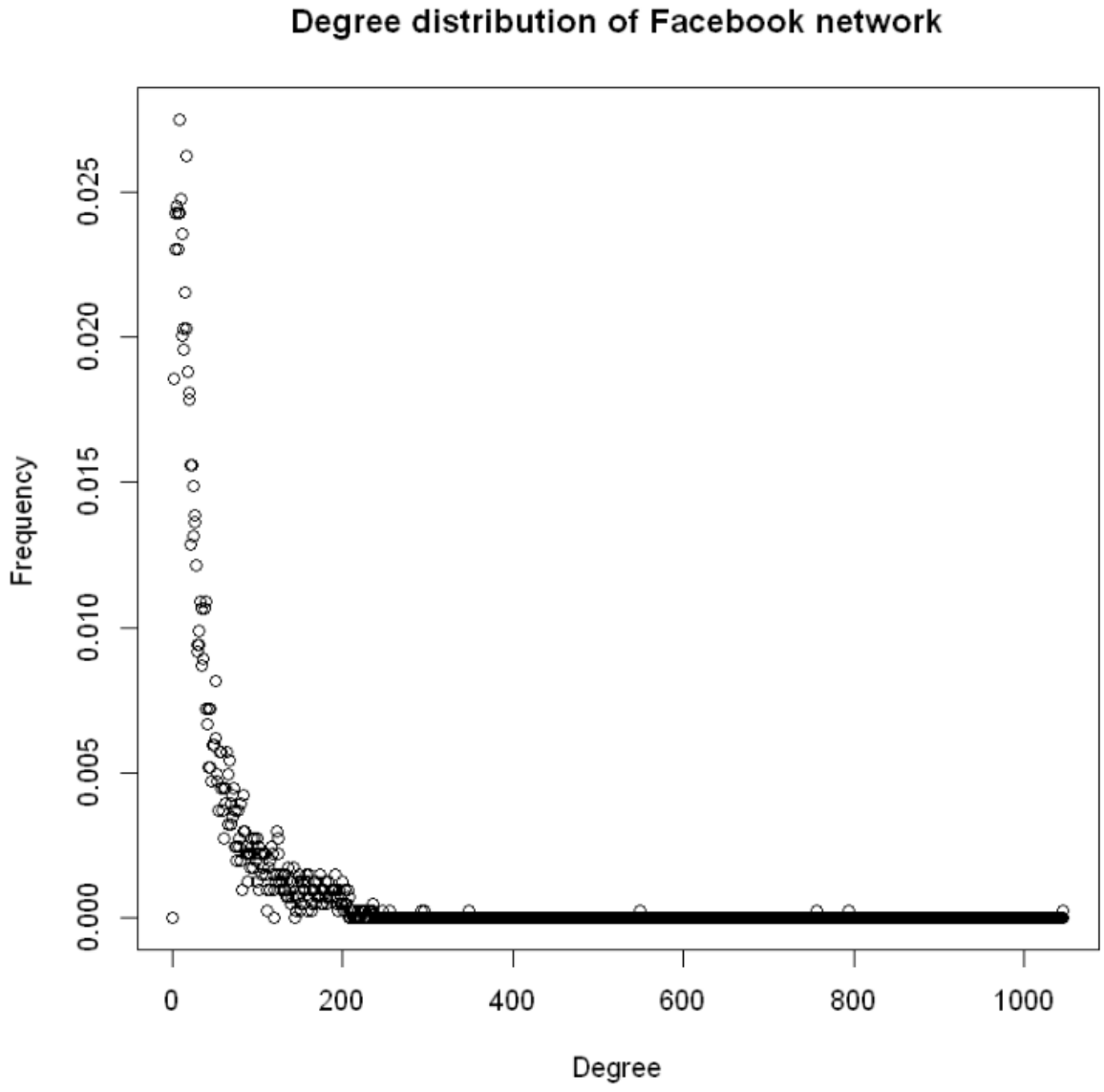
Project 2_Social Network Mining
Haoting Ni (905545789), Yikai Wang (905522085), Yuanxuan Fang (005949389)

Question 1.1:
From the output, we find that there are 4039 nodes and 88234 edges in this network.

