ECE232E: LARGE SCALE SOCIAL AND COMPLEX NETWORKS: DESIGN AND ALGORITHMS

Project 2: Social Network Mining

May 6, 2020

Wanli Gao, UID: 105431975
Yifan Zhang, UID: 805354474
Tianyi Zhao, UID: 804380974

1 Facebook Network

1.1 Structural Properties of the Facebook network

Question 1: We created the network from the facebook_combined.txt, the network we generated is shown below:

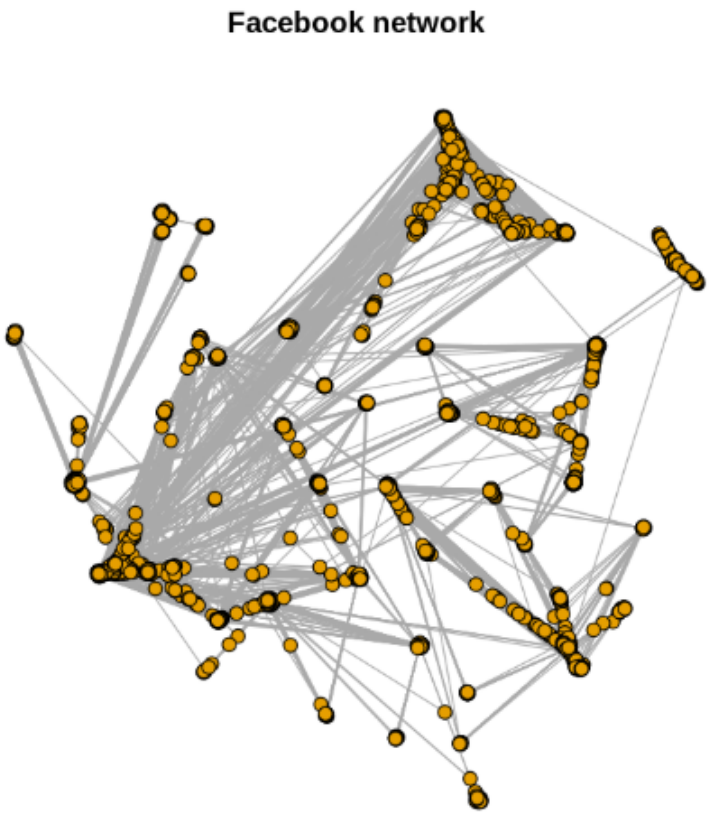

Figure 1.1. Facebook Network

Question 1.1: The number of nodes of the Facebook network is 4039, and the number of edges is 88234.
Question 1.2: The Facebook network is connected.
Question 2: The diameter of the Facebook network is 8.
Question 3: We plot the degree distribution of the Facebook network:

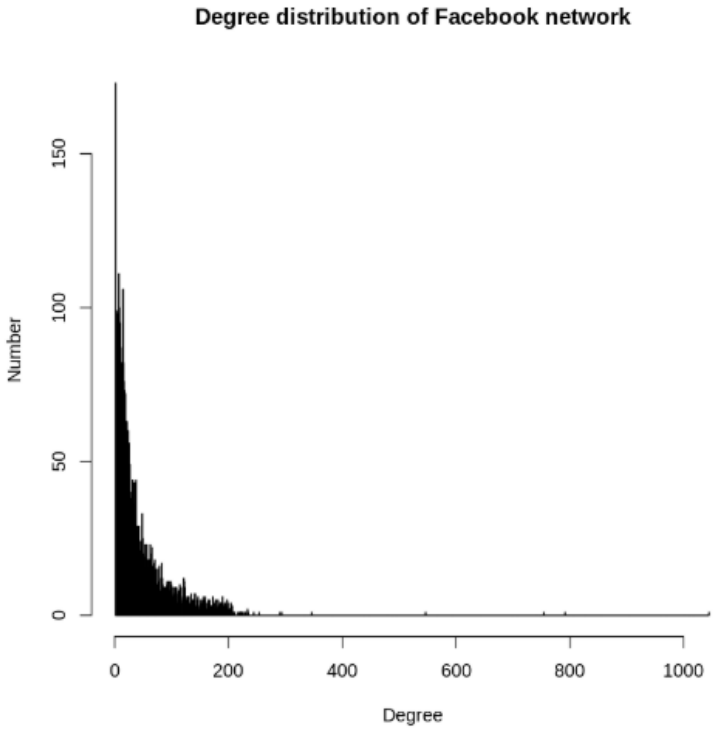

Figure 1.2. Degree distribution of the Facebook network (histogram)

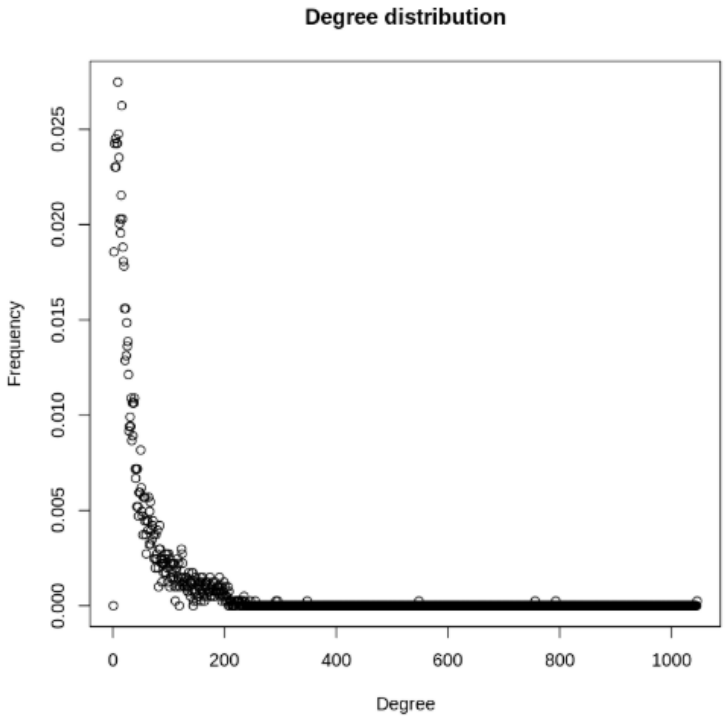

Figure 1.3. Degree distribution of the Facebook network

The average degree is 43.69101.

Question 4: We plot the degree distribution of the Facebook in log-log scale:

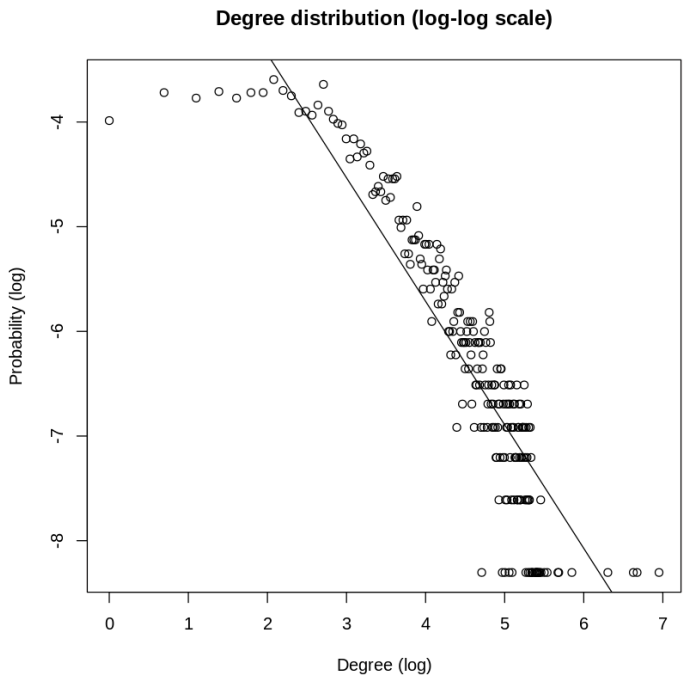

Figure 1.4. log-log scale degree distribution of the Facebook network

We fit a line to the plot, the slope of this line is -1.1802 in log scale.

1.2 Personalized network

We study some of the structural properties of the personalized network of the user whose graph node ID is 1.

Question 5: We create a personalized network of the user whose ID is 1. The personalized network has 348 nodes and 2866 edges.

Question 6: The diameter of the personalized network is 2. A trivial upper bound for the diameter of the personalized network is 2, the trivial lower bound for the diameter of the personalized network is 1.

Question 7: The rough meaning of diameter is the longest shortest path between two vertices. Then the upper bound for diameter should be 2 since the shortest path would be from a ”friend” to another, and only require the original person to be the third node, making the diameter 2. On the other hand, the lower bound for the diameter should be the diameter from the original user to a friend, which means the diameter should be 1 (if the original user has only 1 friend).

1.3 Core node’s personalized network

Question 8: There are 40 core nodes in the Facebook network. The average degree of the core nodes is 279.375.

1.3.1 Community structure of core node's personalized network

Question 9: We study the community structure of the personalized network of the following core nodes:

Node ID 1

Node ID 108

Node ID 349

Node ID 484

Node ID 1087

For each of the above core node's personalized network, we find the community structure using Fast-Greedy, Edge-Betweenness, and Infomap community detection algorithms:

```
Modularity of Fast-Greedy of Node ID = 1 is: 0.4131014
```

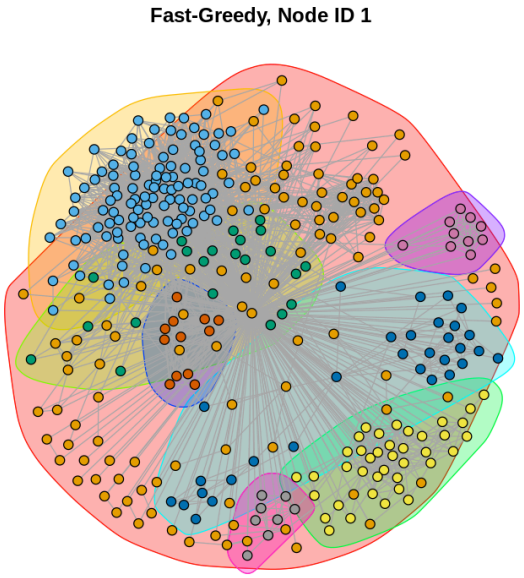

Figure 1.5. Community structure using Fast-Greedy for Node ID 1

Modularity of Edge-Betweenness of Node ID = 1 is: 0.3533022

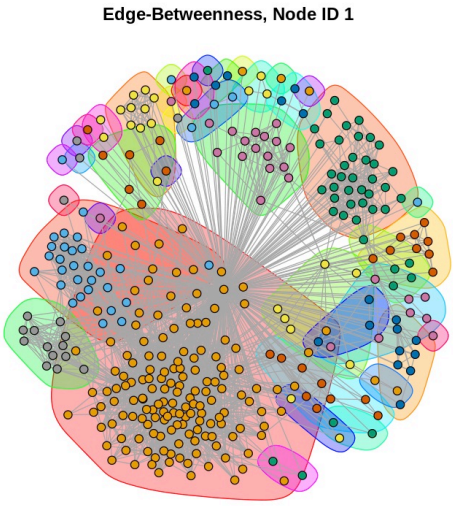

Figure 1.6. Community structure using Edge-Betweenness for Node ID 1

Modularity of Infomap of Node ID = 1 is: 0.3891185

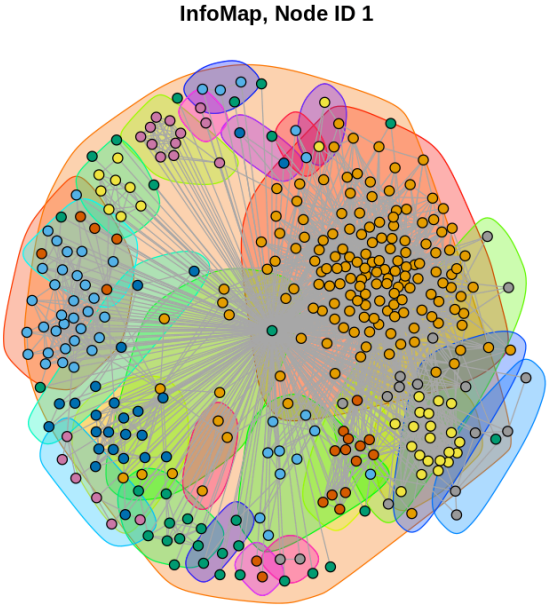

Figure 1.7. Community structure using Infomap for Node ID 1

```
Modularity of Fast-Greedy of Node ID = 108 is: 0.4359581
```

