ECE 232E
Large Scale Data Mining: Models and Algorithms

Project 2
Social Network Mining

Haitao Wang (UID: 504294402)
Xiao Peng(UID:005033608)
Zhao Weng(UID:304946606)

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.

In this question, we are supposed to read the data from the file and tell if it is connected. We use "read.table" function to read the facebook network data and generate graph with the data. As a result, the graph generated by the facebook network is connected.

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

The diameter is 8 for this graph.

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

The degree distribution of the graph can be obtained easily and Figure 3.1 is the plot of it. And the average of the degree of this network is 43.69.

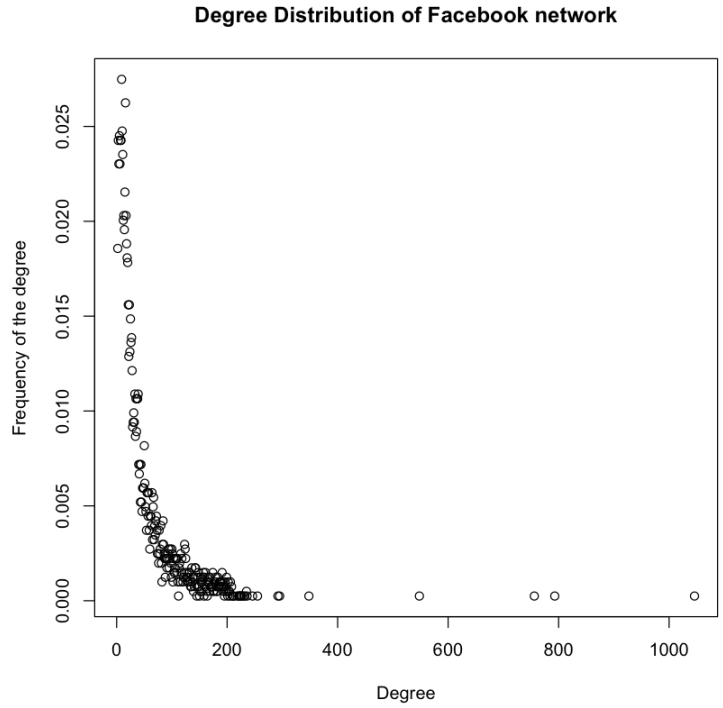


Figure 3.1 Degree distribution of Facebook social network

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.

Since there isn't much useful information in Figure 3.1, we consider to apply some other scale transform on the graph, such as the log-log scale. The degree distribution in the log-log scale is shown below in Figure 4.1. We can observe that some linear property starts to show in the plot. And we use "lm" function to get the slope of the fitting line, which is around -1.18 as shown in Figure 4.2.

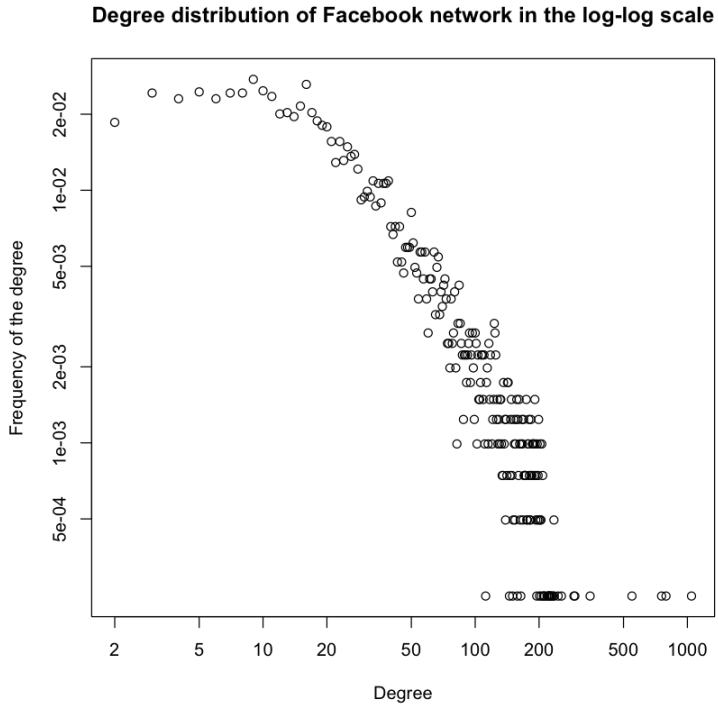


Figure 4.1 Degree distribution of Facebook social network in the log-log scale

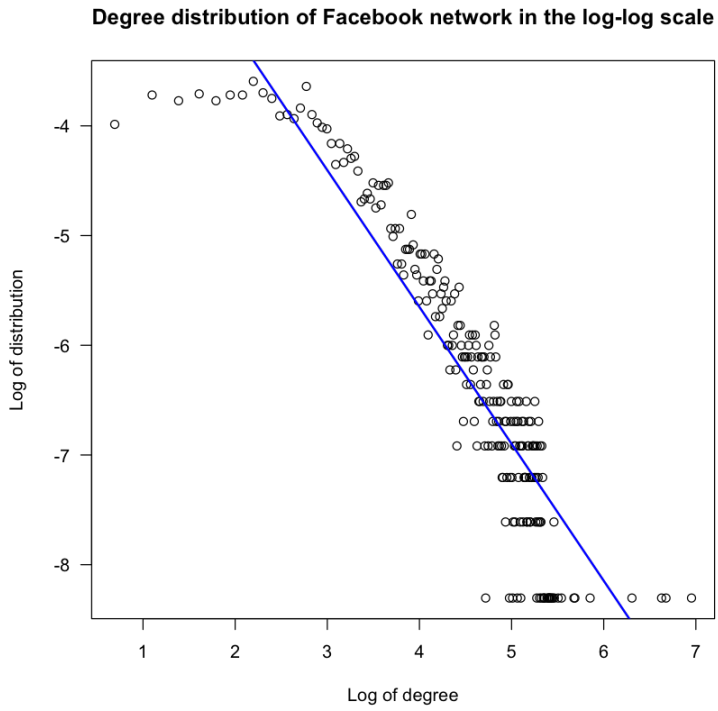


Figure 4.2 Degree distribution in the log-log scale fitted in a line

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?

According to the ID transformation rule, we know that the actual ID is 0 in the graph. We use the "ego" function to list all the neighbors of the vertex 0, then generate the subgraph with

"induced_subgraph" function. In Figure 5.1, we plot the subgraph of personalized network of vertex 0. As the results, the edges and the vertices in this subgraph are 2866 and 348, respectively.

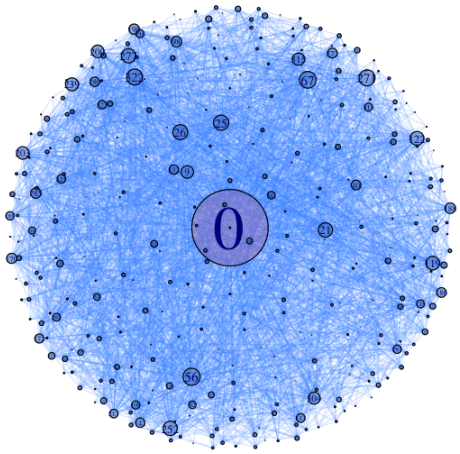


Figure 5.1 The subgraph of personalized network for vertex 0

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 the longest distance among the shortest path of every two nodes in the graph. Since the personalized network is centralized by one nodes, there only two possibilities for choosing two nodes, one center node and one neighbor or two neighbors. The former has fixed distance 1 due to the definition of the personalized network, however for the same reason, the latter choice can have the distance of 2. In addition, there are some exceptions for the normal cases, such as the personalized network of a certain node can have no neighbor, or only one neighbor. In these two exceptions, the distance between every two nodes can be 0.

Generally, we have the trivial upper and lower bound for the diameter of the personalized network are 2 and 0, respectively. In this question, since the graph is connected, the lower bound is 1.

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?

In the last question, we have derive the upper and lower bound for the diameter of the personalized network. When the diameter is 2, it means the personalized network has more than 2 nodes, in another words, the center has more than one neighbor. And the diameter is 1 for facebook social network, it means that the center has only one neighbor.

1.3 Core node's personalized network

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

degree of the core nodes?

For every node in the graph, we calculate the degree of it. And for all the nodes that have more than 200 neighbors, we call them as core nodes. Counting the number of the core nodes and calculating the average degree of all the core nodes, the results are 40 and 279.375, respectively.

Question 9 For each of the above core node's 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 node's 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.

