**Homework 3**   
 EE232E - Graphs and Network Flows

**Team:**

**Rutuja Ubale (UID: 404558257)**

**Usha Amrutha N (UID: 204590772)**

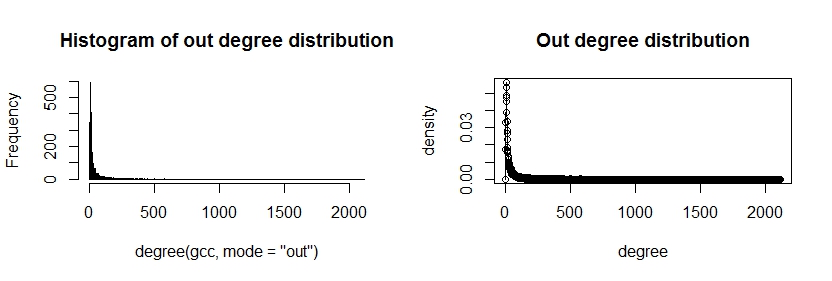
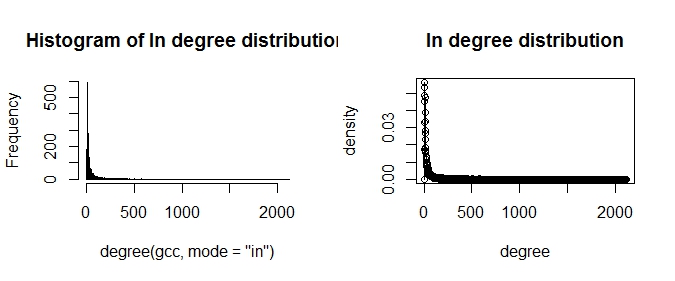
**Pallabi Chakraborty (UID: 404519609)**

**Q1. Determining connectivity of the network and GCC**

We created a directed network from the data given to us in directed edge list format. Each line is the data has three items, node 1, node 2 and the weight of the edge from node 1 to node2. We found that the network was not connected. Next, we extracted the giant connected component from the network and found that the number of nodes is 10487. Also, the number of directed edges is 427472.

**Q2. In and Out degree distributions**

In this question, we measure the degree distribution of in-degree and out-degree of the nodes. The plots are as shown below. We note that both the degree distribution plots are similar and are exponential in nature.

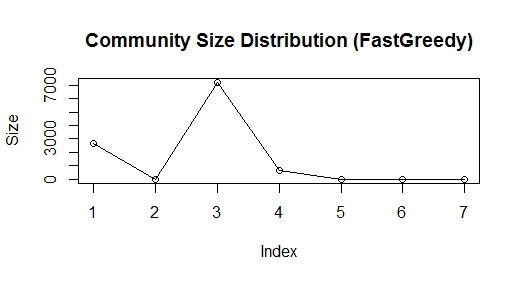


**Q3. Community Structure**

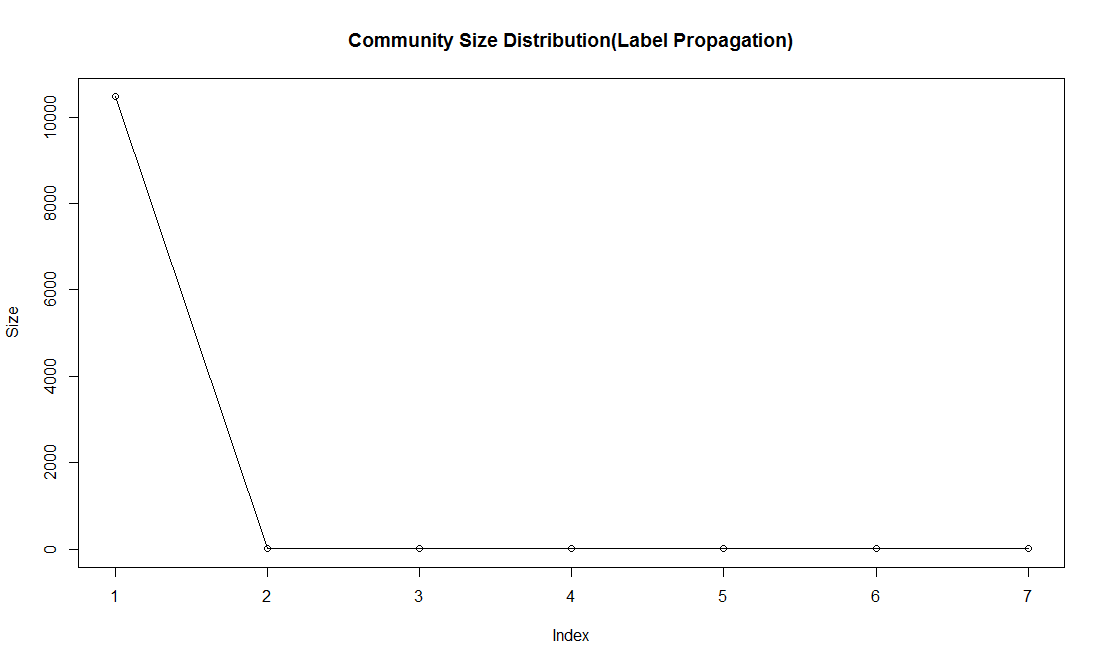
We first converted the graph to an undirected graph. For nodes with multiple edges between them, we collapsed the edges into 1 such that the new weight of this edge was the geometric mean of the weights of the former edges between the nodes. We use fastgreedy.community and label.propagation.community to measure the community structure, and then check if the results of the two methods are similar or not.