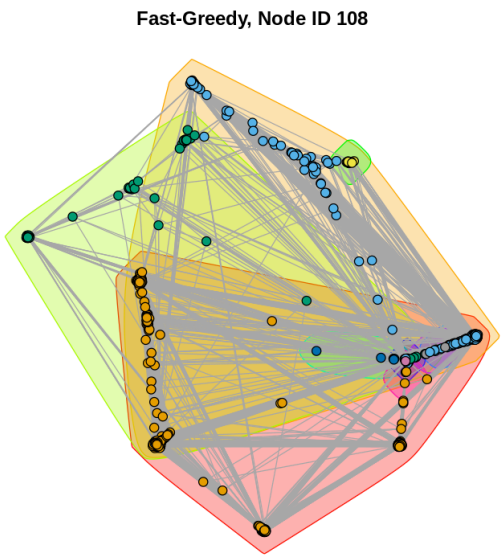

Figure 1.8. Community structure using Fast-Greedy for Node ID 108

```
Modularity of Edge-Betweenness of Node ID = 108 is: 0.5067549
```

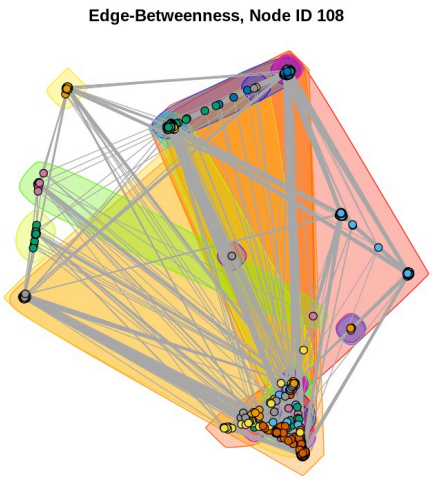

Figure 1.9. Community structure using Edge-Betweenness for Node ID 108

Modularity of Infomap of Node ID = 108 is: 0.5084537

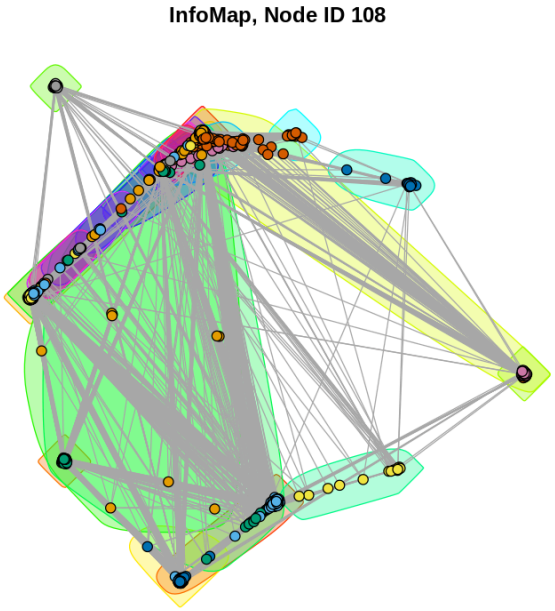

Figure 1.10. Community structure using Infomap for Node ID 108

Modularity of Fast-Greedy of Node ID = 349 is: 0.2503461

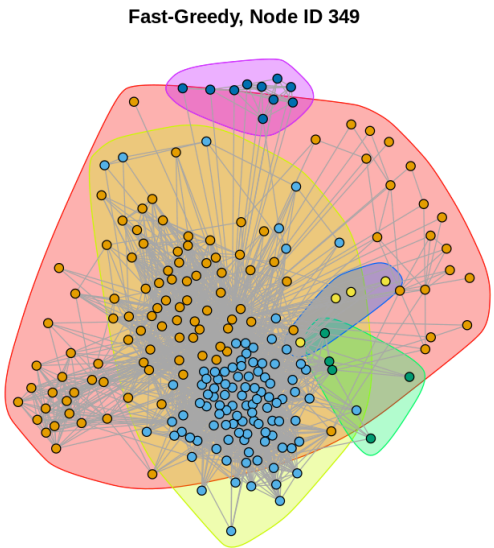

Figure 1.11. Community structure using Fast-Greedy for Node ID 349

```
Modularity of Edge-Betweenness of Node ID = 349 is: 0.133528
```

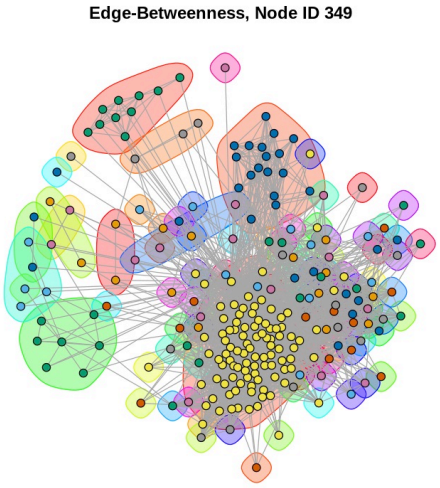

Figure 1.12. Community structure using Edge-Betweenness for Node ID 349

```
Modularity of Infomap of Node ID = 349 is: 0.0954642
```

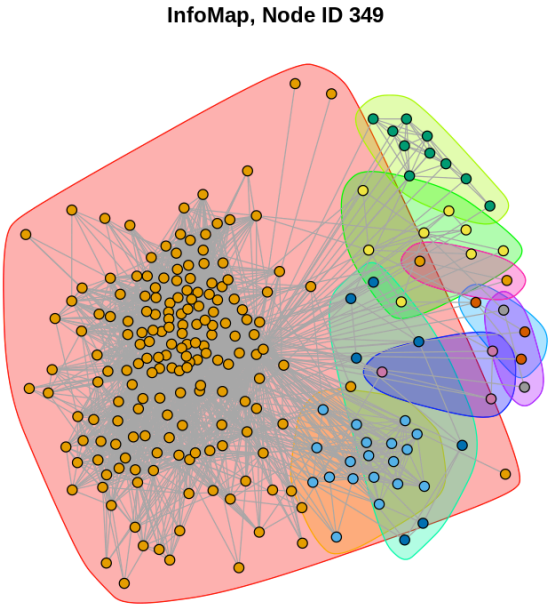

Figure 1.13. Community structure using Infomap for Node ID 349

Modularity of Fast-Greedy of Node ID = 484 is: 0.5070016

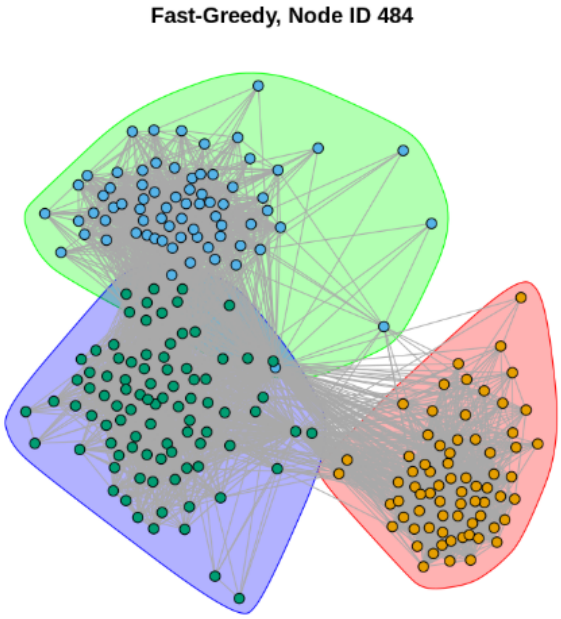

Figure 1.14. Community structure using Fast-Greedy for Node ID 484

Modularity of Edge-Betweenness of Node ID = 484 is: 0.4890952

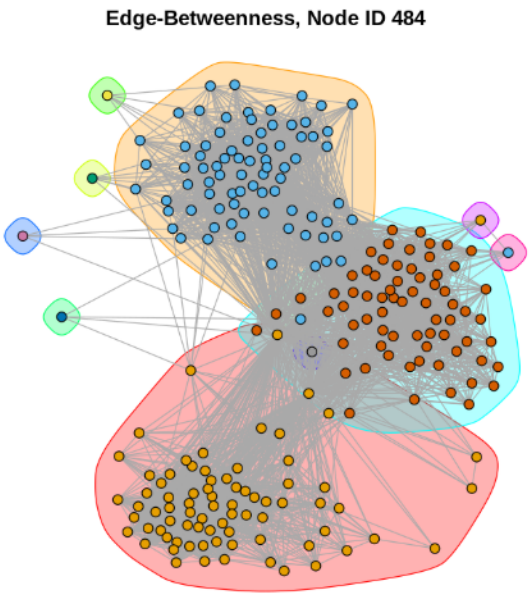

Figure 1.15. Community structure using Edge-Betweenness for Node ID 484

Modularity of Infomap of Node ID = 484 is: 0.5152788

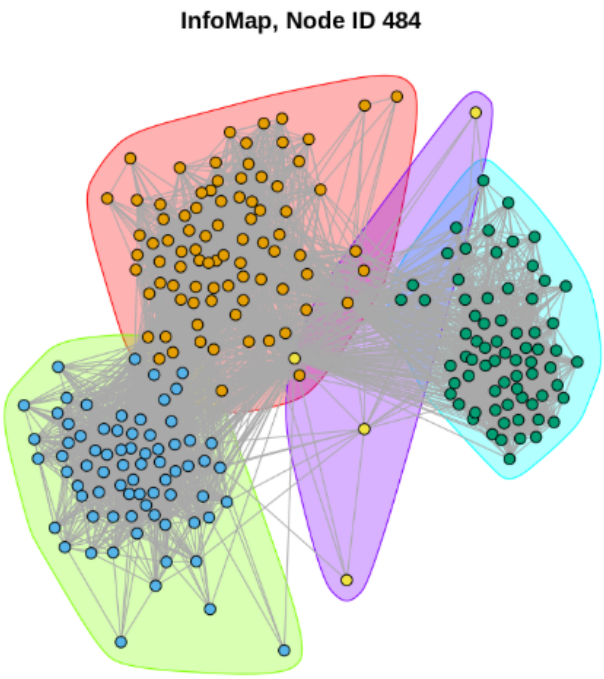

Figure 1.16. Community structure using Infomap for Node ID 484

Modularity of Fast-Greedy of Node ID = 1087 is: 0.1455315

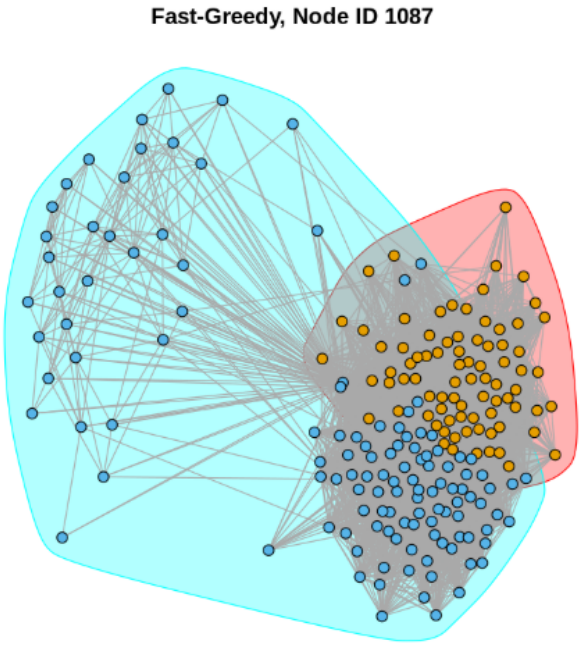

Figure 1.17. Community structure using Fast-Greedy for Node ID 1087

Modularity of Edge-Betweenness of Node ID = 1087 is: 0.02762377

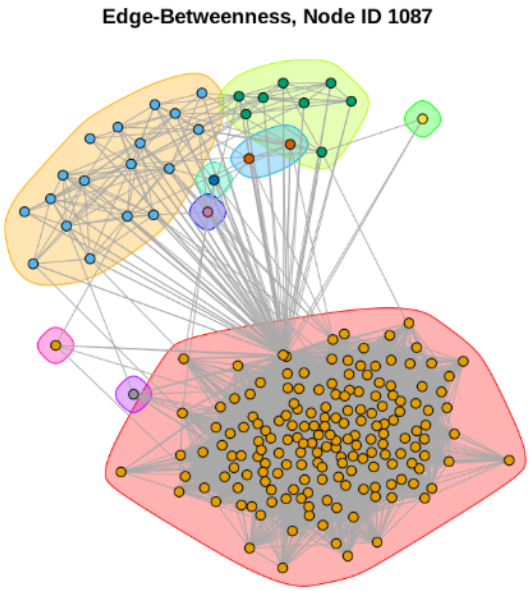

Figure 1.18. Community structure using Edge-Betweenness for Node ID 1087

Modularity of Infomap of Node ID = 1087 is: 0.02690662

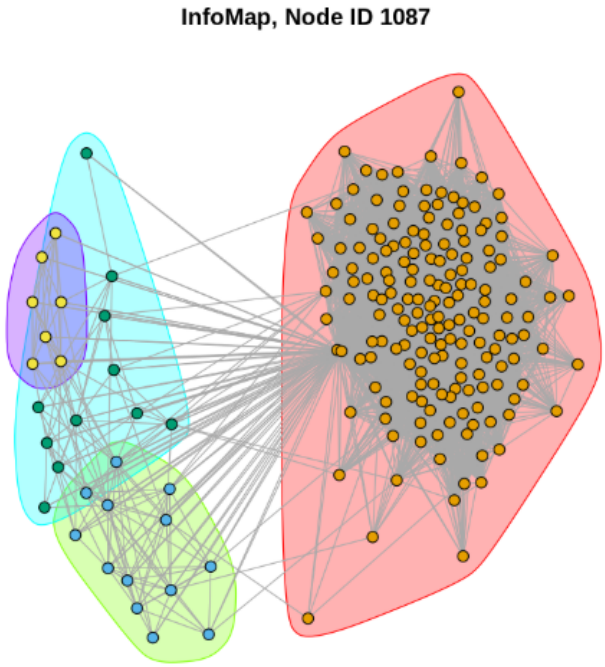

Figure 1.19. Community structure using Infomap for Node ID 1087

Comparison: First of all, we can see the modularities are shown in the following table:

Table 1. Modularity score for 3 community detection algorithms

	Fast-Greedy	Edge-Betweenness	Infomap
Node ID 1	0.436	0.507	0.508
Node ID 108	2.863	3.889	9.405
Node ID 349	0.251	0.134	0.095
Node ID 484	0.507	0.489	0.515
Node ID 1087	0.146	0.028	0.027

After analysing the data in the table above, we can conclude that:
Fast-Greedy performs better than the other 2 methods, since Fast-Greedy has higher modularity. However, for Node ID 108, the other two methods (Edge-Betweenness and Infomp) performs better compared to Fast-Greedy. Since Node ID 108 has the largest number of nodes in the personalized network, we can clearly see that Edge-Betweenness and Infomap have better performance when facing large networks, while Fast-Greedy has better performance when dealing with smaller networks.
On the other hand, the code running time for Fast-Greedy and Infomap are less than Edge-Betweenness.

