ECE 232E Project 2

Social Network Mining

Jui Chang
Wenyang Zhu
Xiaohan Wang
Yang Tang

1. Facebook network

Given the edgelist file, we can create the Facebook network.

1.1 Structural properties of the Facebook network

In this part, we will study structural properties of the Facebook network. Here, we will study connectivity and degree distribution.

Question 1:
The Facebook network we created from the edgelist is connected.

Question 2:
The diameter of Facebook network we created in previous part is 8.

Question 3:
The average degree of the network is 43.69.
The degree distribution of the Facebook network is shown as follows:

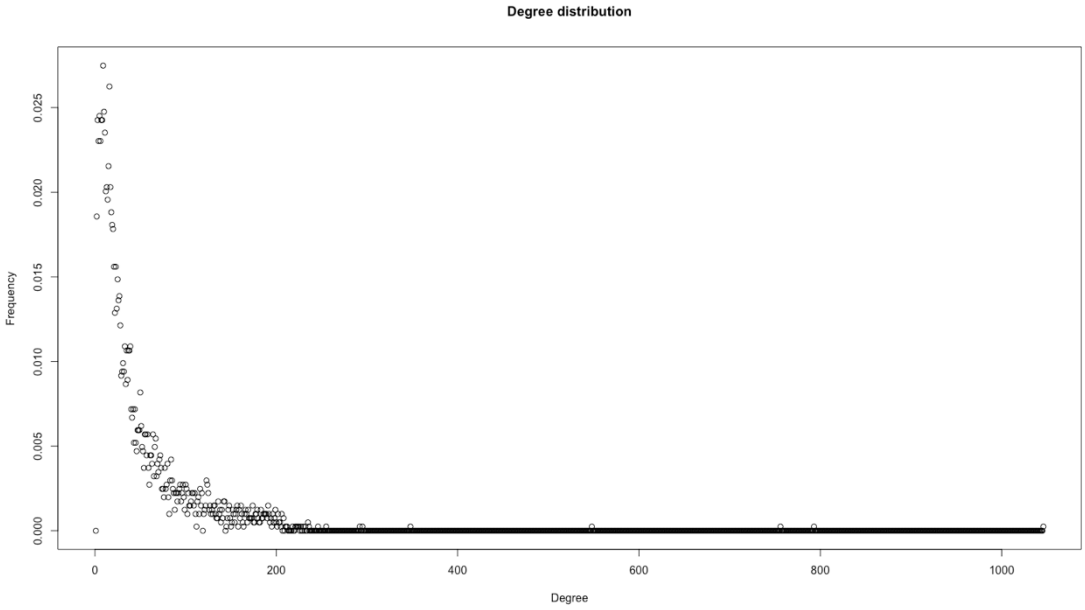


Figure 1. Degree Distribution of the Facebook Network

Question 4:

We plot the degree distribution of question 3 in log-log scale and fit a line to the middle part of the graph, the line is from x=15 to 200, and the function of the line is $y = x^{-1.35}$. The figure is log-log scale, so take log to the equation and it becomes $\log(y) = -1.35 * \log(x)$, the slope of the log scale straight line is -1.35.

The log-log scale degree distribution and the fitted line are shown as follows:

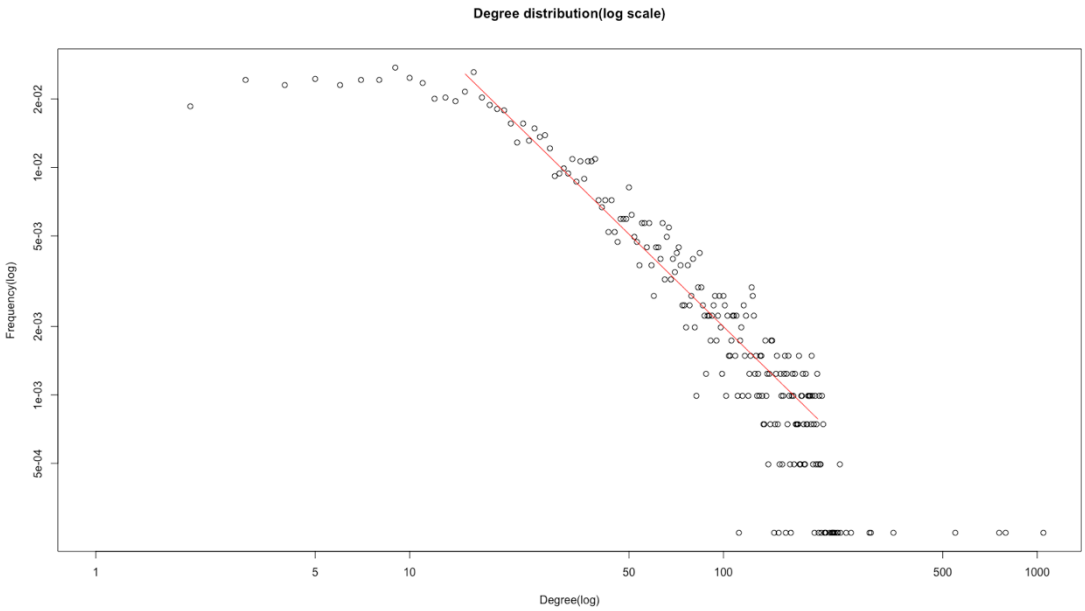


Figure 2. Degree Distribution of the Facebook Network (log-log scale)

1.2 Personalized network

A personalized network for person v_i is defined to be the subgraph from v_i and its neighbors. Here, we will study the structural properties of the personalized network of the user graph node ID is 1. (in edgelist the node ID is 0)

Question 5:

We create a personalized network of the user whose ID is 1. In the personalized network, we find there is 348 nodes and 2866 edges.

Question 6:

The diameter of the personalized network in question 5 is 2. The trivial upper and lower bound for the diameter of the personalized network is $[1, 2]$, i.e. the upper bound is 2, and the lower bound is 1.

Question 7:

The personalized network of node 1 means a graph that consists of node 1 and its neighbors and the edges that have both ends within this set of nodes.

In this context, if the diameter of the personalized network is equal to upper bound "2", it means that there are at least two nodes v_i, v_j, except the center node 1, not having the direct connection. Thus, the shortest path between v_i, v_j is $v_i \rightarrow v_1 \rightarrow v_j = 2$. In the real case, it means that user v_i and user v_j are not mutual friend, but they both have connection with center user v_1.

If the diameter of the personalized network is equal to lower bound "1", it means that any pair of nodes in this subgraph are mutually connected to each other, so the shortest path between any pair of nodes is always "1", and the diameter is also "1". In the real case, it indicates that the users in this group are all mutual friends to each other.

1.3 Core node's personalized network

Question 8:

A core node is the nodes that have more than 200 neighbors. In the Facebook network, we find 40 core nodes. And the average degree of the core nodes is 279.375.

1.3.1 Community structure of core node's personalized network

Question 9:

In this part, we study and plot the community structure of 5 core nodes' personalized network:

- Node ID 1
- Node ID 108
- Node ID 349
- Node ID 484
- Node ID 1087

For each of the core node's personalized network, we use 3 community clustering methods – Fast-Greedy, Edge-Betweenness and Infomap, to detect their community structure and compare the modularity scores. Below are our results:

Table 1. Modularity scores of 3 community detection algorithms (with core node)

