ECE 232E Project 2 Report

Ken Gu, 904-836-308

Yingbo (Max) Wang, 604-593-537

Ziying Feng, 404-395-859

Siqi Huang, 504-490-530

# Part 1: Facebook Network

## 1. Structural Properties of the Facebook Network

### Question 1

The graph of the given undirected Facebook network is connected. This graph has size 88234 (the total number of edges) that equals the size of its Giant Connected Component.

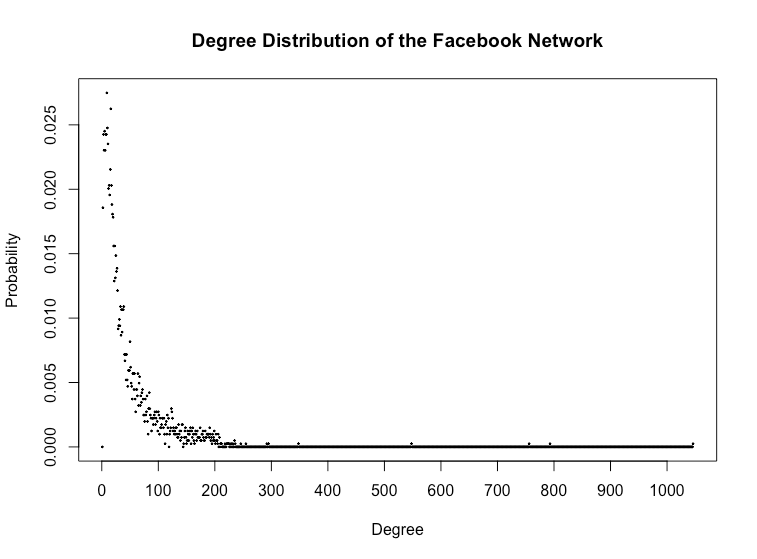
### Question 2

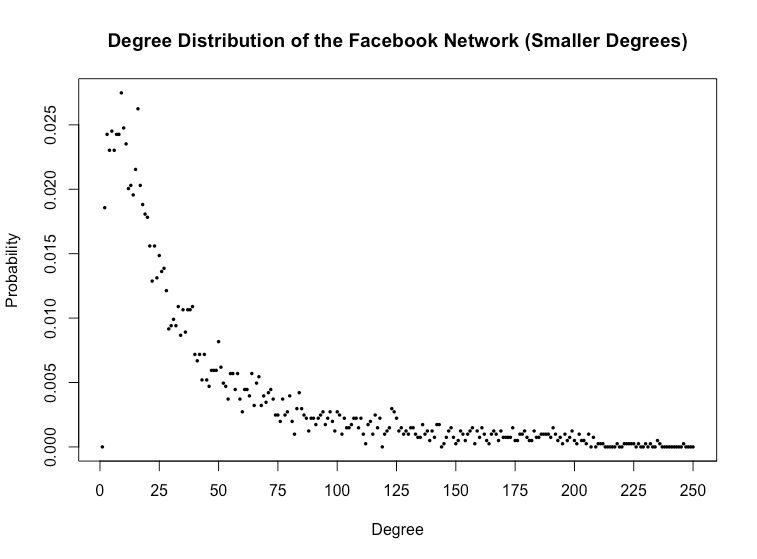
The diameter of the Facebook network is 8. Since the graph is connected, its GCC is just the graph itself, so the GCC also has the same diameter.

### Question 3

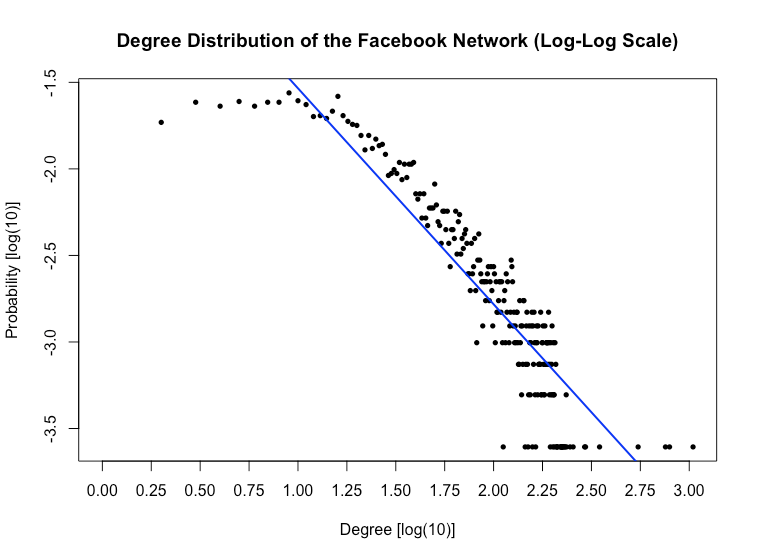
Two degree distribution plots of the Facebook network are shown below. We find that the highest degree of a node in this graph is 1046, meaning that this node has 1046 edges to other nodes. The first plot shows the degree distribution of all nodes. We know from its heavy-tailed nature that very few nodes have a degree of more than 300, as the probability of these high-degree nodes occurring is very close to 0. Thus, we generate a second plot that only consists of nodes with degree less than 250, which better visualizes the degree distribution. It’s seen that the probability that a node has a degree of about 10 is the highest.

The graph’s average degree is 43.691 - it’s higher than the degrees with the highest probability of occurring due to the nodes in the heavy tail with very high degrees.





### Question 4



Both axes of degree and degree probability are changed to log-scales with base 10. For most nodes with a degree in the middle of the range, the log-probability of a particular degree is decreasing almost linearly as the log-degree goes up. Thus, it’s reasonable to fit a linear model to the plot and this line has slope -1.2475 and intercept -0.2871.

## 2. Personalized Network

### Question 5:

This personalized network has 348 vertices and 2866 edges.

### Question 6

The trivial upper bound is 2, and trivial lower bound is 1.

### Question 7

Lower bound is achieved when all the subgraph is fully connected. Upper bound is achieved otherwise since every node can reach any other node within 2 steps.

