E0 259 Assignment-1

Aditya Gupta
SR No: 22205

Question 1

The Girvan Newman algorithm is a clustering algorithm that separates out the nodes of a graph by removing edges one by one. In each iteration of the algorithm, we calculate the edge betweenness centrality of each edge and then remove the highest one. This algorithm was run on two datasets for this assignment. Their details are as follows:

1. Wikipedia vote network: This network contains all the Wikipedia voting data from the inception of Wikipedia till January 2008. The nodes in the network represent the Wikipedia users and directed edges from node A to node B represent that user A voted on user B. However for this assignment, the network was treated as undirected. The network has 7115 nodes and 103689 edges.
2. LastFM Asia Social Network: This is a social network of LastFM users which was collected from the public API in March 2020. The nodes represent users from Asian countries and the edges are mutual follower relationships between them. The network has 7624 nodes and 27806 edges.

Question 2

To determine the set of communities, we will look at connected components in the graph. A connected component is a subgraph in which there is a path between every pair of nodes. Each member of a connected component is a part of the same community. As we keep removing edges from the graph, it will split into more and more connected components. Thus the number of communities will be equal to the number of connected components in the graph.

To determine the stopping criterion, we look at the modularity of the graph. Modularity is a measure of the quality of a partition of a graph into communities. It is calculated as follows:

$$Q = \frac{1}{2m} \sum_{p \in P} \sum_{i \in p} \sum_{j \in p} \left(A_{ij} - \frac{\delta_i \delta_j}{2m} \right) \tag{1}$$

The best partition of a graph is the one with the maximum modularity. Hence we should ideally remove every edge iteratively and then find the global maximum of the modularity. However, this is computationally expensive. Instead, we will stop the algorithm when the modularity of the graph hits a local maximum.