1.3.2 Community structure with the core node removed

Question 10: After removing the core nodes: 1 108 349 484 1087 , respectively for each graph, we get the community structure of the modified personalized network, using the same community detection algorithm as Question 9.
Comparison: First of all, we can see the modularities are shown in the following table:

Table 2. Modularity score after removing the core node

	Fast-Greedy	Edge-Betweenness	Infomap
Node ID 1	0.442	0.416	0.418
Node ID 108	0.458	0.521	0.521
Node ID 349	0.246	0.151	0.245
Node ID 484	0.534	0.515	0.543
Node ID 1087	0.148	0.032	0.027

After analysing data in the table above, we can conclude that:
The modularity score of the personalized networks without core nodes are higher than those with core nodes.
For the personalized networks with core nodes the core node only in one community. However, since all of the rest nodes are connected to the core one, the connection among other communities and community with core node are really dense, which lower the degree of modularity.
On the other hand, for personalized networks without the core node, the connection between communities is comparably sparse, so it receives higher modularity scores. After removing the core nodes, the modularity score without core nodes are still close to the scores of personalized networks with core nodes, though they increase a little bit.

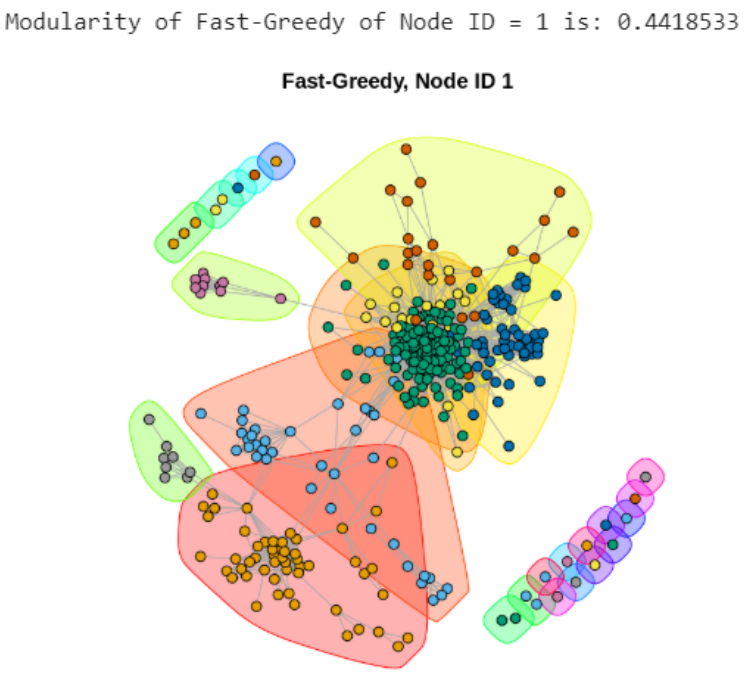

Figure 1.20. Community structure of the modified network, using Fast-Greedy for Node ID 1

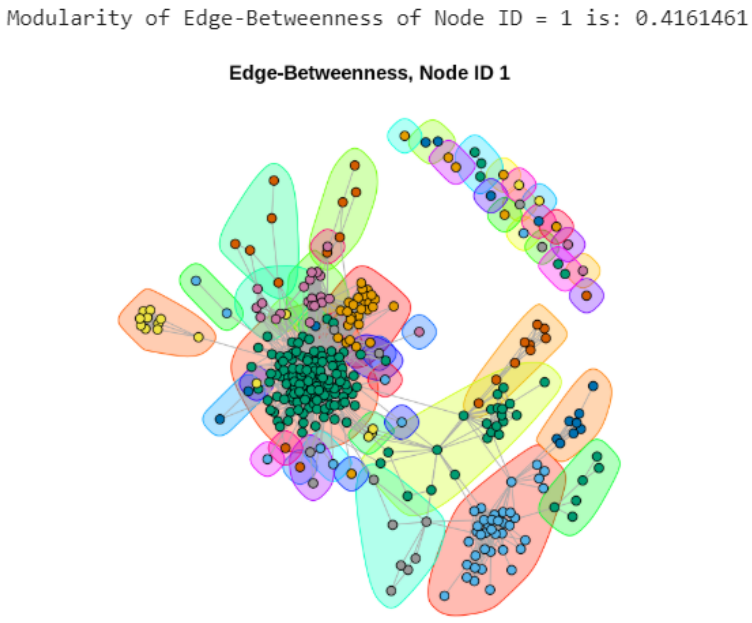

Figure 1.21. Community structure of the modified network, using Edge-Betweenness for Node ID 1

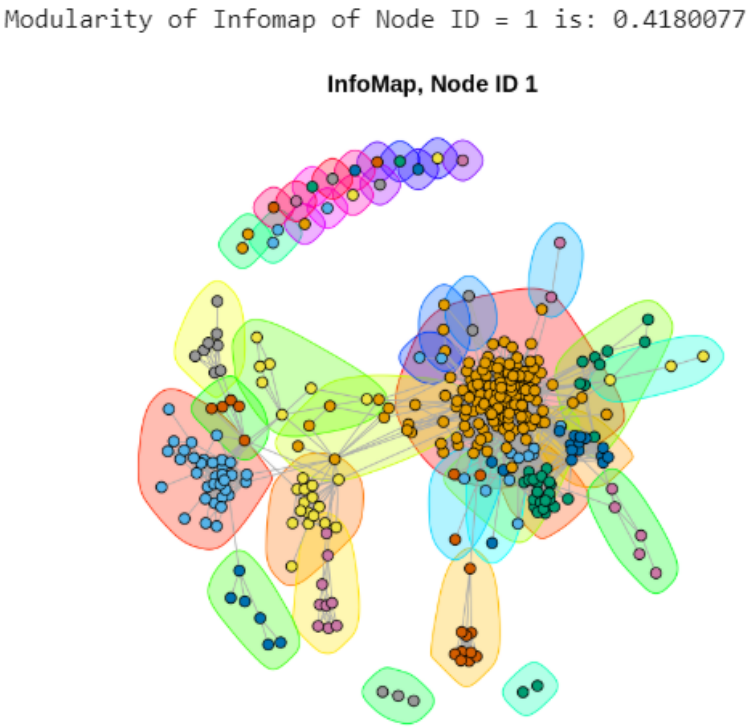

Figure 1.22. Community structure of the modified network, using Infomap for Node ID 1

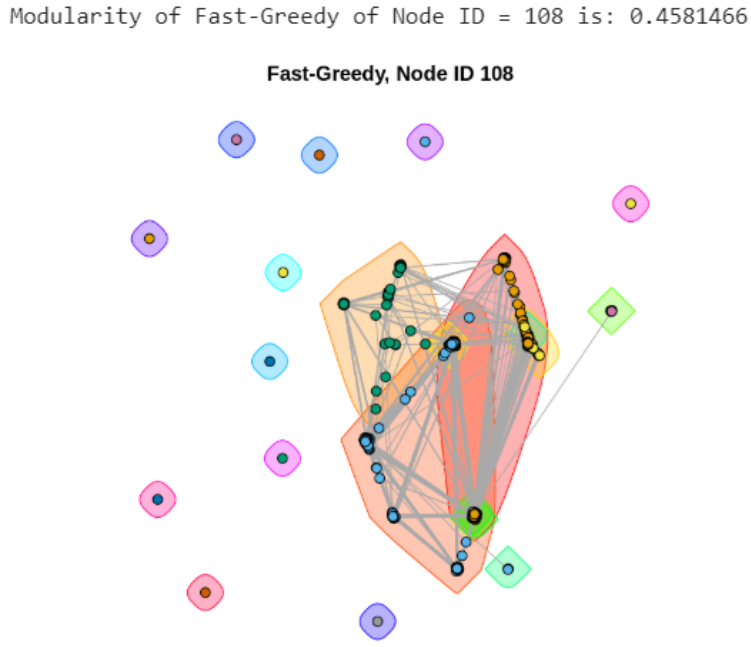

Figure 1.23. Community structure of the modified network, using Fast-Greedy for Node ID 108

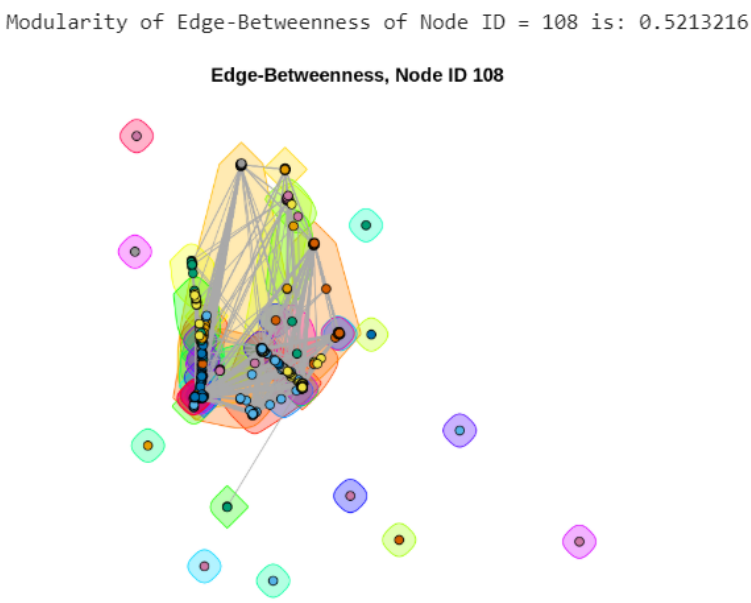

Figure 1.24. Community structure of the modified network, using Edge-Betweenness for Node ID 108

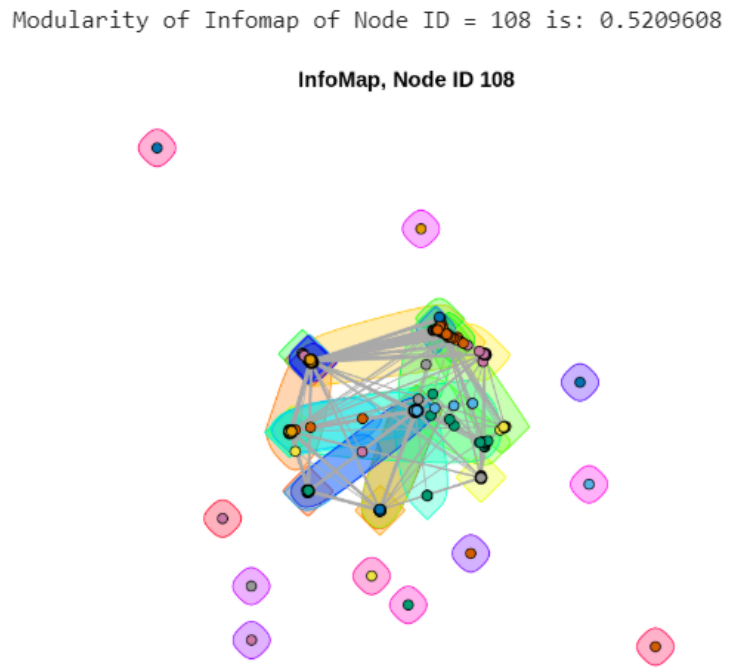

Figure 1.25. Community structure of the modified network, using Infomap for Node ID 108

