Identifying Communities on Static Social Networks

Maliheh Ghasemzadeh[1*], Mohsen Ashourian[2]
1- Department of Electrical Engineering, Majlesi Branch, Islamic Azad University, Isfahan, Iran.
Email: gh_mavad@yahoo.com (Corresponding author)
2- Department of Electrical Engineering, Majlesi Branch, Islamic Azad University, Isfahan, Iran.
Email: ashourian@gmail.com

Received: May 2019 Revised: July 2019 Accepted: August 2019

ABSTRACT:
Many complex natural and social structures can be considered as networks. Internet sites, social networks, organizational communications, family connections, electronic mails, phone calls, and financial transactions are just a few examples of these networks. Nowadays, network analysis is one of the most popular and widely used research branches in the world. One of the most commonly used topics in network analysis is the identification of organizations in the network. In this research, we present the detection of communities in static social networks using the genetic algorithm and its improvement with the label propagation algorithm known as Genetic Algorithm- Label Propagation. The genetic algorithm explores the search space well and converges to the best answer. This algorithm is scalable and our results show that our proposed algorithm performs faster and better than other algorithms.

KEYWORDS: Identifying Communities, Static Social Networks, Genetic Algorithm.

1. INTRODUCTION

Today, social media has many uses and is particularly important among Internet users. For this reason, social network analysis is an important and influential area of research among researchers. Social networking in the social sciences examines the relationships between human beings, human groups and organizations. A social network is a social structure consisting of nodes that are generally an individual or organization, interconnected by one or more specific types of interdependence such as ideas and financial exchanges, friendships, kinship, and so on. [1].

Discovering communities in complex networks or social networks is one of the most important problems in the field of science and social network analysis. Clustering or identifying communities will reveal the structure of groups in social networks and the hidden connections between its components. A community is a set of nodes whose density of communication is greater than any other network entity. The applications of the community recognition consist of improved search engine performance, better understanding of network structure and finding specific groups. The purpose of community recognition is to identify the infrastructures that may exist in the networks. Identifying these groups in social networks has many applications in marketing, social sciences, economics and so on [2].

Identifying communities on social networks can provide information on the structure and performance of networks, but with the increase in the number of users, doing so will cost a lot of time and memory. Therefore, providing an efficient algorithm for recognizing communities with the aim of reducing the running time of the algorithm and the amount of memory consumed can be very useful and valuable. One of the most important issues in social network analysis is the issue of group identification. One of the important features of social networks is how they are formed. This creates groups on the graph surface. The group is a set of nodes that has more inner edges than the outer edges that connect these vertices to the vertices of other groups [3]. Figure 1 shows a graph with three groups, in which the inner group edges are less colored than the inter group edges.

The vertices in this group are very similar and often play the same role in the network. This feature of groups has received much attention from researchers. That is why finding groups is one of the most important issues in network analysis. In this study, a new method called GA-LP (Genetic Algorithm_ Label Propagation) is proposed that aims to detect communities in static networks and to perform better than other algorithms in optimizing the detected communities.

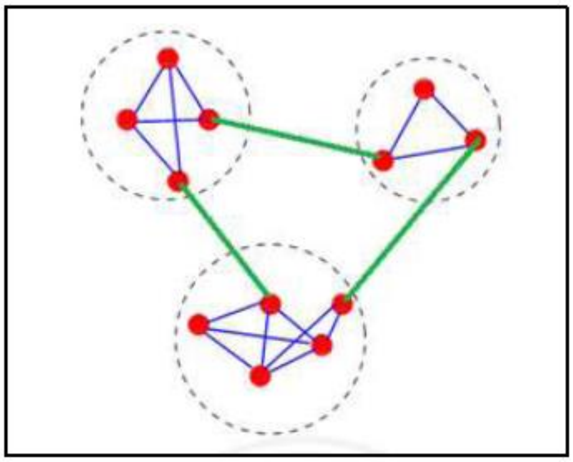

Fig. 1. A simple graph with three distinct groups.

2. BASIC PRINCIPLES AND DEFINITIONS

This section provides an overview of the basic definitions needed.