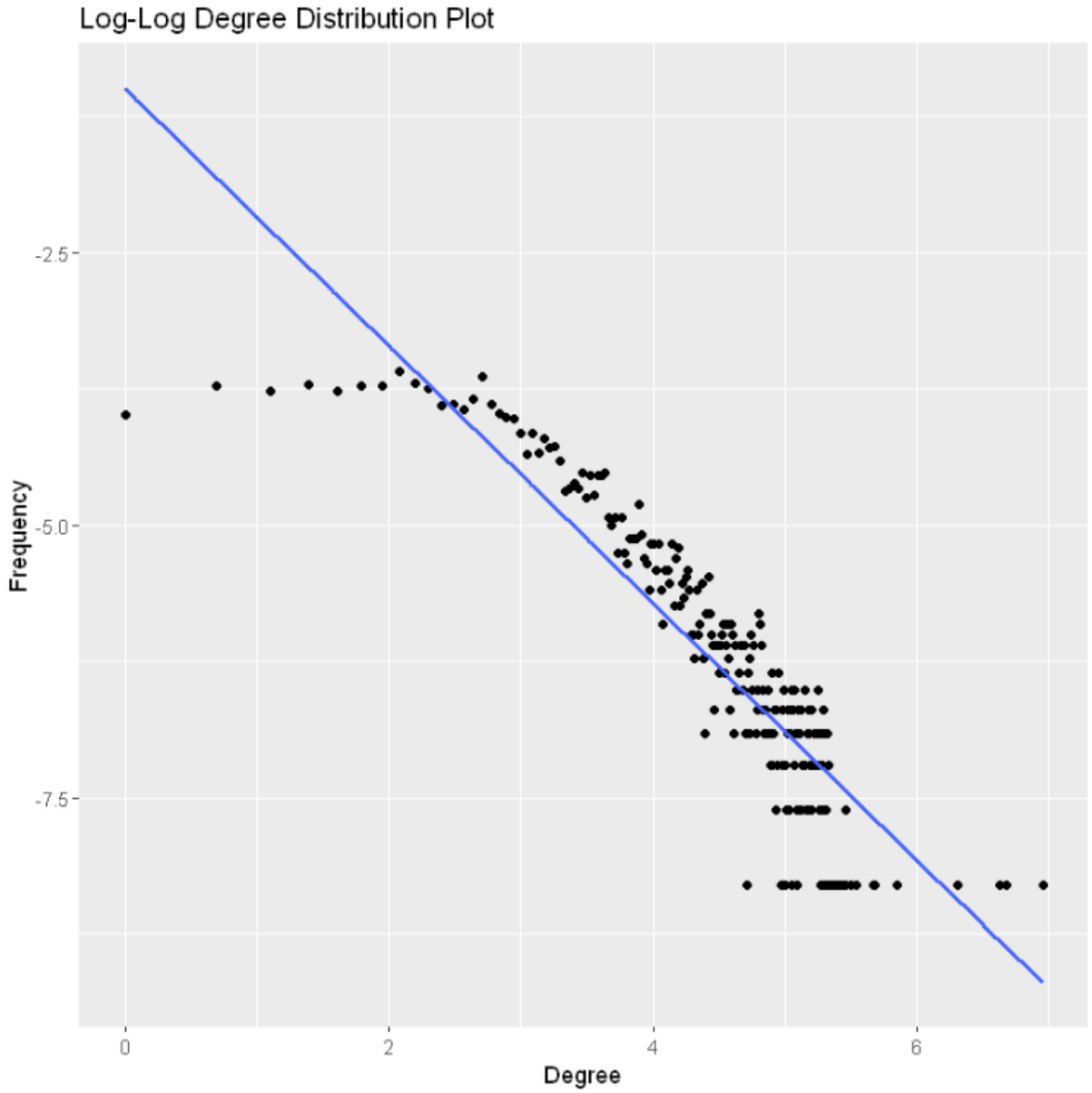
Question 1.2:
Yes, The Facebook network is connected

Question 2:
The Diameter of the network is 8

Question 3:
The average degree is 43.691013



Question 4:
Estimated slope of the line for the log-log scale degree distribution is -1.18



Question 5:
There are 348 nodes and 2866 edges in the personalized network if the user whose ID is 1.