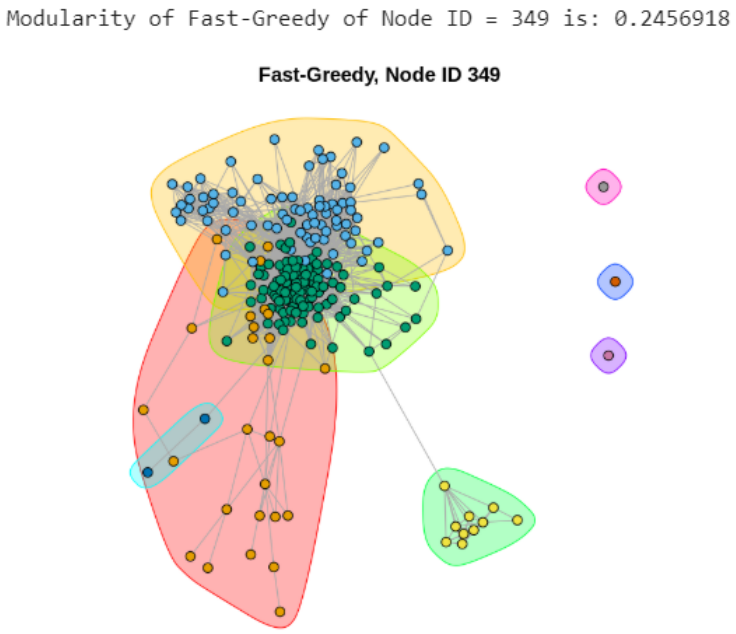

Figure 1.26. Community structure of the modified network, using Fast-Greedy for Node ID 349

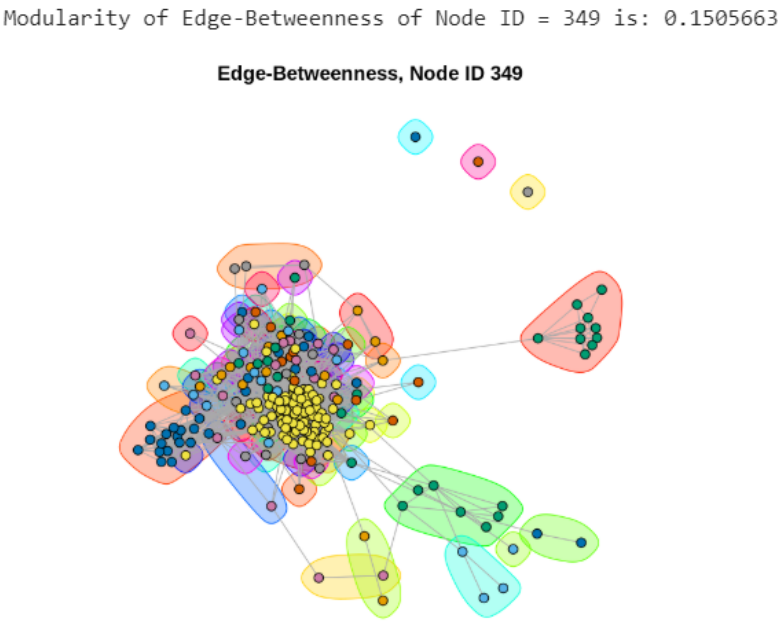

Figure 1.27. Community structure of the modified network, using Edge-Betweenness for Node ID 349

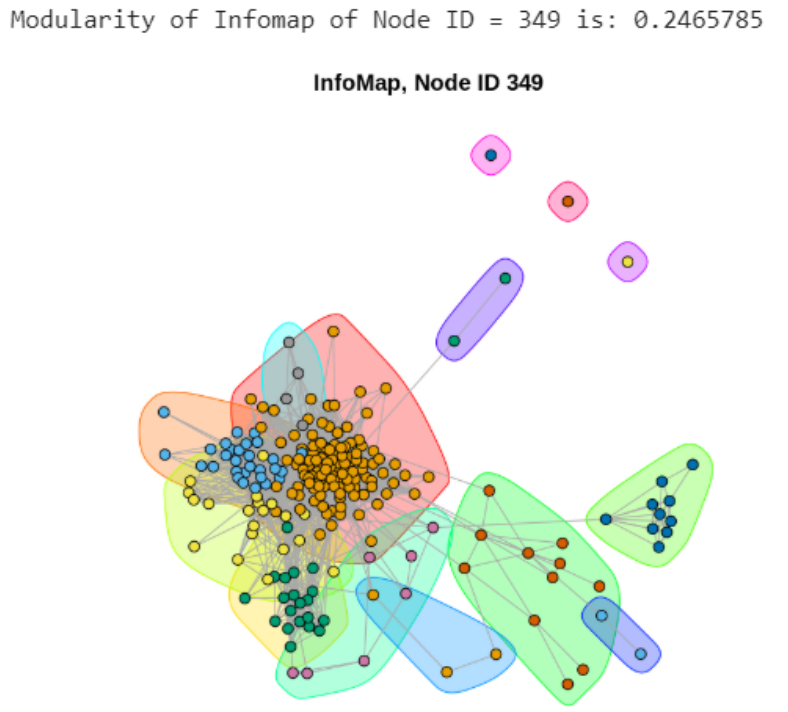

Figure 1.28. Community structure of the modified network, using Infomap for Node ID 349

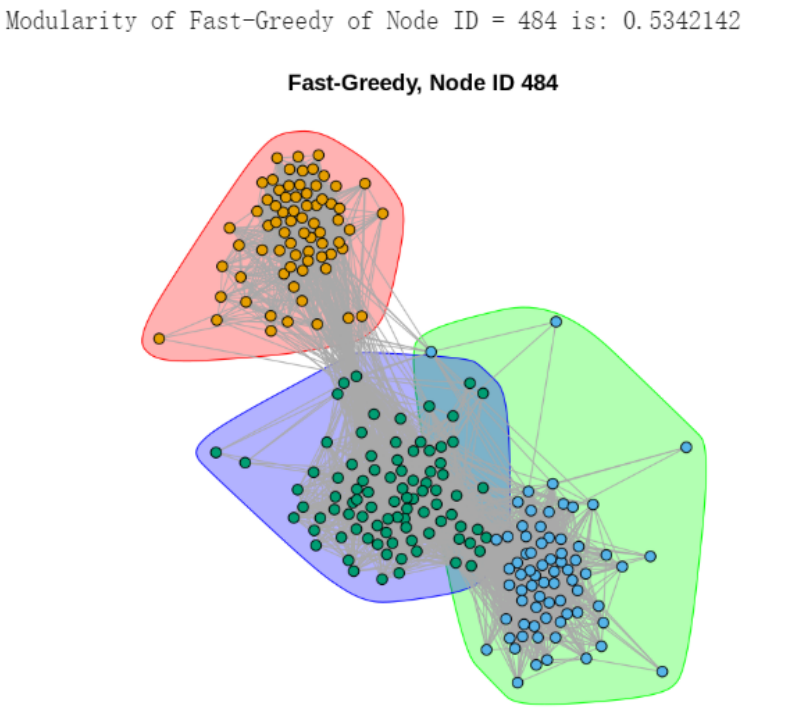

Figure 1.29. Community structure of the modified network, using Fast-Greedy for Node ID 484

Modularity of Edge-Betweenness of Node ID = 484 is: 0.5154413

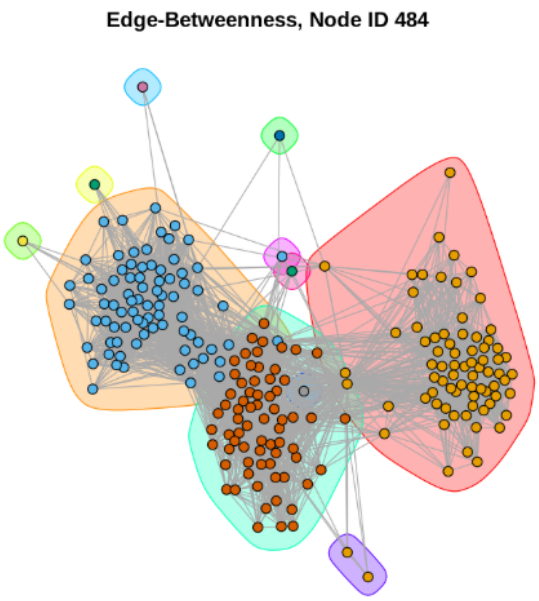

Figure 1.30. Community structure of the modified network, using Edge-Betweenness for Node ID 484

Modularity of Infomap of Node ID = 484 is: 0.5434437

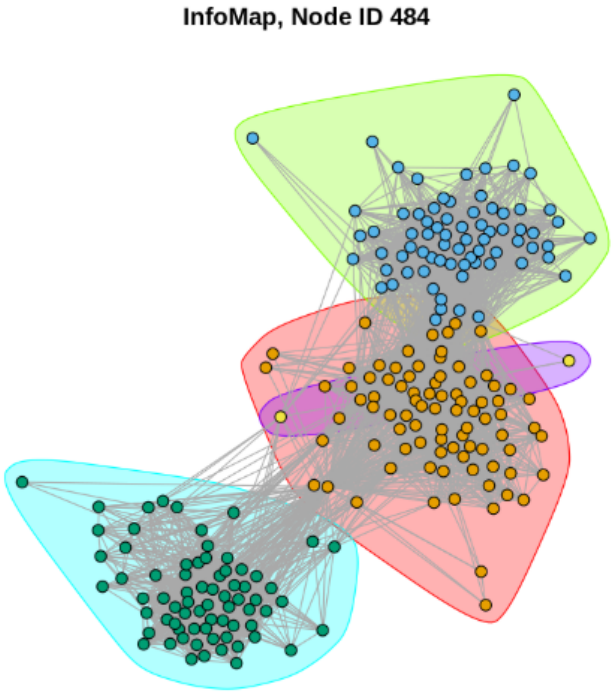

Figure 1.31. Community structure of the modified network, using Infomap for Node ID 484

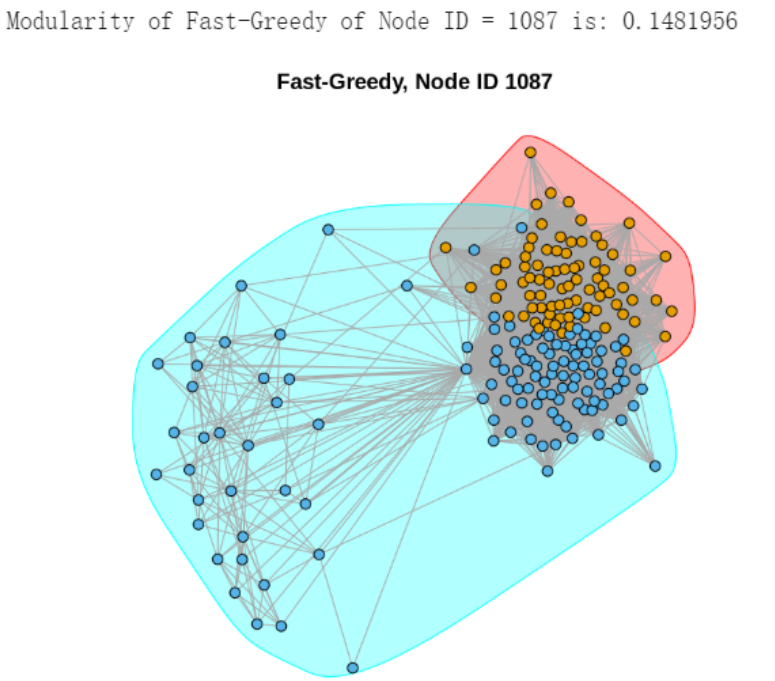

Figure 1.32. Community structure of the modified network, using Fast-Greedy for Node ID 1087

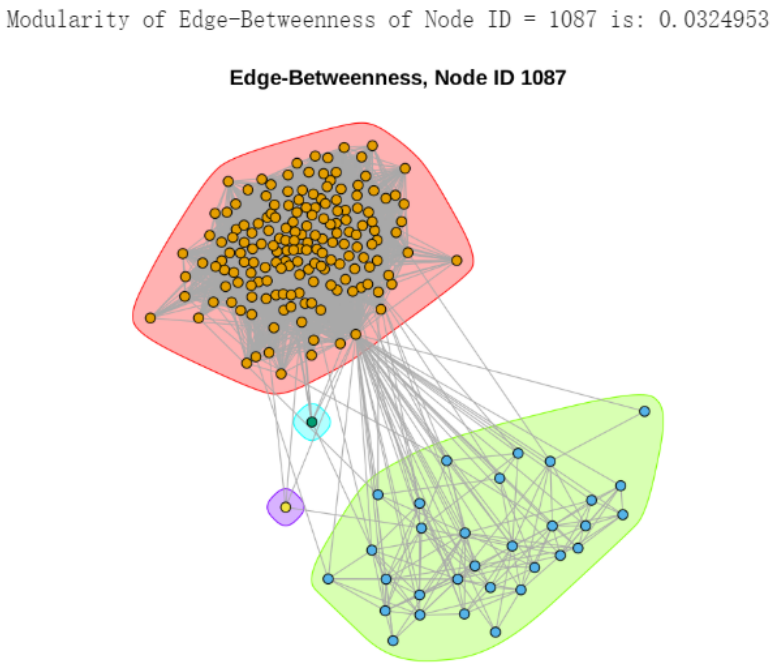

Figure 1.33. Community structure of the modified network, using Edge-Betweenness for Node ID 1087

