ECE232 Project2
Social Network Mining

Juo-Wen Fu 805025223
Alex Hung 705035114
Te-Yuan Liu 805027255
Wesly Tzuo 704946416

May 8, 2018

1 Facebook network

1.1 Structural properties of the facebook network

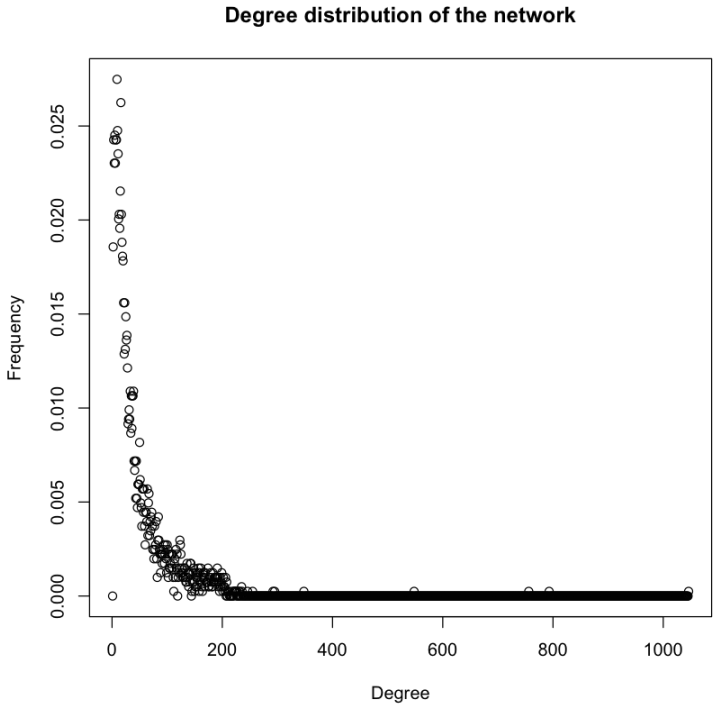
Question 1: Is the facebook network connected? If not, find the giant connected component (GCC) of the network and report the size of the GCC.

Yes, the Facebook network is connected. There are 4039 vertices thus the GCC size is also 4039.

Question 2: Find the diameter of the network. If the network is not connected, then find the diameter of the GCC.

The diameter of the network is 8.

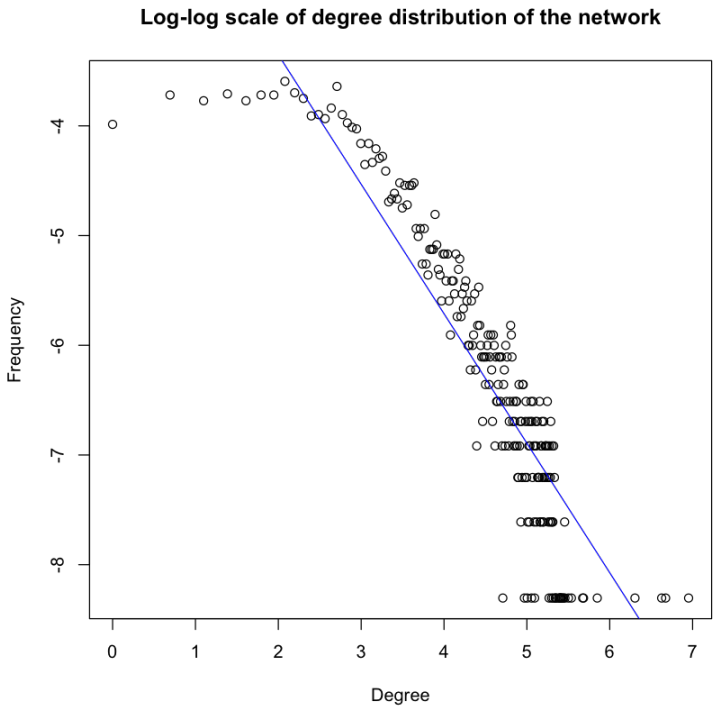
Question 3: Plot the degree distribution of the Facebook network and report the average degree.



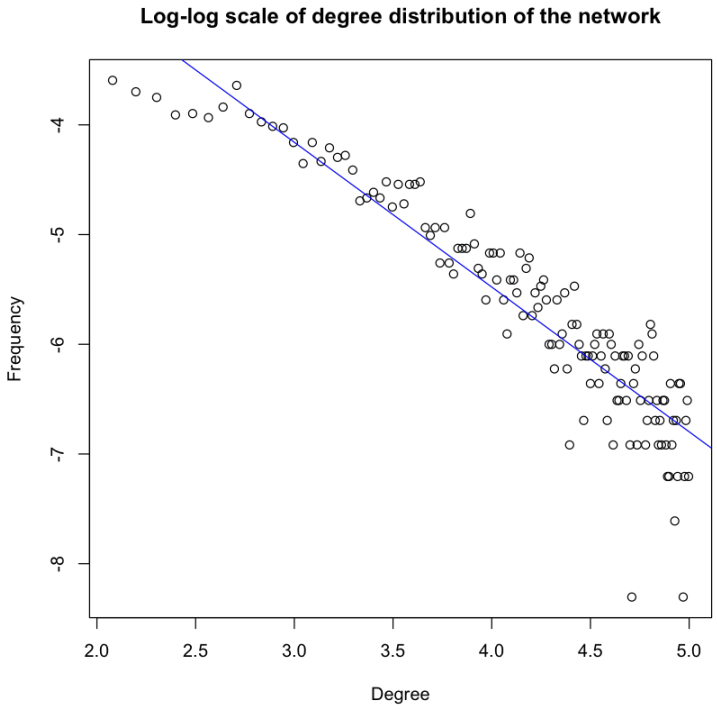
As we can see in the plot, most of the nodes have lower than 200 degrees.
Average degree = 43.69

Question 4: Plot the degree distribution of question 3 in a log-log scale. Try to fit a line to the plot and estimate the slope of the line.

We scale frequency and degree to log then plot the fitted line. Here the slope is -1.1802.

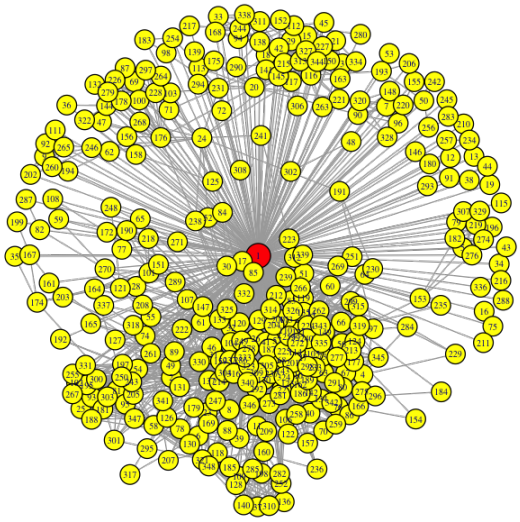


We remove the outliers and use the middle part (from log2 to log5)to fit the line. Then we get slope equals to -1.3194.



1.2 Personalized network

Question 5: Create a personalized network of the user whose ID is 1. How many nodes and edges does this personalized network have?



There are 348 nodes and 2866 edges in this personalized network.

Question 6: What is the diameter of the personalized network? Please state a trivial upper and lower bound for the diameter of the personalized network.

The diameter of the personalized network is 2. The upper bound of a personalized network is 2 and the lower bound is 0.

1.3 Core nodes personalized network

Question 7: In the context of the personalized network, what is the meaning of the diameter of the personalized network to be equal to the upper bound you derived in question 6. What is the meaning of the diameter of the personalized network to be equal to the lower bound you derived in question 6?

When number of nodes in a personalize network is larger than 2, the diameter is always 2. Since all nodes other than the center are neighbor of center, each pair of node has a shortest path of 2. For example, from node a to center and center to node b. This means that every person other than center shares a friend, center. When a personalized network only contain one node, the diameter is 0. This means that the center has no friends.

Question 8: How many core nodes are there in the Facebook network. What is the average degree of the core nodes?

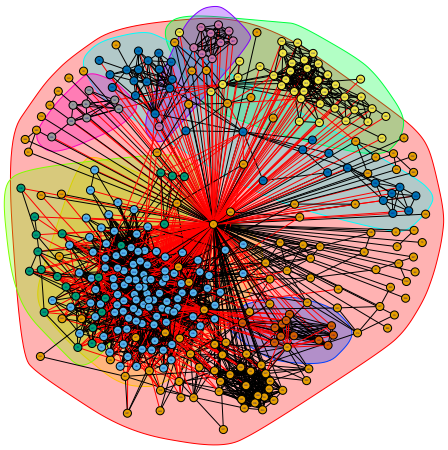
There are 40 core nodes in the Facebook network. The average degree of the core nodes is 279.35.

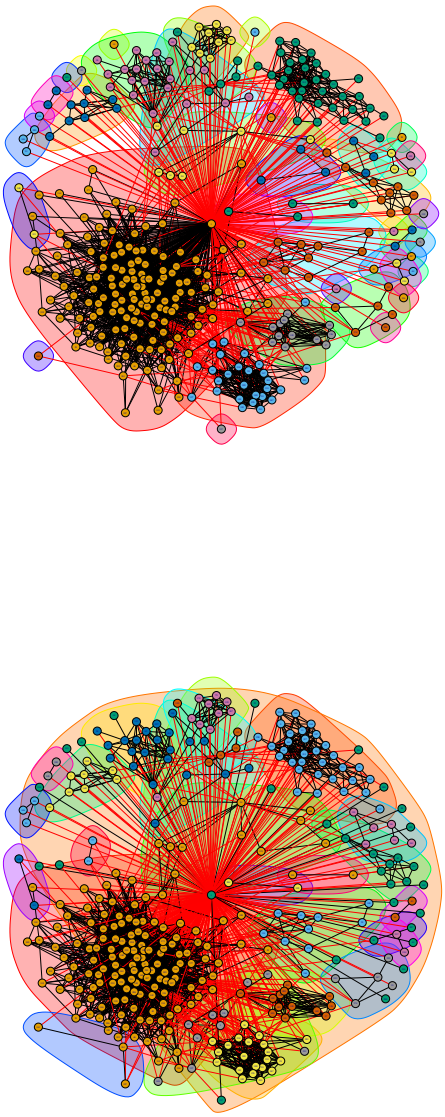
Question 9: For each of the above core nodes personalized network, find the community structure using Fast-Greedy, Edge-Betweenness, and Infomap community detection algorithms. Compare the modularity scores of the algorithms. For visualization purpose, display the community structure of the core nodes personalized networks using colors. Nodes belonging to the same community should have the same color and nodes belonging to different communities should have different color. In this question, you should have 15 plots in total.

