Louvain & Girvan-Newman

Which one is better for a large social network?

Community Detection Performance Comparison

I. Introduction

Undoubtedly, with the rapid development of internet technology, human social activities have gradually shifted from reality to social media. In addition, thanks to the spread of mobile devices and their reduced cost of use, the possibilities for human connection are endless. Online social networking has gone through a number of phases since the late 1990s, with platforms such as Youtube, Facebook, Twitter, Instagram and TikTok being the most representative [1]. The development of online social networks has allowed people to communicate and interact virtually and has given rise to many new industries that have had a significant impact on the development of modern society. As a result, social network plays an important role in social research and business activities, and the use of community detection to analyse social networks is of particular importance. It can be used to discover potential relationships and structures hidden in the network. In business, community detection helps companies to gain insight into their customers and markets and provide better products and services, which in turn improves their competitiveness and economic efficiency. Take social media marketing as an example, community detection can help companies find valuable social groups on social media, understand their interests and needs, and develop more targeted marketing campaigns. In social science， community detection can help researchers to uncover the hidden structures and relationships in social networks and further understand the workings and characteristics of human society. For instance, social influence analysis can help researchers understand the important players and influences in social networks and in turn study the formation and evolution of social decisions and behaviours.

The complexity of today's social networks and the vast amount of data available pose a challenge to community detection, and effective community detection requires appropriate algorithms and methods. In other words, due to the diversity and specificity of social networks, choosing a suitable community detection algorithm is complex and difficult. In the previous algorithm taught in the lecture, the Greedy algorithm, the disadvantage of the Greedy algorithm is mainly that it tends to fall into local optima and is unable to find the global optimal solution [2]. As a consequence, the results of the Greedy algorithm are often not accurate enough, especially for complex network structures [3].

Based on the aforementioned, in this mini-project, I would like to focus on two modern algorithms, Louvain and Girvan-Newman algorithms which were taught in the lecture as well. These two methods have the advantages of high computational efficiency, modularity and low time complexity, and can therefore handle overlapping communities. Although these two algorithms have many advantages, they also make many people hesitant to choose between them, especially students like us, who want to perform community analysis on a complex network, and which of these two algorithms is more suitable. Therefore, the problem I want to solve is to investigate the efficiency and performance of the two algorithms in classifying the nodes of this huge social network into different communities, using the Facebook social network as the database, which has 4039 users and 88,324 links between users. The final conclusions are drawn for future research.

II. Description of methods

Since this experiment is focusing on performance comparison; hence, the methods I will select two algorithms from the lecture，as our main discussion to find the optimal solution to solve the community detection of large social networks.

A. Girvan-Newman

The Girvan-Newman community detection algorithm is a method based on the edge betweenness centrality to identify the community structure in a social network [4]. This algorithm determines the importance of an edge by calculating its edge betweenness centrality, which is the number of shortest paths in the network that pass through the edge. The higher the edge betweenness centrality, the more important the edge is in connecting different parts of the network [5]. The basic idea of the algorithm is to iteratively remove edges with the highest betweenness centrality until the network is divided into different communities. Each removal of an edge will change the community structure in the network, ultimately resulting in multiple distinct communities [6].

B. Louvain Algorithm

The Louvain algorithm is a community detection algorithm based on maximising modularity to identify community structures in social networks. The basic idea of the algorithm is to partition the nodes in the network into different communities, in which the nodes are closely connected to each other while being sparsely connected to nodes in other communities [7]. The optimisation objective of the algorithm is to maximise the connectivity within each community while minimising the connectivity between different communities.

C. Performance Assessment

Based on the study from the lecture, using modularity to evaluate the effectiveness of community partitioning or community detection is a good method because it can quantify the tightness of the nodes within the community and the looseness of the nodes between communities, reflecting the quality of the community partitioning. Modularity can be understood as the difference between the density of connections between nodes within a network and the density

of connections between nodes within a community [3]. It is a value between -1 and 1. If the modularity is close to 1, it means that the nodes within the community are closely connected, and there are fewer connections between communities. If the modularity is close to 0, it means that the connections between nodes within the community are the same as the connections within the network, and the partitioning is not significant. If the modularity is close to -1, it means that there are fewer connections between nodes within the community and more connections between communities. Modularity calculation can comprehensively consider the connection status of nodes within communities and between communities, avoiding the limitations of node indicators such as degree centrality and betweenness centrality. Therefore, it is a more objective and comprehensive evaluation metric for community partitioning. As a result, modularity has become a widely used evaluation metric in the field of community detection.