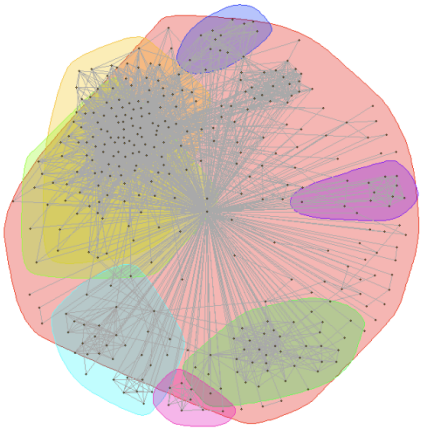
Question 6:
The diameter of the personalized network is 2, then the trivial upper bound is 2 and trivial lower bound is 1 for the diameter of the personalized network because the diameter must be a positive integer.

Question 7:
When the diameter of the personalized network is equal to 2, each pair of nodes can be connected through 2 edges, which means the network will have a high degree of connectivity. If the diameter of the personalized networks is equal to 1, each pair of nodes can have one edge, which will indicate a fully connected network.

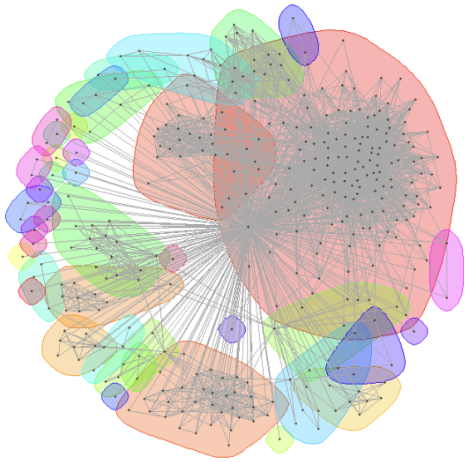
Question 8:
There are 40 core nodes in the Facebook Network.
The Average degree of the core nodes is 279.375.

Question 9:
Node ID 1:
Fast-Greedy: 0.413101
Edge-Betweenness: 0.353302
Infomap community: 0.389118

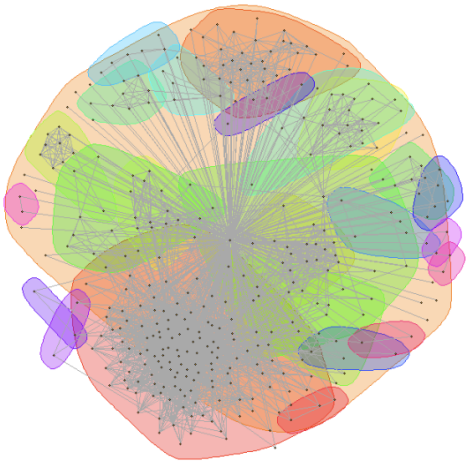
Community Structure of Fast-Greedy for Node ID=1



