ECE 232E Project 2
Social Network Mining
Spring 2018

Cyrus Tabatabai-Yazdi (405029242)

Zongheng Ma (905027293)

Akshay Shetty (905028886)

Bakari Hassan (705035029)

Table of Contents

1. Facebook network
    1.1 Structural properties of the facebook network
    1.2 Personalized network
    1.3 Core node's personalized network
        1.3.1 Community structure of core node's personalized network
        1.3.2 Community structure with the core node removed
        1.3.3 Characteristic of nodes in the personalized network
    1.4 Friend recommendation in personalized networks
        1.4.1 Neighborhood based measure
        1.4.2 Friend recommendation using neighborhood based measures
        1.4.3 Creating the list of users
        1.4.4 Average accuracy of friend recommendation algorithm
2. Google+ network
    2.1 Community structure of personal networks

1. Facebook network

1.1 Structural properties of the facebook network

In this section, several structural properties of the Facebook network were studied, mainly connectivity and degree distribution. The GCC size and diameter were calculated, as well as the degree distribution.

Question 1

Yes, the Facebook network used in this problem is connected. As shown in the code. For given survey, we found that the network is connected as is_connected(graph) is true. However, we are not sure if this holds true for the real world facebook network as there could be groups of friends connected to each other and not connected to the rest of the network at all. So in the case of our survey, the GCC (Giant Connected Component) is the network itself.

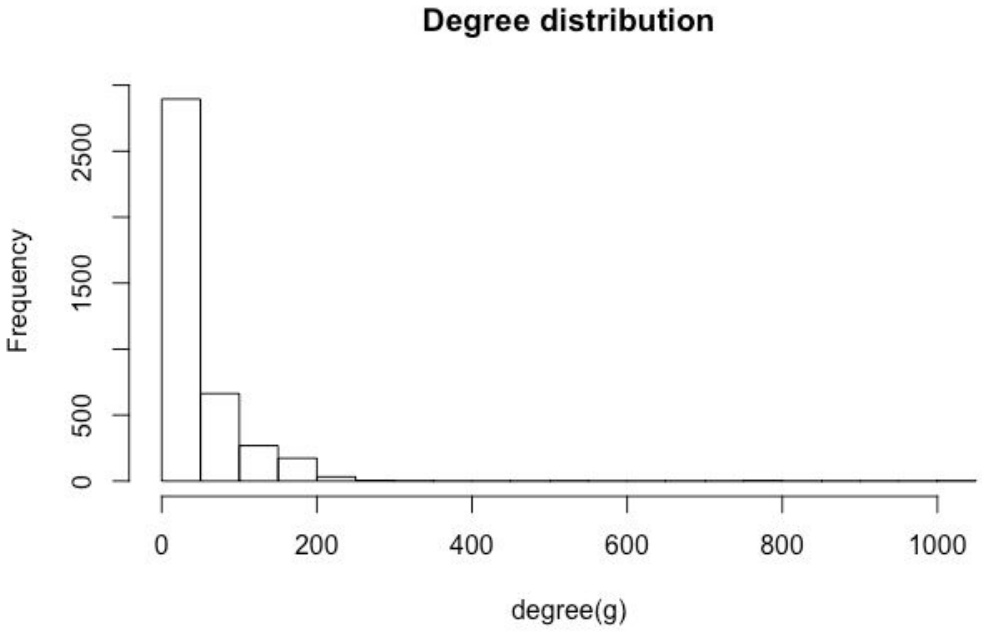
Question 2

The diameter of the GCC is 8. Since the network is connected, the GCC is the entire network. This hints at the general theory of 6 Degrees of Separation. However, in this case, there are 8 degrees of separation, meaning any two randomly selected nodes in the network are separated by a maximum of 8 edges.

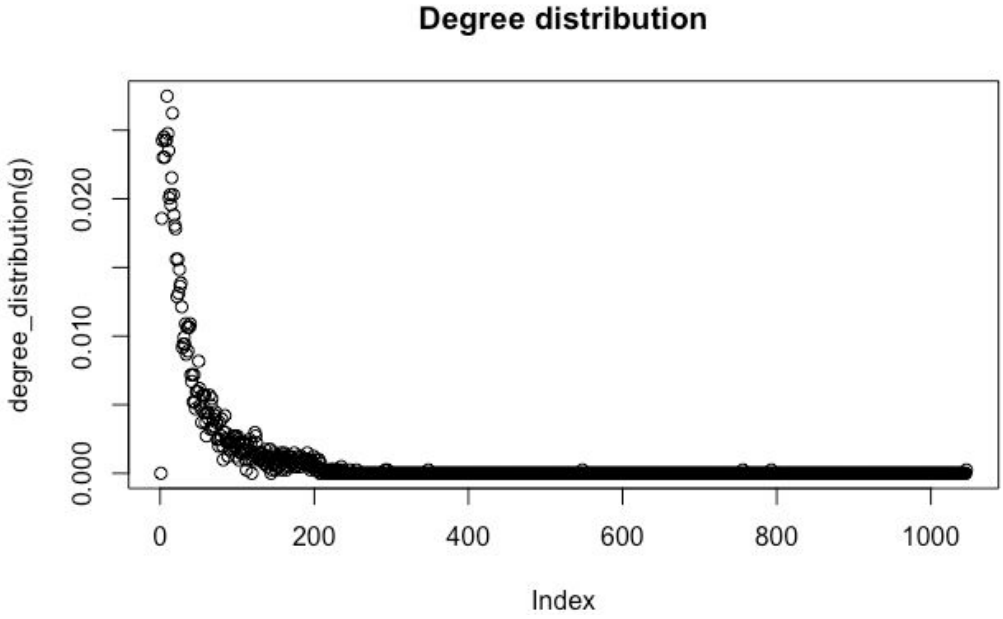
Question 3

The distribution of degrees is show in the histogram(top) and plot(bottom) in the figure below. The average degree is 43.69101. In average, every person in the network has around 44 friends. The degree distribution is exponential. There are fewer people that have many friends(the highest number is around 1000 as shown in the graph), but lots of users that have just a few friends.

The degree histogram indicates that a large number of nodes in the network are of low degree, with just a few nodes having degree greater than 200. This agrees with expectation of network evolution, resulting in core nodes/hubs comprising the backbone of the network.



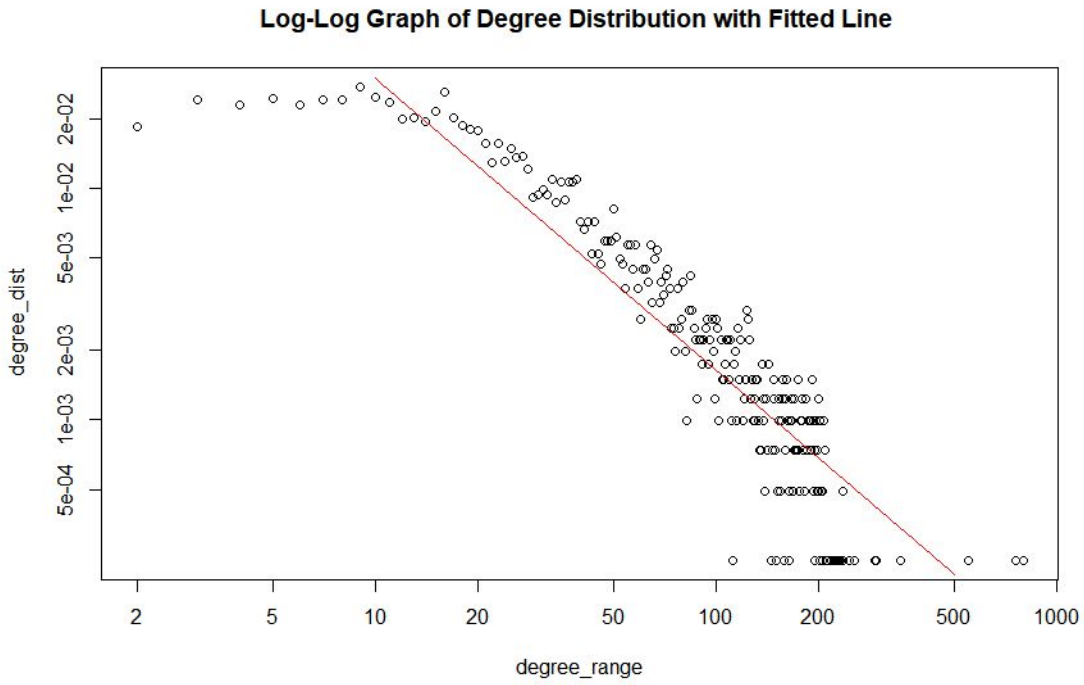
The degree distribution is consistent with the histogram and resembles a power law distribution.



Question 4

The degree distribution shown above is shown below on a log-log scale with linearly fitted line. The line of best fit was attained using least squares regression. The fit for the degree distribution is as follows, where k is the node degree:

$$\ln\left(P\left(k\right)\right) = -1.2\ln k + 0.61$$



1.2 Personalized network

In this section, a personal (ego) network was generated for a specific node. This personalized network was limited to all neighbors within 1 degree of the ego node. All connections between two nodes that are both within the network are maintained.

Question 5

Node 1's personalized network has 348 nodes, including ego node 1, meaning node 1 has 347 friends. This personal network has 2,866 edges, which corresponds to 5,732 degrees in total.

Question 6

The diameter of the personalized network is 2. We would expect the diameter to be 2 because all of the nodes are connected to the core node so, in the worst case, a node can get to any other node by going through the core node and then to any node. If the network is fully connected, the diameter is expected to be 1. The diameter for the personalized network can be a minimum of 1 for lower bound and a maximum of 2 for upper bound.

Question 7

If the diameter of the network is 1 (lower-bound), the personalized is fully connected. It implies that all of the friends of the central node are also friends with all of the other friends of the central node.

If the diameter of the network is 2 (upper-bound), there is at least one of the central node's friend is not a friend of at least one of other friends of the central node.

1.3 Core node's personalized network

Section 1.3 focuses on on the social characteristics of core nodes' personalized networks. This includes community structure, modularity, embeddedness, and dispersion. These characteristics and measures will allow us to make inferences regarding interpersonal relationships between users.