We are supposed to plot the community structure for 5 nodes, every nodes with three different community detection algorithms. Firstly, we generate the subgraph for the nodes with "induced_subgraph" function. Then apply Fast-Greedy, Edge-Betweenness, and Infomap these three community detection algorithms to the subgraphs of these 5 nodes. In Figure 9.1 – 9.5, we plot the community structure with same color for same community. As a measurement, we also record modularity score for every nodes with every algorithms, listed as follow in Table 9.1

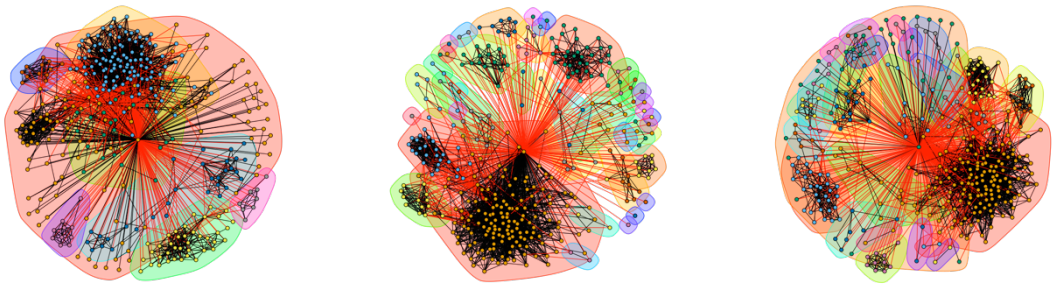


Figure 9.1 Community structure using Fast-Greedy, Edge-Betweenness, and Infomap algorithms for node 1

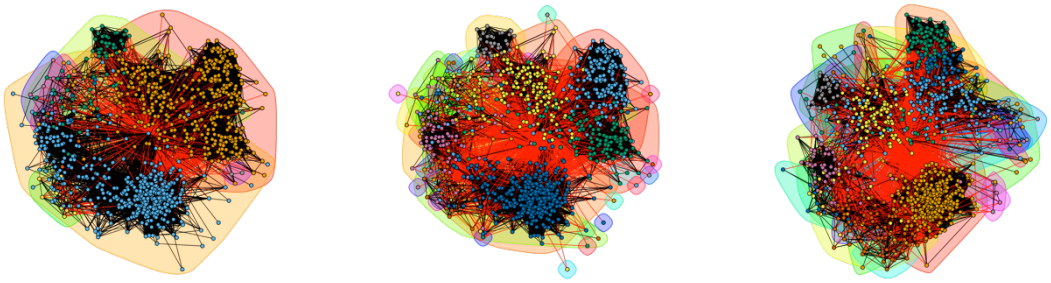


Figure 9.2 Community structure using Fast-Greedy, Edge-Betweenness, and Infomap algorithms for node 108

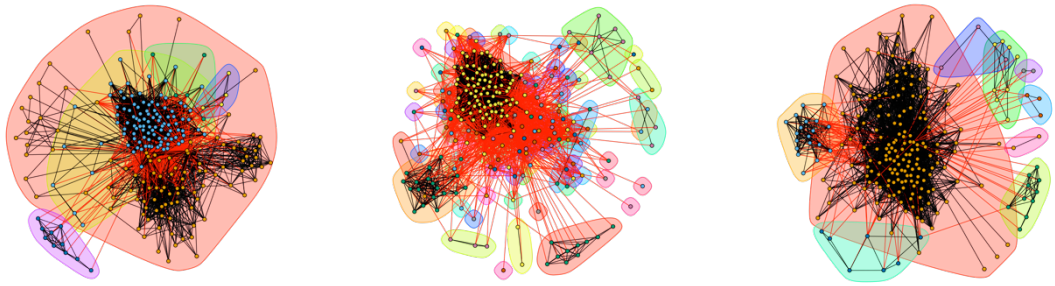


Figure 9.3 Community structure using Fast-Greedy, Edge-Betweenness, and Infomap algorithms for node 349

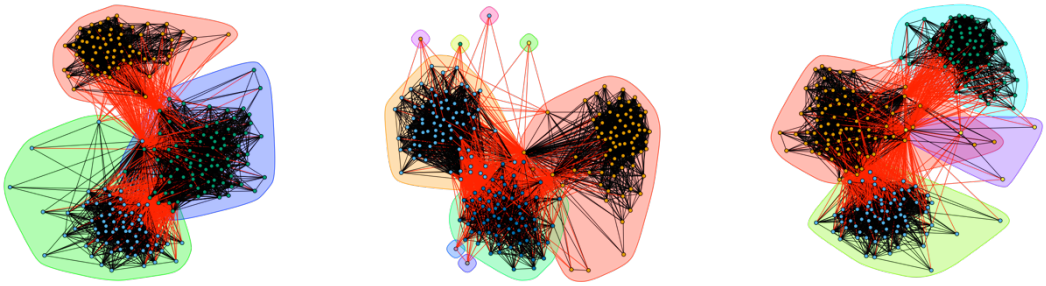


Figure 9.4 Community structure using Fast-Greedy, Edge-Betweenness, and Infomap algorithms for node 484

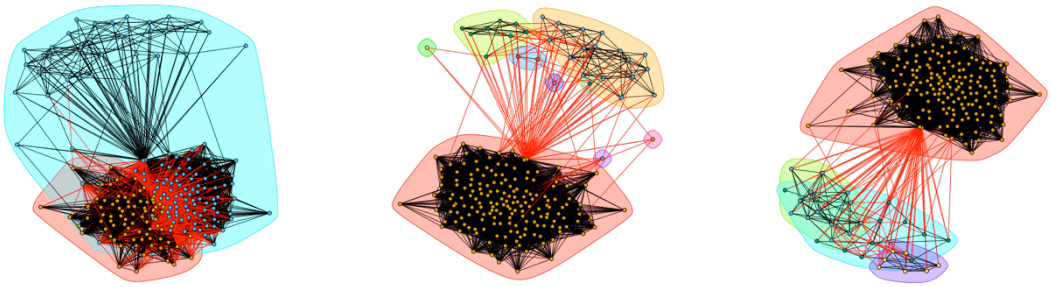


Figure 9.5 Community structure using Fast-Greedy, Edge-Betweenness, and Infomap algorithms for node 1047

Node ID	Fast-Greedy	Edge-Betweenness	Infomap
• **Node ID 1**	0.413	0.353	0.389
• **Node ID 108**	0.436	0.507	0.508
• **Node ID 349**	0.250	0.134	0.095
• **Node ID 484**	0.507	0.489	0.515
• **Node ID 1087**	0.146	0.028	0.027

Table 9.1 Modularity scores of nodes with different community detection algorithms

Question 10 For each of the core node's 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 9.4 Compare the modularity score of the community structure of the modified 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.

The plots are shown in Figure 9.

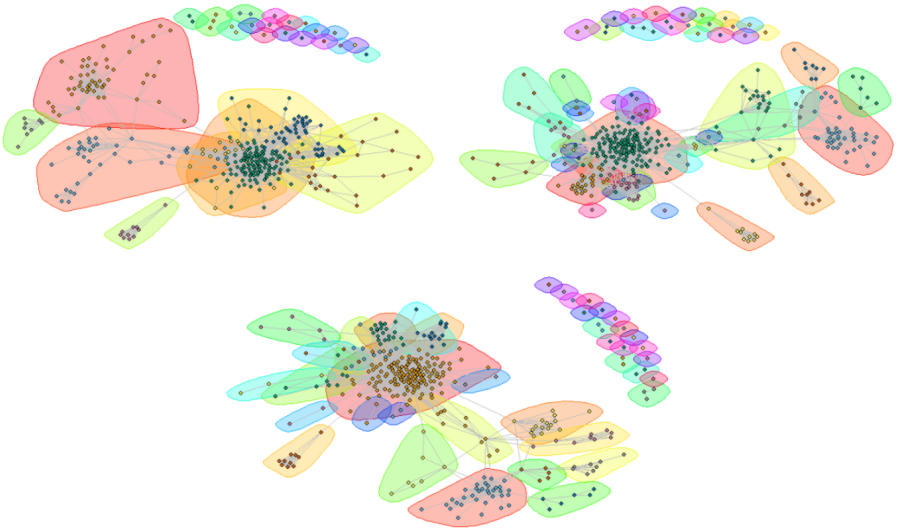


Figure 9.1: Community structure of node 1

Modularity of core node 1's personalized network:

Fast	Greedy:	0.4418533
Edge	Betweenness:	0.4161461

Infomap: 0.4180077

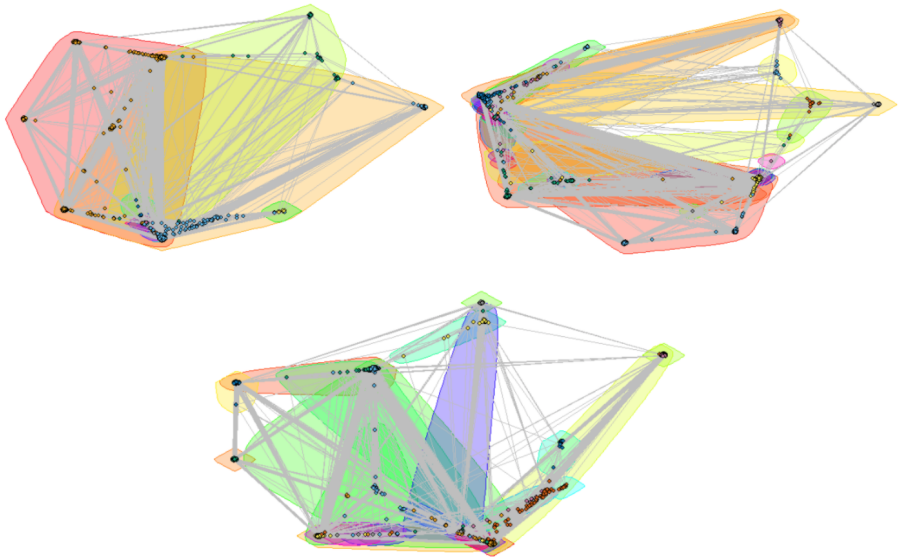


Figure 9.2: Community structure of node 108

Modularity of core node 108's personalized network:

