Memory-based label propagation algorithm for community detection in social networks

Razieh Hosseini
Operating System Security Lab (OSSL)
Computer Engineering Department
Alzahra University, Tehran, Iran
r.hosseini@student.alzahra.ac.ir

Reza Azmi
Operating System Security Lab (OSSL)
Computer Engineering Department
Alzahra University, Tehran, Iran
azmi@alzahra.ac.ir

***Abstract*— Community detection in social network is a significant issue in the study of the structure of a network and understanding its characteristics. A community is a significant structure formed by nodes with more connections between them. In recent years, several algorithms have been presented for community detection in social networks among them label propagation algorithm is one of the fastest algorithms, but due to the randomness of the algorithm its performance is not suitable. In this paper, we propose an improved label propagation algorithm called memory-based label propagation algorithm (MLPA) for finding community structure in social networks. In the proposed algorithm, a simple memory element is designed for each node of graph and this element store the most frequent common adoption of labels iteratively. Our experiments on the standard social network datasets show a relative improvement in comparison with other community detection algorithms.**

***IndexTerms*—label propagation algorithm, community detection, social networks, complex networks.**

I. Introduction

Most of real networks such as technological and social networks can be presented as network systems which have a several common characteristics such as small world, scale-free and modular structure (also called community structure) [1]. In recent years, some attempts tried to show that community structures are one of the significant characteristic in the most complex networks such as social networks due to numerous trends of human being to forming groups or communities. The community structure is a set of nodes with more internal links than external. Finding community structure of social networks reveals useful information about structural and functional of such networks to analysis them with many applications [2–5]. Community detection plays an important role in social network analysis, because it can help researchers to understand the organization and function of the networks, the dynamics and evolution of the network, and so on. Recently community detection techniques widely used in various domains such as epidemiology networks, biological networks, metabolic networks, ecological webs and especially online social networks [6].

A social network can be modeled as a graph G=(V, E), where V is a set of nodes and E is a set of edges that represent the interaction between the nodes. A community in a network is defined as a group of nodes that have more edges among themselves than those vertices outside the group. Due to the significant applications of community detection, several community detection approaches have been presented in literature which can be classified into six categories: spectral and clustering methods, hierarchical algorithms, modularity-based methods, model-based methods, local community detection methods, and feature-based assisted methods [2]. In [7], a novel algorithm using artificial immune system and clonal selection mechanism was presented for clustering. In this algorithm they proposed an adaptive spectral clustering algorithm which could determine the number of cluster automatically and obtain the corresponding cluster centers. In [8], the author developed a new spectral clustering algorithm to approach the community detection problem in networks which consists of analyzing the neighborhood structure of the clusters. This heuristic was investigated over a maximization problem based on an adapted version of the clustering coefficient measure. In [9], researchers proposed a new method that utilizes social interaction data (e.g., users' posts on *Facebook*). Since many relationships are missing or not recorded, so they attempted to identify hidden communities. Their method also finds the multiple social groups of social networks and does not depend on structural information. There are several approaches in order to optimizing a quality function for a good graph partitioning such as maximizing modularity which called modularity based methods. In [10], researchers proposed a polynomial-time approximation algorithm for the modularity maximization problem in the context of scale-free networks. A comprehensive survey on modularity-based and non-modularity based approaches for community detection can be found in [11].

One of the important centrality measures that can be used for finding community is edge betweenness [12] which is define as the number of shortest paths between all nodes to all others that pass through that edge. In [13], an improved version of betweenness algorithm has been presented to detect communities in networks and it applied on both weighed and unweighted networks. According to this algorithm, the edge betweenness, dissimilarity index and edge-clustering coefficient converted into the weight of edges so it will lead to the conclusion that the inter-community edges have greater