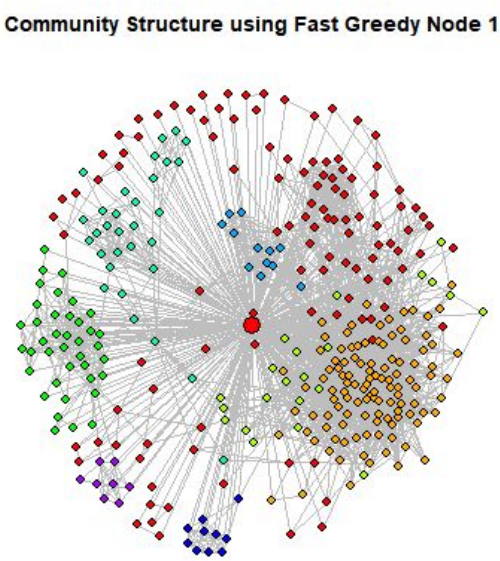
Question 8

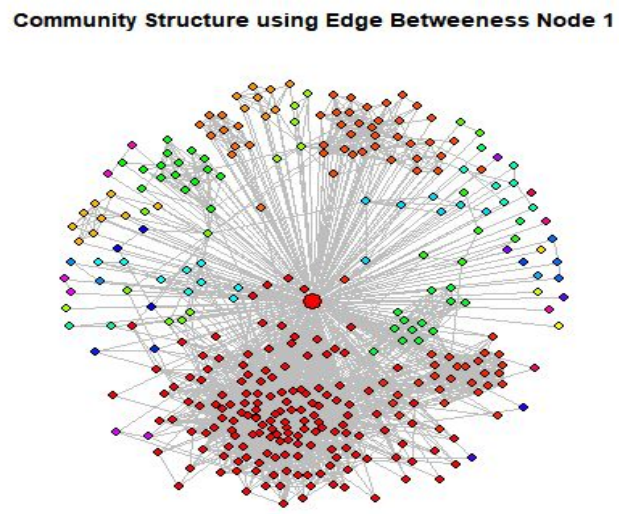
There are 40 core nodes in the facebook network with an average degree of 279.375. If we multiply 40 by the average degree we got, we found out that number is around 12% of the total number of edges of the entire graph. Although this number is not accurate due to the fact that those nodes may be connected to each other so there will be edges that we count twice, this can provide a roughly estimate.

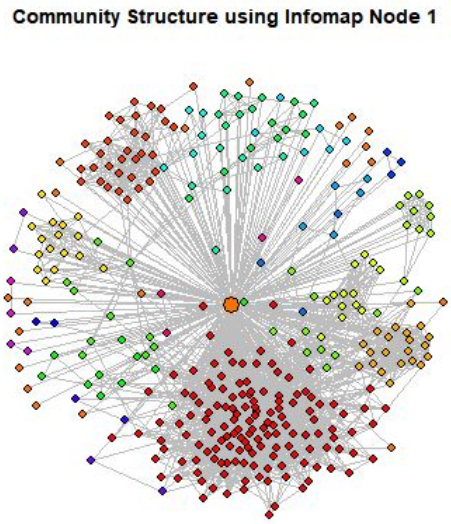
1.3.1 Community structure of core node's personalized network

Question 9:

Community structures for core node 1's personalized network Shown below



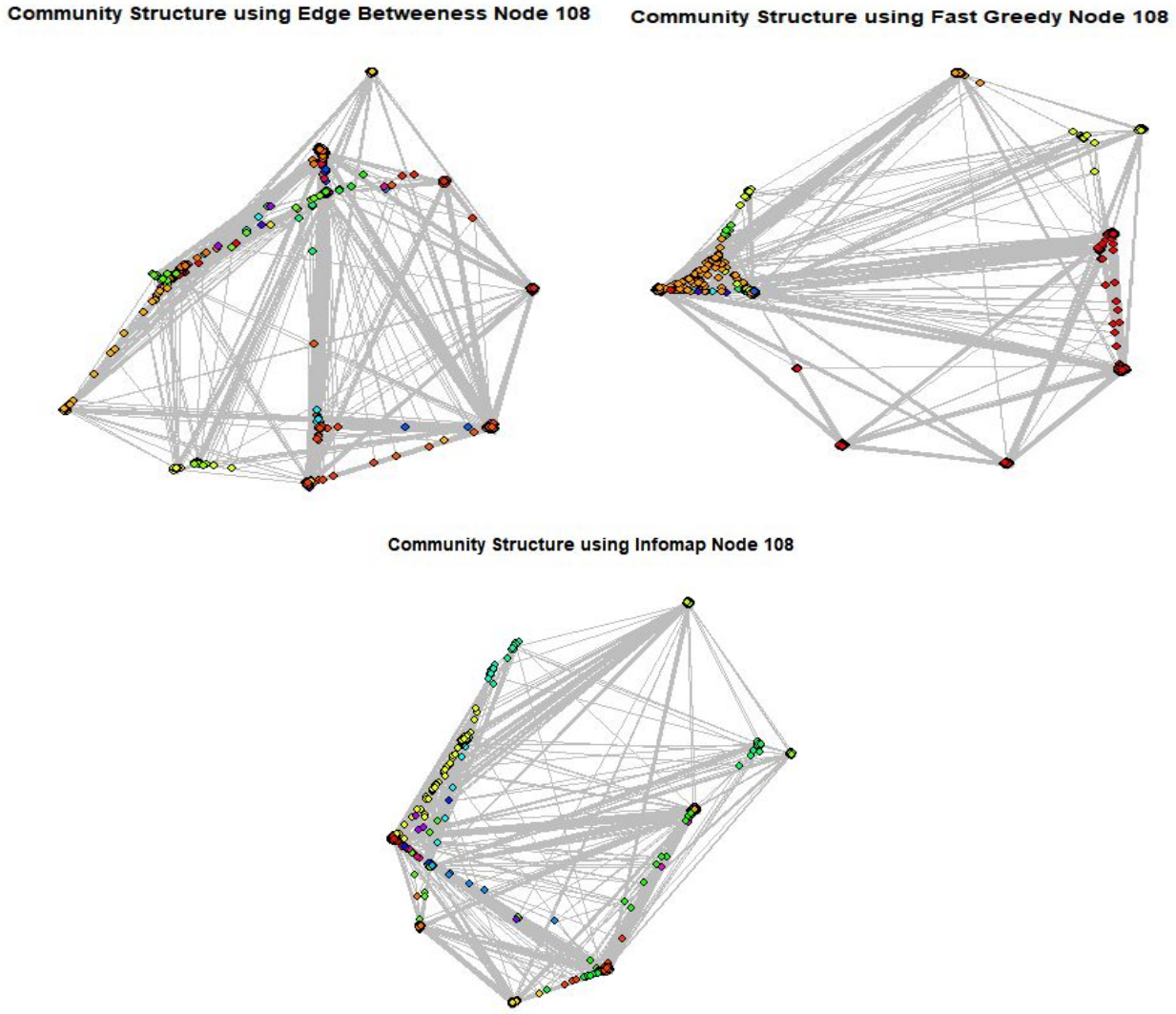




Modularity and Number of Communities for Ego Network of Node 1 for Different Algorithms

Community Algorithm Used	**Modularity**	**Number of Communities**
Fast Greedy	0.4131	8
Edge Betweenness	0.35330	41
InfoMap	0.38911	26

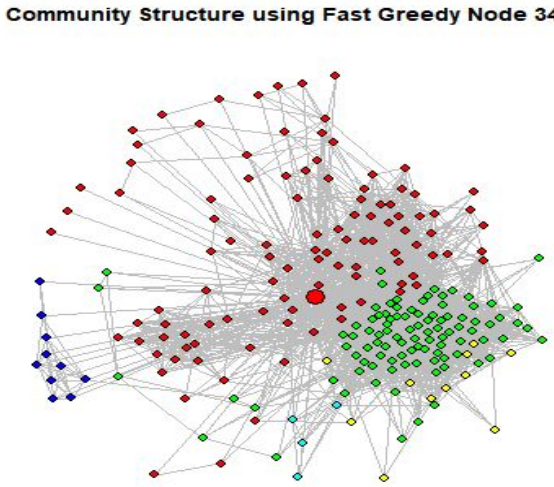
Community structures for core node 108's personalized network Shown below

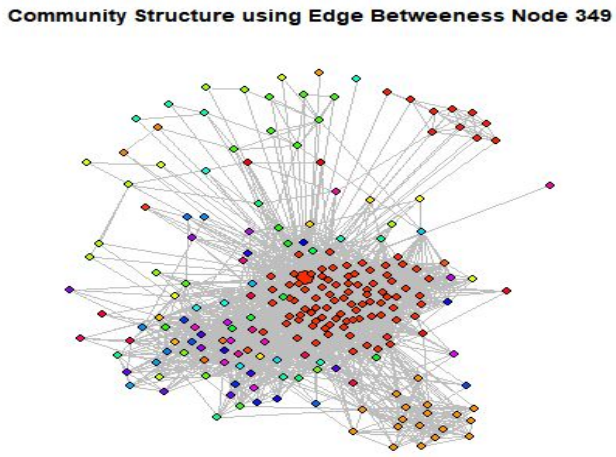


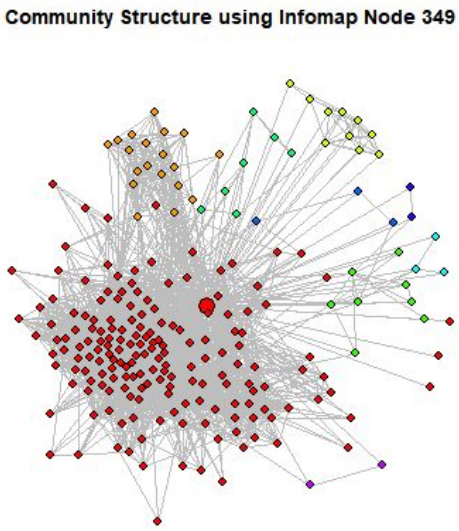
Modularity and Number of Communities for Ego Network of Node 108 for Different Algorithms

Method	**Modularity**	**Number of Communities**
Fast Greedy	0.4359293	9
Edge Betweenness	0.50675	52
InfoMap	0.50824	27

Community structures for core node 349's personalized network Shown below



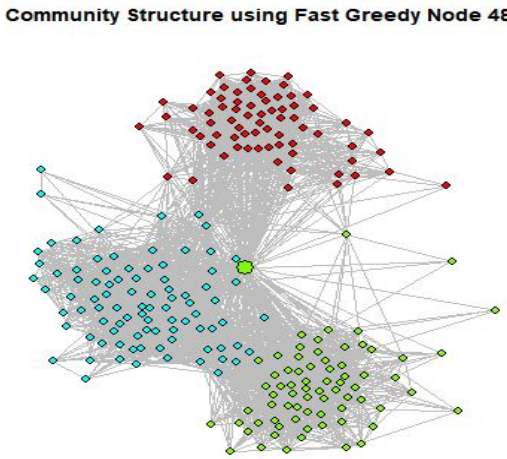


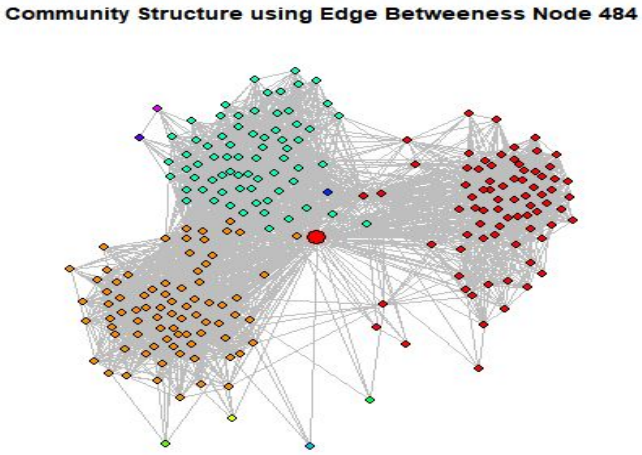


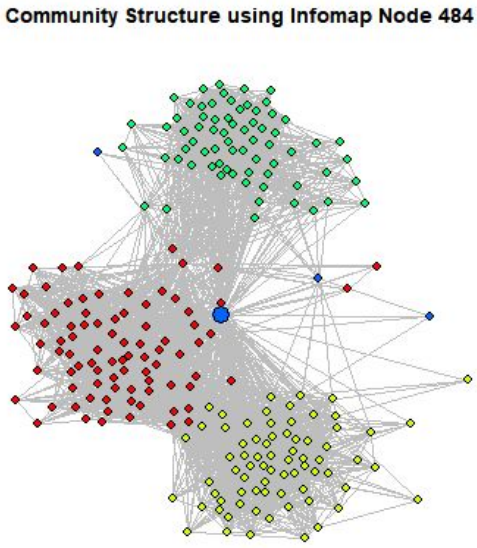
Modularity and Number of Communities for Ego Network of Node 349 for Different Algorithms