**Fast Greedy Community Detection -**

Using this algorithm, the number of communities detected turned out to be 8. Sizes of the 8 communities were 1856, 1666, 1022, 2266, 731, 1236, 633, 1077. The modularity is 0.328. The graph for the community sizes of each of the 8 communities is as shown below.



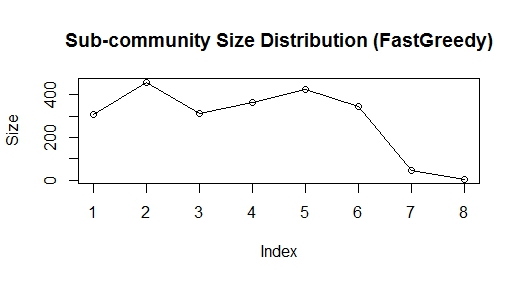
**Label Propagation Community Detection -**

Using this algorithm the number of communities detected were 6. Sizes of the 6 communities were 10469, 4, 5, 3, 3, 3. The graph for community size of each of the 6 communities is as shown below. 

We note that both the two community detection algorithms produced very different results. The Fast Greedy method optimizes a modularity score to find dense sub-graphs/ communities. Label propagation labels the vertices with unique labels and then updates the labels by majority voting in the neighborhood of the vertex. In the case of Fast Greedy community detection, the communities are found such that they have higher intra connectivity than inter connectivity between communities thus leading to a substantially different set of community sizes. However, for Label Propagation community detection, since the network is highly connected and the algorithm works on majority voting we can see that for this method most of the nodes in the network fall into 1 community.

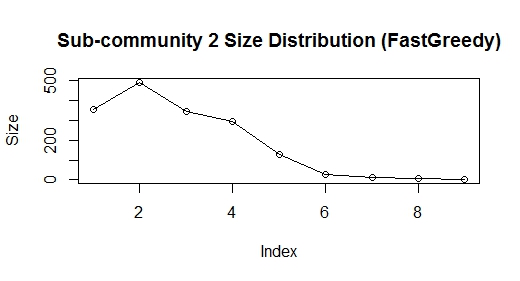
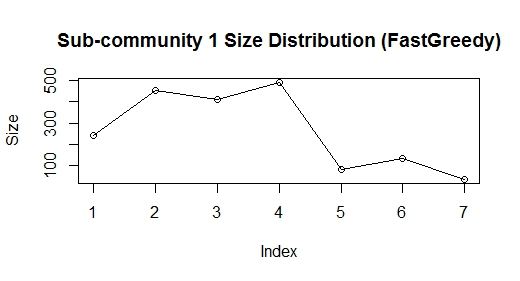
**Q4. Sub community Structure Detection**

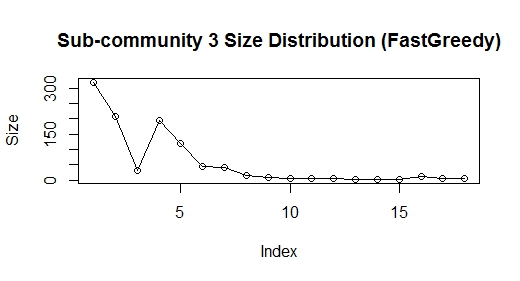
For this section, we found the largest community using Fast Greedy Community Detection. We then isolated the community from the whole network to form a new network and found the sub community structure on the new network using Fast Greedy algorithm. In the largest community detected, 8 sub communities were found. The sizes of the 8 sub communities were 306, 457, 313, 365, 426, 347, 47, 5. The modularity is 0.29. The plot of sub community size of the 8 sub communities is shown below.