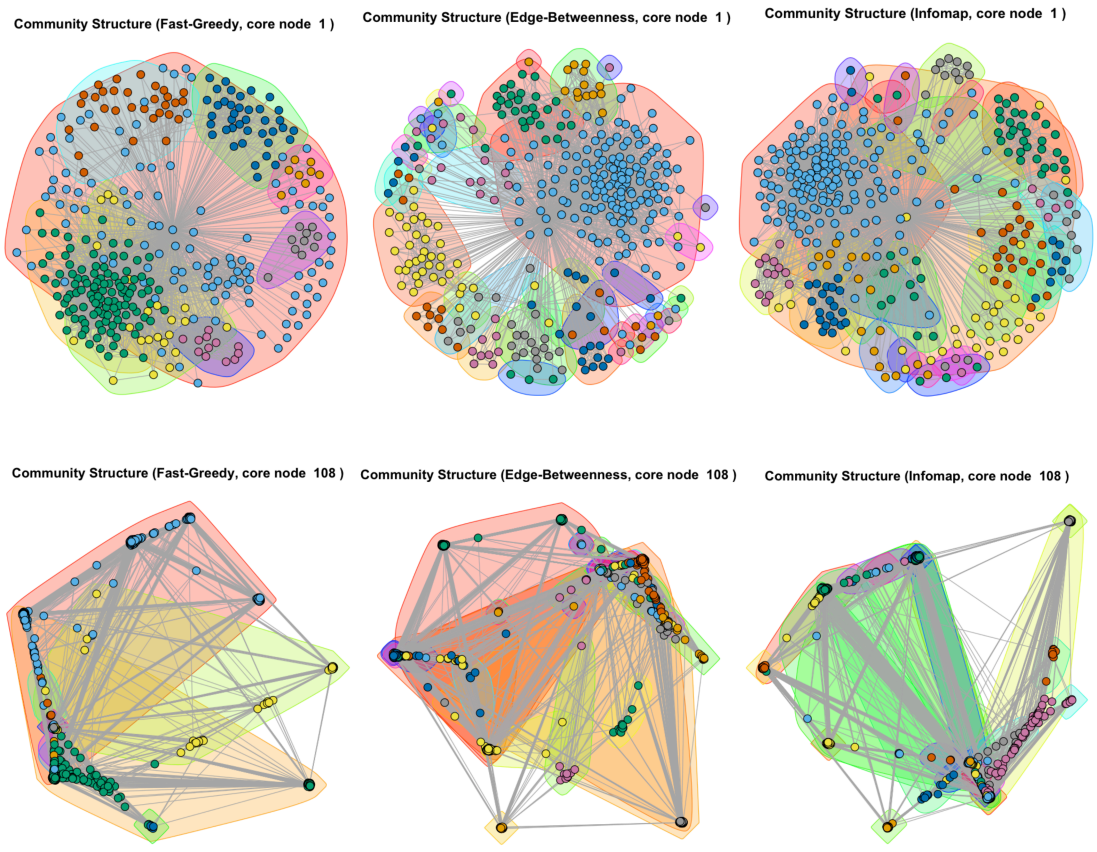
	Fast-Greedy	Edge-Betweenness	Infomap
Node ID 1	0.413	0.353	0.389
Node ID 108	0.436	0.507	0.508
Node ID 349	0.252	0.134	0.095
Node ID 484	0.507	0.489	0.515
Node ID 1087	0.146	0.028	0.027

From the above table, we find that:

1) For community detection algorithms, in general, Fast-Greedy has relatively better performances against the other two methods because of the higher modularity. While for Node ID 108, Edge-Betweenness and Infomap algorithms have higher modularity scores than that of Fast-Greedy. Since node 108 has the largest number of nodes in the personalized network, we can conclude that Fast-Greedy can have better performance on smaller networks, while Edge-Betweenness and Infomap are suitable for larger networks. Besides, when we run the codes, we also find that Fast-Greedy and Infomap are much faster than Edge-Betweenness.

2) For core nodes' personalized networks, it indicates that Node ID 484 achieves the highest modularity scores than other nodes, and Node ID 1087 has the lowest modularity score, which indicates that Node 484's personalized network has clear community structure and Node 1087 has personalized network with more ambiguous community structure.

Below are community structures of those five core nodes' personalized networks:



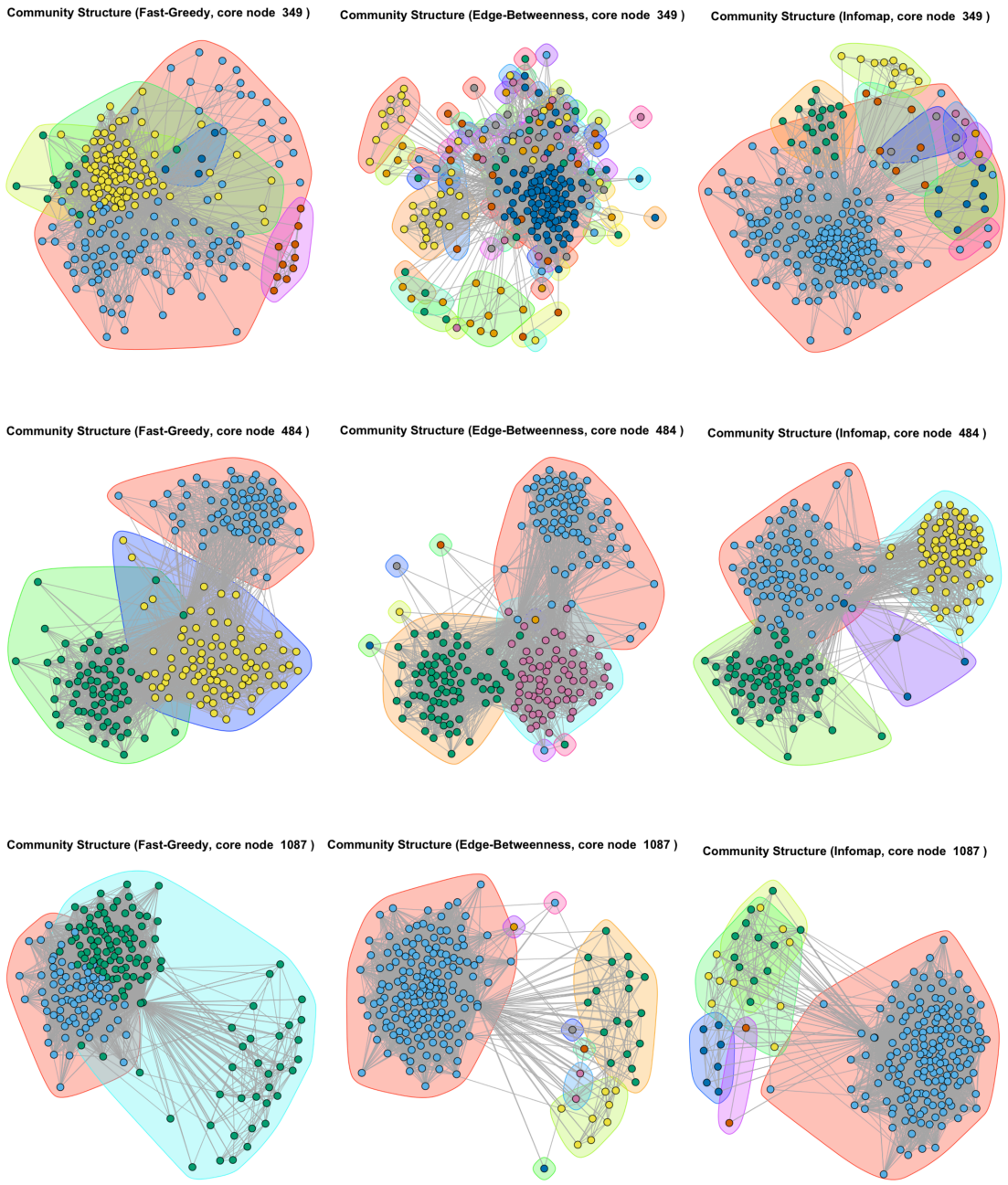


Figure 3. Community structure with 3 algorithms (with core node)

1.3.2 Community structure with the core node removed

Question 10:

In this part, we removed the core nodes in each personalized network and explored the effect on the community structure. Below are our results:

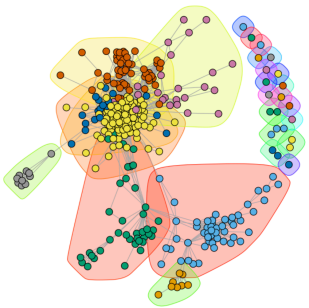
Table 2. Compare modularity score with and without core nodes

		Fast-Greedy	Edge-Betweenness	Infomap
Node ID 1	With core	0.413	0.353	0.389
	Without core	0.442	0.416	0.418
Node ID 108	With core	0.436	0.507	0.508
	Without core	0.458	0.521	0.521
Node ID 349	With core	0.252	0.134	0.095
	Without core	0.246	0.151	0.245
Node ID 484	With core	0.507	0.489	0.515
	Without core	0.534	0.515	0.543
Node ID 1087	With core	0.146	0.028	0.027
	Without core	0.148	0.032	0.027

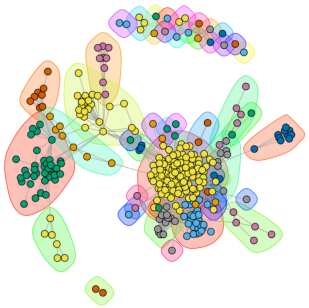
From the above table, we find that the modularity scores of the personalized networks without core nodes are higher than those with core nodes. For those personalized networks with core nodes, the core node can only be in one community. But since all other nodes connect to the core node, the connection between other communities and community with core node is very dense, which decrease the degree of modularity. While for personalized networks without the core node, the connection between communities is much sparser, thus achieving higher modularity scores. Besides, though the modularity scores increase after removing the core nodes, they are still close to the scores of personalized networks with core nodes, i.e. the scores only fluctuate in a small range.