Method	**Modularity**	**Number of Communities**
Fast Greedy	0.2517148	5
Edge Betweenness	0.13335	104
Infomap	0.0954	9

Community structures for core node 484's personalized network Shown below



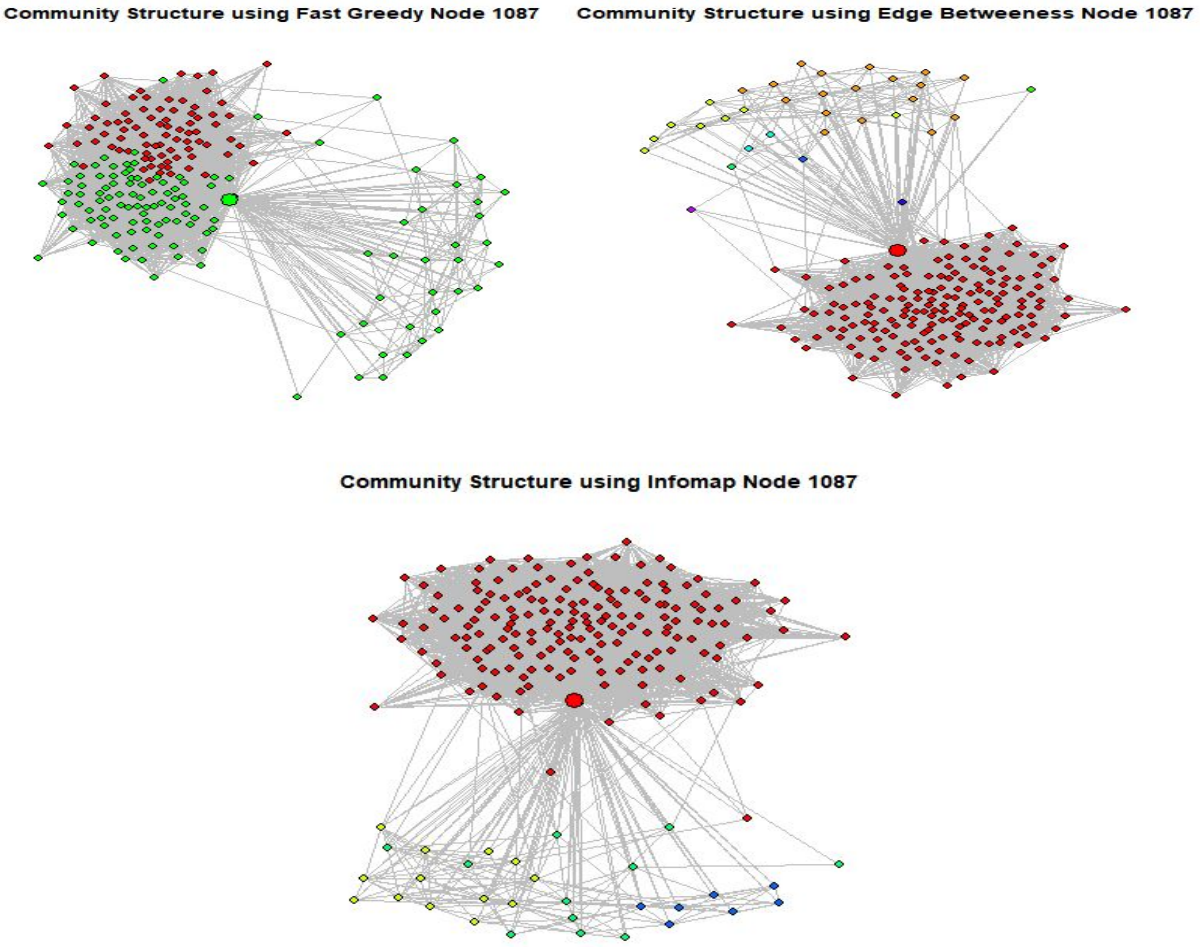




Modularity and Number of Communities for Ego Network of Node 484 for Different Algorithms

Method	**Modularity**	**Number of Communities**
Fast Greedy	0.50700	3
Edge Betweenness	0.48909	10
InfoMap	0.51528	4

Community structures for core node 1087's personalized network Shown below



Modularity and Number of Communities for Ego Network of Node 1087 for Different Algorithms

Method	**Modularity**	**Number of Communities**
Fast Greedy	0.14553	2
Edge Betweenness	0.027623	9
Infomap	0.02690	4

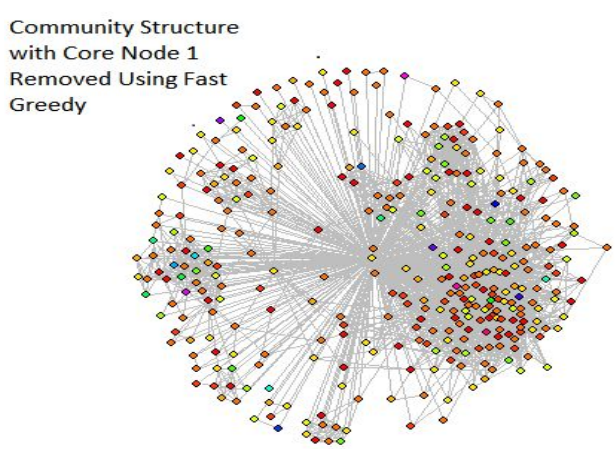
Analysis of Results:

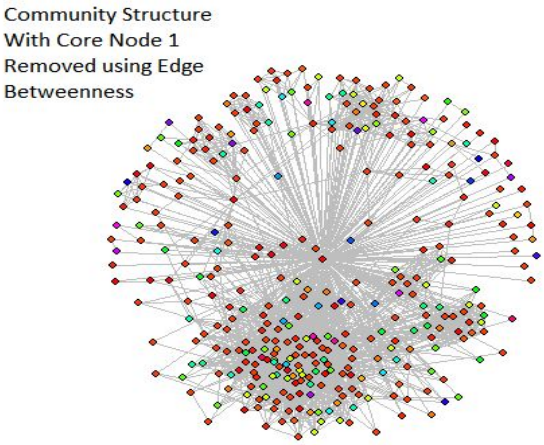
1. It seems that edge-betweenness algorithm divides the graph up into many more communities compared to fast greedy and infomap as can be seen in the tables above that summarizes the number of communities generated by each algorithm for each core node.
2. Due to the fact that edge-betweenness is a divisive algorithm, the high modularity of the network is a result of the high number of communities generated by the semantics of the algorithm.
3. It seems that for core node 1087 and core node 349, they appear in the maximal community, in other words the community with the highest number of nodes.

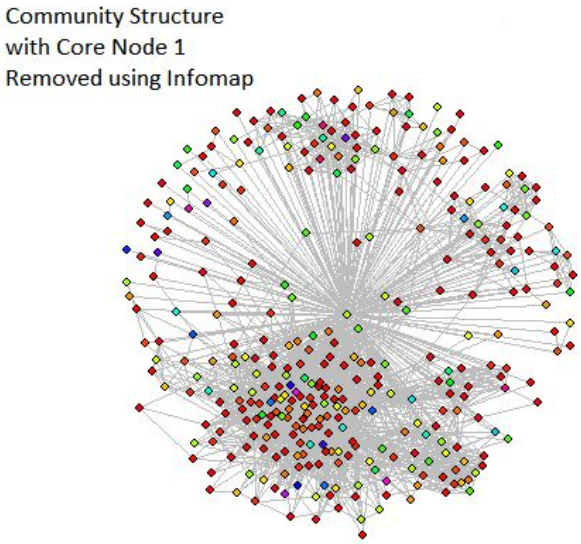
1.3.2 Community structure with the core node removed

Question 10:

Community Structures with Node 1 Removed Shown Below



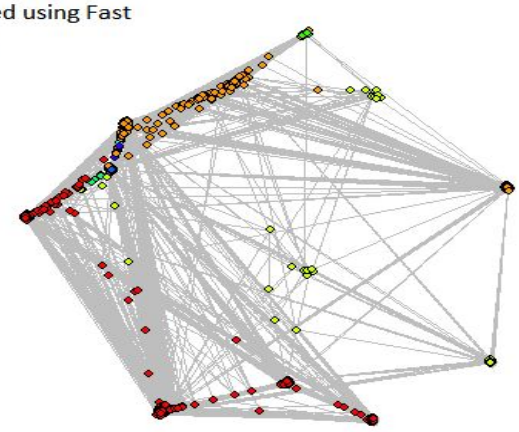




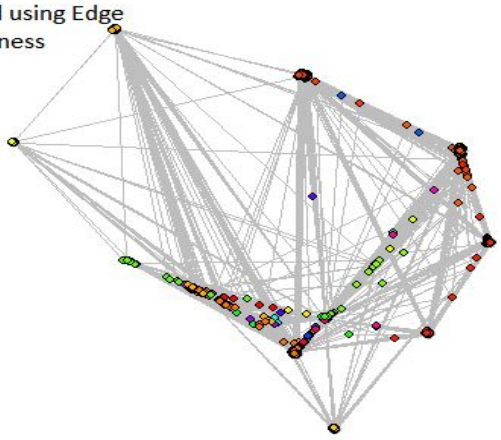
Method	**Modularity**	**Number of Communities**
Fast Greedy	0.44185	26
Edge Betweenness	0.41615	50
Infomap	0.41801	40

Community Structures with Node 108 Removed Shown Below

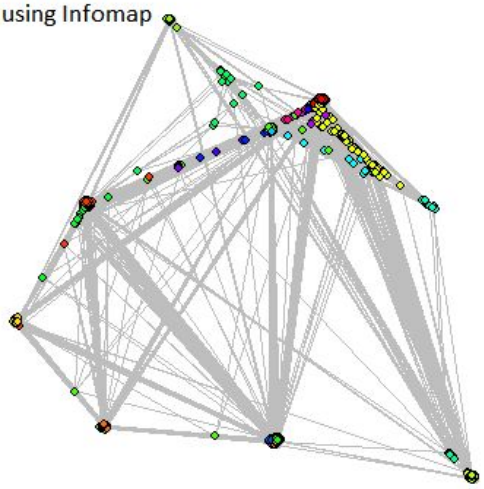
Community Structure with Core Node 108 Removed using Fast Greedy