For Node ID 1, and the modularity scores are

Algorithm	Fast-Greedy	Edge-Betweenness	Infomap
Modularity score	0.4131	0.3533	0.3891

The community structures of the Node ID 1 personalized networks from Fast-Greedy, Edge-Betweenness, and Infomap community detection algorithm are shown below:

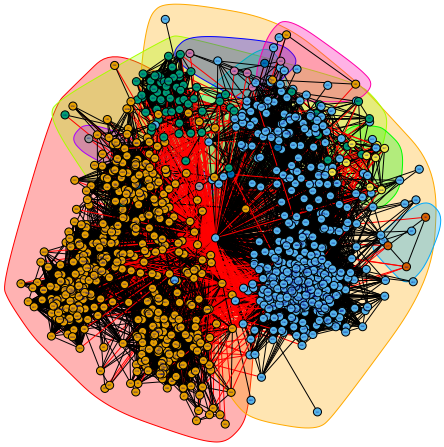


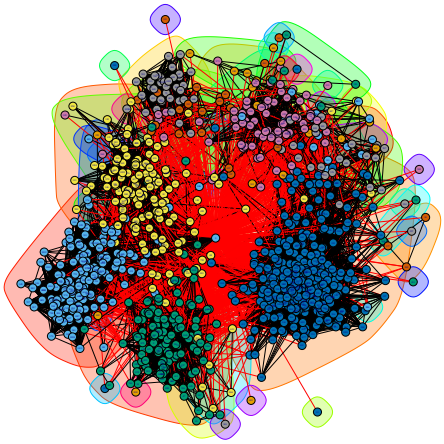


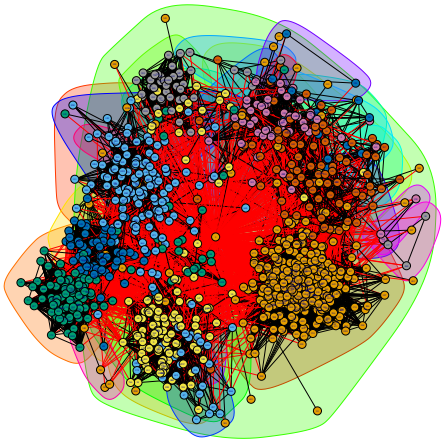
For Node ID 108, and the modularity scores are

Algorithm	Fast-Greedy	Edge-Betweenness	Infomap
Modularity score	0.4359	0.5068	0.5082

The community structures of the Node ID 108 personalized networks from Fast-Greedy, Edge-Betweenness, and Infomap community detection algorithm are shown below:



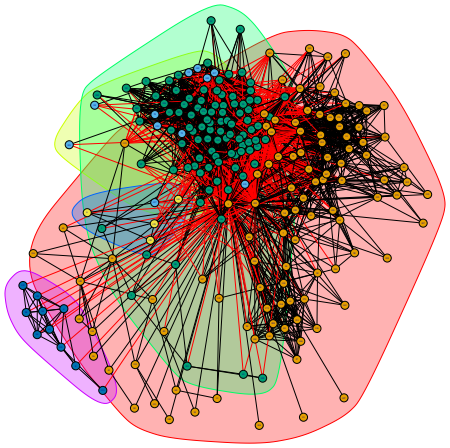


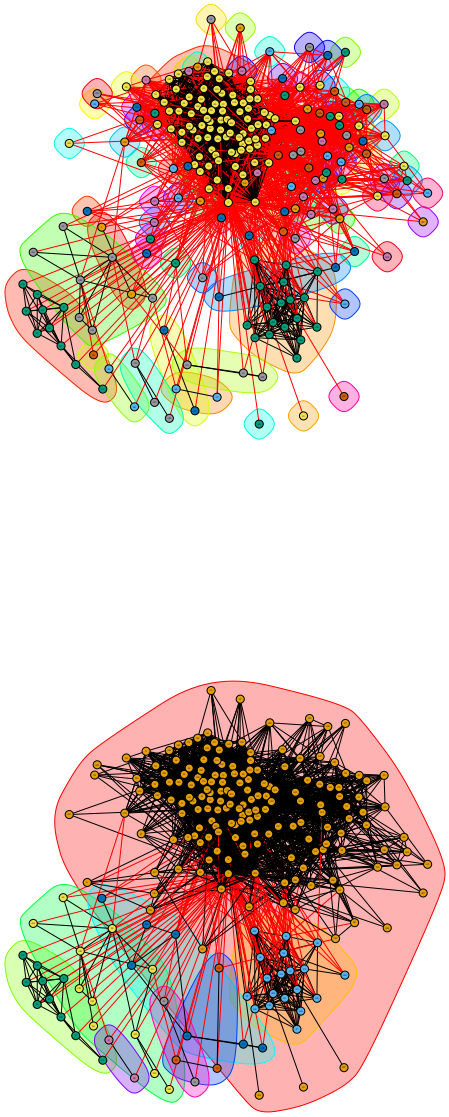


For Node ID 349, and the modularity scores are

Algorithm	Fast-Greedy	Edge-Betweenness	Infomap
Modularity score	0.2517	0.1335	0.0960

The community structures of the Node ID 349 personalized networks from Fast-Greedy, Edge-Betweenness, and Infomap community detection algorithm are shown below:

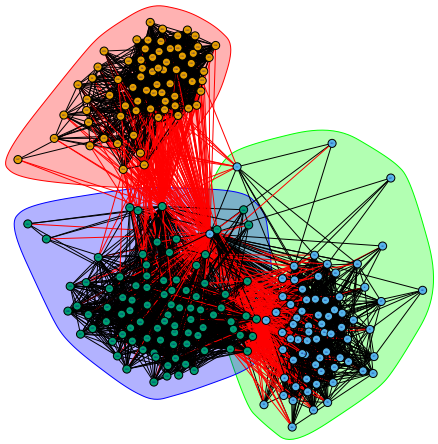


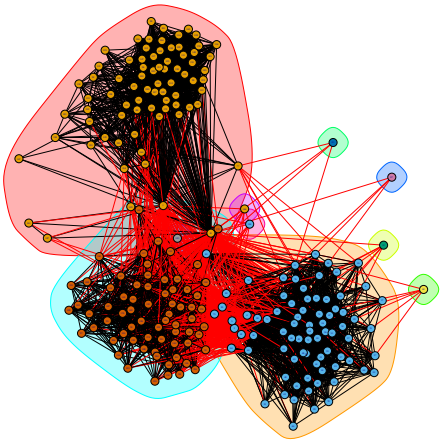


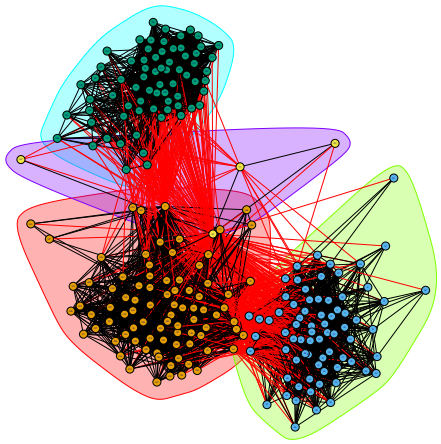
For Node ID 484, and the modularity scores are

Algorithm	Fast-Greedy	Edge-Betweenness	Infomap
Modularity score	0.5070	0.4890	0.5153

The community structures of the Node ID 484 personalized networks from Fast-Greedy, Edge-Betweenness, and Infomap community detection algorithm are shown below:



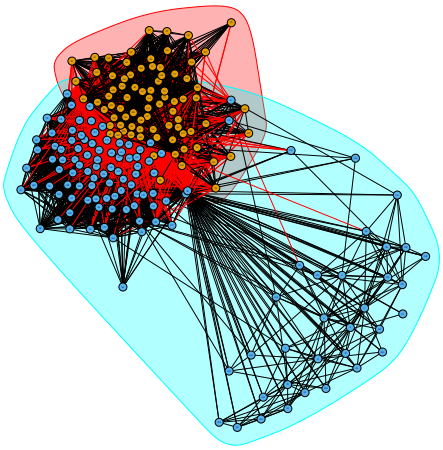


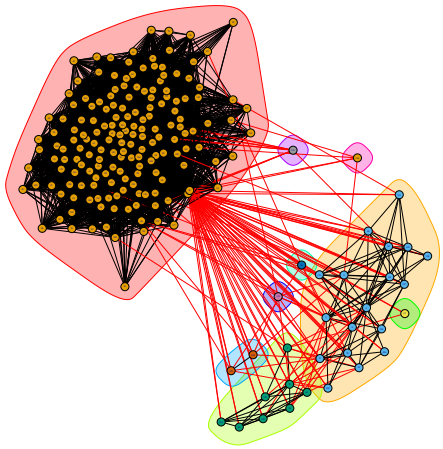


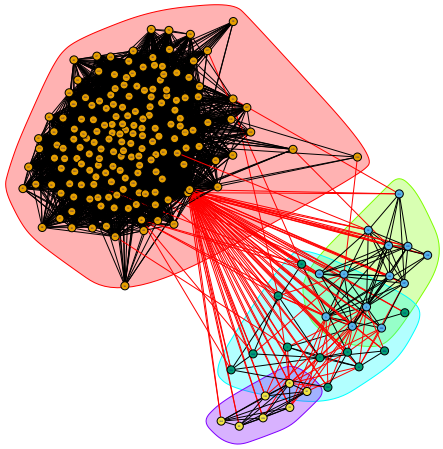
For Node ID 1087, and the modularity scores are

Algorithm	Fast-Greedy	Edge-Betweenness	Infomap
Modularity score	0.1455	0.0276	0.0269

The community structures of the Node ID 1087 personalized networks from Fast-Greedy, Edge-Betweenness, and Infomap community detection algorithm are shown below:







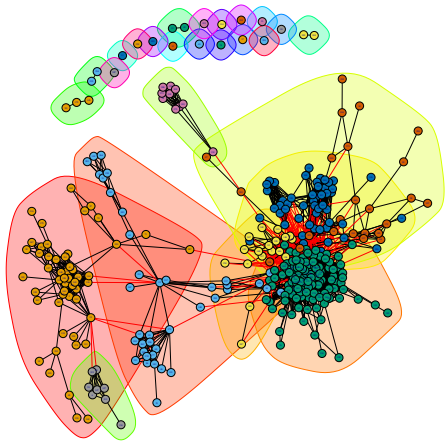
The community structures for each of the core node's personalized network are displayed using different colors. From the plots, it was observed that there are some overlaps between the detected communities and this reflects in the modularity scores which all have values less than 1. Furthermore, it can be seen from the plots that in general Edge-Betweenness tend to break graph into more communities than the other two algorithms and this results in the modularity scores computed from Edge-Betweenness to be lower.