Below are community structures of those five personalized networks without core nodes:

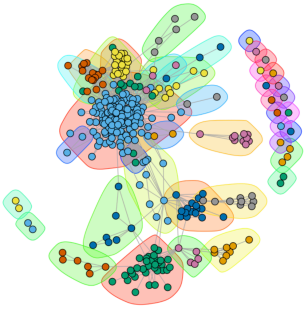
Community Structure (Fast-Greedy, Removed core node 1)



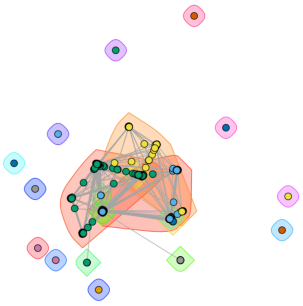
Community Structure (Edge-Betweenness, Removed core node 1)



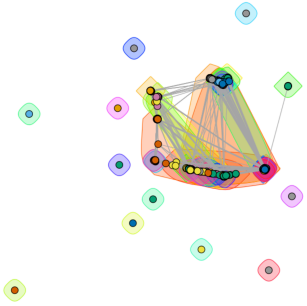
Community Structure (Infomap, Removed core node 1)



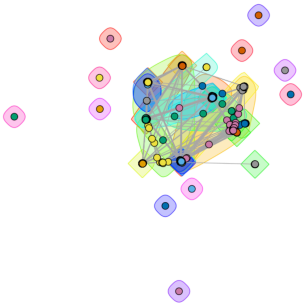
Community Structure (Fast-Greedy, Removed core node 108)



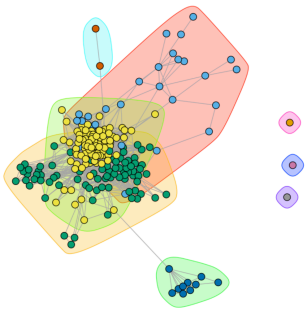
Community Structure (Edge-Betweenness, Removed core node 108)



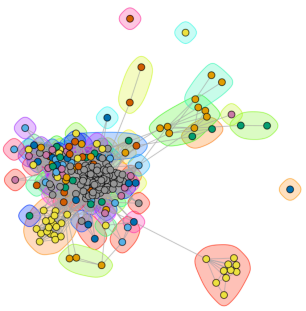
Community Structure (Infomap, Removed core node 108)



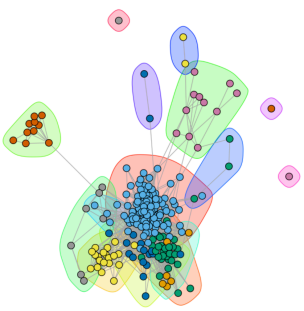
Community Structure (Fast-Greedy, Removed core node 349)



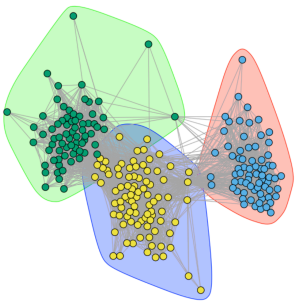
Community Structure (Edge-Betweenness, Removed core node 349)



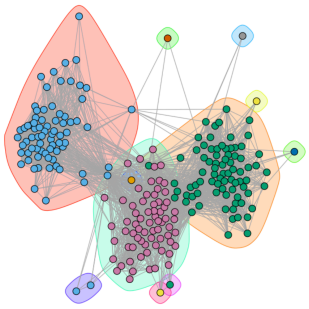
Community Structure (Infomap, Removed core node 349)



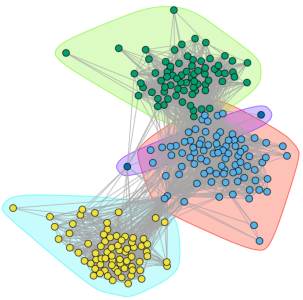
Community Structure (Fast-Greedy, Removed core node 484)



Community Structure (Edge-Betweenness, Removed core node 484)



Community Structure (Infomap, Removed core node 484)



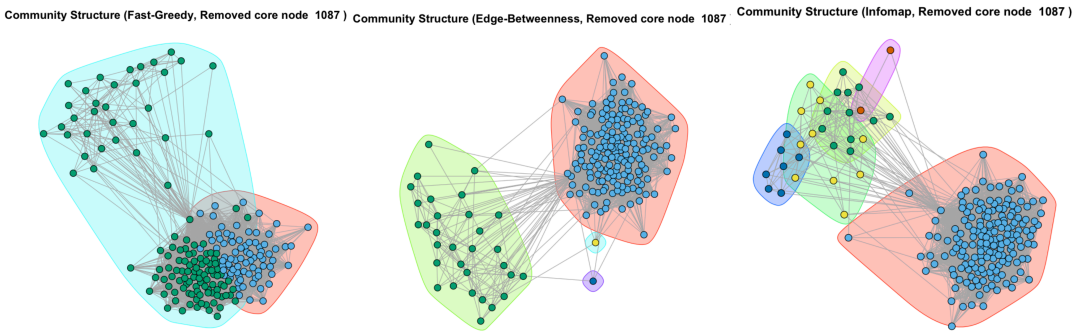


Figure 4. Community structure with 3 algorithms (without core node)

1.3.3 Characteristic of nodes in the personalized network

In previous parts, we have explored the properties of the personalized networks. In this part, we explore the characteristics of nodes in those networks measured by Embeddedness and Dispersion values.

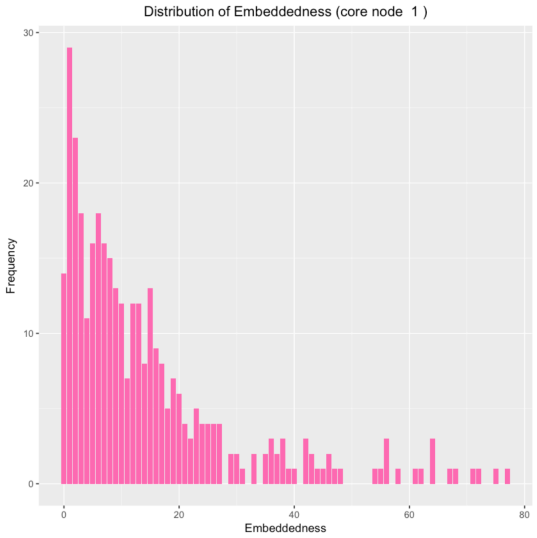
Question 11:

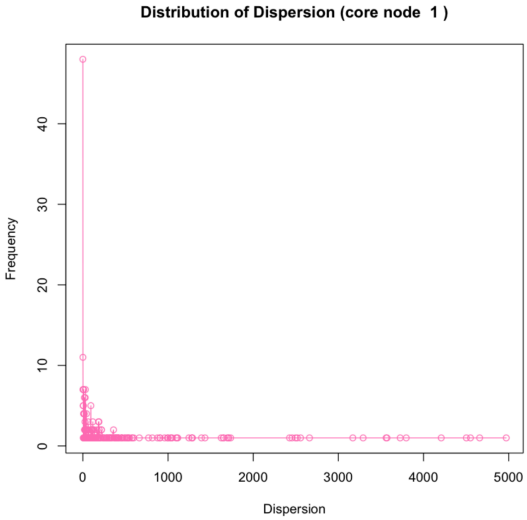
Embeddedness of a node is defined as the number of mutual friends a node shares with the core node. In the personalized network of a core node, since all other nodes are neighbors to the core node, the neighbors to a target node can be divided into two parts, one is the core node, the other is the group of common friends with the core node. Thus, the expression relating the Embeddedness of a node v_i in the personalized network of core node v_{core} is