Community Structure with Core Node 108 Removed using Edge Betweenness



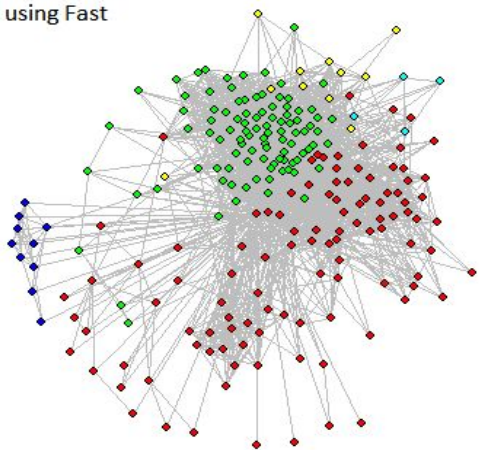
Community Structure with Core Node 108 Removed using Infomap



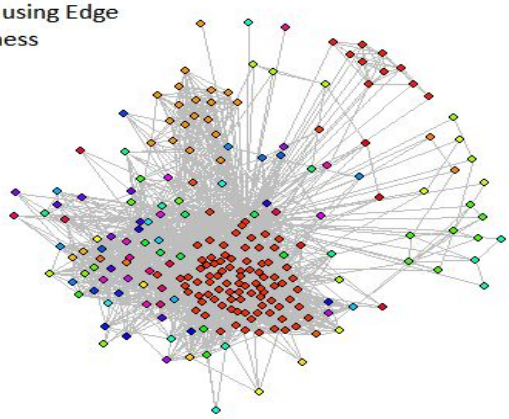
Method	**Modularity**	**Number of Communities**
Fast Greedy	0.43592	9
Edge Betweenness	0.50675	52
Infomap	0.50822	27

Community Structures with Node 349 Removed Shown Below

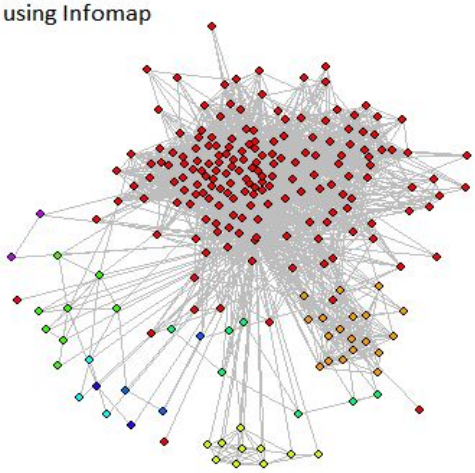
Community Structure with Core Node 349 Removed using Fast Greedy



Community Structure with Core Node 349 Removed using Edge Betweenness

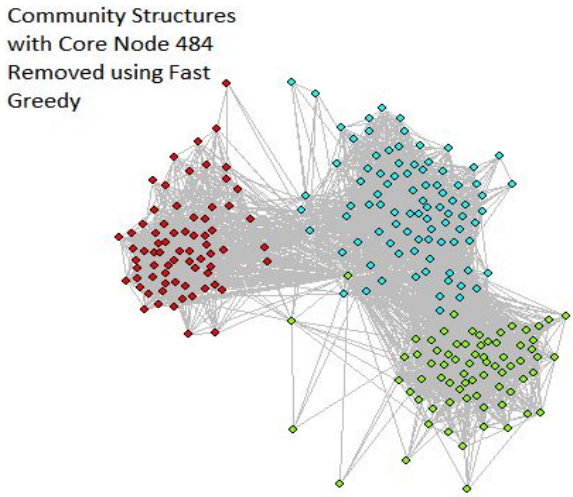


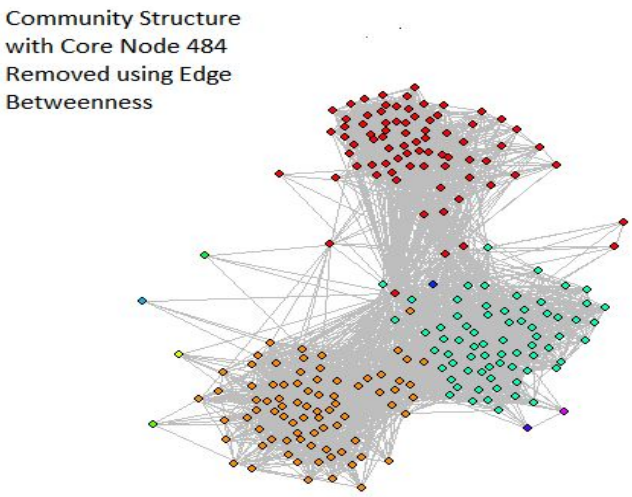
Community Structure with Core Node 349 Removed using Infomap

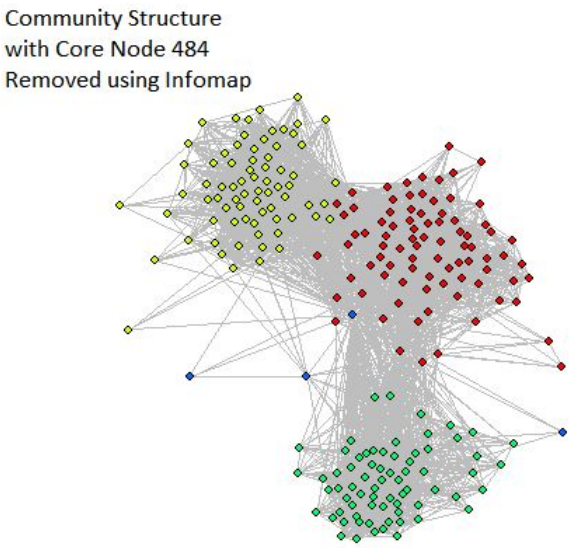


Method	**Modularity**	**Number of Communities**
Fast Greedy	0.25171	5
Edge Betweenness	0.13352	194
Infomap	0.09546	9

Community Structures with Node 484 Removed Shown Below



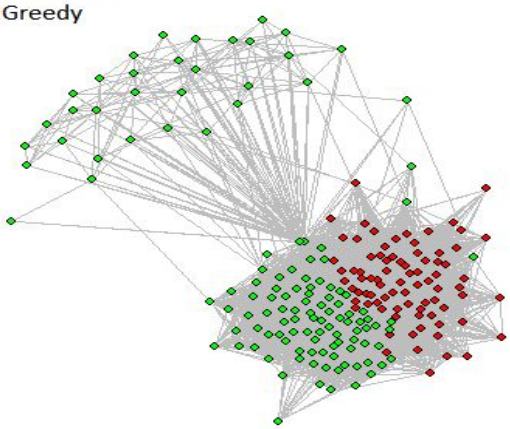




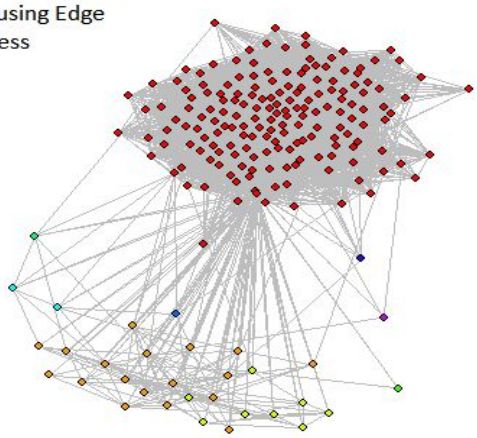
Method	**Modularity**	**Number of Communities**
Fast Greedy	0.50700	3
Edge Betweenness	0.48909	10
Infomap	0.51528	4

Community Structures with Node 1087 Removed Shown Below

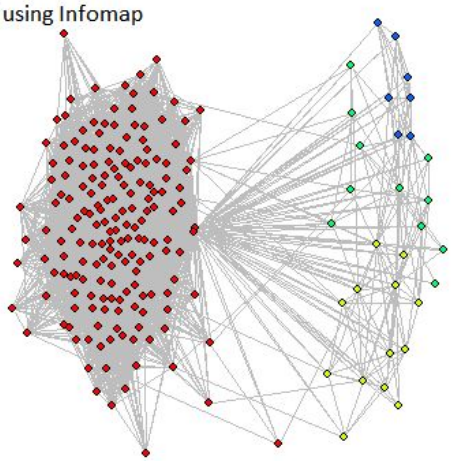
Community Structure with Core Node 1087 using Fast Greedy



Community Structure with Core Node 1087 Removed using Edge Betweenness



Community Structure with Core Node 1087 Removed using Infomap



Method	**Modularity**	**Number of Communities**
Fast Greedy	0.14553	2
Edge Betweenness	0.02762	9
Infomap	0.02691	4

Analysis of Results

1. Removing the core node decreases the number of communities for the fast greedy and edge betweenness algorithms but the the number of communities stays the same for the infomap algorithm when the core node is removed
2. Despite the communities decreasing, the modularity is broadly similar showing that the individual communities still have the same distribution of nodes.

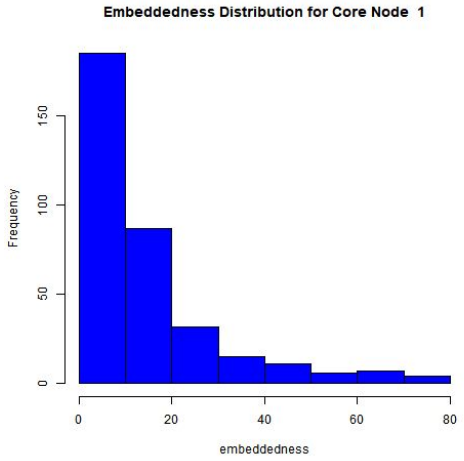
1.3.3 Characteristic of nodes in the personalized network

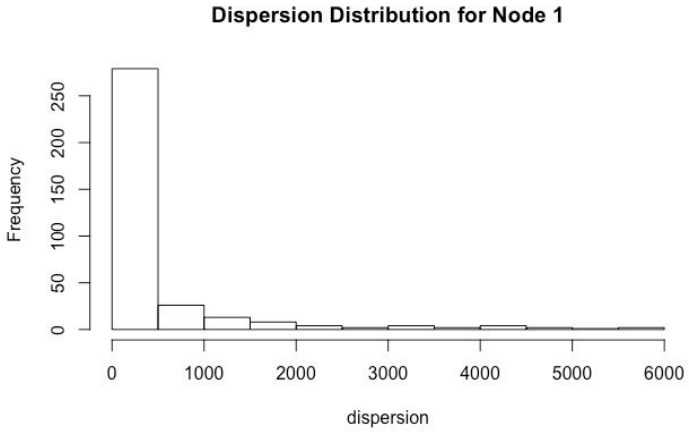
Question 11:

Embeddedness(Node, Core Node) = degree(Node) - 1

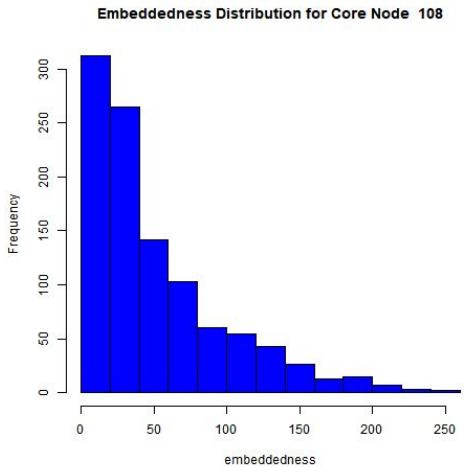
Question 12:

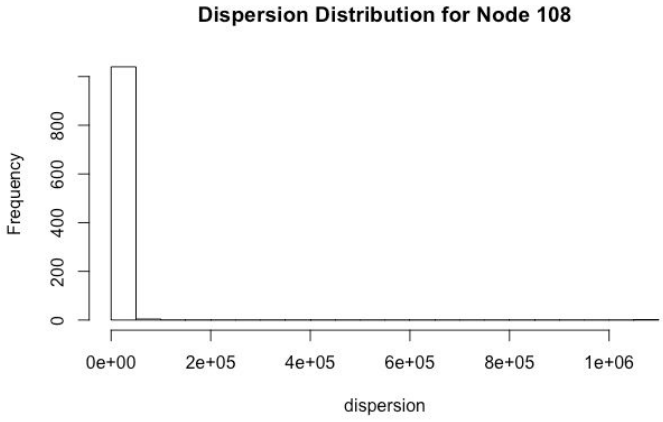
Node 1's personalized network (shown below) has the lowest degree of embeddedness and dispersion out of the other nodes observed in this study. This group of people is hardly cohesive with relatively low numbers of mutual friends and a long dispersion tail, indicating the sums of distances between pairs of mutual friends are large.



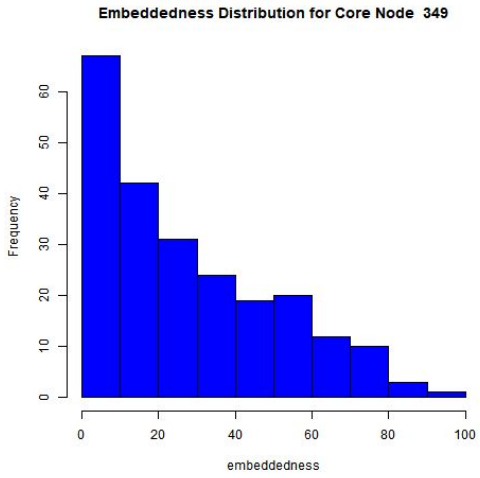


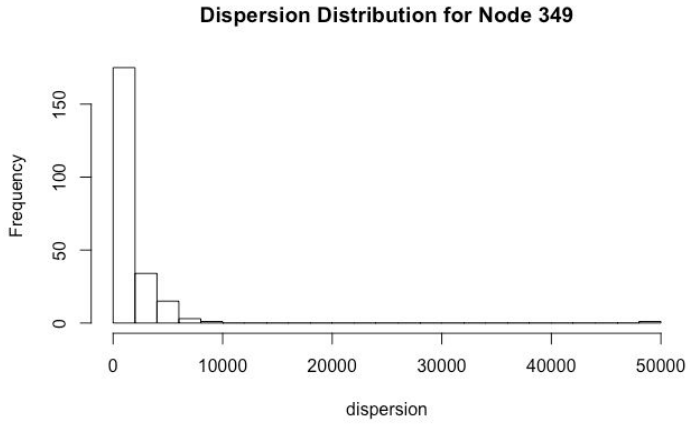
Node 108's personalized network (shown below) has the fourth highest degree of embeddedness and the largest dispersion. This node is highly integrated with various friend groups in the network.



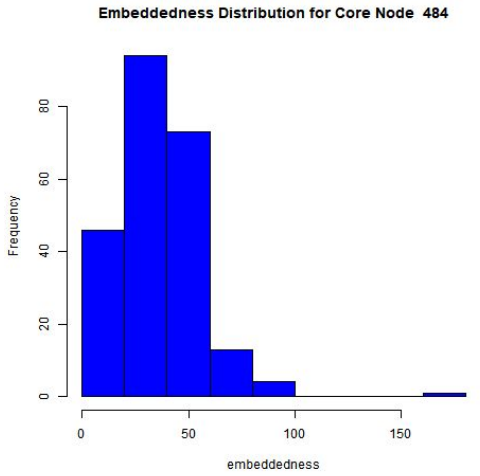


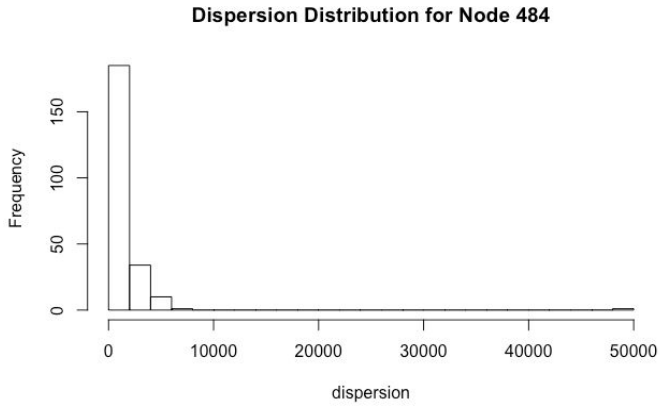
Node 349's personalized network (shown below) has the 2nd highest degree of embeddedness and dispersion out of the other nodes observed in this study. The dispersion distribution is similar to node 484's distribution. However, the embeddedness distributions differ significantly. Its average embeddedness score is lower than that of node 484.



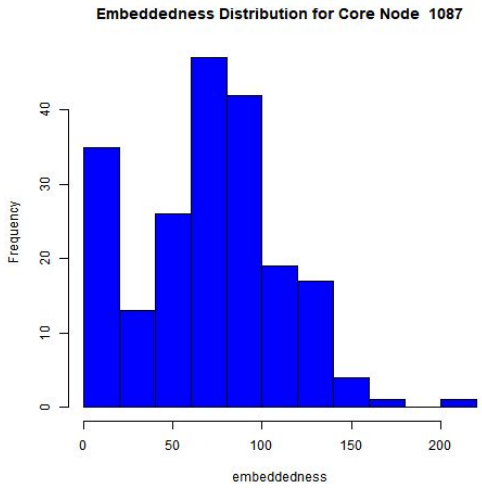


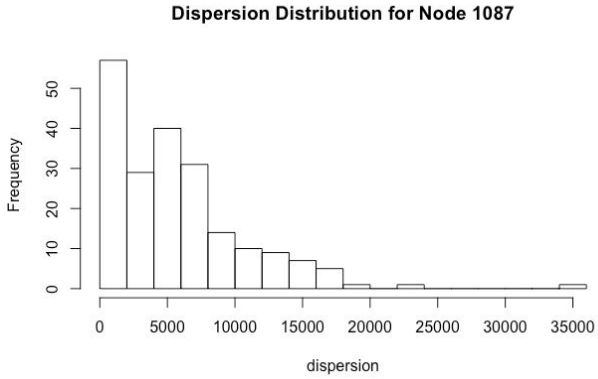
Node 484's personalized network (shown below) has the third highest degree of embeddedness and a dispersion distribution similar to that of node 349. Although it shares a dispersion distribution similar to 349, its average embeddedness score is higher, with a high concentration around a score of 20. However, as stated in "Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook," the embeddedness score did was not a great indicator for romantic relationships. So in general, this personal network has characteristics highly similar to node 349's personalized network.





Node 1087's personalized network (shown below) has the highest degree of embeddedness and the 2nd highest dispersion out of the other nodes observed in this study. Its high level of embeddedness indicates that the people in its network share a large number of mutual friends and that the group is more cohesive.



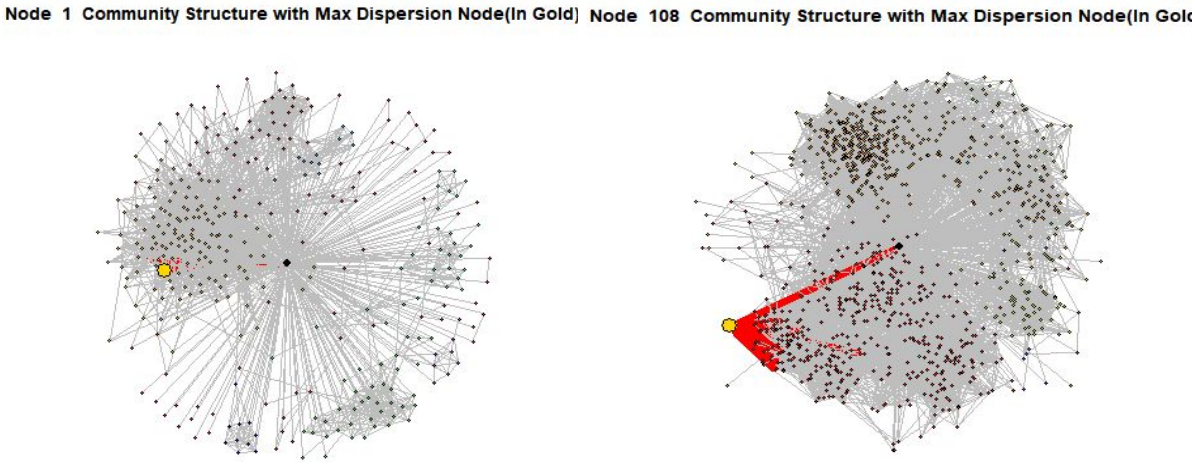


Question 13

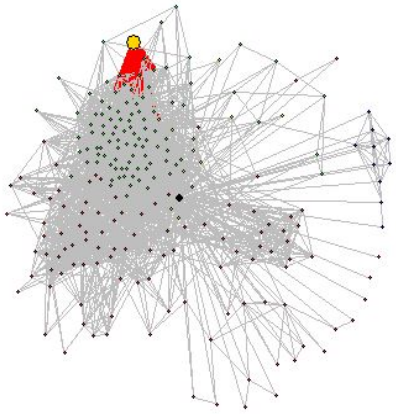
For this question, our goal was to calculate t

For this question, we calculated the node with the max dispersion with respect to each of the specified core nodes. The max dispersion node is highlighted in gold. We used the shortest path measure to measure dispersion and if two nodes were not connected, meaning the shortest path was Inf, we would set the dispersion to be 0.