Fast	Greedy:	0.4359284
Edge	Betweenness:	0.5067549

Infomap: 0.5082493

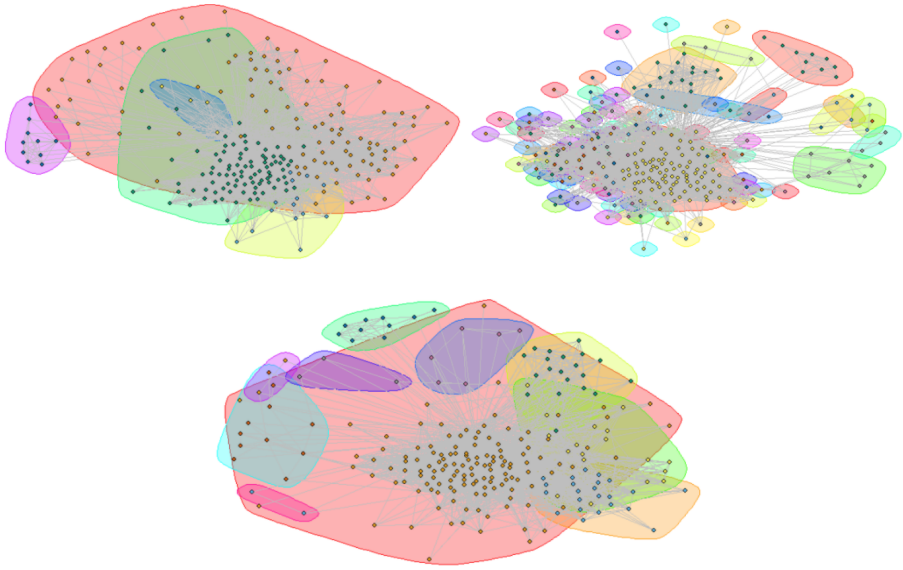


Figure 9.3: Community structure of node 349

Modularity of core node 349's personalized network:

Fast	Greedy:	0.2456918
Edge	Betweenness:	0.1505663
Infomap:0.2377727		

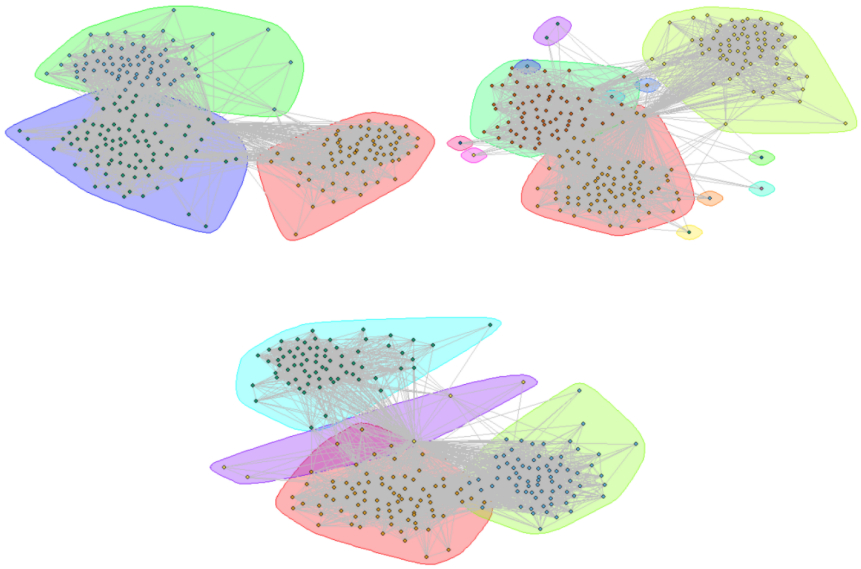


Figure 9.4: Community structure of node 484

Modularity of core node 484's personalized network:

Fast	Greedy:	0.52114539
Edge	Betweenness:	0.50838308
Infomap:0.53148912		

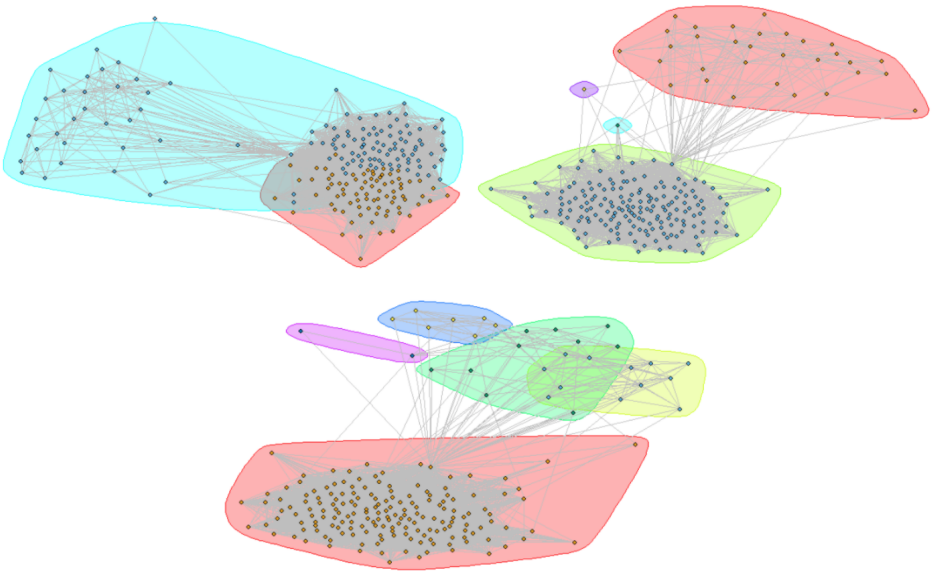


Figure 9.5: Community structure of node 1087

Modularity of core node 1087's personalized network:

Fast Greedy: 0.1477131
Edge Betweenness: 0.032194672
Infomap: 0.02736246

Question 11 Write an expression relating the Embeddedness of a node to it's degree.

Since the embeddedness of the node is defined as the number of mutual friends a node shares with the core node. Embeddedness of a node < it's degree.

Question 12 For each of the core node's 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

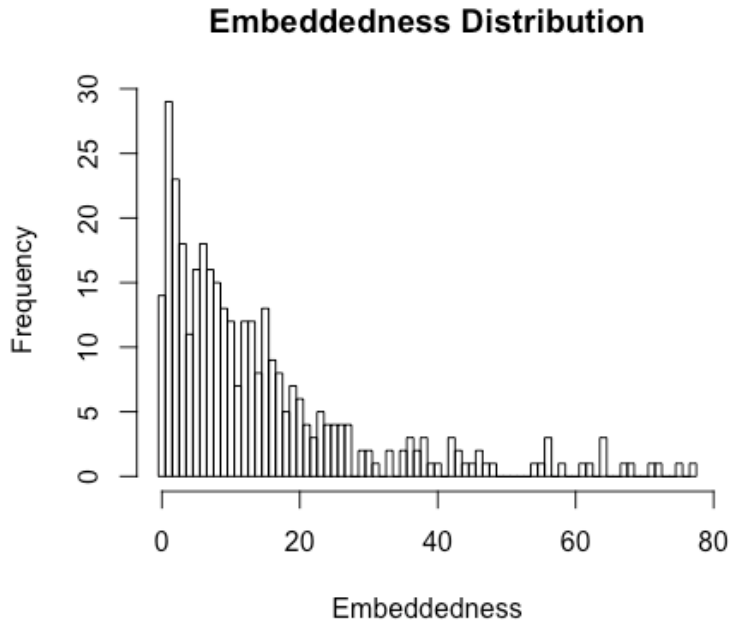


Figure 12.1
for coreNode 1's personalized network, distribution of embeddedness is shown above

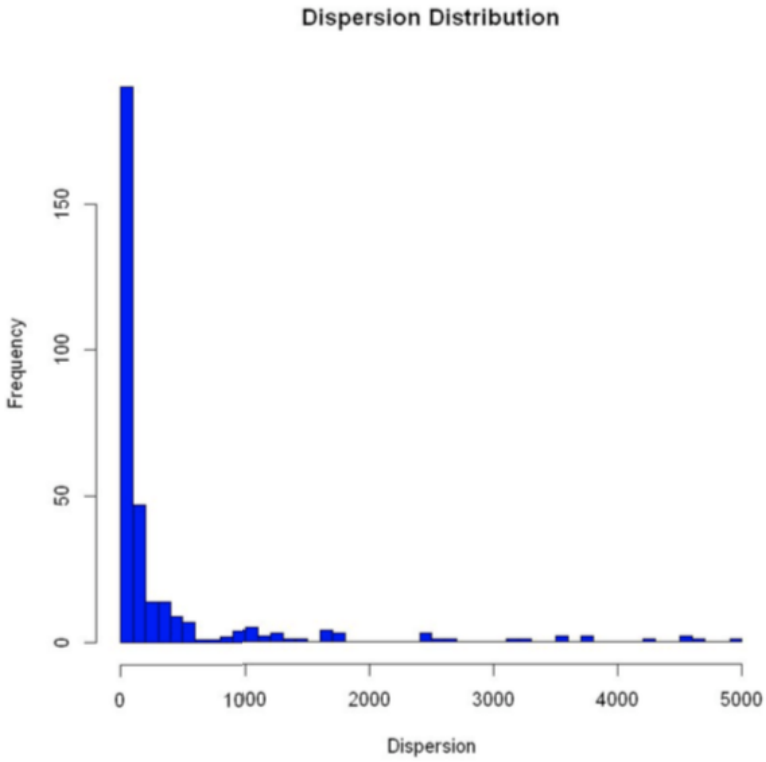


Figure 12.2
for coreNode 1's personalized network, distribution of dispersion is shown above