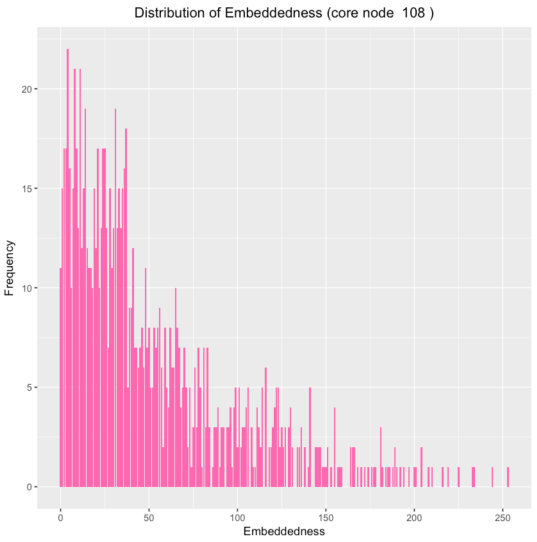
$$\text{embed}(v_i, v_{core}) = \deg(v_i) - 1$$

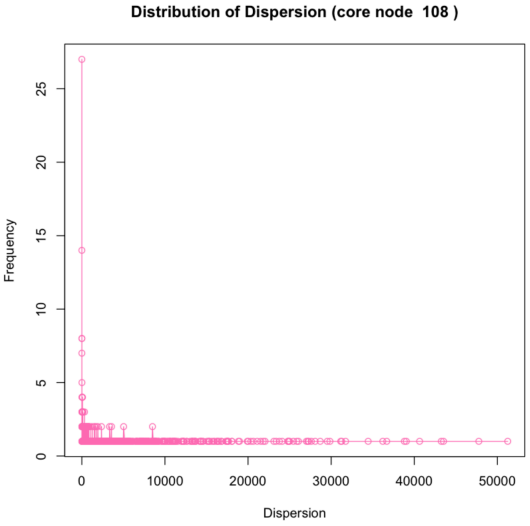
Question 12:

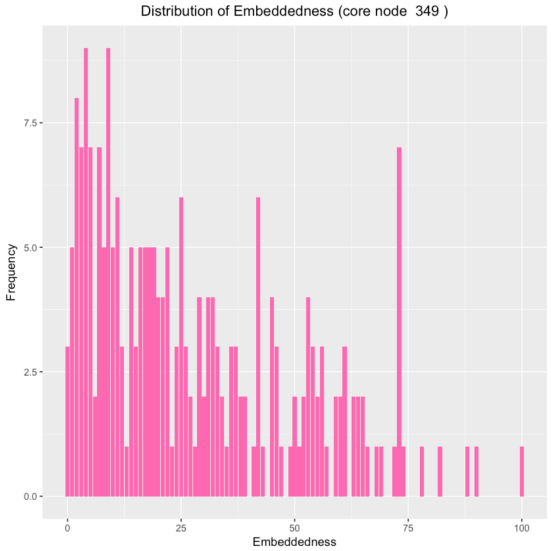
We use the same core nodes to analyze the distribution of embeddedness and dispersion. To be specific, by the definition of dispersion, we need to delete the core node and the target node in the personalized network. But after the deletion, the ids of nodes are rearranged from 1 again and we may lose the information of the original network. Here, we also record and set the labels of the nodes in the new graph, and use "label" attribute to position the node instead of the "id" attribute. Below are the distribution of embeddedness and dispersion in each core nodes' personalized network:

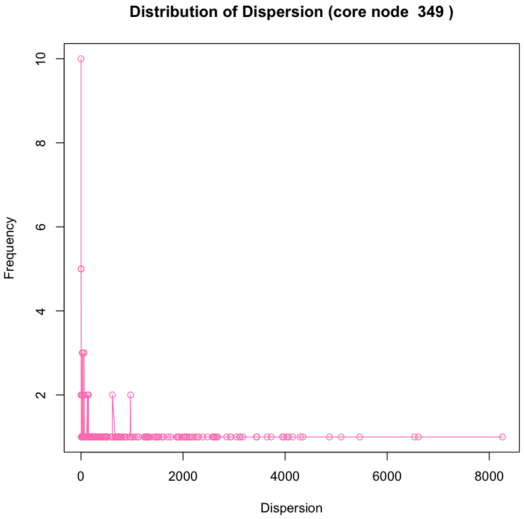

Distribution of Embeddedness (core node 1)
Frequency
Embeddedness
0
10
20
30
0
20
40
60
80


Distribution of Dispersion (core node 1)
Frequency
Dispersion
0
10
20
30
40
0
1000
2000
3000
4000
5000


Distribution of Embeddedness (core node 108)
Frequency
Embeddedness
0
5
10
15
20
0
50
100
150
200
250


Distribution of Dispersion (core node 108)
Frequency
Dispersion
0
5
10
15
20
25
0
10000
20000
30000
40000
50000


Distribution of Embeddedness (core node 349)
Frequency
Embeddedness
0.0
2.5
5.0
7.5
0
25
50
75
100


Distribution of Dispersion (core node 349)
Frequency
Dispersion
2
4
6
8
10
0
2000
4000
6000
8000