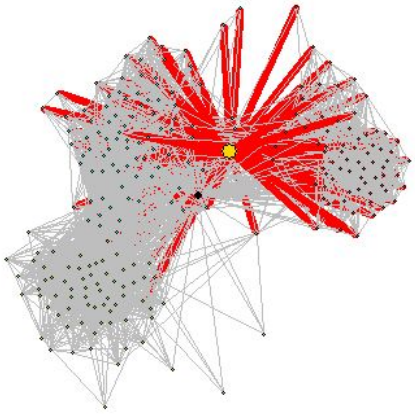
Community Structures of core nodes with max dispersion node are highlighted in gold.



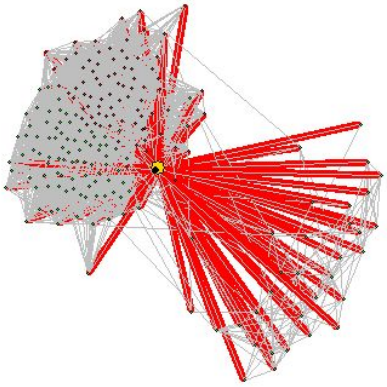
Node 349 Community Structure with Max Dispersion Node(In Gol



Node 484 Community Structure with Max Dispersion Node(In Gol



Node 1087 Community Structure with Max Dispersion Node(In Gol

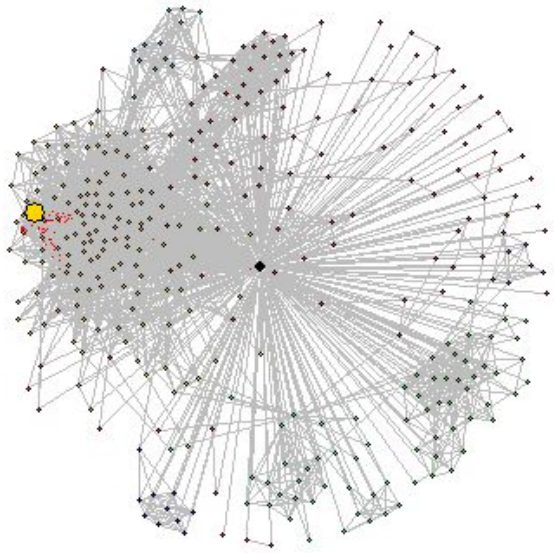


Question 14

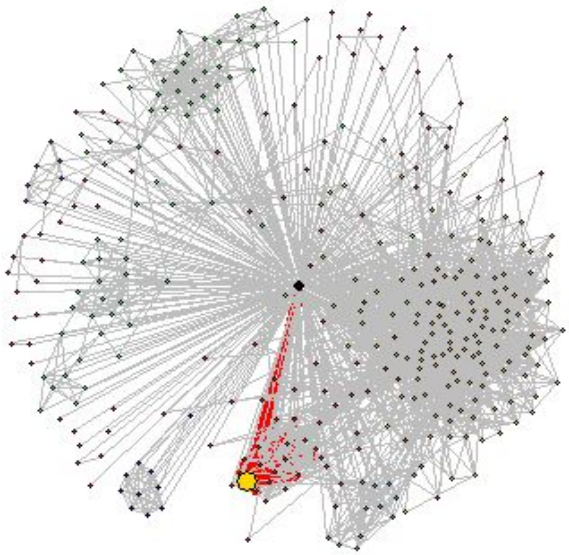
For this question, similar to the above, the nodes with the maximum embeddedness and max ratio of dispersion to embeddedness are highlighted in gold with the core node shown in black. The embeddedness between two nodes is defined as the number of mutual friends shared between two nodes. It basically is a measure of social tie. Below, we plot the community structures for the core nodes and highlight the nodes with maximum embeddedness and maximum $\frac{embeddedness}{dispersion}$

Community structures with max embeddedness and max dispersion/embeddedness highlighted in gold

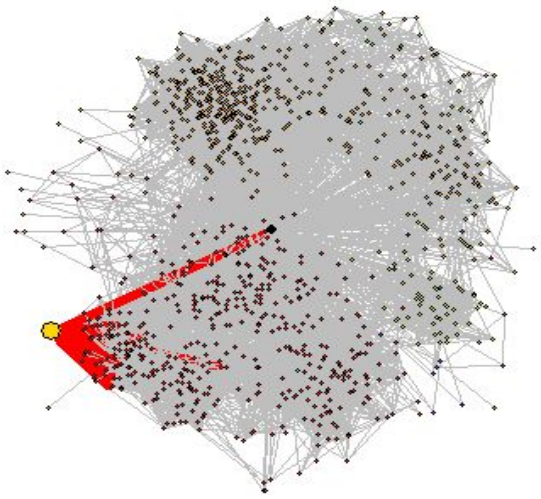
Node 1 with Max Embeddedness Node



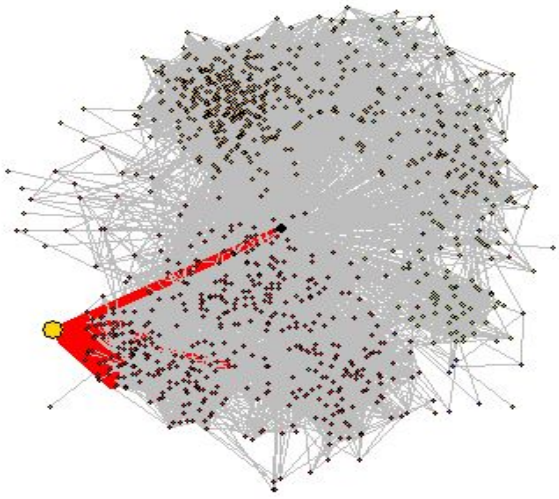
Node 1 Community Structure with Max Ratio Node



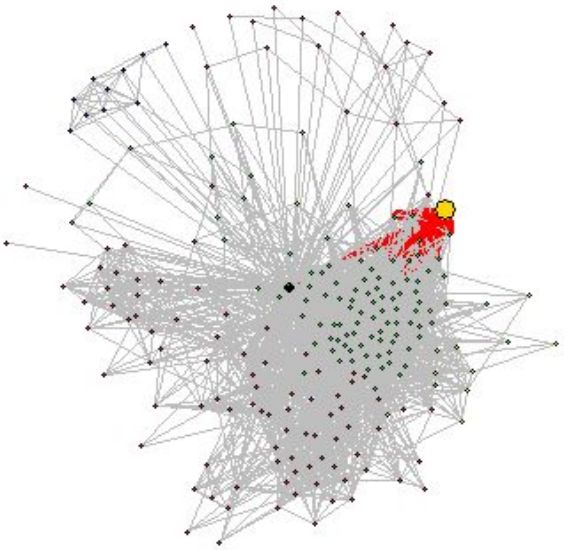
Node 108 with Max Embeddedness Node



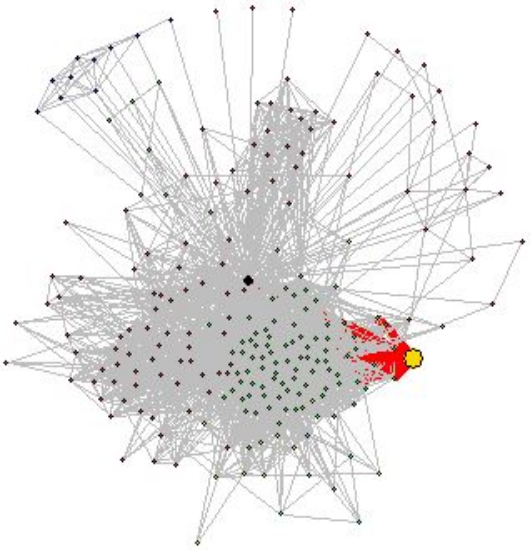
Node 108 Community Structure with Max Ratio Node



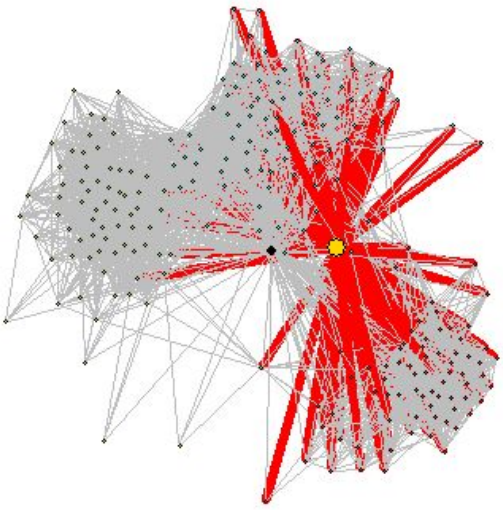
Node 349 with Max Embeddedness Node



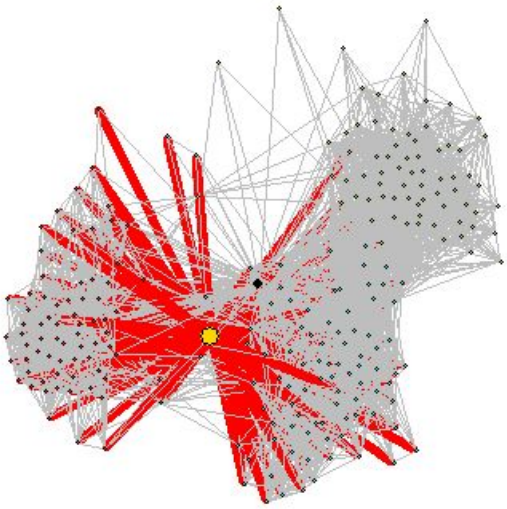
Node 349 Community Structure with Max Ratio Node



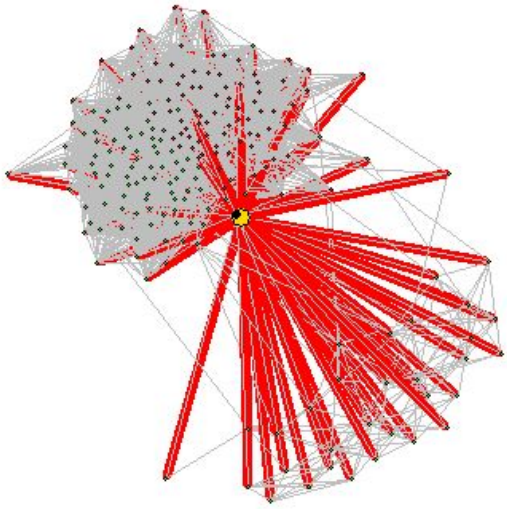
Node 484 with Max Embeddedness Node



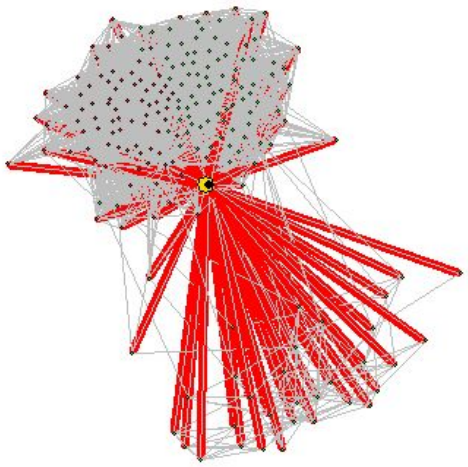
Node 484 Community Structure with Max Ratio Node



Node 1087 with Max Embeddedness Node



Node 1087 Community Structure with Max Ratio Node



Question 15

Dispersion: In the plots in question 13, the gold node has the maximum dispersion with respect to the core node. This means that that node is not well acquainted with the core node. In other words, mutual friends of the core node and the maximum dispersion node are not well connected with each other. Dispersion decreases as the size of the core nodes ego network increases.