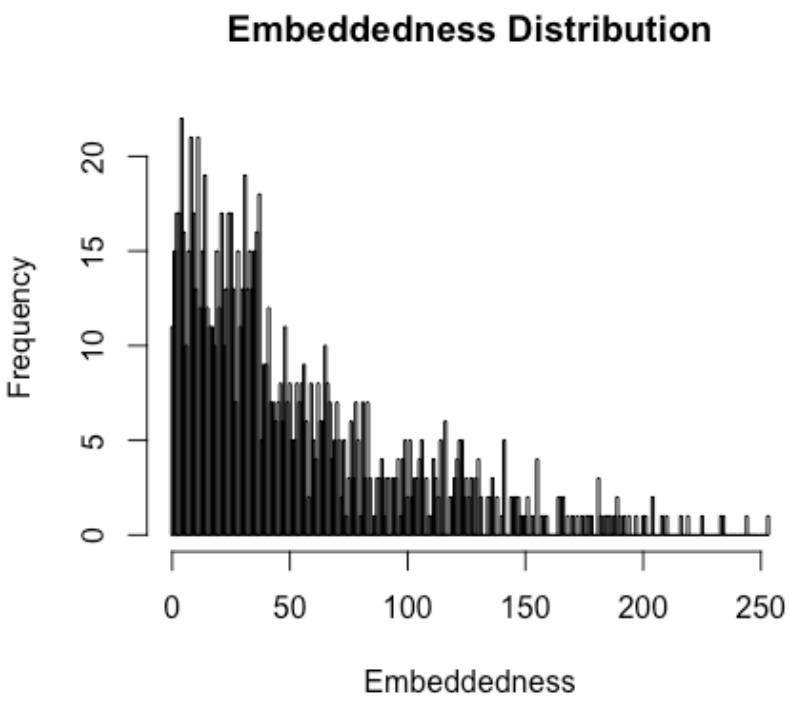


Figure 12.3
for coreNode 108's personalized network, distribution of embeddedness is shown above

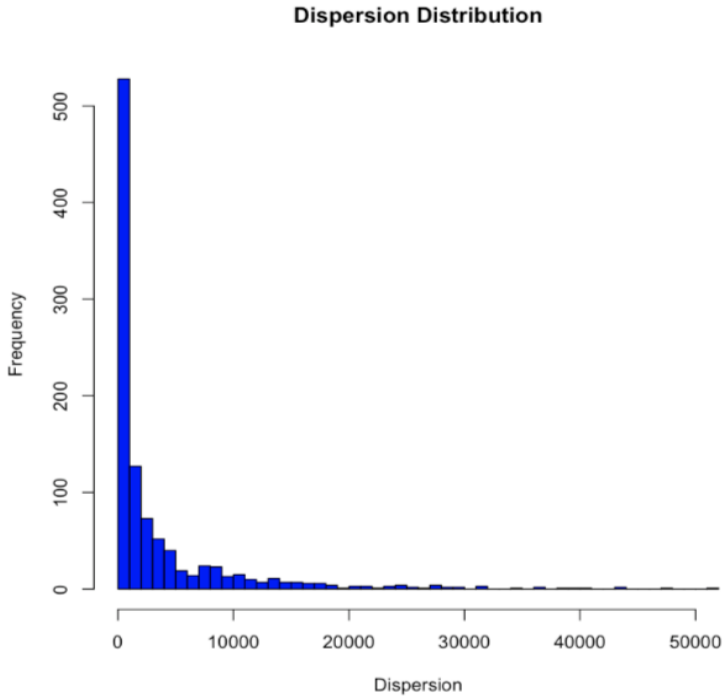


Figure 12.4
for coreNode 108's personalized network, distribution of dispersion is shown above

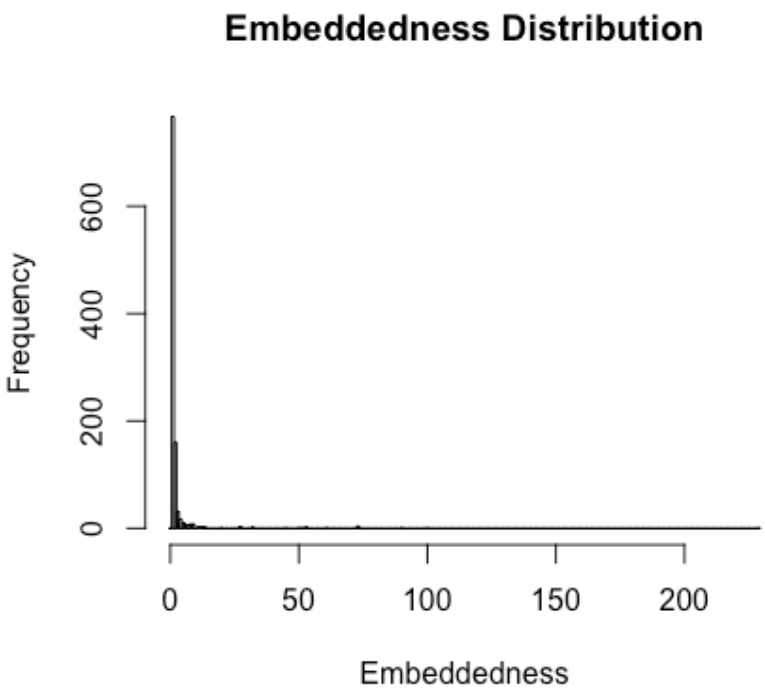


Figure 12.5
for coreNode 349's personalized network, distribution of embeddedness is shown above

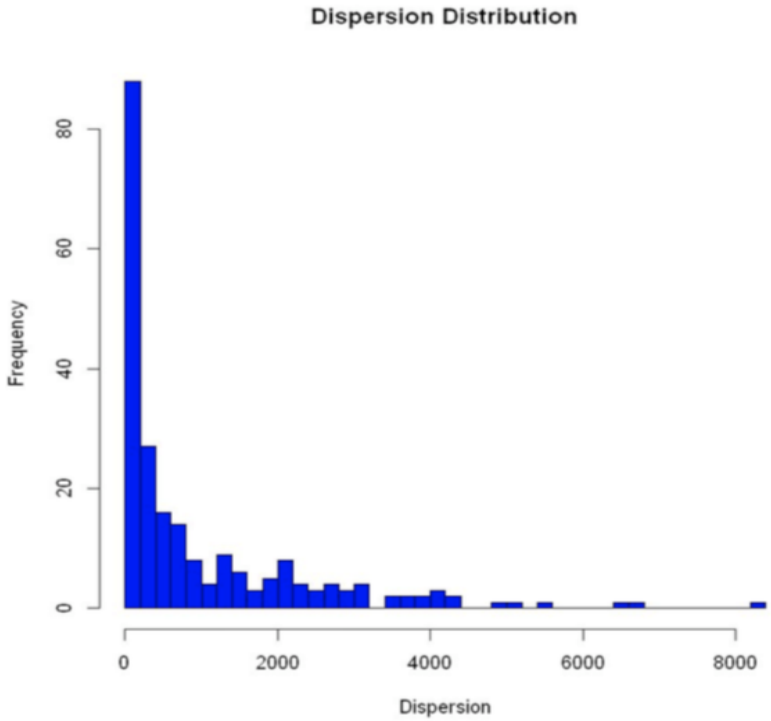


Figure 12.6
for coreNode 349's personalized network, distribution of dispersion is shown above

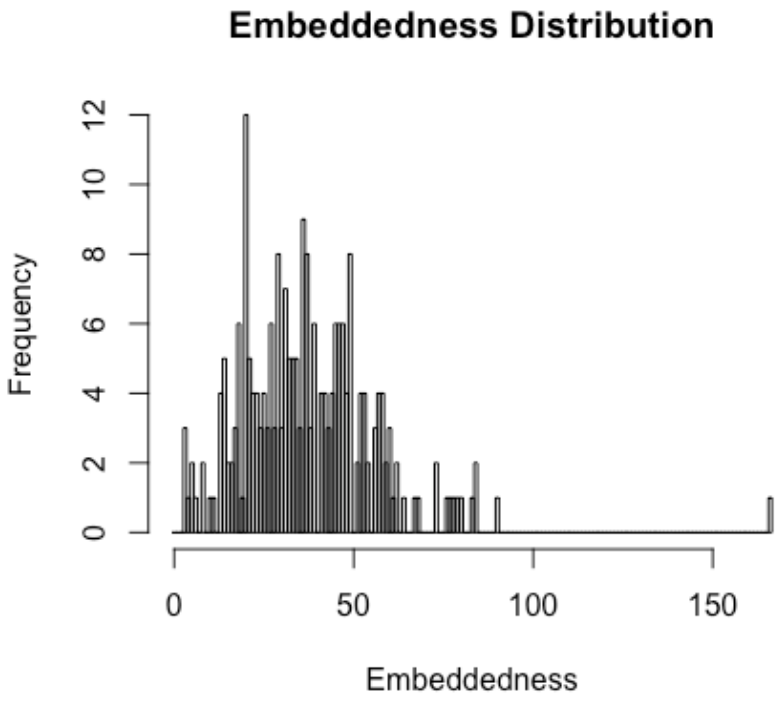


Figure 12.7
for coreNode 484's personalized network, distribution of embeddedness is shown above