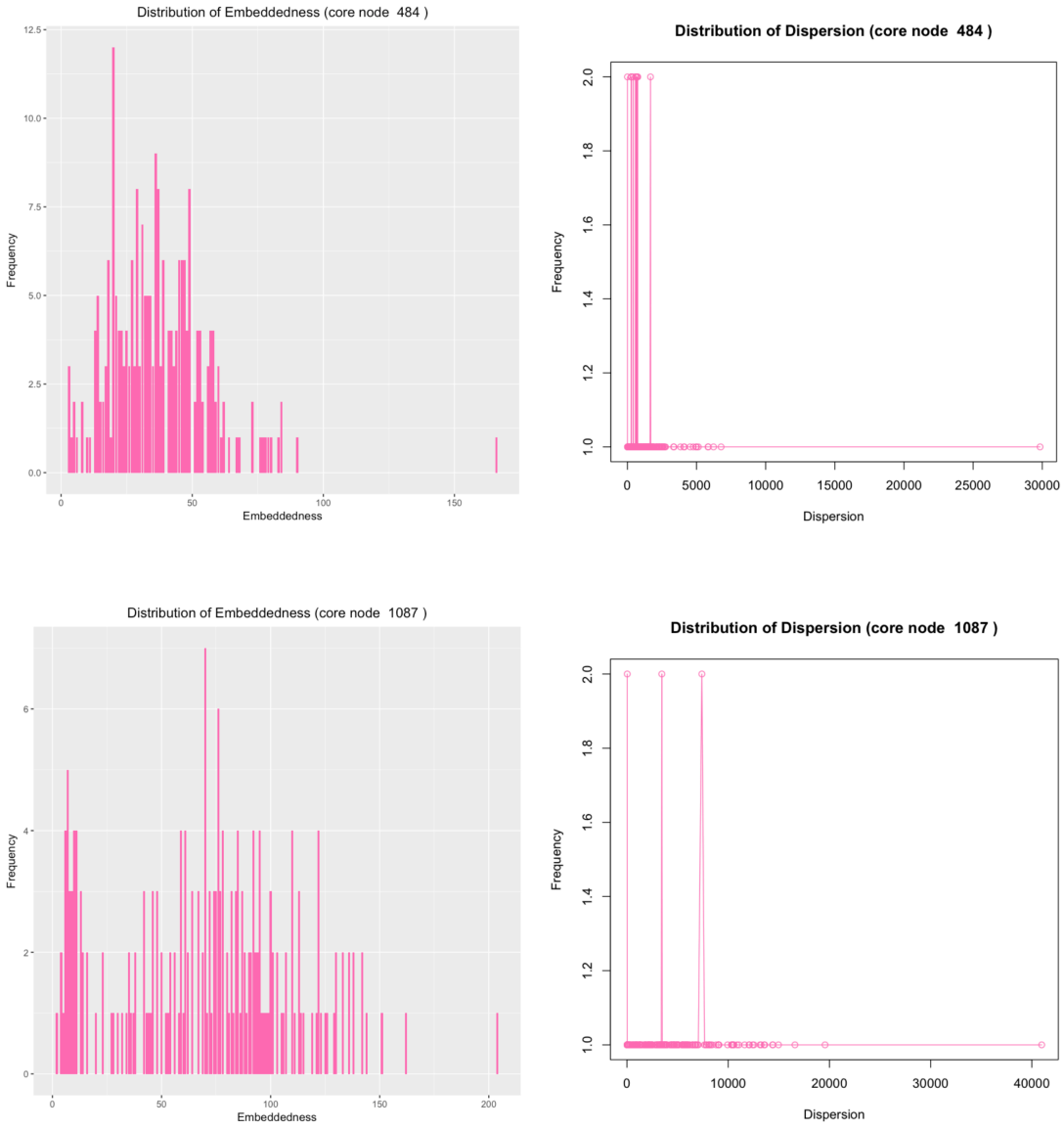


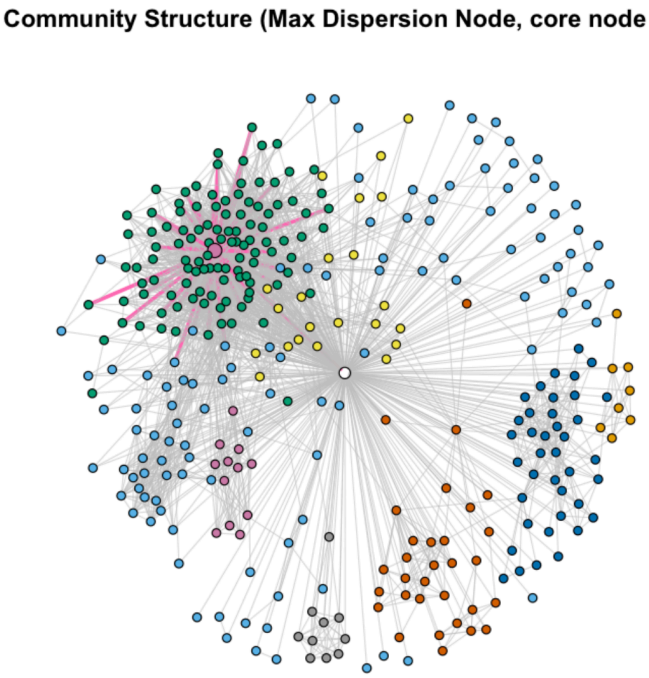
Figure 5. Distribution of embeddedness and dispersion

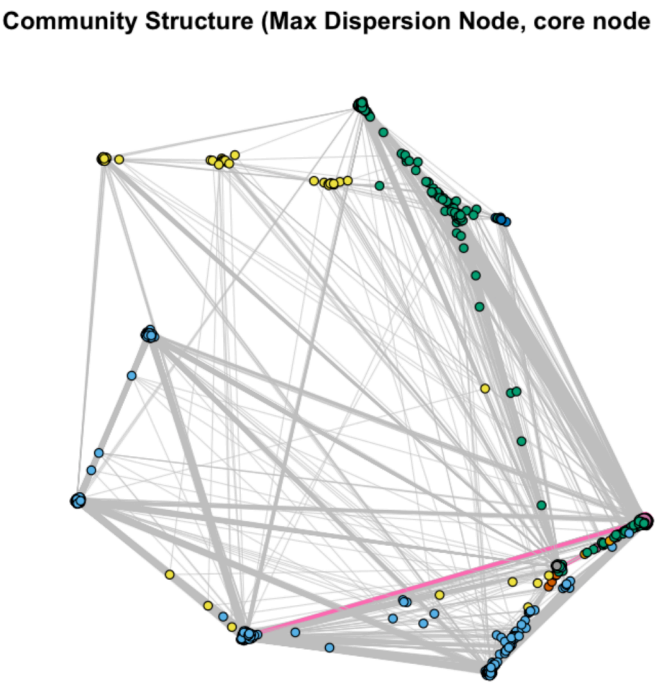
From the above figures, we find that the embeddedness of nodes in core node 108's and 1087's personalized networks are relatively larger than the other three with the maximum common friends of more than 200. It indicates that the nodes in these two personalized networks connect closer to each other (have strong ties) or these two personalized networks consist of larger clusters of friends corresponding to well-defined foci of interaction in their lives, e.g. co-workers, schoolmates, etc.

Besides, we also find that the dispersions of nodes in core node 108's, 484's and 1087's personalized networks are much larger than the other two with maximum sum of distance of more than 30000. It shows that the nodes in these three networks have wider personal social scales and their common friends spread more extensive.

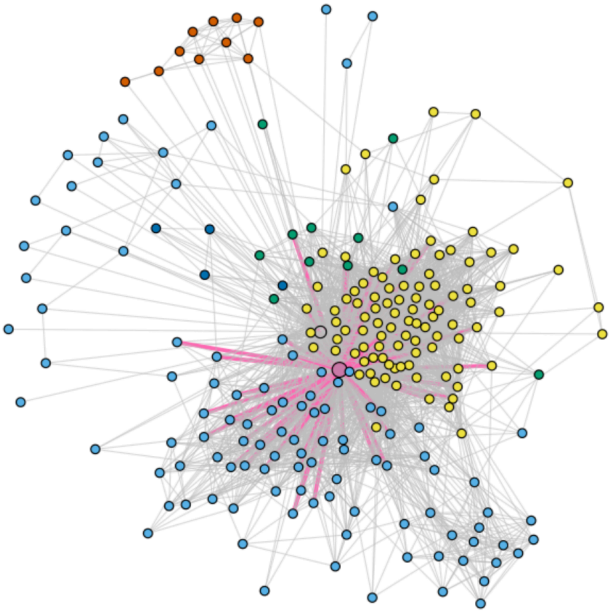
Question 13:

In this part, we use Fast-Greedy algorithm to detect community. Below are community structures of the five core nodes' personalized networks with maximum dispersion node and the incident edges highlighted: (the node with maximum dispersion is highlighted in pink, and the white node is the core node)

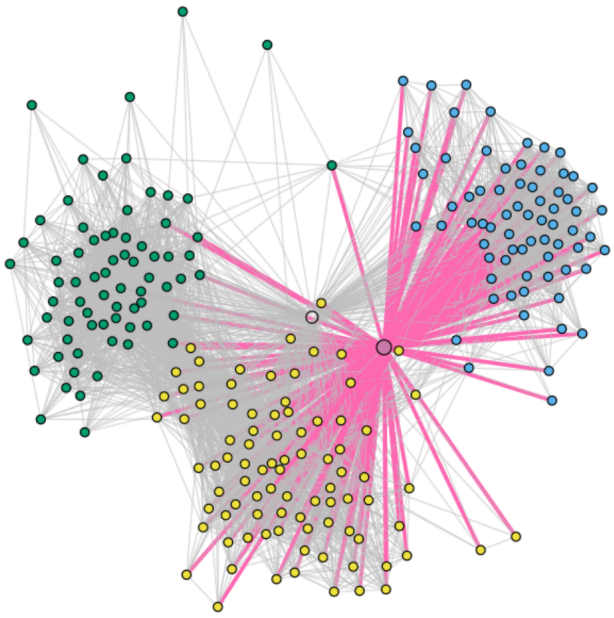




Community Structure (Max Dispersion Node, core node 349)



Community Structure (Max Dispersion Node, core node 484)



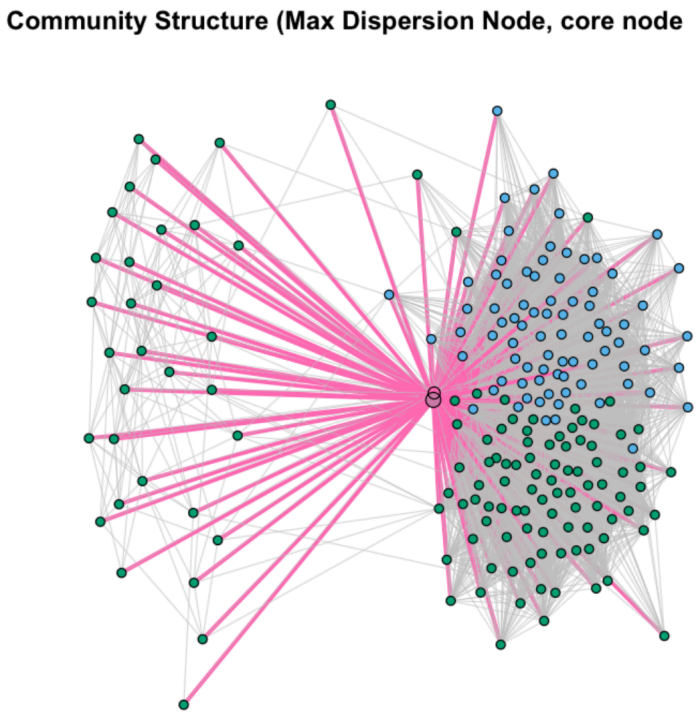
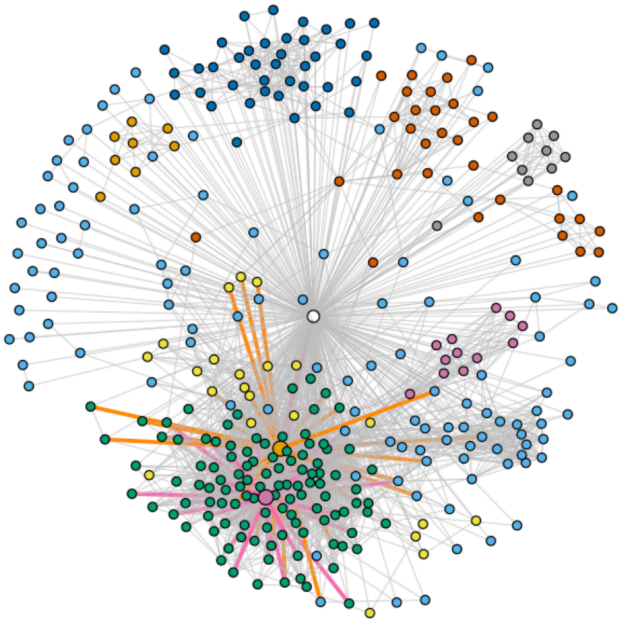