Community Structure of Edge-Betweenness for Node ID=1

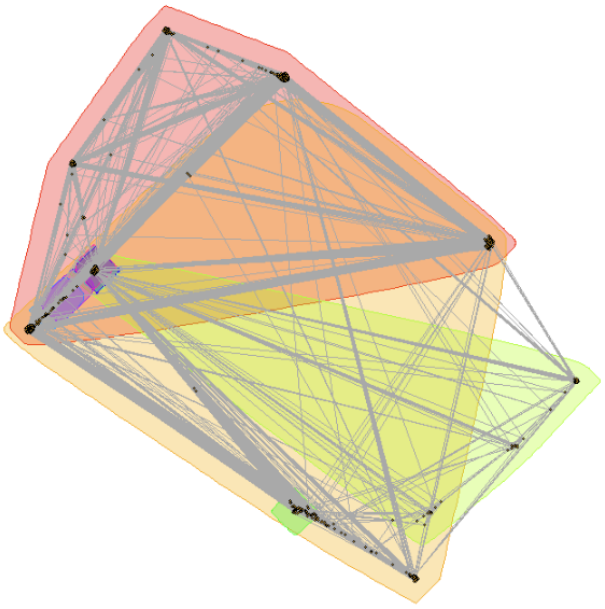


Community Structure of Infomap for Node ID=1

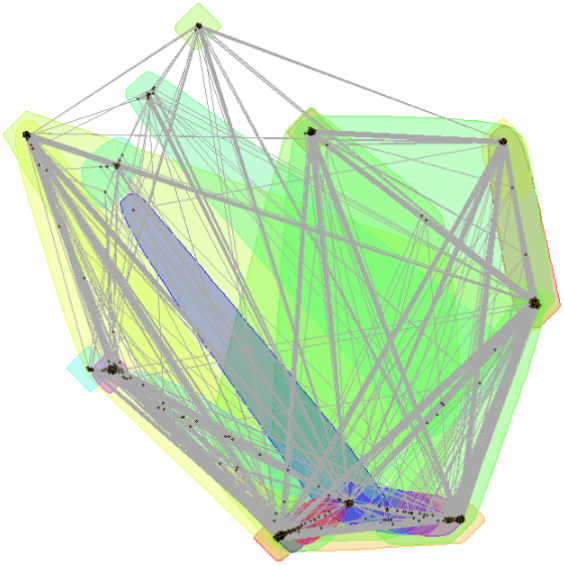


Node ID 108:
Fast-Greedy: 0.435929
Edge-Betweenness: 0.506755
Infomap community: 0.508223

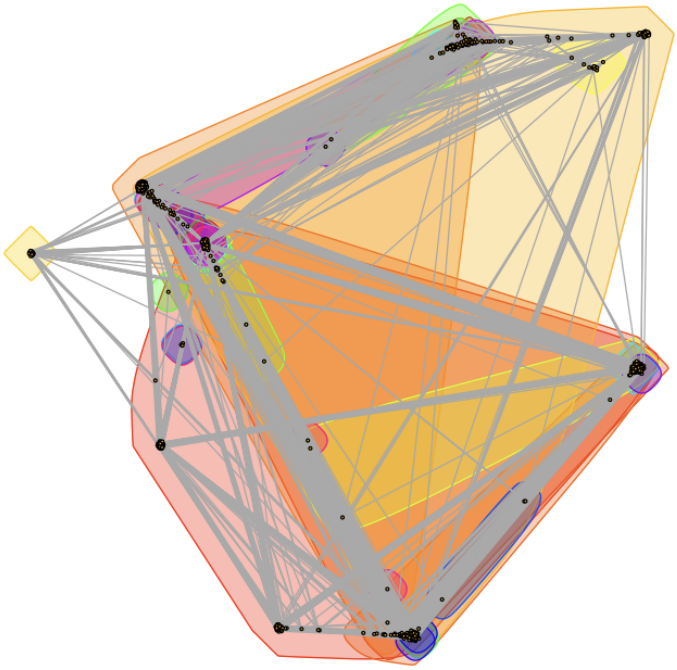
Community Structure of Fast-Greedy for Node ID=108