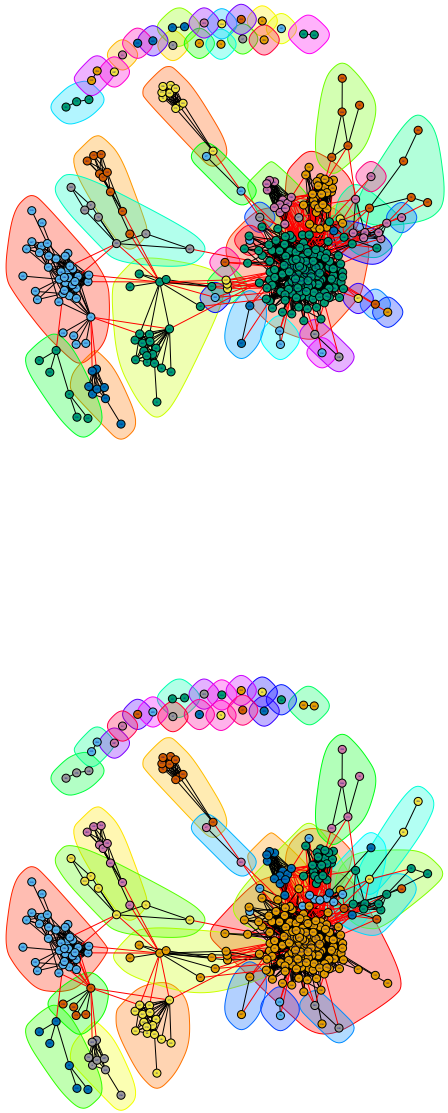
Question 10: For each of the core nodes personalized network(use same core nodes as question 9), remove the core node from the personalized network and find the community structure of the modified personalized network. Use the same community detection algorithm as question 9. Compare the modularity score of the community structure of the modi- fied personalized network with the modularity score of the community structure of the personalized network of question 9. For visualization purpose, display the community structure of the modified personalized network using colors. In this question, you should have 15 plots in total.

For Node ID 1, and the modularity scores are

Algorithm	Fast-Greedy	Edge-Betweenness	Infomap
Modularity score	0.4419	0.4161	0.4180

The community structures of the Node ID 1 modified personalized networks from Fast-Greedy, Edge-Betweenness, and Infomap community detection algorithm are shown below:

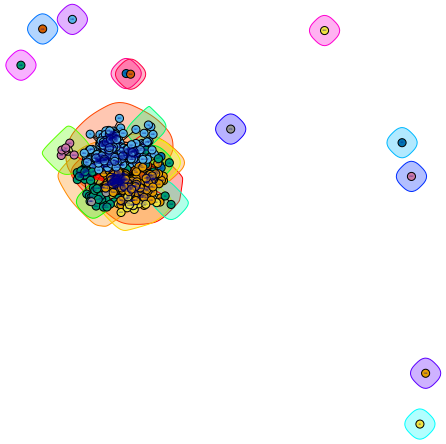


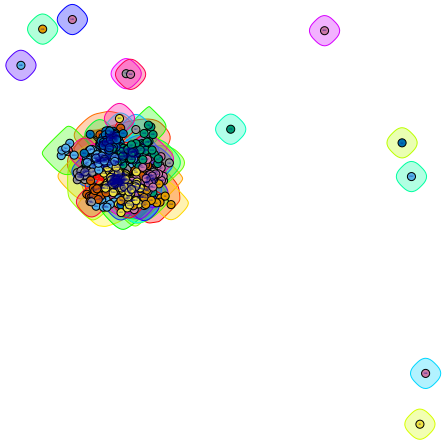


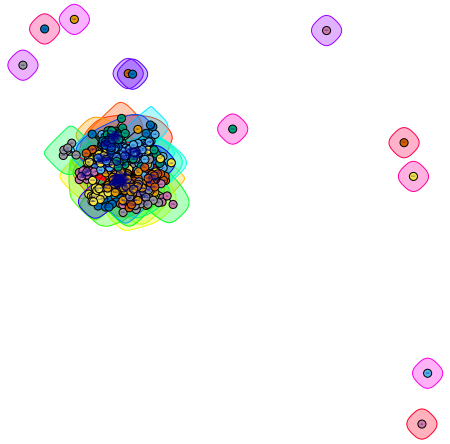
For Node ID 108, and the modularity scores are

Algorithm	Fast-Greedy	Edge-Betweenness	Infomap
Modularity score	0.4581	0.5213	0.5188

The community structures of the Node ID 108 modified personalized networks from Fast-Greedy, Edge-Betweenness, and Infomap community detection algorithm are shown below:



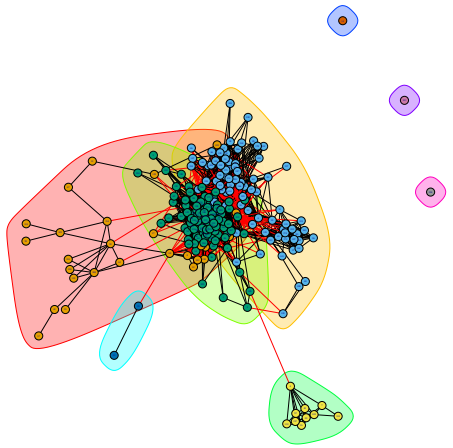


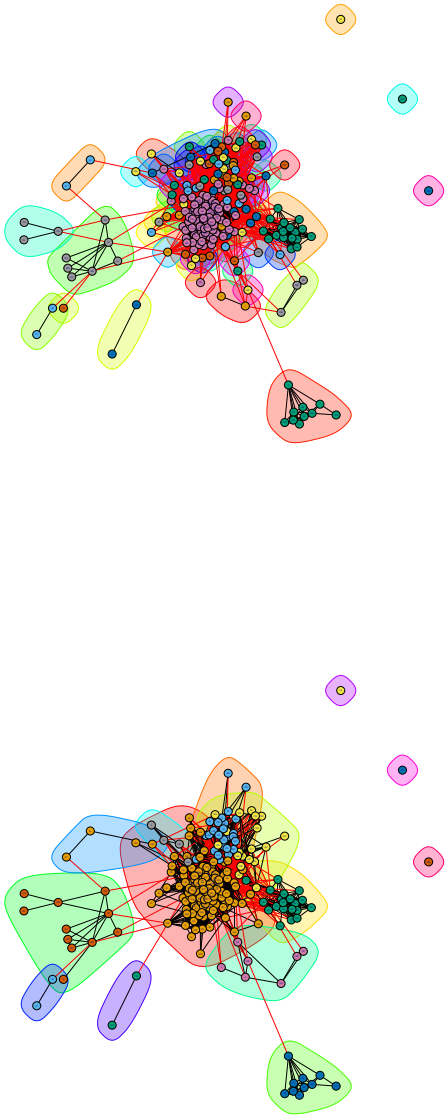


For Node ID 349, and the modularity scores are

Algorithm	Fast-Greedy	Edge-Betweenness	Infomap
Modularity score	0.2457	0.1506	0.2338

The community structures of the Node ID 349 modified personalized networks from Fast-Greedy, Edge-Betweenness, and Infomap community detection algorithm are shown below:

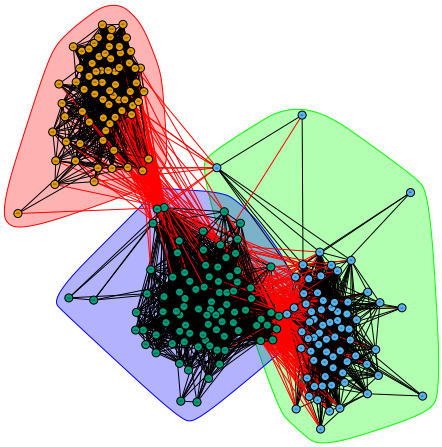


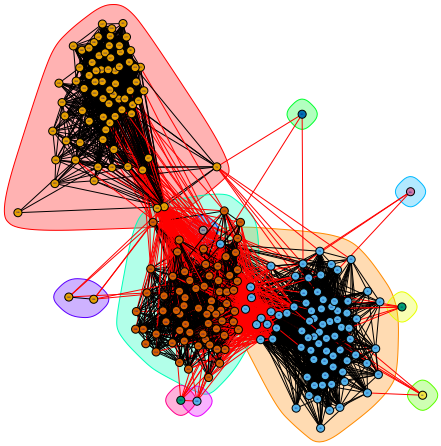


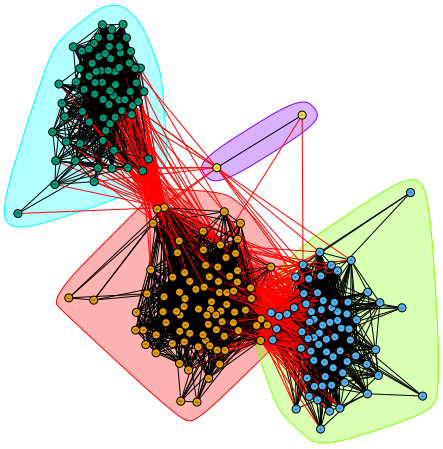
For Node ID 484, and the modularity scores are

Algorithm	Fast-Greedy	Edge-Betweenness	Infomap
Modularity score	0.5342	0.5154	0.5434

The community structures of the Node ID 484 modified personalized networks from Fast-Greedy, Edge-Betweenness, and Infomap community detection algorithm are shown below:



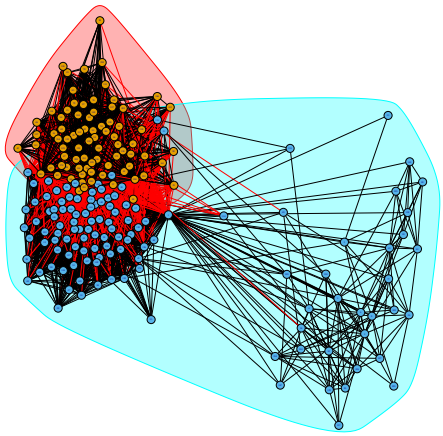


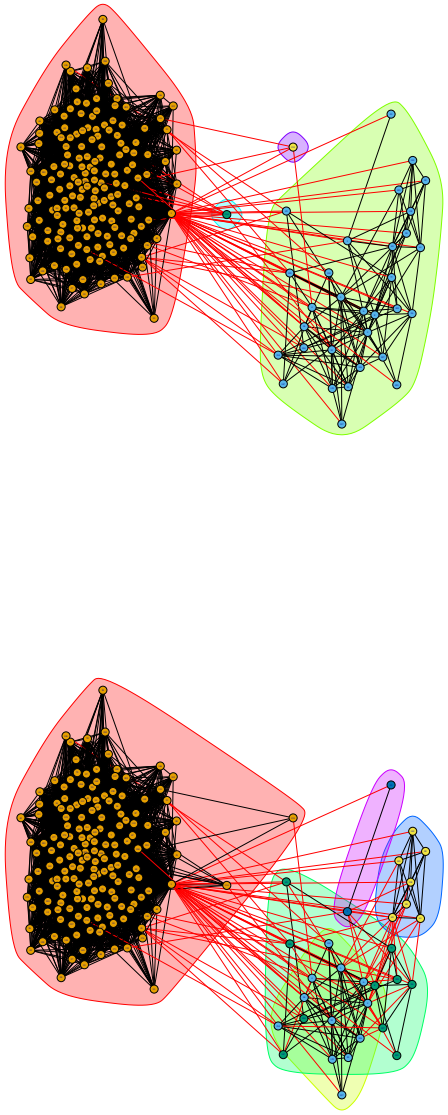


For Node ID 1087, and the modularity scores are

Algorithm	Fast-Greedy	Edge-Betweenness	Infomap
Modularity score	0.1482	0.0325	0.0274