Figure 6. Community structure with maximum dispersion nodes and edges

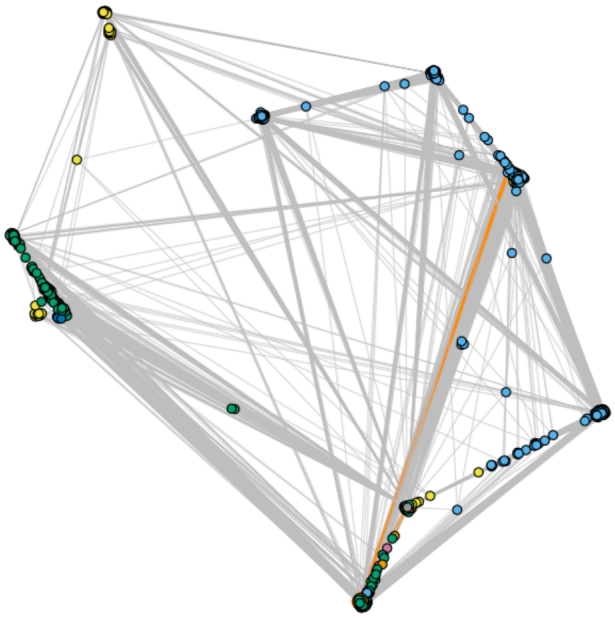
Question 14:

In this part, we also use Fast-Greedy algorithm to detect community. Here, we highlight the node with maximum embeddedness and the node with maximum $\frac{dispersion}{embeddedness}$, and also the edges incident to these nodes: (the pink node is the node with maximum embeddedness, the orange node with maximum $\frac{dispersion}{embeddedness}$, and the white node is the core node)

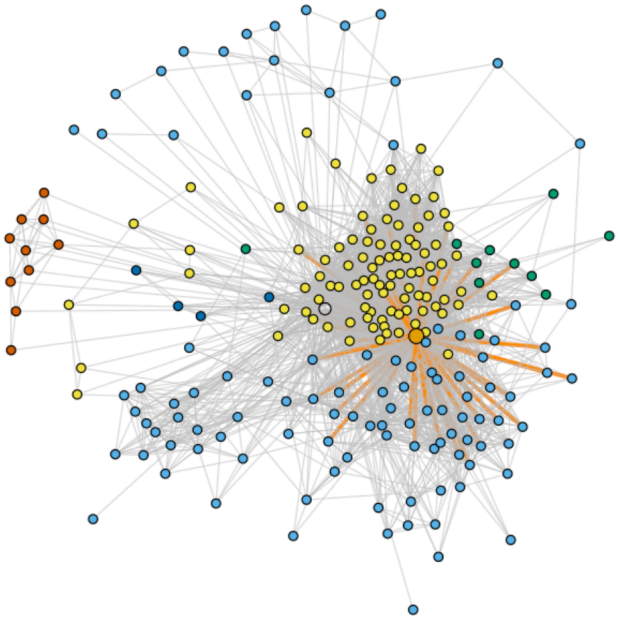
Community Structure (Max Embed & Disp/Embed Nodes, core node 1



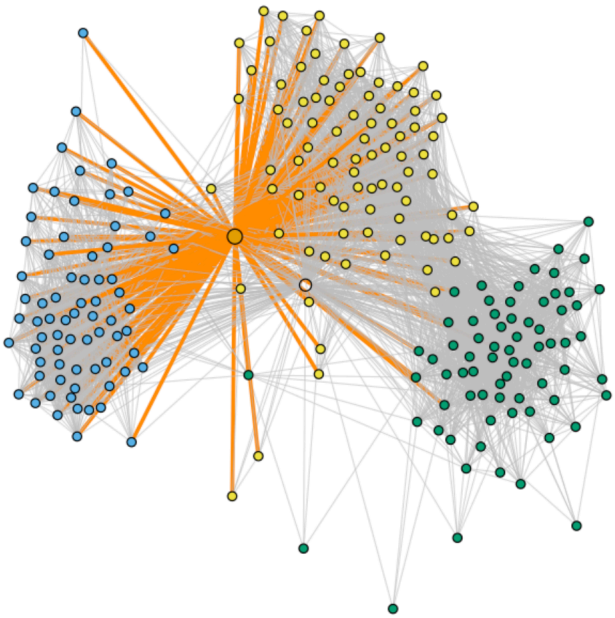
Community Structure (Max Embed & Disp/Embed Nodes, core node 10



Community Structure (Max Embed & Disp/Embed Nodes, core node 349



Community Structure (Max Embed & Disp/Embed Nodes, core node 484



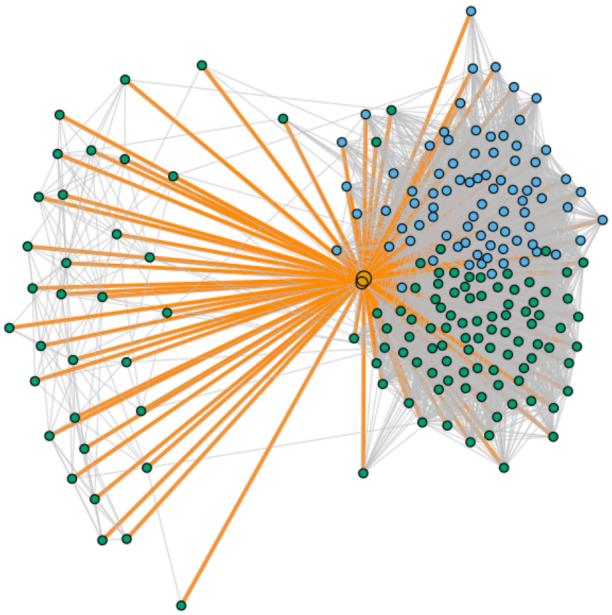


Figure 7. Community structure with maximum embeddedness and $\frac{dispersion}{embeddedness}$ nodes and edges

Question 15:

From the results from question 13 and 14, we can conclude that:

1) Node with maximum dispersion

   The nodes with maximum dispersion indicate that they are dispersed, which means that the common friends shared by the target node and the core node are not well connected or from different circles. We can compare it to some real cases in our daily life. There are two friends from the same primary school and they are connected on Facebook. While they didn't take same high school and university later. Then we can induce that even though they have some common friends except their primary schoolmates, the common friends may from different areas or fields, and they don't know each other.

2) Node with maximum embeddedness

   The embeddedness value reflects the strength of ties between the target node and

the core node. The node with maximum embeddedness is generally closed related to the core node because they share many common friends. Besides, we can also induce that the node with maximum embeddedness is likely from a large group and nodes in this group are well-connected. This can be verified from the figures in question 14 that the node with maximum embeddedness almost from the largest communities.

However, we cannot tell how close the two nodes are only from embeddedness value, because they may just from the same circle in which all nodes are well-connected. For example, two students from the same college may share 100 common friends since they are in the same social circle and people in this circle know each other well. But they may be just-known classmates, not close friends.