III. Discussion about experiments

In this section, I will introduce my dataset in detail and describe how I do my experiment. Finally, I will analyse the results and give the table and plot to prove the better algorithm.

A. Dataset

The Facebook social network was chosen as the experimental data for community detection because it is a real social network and contains a large number of users and connections. The dataset was extracted from real Facebook user data; therefore more closely reflects the characteristics of real social networks. The detail of the dataset shows as follow:

1) 10 subnets: Each net can represent a small community.
2) 4039 nodes: Each node has its own ID, which represents the real users on Facebook.
3) 8823 edges: Including the edges between the nodes in the sub-network. The edges on this whole network are undirected for Facebook, which means the pure relationships between users.
4) Egonets: Edges from all subnets combined, which means the edges connecting different sub-networks.
5) Features: The features for each of the nodes. However, due to privacy reasons, while feature vectors from this dataset have been provided, the interpretation of those features has been obscured. For example, using anonymised data, it is possible to determine whether two users have the same political affiliations but not what their individual political affiliations represent.

Overall, this dataset represents real human relationships and is large enough to be used to validate the efficiency and usability of the algorithms.

B. Experiment Progress

In this experiment, there are basically three steps which show as follows:

1) Step1. Data Pre-Process

Due to the size of the dataset I am using, it is necessary to pre-process the dataset. Here are some important things I need to do.

- Removing duplicate edges: Duplicate edges have no real meaning in the graph and may affect the result of community detection. Therefore, duplicate edges should be removed from the graph, leaving only one edge.
- Removing self-loop: Self-loops refer to edges with the same start and end nodes, which also do not belong to any community and need to be removed.
- Handling edge weights: Some graphs have weighted edges, which represent the similarity or importance between two nodes, and these weights can be considered in community detection. If edge weights need to be considered, they should be saved in the graph. However, since this experiment is focusing on the Facebook social network, the edge represents a human connection, so there is no need to consider the weighting issue.
- Feature extraction: For graphs with node features, feature extraction needs to be performed before community detection. In this experiment, I need to match all the features to the nodes that can be used in community detection.

2) Step2. Using algorithms to do community detection

The algorithms I have already mentioned above in Section II. The following are the specific steps of the two algorithms.

- Louvain Algorithm:
  (1) First, each node is considered as a community. The total modularity value of the network is calculated.
  (2) Iterate through each node and move the node to the community where its neighbouring nodes belong if this movement can increase the modularity value.
  (3) For all newly formed communities, recalculate the modularity value. If the modularity value increases, continue node movement until no further improvement can be made.
  (4) For all new communities, they are merged into a new node to form a new graph. Repeat the above process until no further merging can be made.
  (5) Complete communities division.
  (6) Calculate the total weight (in this experiment, is 0) of the entire network and the degree of each node. Then, for each community, calculate the total weight between the nodes in the community and the sum of the degrees of the nodes in the community.
  (7) Finally, calculate the modularity of each community in the network and add them up.
- Girvan-Newman Algorithm:
  (1) Calculate the betweenness of all edges in the network, which is the number of times the edge appears in all shortest paths.

(2) The edge with the highest betweenness is found and removed . The deletion will result in two or more connected subgraphs.

(3) For each connected subgraph, calculate the sum of the meshes of all the edges within it and select the edge with the largest meshes to delete, again creating two or more connected subgraphs. And repeated until all the nodes have been grouped into a connected subgraph.

(4) Based on the different connected subgraphs decomposed, the network is hierarchically clustered, with each connected subgraph considered as a cluster, and all clusters are then combined.

(5) The best division of the clustering result is selected as the community division of the network according to the features Facebook provided.

(6) It should be noted that since the Girvan-Newman algorithm is a bottom-up algorithm, it may divide some small communities into smaller sub-communities. Also, during the experiment, I found that the time complexity of the algorithm was high and took quite a long time, which made it impossible for my experiment to proceed smoothly; therefore, after dividing the community, I chose to use the Louvain algorithm to calculate the modularity.