## 3. Core Node’s Personalized Network

**Question 8:**

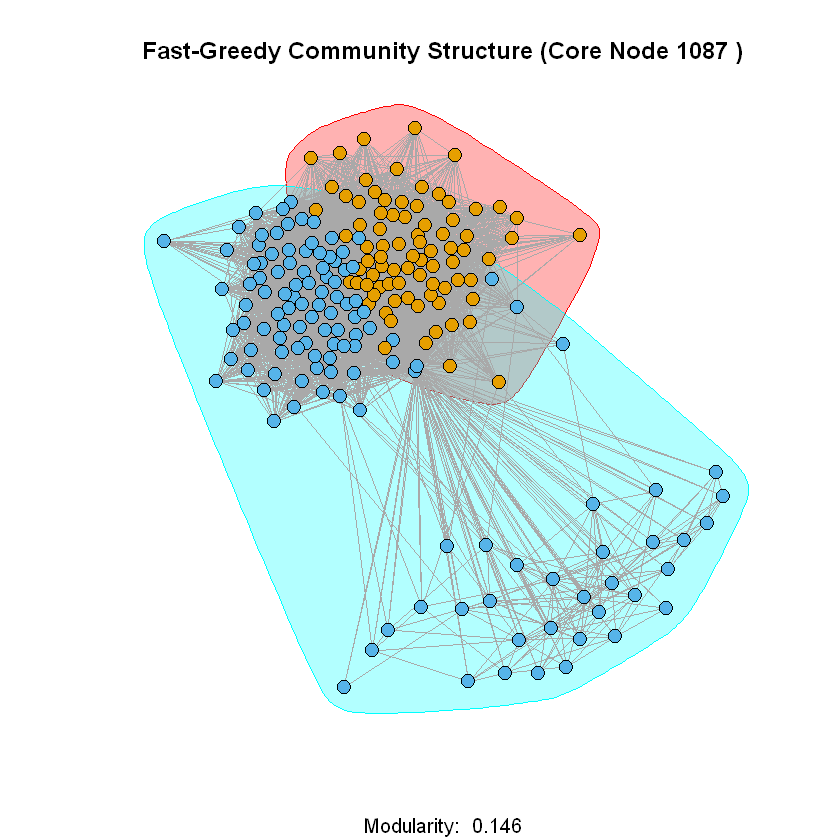
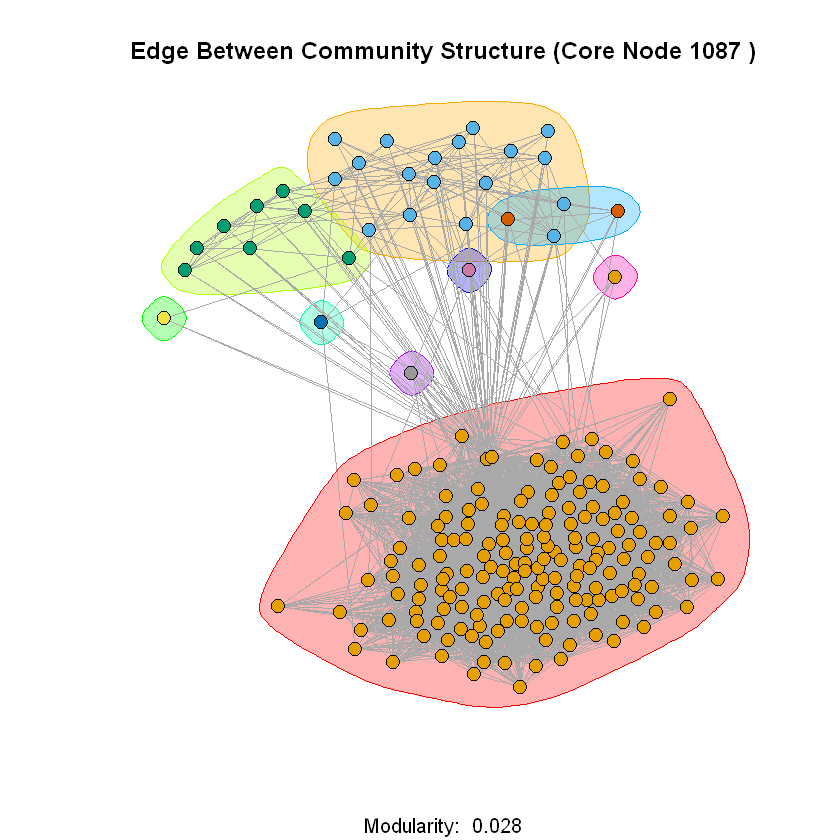
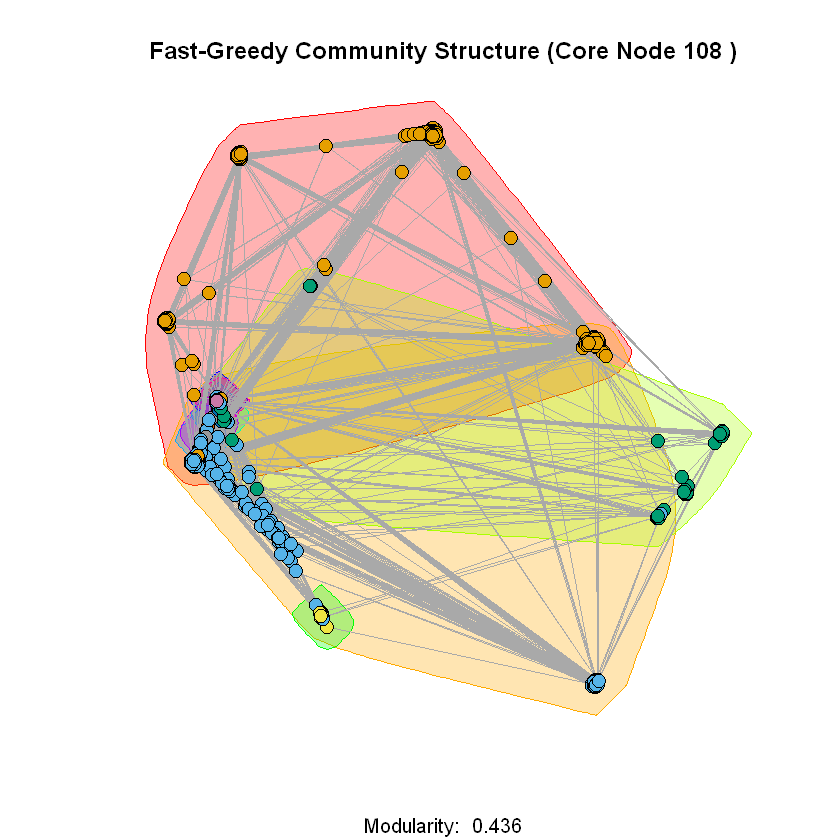
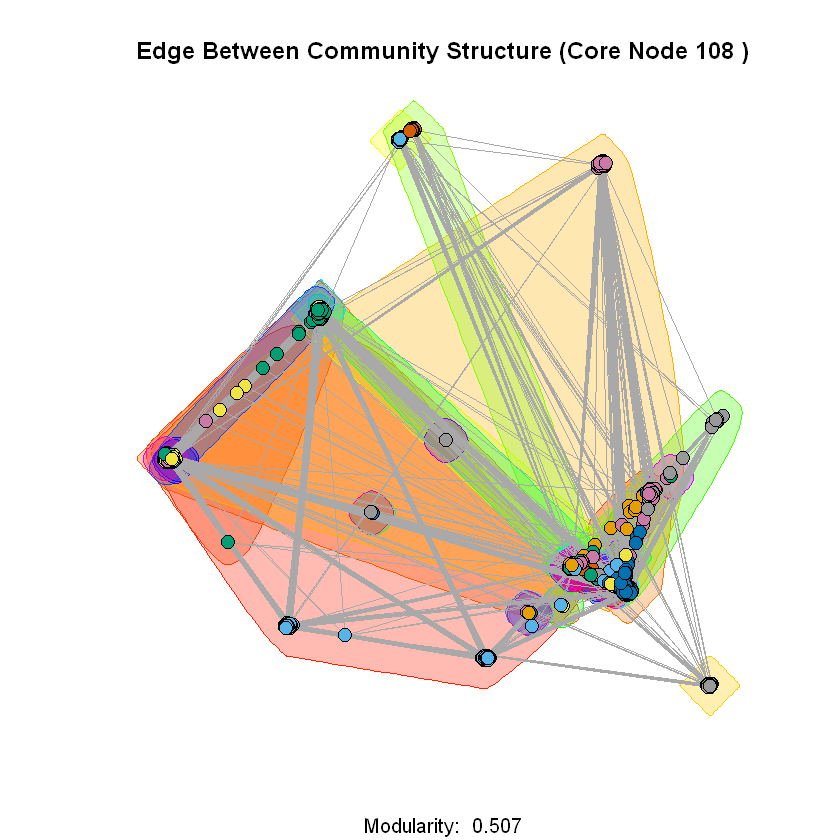
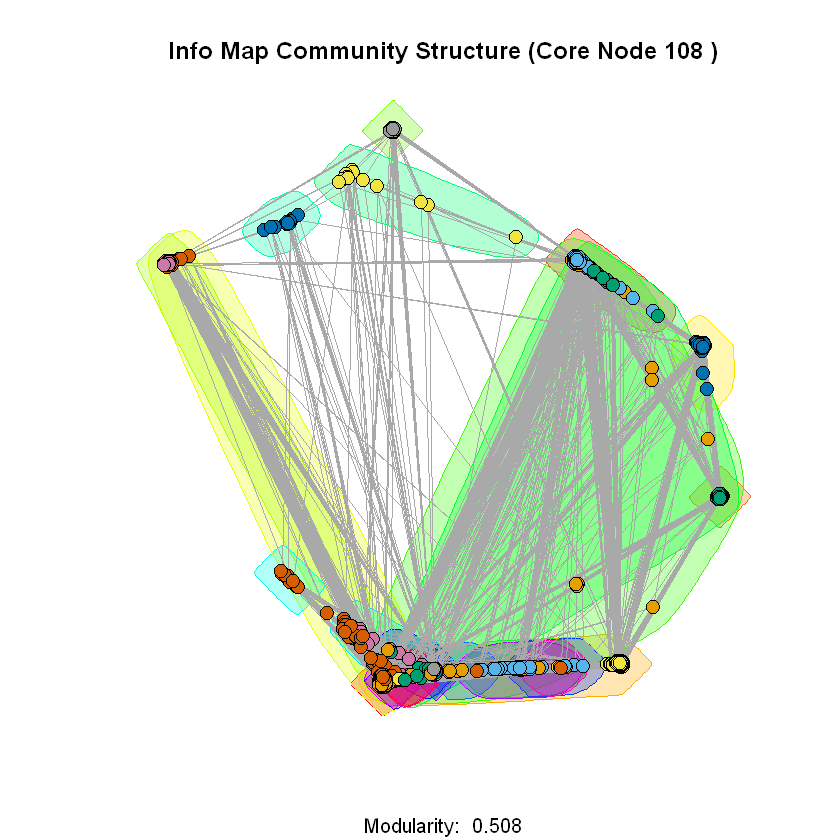
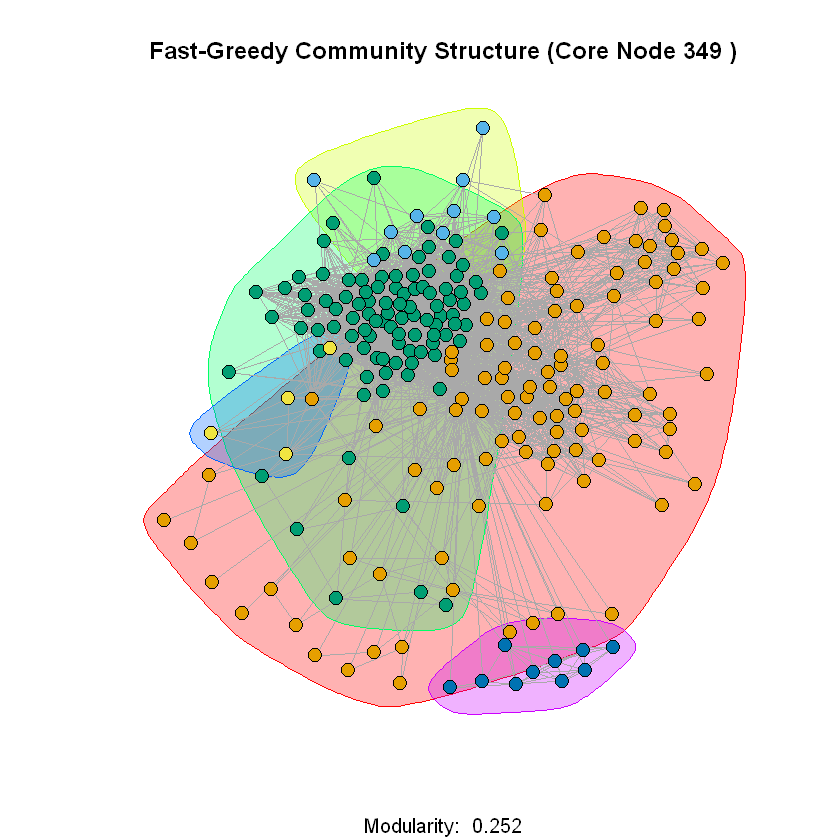
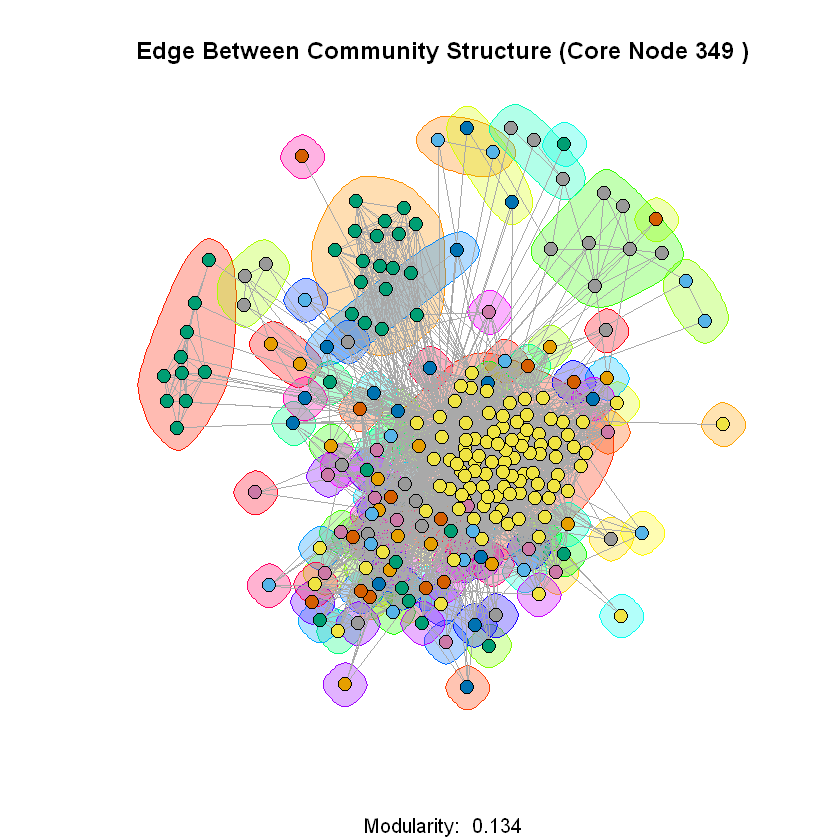