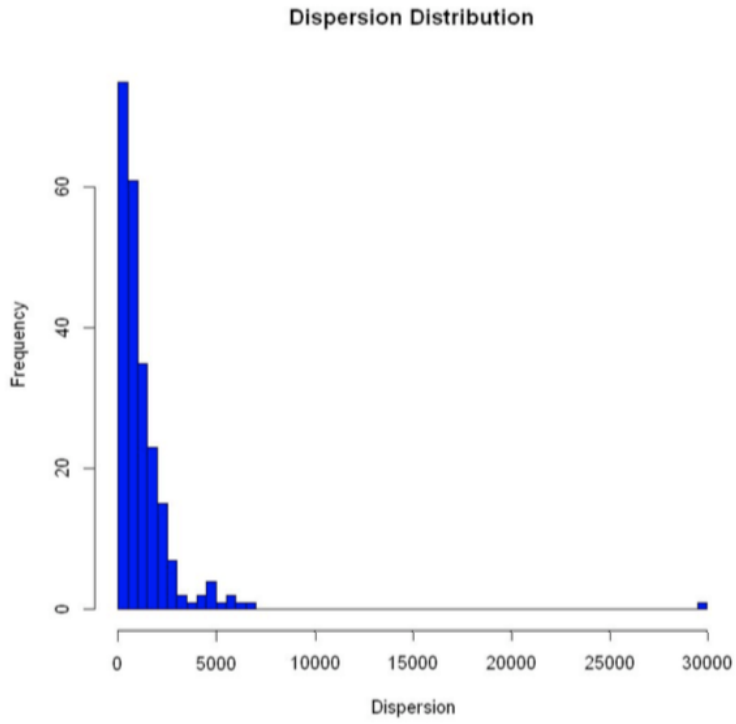


Figure 12.8
for coreNode 484's personalized network, distribution of dispersion is shown above

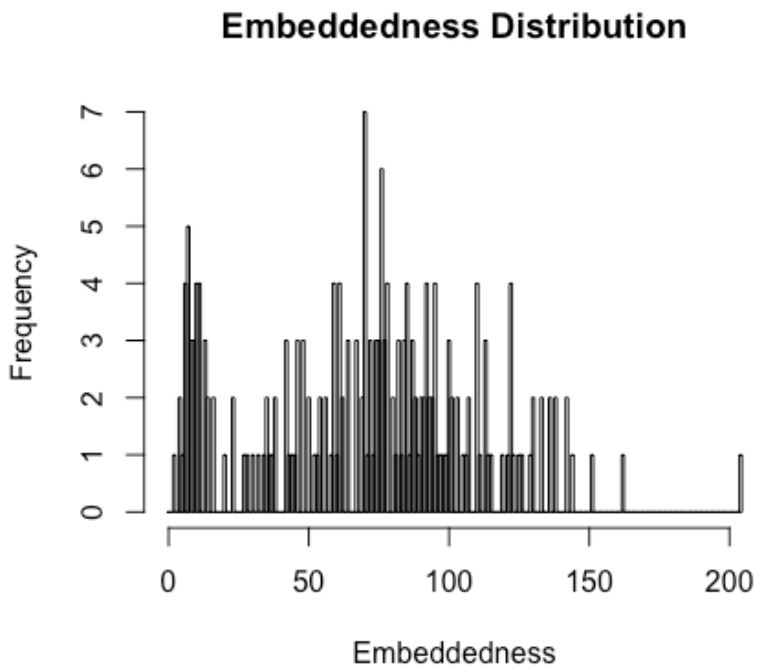


Figure 12.9
for coreNode 1087's personalized network, distribution of embeddedness is shown above

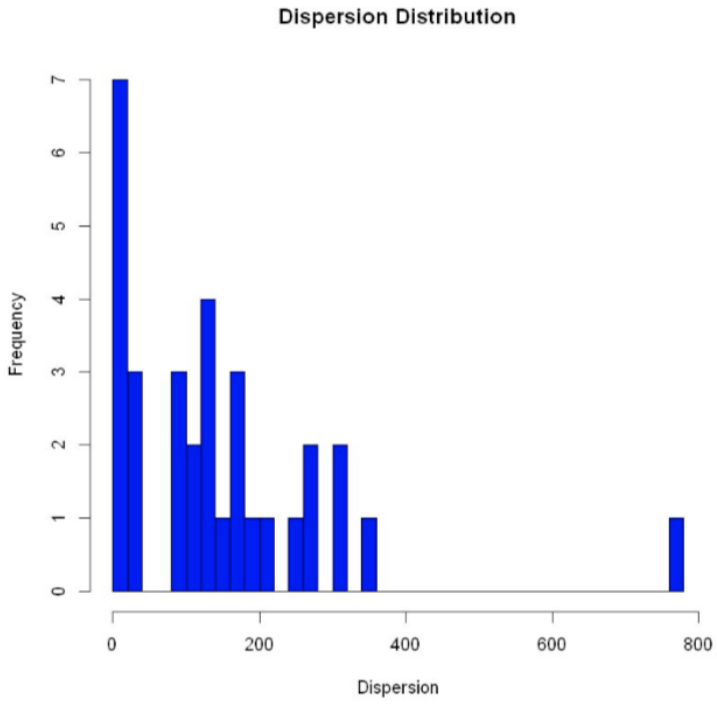


Figure 12.10
for coreNode 1087's personalized network, distribution of dispersion is shown above

Question 13 For each of the core node's personalized network, plot 5 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 FastGreedy algorithm. In this question, you will have 5 plots.

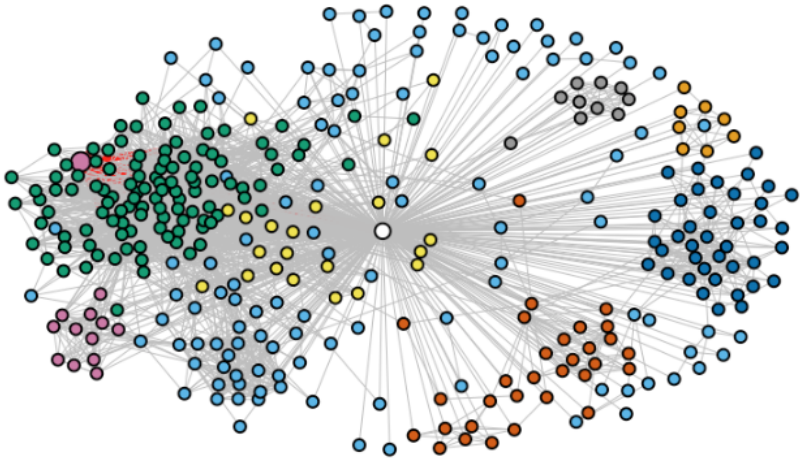


Figure 13.1
Community structure of the personalized network with coreNode 1, and the node with larger size than the surrounding nodes is the one with maximum dispersion. The incident edges are thicker than other edges

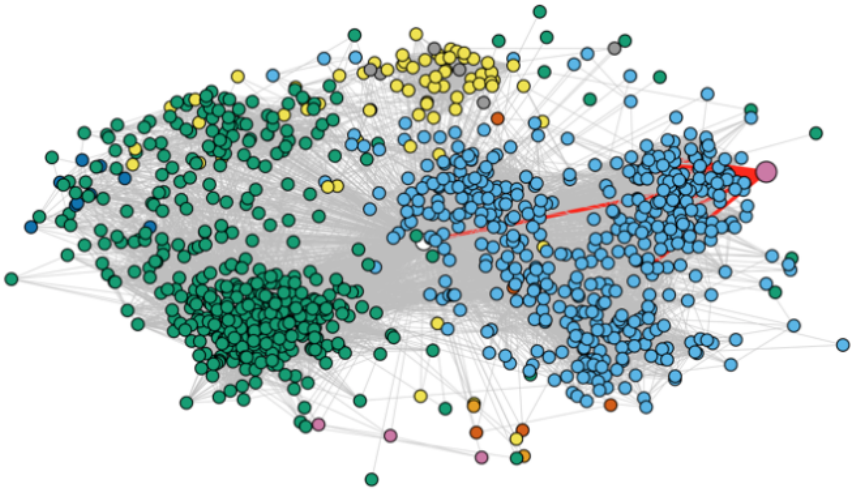


Figure 13.2
Community structure of the personalized network with coreNode 108, and the node with larger size than the surrounding nodes is the one with maximum dispersion. The incident edges are thicker than other edges

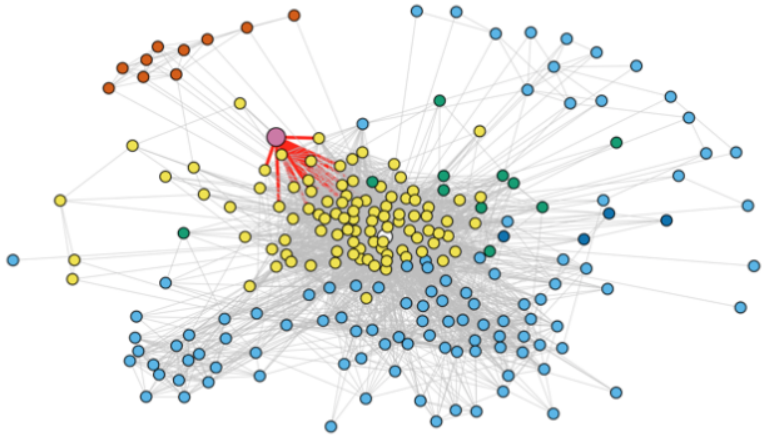


Figure 13.3
Community structure of the personalized network with coreNode 349, and the node with larger size than the surrounding nodes is the one with maximum dispersion. The incident edges are thicker than other edges