Community Structure of Infomap for Node ID=108

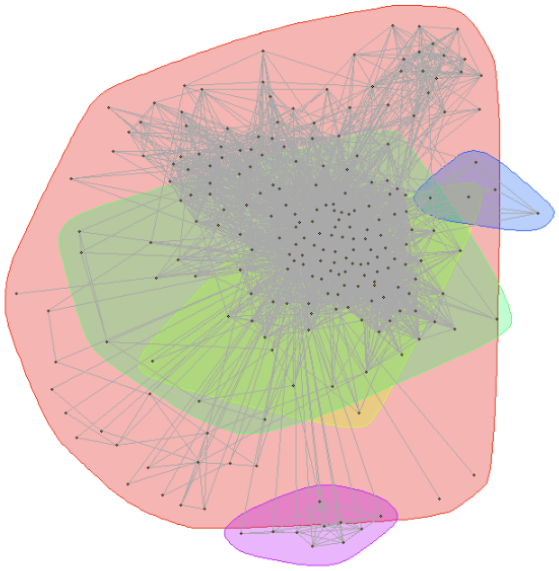


Community Structure of Edge-Betweenness for Node ID=108

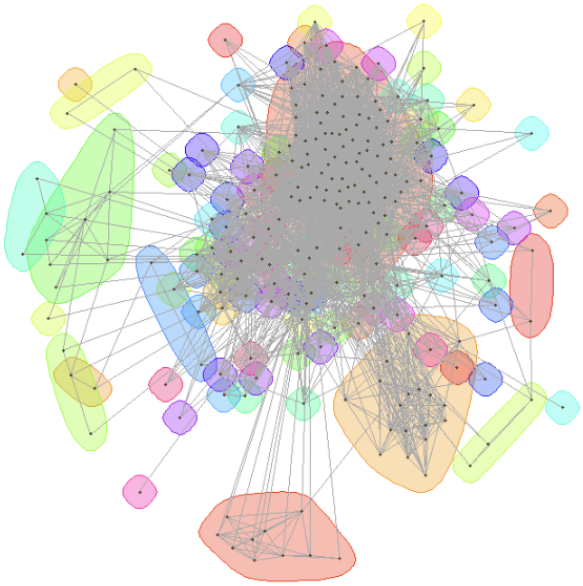


Node ID 349:
Fast-Greedy: 0.251715
Edge-Betweenness: 0.133528
Infomap community: 0.096029

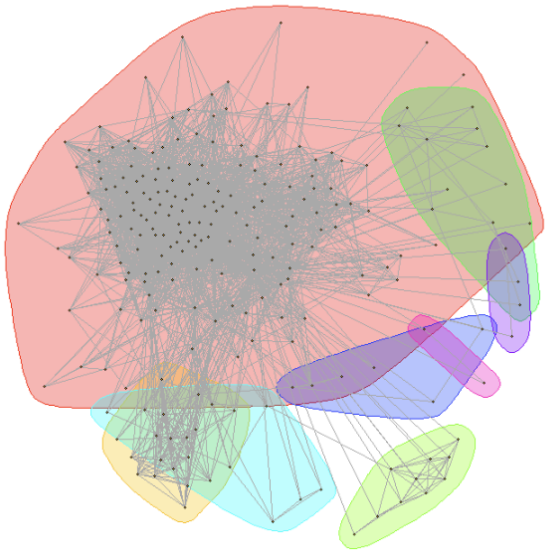
Community Structure of Fast-Greedy for Node ID=349



Community Structure of Edge-Betweenness for Node ID=349

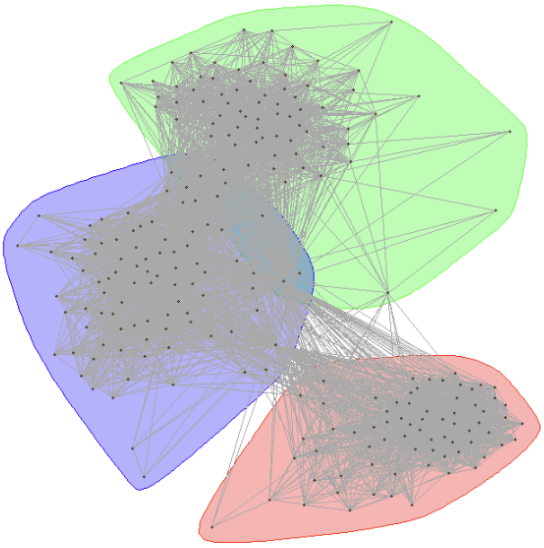


Community Structure of Infomap for Node ID=349

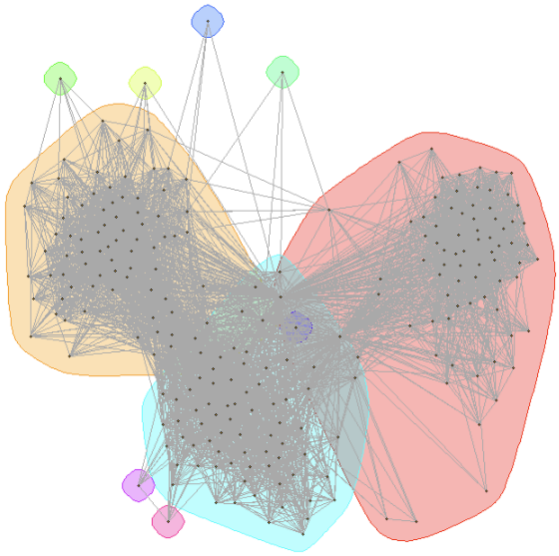


Node ID 484:
Fast-Greedy: 0.507002
Edge-Betweenness: 0.489095
Infomap community: 0.515279