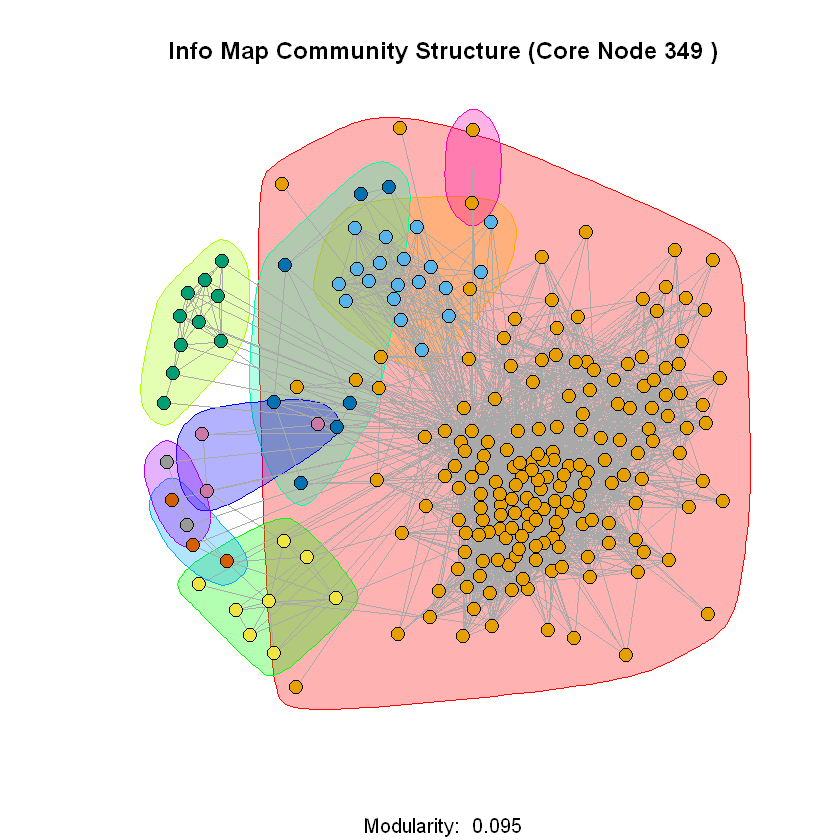
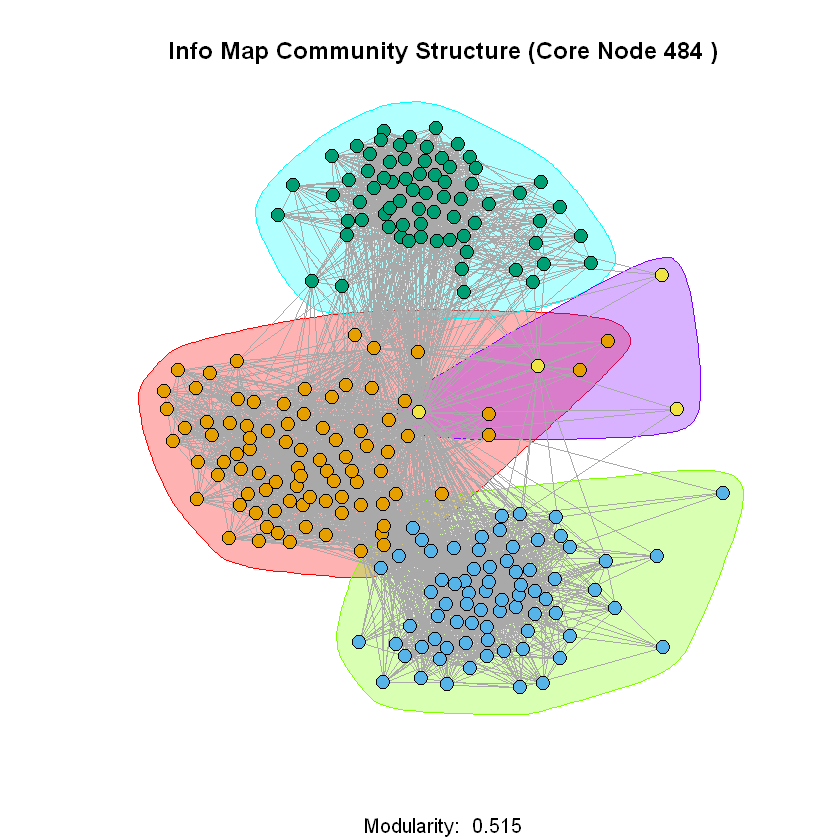
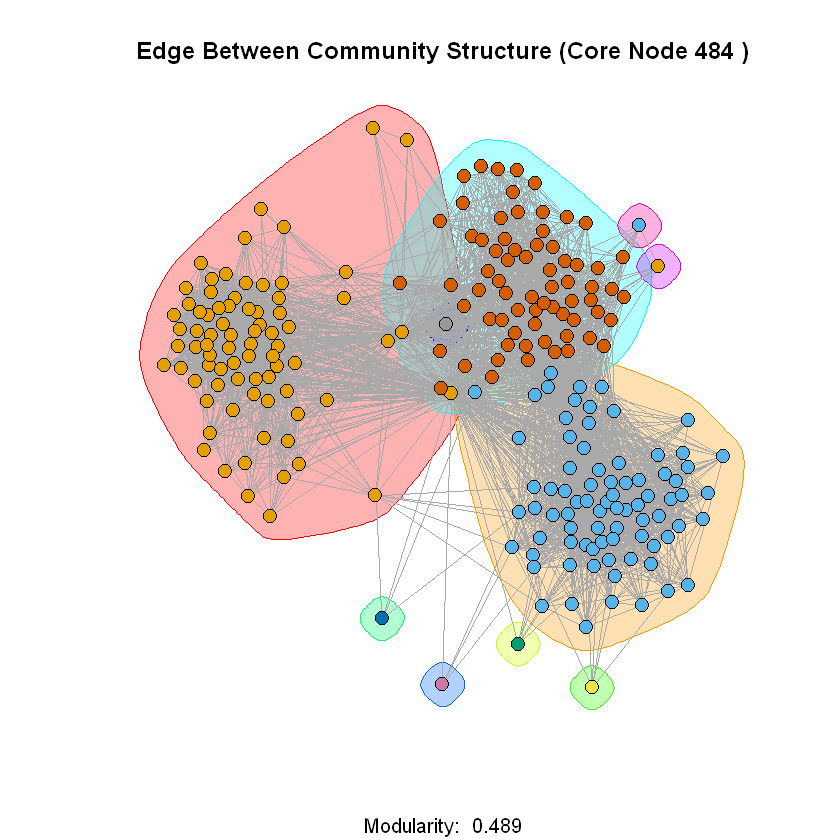
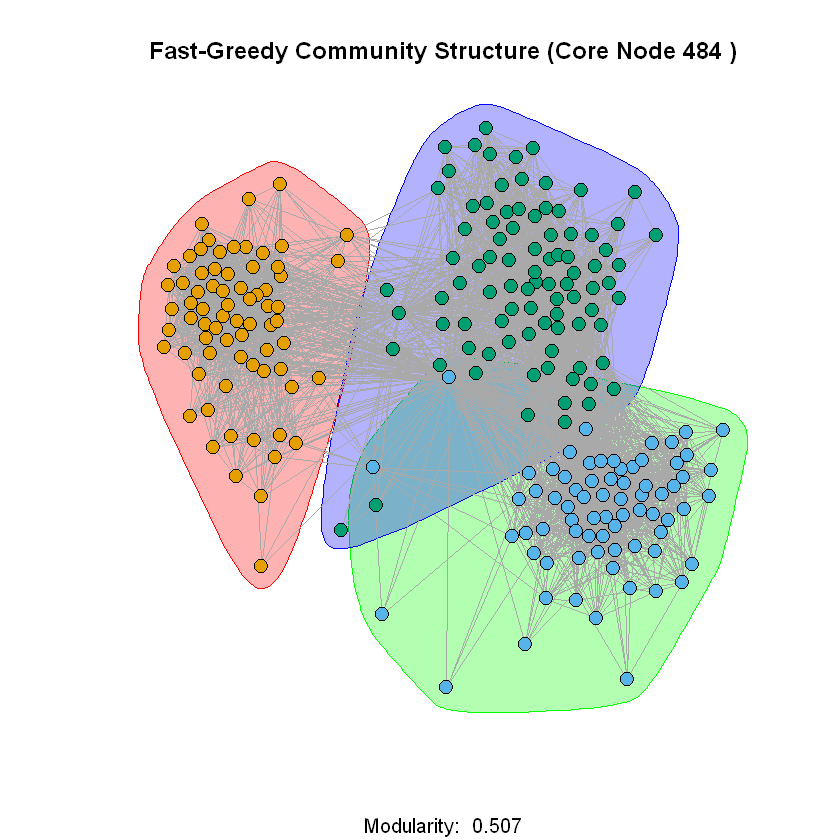
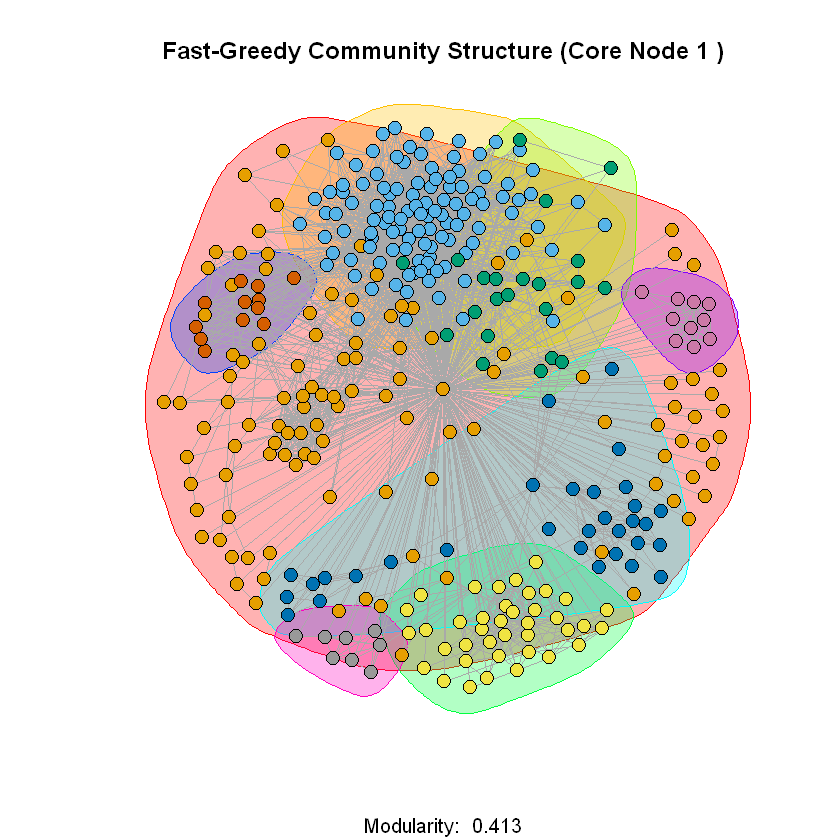
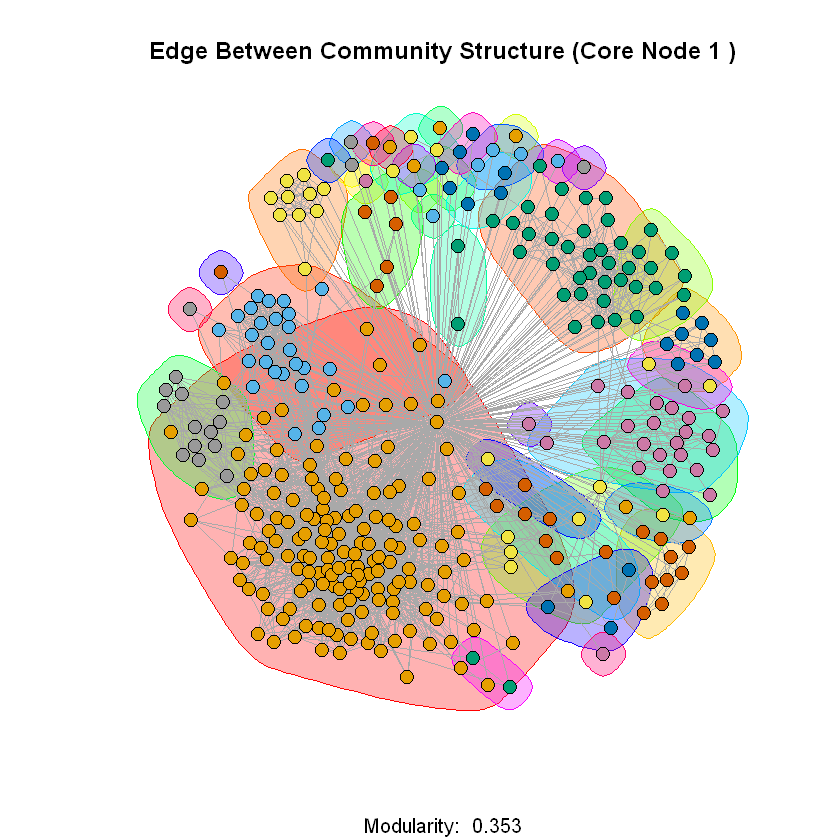
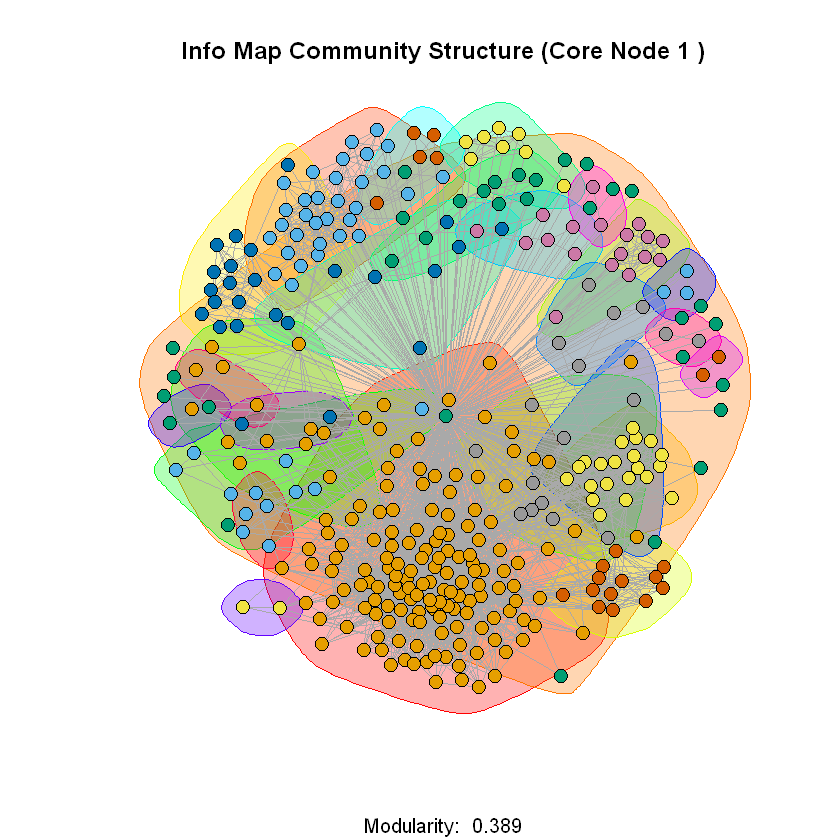
There are 40 core nodes with an average degree of 279.375