Embeddedness: As can be seen in the plots, the maximum embeddedness, highlighted in gold, belongs to the core nodes largest community. This means that for the maximal community in terms of size, within that community there will exist a node that shares the most mutual friends with the core node, measuring the number of direct neighbors of the core node that belong to the largest community.

$Dispersion / Embeddedness$: Having a high dispersion to embeddedness ratio is used to predict the romantic partner of a specific person. Dispersion can be defined as the number of mutual friends (s,t) of two nodes u and v so that (s,t) have no common neighbors besides u and v. This means the two communities that the core node and potential romantic partner belong to are linked only via the core node and the maximum ratio node.

1.4 Friend recommendation in personalized networks

In this section, three algorithms for friend recommendation were explored. First, an ego graph was generated for node 415. Then, all of node 415's friends with 24 degrees were selected. For each node, its friends were randomly deleted with probability 0.25 and then replaced with new friends within the ego network based on the chosen similarity measure. The measures explored include Adamic Adar, Jaccard, and Common Neighbors.

1.4.1 Neighborhood based measure

(No questions in this part)

1.4.2 Friend recommendation using neighborhood based measures

(No questions in this part)

1.4.3 Creating the list of users

Question 16

When the number of nodes with degree

The number of nodes with degree 24 is 11. We iterated through all the vertices and counted the number that had degree 24.

1.4.4 Average accuracy of friend recommendation algorithm

Question 17

Method	**Accuracies**
Adamic-Adar	0.84127
Common Neighbors	0.84245
Jaccard	0.81664

Based on our simulation, the highest accuracy was with Common Neighbors with 0.8425 percent accuracy. We believe the reason Common Neighbors performs so well is because of the size of the network is sufficiently small that only looking at mutual friends would be a good heuristic. Also, in reading the paper “A Survey of Link Recommendation for Social Networks: Methods, Theoretical Foundations, and Future Research Directions”, the most common algorithm used by many social networks to recommend friends is Common Neighbors. The Adamic-Adar assigns less weight to more connected common neighbors. The reason the Jaccard measure is the entropy of the network is not high enough.

2. Google+ network

In this section, the Google+ network is explored and analyzed to gain an understanding of its community structure. Additionally, network homogeneity and completeness measures (two commonly used measures for clustering performance assessment) were implemented in the context of network graphs.

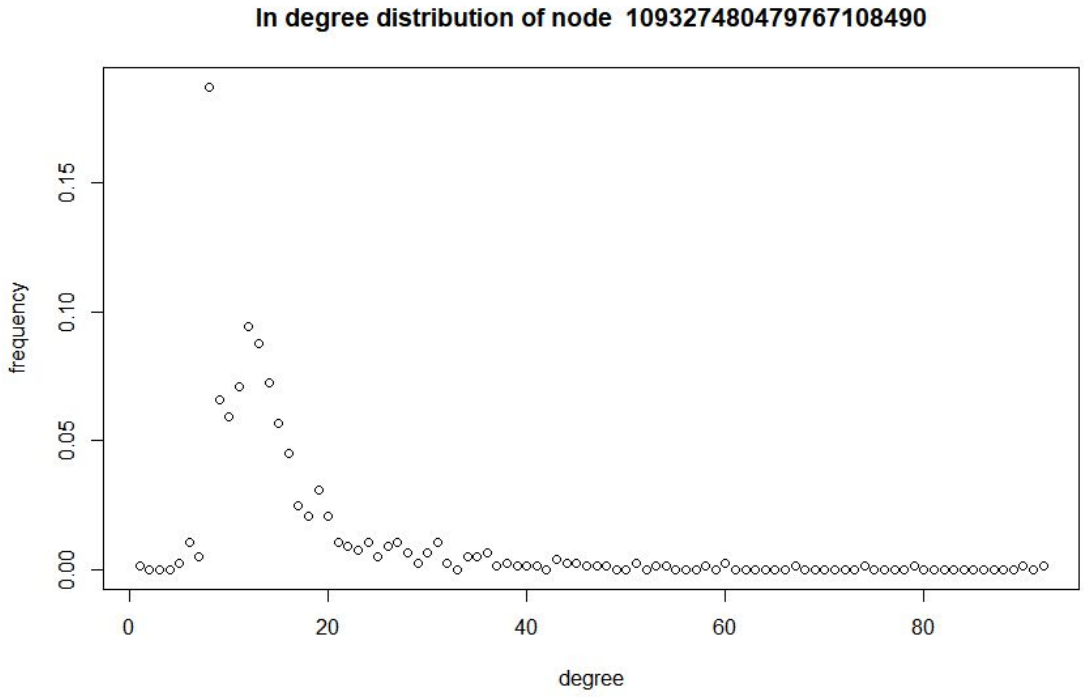
Question 18

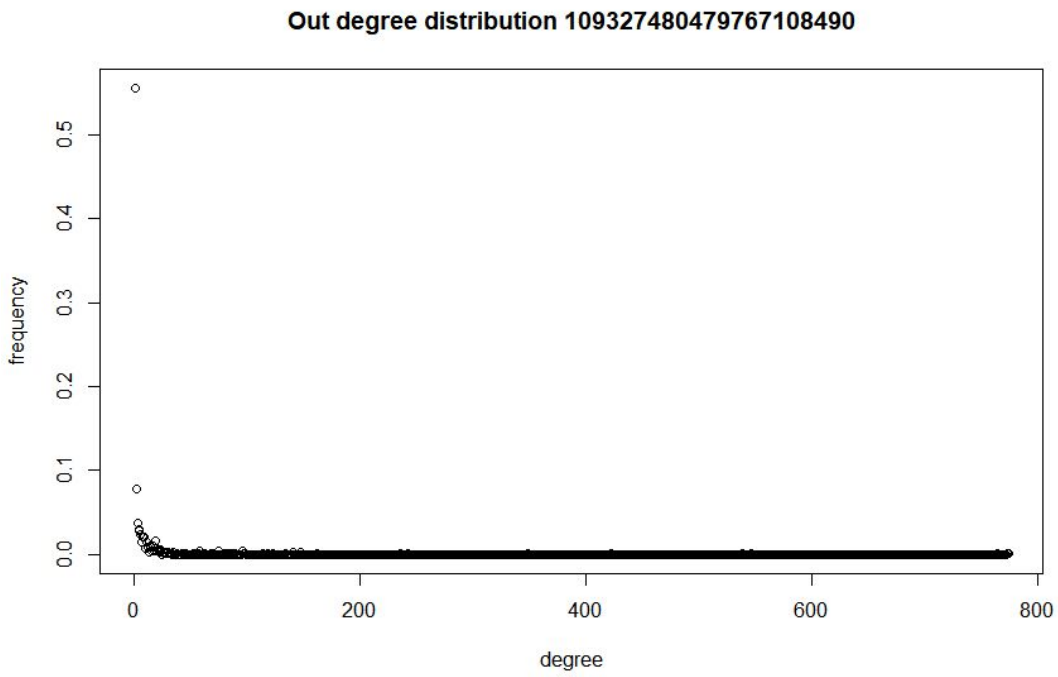
The Google plus dataset has 132 unique nodes/ Ego nodes, and the number of ego nodes with more than 2 circles is 57.

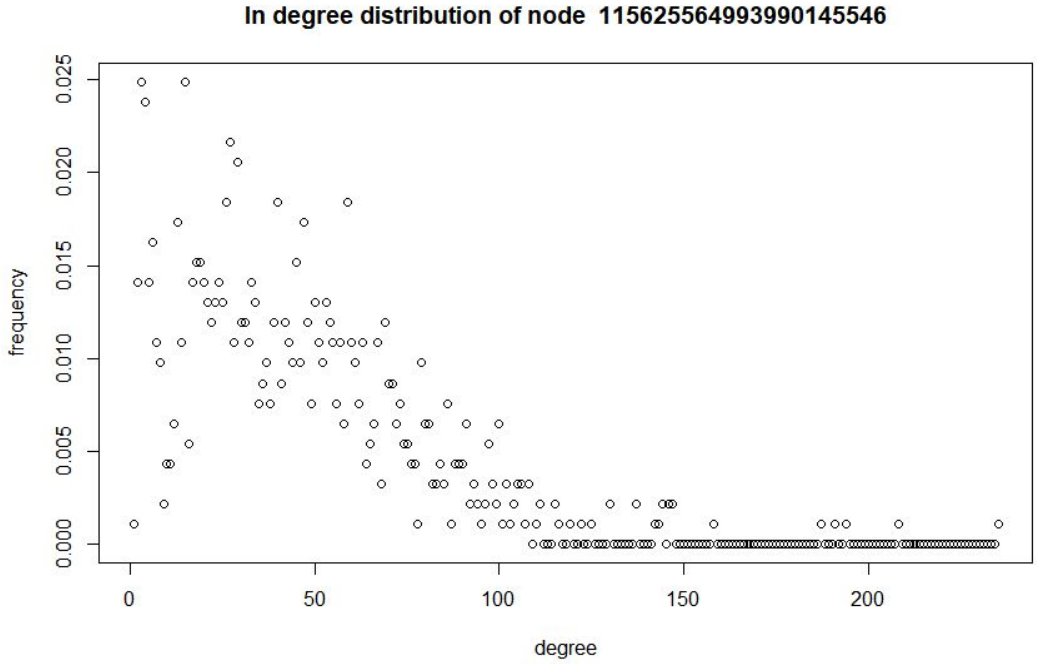
Every node has a circle file where each line corresponds to a circle with nodes in that circle separated by tab. We read the number of lines in the circles file and if the number is greater than 2 , we take those ego nodes into consideration.