978-1-4799-8818-1/15/$31.00 ©2015 IEEE

discriminating values (or weighted betweenness) than intra-community edges, and hence the inter community edges can be more easily identified than before and strength of relations will be more obvious. A novel measure proposed by *yang et al.*[14] that integrates both the concept of closed walks and clustering coefficients to replace the edge betweenness in the well-known divisive hierarchical clustering algorithm [15]. The edges with the lowest value are removed iteratively until the network is degenerated into isolated nodes. Some of the works are focus on finding the optimal community detection such as density drop of sub graphs method [16]. In this paper a new algorithm for community selection has been proposed. The intuition of their algorithm is based on drops of densities between each pair of parent and child nodes on the dendogram, the higher the drop in density, the higher probability the child should form an independent community. Based on the Max-Flow Min-Cut concept, they proposed a novel algorithm which can output an optimal set of local communities automatically and it is also suitable for the case that the dendogram is a tree. However, most of the existing community detection algorithms focus on binary networks; most of the networks are naturally weighted such as delay tolerant networks (DTN) or online social networks (OSN). In [17], the authors illustrated the problems of community detection in weighted networks and used weighted community for data forwarding in DTN and worm containment in OSN, and they also introduced two metrics: intra-centrality and inter-centrality, to characterize nodes in communities.

In this paper, we proposed memory-based label propagation algorithm (MLPA) for finding community structure in social networks. In the proposed algorithm each node of network equipped with a simple memory element and then iteratively each node adopted several labels and simultaneously most adoption of labels is stored for each node by the memory element. At the end of the algorithm, most frequent adoption of labels for each node is extracted to form proper communities. The proposed algorithm does not aware about the number of community and easily executed in a near linear time complexity. Experimental results on the proposed algorithm in comparison with standard label propagation algorithm and several algorithms reveal the superiority of the proposed algorithm.

The rest of this paper is organized as follows: section 2 describes the standard label propagation algorithm and memory-based label propagation algorithm. Experimental simulations presented in section 3, and finally section 4 concludes this paper.

II. Proposed algorithm

Since our proposed algorithm improves label propagation algorithm, in this section at first we describe the standard label propagation algorithm (LPA) and then we describe our proposed algorithm called memory-based label propagation algorithm (MLPA).

A. Standard Label propagation algorithm (LPA)

Label propagation algorithm was proposed by *Raghavan* [18] in 2007. He proposed a localized community detection algorithm based on label propagation which uses the network structure alone as its guide. Each node is initialized with a unique label, and in every iteration of the algorithm each node adopts a label that a maximum number of its neighbors have. As the labels propagate through the network in this manner, densely connected groups of nodes form a consensus on their labels. At the end of the algorithm, nodes having the same labels are grouped together as communities. It is noted that in this method the number of communities and their sizes are not known a priori because these parameters are determined at the end of the algorithm. Standard Label propagation algorithm is described as the following steps:

(1) Initialize the labels at all nodes in the network. For a given node *i*, its label is *iL*. Label *iL* should be a unique label. The number of nodes does not need be known.

(2) Arrange the nodes in the network in a random sequence *S*. For a given node *i*, its order is *iS*.

(3) Each node changes its label to maximum number of the same label among its neighbors in the order of sequence *S*.

(4) Iterate steps (2) and (3) above till no labels can be changed.

The label propagation algorithm is presented to be useful with static networks

B. Memory-based label propagation algorithm (MLPA)

In this section, we describe the proposed algorithm called memory-based label propagation algorithm (MLPA) for finding community structure in social networks. At the beginning of the proposed algorithm, each node of the given network equipped with a simple memory element. The following step repeats several times with several label adoptions. The memory element could store the label of each node on each iteration of the proposed algorithm. In fact, the memory element of nodes reflects the partition of network for each iteration. At the end of algorithm most frequent common adoption of labels for each node is extracted. This process causes to the final label of each node replace with maximum number of the same neighboring labels. Finally the community structure of network is returned.

It is noted that if the algorithm is repeated for *k* times it need a memory element with capacity of *k* different labels. Moreover, the proposed algorithm does not aware about the number of community and easily execute in a near linear time complexity.