**3.1 Community Structure of Core Node’s Personalized Network**

**Question 9**

For each of the core nodes we explore the results of three community detection algorithms, namely Fast-Greedy, Edge-Betweenness and Info Map. Below are the measures of the modularity for each of the algorithms on each of the 5 core nodes. Averaging the results for each algorithm we can see that in general Fast-Greedy tends to produce results with a higher modularity score meaning the Fast-Greedy is a good algorithm for finding communities.

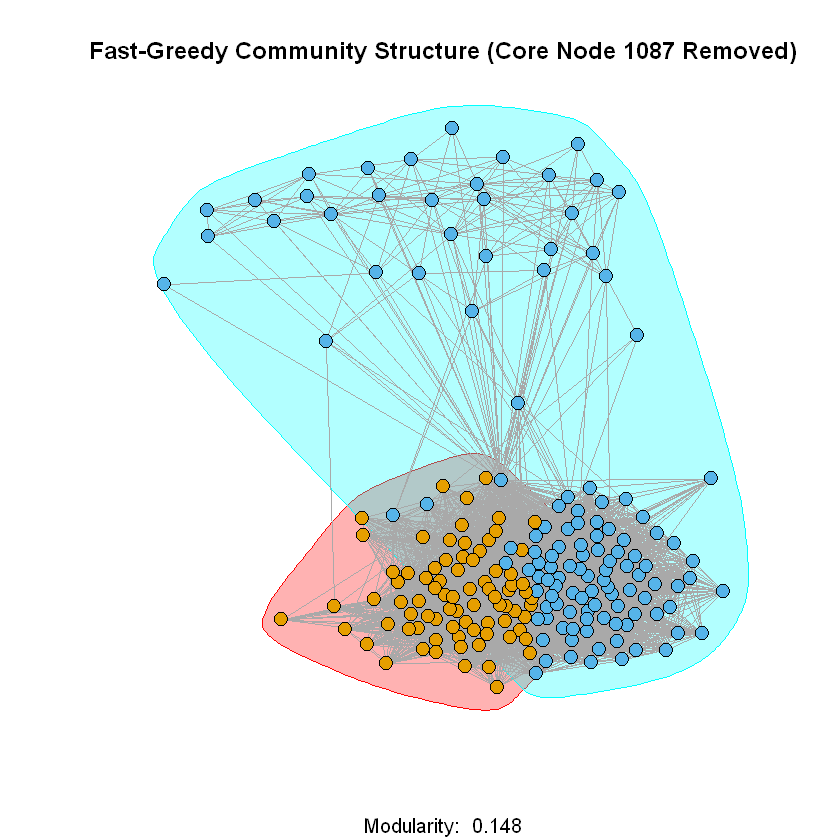
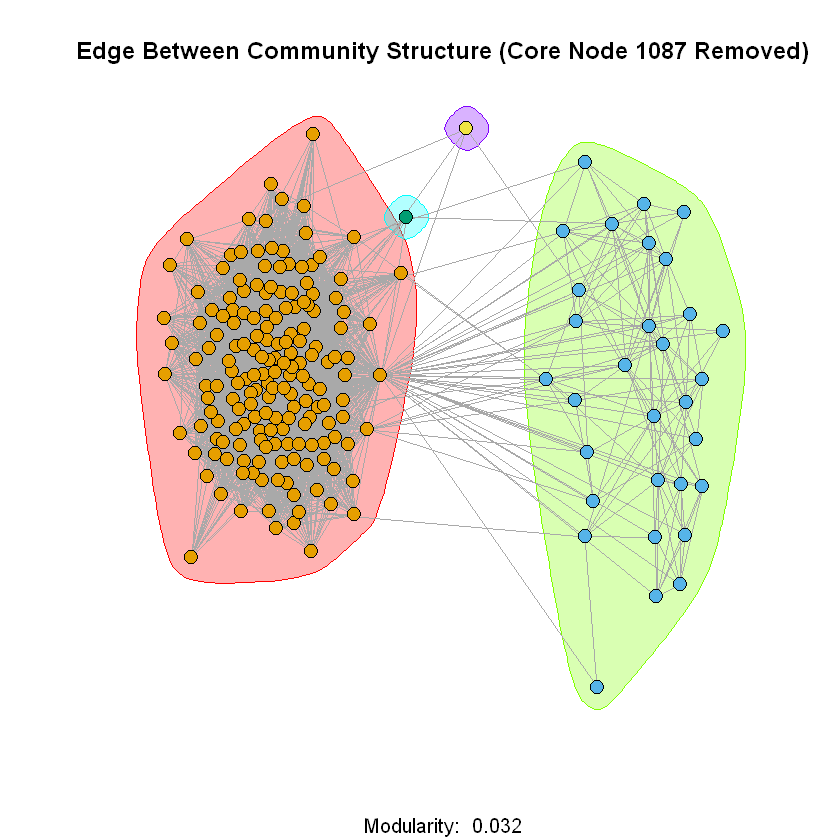
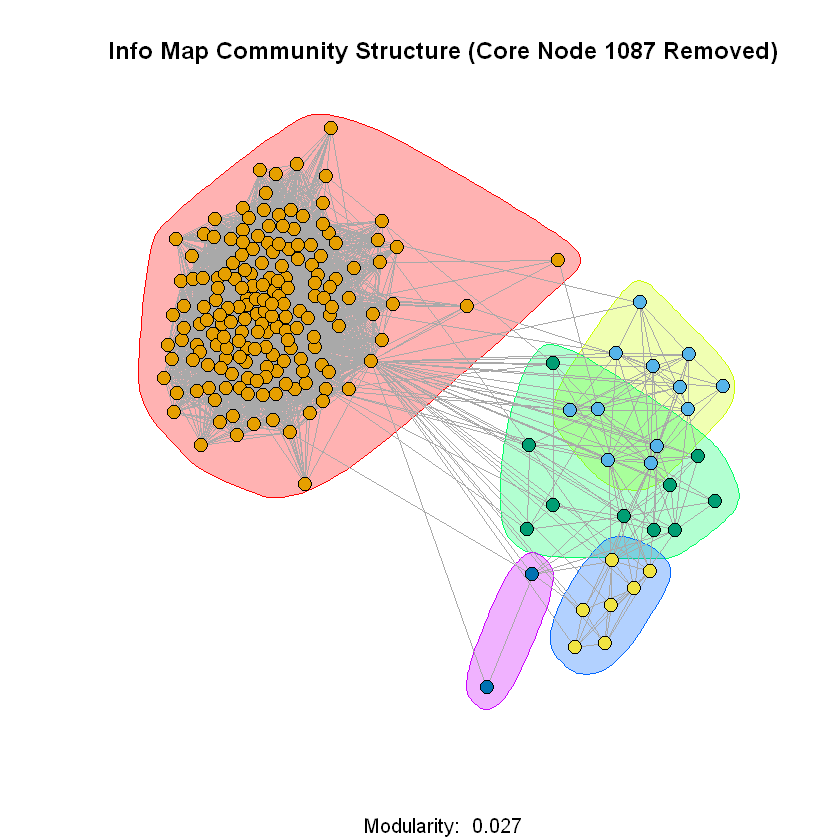
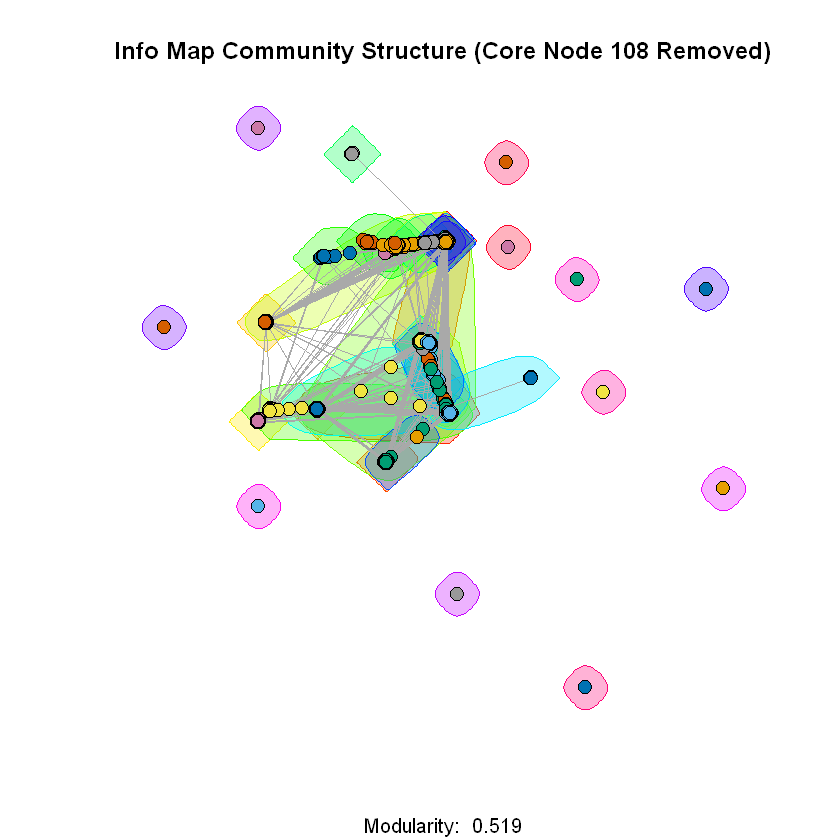
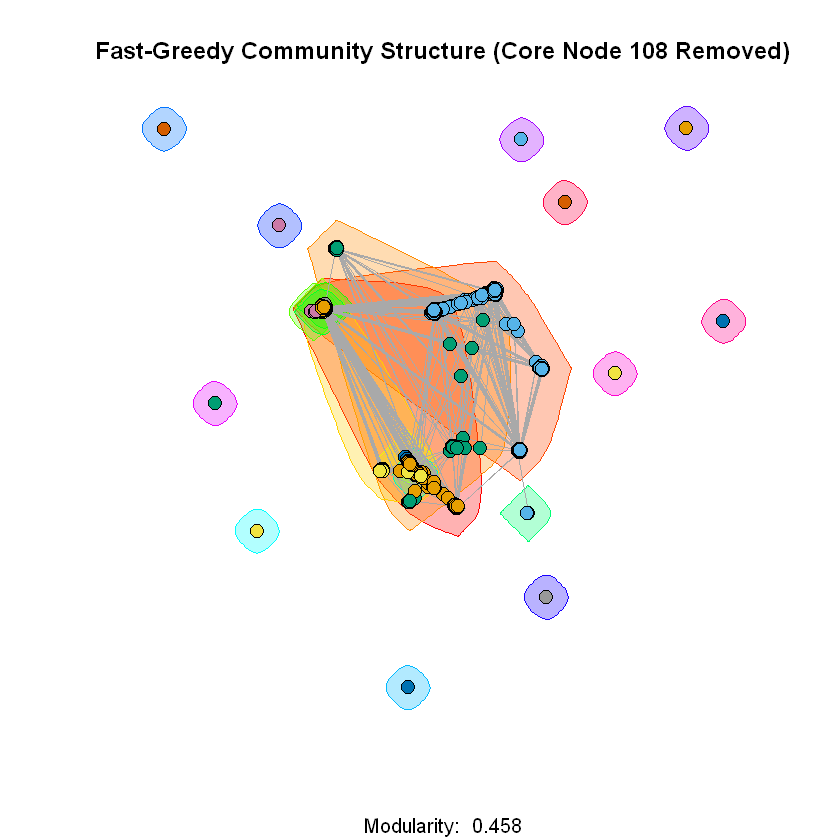
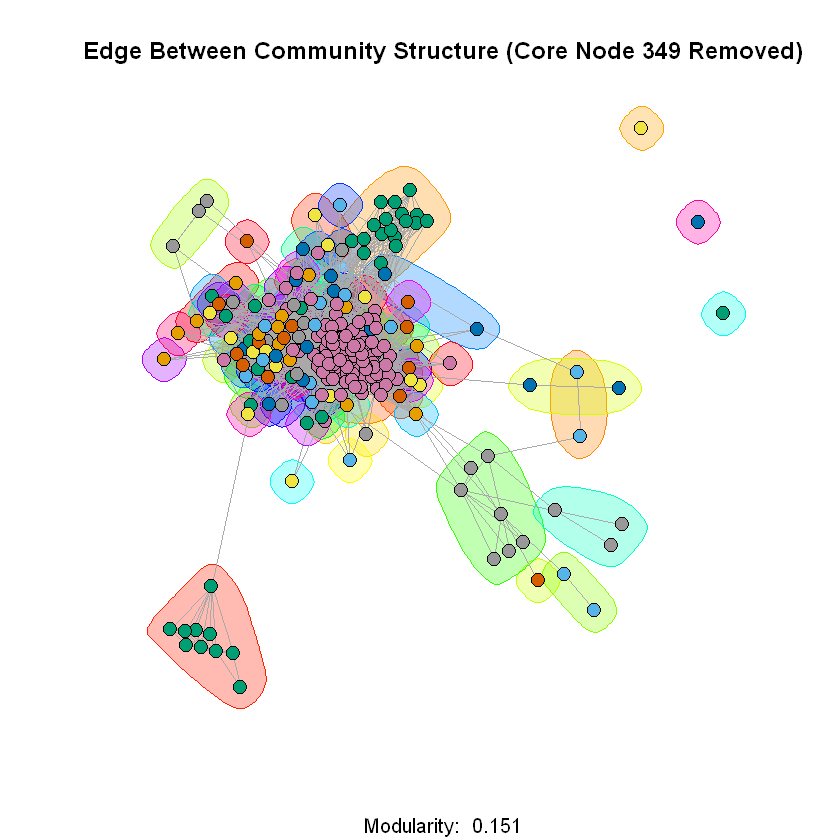
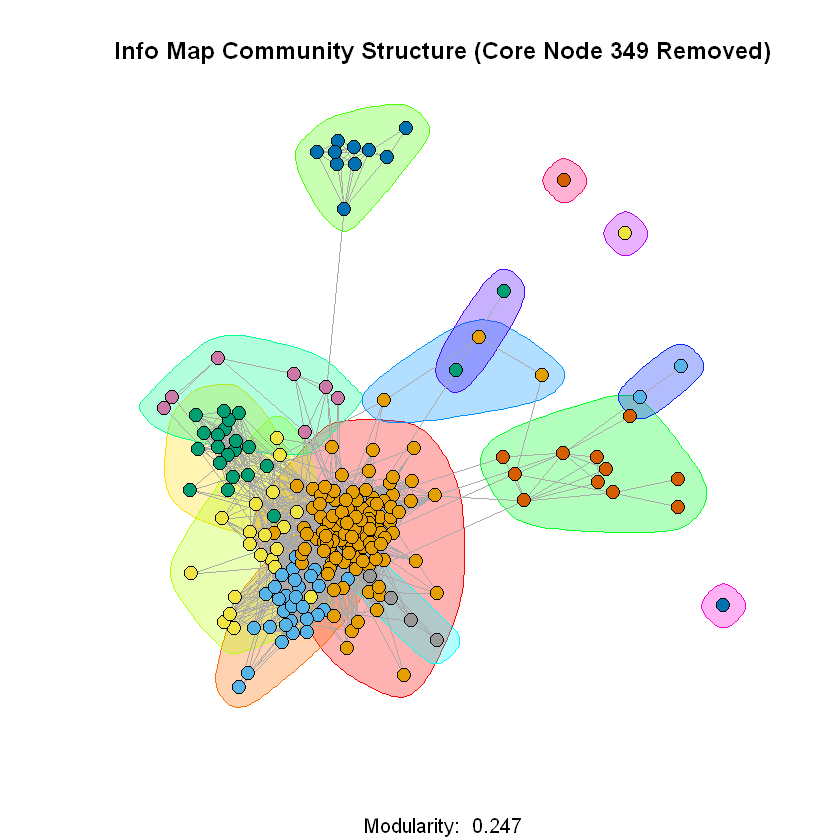
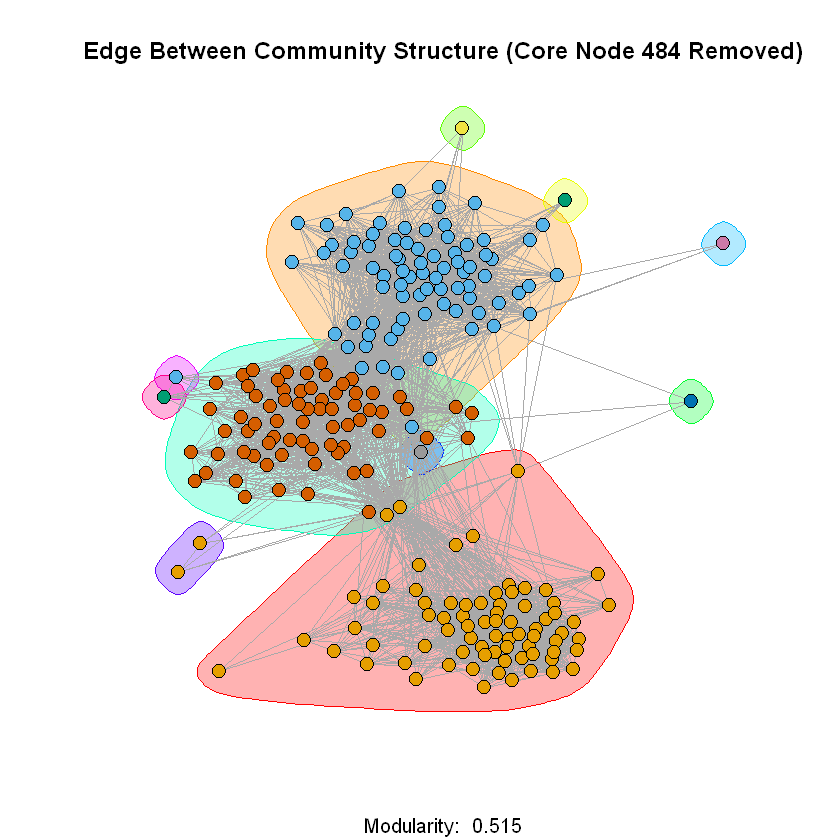
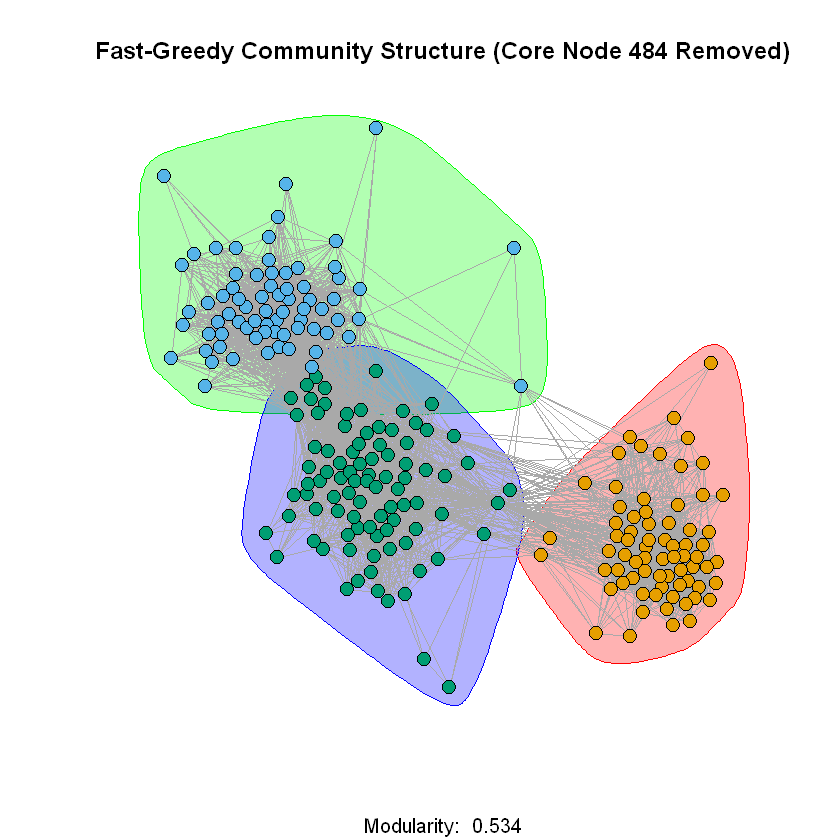
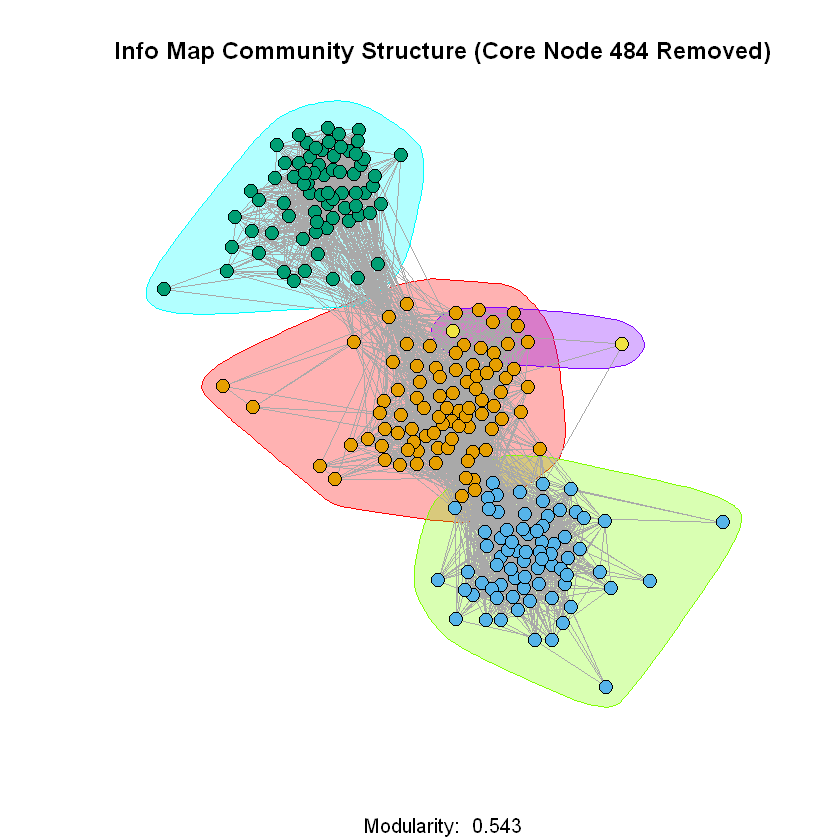
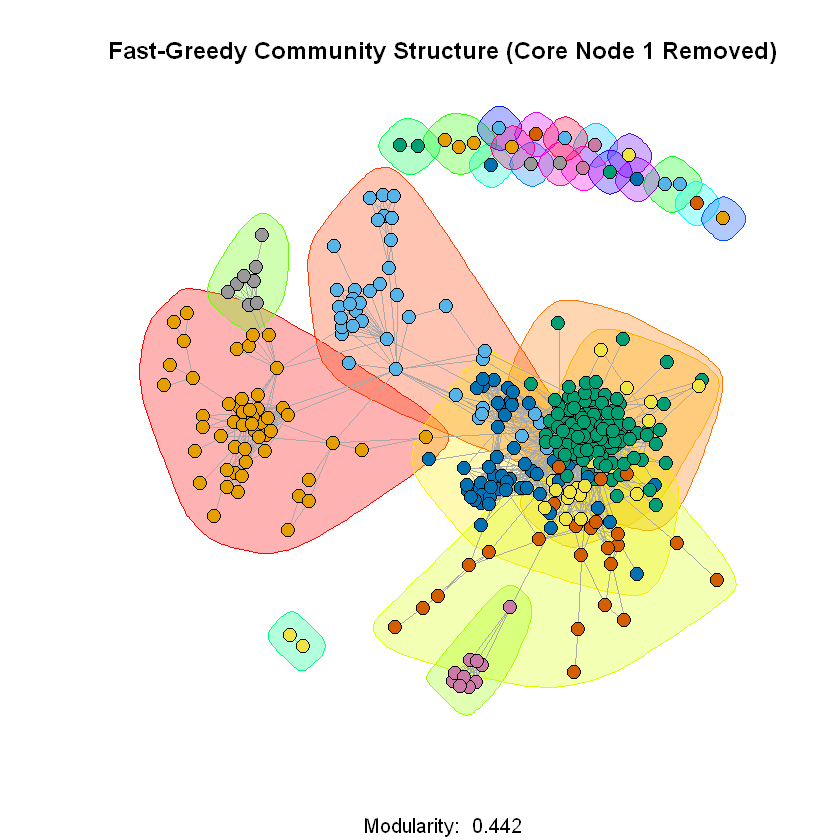
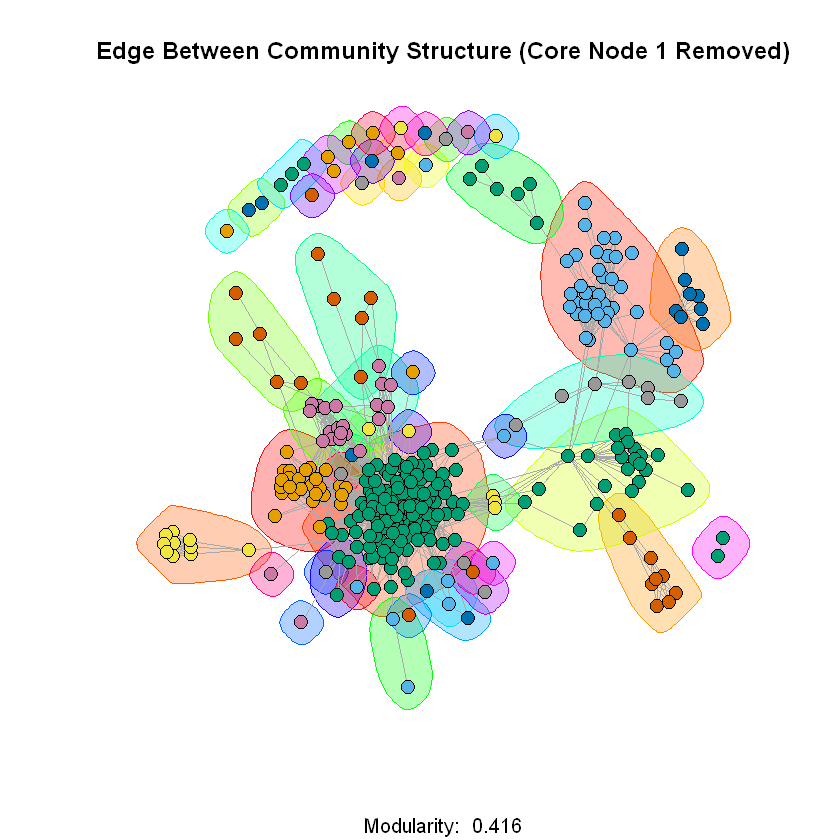
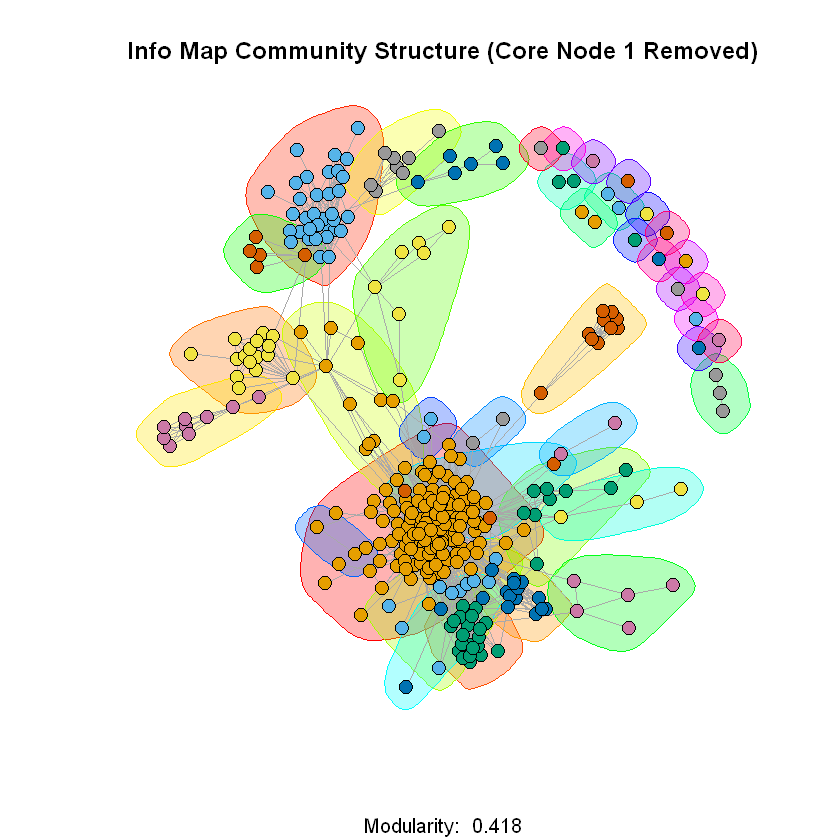
|  |  |  |  |
| --- | --- | --- | --- |
| Node ID | Modularity Fast-Greedy | Modularity Edge-Betweenness | Modularity Infomap |
| 1 | 0.413 | 0.353 | 0.389 |
| 108 | 0.436 | 0.507 | 0.508 |
| 349 | 0.252 | 0.134 | 0.095 |
| 484 | 0.507 | 0.489 | 0.515 |
| 1087 | 0.146 | 0.028 | 0.027 |
| **Avg** | 0.351 | 0.302 | 0.307 |



