferring this label to receiver node does not have any effect on calculating the score between non-existent links. Hence, according to these conditions, the shortest paths between node5 and label of nodes 1,2,4 are calculated. Because the distance of the shortest path between label 3 and 5 is ‘1’, node5 randomly selects one the labels and sends for node3. Here, consider that node5 selects label4 and sends it for node3. Nodes 1,2 also because have one label in their memory, select label 1 and 2 for node3, respectively.

Due to degree centrality of nodes, node2 is considered as the next node. According to Fig.4, adjacent node3 includes labels {3,1,2,4} in its memory where labels 2,3 are not considered for calculating the shortest path in proposed method and the distance of shortest path from node3 to other labels 3,4 is calculated. The distance of the shortest path from node3 with labels1,4 are 1 and 2 respectively. Consequently, node3 sends label1 to node2 because has the minimum distance of the shortest path than other labels.

Node5 also randomly selects label1 from labels {1,3,4} and send for node2. Node1 sends label1 for node2, too. As the result, memory of node2 includes labels {2,4,1,1}.

According to degree centrality of nodes, next node is node4. Node2 includes labels {2,4,1,1} in its memory where the distance of the shortest path from node2 with labels1 and1 are 2 and2 respectively. Therefore, node2 sends label1 for node4. Node5 due to equality of distance of the shortest path with other labels, randomly selects label3 from labels 1,2,3 and sends it for node4. Memory labels of node4 are figured in Fig.4.

As figure4 shows, node1 receives label2 from neigbor3 and label3 from neighbor5, respectively.

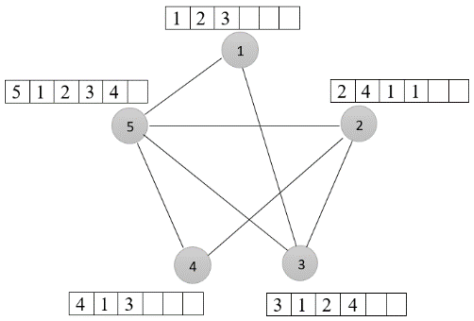

Fig4. *Step of Completing Memory of Nodes*

First iteration will end after all nodes in network complete their memory. This step iterates t times and each time more labels are added into memory. Here in order to characterize the method of calculating non-existent links in network, label propagation process is stopped after this iteration.

3.2.3. Calculating Scores for Non-Existent Links in Network

Calculation the scores of non-existed links is based on maximum in-memory information. Fig5. Shows three non-existed links(1,2),(1,4) and (4,3). The number of in-memory label similarity is used in order to obtain scores for non-existed links in network. Therefore, to calculate non-existed link probability (1,2), the number of labels of node2 in memory of node1 and number of labels of node1 in memory of node2 are investigated. So, the probability of making a connection between two nodes is explained as:

$$S(x,y) = \text{MAX}\left\{ f_y\left(L_x(t)\right) , f_x\left(L_y(t)\right)\right\} \quad (1)$$

Where, S(x,y) is the probability of making a link between two nodes x,y in future. f_x is the number of labels of node x in memory of node y. $L_y(t)$ is all current labels in memory of node y in t^{th} iteration.

Therefore, the probability of making a link between node1 and 2 is the maximum label frequency. Hence, the score of link (1,2) is MAX {2,1} =2. Also, in order to obtain the score of non-existed link (1,4), the number of similar labels in the memory of nodes 1 and 4 are checked. In memory of node4 there is a labels of node1 but memory of node1 lacks of any labels of node4. So, the score of link (1,4) is MAX {1,0} =1. Similarly, the score of non-existed link (3,4) is MAX {1,1} =1.

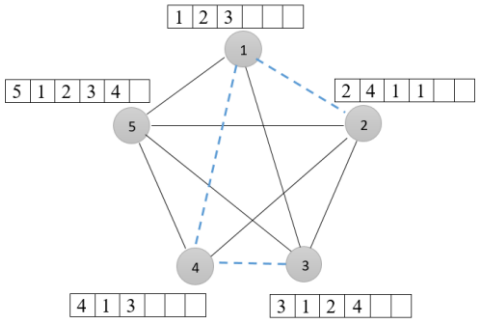

Fig5. *Calculating scores for non-existent links*

3.2.4. Link Prediction Based on Score

In this step, links are ordered decently in a list based on their score. High score for each pair of disconnected nodes refers to high probability of connection between these two nodes in future. Whereas, low score refers to low probability of connection between these two nodes. Fig.5 according to the scores of non-existent links, shows that links (1,2) have the highest scores than other non-existent links, therefore the prediction probability for the future link between nodes 1,2 is high. So, it is possible to predict future newborn or non-existent links or invisible links in current networks, through the rank of similar scores.

4. EXPERIMENTS & RESULTS

4.1. Data Set

Experimental data in this research include American college football network [14], Net Science [15], Email network [16] and Power network [17]. Table1 presents the specification of these data.