Pseudo-code of the memory-based label propagation algorithm is presented as figure 1.

```
Memory-based label propagation algorithm
Begin Algorithm
  Input: network $G=(V, E)$
  Output: set of communities
    Assign a memory element for each node of network
   While (Iteration number of algorithm is lower than $T$) Do
    Initialize the labels of each node with a unique label
    Repeat
         Change the labels of each node with maximum
          number of the same neighboring labels
    Until (no labels can be changed)
    Store label of each node in memory element
   End While
  For each node extract the most frequent common labels
   Change the labels of each node with maximum number
        of the same neighboring labels
End Algorithm
```

Fig. 1. Pseudo-code of the memory-based label propagation algorithm (MLPA)

III. Simulation results

In order to study the performance of proposed algorithm, at first, in section *A*, we introduce the well-known social network datasets and then in section *B*, we describe the performance metric for evaluation of community detection algorithms. The results of Experiments are given in section *C*.

A. Social network datasets

In this section we describe for well-known datasets in details and description of all datasets is given in table 1.

Zachary's karate club is a social network of interactions between people in a karate club which has 34 members of a karate club over a period of 2 years. A split appeared in the club, and the club's instructor (node 33) took away a half of the members in the club and built a new one, since a disagreement arose between the club's instructor and the administrator of the club. *Zachary* [19] studied the members in a karate club for many years and constructed a relationship network for the 34 members in the original club, which contains 78 links.

The second well-known network is Dolphin which indicates the social network of frequent associations between the dolphins. There are 62 dolphins and edges were set between animals which were seen together more often than expected by chance. *David Lusseau*, a biologist investigated 62 dolphins for many years [20]. During the study, he observed that the 62 dolphins can be roughly classified into two groups, since dolphins in a same group are closely related to each other in activities of daily living. *Lusseau* found that dolphins of same groups actually share a similar age. *Lusseau* constructed a relationship network for the 62 dolphins that contains 159 links [20]. Due to the natural classification, *Lusseau*'s dolphin network is often used to test algorithms for community detection.

The third dataset is network of US College Football Teams [2] with 115 vertices representing the teams and 616 links. In this real network two vertices being connected when their teams play against each other. The teams are divided roughly into 12 conferences. Games between teams in the same conference are more frequent than games between teams of different conferences, so one has a natural partition where the communities correspond to the conferences

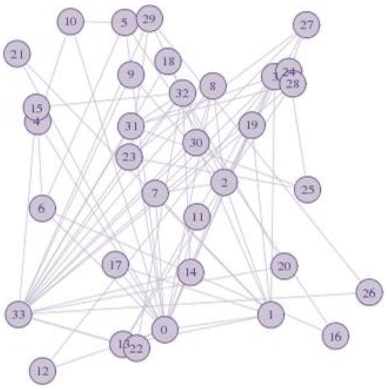
The last network is Netscience which is a coauthor-ship network of scientists working on network theory and experiment, as compiled by *M. Newman* in May 2006. The network was compiled from the bibliographies of two review articles on networks. The network contains all components for a total number of 1589 scientists and not just the largest component of 379 scientists [21].

Description of the mentioned social network dataset is summarized in Table 1

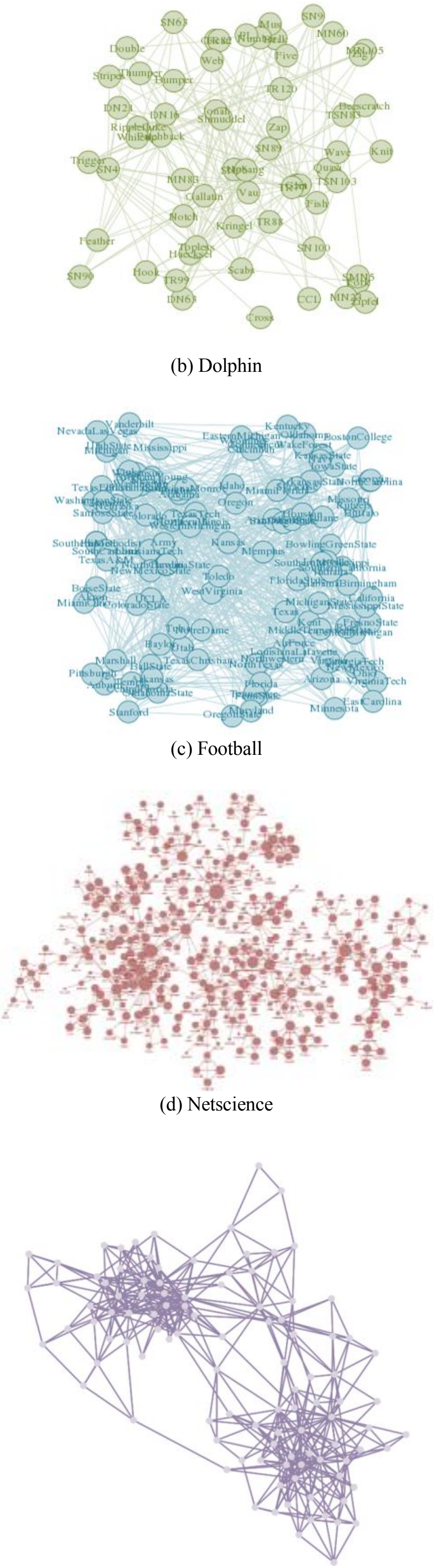
TABLE I. Description of social network datasets for experiments