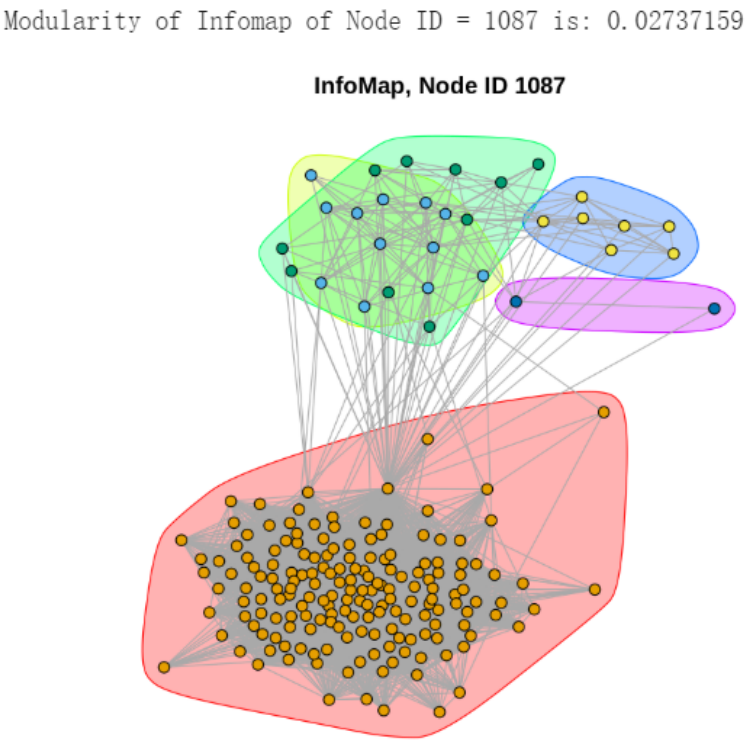

Figure 1.34. Community structure of the modified network, using Infomap for Node ID 1087

1.3.3 Characteristic of nodes in the personalized network

We explore characteristics of nodes in the personalized network using two measures, Embeddedness and Dispersion.

Question 11:
As the core node connects with every node in its personalized network, it connects with every non-core node that that the non-core node i we are considering connects to. So the embeddedness of this non-core node i can be simply expressed by the following expression:

$$Embeddedness(i) = degree(i) - 1$$

Question 12: For each of the core node's personalized network in Question 9, we plot the distribution histogram of embeddedness and dispersion:

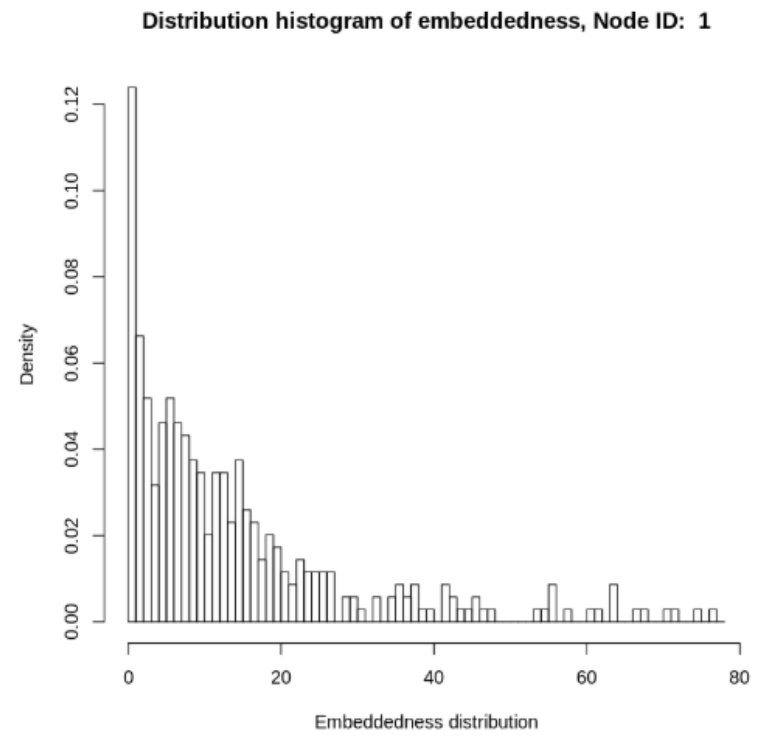

Figure 1.35. Distribution histogram of embeddedness for Node ID 1

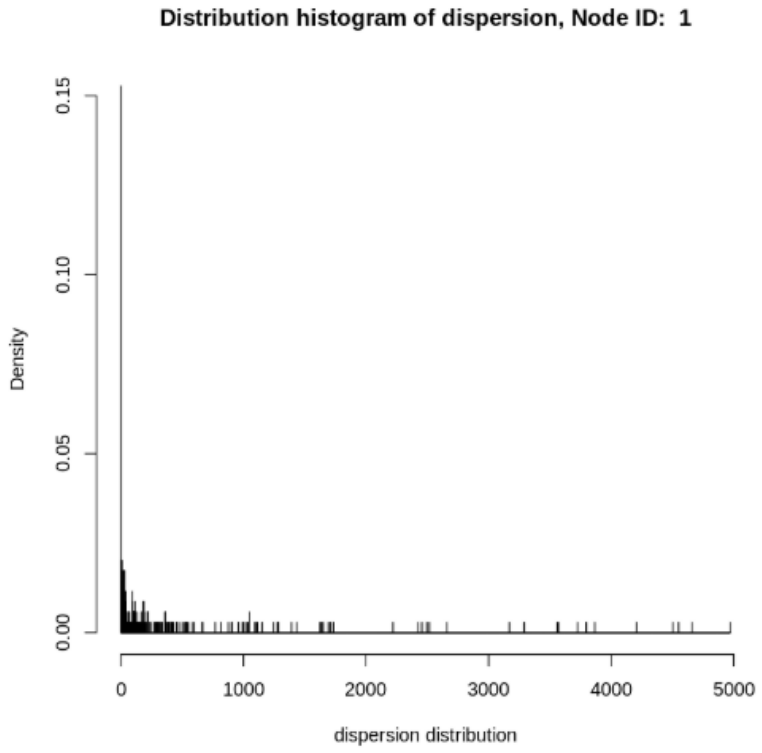

Figure 1.36. Distribution histogram of dispersion for Node ID 1

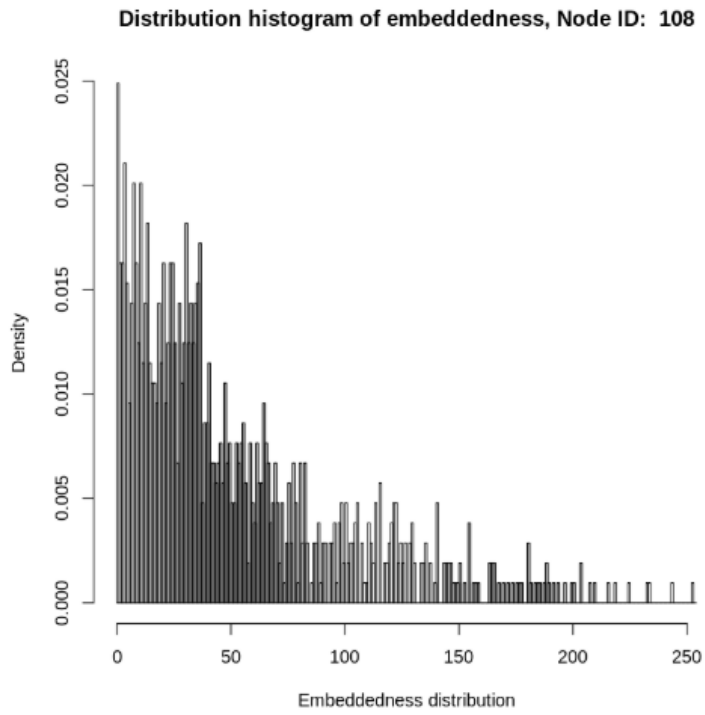

Figure 1.37. Distribution histogram of embeddedness for Node ID 108

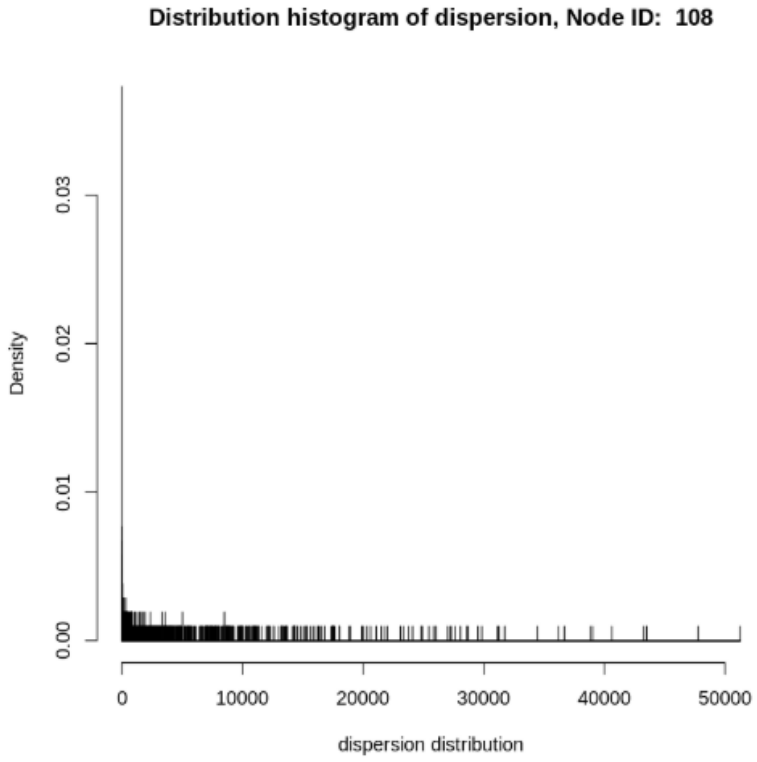

Figure 1.38. Distribution histogram of dispersion for Node ID 108

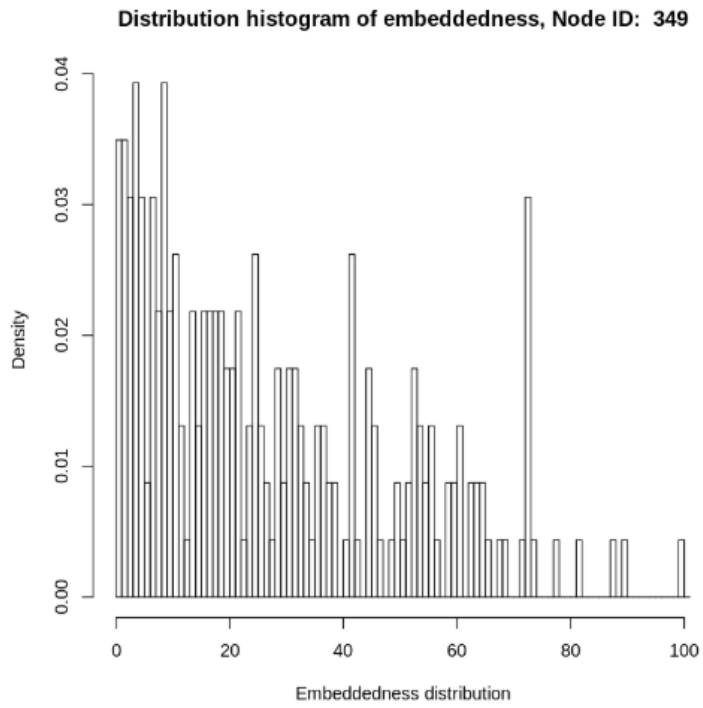

Figure 1.39. Distribution histogram of embeddedness for Node ID 349

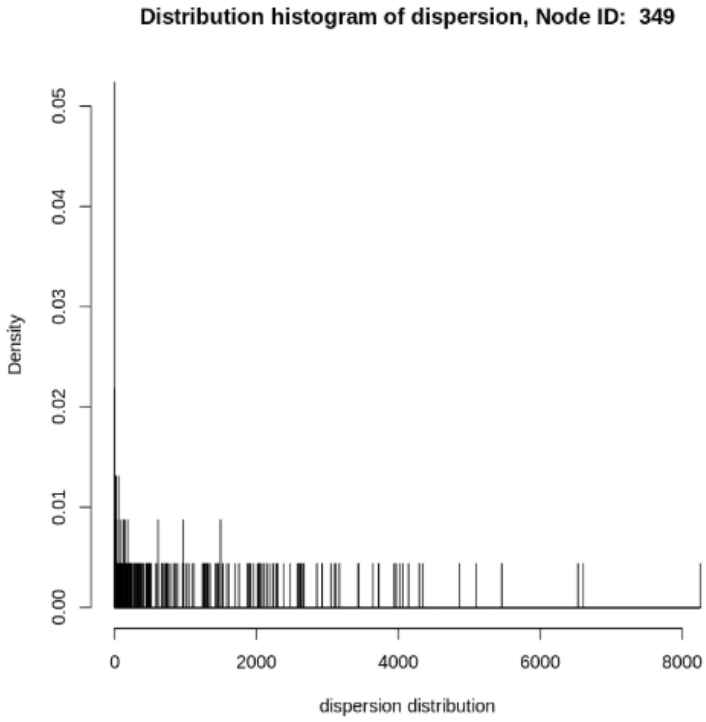

Figure 1.40. Distribution histogram of dispersion for Node ID 349

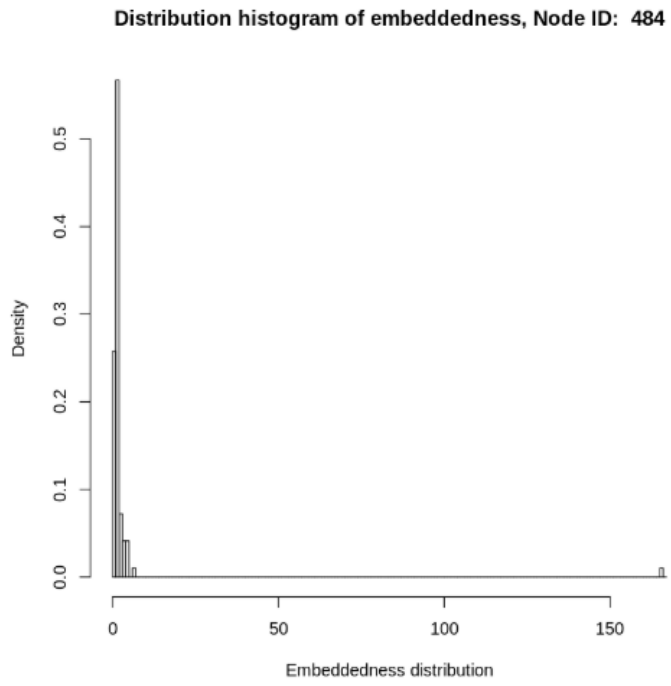

Figure 1.41. Distribution histogram of embeddedness for Node ID 484

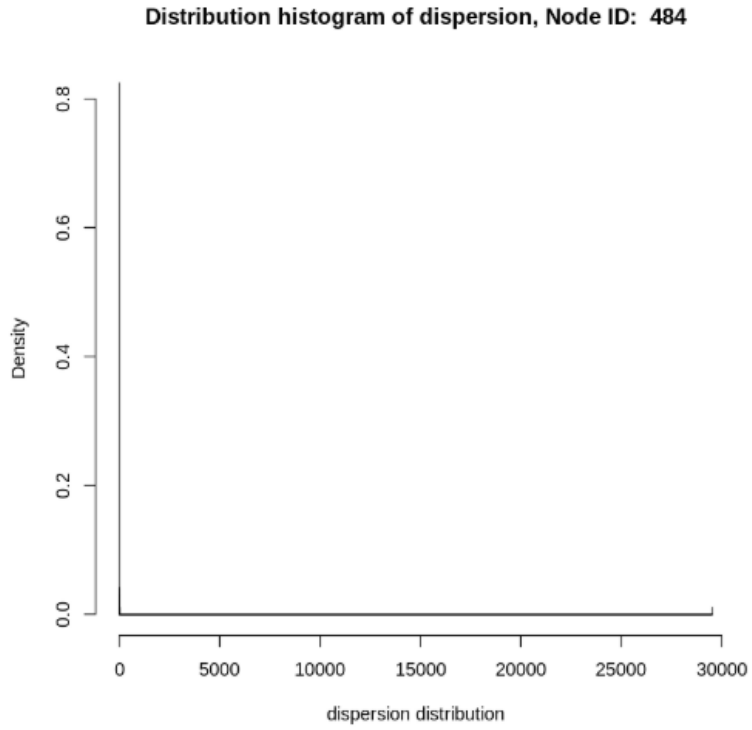

Figure 1.42. Distribution histogram of dispersion for Node ID 484

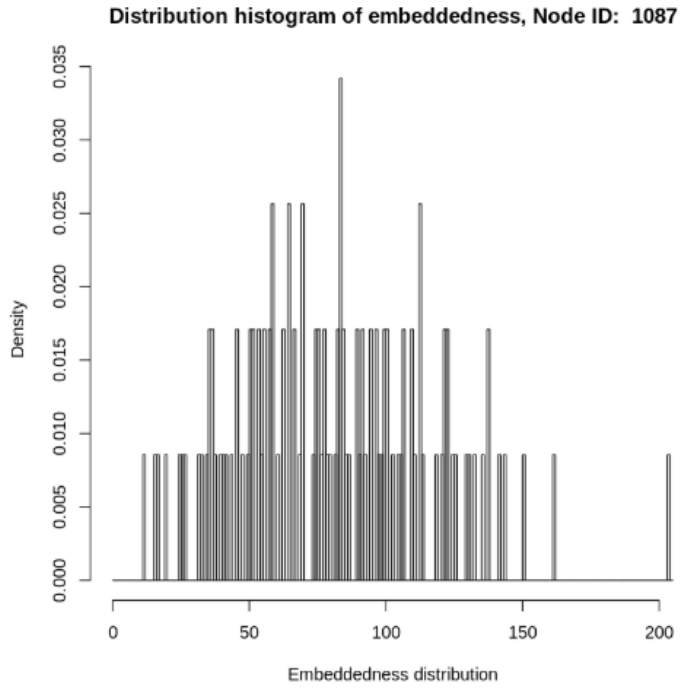

Figure 1.43. Distribution histogram of embeddedness for Node ID 1087

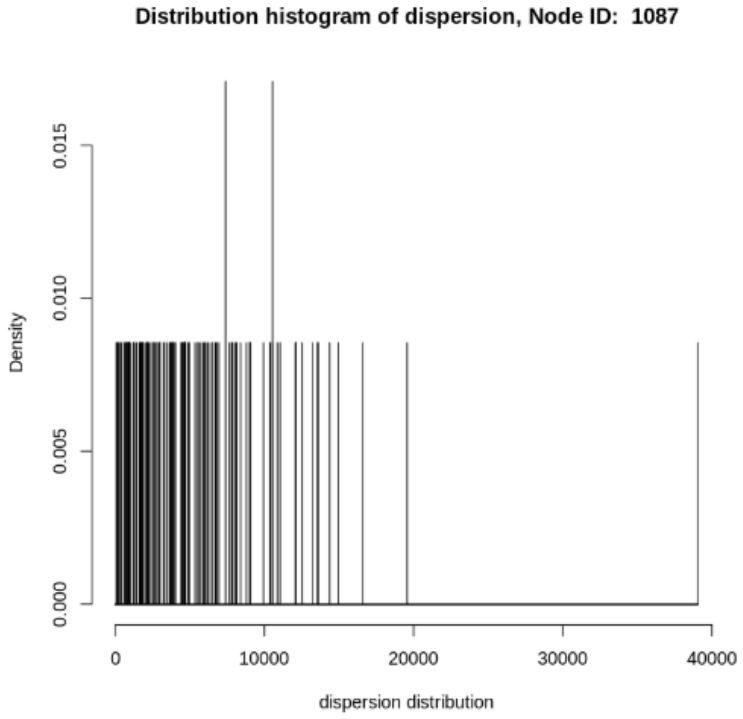

Figure 1.44. Distribution histogram of dispersion for Node ID 1087

Question 13: We plot the community structure of the personalized network using colors and highlight the node with maximum dispersion, we also highlight the edges incident to this node. The plots for them are shown below:

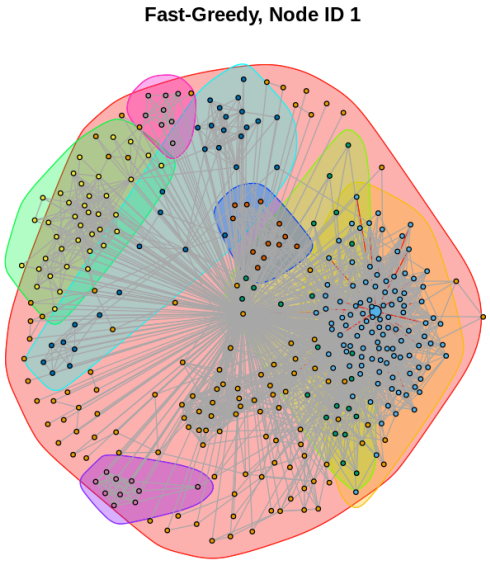

Figure 1.45. Fast-greedy with highlight for node id 1

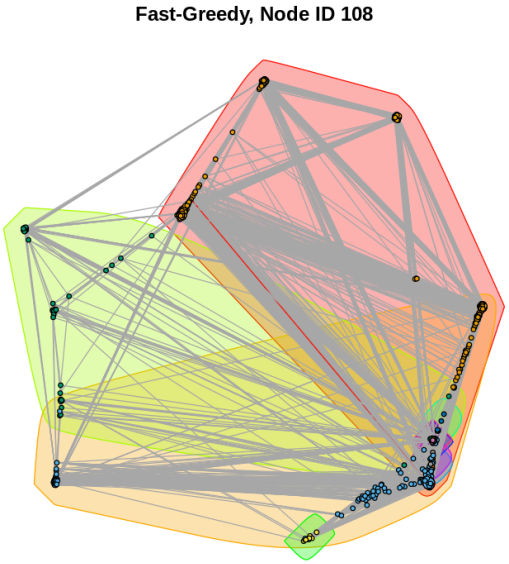

Figure 1.46. Fast-greedy with highlight for node id 108

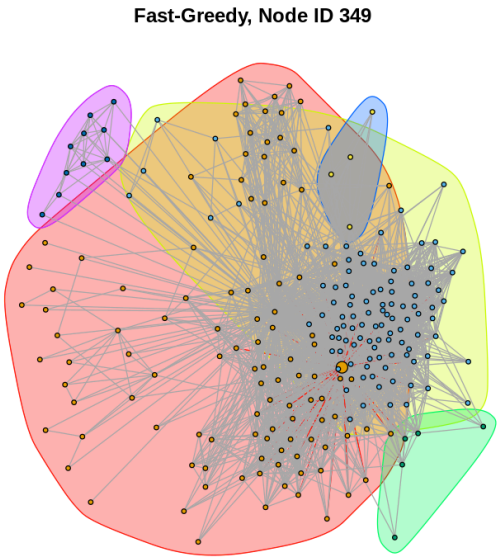

Figure 1.47. Fast-greedy with highlight for node id 349

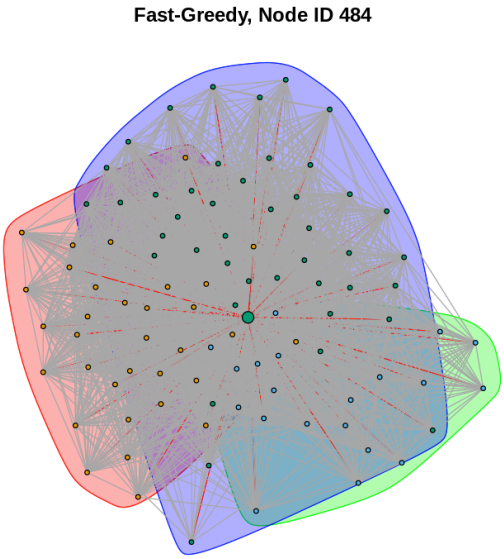

Figure 1.48. Fast-greedy with highlight for node id 484

```
Modularity of Fast-Greedy of Node ID = 1087 is: 0.1455315
The node with maximum embeddedness is ID: 108
```

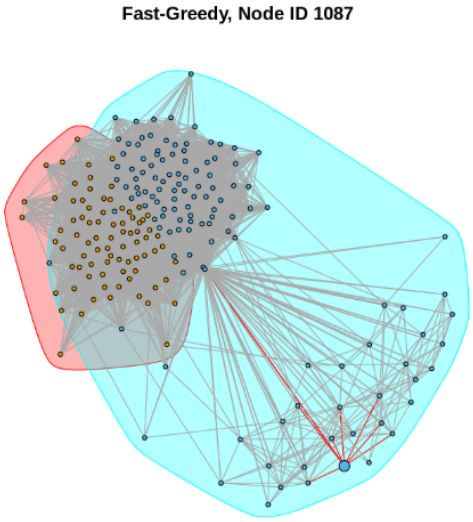

Figure 1.49. Fast-greedy with highlight for node id 1087

Question 14:

```
Modularity of Fast-Greedy of Node ID = 1 is: 0.4131014
The node with maximum dispersion (red) is ID: 57
The node with maximum divation (green) is ID: 16
```