**Question 10**

From the following table we see that the modularity scores consistently increase slightly for the personalized networks without the core node. This makes sense as the core node was connected to every other node in the personalized network so by removing it, there are less connections between communities. Moreover, we can also see that the info map algorithm sees that largest increase in modularity without the core node, especially for the personalized network of node 349.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Modularity** | | | | | |
| Node ID | Fast-Greedy  With Core | Fast-Greedy  Without Core | Edge-Betweenness  With Core | Edge-Betweenness  Without Core | Infomap  With Core | Infomap  Without Core |
| 1 | 0.413 | 0.442 | 0.353 | 0.416 | 0.389 | 0.418 |
| 108 | 0.436 | 0.458 | 0.507 | 0.521 | 0.508 | 0.519 |
| 349 | 0.252 | 0.246 | 0.134 | 0.151 | 0.095 | 0.247 |
| 484 | 0.507 | 0.534 | 0.489 | 0.515 | 0.515 | 0.543 |
| 1087 | 0.146 | 0.148 | 0.028 | 0.032 | 0.027 | 0.027 |
| **Avg** | 0.351 | 0.366 | 0.302 | 0.327 | 0.307 | 0.351 |

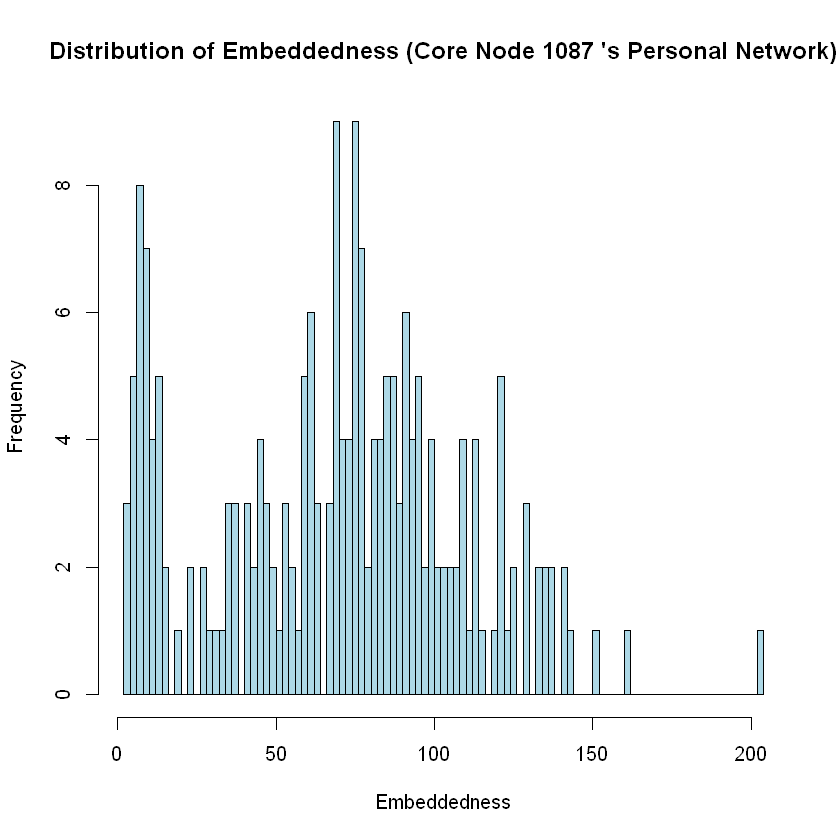
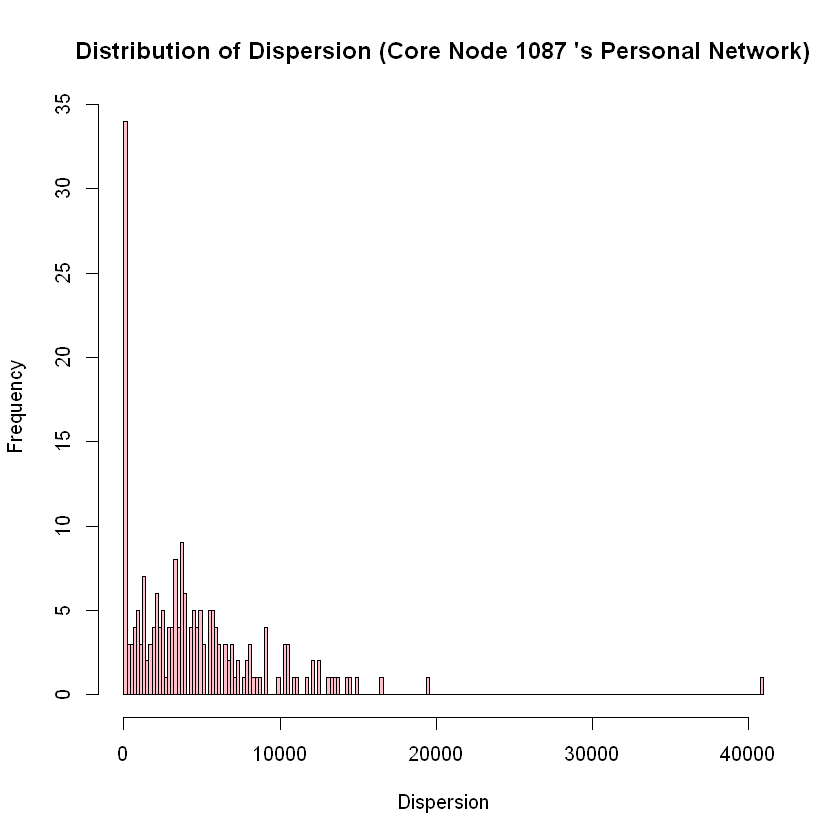
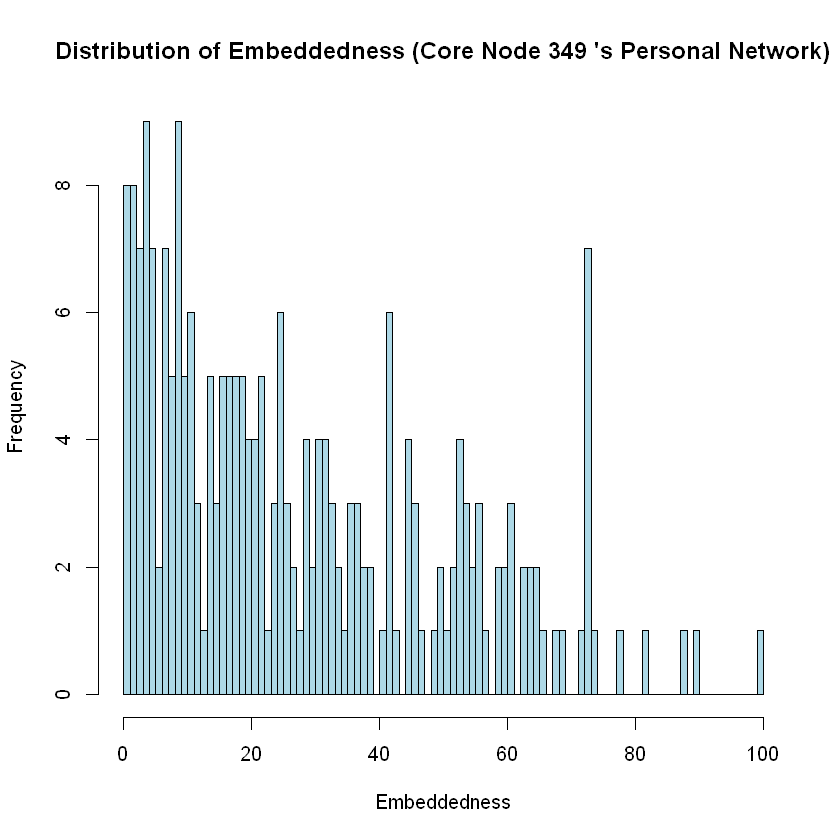
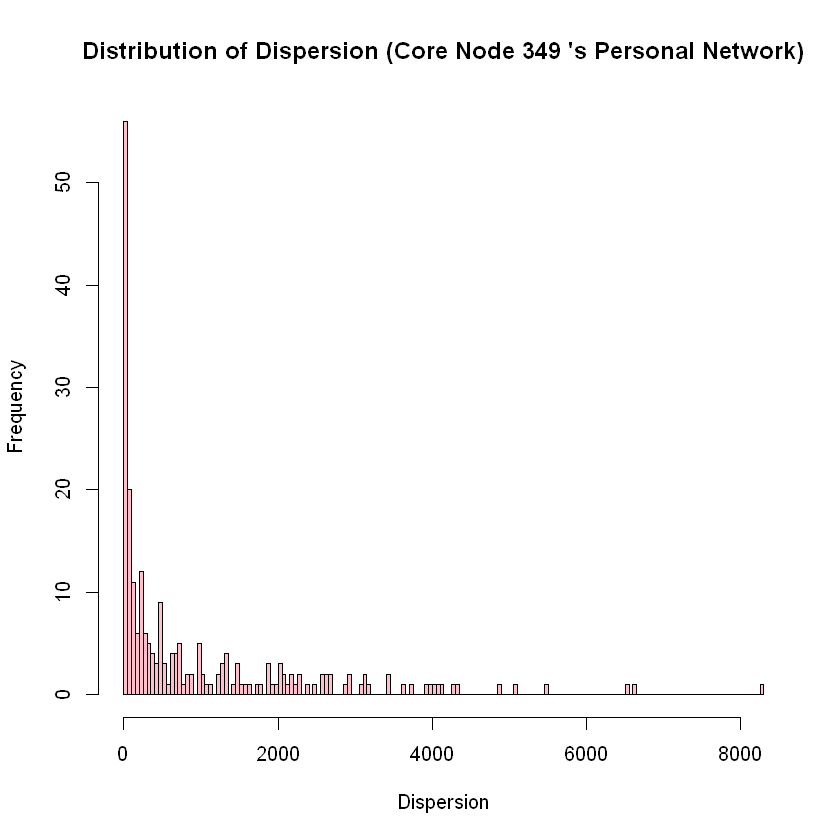
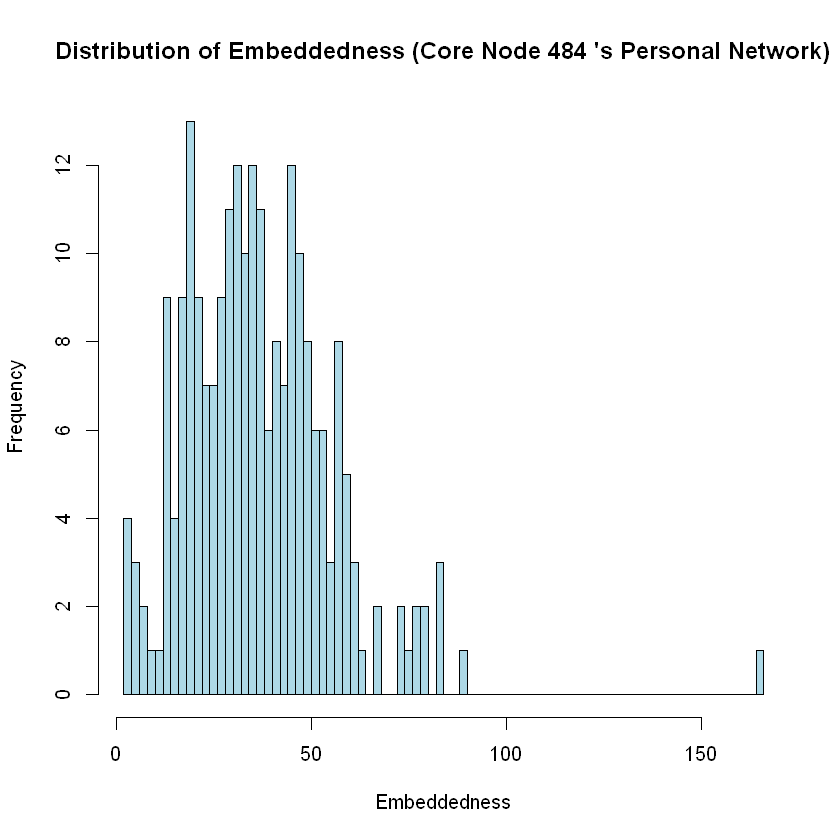
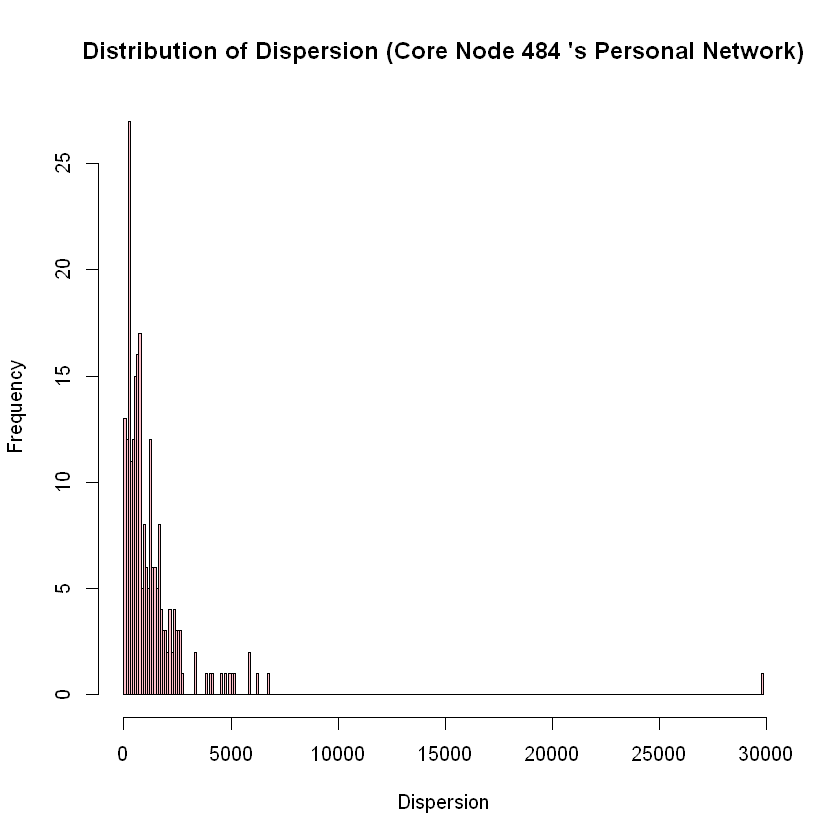
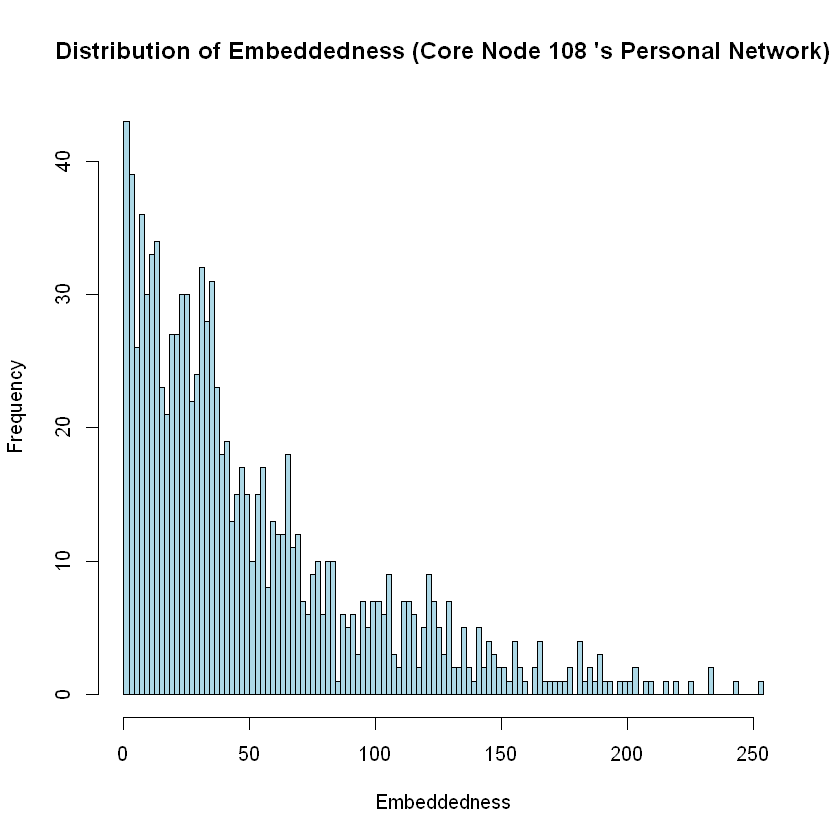
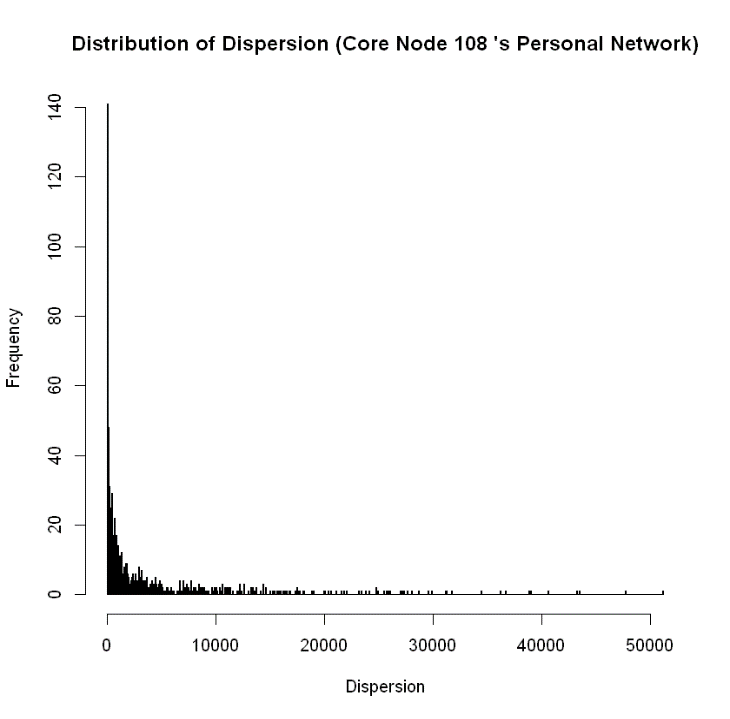
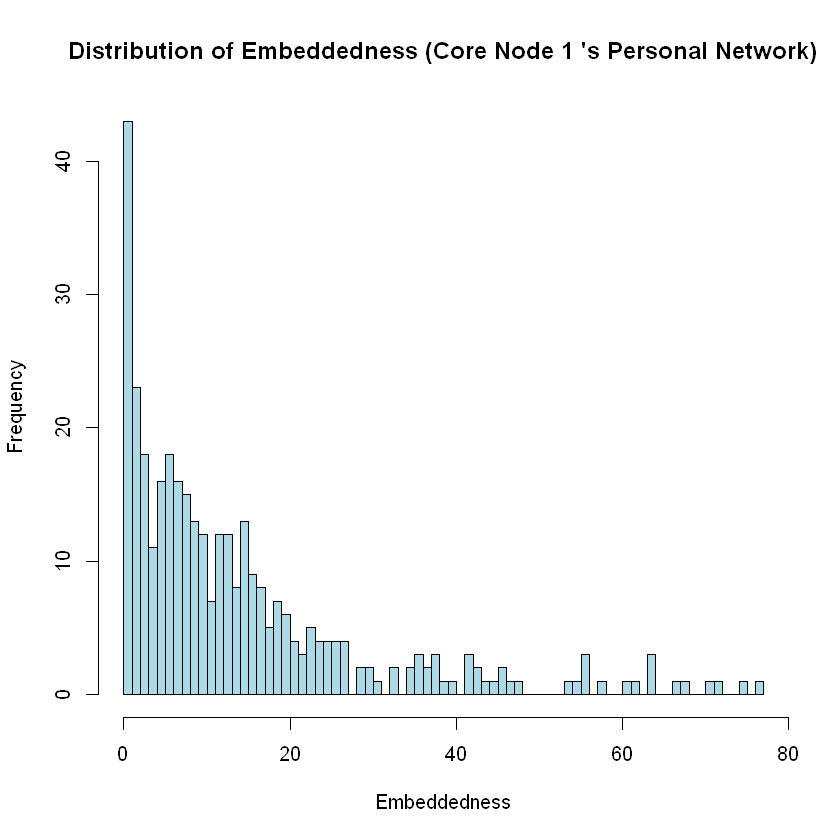
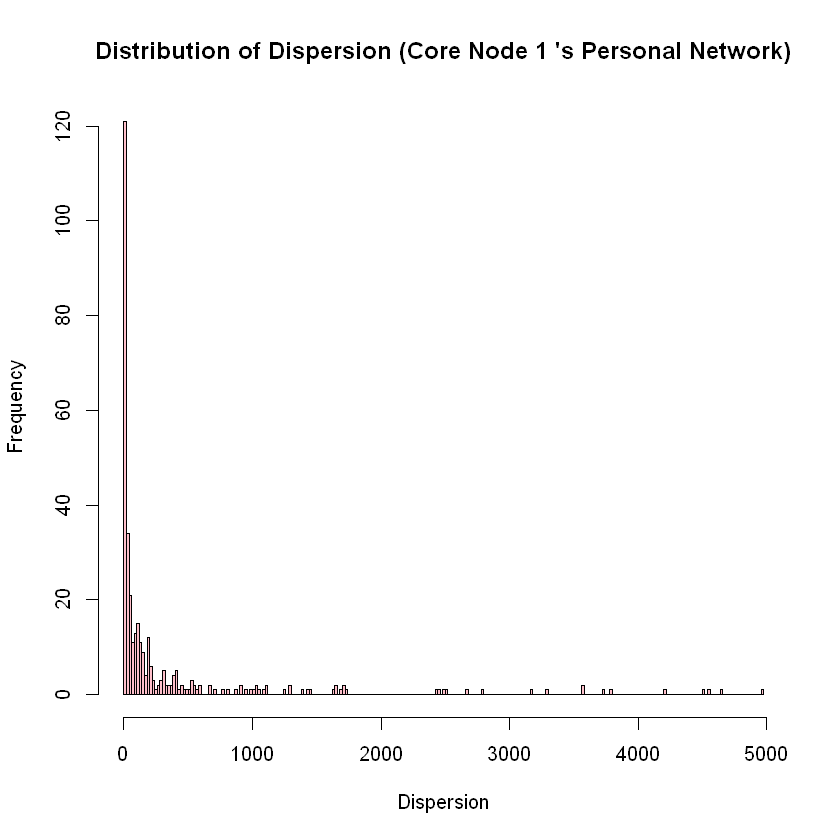