Dataset	Number of nodes	Number of edges
Karate	34	78
Football	115	613
Dolphin	62	159
Netscience	1589	2742
PGP	10680	24316
Political Books	105	441
Email	1133	5451
Blogs	3982	6803

Visualization of some small social network datasets is presented in Figure 2.



(a) Karate



(b) Dolphin

(c) Football

(d) Netscience

(e) Political Book

Fig. 2. Visualization of small social network datasets

B. Evaluation metric: Modularity

In this study, we use a famous measure is presented by *Girwan-Newman*[21] called *Modularity* in order to evaluate the quality of the communities found through our proposed algorithm. *Modularity Q* is defined as follows:

$$Q=\frac{1}{2m}\sum_{i,j}\left(A_{ij}-\frac{k_i k_j}{2m}\right)\delta\left(C_i,C_j\right) \qquad (1)$$

whereA_{ij} indicates the adjacent matrix of the input network, m indicates the total number of edges of the input network, $k_i$denotes the degree of node i and δ delta function yields one if node i and node j are in the same community and zero otherwise.

C. Experimental results

In this section, we report the results of experiments on the social network datasets described in table 1. The experiments are conducted on a system with windows 7 platform having configuration Intel® Core i53337U CPU 1.80 GHz and 4 GB RAM.

Our results on the social network datasets are presented in Table 2. We set limited capacity for each memory element is 10. The results of modularity measure for proposed algorithm as MLPA in comparison with other algorithm such as LPA [9], LPA-CNPE [16], LPA-CNP1 [16] and KLPA [15] are listed in table 4.

TABLE II. RESULTS OF MODULARITY OF COMMUNITY DETECTION ALGORITHMS ON SOCIAL NETWORK DATASETS

Dataset	LPA	LPA-CNPE	LPA-CNP1	KBLPA	MLPA
Karate	0.296	0.302	0.284	0.073	**0.303**
Football	0.582	**0.600**	**0.600**	0.573	0.584
Dolphins	0.465	0.463	0.457	**0.489**	0.485
Netscience	0.871	-	-	0.883	**0.884**
PGP	0.806	-	-	0.775	**0.808**
Political Books	0.489	0.451	0.451	0.449	0.455
Email	**0.380**	-	-	0.183	0.379
Blogs	0.791	0.423	0.423	0.808	**0.805**

Also, the results of this experiment are demonstrated as bar chart diagram for karate, football, dolphin, political books and Blogs.

As shown in Tables 2 and figure 3, the results show a relative improvement in comparison with other community detection algorithms.

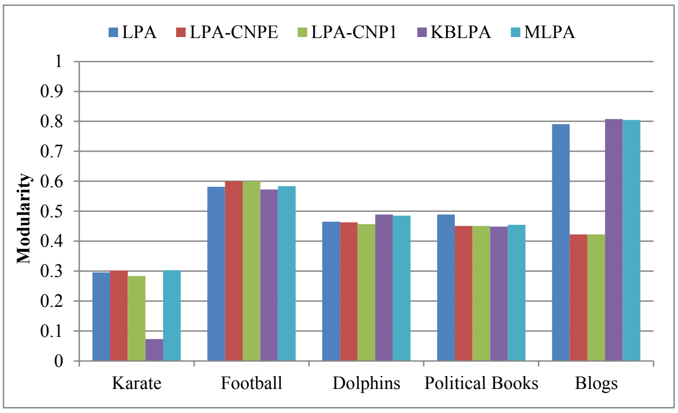


Fig. 3. Results of modularity for comparison algorithms

In the future work, the performance of the proposed algorithm can be improved with the aid of meta-heuristics and evolutionary algorithms.

