cmpe556- Spring 2021

Final Report
Comparison of Different Community Detection Algorithms on Real-World Networks

Barış Büyüktaş

2021-07-05

1 Introduction

In that project, I analyze 4 different algorithms for community detection such as Girvan-Newman, Clauset-Newman-Moore greedy modularity maximization, label propagation, and k-means. These methods are evaluated in terms of speed, and accuracy. The novelty of this work is that I adapted k-means clustering method for community detection. Although there is a k-means approach [1] for community detection, I used the method in a different way. The purpose of the k-means is partition the points into clustering so that the Euclidean distance between data points and nearest cluster center is minimized. Since there is no Euclidean distance in real-world networks, we used some other property of graphs to specify the feature vector of each data point, which is explained in Section 2.

Community detection is a hot topic, which helps to find densely connected components in networks. There are many approaches are proposed to solve this problem. One of the most used methods that I implemented for my project is the Girvan-Newman algorithm [2]. To detect the communities, this method removes the edges from the network one by one, and the nodes, which remain connected are considered as a community. Another method, which I implemented and compared with other methods is Clauset-Newman-Moore greedy modularity maximization method [3]. It is a greedy method that maximizes the modularity to find the community of the node. Also, I implemented label propagation algorithm [4]. It is a fast and accurate algorithm that does not require prior information about the communities. The first method that comes to my mind in clustering problems is k-means [5]. K-means uses Euclidean distance between the position of the point and the center of a cluster, however, there is no distance in simple graphs. Therefore, I used some other property of the graphs to adapt the k-means on community detection.

One of the purpose of the project is that although there are many studies in the literature, I did not see any work analyzing the differences and advantages/disadvantages of these studies between each other. Moreover, they are categorized based on the agglomerative and divisive methods. However, there are some well-known machine learning methods like k-means that can be applied to detect the community of each node.

K-means is one of the most used algorithms for clustering due to its speed and simplicity. In this project, in addition to the classical methods, I adapted the k-means algorithm to community detection.

2 Our method

In this section, I compared 4 different community detection methods, which are Girwan-Newman, label propagation, Clauset-Newman-Moore greedy modularity maximization and k-means. These methods are compared based on speed and accuracy. Also, the adaptation of k-means on community detection is examined.

2.1 Datasets

I used 2 different datasets for the project. These are Zachary's karate club [6], and Physicians [7] networks. Zachary observed 34 members of a karate club and constructed a network of friendships between members of the club. Each node represents the person in the graph, and each edge represents the relationship between two people.

There are 246 physicians in the directed Physicians network. A node represents a physician and an edge between two physicians shows that they are friends or one physician asks for advice or interested in a discussion with other physician.

2.2 Utilizing k-means to community detection

K-means algorithm aims to divide the dataset into k different clusters. Normally, the data points are n-dimensional vectors and clustering operation is completed using Euclidean distance. However, in real-world networks, nodes do not have feature vectors. To specify and assign the nodes to a feature vector, we used the distance from one node to another. The distance between 2 vertices is minimum number of edges between 2 vertex. Let v_i be the vertex of the graph. $d(v_i, v_j)$ represents the distance between v_i and v_j. The attributes of the vertex a_i is as follows:

$$a_{i1} = d(v_i, v_1) \tag{1a}$$
$$a_{i2} = d(v_i, v_2) \tag{1b}$$
$$a_{ik} = d(v_i, v_k) \tag{1c}$$

where a_i represents the node, a_{ik} represents the kth attribute of vertex a_i. After we assigned a feature vector to nodes, k-means operation is applied.

3 Performance Evaluation

In this section, we evaluated the performance of 4 different community detection methods on Zachary's karate club and Physicians networks in terms of speed and accuracy.

The accuracy of the community detection means that how well the each node is classified to the correct communities. For the Zachary's karate club, we know the ground-truth representation of the network. After our methods predict the communities of the nodes, the ground-truth and predicted communities are compared. Since we know which nodes belong to which community on Zachary's karate club, we can use this information to calculate accuracy. Since there are more nodes in Physicians network than Zachary's karate club, it is used for speed comparison. The ground-truth network of Zachary's karate club is seen in Figure 1. According to the Figure 1, circles and squares are 2 different factions, which we can consider as community.