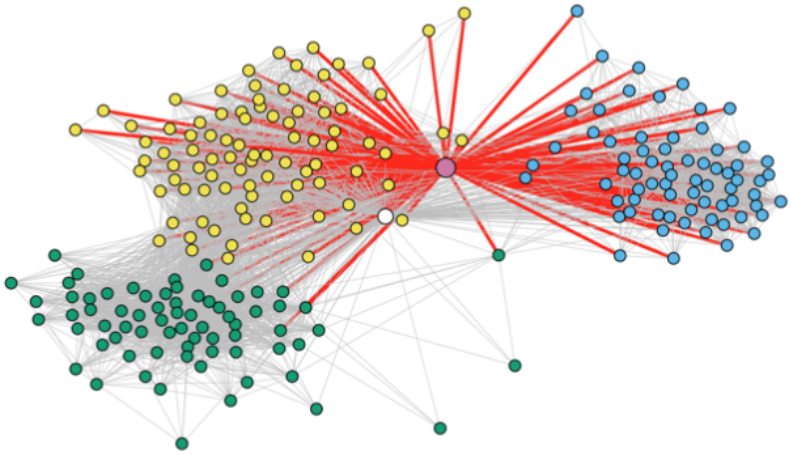


Figure 13.4
Community structure of the personalized network with coreNode 484, and the node with larger size than the surrounding nodes is the one with maximum dispersion. The incident edges are thicker than other edges

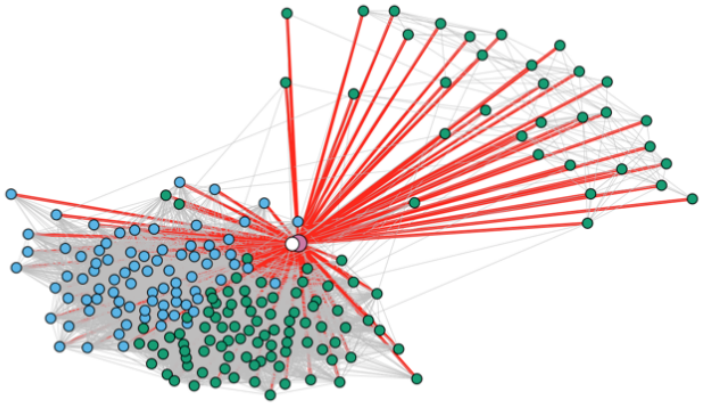


Figure 13.5

Community structure of the personalized network with coreNode 1087, and the node with larger size than the surrounding nodes is the one with maximum dispersion. The incident edges are thicker than other edges

Question 14 Repeat question 13, but now highlight the node with maximum embeddedness and the node with maximum dispersion embeddedness . Also, highlight the edges incident to these nodes

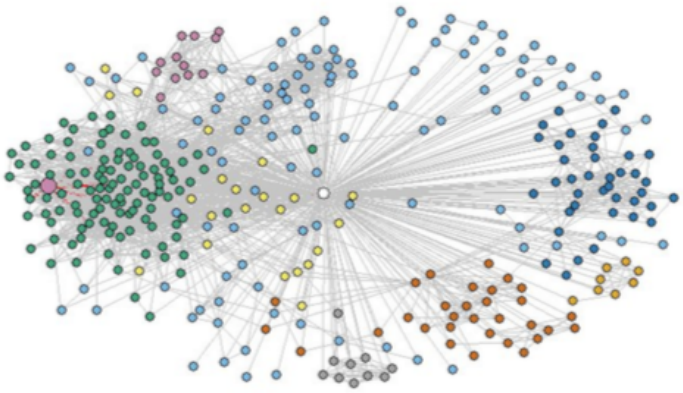


Figure 14.1

Community structure of the personalized network with coreNode 1, and the node with larger size than the surrounding nodes is the one with maximum embeddedness. The incident edges are thicker than other edges

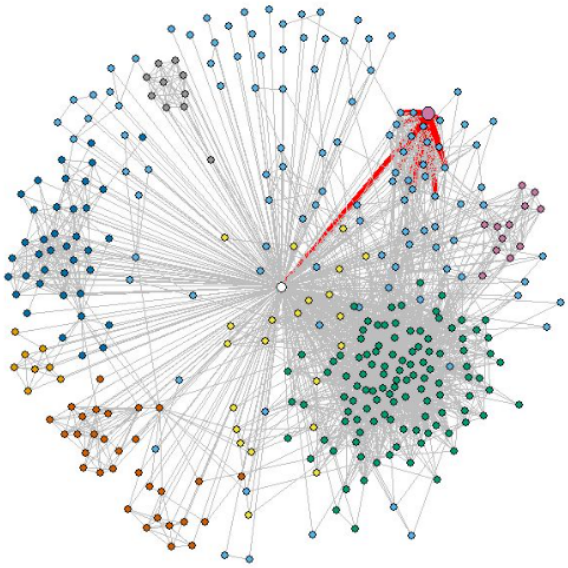


Figure 14.2

Community structure of the personalized network with coreNode 1, and the node with larger size than the surrounding nodes is the one with maximum dispersion/embeddedness. The incident edges are thicker than other edges

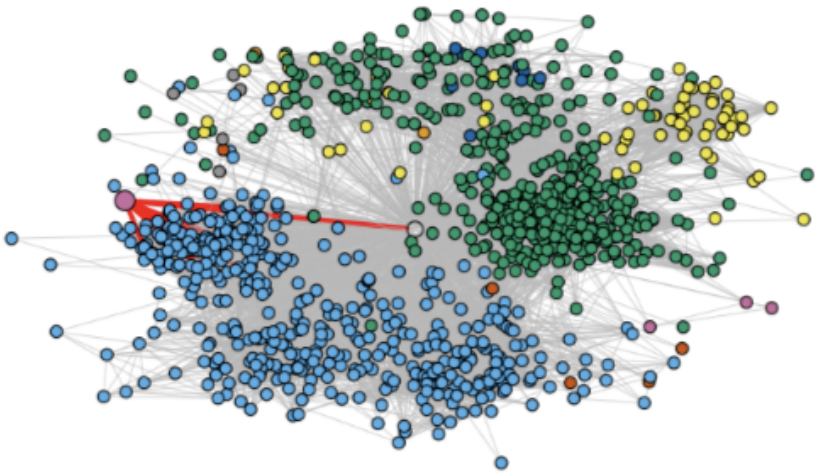


Figure 14.3

Community structure of the personalized network with coreNode 108, and the node with larger size than the surrounding nodes is the one with maximum embeddedness. The incident edges are thicker than other edges.

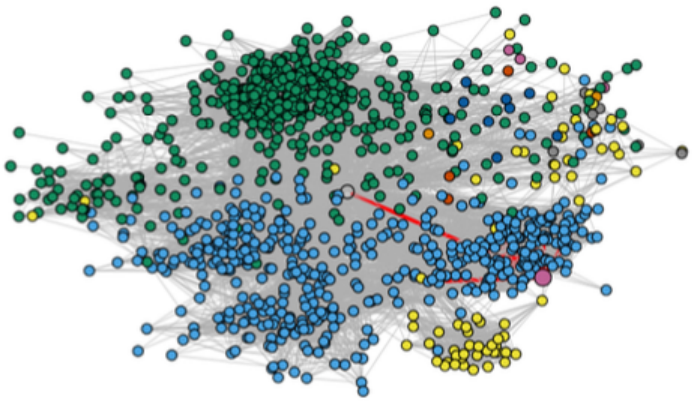


Figure 14.4
Community structure of the personalized network with coreNode 108, and the node with larger size than the surrounding nodes is the one with maximum dispersion/embeddedness. The incident edges are thicker than other edges.

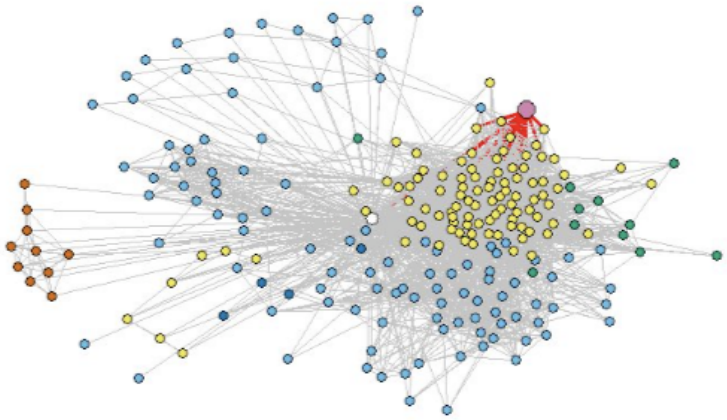


Figure 14.5
Community structure of the personalized network with coreNode 349, and the node with larger size than the surrounding nodes is the one with maximum embeddedness. The incident edges are thicker than other edges.

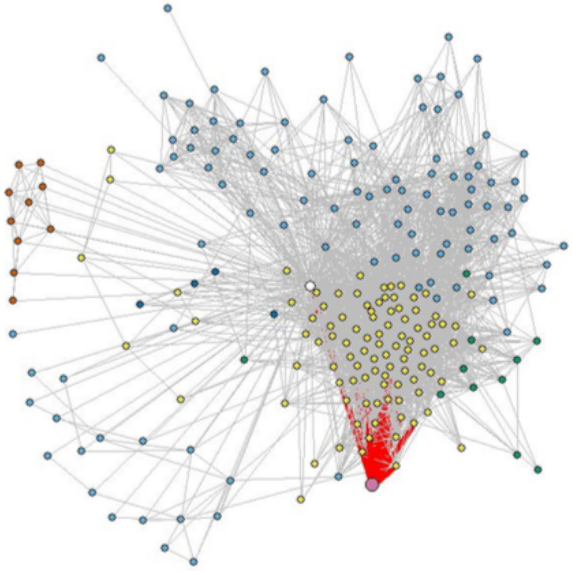


Figure 14.6
Community structure of the personalized network with coreNode 349, and the node with larger size than the surrounding nodes is the one with maximum dispersion/embeddedness. The incident edges are thicker than other edges.