**Question 11**

Since every node, a given node is friends with, the core node is also friends with (minus the core node itself), the number of mutual friends a node shares with the core node or embeddedness is defined as

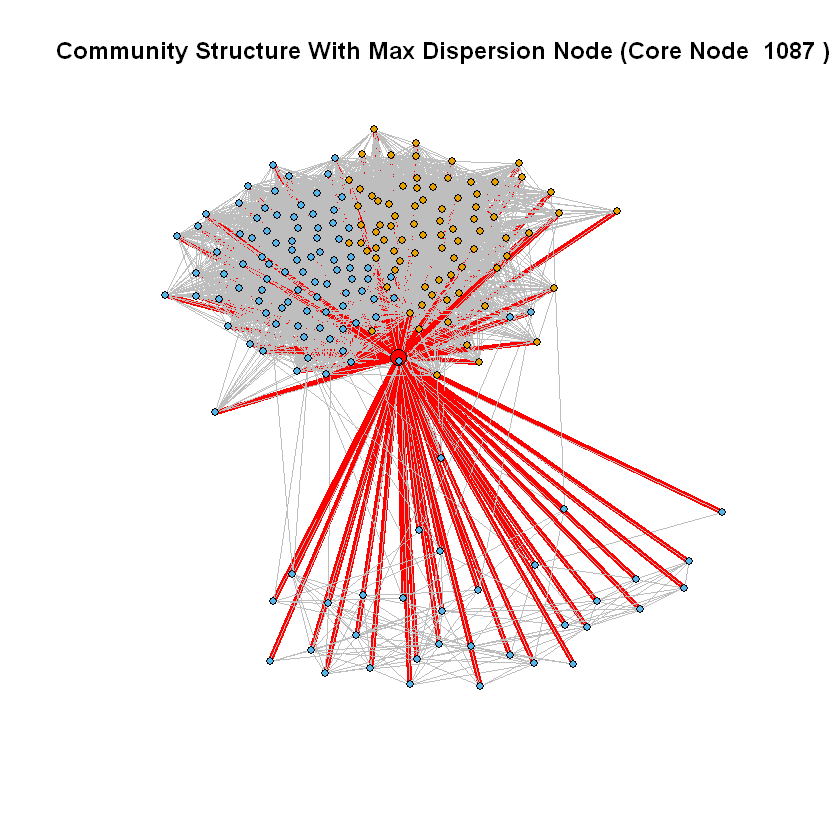
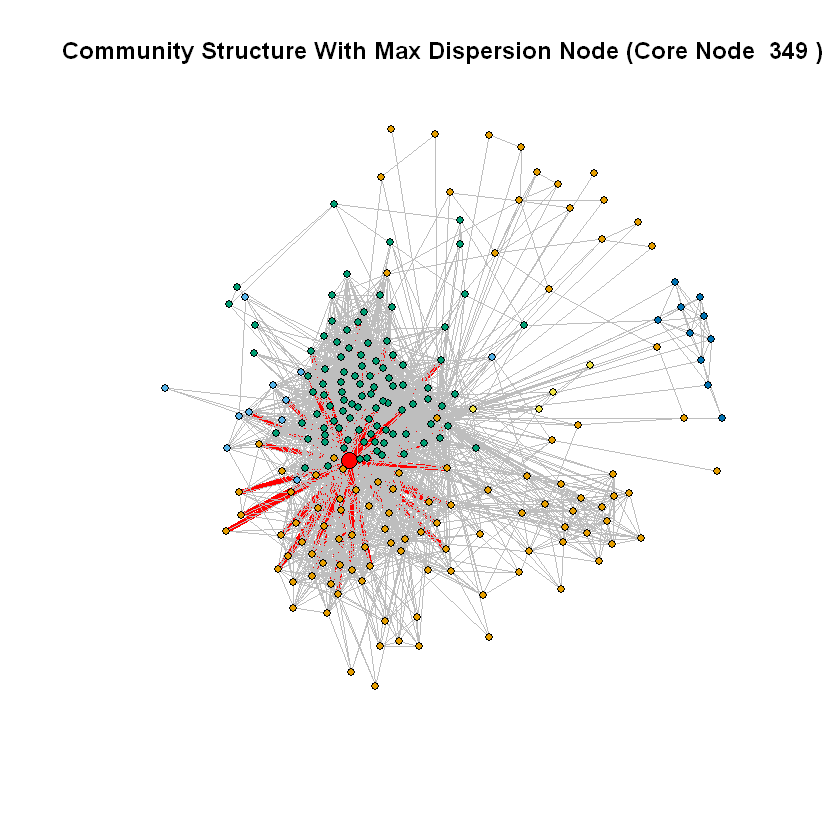
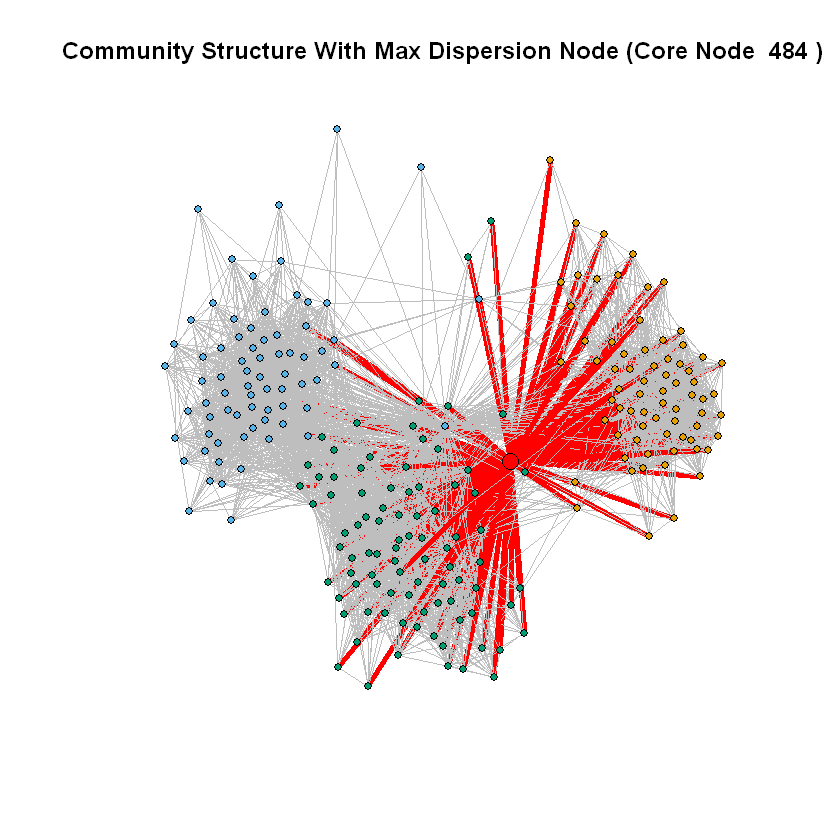
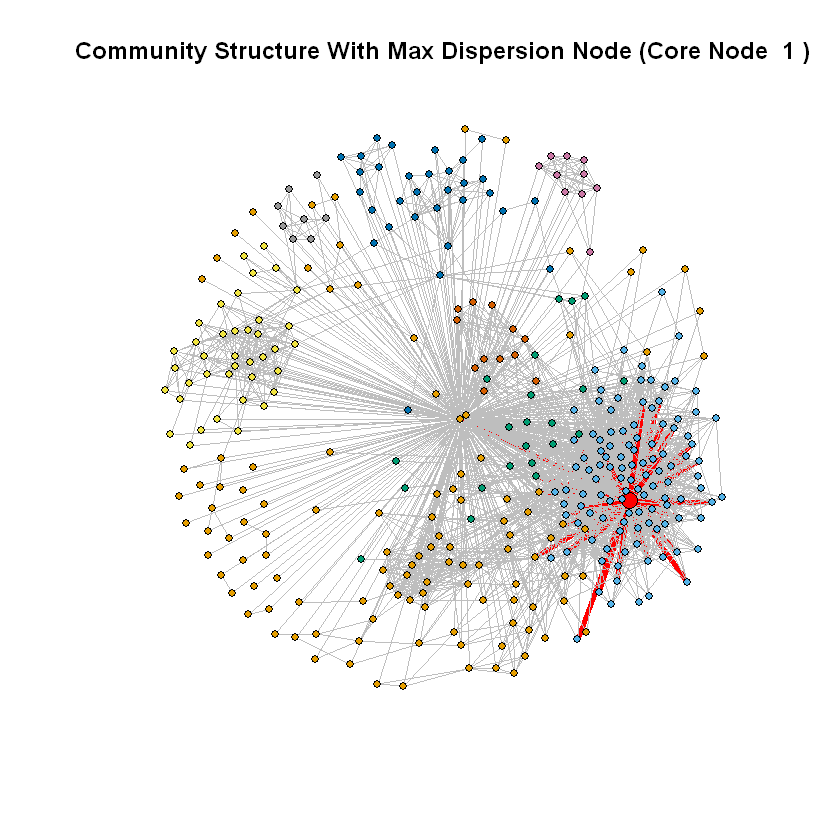
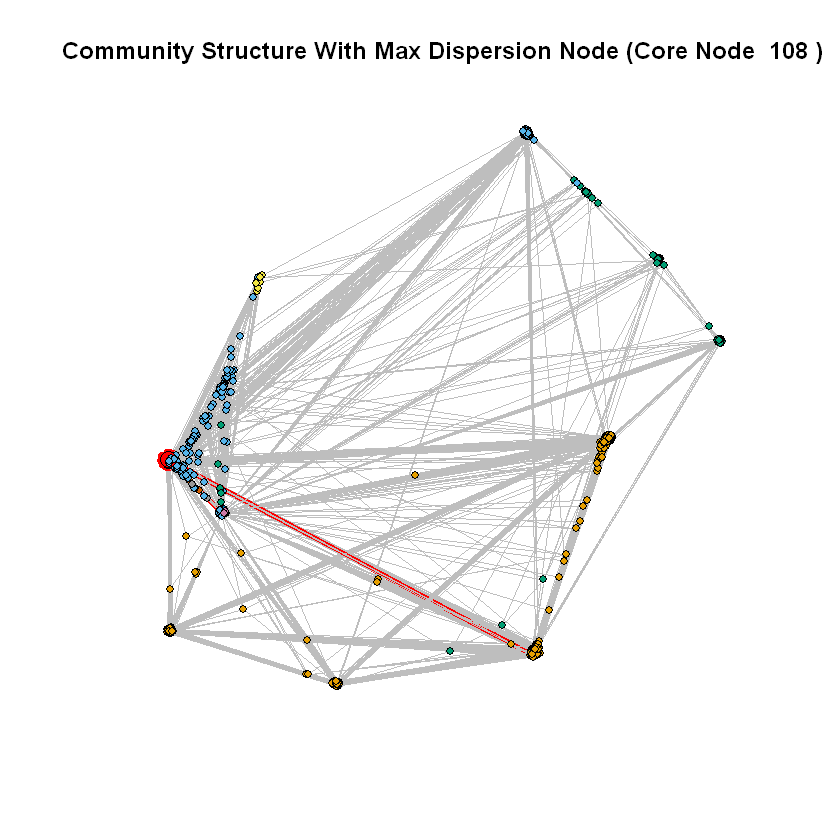
**Question 12**