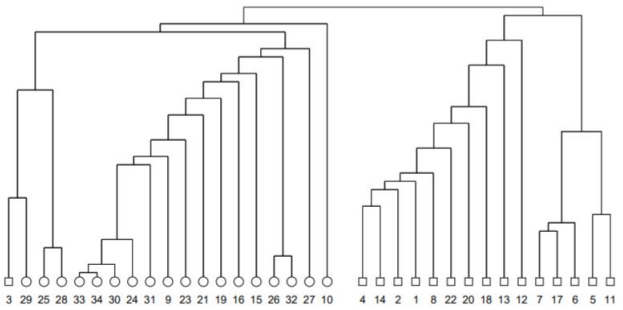

Figure 1: The friendship network of Zachary's karate club [2].

- **Girvan-Newman**

  I implemented Girvan-Newman method to detect the communities. You can see the detection result in Figure 2. According to the results, only the Node 2 was wrongly detected. As expected, the completion time was longer than the others. It lasted 24.7 seconds. It is predictable because the worst-case complexity of the algorithm is square of number of edges times number of vertices.

- **Label Propagation** It is a fast algorithm but it did not give the good result on Zachary's karate club. It detected 3 communities. Node 4, 5, 6, 10, 16 are detected as a community even though it is not a separate community. Also, Node 9 was wrongly detected. You can see the detection result in Figure 3. The total completion time was 0.1 seconds.

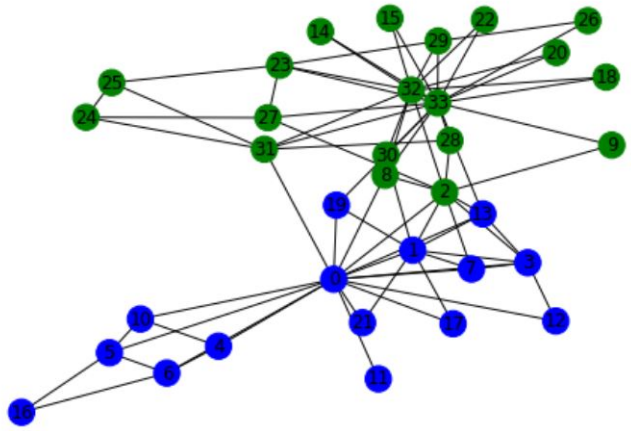

Figure 2: Community detection result on Zachary's karate club using Girvan-Newman method.

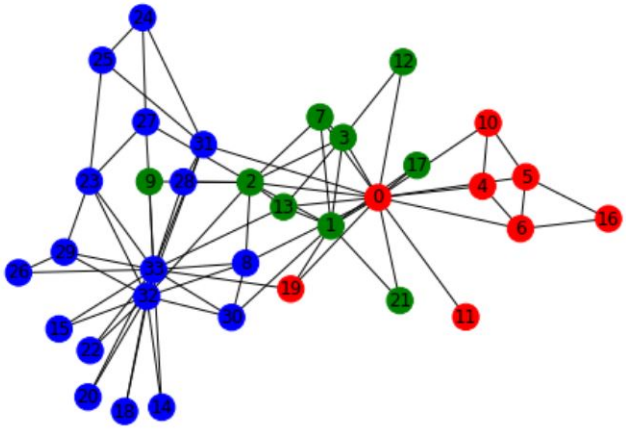

Figure 4: Community detection result on Zachary's karate club using Clauset-Newman-Moore greedy modularity maximization method.

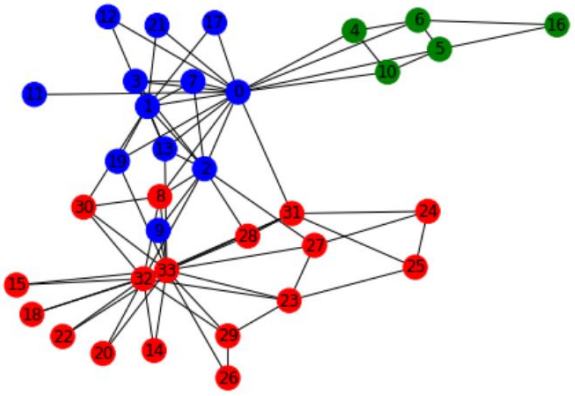

Figure 3: Community detection result on Zachary's karate club using label propagation method.

- **Clauset-Newman-Moore** I observed the worst accuracy using Clauset-Newman-Moore method. It wrongly detected the Node 9 and found 3 different communities. It lasted 0.12 seconds, which is close to the label propagation. It is a fast algorithm because worst-case complexity is number of edges times number of nodes times depth of the dendrogram. You can see the detection result in Figure 4.