Figure 14.7
Community structure of the personalized network with coreNode 484, and the node with larger size than the surrounding nodes is the one with maximum embeddedness. The incident edges are thicker than other edges.

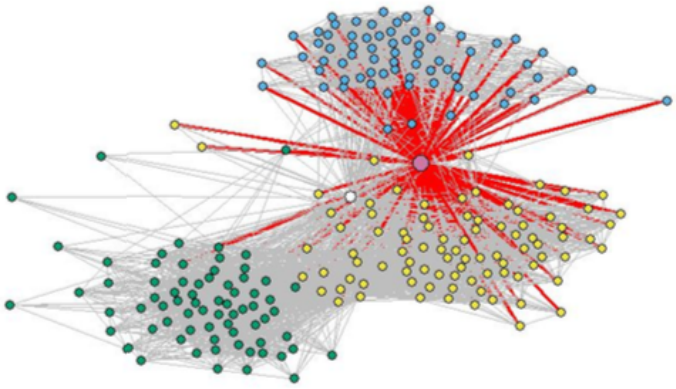


Figure 14.8
Community structure of the personalized network with coreNode 484, and the node with larger size than the surrounding nodes is the one with maximum dispersion/embeddedness. The incident edges are thicker than other edges.

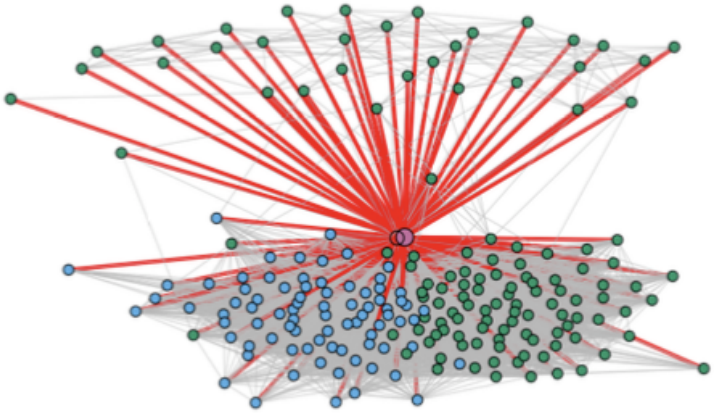


Figure 14.9
Community structure of the personalized network with coreNode 1087, and the node with larger size than the surrounding nodes is the one with maximum embeddedness. The incident edges are thicker than other edges.

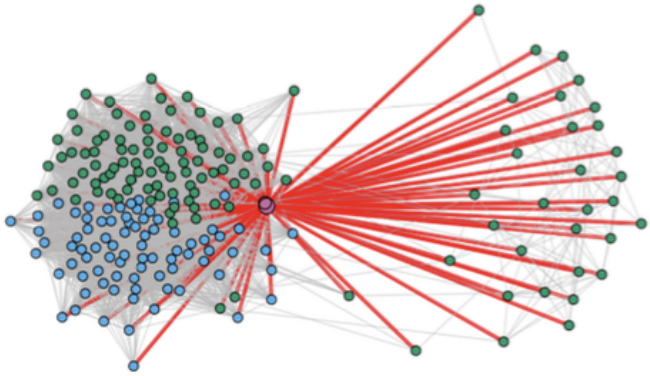


Figure 14.10
Community structure of the personalized network with coreNode 1087, and the node with larger size than the surrounding nodes is the one with maximum dispersion/embeddedness. The incident edges are thicker than other edges.

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

After getting the results above, I can conclude that the embeddedness indicates the closeness between two people, that is, core node and each of its friend. From the above plots, personalized network with larger embeddedness indeed shows lots of common connected nodes.

Dispersion indicates that how far away between core nodes' friends. From the above plots, personalized network with larger dispersion means that the core node's friends distance in the graph is quite large. That scenario is confirmed by the above plots. Larger dispersion indicates that core node's friends don't know each other.

The dispersion/embeddedness rate indicates the combination of previous two metrics. Larger dispersion/embeddedness rate means that the relationship of this personalized network formed by current core node is not close. From the observation of above plots, personalized network with larger rate has quite loose connection between each node in the network Smaller rate means that the relationship of this personalized network formed by current core node is close.

1.4 Friend recommendation in personalized networks

Create an induced subgraph of personalized network of node ID 415.

Question 16 What is $|N_r|$?

The list of users who we want to recommend new friends to is created by picking all nodes with degree 24. The number of users in the list is 11. The list of users' ID is shown as follows:

"497" "579" "601" "616" "619" "628" "644" "659" "660" "662" "663"

Question 17 Compute the average accuracy of the friend recommendation algorithm that uses:
• Common Neighbors measure
• Jaccard measure

• Adamic Adar measure

Based on the average accuracy values, which friend recommendation algorithm is the best?

Follow the instruction provided in the tutorial. The results are shown as follows. Table 1.4.1 shows the average accuracy of three friend recommendation algorithms for each user. The average accuracies are obtained from running the experiment in the same setting by 10 times.

Index of user	Common Neighbor	Jaccard	Adamic Adar
1	0.3894048	0.1813095	0.3620238
2	1.0000000	0.9633333	0.9833333
3	0.9264286	0.9264286	0.9264286
4	0.8301587	0.8301587	0.8301587
5	0.4133333	0.5642857	0.4133333
6	1.0000000	0.9633333	1.0000000
7	0.8952778	0.8809921	0.8952778
8	0.9875000	0.9341667	0.9875000
9	0.9833333	0.9323810	0.9833333
10	0.9427778	0.8927778	0.9427778
11	0.9800000	0.9600000	0.9800000

Table 1.4.1: Average accuracy of three different friend recommendation algorithms for the users in the list.

Compute the average accuracy of each algorithm by averaging across the accuracies of the users in the list. The result is shown in Table 1.4.2.

Common Neighbor	Jaccard	Adamic Adar
0.8498377	0.8208333	0.8458333

Table 1.4.2: Average accuracy of each algorithm.

Based on the average values of accuracies for the three different measures, the Common Neighbors measure achieves the best performance. But, in fact, three different measures achieved similar accuracy scores.

2 Google+ network

Question 18 How many personal networks are there?

Create directed personal networks for users who have more than 2 circles. The number of the personal networks is 57.

Question 19 For the 3 personal networks (node ID given below), plot the in-degree and out-degree distribution of these personal networks. Do the personal networks have a similar in and out degree distribution. In this question, you should have 6 plots.

• 109327480479767108490

• 115625564993990145546

• 101373961279443806744

Extract the three personal networks of the given node ID. The in-degree and out-degree distributions of the three personal networks are shown in Figure 2.1.1.