The community structures of the Node ID 1087 modified personalized networks from Fast-Greedy, Edge-Betweenness, and Infomap community detection algorithm are shown below:





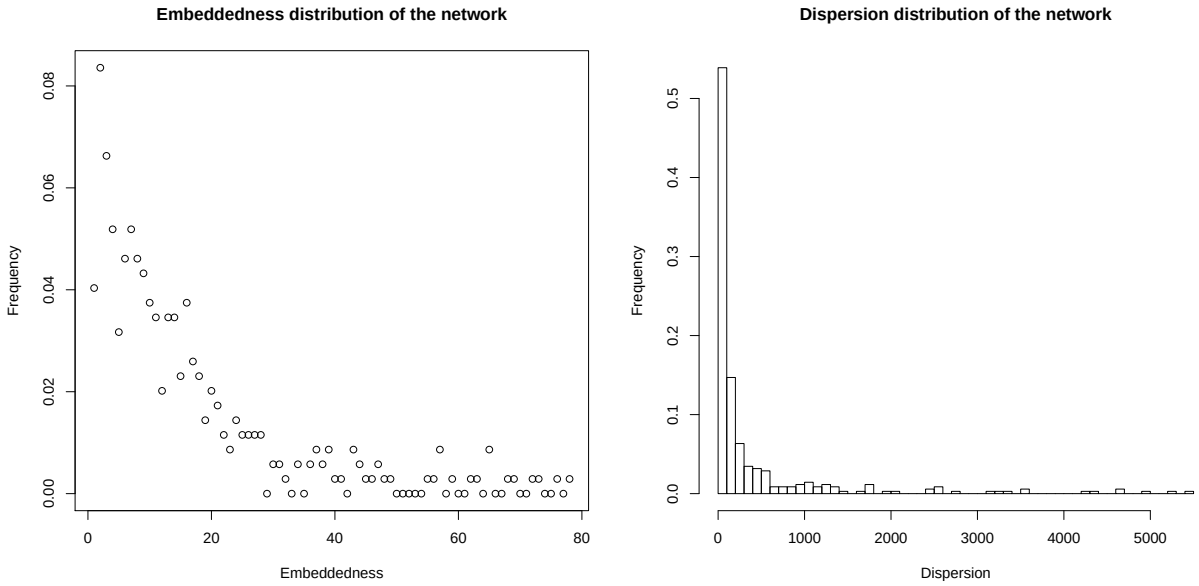
From the above plots, it can be observed that with the core nodes removed, there tends to be more isolated nodes. This is particularly evident in nodes with large number of neighbors, such as Node 108. On the other hand, for core nodes with high modularity scores, such as Node 484, the community structure with the core node remove tend to have similar appearance. Nonetheless, the modularity scores with and without the core nodes removed agree approximately within 10% and the modularity socres with the core node removed tend to be higher.

Question 11: Write an expression relating the Embeddedness of a node to its degree.

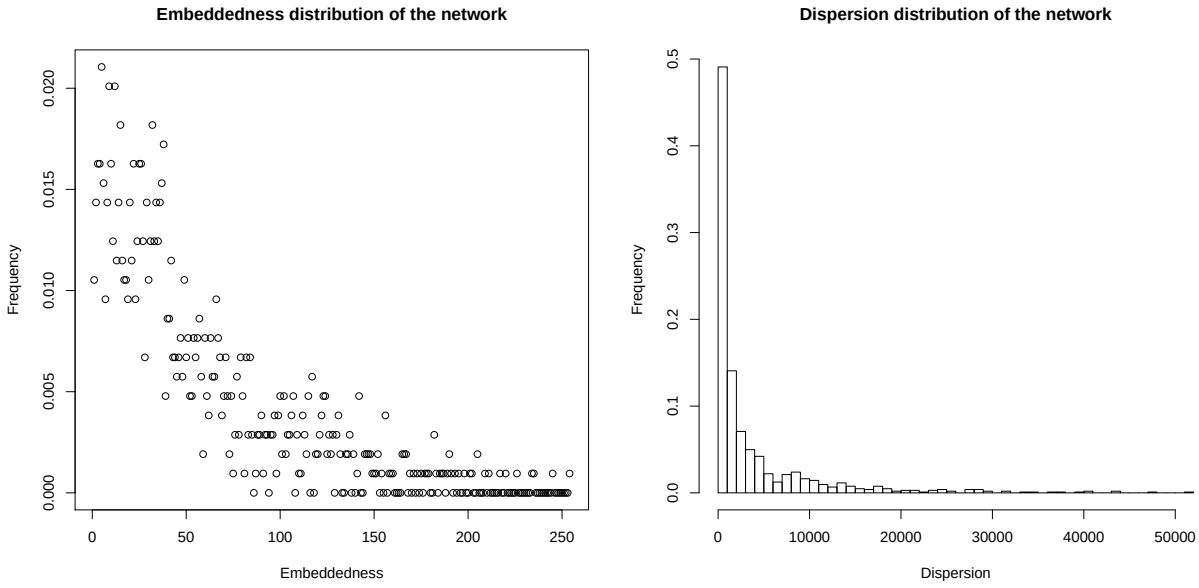
In a personalized network, the embeddedness of a node is its degree - 1. This is because the number of friends of a node is simply its degree and the number of mutual friends with the core node is all its friends excluding the core node.

Question 12: For each of the core nodes personalized network (use the same core nodes as question 9), plot the distribution of embeddedness and dispersion. In this question, you will have 10 plots.

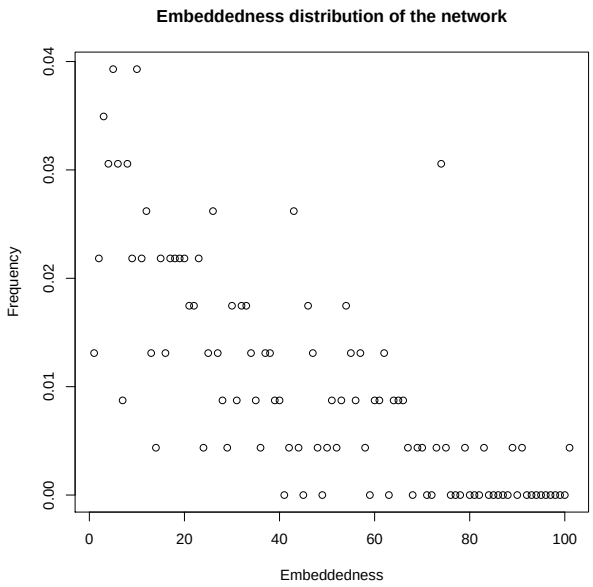
The embeddedness and dispersion of Node ID 1 are shown below

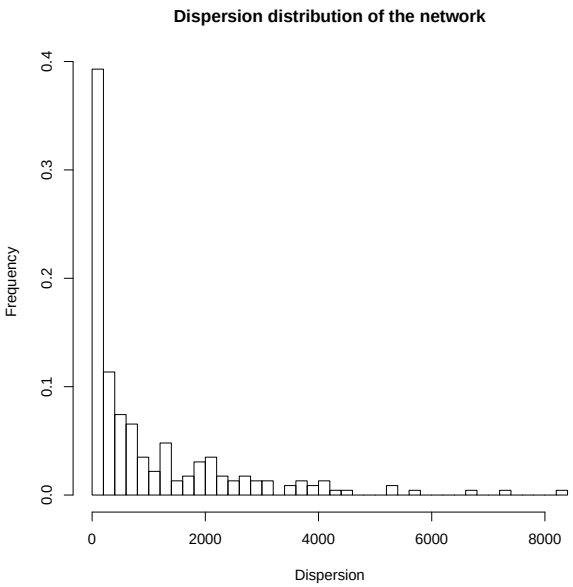


The embeddedness and dispersion of Node ID 108 are shown below

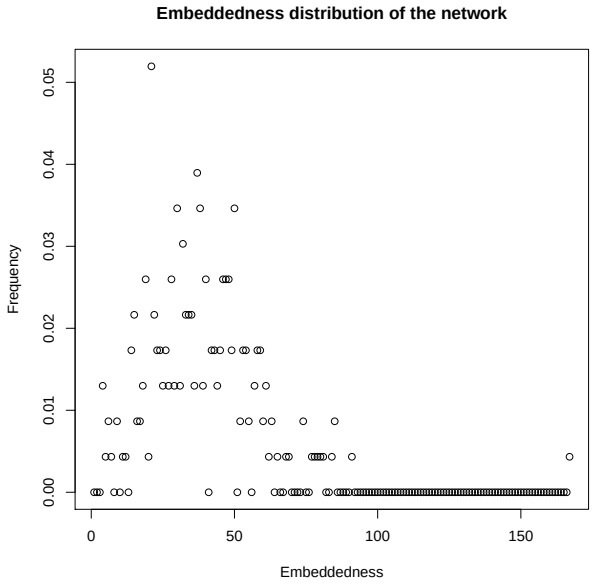


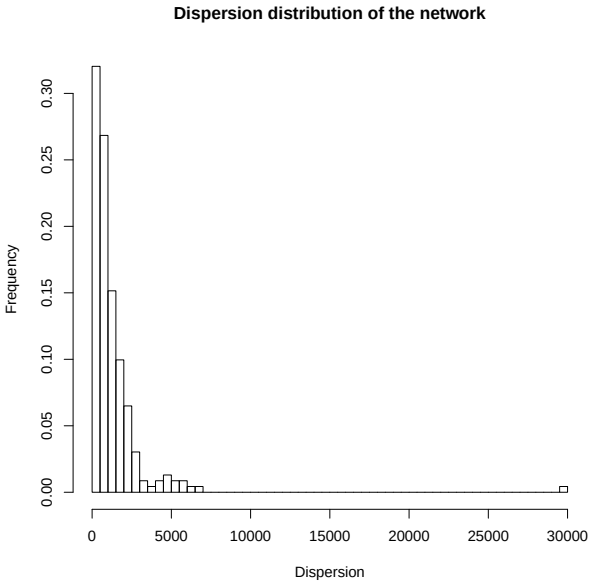
The embeddedness and dispersion of Node ID 1 are shown below