- **K-means**

  As I explained in Section 2.2, I adapted k-means on community detection. Firstly, I tried with the network in Figure 5, which is a South African company network contains only 18 nodes. I tried it on first to check if my approach is correct. It could be seen that there are 3 clusters. K-means found all the communities of the nodes correctly. Then I tried with the Zachary's karate club network. When I checked the accuracy, I obtained that only 3 nodes were wrongly detected.

  Utilizing k-means on community detection is the fastest method by far. It is completed less than 0.03 seconds, which is 3 times faster than label propagation and Clauset-Newman-Moore algorithms. Hence, we can say that it is an accurate and fast method. However, one of the drawback is data preparation step. The data comes with only the edges written in the file. That's why, we need to calculate the distances between nodes.

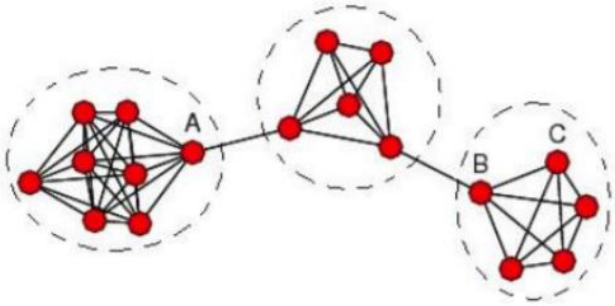

Figure 5: Community structure of the South African company [8].

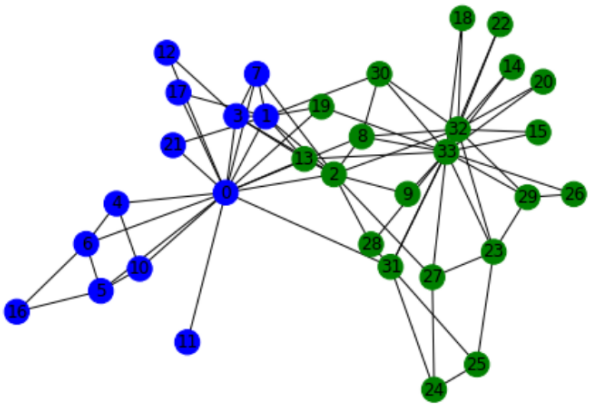

Figure 6: Community detection result on Zachary's karate club using k-means method.

You can see all the results in Table 1. According to it, K-means and Girvan-Newman are the most accurate methods, which the number of wrong detected nodes are 1 and 3 respectively. However, Girvan-Newman has a major drawback, which is the long completion time.

Table 1: Number of wrong detected nodes and completion time of the 4 methods.

Method	Number of Wrong Detections	Completion Time (s)
Girvan-Newman	1	27.4
Label Propagation	6	0.10
Clauset-Newman -Moore	9	0.12
K-means	3	0.03

4 Conclusion

In that project, I compared 4 different methods for community detection, such as Girvan-Newman, Clauset-Newman-Moore greedy modularity maximization, label propagation, and k-means. Adapting k-means on community detection is a novel approach that I proposed. In conclusion, I observed that Girvan-Newman is an accurate algorithm, however, the completion time is too long. There are some faster algorithms but they did not give the accurate results on Zachary's karate club. Since k-means is the fastest and accurate algorithm, it can be used on community detection.

References

[1] A. Bóta, M. Krész, and B. Zaválnij, "Adaptations of the k-means algorithm to community detection in parallel environments," in *2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*. IEEE, 2015, pp. 299–302.

[2] 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.

[3] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, "Generalized louvain method for community detection in large networks," in *2011 11th International Conference on Intelligent Systems Design and Applications*. IEEE, 2011, pp. 88–93.

[4] J. Xie and B. K. Szymanski, "Community detection using a neighborhood strength driven label propagation algorithm," in *2011 IEEE Network Science Workshop*. IEEE, 2011, pp. 188–195.

[5] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognition*, vol. 36, no. 2, pp. 451–461, 2003.

[6] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of anthropological research*, vol. 33, no. 4, pp. 452–473, 1977.

[7] "Physicians dataset," http://konect.cc/networks/moreno_innovation/, accessed: 2021-06-23.

[8] I. Durbach, D. Katshunga, and H. Parker, "Community structure and centrality effects in the south african company network," *South African Journal of Business Management*, vol. 44, no. 2, pp. 35–43, 2013.