IV. Conclusion

Finding community structures in many complex networks such as technological and social networks plays an important role for analysis of such networks. In this paper we improved the label propagation algorithm as the one of the fastest algorithm for community detection with the aid of a memory element for each node of network. The proposed algorithm called memory based label propagation algorithm tried to store several label of each node. Finally, the algorithm extracts the most frequent common label for each node to form communities of the networks. The experimental results showed a relative improvement in comparison with other community detection algorithms.

References

[1] D. Easley and J. Kleinberg, Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press, 2010.

[2] S. Fortunato, "Community detection in graphs," Physics Reports, vol. 486, no. 3–5, pp. 75–174, 2010.

[3] F. Amiri, N. Yazdani, H. Faili, and A. Rezvanian, "A Novel Community Detection Algorithm for Privacy Preservation in Social Networks," in Intelligent Informatics, vol. 18, A. Abraham, Ed. 2013, pp. 443–450.

[4] A. Rezvanian, and M. R. Meybodi, "Sampling social networks using shortest paths," Physica A: Statistical Mechanics and its Applications, vol. 424, pp. 254-268, 2015.

[5] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Modeling blogger influence in a community," Social Network Analysis and Mining, vol. 2, no. 2, pp. 139–162, 2012.

[6] F. D. Malliaros and M. Vazirgiannis, "Clustering and community detection in directed networks: A survey," Physics Reports, vol. 533, no. 4, pp. 95–142, 2013.

[7] H. Wang, J. Chen, and K. Guo, "An Adaptive Spectral Clustering Algorithm," Journal of Computational Information Systems, vol. 8, no. 2, pp. 895–904, 2012.

[8] M. C. V. Nascimento, "Community detection in networks via a spectral heuristic based on the clustering coefficient," Discrete Applied Mathematics, vol. 176, pp. 89–99, 2014.

[9] Y.-L. Chen, C.-H. Chuang, and Y.-T. Chiu, "Community detection based on social interactions in a social network," Journal of the Association for Information Science and Technology, vol. 65, no. 3, pp. 539–550, 2014.

[10] T. N. Dinh and M. T. Thai, "Community detection in scale-free networks: approximation algorithms for maximizing modularity," IEEE Journal on Selected Areas in Communications, vol. 31, no. 6, pp. 997–1006, 2013.

[11] S. Fortunato and C. Castellano, Community structure in graphs. Encyclopedia of Complexity and System Science.Springer, 2008.

[12] M. E. J. Newman, "A measure of betweenness centrality based on random walks," Social networks, vol. 27, no. 1, pp. 39–54, 2005.

[13] B. Chen, J. Xiang, K. Hu, and Y. Tang, "Enhancing betweenness algorithm for detecting communities in complex networks," Modern Physics Letters B, vol. 28, no. 09, 2014.

[14] Y. Yang, P. G. Sun, X. Hu, and Z. J. Li, "Closed walks for community detection," Physica A: Statistical Mechanics and its Applications, vol. 397, pp. 129–143, 2014.

[15] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," Proceedings of the National Academy of Sciences, vol. 99, no. 12, pp. 7821–7826, 2002.

[16] X. Qi, W. Tang, Y. Wu, G. Guo, E. Fuller, and C.-Q. Zhang, "Optimal local community detection in social networks based on density drop of subgraphs," Pattern Recognition Letters, vol. 36, pp. 46–53, 2014.

[17] Z. Lu, X. Sun, Y. Wen, G. Cao, and T. La Porta, "Algorithms and Applications for Community Detection in Weighted Networks."

[18] 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.

[19] W. W. Zachary, "An information flow model for conflict and fission in small groups," Journal of anthropological research, pp. 452–473, 1977.

[20] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations," Behavioral Ecology and Sociobiology, vol. 54, no. 4, pp. 396–405, 2003.

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