2.1. Social Network

Each social network can be modeled as a graph G (V, E) in which individuals or web pages form the graph nodes and the relationship between them are the graph edges. As stated earlier, there is no unique definition for the group. Here, a group, category, or module is referred to as a set of vertices in a graph that has more inner edges than outer edges, resulting in a density within the group and a small distance between the vertices of a graph. According to this definition, the group is not closed to the community and is open to sharing.

2.2. The Structure of Communities

The main focus of this research is on the structure of communities. In this study, a comprehensive definition of the structure of communities presented by Newman and Girvan is given. The definition is as follows [4]:

"Dividing the nodes of a network into groups that there is high network communication density within each group but between groups this density is low."

An example of a network with community structure is shown in Figure 2. Nodes in a community must have many relationships with each other compared to another outside the community. In this example, the nodes within the community are fully connected. This means that all possible links exist within the community, while there are few links between communities [5].

Fortunato and Castellano presented a review of the structure of communities in networks [6]. Communities can have different levels of organization, where communities include several sub-communities. This phenomenon is known as hierarchy. Hierarchies are used in community recognition algorithms, such as fast greedy community detection. Since, social recognition algorithms, such as label propagation and Spinglass, do not use this hierarchy, social recognition algorithms are not considered in the analysis of crawling methods. Online social networks analyzed by community recognition algorithms have shown that there is no mediation for community size, but there is a power-law distribution for community size [7]. Small communities seem to exist alongside large ones. However, in the networks used, relationships or links are formed based on a shared interest. However, a research by Leskovac et al showed that in real-world communities, communities gradually lose their shape and become more integrated into the whole network. They showed that with communities of 100 nodes size, the quality of communities becomes worse and worse [8].

Many algorithms have been proposed to identify communities so far. These algorithms can be classified into two general and local categories [9].

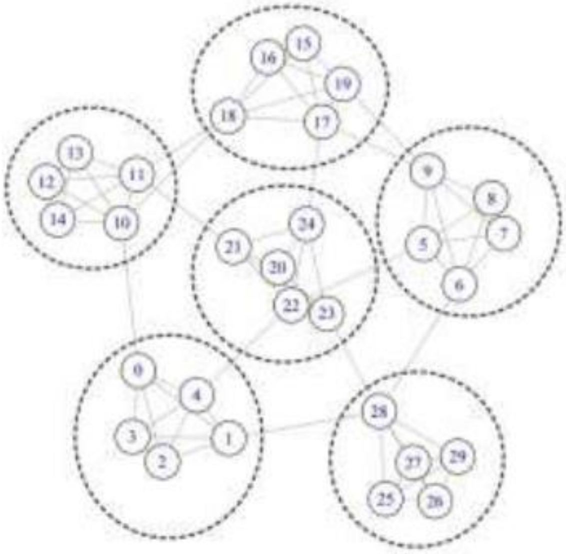

Fig. 2. A network with 6 communities marked in circles.

2.3. Collecting Social Networks Methods

In this research, two types of social networks are defined based on the method of collection [10].

- Traditional Social Networks
- Online Social Networks

Social networks are manually assembled by social researchers to examine interactions between a group of people over a period of time. These types of networks are called traditional social networks. Well-known examples of this type of social networks include the Zachary Karate Club Network and the Girvan and Newman Football Network. In the Zachary Karate Club Network, the nodes represent the membership of a karate club and the links represent the social interactions between them. On the football network, the nodes represent the football teams and the links between them represent the matches held between them.

2.4. Providing Solutions and Proposed Methods

There are several methods for identifying communities in social networks that we have proposed a new method in this study and compared it with other algorithms. Based on the results, the algorithm performs better than other algorithms. The proposed algorithm for community detection is Genetic Algorithm that we used the label propagation algorithm, which has linear temporal complexity and is one of the well-known algorithms in the field of community detection [13] to improve the exploring of the search space. Thus, after each iteration of the genetic algorithm, the label propagation algorithm modifies the set of solutions obtained.

2.5. Genetic Algorithms

Genetic algorithm is the most common method in evolutionary computation. This algorithm, known as one of the random optimization methods, was invented by John Holland in 1967. Genetic algorithm is a useful tool for search and optimization problems. The space of all possible solutions is called the search space. Every dot in the search space represents a possible solution. Therefore, each possible solution is determined by the amount of fitting defined in the problem. The genetic algorithm seeks to find the best solution out of all possible search space solutions.