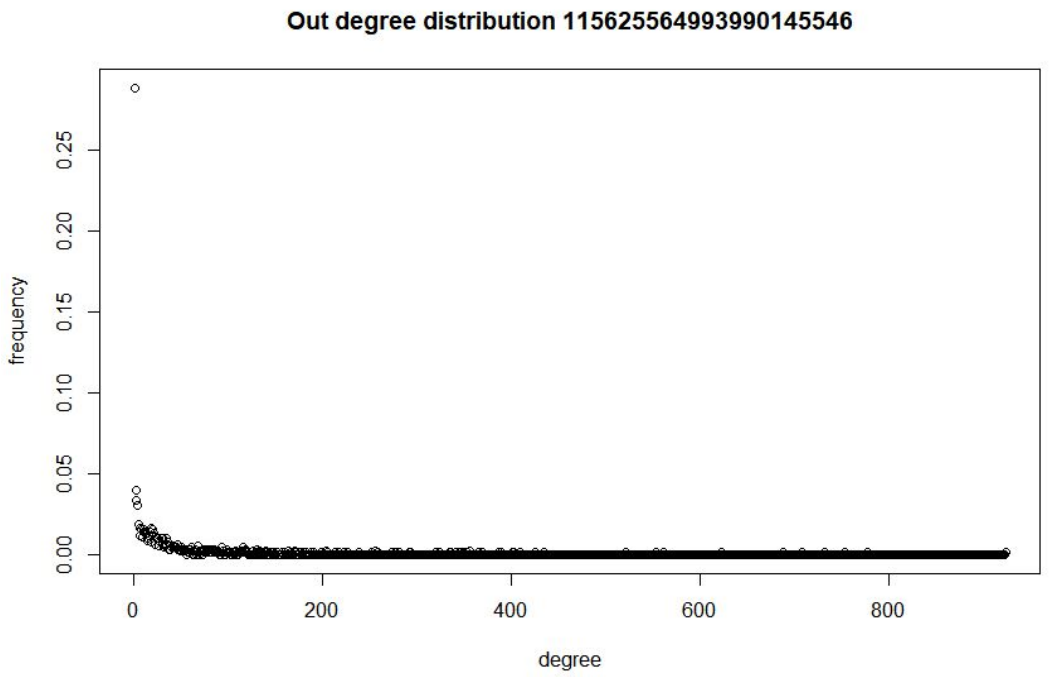
Question 19

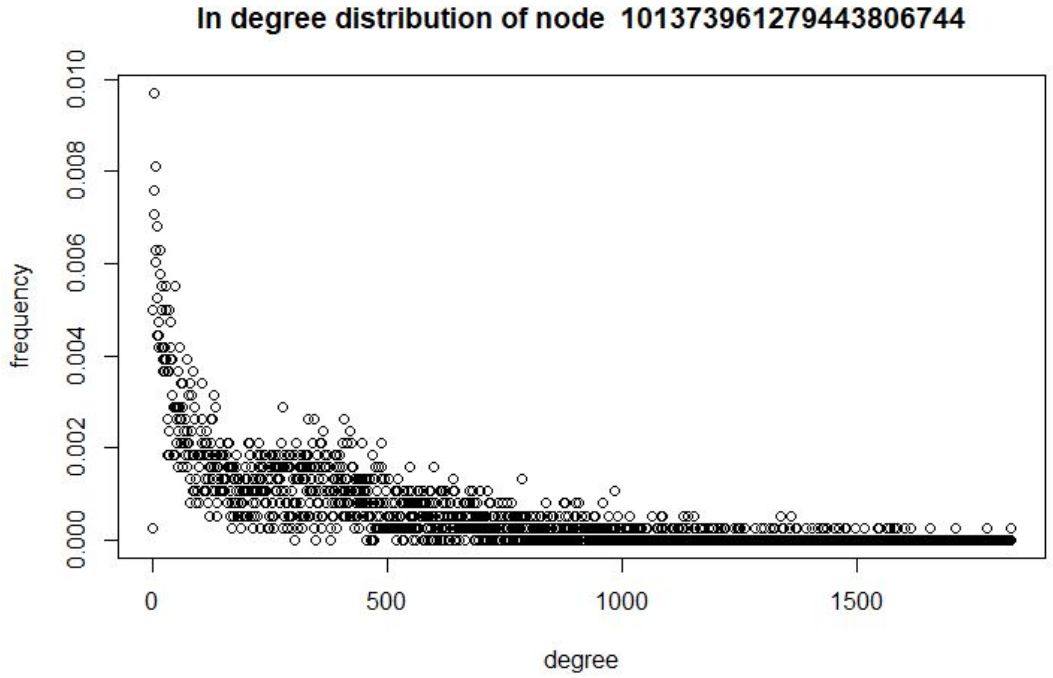
The Google+ Network in-degree and out-degree distributions are shown below. A summary of the distributions is included below the final plot.

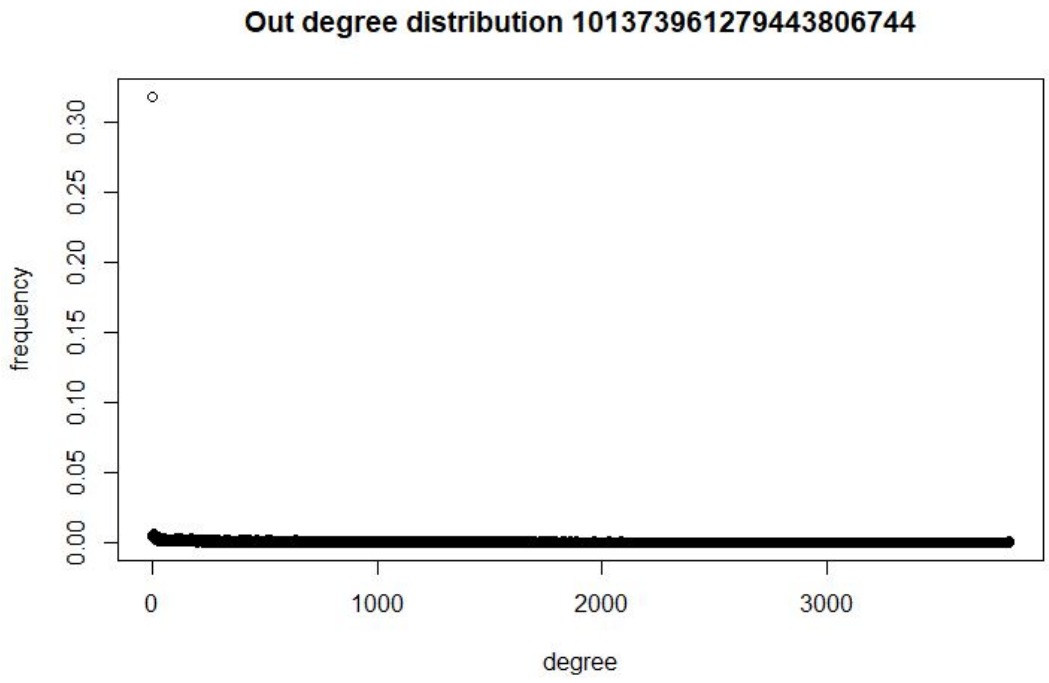



Out degree distribution 109327480479767108490
frequency
0.0
0.1
0.2
0.3
0.4
0.5
0
200
400
600
800
degree


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 11562556499399014554 6
frequency
0.00
0.05
0.10
0.15
0.20
0.25
0
200
400
600
800
degree


In degree distribution of node 101373961279443806744
frequency
0.000
0.002
0.004
0.006
0.008
0.010
0
500
1000
1500
degree



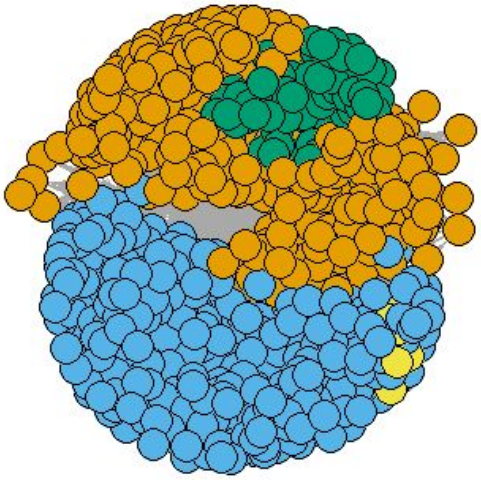
Based on the above graphs, it seems both the in degree and out degree follow an exponential distribution. However, the indegree seems to have more variation. We believe this is because in a directed network, your number of followers is more random than the number of people you follow.

2.1 Community structure of personal networks

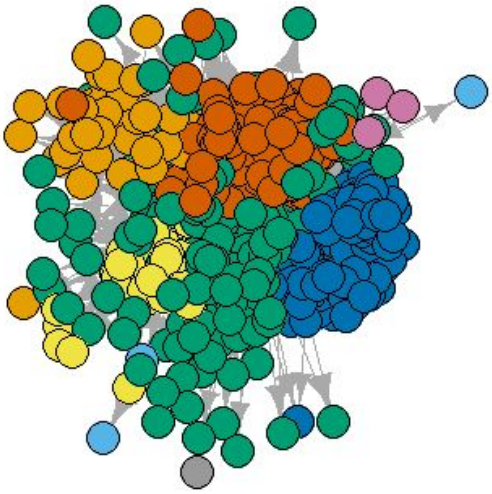
Question 20

Node	**Modularity**
109327480479767108490	0.2527654
115625564993990145546	0.3194726
101373961279443806744	0.1910903

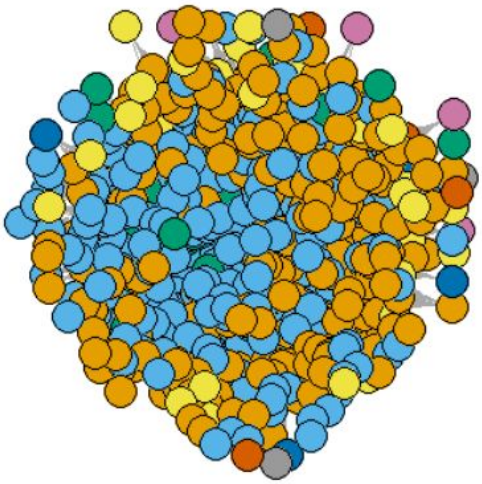
Walk trap community for node
10932748047976710849O



Walk trap community for node
11562556499399014554 6



Walk trap community for node
10137396127944380674 4



Since the edges of the network are so well connected, finding disparate communities is a difficult task, leading to the low modularity values above.

Question 21

Homogeneity measures the ratio of a node belonging to a community. The higher the number is, the fewer communities are in the graph that is every cluster starts having elements from a single class with higher homogeneity. In the least homogeneous clusters, H(C|K) is maximum i.e H(C|K) = H(C).In this case h reaches its minimum value, which is 0. As homogeneity increases H(C|K) values decrease. For the most homogeneous clustering, H(C|K) = 0 and h = 1

Completeness measure the size of a single community. The higher the number is, the larger the communities are. The least complete clustering ie c=0 is when H(K|C) = H(K) and the most complete cluster has H(K|C) = 0 thereby c=1.

Question 22

Node	**Homogeneity**	**Completeness**
109327480479767108490	0.8640041	0.3456923
115625564993990145546	0.4429445	-3.375718
101373961279443806744	0.001839249	-1.609263

Node #109327480479767108490 has the highest homogeneity and completeness among all three. This makes sense because as shown in the community graph for Question 20, this node has the lowest number of communities and the average community size is the highest as well.

Node #115625564993990145546 has the second-highest homogeneity while the lowest completeness. It has the second-highest number of communities, and it also has a few community with very small number of nodes.

Node #101373961279443806744 has the lowest homogeneity and the second-lowest completeness. As we can see there are lots of communities in the graph, but there are some communities with relatively large size bringing the completeness higher a little compared with node #115625564993990145546.