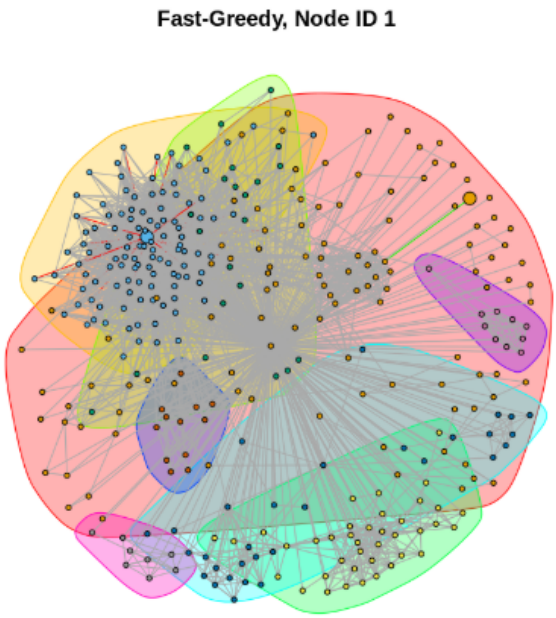

Figure 1.50. Fast-greedy with highlight for node id 1

```
Modularity of Fast-Greedy of Node ID = 108 is: 0.4359581
The node with maximum dispersion (red) is ID: 1889
The node with maximum divation (green) is ID: 1878
```

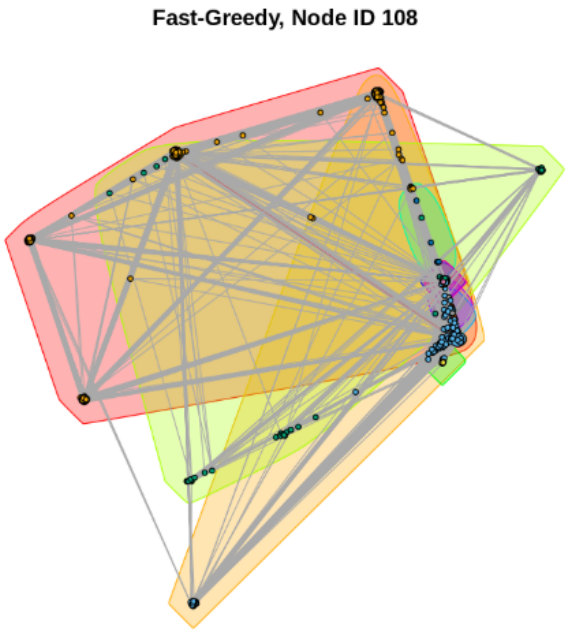

Figure 1.51. Fast-greedy with highlight for node id 108

```
Modularity of Fast-Greedy of Node ID = 349 is: 0.2503461
The node with maximum dispersion (red) is ID: 377
The node with maximum divation (green) is ID: 377
```

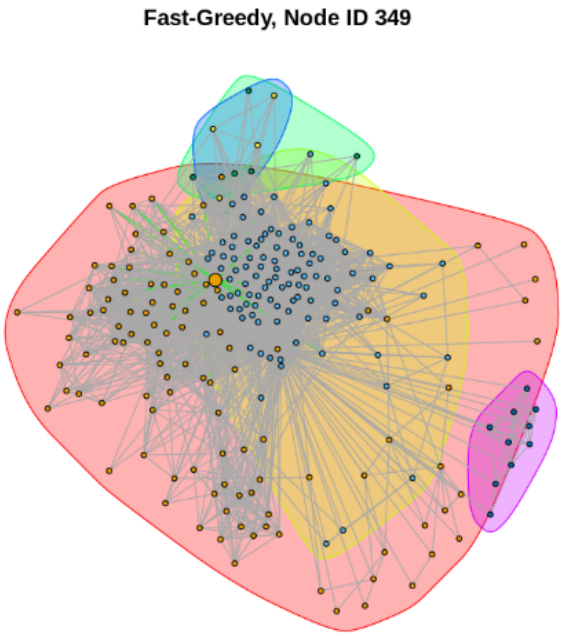

Figure 1.52. Fast-greedy with highlight for node id 349

```
Modularity of Fast-Greedy of Node ID = 484 is: 0.1008014
The node with maximum dispersion (red) is ID: 108
The node with maximum divation (green) is ID: 108
```

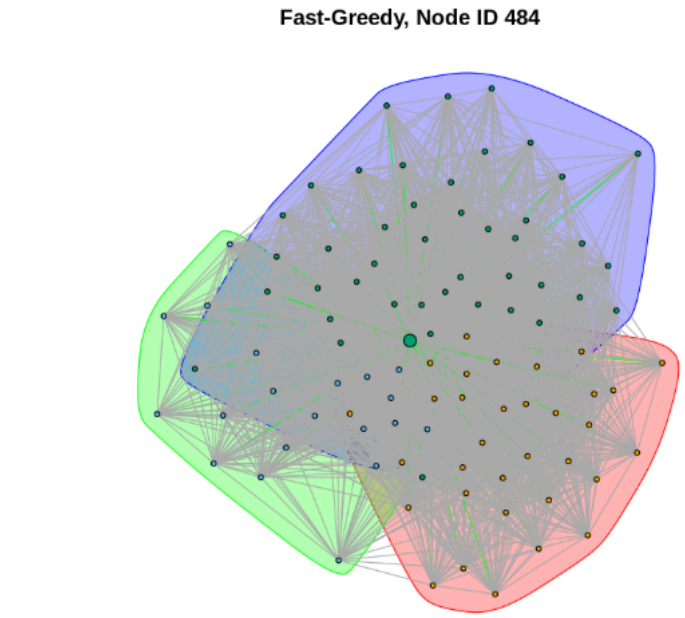

Figure 1.53. Fast-greedy with highlight for node id 484

```
Modularity of Fast-Greedy of Node ID = 1087 is: 0.1103563
The node with maximum dispersion (red) is ID: 108
The node with maximum divation (green) is ID: 108
```

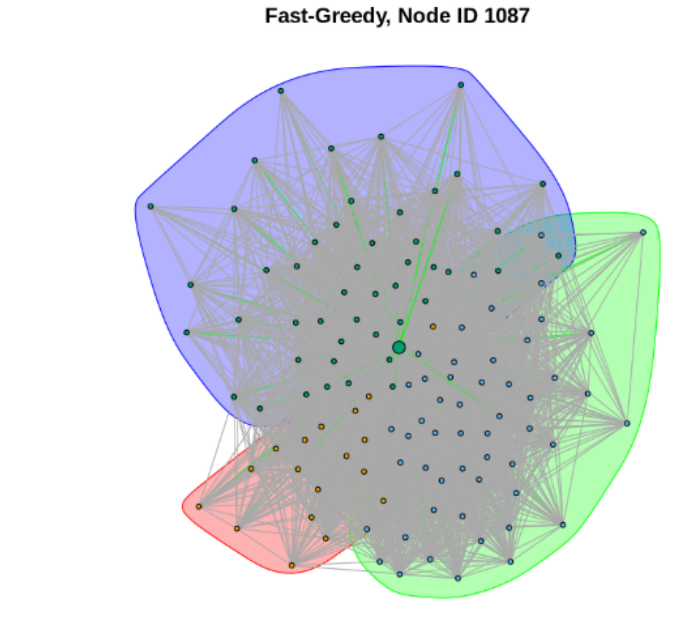

Figure 1.54. Fast-greedy with highlight for node id 1087

Question 15:

The maximum dispersion node has the greatest sum of distances between mutual friends with the core node, where a distance which is greater than 2 would be considered as a distance value. Since a farther

distance apart would mean that the dispersion value is greater, so the node and the core node are not always in the same community.
The maximum embeddedness node is the node that shares the maximum number of friends with the core node. As we can see in the plots above, a large number of plots has its node is in the center of the communities. Since the core node is a neighbor of each node in the personalized network, and is located in the center of the easiest graph structure. Since the maximum embeddedness node should have the biggest amount of neighbors, so it's right to be placed centralized in the communities.
When the core node and the node have a distance between them, and they share a large number of mutual friends, then the maximum dispersion node as well as the maximum embeddedness node can be the same node.
The maximum dispersion/embeddedness node represents the nodes which have small embeddedness and large dispersion. It means that there is not so many friends for the core node, and its friends are not strongly connected. Node ID 484 and node ID 1087 have the same node for dispersion/embeddedness and embeddedness, as well as dispersion, which means that the dispersion value should be so large to counteract the comparatively large embeddedness values.

1.4 Friend recommendation in personalized networks

Question 16:
We generate the personalized graph for Node ID 415, we counted that the nodes had a degree of 24.
$|N_r| = 11$
Question 17:
Using Common Neighbors measure: 0.8525899
Using Jaccard measure: 0.8684343
Using Adamic Adar measure: 0.8591657
Based on the average accuracy values, the three friend recommendation algorithms have similar performance, with Jaccard measure slightly better that the others.

2 Google+ Network

Question 18: There are 57 personal networks for users who have more than 2 circles.
Question 19:
The plots for 3 personal networks are shown below:

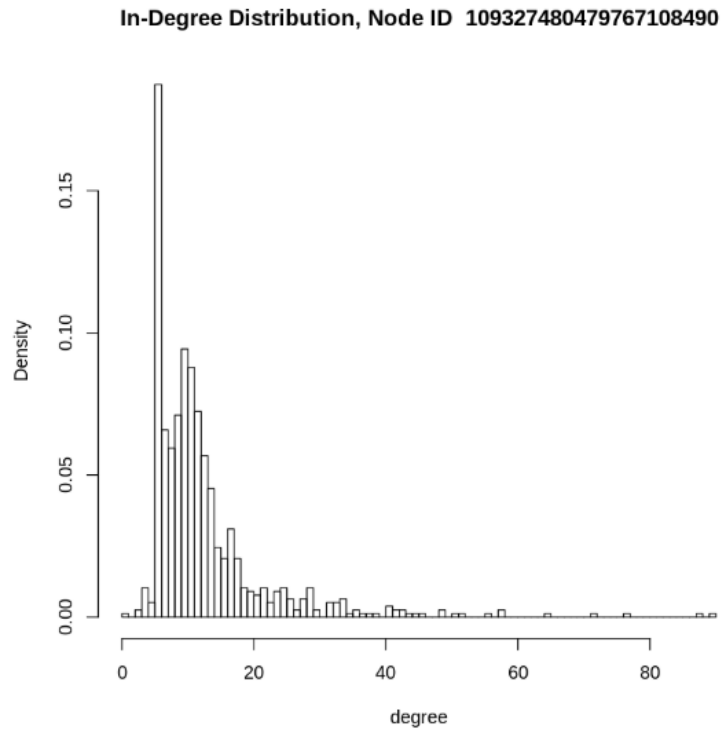

Figure 2.1. In-degree distribution of node 10932748047976710849

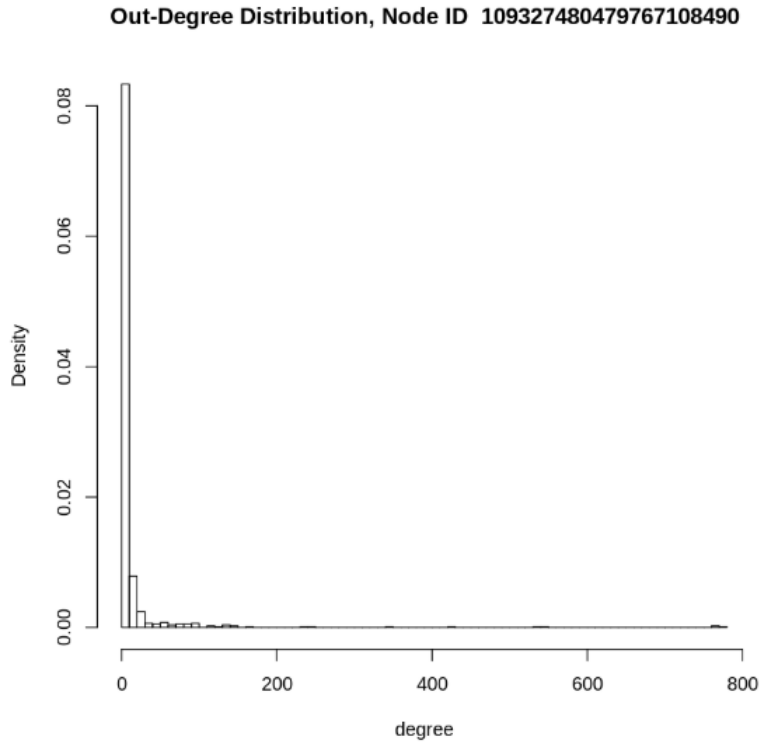

Figure 2.2. Out-degree distribution of node 10932748047976710849

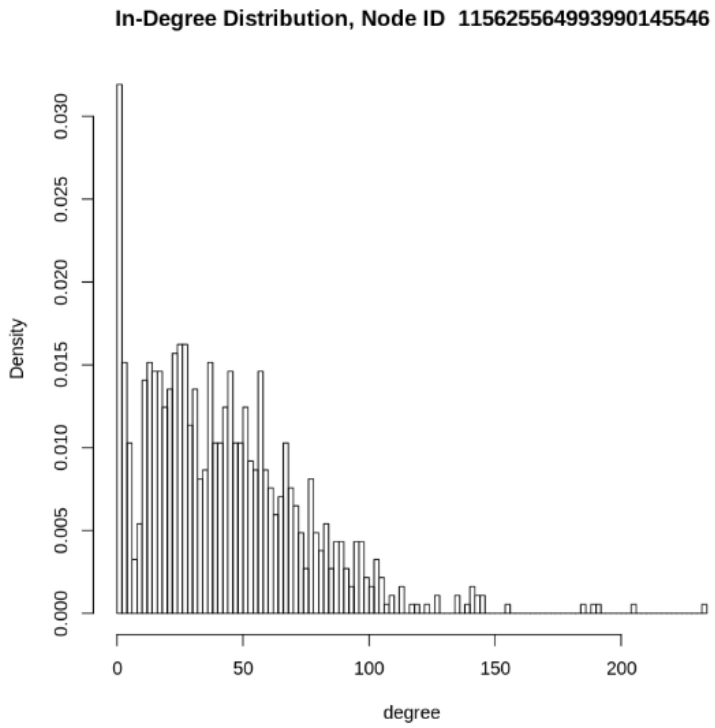

Figure 2.3. In-degree distribution of node 11562556499399014554 6

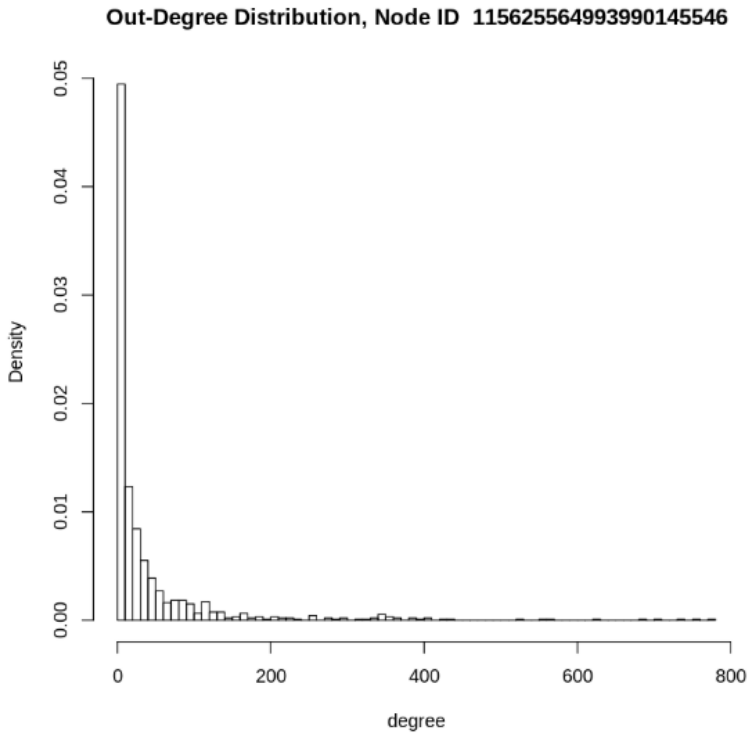

Figure 2.4. Out-degree distribution of node 11562556499399014554 6

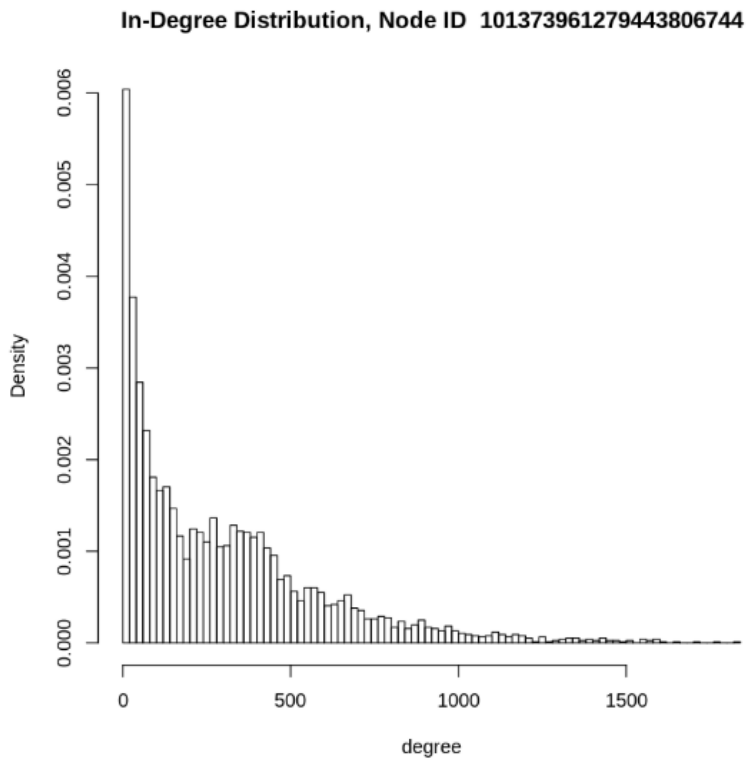

Figure 2.5. In-degree distribution of node 10137396127944380674

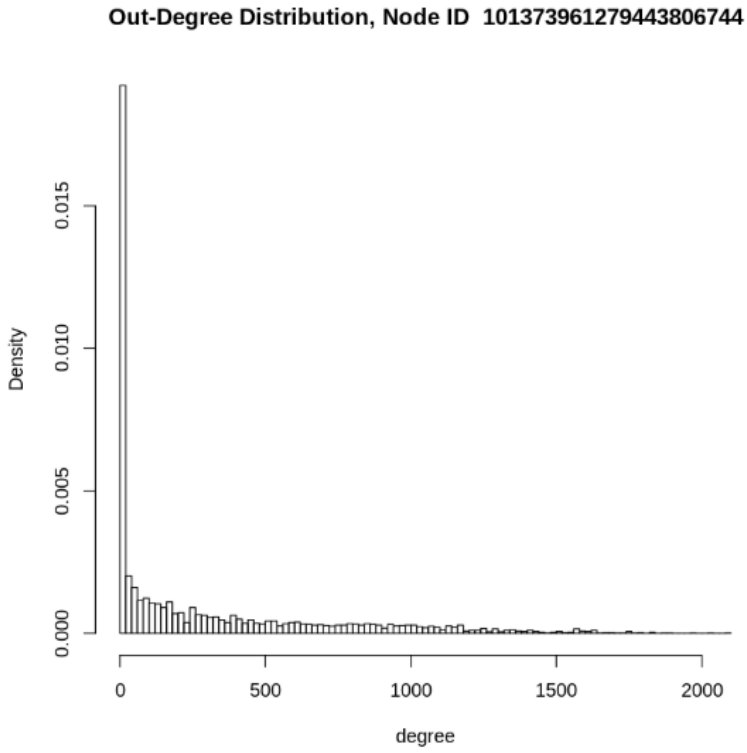

Figure 2.6. Out-degree distribution of node 10137396127944380674

The in-degree distributions shows difference, we can clearly see that the Node ID 10137396127944380674 and Node ID 109327480479767108490 have personalized networks which fall off while degree in-

creases. However, for Node ID 115625564993990145546, the personalized network have more linear fall off, therefore compared to other two Node IDs, there is a large portion of users have high in-degree. The out-degree distribution apears the same for these 3 personalized networks.

2.1 Community structure of personal networks

Question 20: After extract the community structure of each personal network using Walktrap community detection algorithm, the modularity scores and the communities with colors are shown below:

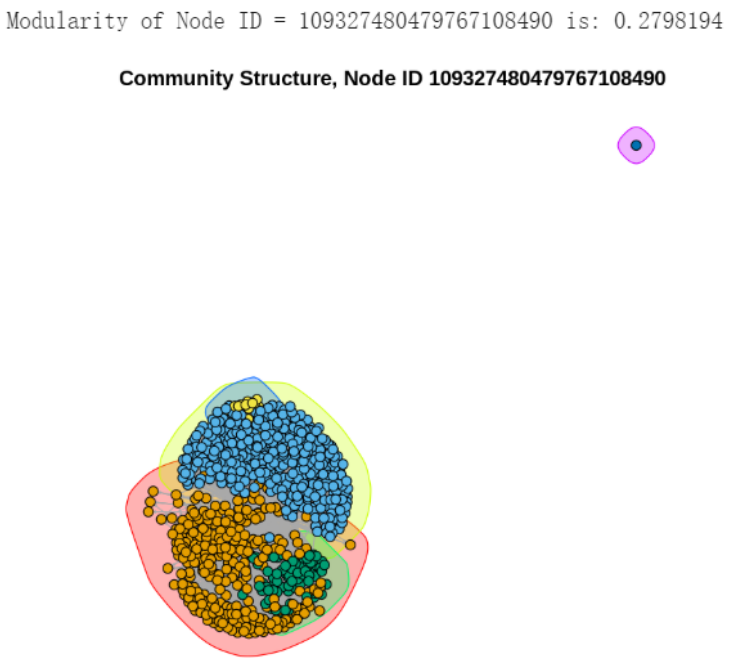

Figure 2.7. Community structure for Node ID 109327480479767108490

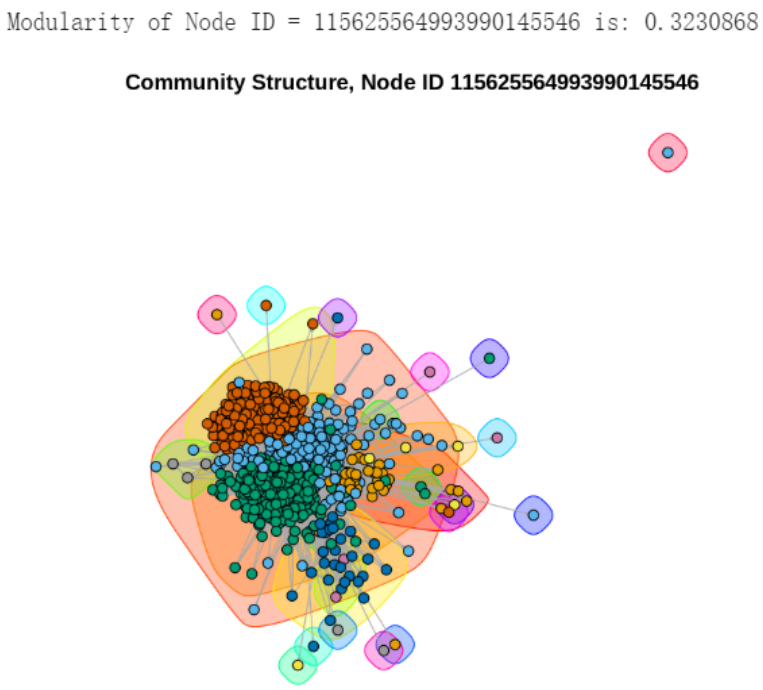

Figure 2.8. Community structure for Node ID 11562556493990145546

Figure 2.9. Community structure for Node ID 101373961279443806744

The modularity score is not similar. The second node has the highest modularity score and the third one has the lowest score.

Question 21: Homogeneity measures the variation in the community structure. If every community contains members that all in the same circle, then the homogeneity reaches higher score.

Completeness means to measure the variation in each circle member's community assignment. When the circle assigns nodes to the same community, then the completeness reaches higher score.

Question 22:
The homogeneity and completeness values for the three community structures are shown below:

```
Homogeneity of Node ID = 109327480479767108490 is: 0.8640041
Completeness of Node ID = 109327480479767108490 is: 0.3456923
Homogeneity of Node ID = 115625564993990145546 is: 0.4429445
Completeness of Node ID = 115625564993990145546 is: -3.375718
Homogeneity of Node ID = 101373961279443806744 is: 0.001839249
Completeness of Node ID = 101373961279443806744 is: -1.609263
```

Figure 2.10. Statistics for 3 community structures

As we can see in the statistics above, the first node has the highest homogeneity and completeness values, while the third one has the lowest homogeneity and completeness values. We already state the meaning for homogeneity and completeness in Question 21, based on what we declared in that question, we can conclude that:
The first node forms communities within the node circle best, and the nodes from this circle form the same community. However, the third node performs bad when forms the communities with nodes coming from that circle, and nodes from the circle form different communities, that is why we have negative values for the third node.