3) Node with maximum $\frac{dispersion}{embeddedness}$

From the results in question 14, we find that the personalized network of core node 1 has different nodes with maximum embeddedness and $\frac{dispersion}{embeddedness}$, while the networks of other 4 core nodes have the same nodes with maximum embeddedness and $\frac{dispersion}{embeddedness}$. Thus, the $\frac{dispersion}{embeddedness}$ value is a normalized feature to describe the relation between two people.

The node with maximum $\frac{dispersion}{embeddedness}$ has large dispersion value and small embeddedness value. In other words, the node doesn't share many common friends with the core node, and their common friends may from different areas and not well-connected. This value can be used to judge how likely two nodes to be in romantic relationship. The higher the value is, the more likely two nodes to be potential romantic tie.

1.4 Friend Recommendation in Personalized Networks

Question 16:

In this part, the network we are working with is the personalized network of node ID 415 within the Facebook Network. Therefore, we firstly get the neighbors of ID 415, and we create a subgraph with all these neighbor nodes and node ID 415 as well. Then, we construct a list of users who we will recommend friends to and test the accuracy of friend recommendation algorithms in the next question. This nodes in the list satisfy the requirement that their degrees are all 24.

The result shows that the length of the list is **|N_r| = 11**.

Question 17.

For each user in the list N_r above, we randomly remove its neighbor with the probability 0.25. The removed neighbors form a list R_i. Then, with the three friend recommendation algorithms, we recommend |R_i| friends, which form a list of P_i, to the current user i. Finally, the accuracy for the user i for the iteration can be calculated. We do this procedure 10 times and get the average accuracy for each user. The final accuracy is the average of all the users' average accuracy results.

The results for the three algorithms are:

Common Neighbors Measure: accuracy_common_neighbours_measure = **0.8599**

Jaccard Measure: accuracy_jaccard_measure = **0.8647**

Adamic Adar Measure: accuracy_adamic_adar_measure = **0.8715**

With all the average accuracy values above, we can find that these three methods are all very effective and have relatively high accuracy values. Based on the average accuracy values, the Adamic Adar Measure has the best performance.

2. Google+ Network

Question 18:

In this part, we will use the directed Google+ Network. Since the required directed personal networks for users should have more than two circles, we need to check all the ".circle" files in "gplus.tar.gz" and see if there are more than two lines in each file.

The result shows the total number of personal networks is **57**.

Question 19:

In this question, we need to plot the in and out degree distribution for three given nodes. Using the ".edges" files, we can create the subgraph for each personal network. In this way, the indegree and outdegree distribution can be plotted using the functions in "igraph" library.

The plots for the node "109327480479767108490" are:

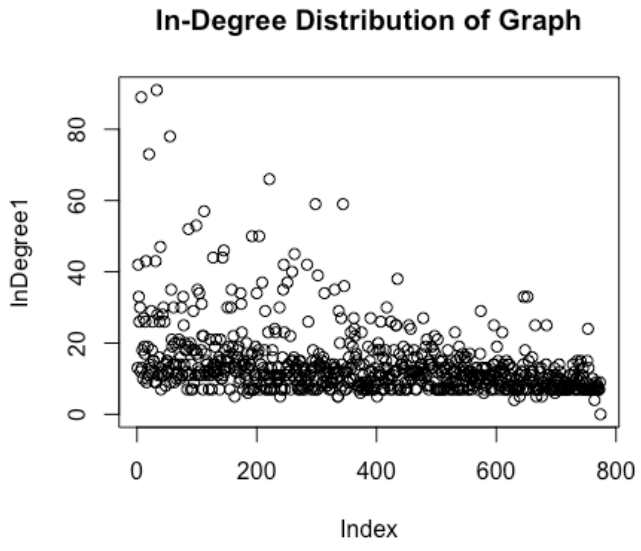


Figure 8. Indegree Distribution for node "109327480479767108490"

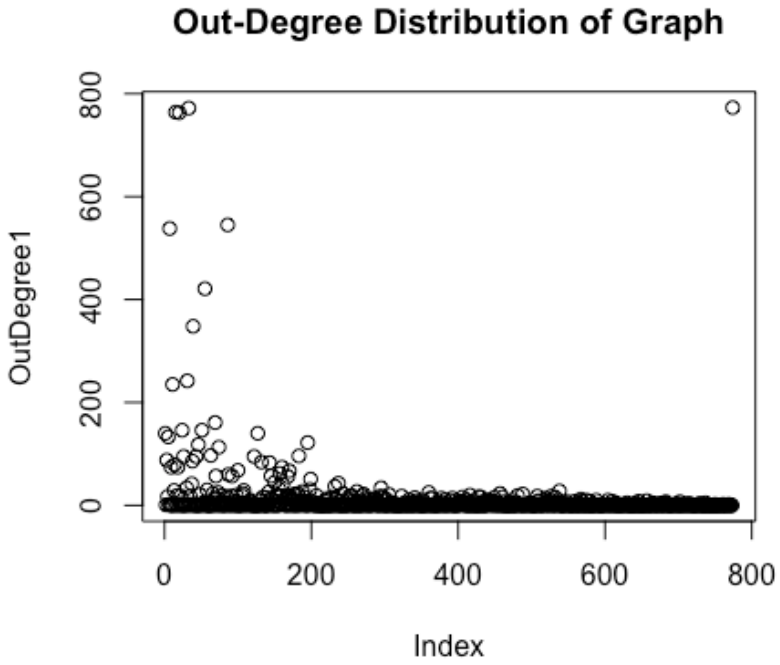


Figure 9. Outdegree Distribution for node "109327480479767108490"

The plots for the node "115625564993990145546" are:

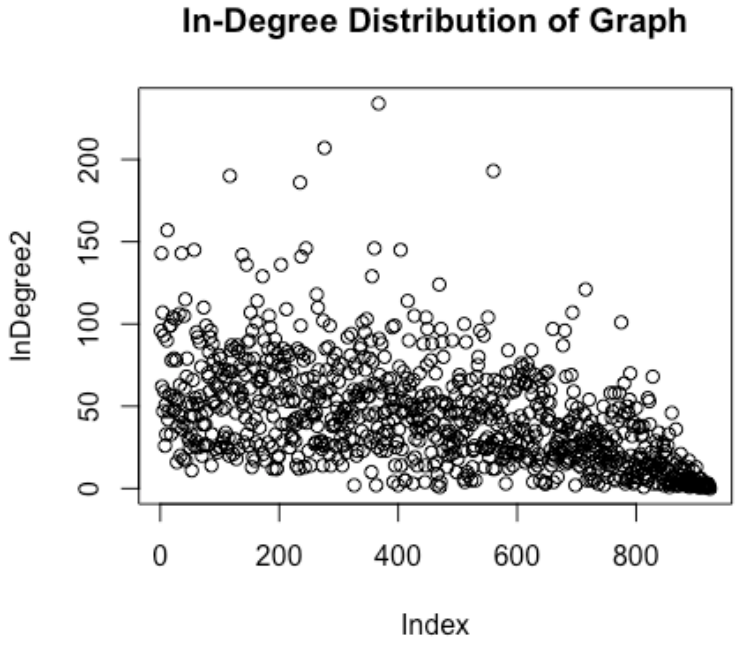


Figure 10. Indegree Distribution for node "115625564993990145546"

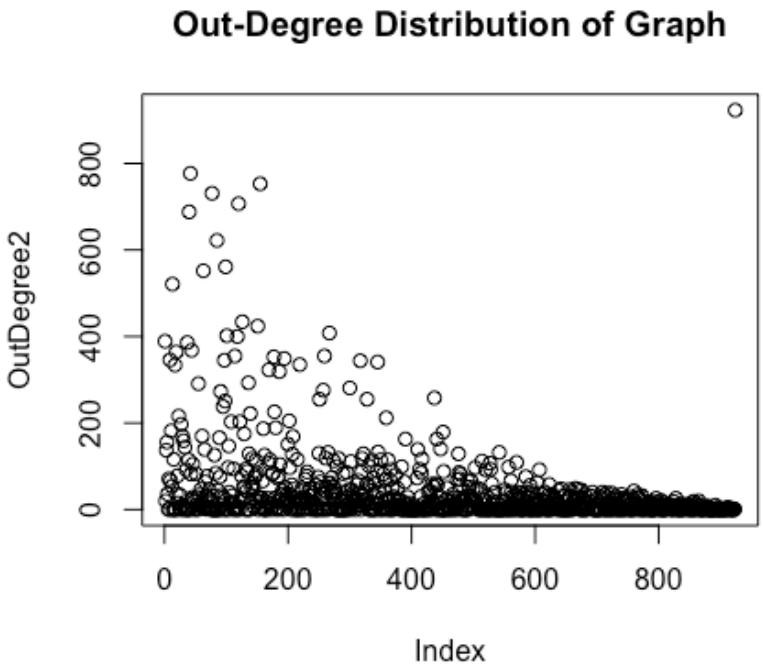


Figure 11. Outdegree Distribution for node "11562556499399014554 6"