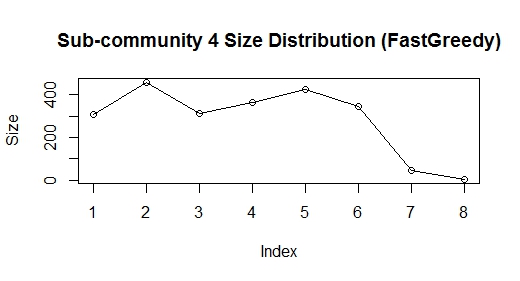


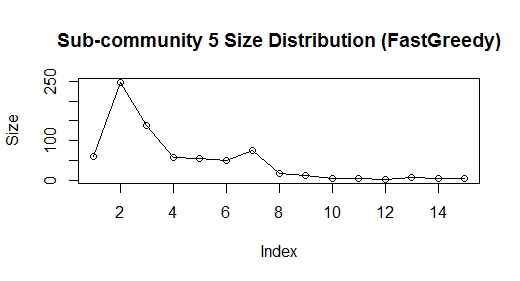
**Q5. Sub Community Structures of all the communities greater than size 100**

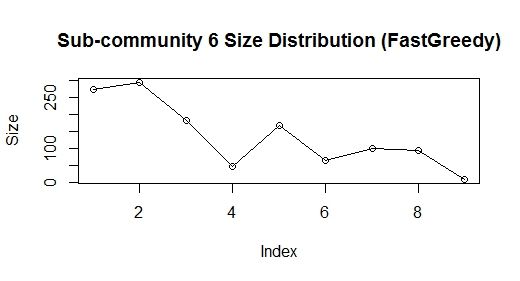
In this question, we find all the sub-community structures of the communities with size larger than 100. We see that all the communities have a size greater than 100 and hence plot the sub-community structure of all eight communities.

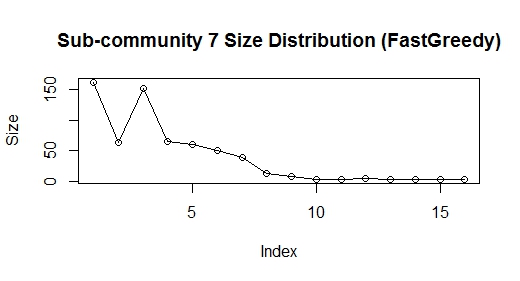


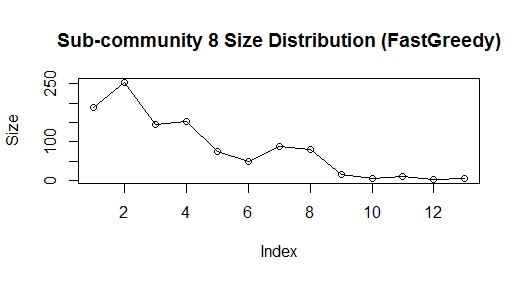












**Q6. Memberships of Nodes between multiple communities**

From the previous question we know that there are 8 communities with number of nodes greater than 100. We performed random walk experiment on every node of eight communities. We chose a threshold of 0.2. From the graphs, we see that the membership probabilities of community number 2 are very high. This is followed by community number 1 and then 4. It was observed that this follows the order of number of vertices in the 8 communities which share a strong correlation with the density of nodes.. The sizes of the 8 communities are 1856, 1666, 1022, 2266, 731, 1236, 633, 1077. The node numbers of the 8 nodes are 2, 7, 11, 23, 28, 29, 37, 41. The membership probabilities vs index of community plots for the first seven communities are shown below. Most of the nodes share multiple communities. However, we have listed only a few. The thick line across the graphs denotes the threshold line (0.2).