Table1. *Data Set Specification*

name	Nodes	Links	Average shortest path distance	Graph diameter	Average clustering coefficient	Average degree
Footbal	115	613	2.508	4	0.403	10.661
NetScience	1589	2742	5.823	17	0.878	3.451
Email	1133	5451	3.606	8	0.22	9.622
Power network	4941	6594	18.989	46	0.107	2.669

4.2. Dataset Preparation and Accuracy Metric

In order to test the accuracy of the algorithm, current network links are divided randomly into two parts: training set and probe set. Training set acts as known information and probe set acts as unknown set and any information in probe set are not permitted to be used in prediction operation [4]. In link prediction method, a similarity score is considered for each non-existent or invisible link in training set which determines the possibility of link creation. Therefore, non-existent links with higher similarity score are predicted as future links. AUC metric is used for accuracy measurement. In a custom, undirected and un-weighted network G(*V, E*) where V is a set of nodes and E is a set of links, possibility AUC is interpreted as: the score of a missed and randomly selected link (a link in probe set) is higher than a non-existent and randomly selected link (member of U-E). If *n* independent comparisons are performed, the number of missing links having a higher score than the nonexistent links is n_{-1}, the number of missing links having the same score as the nonexistent links is n_{-2} , and the number of missing links having a smaller score than the nonexistent links is n_{-3}. The *AUC* value is defined [13]

$$AUC = \frac{1\times n_{-1}+0.5\times n_{-2}+0\times n_{-3}}{\text{n}} \quad (2)$$

Obviously, if all links are segmented and all scores are selected randomly then $n_{-1} = n_{-2} = n_{-3} = {}^{1}/_{3}\, n$ and AUC is 0.5. So, degree higher than 0.5 refers to better performance of algorithm than pure chances.

4.3. Experiments results

To do a sound and ensured investigation of experimental results, the experiment is ten then times and average results are obtained between cost and performance. We considered t iterations to 100 iterations for each network. In figures 7,8,9 and 10, horizontal axis shows the percent of observed links in network. If all current links is data set are considered 100%, 90% (or 0.9) of links means that 90% of all links are placed randomly in training set and are known as observed links in network and 10% remained links are placed in probe set and are known as missed links in network. Hence, initially and in 0.1 of observed links, network includes many of missed links.

The input of implementation method based on proposed idea are observed links or the same current links in training set. Then, according to the proposed method, a score is considered for each non-existent link in this set. Following, current links in probe set are used for accurate evaluation of proposed method with AUC metric.

Figure.6 shows the improvement scale of implemented method based on proposed idea through basic method on football network data set that is evaluated with AUC metric. As it is illustrated in this figure, proposed method about the links that are more missed, has a close accuracy to basic algorithm but by gradually raising the curve trend, it is clear that prediction accuracy of proposed method is better than basic method.

Figure.7 shows the improvement rate of implemented method based on proposed method through basic method on NetScience data set that is evaluated with AUC metric. As it is figure shows, proposed method about the links that are more missed, has a close or lower accuracy to/than basic algorithm but by gradually raising the curve trend and lowering the missed links, it is obvious that prediction accuracy of proposed method is better than basic method in 80% of observed links.

Figure.8 shows the improvement rate of implemented method based on proposed method through basic method on Email data set. As this figure shows, the accuracy of proposed method is in most of cases in curve trend, better than basic algorithm.

Figure.9 presents the comparison between the performances of proposed method based on proposed idea through basic method on Power network set. According to this figure, curve is mostly flat. Because

power network has the maximum mean of the shortest length of path and minimum network clustering coefficient.

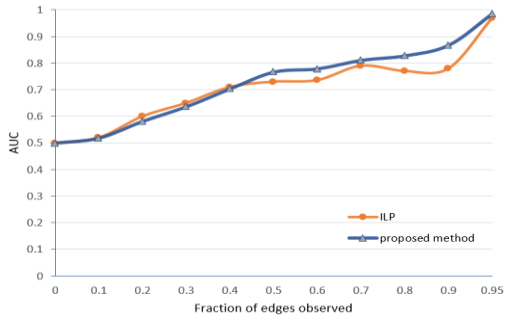

Fig6. *Comparison of the performance of proposed method and basic algorithm(ILP) in fottbal data set*

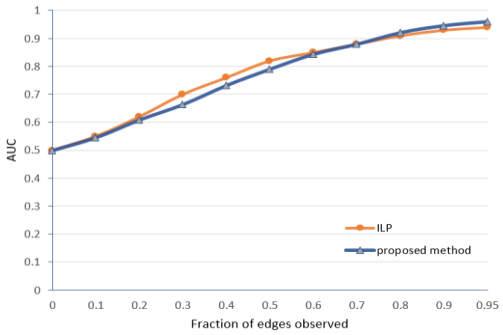

Fig7.*comparison of the performance of proposed method and basic algorithm(ILP) in Net data set*