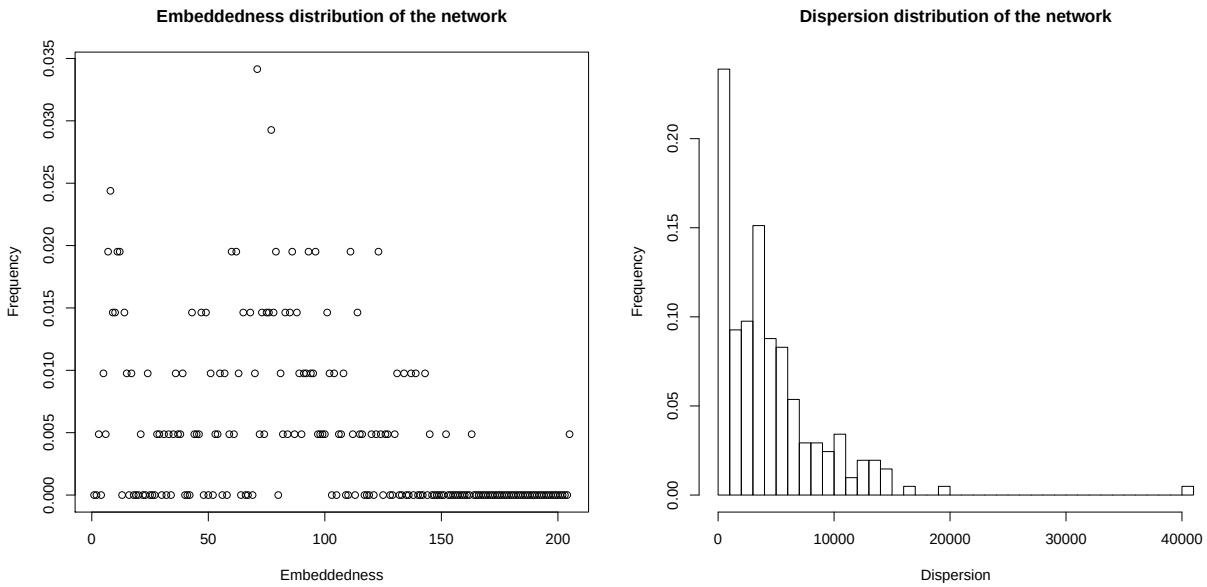


The embeddedness and dispersion of Node ID 484 are shown below





The embeddedness and dispersion of Node ID 1087 are shown below

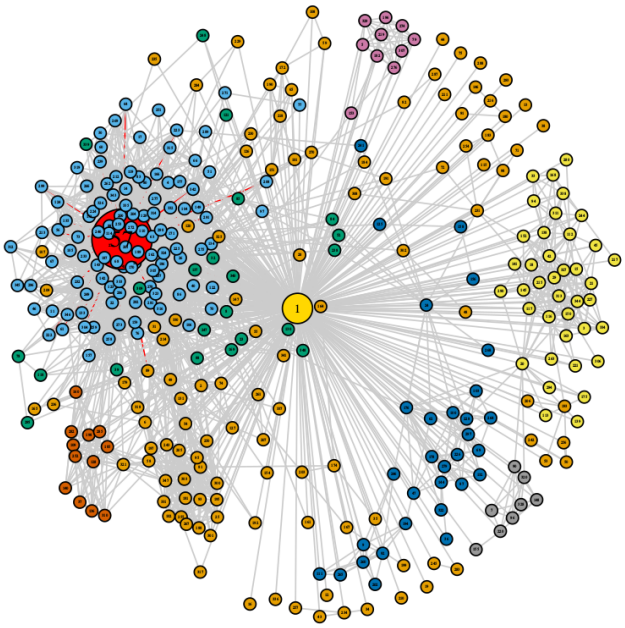


Question 13: For each of the core nodes personalized network, plot the community structure of the personalized network using colors and highlight the node with maximum dispersion. Also, highlight the edges incident to this node. To detect the community structure, use Fast-Greedy algorithm. In this question, you will have 5 plots.

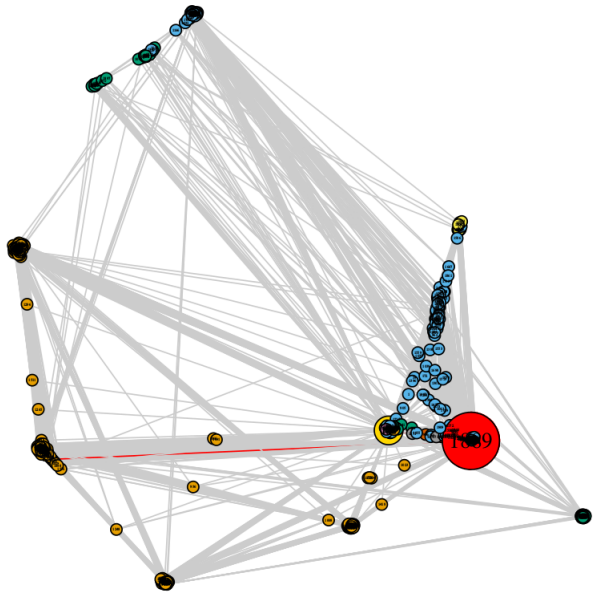
For nodes having zero or one mutual friend with the core node, we define their dispersion to be zero. Furthermore, a node having unconnected mutual friends is defined to have dispersion equals the diameter of the entire network plus one.

In the following graphs, we highlight the node with maximum dispersion by making its size five times greater than normal nodes and painting it red. We also highlight the incident edges by painting them red. Note that the core node is also highlighted by making its size 2.5 times larger than normal nodes and painting it gold.

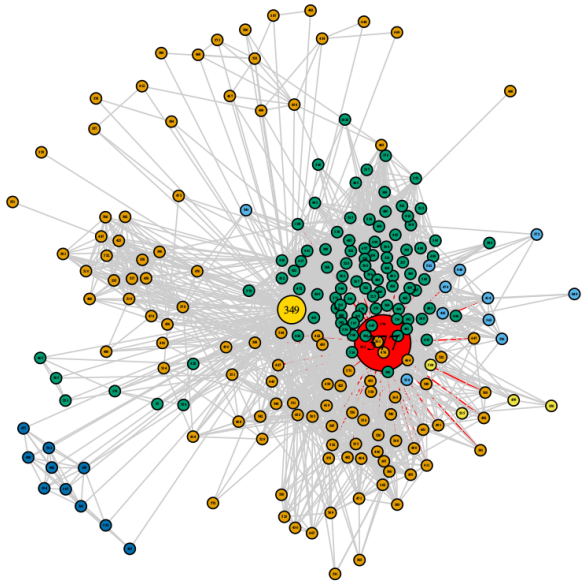
The community structure of Node ID 1's personalized network is shown below. Node ID 57 is the node with the maximum dispersion in this network.



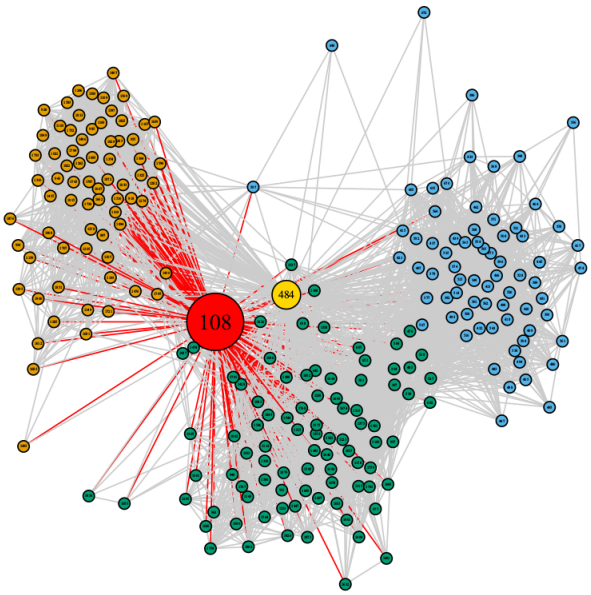
The community structure of Node ID 108's personalized network is shown below. Node ID 1889 is the node with the maximum dispersion in this network.



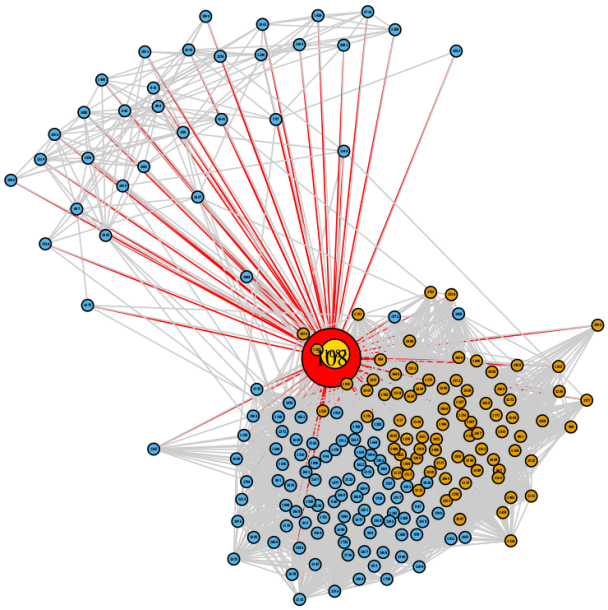
The community structure of Node ID 349's personalized network is shown below. Node ID 377 is the node with the maximum dispersion in this network.



The community structure of Node ID 484's personalized network is shown below. Node ID 108 is the node with the maximum dispersion in this network.



The community structure of Node ID 1087's personalized network is shown below. Node ID 108 is the node with the maximum dispersion in this network.