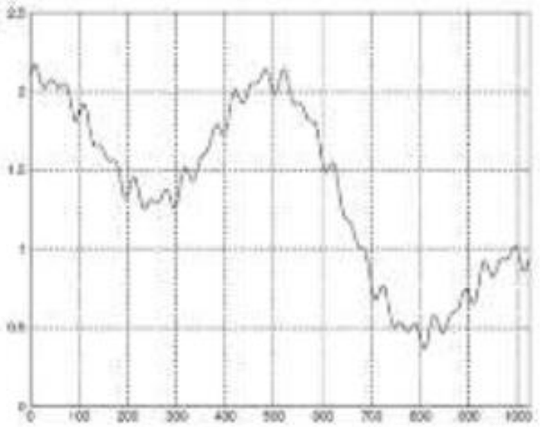

Fig. 3. An example of a search space.

Different search and optimization methods include:
1. Gradient-Based Local Optimization Method
2. Random search
3. Random hill climbing
4. Symbolic Artificial Intelligence.

2.6. label propagation algorithm

The label propagation algorithm was first used by Raghavan in the social recognition problem. The purpose of this algorithm is to divide the network without the knowledge of the size and number of communities. The steps of the standard algorithm are as follows: At first, all nodes are given a unique initial label. Then a random checklist is generated for all nodes. Each node label is updated according to the neighbor nodes label. The label that has the highest number of repetitions in the neighborhood is given to it and if there are multiple labels with the same number of repetitions, the label is randomly selected. This label updating operation will continue until the label of each node is equal to that of most of its neighbors. Finally, nodes with equal labels are placed in a community. The time complexity of the label propagation algorithm is close to linear and hence is a good candidate for community recognition in social networks.

3. DATASETS

The data used are the data that are used and standardized in almost all community recognition methods, such as the Karate Zacharyi Club, the American college Football League, the Dolphin Network. The data sets used in this study are as follows.

Zachary Karate dataset

In the early 2020s, Wayne Zachary studied at a karate club for two years at an American university and recorded their social interactions. Based on their social interactions, he created a network dataset with 22 nodes and 21 edges. In this dataset, students are marked as vertices and two students are connected if they are good friends. Coincidentally, there was a dispute between the director of the club and the karate teacher during their study. As a result, the club divided into two smaller communities, respectively with 2 heads; the director of the club and the karate teacher. The division of the club into two communities is shown in Figure 4.

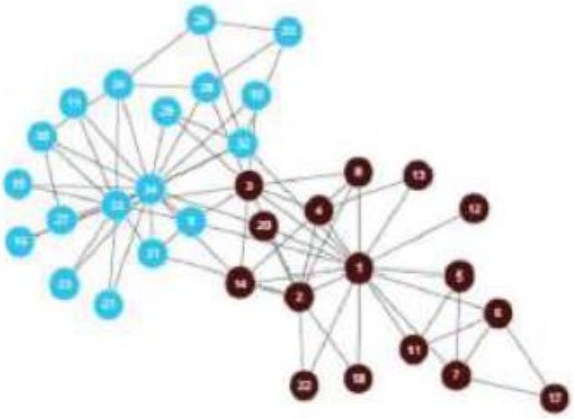

Fig. 4. Dividing the Zachary Karate Club Network into two communities.

American College Football

The American College Football network dataset has derived from American Football School game data. The IA School Inter-School Match Table is shown by the network in the autumn of 2000. In this network, the games are shown by the vertices and the games between the two teams in this season are the edges. The number of vertices in this dataset is 222 and the number of edges is 121. Teams are divided into categories. Each team in each group has an average of two games with its peers and two with other teams.

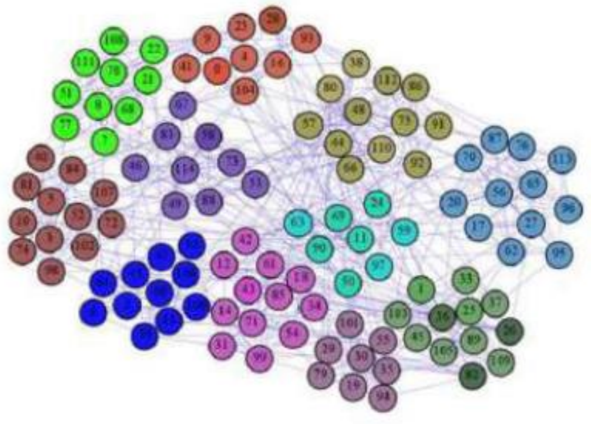

Fig. 5. The Real Structure of Communities shows this dataset before running the algorithm.