The following shows the plots for the distribution of embeddedness and distribution of dispersion for each of the core nodes in question 9. The distance function used for dispersion was the shortest path distance between two nodes with the distance being the max distance plus 1 for nodes who are not connected once we remove from the personalized network.

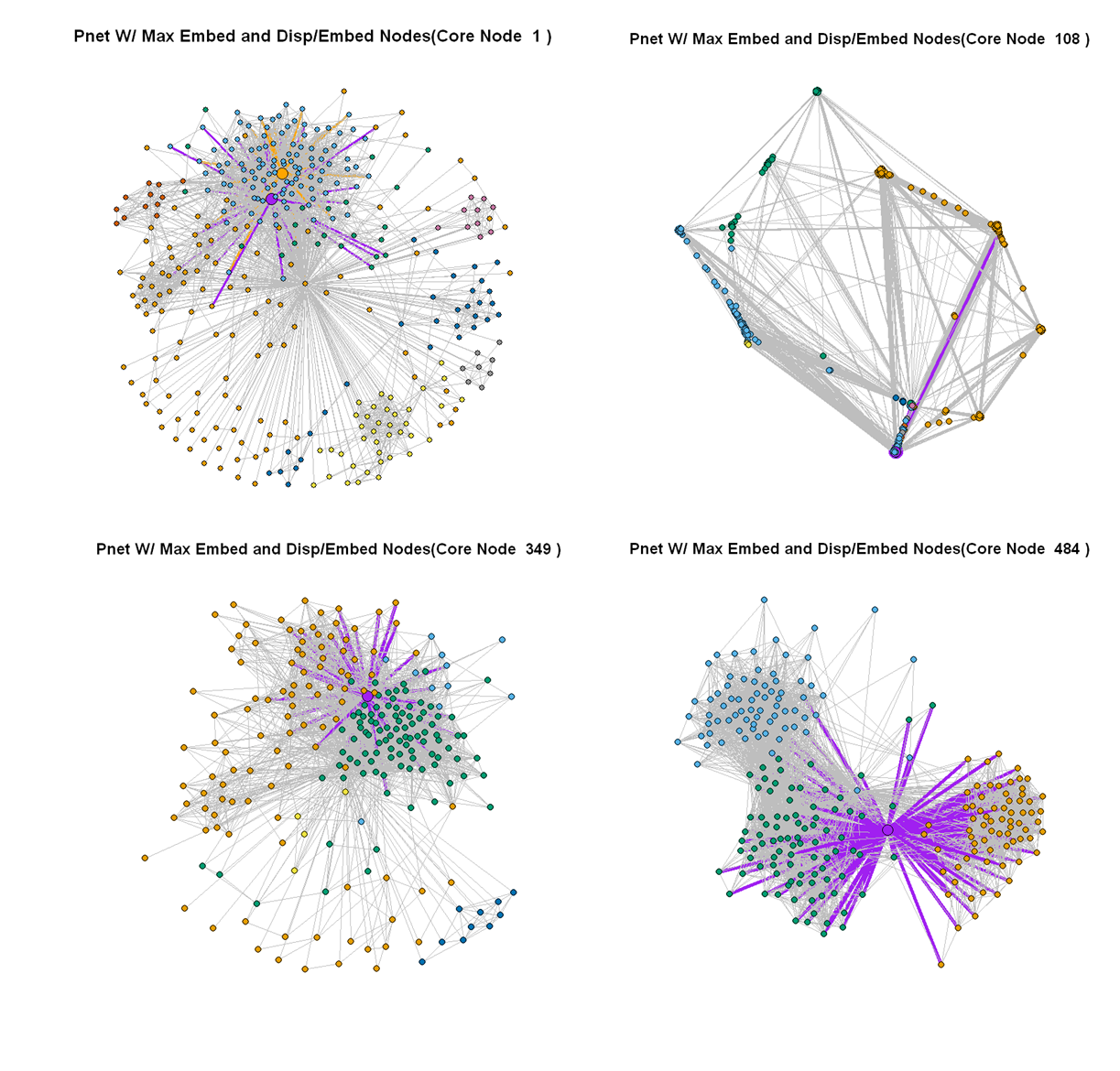


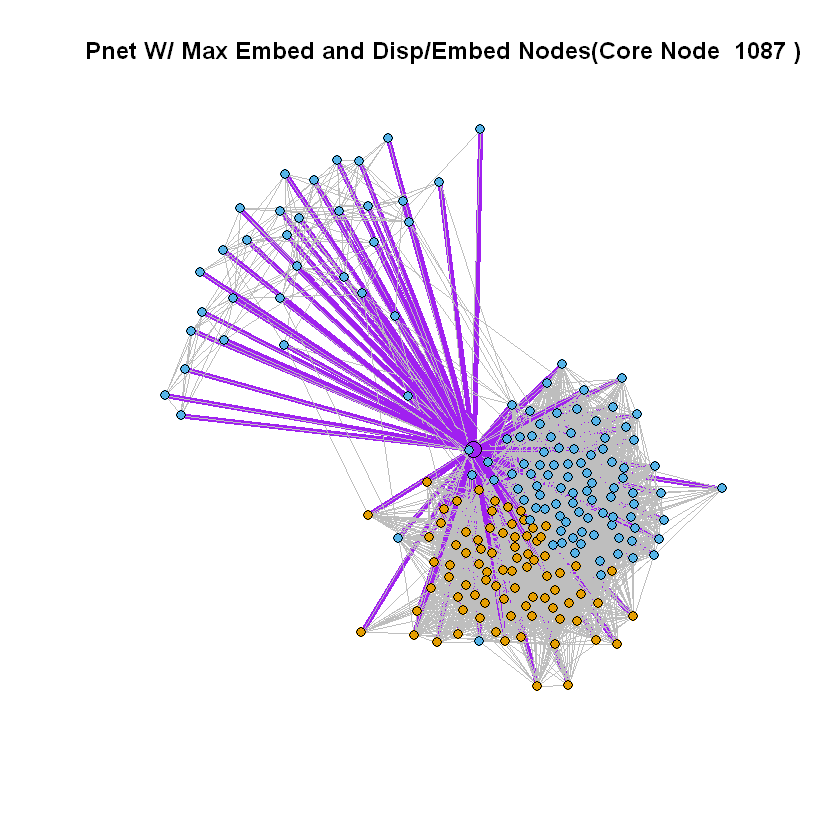
**Question 13**

For this question the node with max dispersion is enlarged and colored red. All edges incident to this node are also colored red.



**Question 14**

For this question the max embeddedness node of each personalized network is colored orange with nodes incident to it also colored orange. For the max dispersion/embeddedness node of each personalized network, it is colored purple with incident edges also colored purple. If only the purple node is seen that means that that node is both the node with max embeddedness and max dispersion/embeddedness.



**Question 15**

From the plots from question 13 we can see that the nodes with maximum dispersion indicate nodes connected to the core node which are connected to many nodes from separate communities. This is more evidently seen in the personalized network of node 349 in which the max dispersion node is connected to nodes colored as green, orange, and blue. Moreover, max dispersion nodes also have the characteristic of being the gateway for nodes of different communities such as the personalized network of node 484 where we have three distinct communities and the max dispersion node, like the core node acts as a connector between these communities. In addition, nodes from the same community who are connected to the max dispersion node are likely to be not very connected with each other. This is seen in core node 349’s personal network. In the context of Facebook friend’s max dispersion nodes can be thought of as friend of mine who is mutual friends with many of my friends whom I met through many different communities (maybe someone who has had a shared life experience as me). Since we share so many experiences being part of different communities, dispersion may be a good indicator of strong social ties.

Similarly, using plots from question 14, nodes with maximum embeddedness have characteristics of being connected with a lot of the nodes in the personal network. By the definition achieved in question 11 max dispersion should be the node with the second highest degree (first is core node) in the personalized network. However as seen in the personalized network for node 1, the max embeddedness node may share a bunch of mutual node connections with the core node with nodes from the same community. Thus, in the context of Facebook friends, this can be someone I met in college who happens to be in the same club I am in. Since I get more Facebook friends in college as opposed to middle school, a close middle school friend would not have high embeddedness. Thus, this may not be the strongest indicator of social ties. Nevertheless, compared with the dispersion/embeddedness measure, 4 of the 5 plots revealed that the max embeddedness node and the max dispersion/embeddedness node was the same node.

Finally, the max dispersion/embeddedness behaves for the most part like embeddedness. From the paper linked in the problem spec, it is specified that this metric is a normalized dispersion in which we try to find nodes in a personalized network who are connected to different communities but does not share a super large amount of mutual friends with the core node as this naturally leads to an increase in embeddedness as well (more pairs of nodes to sum together non-zero distance). In the cases of our plots we can see that even with this normalizing factor, the results of dispersion/embeddedness are like embeddedness for our personal networks that stem from human relationships.

Essentially all these three measures as shown by the plots in question 13 and 14 try to show nodes who are friends with the core nodes’ friends but are not more part of a particular community within the core node’s personal network than the core node itself. In the context of Facebook friends, this is like a friend who shares many friends and communities as me but is not more involved with a specific community (friends with everyone in that community) than with me.

**4. Friend Recommendation**

**Question 16**

is the number of nodes with degree 24 in node 415’s personal network.

**Question 17**

From the table below we see that all algorithms perform reasonably well in terms of average accuracy. Based on the best average accuracy, the common neighbors algorithm would be the best.

|  |  |
| --- | --- |
| **Algorithm** | **Average Accuracy** |
| Common Neighbors Measure | 0.837067 |
| Jaccard Measure | 0.807071 |
| Adamic Adar Measure | 0.826358 |