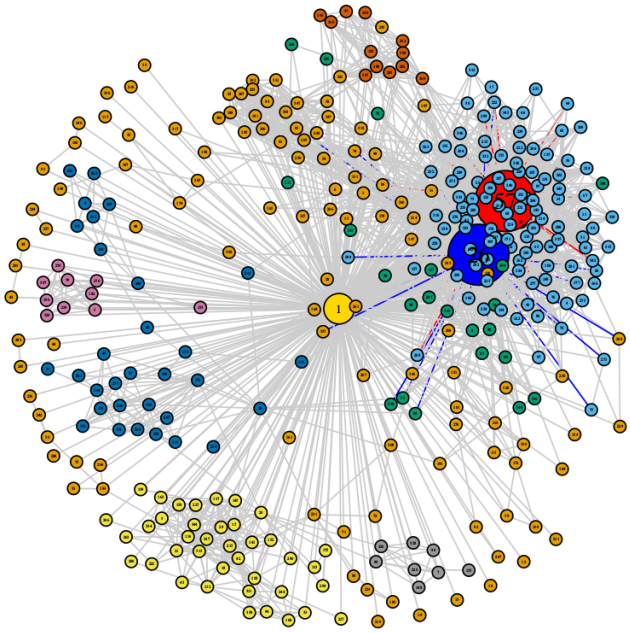


Question 14: Repeat question 13, but now highlight the node with maximum embeddedness and the node with maximum $\frac{dispersion}{embeddedness}$. Also, highlight the edges incident to these nodes.

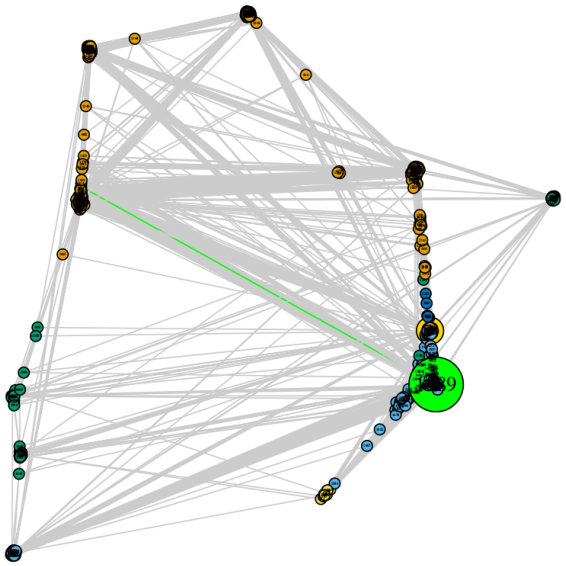
For nodes having no mutual friend with the core node, we define their $\frac{dispersion}{embeddedness}$ ratio to be zero.

In the following graphs, we highlight the node with maximum embeddedness by making its size five times greater than normal nodes and painting it red. We also highlight the incident edges by painting them red. On the other hand, we highlight the node with maximum $\frac{dispersion}{embeddedness}$ ratio by making its size five times greater than normal nodes and painting it blue. We also highlight the incident edges by painting them blue. However, once the node with maximum embeddedness is the same as the node with maximum $\frac{dispersion}{embeddedness}$ ratio, the node and its incident edges are painted green. Note that the core node is also highlighted by making its size 2.5 times larger than normal nodes and painting it gold.

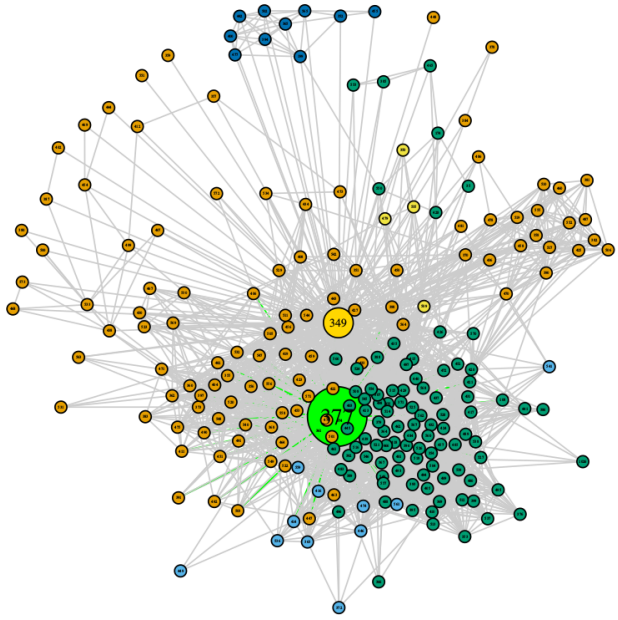
The community structure of Node ID 1's personalized network is shown below. Node ID 57 is the node with the maximum embeddedness in this network. Node ID 26 is the node with the maximum $\frac{dispersion}{embeddedness}$ ratio in this network.



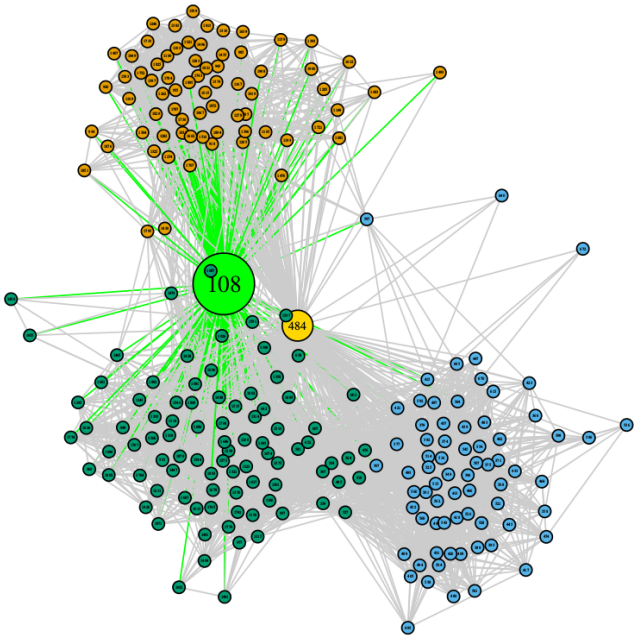
The community structure of Node ID 108's personalized network is shown below. Node ID 1889 is the node with both the maximum embeddedness and the maximum $\frac{dispersion}{embeddedness}$ ratio in this network.



The community structure of Node ID 349's personalized network is shown below. Node ID 377 is the node with both the maximum embeddedness and the maximum $\frac{dispersion}{embeddedness}$ ratio in this network.

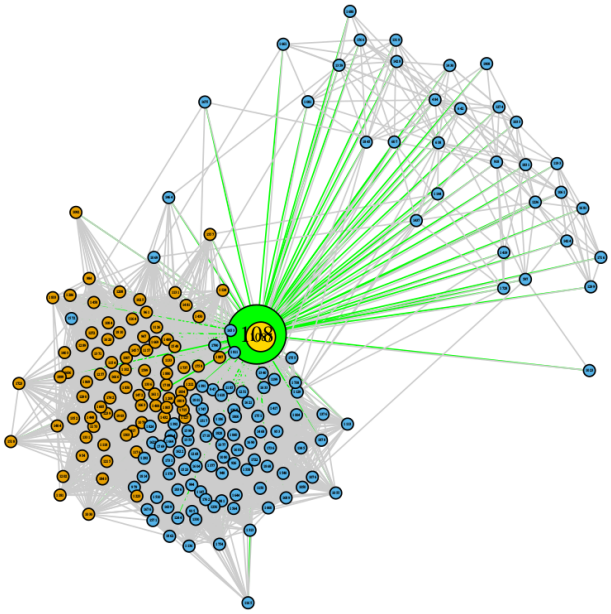


The community structure of Node ID 484's personalized network is shown below. Node ID 108 is the node with both the maximum embeddedness and the maximum $\frac{dispersion}{embeddedness}$ ratio in this network.



The community structure of Node ID 1087's personalized network is shown below. Node ID 108 is the node with both the maximum embeddedness and the maximum $\frac{dispersion}{embeddedness}$ ratio in this

network.



Question 15: Use the plots from questions 13 and 14 to explain the characteristics of a node revealed by each of this measure.

The characteristics of nodes revealed by the maximum dispersion, maximum embeddedness, and maximum $\frac{dispersion}{embeddedness}$ ratio measure are stated below.

It is good to know that the goal of the paper given is to recognize a person's romantic partner from the network structure given all the connections among a person's friends. Consequently, we will discuss the three measures step by step in the following and find out which one is more suitable to determine a person's romantic partner.

Embeddedness shows the number of mutual friends a node share with the core node, so we expect a node with high embeddedness to have edges connecting to the core node's neighbors. An obvious visualization lies in the last plot of question 14, in which Node ID 108 has lots of edges connecting to the core node, Node Id 1087's neighbors. When it comes to human social network, a node with high embeddedness may apply the person represented by the node is a lover, family member, roommate, club associate, religion associate, colleague, classmate, supervisor, or teacher of the person represented by the core node. There are a lot of possible situations in which two individuals have many mutual friends. Therefore, considering only embeddedness is apparently not enough to figure out the relationship between a node and the core node.

Dispersion indicates whether the mutual friends a node share with the core node know each other or not. A node with high dispersion is considered to live a life close to the core node because it has connections to the core node's different groups of friend. An obvious visualization lies in the fourth plot of question 13, in which Node ID 108 has edges connecting to the core node, Node Id 484's different clusters of neighbors. In human society, a node with high dispersion may apply the

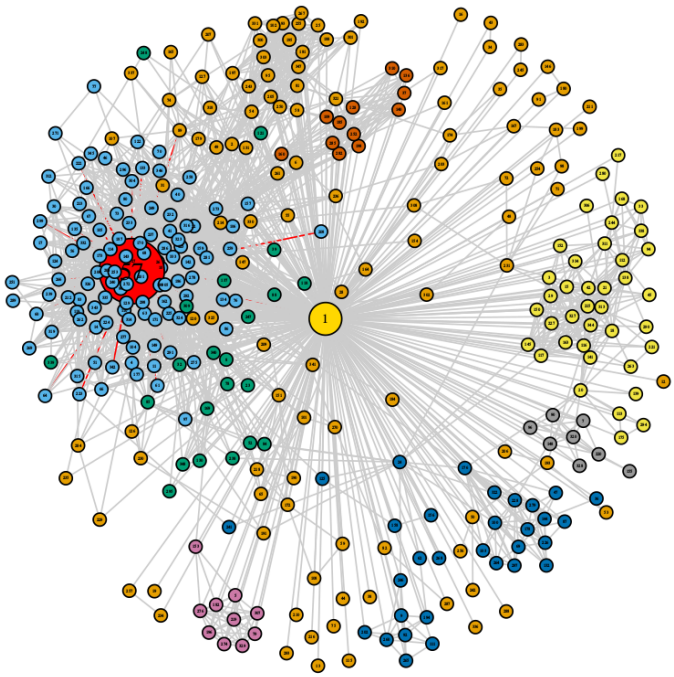
person represented by the node is a lover, family member or roommate who have close relationship with the person represented by the core node. Thus, there are still uncertainties regarding a node's relationship with the core node when just considering dispersion.