Question 3

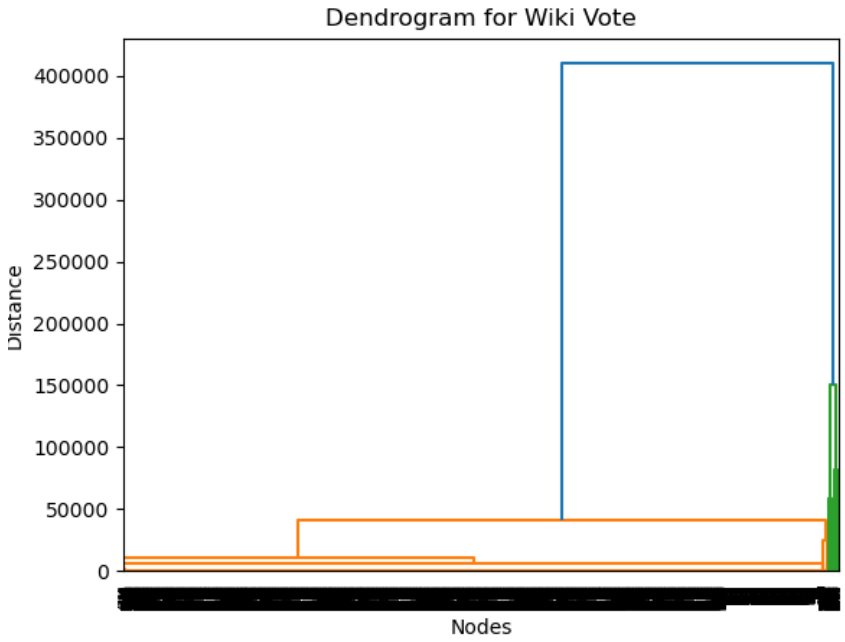

Figure 1: Wikipedia vote network

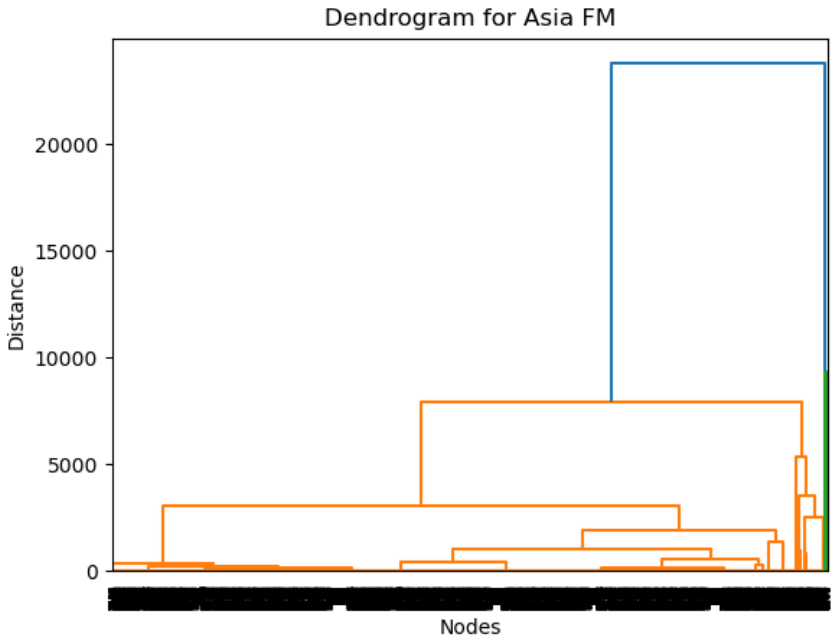

Figure 2: LastFM Asia Social Network

Question 4

The Louvain algorithm is another algorithm which can be used to detect communities in a graph. It starts by assigning each node to its own community. Then it iteratively moves nodes to communities where they get the maximum modularity gain. The algorithm stops when the modularity of the graph hits a local maximum. This is a non deterministic algorithm since the order in which the nodes are considered for merging will affect the final result. However for large enough graphs, the results are usually stable.

For the Louvian algorithm, we measure the change in modularity for moving node i from its current community X to a new community Y as follows:

$$\Delta Q = \frac{1}{2m}\left(2(deg(i,Y) - deg(i,X)) - \frac{deg(i)}{m}(deg(Y) - deg(X))\right) \tag{2}$$

where $deg(i,Y)$ is the number of edges from node i to nodes in community Y, $deg(i,X)$ is the number of edges from node i to nodes in community X, $deg(i)$ is the degree of node i, $deg(Y)$ is the total degree of nodes in community Y and $deg(X)$ is the total degree of nodes in community X.

Question 5

To pick the best decomposition, we can use the modularity of the graph as a metric. We assign communities by looking at connected components. This approach makes sense because nodes which do not have a path between them should not be in the same community. To partition the graph quickly, we remove the edges with the highest betweenness centrality. This is because the edges with the highest betweenness centrality are the ones which are most likely to connect different communities. Finally, we evaluate the modularity of the graph to determine the best decomposition.

Question 6

The running time for the Girvan Newman algorithm was about 6 hours for the Wikipedia vote network and 8 hours for the LastFM Asia Social Network. This algorithm is very slow since it has to calculate the betweenness centrality of all the edges in the graph in each iteration. This is computationally expensive and hence the algorithm is slow. To speed up the algorithm, parallel computation was used since the betweenness centrality of each edge can be calculated independently. However, the algorithm was still slow.

The Louvain algorithm was much faster. It took about 3 seconds for the Wikipedia vote network and less than 1 second for the LastFM Asia Social Network. This algorithm is faster because it only checks the neighbours of a node to determine the change in modularity. This is much faster than calculating the betweenness centrality of all the edges in the graph.

Question 7

Upon running both the community detection algorithms, the Girvan Newman algorithm found 56 communities in the Wikipedia vote network and 24 communities in the LastFM Asia Social Network. The Louvain algorithm found 55-60 communities in the Wikipedia vote network and 700-900 communities in the LastFM Asia Social Network. The number of communities for Louvain changes due to inherent randomness in the algorithm. The Louvain algorithm was also run only for one iteration. If it was run for more iterations, the number of communities would have been more stable and closer to the value obtained by the Girvan Newman algorithm.

In terms of giving rise to better communities, the Girvan Newman algorithm is the winner since it does an exhaustive search of the graph to find the best decomposition. It is also a deterministic algorithm, unlike the Louvain algorithm which is prone to suboptimal solutions due to its randomness. However, the Girvan Newman algorithm is very slow and hence not practical for large graphs. The Louvain algorithm is much faster and hence more practical for large graphs.