## 

## 

## 

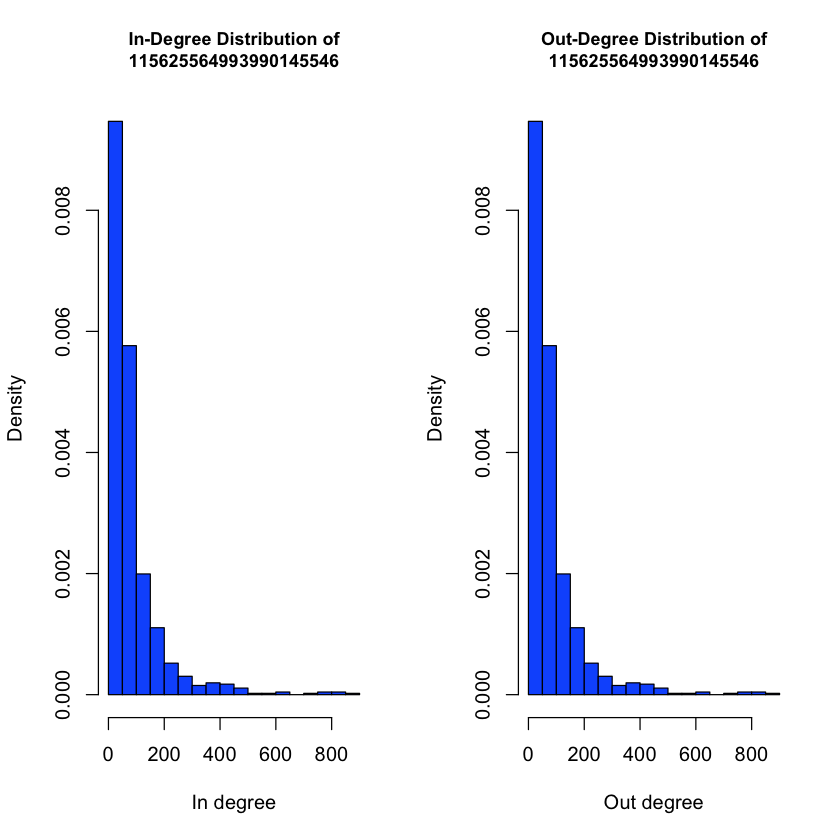
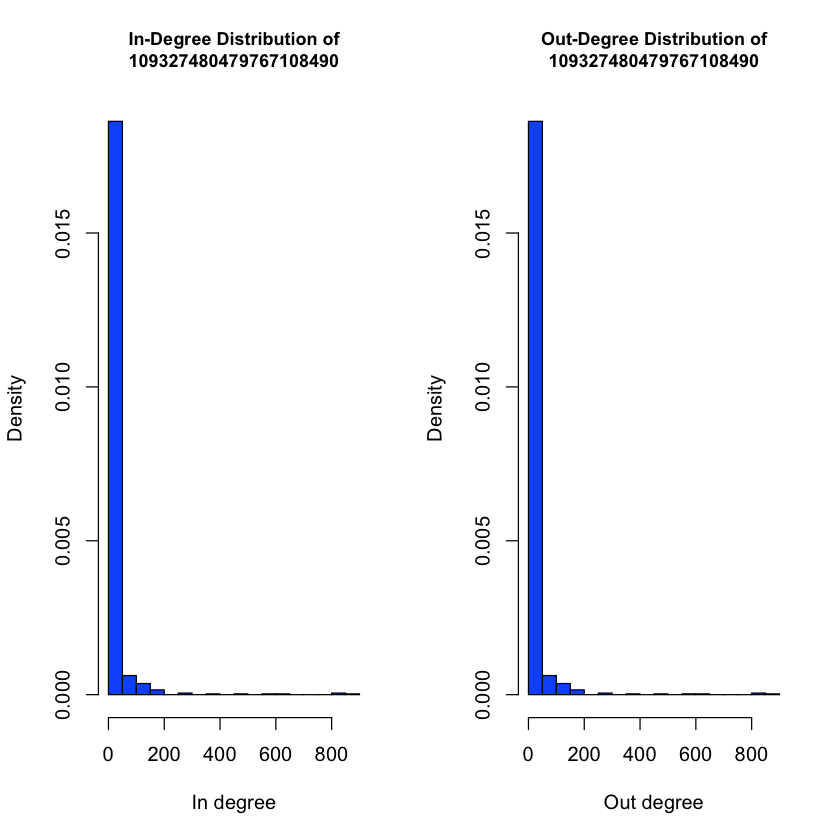
# 

# Part 2: Google+ Network

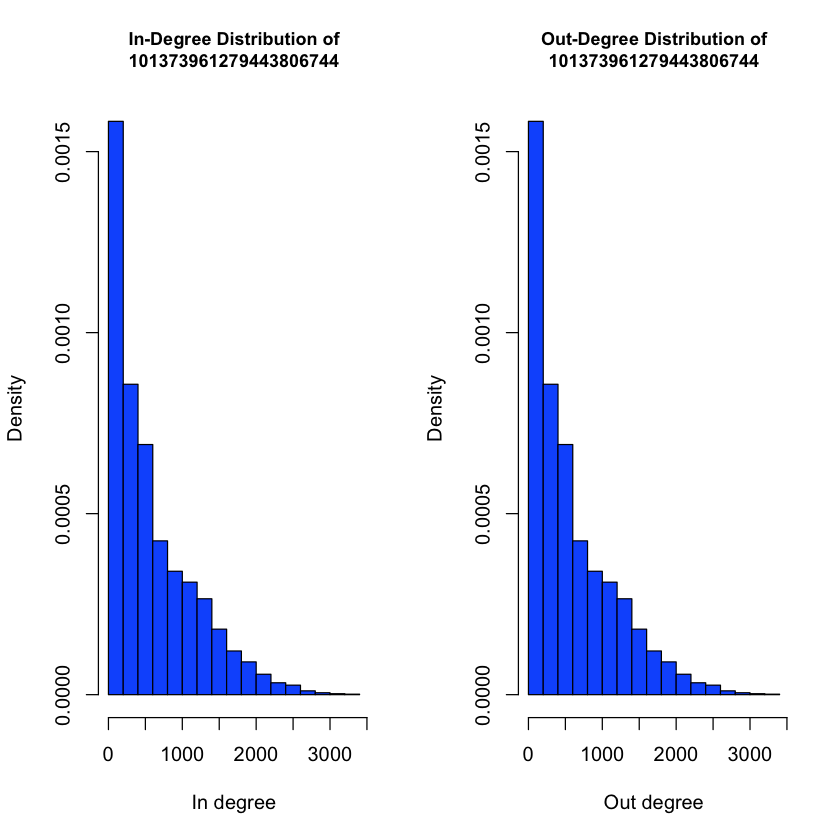
### Question 18

Total number of personalized network is 132, and the total number of personalized network with more than 2 circles are 57.

### Question 19



In-degree & Out-degree distribution of node\_1 In-degree & Out-degree distribution of node\_2

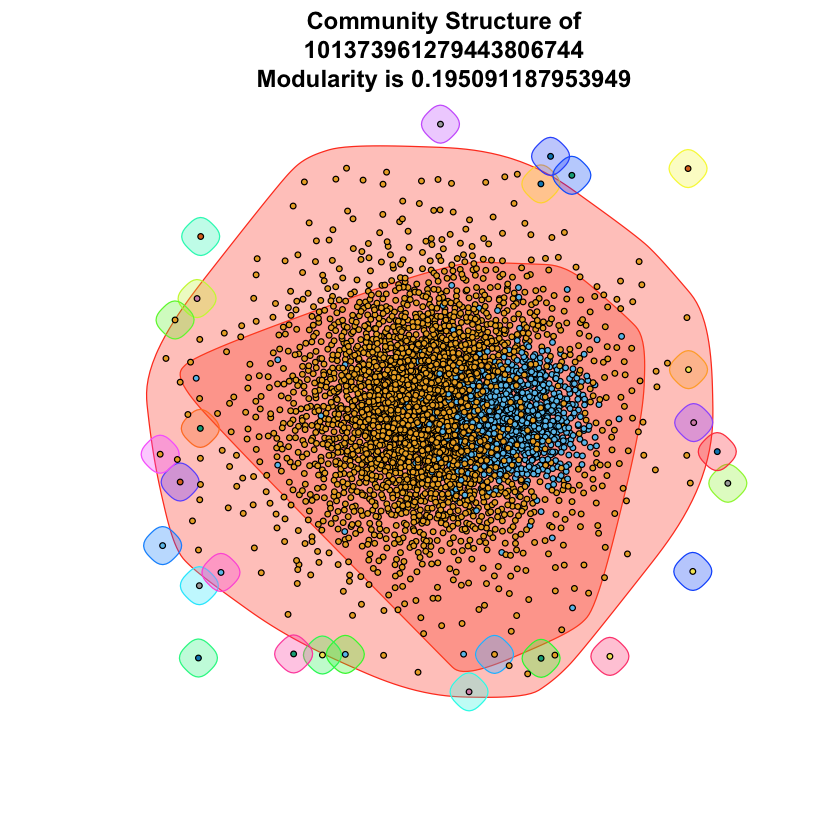
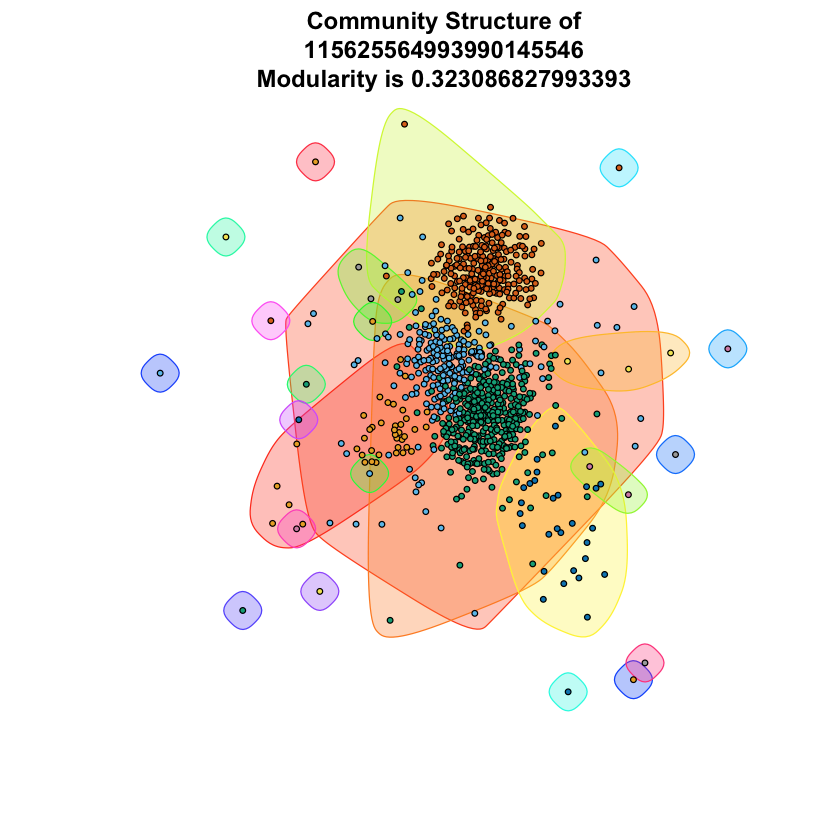
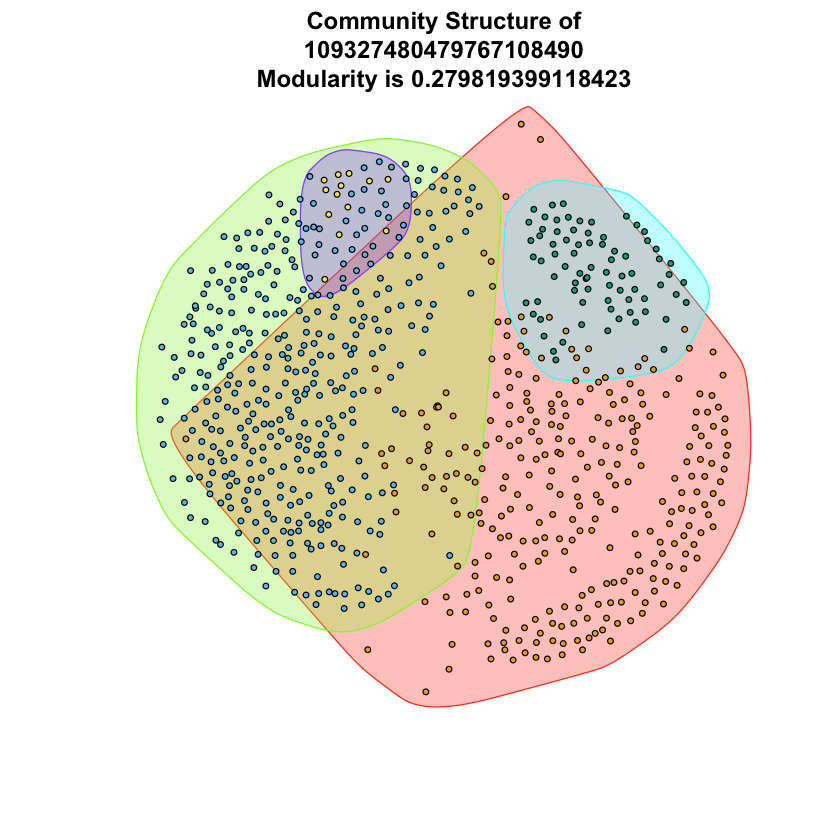


In-degree & Out-degree distribution of node\_3

For all the networks, both the in-degree and out-degree distributions are similar.

Node\_1 network and node\_2 network have similar degree distribution, most of their nodes have low degree. Node\_3 network has more distribution in higher degrees, meaning that node\_3 network has more nodes and might have more hubs, which leads to the sparser of the node\_3 network. The modularity in the next question could explain this further.

### Question 20



Community Structure with Walktrap Method of 3 Personal Networks

Modularity of 3 Personal Networks

|  |  |  |  |
| --- | --- | --- | --- |
|  | node\_1 | node\_2 | node\_3 |
| Node ID | 109327480479767108490 | 115625564993990145546 | 101373961279443806744 |
| Modularity | 0.279819399118423 | 0.323086827993393 | 0.195091187953949 |

The modularity of node\_3 is the smallest, and the modularity of node\_2 is the largest. Modularity measures the density of the connections between the nodes within communities. Based on this definition, node\_3 having the smallest modularity means that node\_3 has sparse connections to the other nodes within the community, weaker tie with friends within the network, which corresponds to the last question, where node\_3 network have sparser degree distribution.

On the other hand node\_2 having the largest modularity means that node\_2 has dense connections to the other nodes within the community, stronger tie with friends within the network.

### Question 21

Based on the definitions, homogeneity h means the purity of the circle within the community when compared to that in the whole graph, and completeness c means the purity of the community within the circle when compared to that in the whole graph.

That is, h means the possibility of the nodes within the same community also lie in the same circle, when compared to that in the whole graph context. And c means the possibility of the nodes within the same circle also lie in the same community, when compared to that in the whole graph context

Entropy measures the extent of disorder, so H(C) means the diversity of circle in the whole graph, H(K) means the diversity of community in the whole graph, H(C|K) means the diversity of circle within the community, H(K|C) means the diversity of community within the circle. Then H(C|K)/H(C) means the extent of diversity of circle within the community when compared to that in the whole graph, and H(K|C)/H(K) means the extent of diversity of community within the circle when compared to that in the whole graph. The expressions of h and c are h = 1 - H(C|K)/H(C), c = 1 - H(K|C)/H(K), so h means the purity of the circle within the community when compared to that in the whole graph, and c means the purity of the community within the circle when compared to that in the whole graph.

### Question 22

Entropy, Homogeneity, Completeness of 3 Personal Networks

|  |  |  |  |
| --- | --- | --- | --- |
|  | node\_1 | node\_2 | node\_3 |
| Node ID | 109327480479767108490 | 115625564993990145546 | 101373961279443806744 |
| N | 764 | 727 | 521 |
| H(C) | 1.05077934757594 | 8.46514668159249 | 0.384319958960942 |
| H(K) | 1.00699108804102 | 1.05715052927858 | 0.0659438893838449 |
| H(C|K) | 0.142901677053362 | 4.71555621215019 | 0.00304361110521456 |
| H(K|C) | 0.658882001908203 | 4.62579231827161 | 0.175898307700917 |
| **h** | **0.864004105730642** | **0.442944536046352** | **0.992080528127024** |
| **c** | **0.345692320683815** | **-3.37571773381067** | **-1.66739358785848** |

The h and c values of the three nodes show that for the h values, node\_3 has the highest h value and node\_2 has the lowest h value. For the c values, node\_1 has the highest c value and node\_2 has the lowest c value, and both node\_2 and node\_3 have negative c value.

1. For the h values, node\_3 has the highest h value and node\_2 has the lowest h value. This means that the nodes in the node\_3 network have higher possibility to be in the same circle. And the nodes in the node\_2 network have higher possibility to be in different circles.
2. These h values make sense, we could look back at the modularity at Q20, node\_2 network has the largest modularity and node\_3 network has the smallest modularity. Since node\_2 network is denser than the node\_3 network, each community in the node\_2 network would have more nodes, leading to higher possibility of the nodes within the same community are in different circles.
3. For the c values, node\_1 has the highest c value and node\_2 has the lowest c value. This means that the nodes in the node\_1 network have higher possibility to be in the same community. And the nodes in the node\_2 network have higher possibility to be in different communities.
4. All the h values are positive, which means that H(C|K) < H(C) for the networks of all the three nodes. This is because the circles are “larger” than the communities, there are more nodes in each circle on average than in the community, that is, in most of the cases a community is only a subset of the circle, so that is why the circle diversity within the community is less than that in the whole graph.
5. The c values of node\_2 and node\_3 are negative and that of node\_1 is positive. This means that for node\_2 and node\_3, H(K|C) > H(K). This case is possible, because as stated in (4), the circles are “larger” than the communities, so when the communities crowd in the same circle, then the community diversity of this circle is larger than that of the whole graph.