The ratio $\frac{dispersion}{embeddedness}$ implies the consideration of both dispersion and embeddedness of a node. We understand this measure by thinking of the fact that the lover of a person may not know about the person's friends that much. An obvious visualization lies in the first plot of question 14, in which Node ID 26 has edges connecting to the core node, Node Id 1's different groups of neighbors but the connections to each group are not dense. Back to the relationship network in human society, for example, Alice, the girlfriend of Bob may not know Bob's colleagues in the company. Likewise, Bob may not know Alice's classmates in the yoga class. Hence, $\frac{dispersion}{embeddedness}$ ratio provides a more reasonable way to determine the romantic relationship among people.

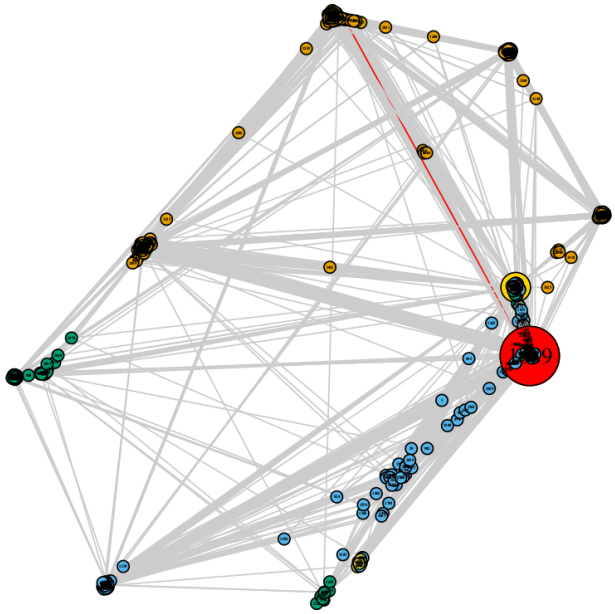
Besides the above three measures, we adopt the product of dispersion and embeddedness and conduct experiments to see the effect of this new measure. Because this new measure is proportional to both dispersion and embeddedness and we know from the above graphs that for each core node, the node with the highest dispersion and embeddedness is the same. Therefore, we expect to get the same nodes as in the dispersion plots or embeddedness plots.

In the following graphs, we highlight the node with maximum product of dispersion and embeddedness by making its size five times greater than normal nodes and painting it red. We also highlight the incident edges by painting them red. Note that the core node is also highlighted by making its size 2.5 times larger than normal nodes and painting it gold.

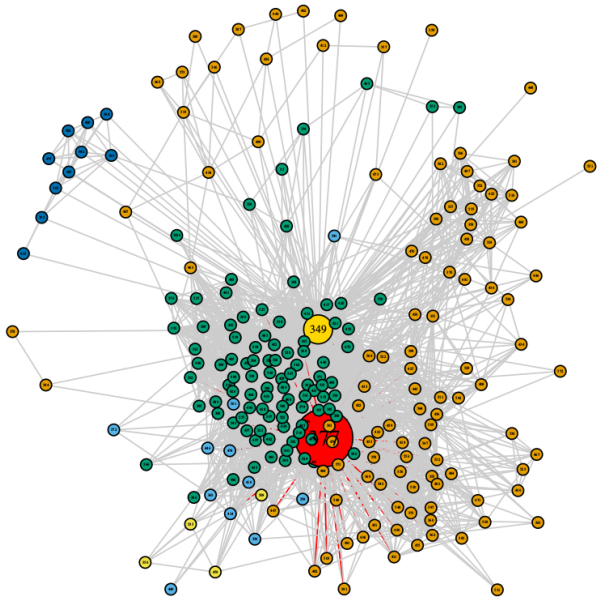
The community structure of Node ID 1's personalized network is shown below. Node ID 57 is the node with the maximum product of dispersion and embeddedness in this network.



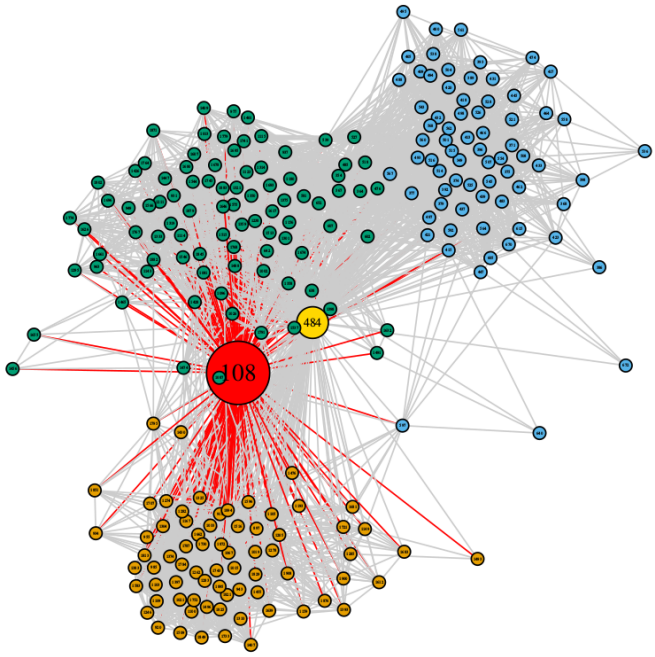
The community structure of Node ID 108's personalized network is shown below. Node ID 1889 is the node with the maximum product of dispersion and embeddedness in this network.



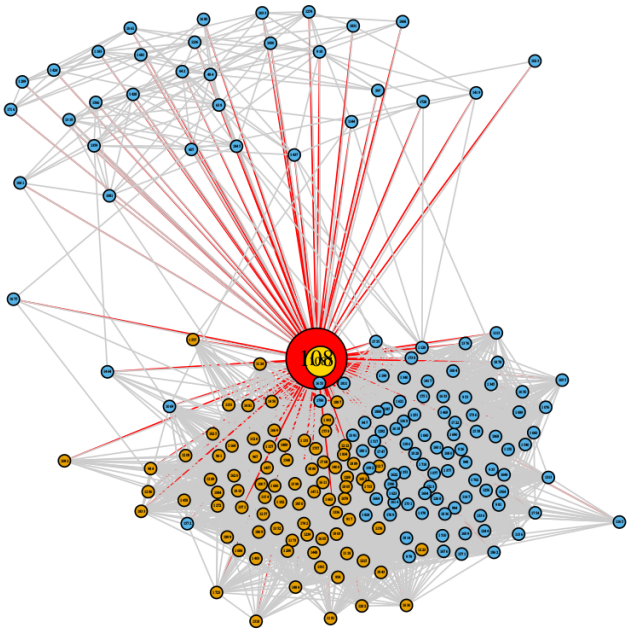
The community structure of Node ID 349's personalized network is shown below. Node ID 377 is the node with the maximum product of dispersion and embeddedness in this network.



The community structure of Node ID 484's personalized network is shown below. Node ID 108 is the node with the maximum product of dispersion and embeddedness in this network.



The community structure of Node ID 1087's personalized network is shown below. Node ID 108 is the node with the maximum product of dispersion and embeddedness in this network.



In overall, the results correspond to our expectation and we get the same nodes as in the dispersion plots or embeddedness plots.

1.4 Friend recommendation in personalized networks

Question 16: What is $|N_r|$?

$|N_r|$ is the number of users to whom we are going to recommend friends.

Question 17: Compute the average accuracy of the friend recommendation algorithm that uses Common Neighbors measure, Jaccard measure, and Adamic Adar measure separately. Based on the average accuracy values, which friend recommendation algorithm is the best?

The friend recommendation algorithms using the Common Neighbors measure, Jaccard measure, and Adamic Adar measure separately achieve 0.826, 0.812, and 0.835 on the average accuracy accordingly. Apparently, the Adamic Adar measure achieves the best accuracy among the three in this case.

The concept behind the Common neighbor measure is that nodes in the same neighborhood should share many mutual friends. This intuition is straightforward and it has moderate performance in this case. The Jaccard measure takes the number of neighbors of the user node and the recommended node into consideration but it does not perform well in this case. The Adamic Adar measure considers not only the number of mutual friends but also the neighborhood structures of these mutual friends, and it is shown to be the best among the three measures in this case.

2 Google+ network

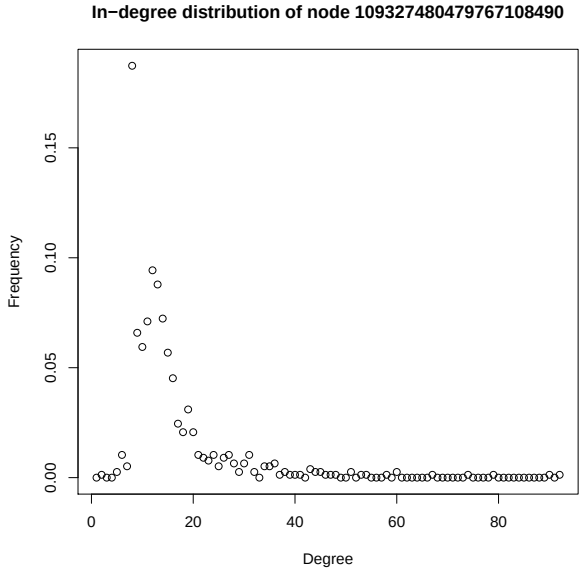
Question 18: How many personal networks are there?

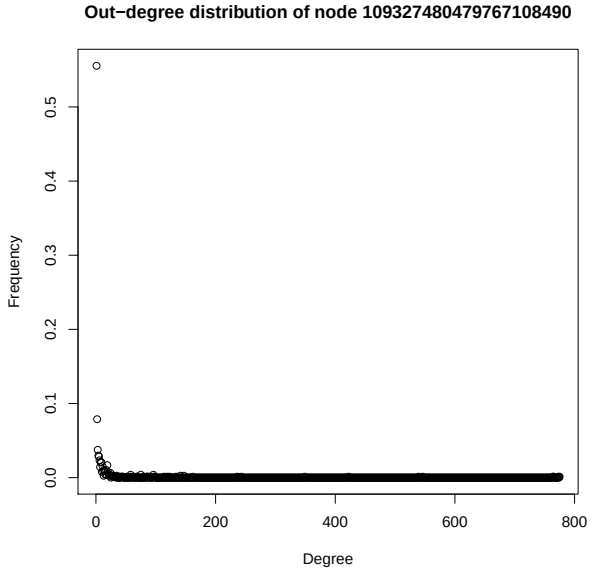
There are **57** personal networks.
Personal networks are created only for users who have more than 2 circles. Here we can read all the *.circles* files. If there are more than 2 circles in the files, we increase the count of personal networks. Note that if the *.circles* is empty, the *read.table* function will give us error so we have to skip those files which file size equals to zero.