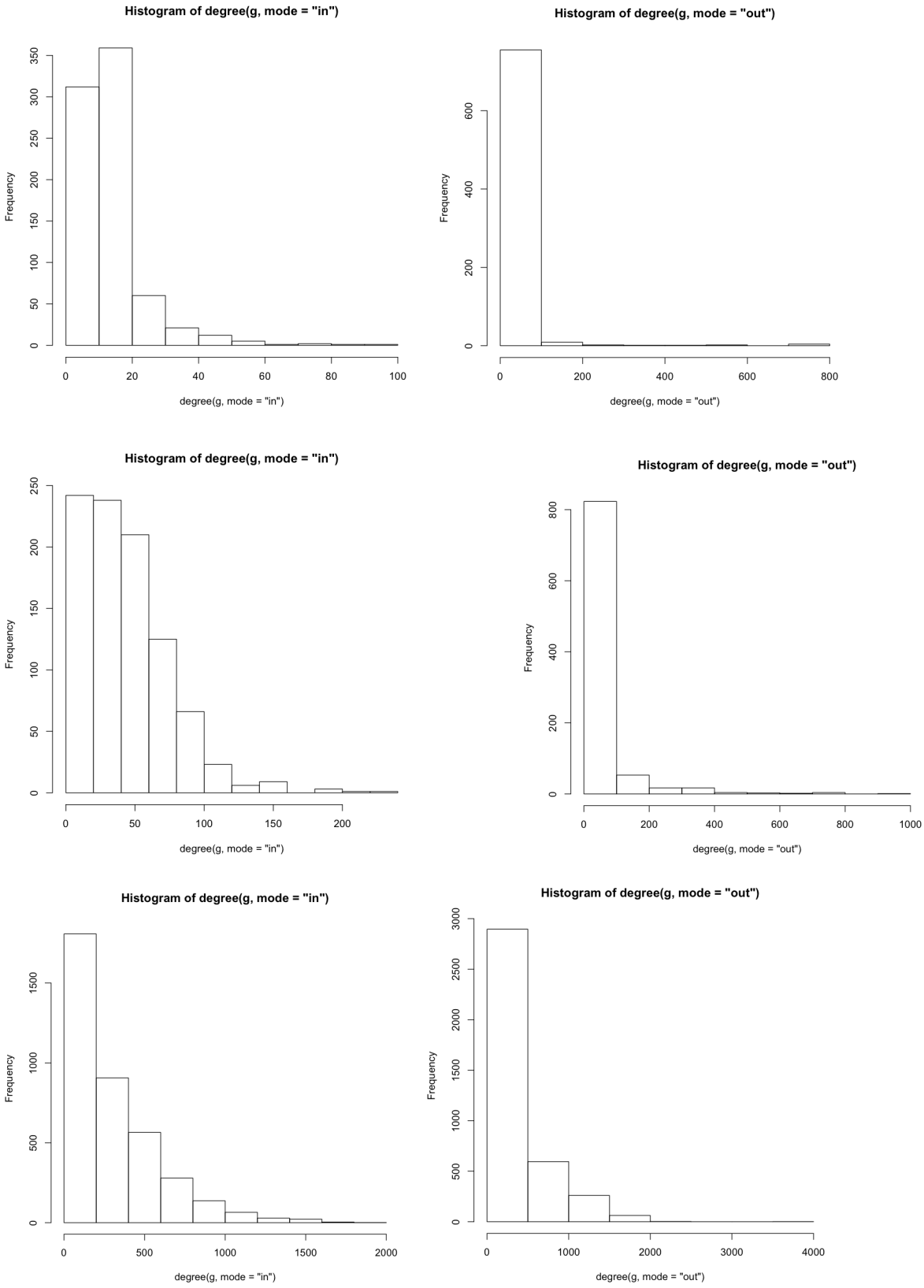


Figure 2.1.1: Degree distribution of the personal networks.

According to Figure 2.1.1, the out-degree distributions of the first two personal networks are pretty similar, but the in-degree distributions vary. For the firs personal network, most of its nodes have out-degree in the range from 0 to 20, while for the second personal network, the

out-degrees of most of its nodes are within the range from 0 to 100. For the third personal network, the network seems to be much larger than the other two, so its out-degree and in-degree distributions have relatively larger spans. The pattern of its out-degree distribution is similar to the other two networks, where most of its nodes, however, have out-degree of less than 500. The in-degree distribution is more like an exponential decay pattern than the other two networks.

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? In this question, you should have 3 plots.

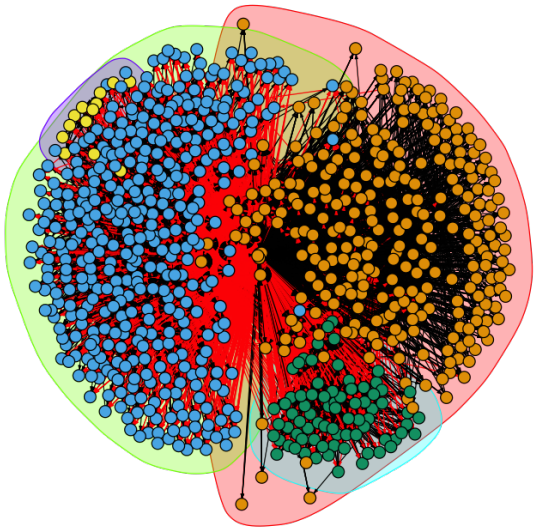
Personal Network 1: 109327480479767108490

Personal Network 2: 115625564993990145546

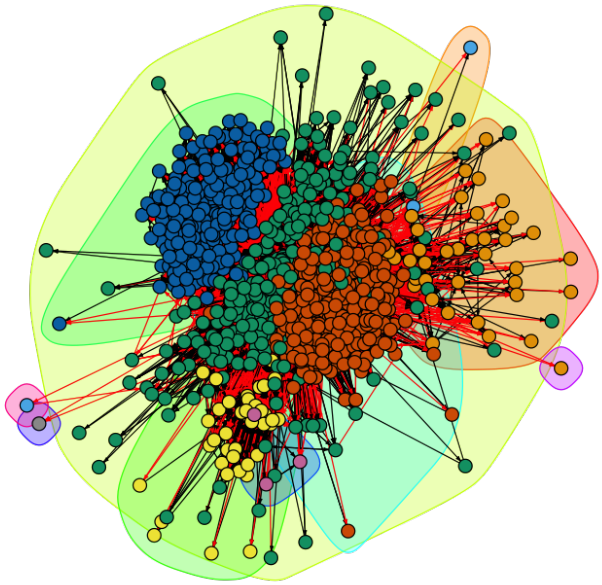
Personal Network 3: 101373961279443806744

The graphs for the personal networks are demonstrated in Figure 2.1.2. The modularity scores are listed in Table 2.1.1.

Personal Network 1



Personal Network 2



Personal Network 3

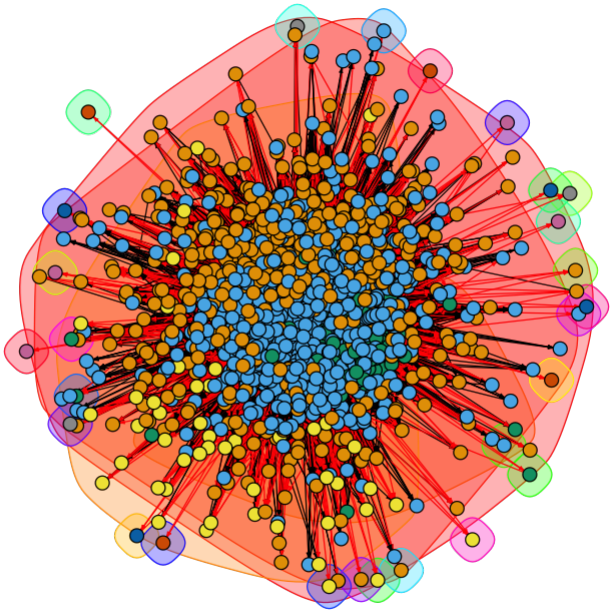


Figure 2.1.2: Plots of Personal Networks 1,2 and 3

Personal Network	Personal Network 1	Personal Network 2	Personal Network 3

Modularity score	0.2527654	0.3194726	0.1910903

Table 2.1.1: Modularity scores of the personal networks.

The modularity scores of the three personal networks are not similar, also proved by Figure 2.1.2. The personal network 3 has the lowest modularity score, meaning that it doesn't have dense connections between the nodes within communities but sparse connections between nodes in different communities. Relatively, the personal network 2 is most likely to have strong connection within the communities and weak connections in-between the communities

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

The homogeneity formula is:

$$h = 1 - \frac{H(C|K)}{H(C)}$$

H(C|K) describes the conditional entropy of circle given community, while H(C) describes the entropy of circle. If all communities contain only elements of a single circle, H(C|K)/ H(C) will be zero. Then h will be 1, standing for pure homogeneity.

In the same way, the completeness formula is shown as follows:

$$c = 1 - \frac{H(K|C)}{H(K)}$$

If all elements of a circle are members of the same community, H(K|C)/ H(K) will become zero. Then, c will be 1, standing for perfect completeness.

In other words, a clustering result satisfies homogeneity if all of its communities contain only data points which are members of a single circle. A clustering result satisfies completeness if all the data points that are members of a given circle are elements of the same community.

Question 22 Compute the h and c values for the community structures of the 3 personal networks (same nodes as question 19). Interpret the values and provide a detailed explanation.

In general, the homogeneity score can hardly be negative, while the completeness score can easily be negative. According to the given definition, the entropy of circle can be high, since there can be a couple of circles, where most of circles overlap with each other in their members, since a member can belong to multiple circles simultaneously. The entropy of community cannot be high, since a member cannot belong to two communities at the same time. In other word, sum of the number of people in circle C_i, a_i, is not only not equal to the total number of people with circle information, N, but is much larger than N in some cases. Therefore, if there are many circles which contain many members and some circles share similar members with each other, then the conditional entropy of community given circle can be higher, compared with the entropy of community. Then, the c-value, completeness score, might be negative. The homogeneity scores and completeness scores of the three personal networks are listed in Table 2.1.2.

Personal Network	Personal Network 1	Personal Network 2	Personal Network 3
Homogeneity	0.85188544	0.45189098	0.00386597
Completeness	0.3298743	-3.4239577	-1.5042390

Table 2.1.2: Homogeneity and completeness scores of the three personal networks.

For the personal network 1, its homogeneity score is very high. It is because most of its communities almost only contain data points belonging to a single circle. Its completeness is the highest, since most of data points of a circle belong to the same community.

For the personal network 2, its homogeneity score is medium. It is because some of its communities almost only contain data points belonging to a single circle, while the other communities contain data points, which are members of different circles. Its homogeneity score is the lowest. It has high entropy of circle, since it has many circles, where most of circles overlap with each other in their members. The entropy of community cannot be high, since a member cannot belong to two communities at the same time. Then, it has high conditional entropy of community given circle. Thus, the completeness score is negative. This also suggests circles in personal network 2 have data points belonging to different communities.

For the personal network 3, its homogeneity score is extremely low. The reason is that most of people with circle information appear simultaneously in different circles. It is to say multiple circles overlap with each other in their members. So, for any single community, it is very likely that the members with circle information belong to several circles at the same time. Then, it is very hard to achieve homogeneity. Thus, the homogeneity score is very low in the case. It also has a negative completeness score. It has few communities whose members possess circle information, making the entropy of community is relatively low. Since it has highly-overlapping circles and meanwhile, each circle contain elements, not achieving perfect completeness, this yields to a comparably high conditional entropy of community given circle. This also suggests circles in personal network 3 have data points belonging to different communities. But, the case is better than that for the personal network 2.