3) Step3. Visualise the results

Through the visual results, you can more intuitively understand the performance difference between the two algorithms for complex network analysis.

C. Result and Analysis

After the process of the experiment, we can separate the result into two parts, value and graph. One is a quantitative comparison of results, and the other is a visual comparison of graphs.

1) Modularity

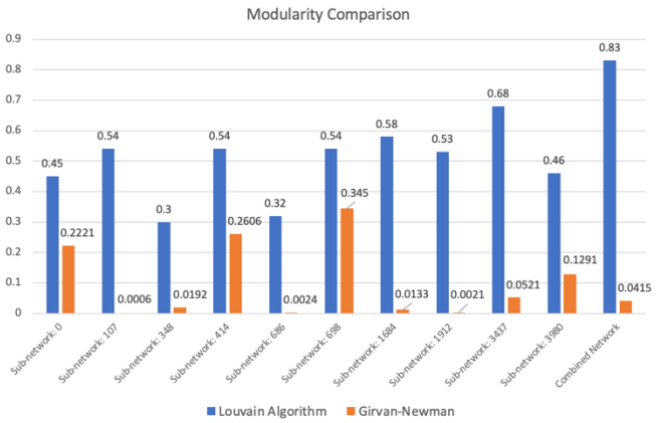

Fig. 1. The modularity of two algorithms

According to Fig.1, we can see that the modularity of the Louvain algorithm is far better than the Girvan-Newman algorithm. In the full network part (combined network), the performance of the Louvain algorithm is 20 times higher than the Girvan-Newman algorithm.

2) Plot

Fig. 2. Full network (Louvain).

Fig. 3. Full network (Girvan-Newman)

From the observations in Fig. 2 and Fig. 3, it can be concluded intuitively that different colours represent different communities, and the Louvain algorithm can achieve better results in classifying nodes and dividing communities, while the Girvan-Newman classification results in most of the nodes being classified in the same community.

Through the study of the comparative results of the algorithms in this experiment, it can be learned that the Girvan-Newman algorithm is a bottom-up algorithm that may divide some small communities into smaller sub-communities, and therefore a suitable threshold needs to be set to control the number of communities. Also, the algorithm has a high time complexity and is therefore not suitable for dealing with large-scale networks. The community detection problem is essentially an NP-hard problem, although the output of Louvain's algorithm may not be a globally optimal solution. However, in practice, compared with the Girvan-Newman algorithm, the Louvain algorithm has proved to be an efficient and effective algorithm for the fast discovery of community structures in large-scale networks.

IV. Conclusions

In conclusion, through this mini-project, we have solved the problem of selecting algorithms from Louvain and Girvan-Newman algorithms for community detection and node classification in large-scale networks so that users or other students can have a quantitative data reference to choose the best community detection method when they research of network analysis. This report will provide a valuable reference for future research. However, there are still some problems that have not been addressed in this experiment, such as the lack of comparison of correct rates, so there is still much work to be done in the future, and I hope I can still work in this field.

REFERENCES

[1] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?," in Proceedings of the 19th international conference on World wide web, 2010, pp. 591-600.

[2] J. Sanchez-Oro and A. Duarte, "Iterated Greedy algorithm for performing community detection in social networks," Future Generation Computer Systems, vol. 88, pp. 785-791, 2018.

[3] 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.

[4] X. Liu, C. Hou, Q. Luo, and D. Yi, "Uncovering community structure in social networks by clique correlation," in Modeling Decision for Artificial Intelligence: 8th International Conference, MDAI 2011, Changsha, Hunan, China, July 28-30, 2011, Proceedings 8, 2011: Springer, pp. 247-258.

[5] S. Zhang, H. Zou, and J. Sun, "Knowledge mapping analysis of manufacturing product innovation based on CiteSpace," Journal of Circuits, Systems and Computers, vol. 31, no. 07, p. 2250121, 2022.

[6] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," Physical review E, vol. 69, no. 2, p. 026113, 2004.

[7] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," Journal of statistical mechanics: theory and experiment, vol. 2008, no. 10, p. P10008, 2008.