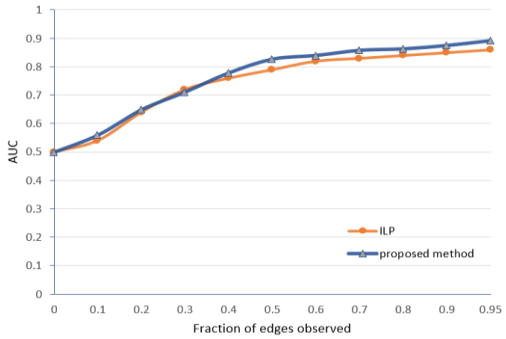

Fig8. *Comparison of the performance of propossed method and basic algorithm(ILP) in Email data set*

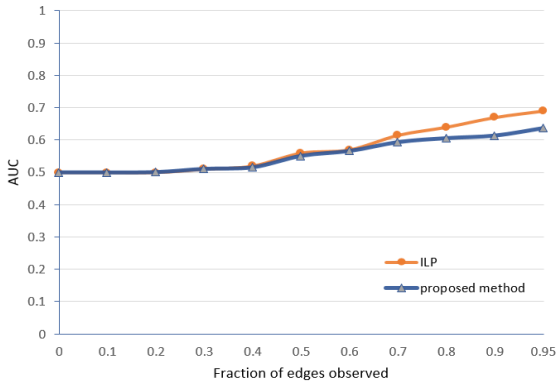

Fig9. *Comparison of the performance of propossed method and basic algorithm(ILP) in Power network data set*

5. CONCLUSION

This paper presents a method for improving link prediction accuracy in social networks based on label propagation where uses the distance of shortest path for label propagation among memories. In compare with basic algorithm, this paper invented the using of the shortest path for selecting in-memory labels that results to improvement of the prediction accuracy. In this method, the improvement rate of proposed

method by using AUC evaluation metric in football network was 26.4%, in NetScience network was 4.4% and in Email network was 20.6% that totally the improvement mean of proposed method is 17.133%. It is recommended that in future researches, this method to be evaluated on directed and weighted networks. Also, as the process of running algorithm needs high computing time, it is recommended that the parallelism techniques to be evaluated.

REFERENCES

[1] Ferreira and T. Du Plessis, "Effect of online social networking on employee productivity," South African Journal of Information Management, vol. 11, no. 1, pp. 1-11, 2009.

[2] L. Garton, C. Haythornthwaite, and B. Wellman, "Studying online social networks," Journal of computer-mediated communication, vol. 3, no. 1, p. JCMC313, 1997.

[3] Mohan, R. Venkatesan, and K. Pramod, "A scalable method for link prediction in large real world networks," Journal of Parallel and Distributed Computing, vol. 109, pp. 89-101, 2017

[4] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," Physica A: statistical mechanics and its applications, vol. 390, no. 6, pp. 1150-1170, 2011.

[5] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," Physical review E, vol. 76, no. 3, p. 036106, 2007.

[6] F. Li, J. He, G. Huang, Y. Zhang, Y. Shi, and R. Zhou, "Node-coupling clustering approaches for link prediction," Knowledge-Based Systems, vol. 89, pp. 669-680, 2015.

[7] W. Cui, C. Pu, Z. Xu, S. Cai, J. Yang, and A. Michaelson, "Bounded link prediction in very large networks," Physica A: Statistical Mechanics and its Applications, vol. 457, pp. 202-214, 2016.

[8] Q. Sun, R. Hu, Z. Yang, Y. Yao, and F. Yang, "An improved link prediction algorithm based on degrees and similarities of nodes," in 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), 2017: IEEE, pp. 13-18.

[9] L. Li and W. Liu, "Link Prediction via Factorization Machines," in Australasian Joint Conference on Artificial Intelligence, 2018: Springer, pp. 681-691.

[10] J. Ding, L. Jiao, J. Wu, and F. Liu, "Prediction of missing links based on community relevance and ruler inference," Knowledge-Based Systems, vol. 98, pp. 200-215, 2016.

[11] S. YounessZadeh and M. R. Meybodi, "A Link Prediction Method Based on Learning Automata in Social Networks," Journal of Computer & Robotics, vol. 11, no. 1, pp. 43-55, 2018.

[12] N. M. Ahmed, L. Chen, Y. Wang, B. Li, Y. Li, and W. Liu, "Sampling-based algorithm for link prediction in temporal networks," Information Sciences, vol. 374, pp. 1-14, 2016.

[13] J. Liu, B. Xu, X. Xu, and T. Xin, "A link prediction algorithm based on label propagation," Journal of computational science, vol. 16, pp. 43-50, 2016.

[14] M. Girvan and M. E. Newman, "Community structure in social and biological networks," Proceedings of the national academy of sciences, vol. 99, no. 12, pp. 7821-7826, 2002.

[15] M. E. Newman, "Finding community structure in networks using the eigenvectors of matrices," Physical review E, vol. 74, no. 3, p. 036104, 2006.

[16] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," Physical review E, vol. 68, no. 6, p. 065103, 2003.

[17] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," nature, vol. 393, no. 6684, p. 440, 1998.

Citation: *Shadi sharifpour & Mehdi bateni, (2019) Improved Link Prediction in Social Networks using Label Propagation. International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), 6(3), pp.31-38. http://dx.doi.org/10.20431/2349-4859.0603004*