The plots for the node "101373961279443806744" are:

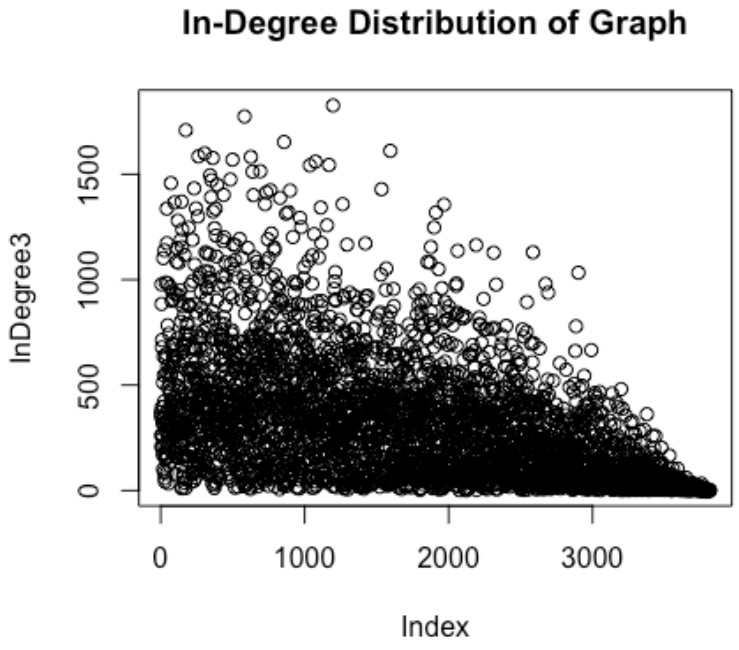


Figure 12. Indegree Distribution for node "101373961279443806744"

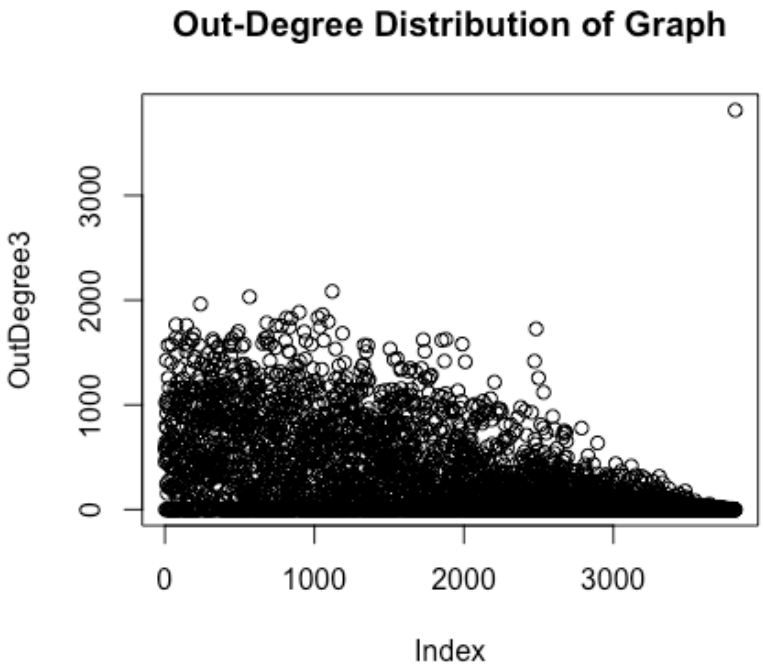


Figure 13. Outdegree Distribution for node "10137396127944380674 4"

The three distributions are a bit different in the following aspect. We can also find that most of the nodes have higher indegree and outdegree values in the second network than in the first network. Also, the nodes in the third network have the highest indegree and outdegree values. Therefore, the second network has stronger connections between the nodes than the first network, and the third network has the strongest connections between the nodes. So, the third network is the most complex, while the first network is the simplest.

However, we can find from the previous graphs that the three personal networks have quite similar indegree and outdegree distribution for each node ID (the last node index in the graphs) of the personal network. The node ID has a very high out-degree and an approximate zero in-degree. This is because for each node ID of that personal network, the node is following all the other nodes in the graph, so the outdegree is equal to the number of nodes in that network.

Question 20:

After extracting the community structures for the three nodes in Question 19, we use walk-trap community detection algorithm to get the modularity scores. The modularity scores for the three nodes are shown below:

Modularity score for node "109327480479767108490" is: **0.2528**.

Modularity score for node "115625564993990145546" is: **0.3195**.

Modularity score for node "101373961279443806744" is: **0.1911**.

Therefore, the modularity score for the second node is the highest, while the score for the third node is the lowest.

A higher modularity score stands for the meaning that the nodes belonging to the same community have stronger connections, and the nodes belonging to different community have weaker connections.

The modularity of these three nodes can be interpreted as below. The second node has the highest modularity score because the connections within communities are strong and the connections between communities are weak, which results into a better community clustering result. The third node, however, has a lowest score. This is because its connections are too complex (as described in Question 19), and therefore the connections between the communities are also strong, thus resulting in a low modularity score.

Then, we also plot the three communities with colors below.

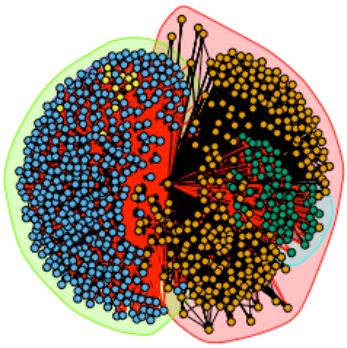


Figure14. Community Structure for node "109327480479767108490"

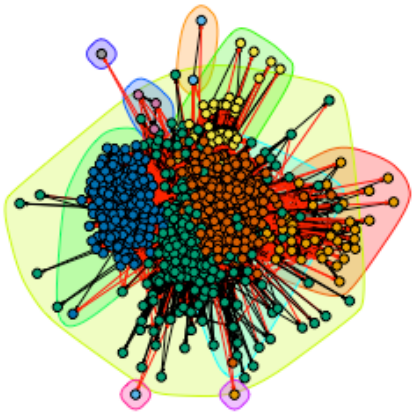


Figure 15. Community Structure for node "115625564993990145546"

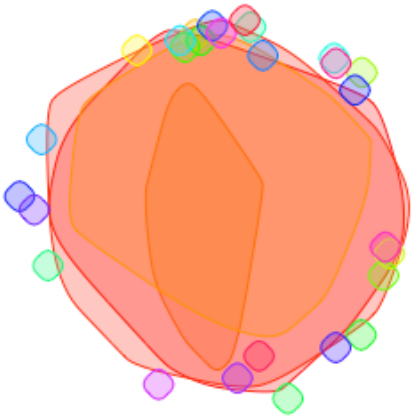


Figure 16. Community Structure for node "101373961279443806744"

Question 21:

The question aims at explaining the meaning of both homogeneity and completeness based on the formula given in the project description.

Homogeneity is to measure the purity of the community structure. When the community is composed of nodes coming from the same circle, the homogeneity reaches a higher score.

Completeness is to measure the purity of the circle. When the circle assigns nodes to the same community, the completeness scores become higher.

Question 22:

The homogeneity and completeness values for the three community structures are:

For node "109327480479767108490", the homogeneity value is **0.5249**, the completeness value is **0.5497**.

For node "115625564993990145546" the homogeneity value is **0.1060**, the completeness value is **0.3822**.

For node "101373961279443806744", the homogeneity value is **0.0003745**, the completeness value is **0.0008590**.

Therefore, the first node has the highest homogeneity and completeness values, while the third node has the lowest homogeneity and completeness values. Based on the meaning of homogeneity and completeness in Question 21, this means the first node best forms communities within the node circle, and nodes from this circle also tends to form the same community. On the contrary, for the third node, it worst forms the communities with nodes coming from that circle, and nodes from the circle form different communities as well.