Dolphin Bottlenose Network

For seven years, biologist David Lusseau analyzed the behavior of bottlenose dolphins living in the Sound of Doubtful Sound (New Zealand) and created this network dataset. Based on frequent communication, a link will be formed if there is a connection between the two dolphins. The total number of dolphins used in this study were 14 and there were 225 edges between these dolphins, among those, which were most likely to be seen together. Figure 6 shows the main structure of the dolphin network community.

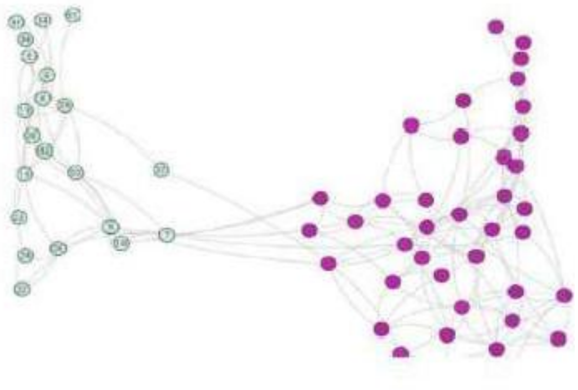

Fig. 6. The social structure of the bottlenose dolphin community

3.1. Evaluation Criteria

The evaluation criterion used in these experiments is modularity. The quality of the communities obtained by the algorithm is obtained using the modular or modularity criterion provided by Girvan-Neumann:

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

Where, "A" is the adjacent matrix of the graph, "m" is the total number of edges of the graph and "ki" represents the "i" vertex degree. The δ function has one value for two vertices inside a community, otherwise zero.

If the number of extra-cluster edges is as large as the random graphs, then Q will be zero. Q values close to 1 indicate a strong community structure. In practice, this value ranges from 0.3 to 0.7 for strong community structures [17]. With this value, it is difficult to compare graphs that are similar in structure but differ in size. Because the larger graph will naturally have higher modularity [18].

3.2. Test Results

To test the proposed algorithm, a comparison of the performance of this algorithm with that of other comparable algorithms was performed in the MATLAB programming environment for small and medium sized networks. On the three stated datasets, the algorithm is run 30 times independently and in each run, modularity is calculated.The population count is 100, the number of generations per algorithm is 100, the crossover rate is 0.8, the mutation rate is 0.1, and the elite count is set at 10%. The results of simulation of the proposed algorithm are compared with other algorithms based on modularity mean.

Proposed algorithm called GA-LP has been evaluated with popular algorithms including: LPA label propagation algorithm [19], the algorithm of Jing et al., KBLPA [20], the algorithm of Lu et al., LPACNP1 and LPAE [21] and the ILPA algorithm. The results of this evaluation can be seen in Tables 1 and figure 7.

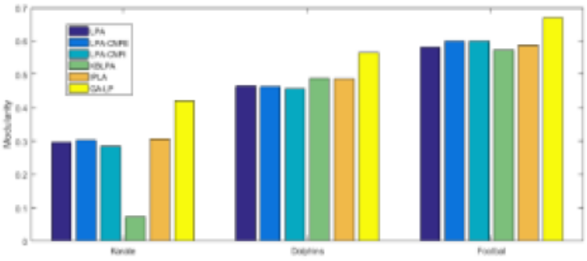

Fig. 7. Comparison results of algorithms.

Table 1. Results of the proposed algorithm with the rest of the algorithms.