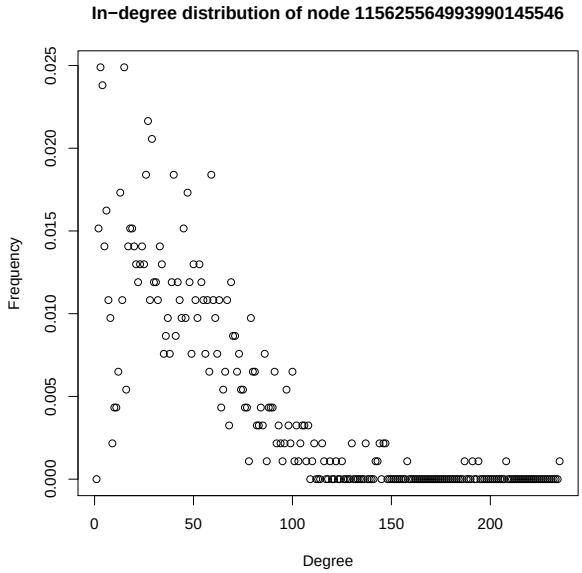
Question 19: For the 3 personal networks, plot the in-degree distribution of these personal networks. Do the personal networks have a similar in and out degree distribution?

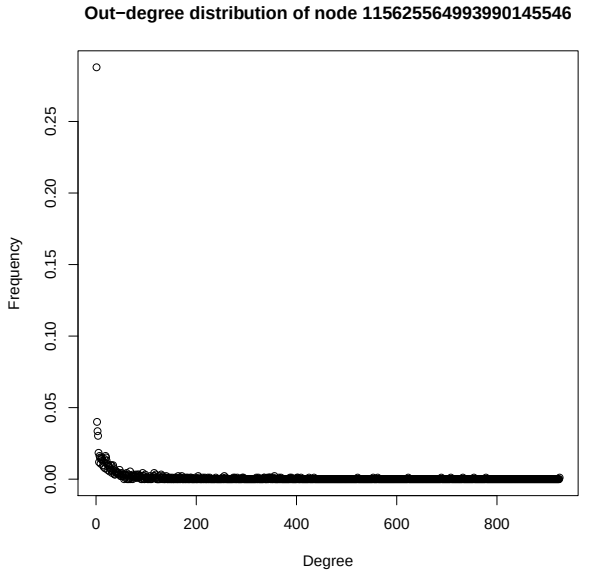
As you can see from the figures below. The personal networks of three different users have similar out-degree distribution but the in-degree distribution is quiet different.

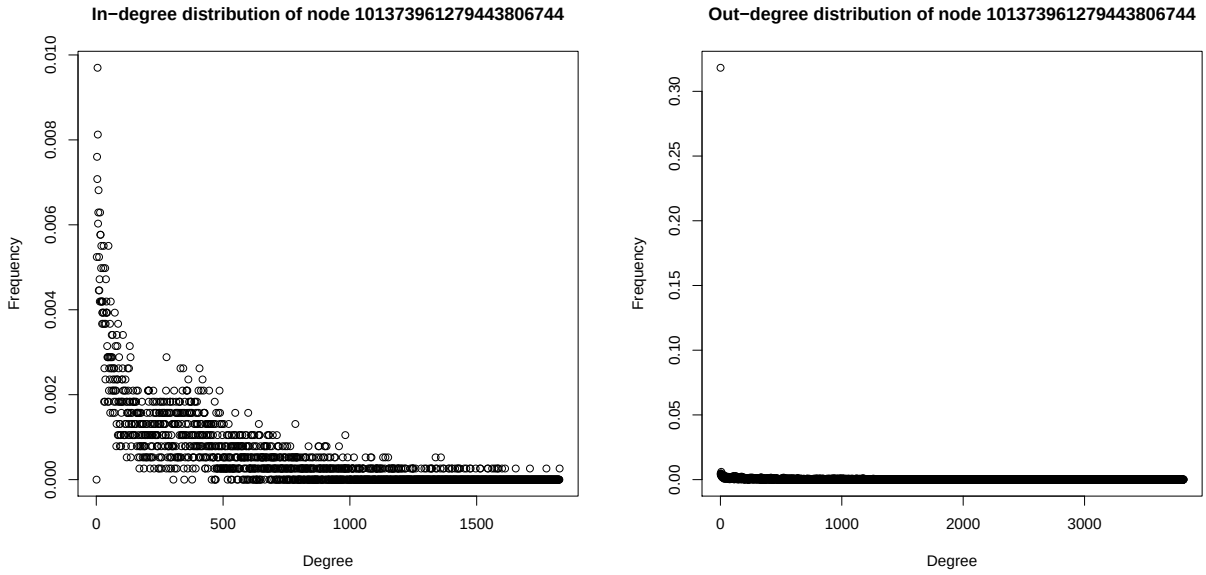
This is reasonable. To illustrate, people in real-life usually follow their friends and a few celebrities or athletes they like, and the difference between the following number is not too much. On the other hand, if a user follow a lot of celebrities, and those celebrities may have hundred to ten thousand or more follower (in-degree), which can cause the difference in the distribution.


In−degree distribution of node 109327480479767108490
Frequency
0.00
0.05
0.10
0.15
0
20
40
60
80
Degree


Out−degree distribution of node 109327480479767108490
Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0
200
400
600
800
Degree


In−degree distribution of node 115625564993990145546
Frequency
0.000
0.005
0.010
0.015
0.020
0.025
0
50
100
150
200
Degree


Out−degree distribution of node 115625564993990145546
Frequency
0.00
0.05
0.10
0.15
0.20
0.25
0
200
400
600
800
Degree

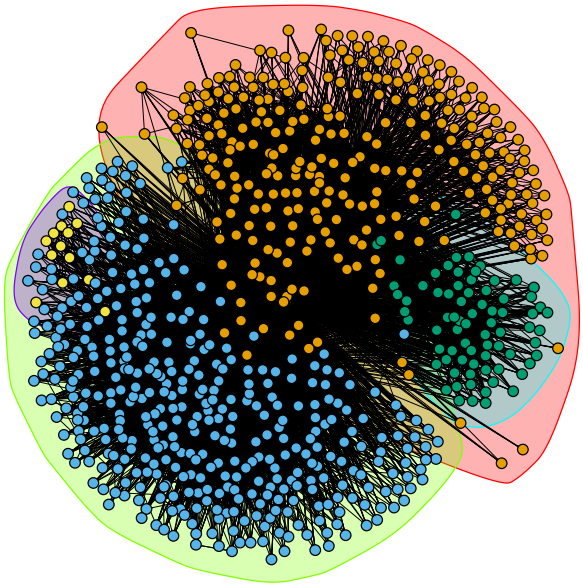


2.1 Community structure of personal networks

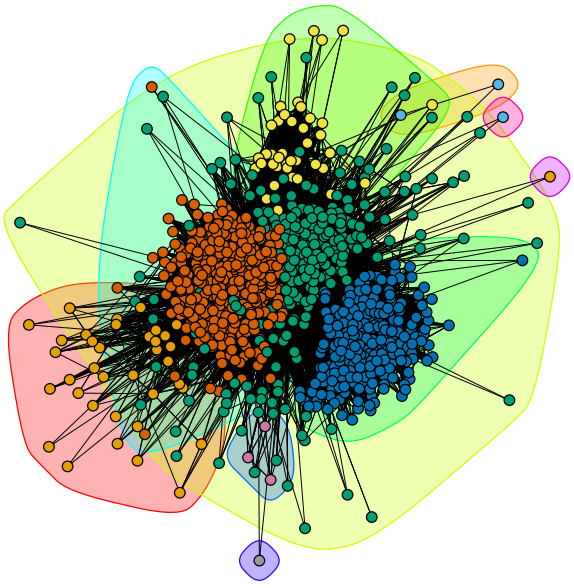
Question 20: For the 3 personal networks picked in question 19, extract the community structure of each personal network using Walktrap community detection algorithm. Report the modularity scores and plot the communities using colors. Are the modularity scores similar?

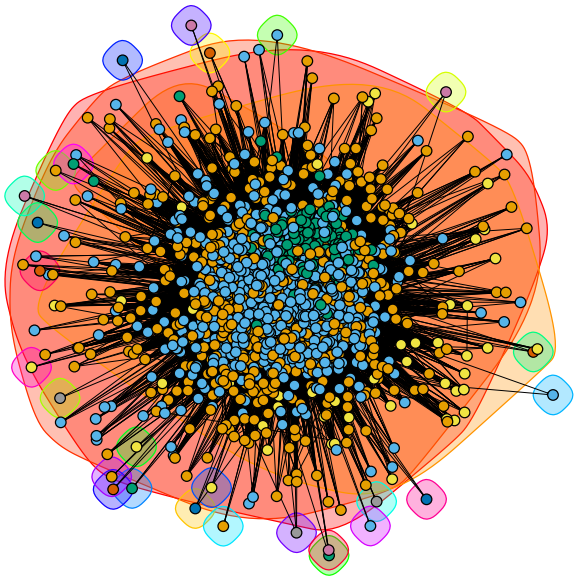
node 1's modularity: 0.252787130836454
node 2's modularity: 0.319486693112148
node 3's modularity: 0.191093274756585

communities of node 109327480479767108490



communities of node 115625564993990145546





It seems that they have similar modularity scores.
That is, all of these personal networks have dense connections between the nodes within the same community but sparse connections between different communities.

Question 21: Based on the expression for h and c, explain the meaning of homogeneity and completeness in words.

First we define *circle* as *class*, and *community* as *cluster*

$H(C|K)$ is maximal (and equals to $H(C)$) when the clustering provides no new information: the class distribution within each cluster equals to the overall class distribution, which causes h equals to zero. $H(C|K)$ is 0 when each cluster contains only members of a single class.
So, the meaning of homogeneity is how pure that each cluster is.

On the other hand, $H(K|C)$ is maximal (and equals to $H(K)$) if each class is represented by every cluster with distribution equal to the distribution of cluster sizes.
That is, in order to get higher completeness score, a clustering must assign all of those data points that are members of a single class to a single cluster.

Question 22: Compute the h and c values for the community structures of the 3 personal networks. Interpret the values and provide a detailed explanation.

Theoretically the value of homogeneity and completeness is between 0 and 1. However, according to the equations in project statement, the probability is not normalized to one: $\sum_{i=1}^{|C|} \frac{a_i}{N} > 1$, since the circles may overlap.

As the reason you can see that here we get negative homogeneity and completeness on node 2 and 3
node 1 homogeneity 0.840701754665236
node 1 completeness 0.143261916873428

node 2 homogeneity 0.565034707846679
node 2 completeness -6.11510322369768

node 3 homogeneity -0.365477515770206
node 3 completeness -1.74250194791754

If we use real-life example, the higher homogeneity means that people in a same cluster tend to have same features (e.g. same gender or same interests.)
The higher completeness means that people have same features form only one huge cluster.