GA-LP	IPLA	KBLPA	LPA CNPI	LPA CNPE	LPA	
0.420	0.306	0.073	0.284	0.302	0.296	Karate
0.5650	0.487	0.489	0.457	0.463	0.465	Dolphins
0.6700	0.588	0.573	0.600	0.600	0.582	Football

4. CONCLUSION

In this research, the basic concepts and definitions of networks are discussed first, and then the issue of recognizing communities, which is one of the practical areas in network analysis, is introduced. The most important algorithms were then compared with different evaluation criteria after performing experiments on valid data sets. In the next step, the label propagation method, which is currently one of the best algorithms available, was further explored and a proposed method was introduced to improve its efficiency and this was determined by presenting the results of the experiments and their analysis.

REFERENCES

[1] Samimian, L. And M. Sadeghzadeh, **“Identification of Societies in Social Networks”**, *Second National Conference* on Computer Engineering and Information Technology. Young Researchers Club and Elite Shushtar Branch, 2014.

[2] Hosseinzadeh, R., H. Alizadeh, et al. Nazemi, “**Identifying Communities with a Mixed Approach in Social Networks”**, *11th National Conference on Intelligent Systems*. Iranian Intelligent Systems Association, 2012.

[3] Barber, M.J., “**Modularity and Community Detection in Bipartite Networks”**. *Physical Review E*, Vol. 76(6), pp. 06610, 20072.

[4] Girvan, M. and M.E. Newman, “**Community structure in Social and Biological Networks”**. *Proceedings of the national academy of sciences*, Vol. 99(12), pp. 7821-7826, 2002.

[5] Zhao, Z., et al., “**Topic Oriented Community Detection through Social Objects and Link Analysis In Social Networks”**. *Knowledge-Based Systems*, Vol. 26, pp. 164-173, 2012.

[6] Fortunato, S. and C. Castellano, “**Community structure in graphs, in Computational Complexity”**. *Springer*. pp. 490-512, 2012.

[7] Clauset, A., M.E. Newman, and C. Moore, “**Finding community structure in very large networks”**. *Physical review E*, Vol. 70(6), pp. 066111, 2004.

[8] Leskovec, J., et al., “**Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters”**. *Internet Mathematics*, Vol. 6(1), pp. 29-123, 2009.

[9] Plantié, M. and M. Crampes, “**Survey on social community detection”**, *in Social media retrieval. Springer*. pp. 65-85, 2013.

[10] Steinfield, C., et al. “**Bowling online: social networking and social capital within the organization”**. *In Proceedings of the fourth international conference on Communities and technologies*. 2009. ACM.

[11] Zachary, W.W., “**An information flow model for conflict and fission in small groups”**. *Journal of anthropological research*, Vol. 33(4), pp. 452-473, 1977.

[12] Girvan, M. and M.E. Newman, “**Community structure in social and biological networks**.” *Proceedings of the national academy of sciences*, Vol. 99(12), pp. 7821-7826, 2002.

[13] Zhu, X. and Z. Ghahramani, “**Learning from labeled and unlabeled data with label propagation”**. 2002.

[14] Sivanandam, V.S. & Deepa. N. (2007). “**Introduction to Genetic Algorithms**” *Springer Berlin Heidelberg New York.* ISBN 978-3-540-73189-4.

[15] Raghavan, U.N., R. Albert, and S. Kumara, “**Near linear time algorithm to detect community structures in large-scale *networks”***. *Physical review E*, Vol. 76(3), pp. 036106, 2007.

[16] Newman, M.E. and M. Girvan, “**Finding and evaluating community structure in networks”**. *Physical review E*, Vol. 69(2), pp. 026113, 2002.

[17] Good, B.H., Y.-A. de Montjoye, and A. Clauset, “**Performance of modularity maximization in practical contexts”**. *Physical Review E*, Vol. 81(4), pp. 046106, 2010.

[18] U. N. Raghavan, R. Albert, and S. Kumara, “**Near linear time algorithm to detect community structures in largescale networks**,*” Physical Review E*, Vol. 76, No. 3, pp. 036106, 2007.

[19] Y. Xing, F. Meng, Y. Zhou, M. Zhu, M. Shi, and G. Sun, “**A Node Influence Based Label Propagation Algorithm for Community Detection in Networks**,” *The Scientific World Journal*, 2014.

[20] H. Lou, S. Li, and Y. Zhao, “**Detecting community structure using label propagation with weighted coherent neighborhood propinquity**,” *Physica A: Statistical Mechanics and its Applications*, Vol. 392, No. 14, pp. 3095–3105, 2013.