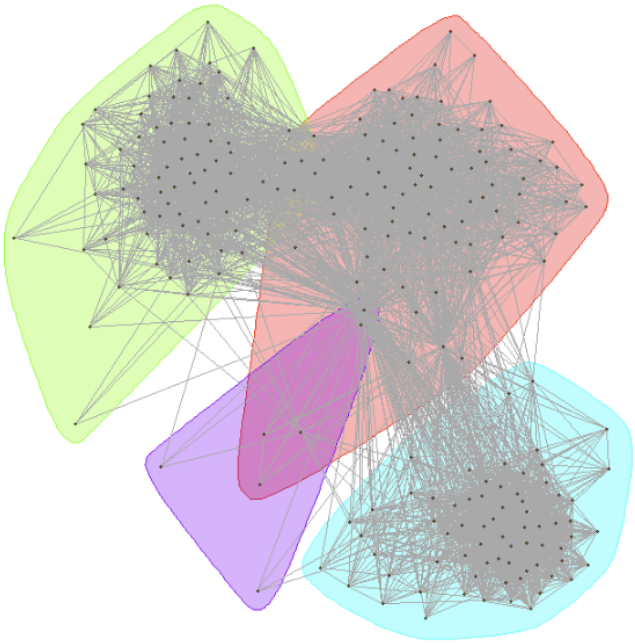
Community Structure of Fast-Greedy for Node ID=484



Community Structure of Edge-Betweenness for Node ID=484

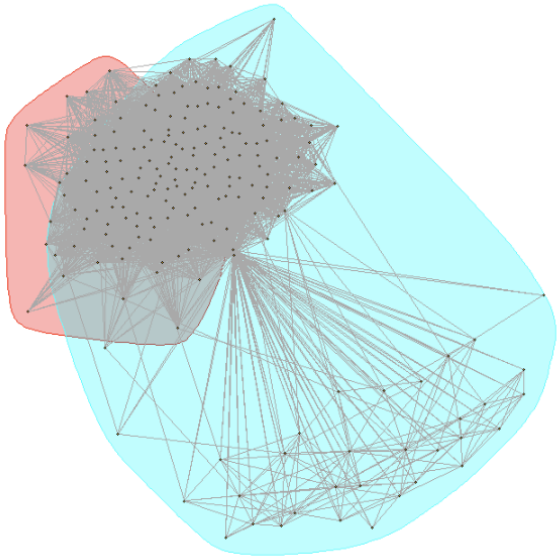


Community Structure of Infomap for Node ID=484

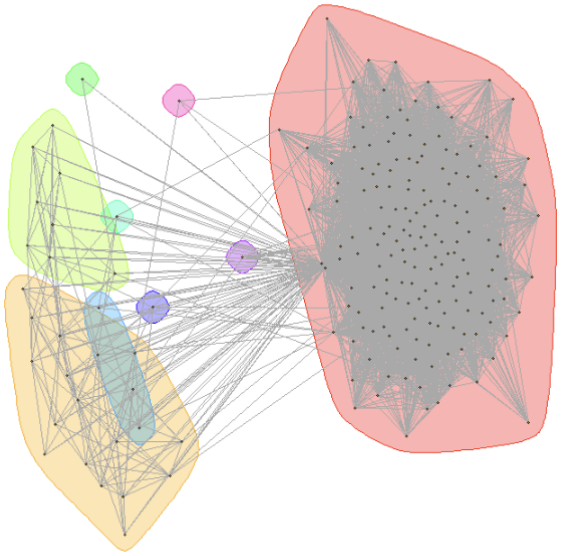


Node ID 1087:
Fast-Greedy: 0.145531
Edge-Betweenness: 0.027624
Infomap community: 0.026907

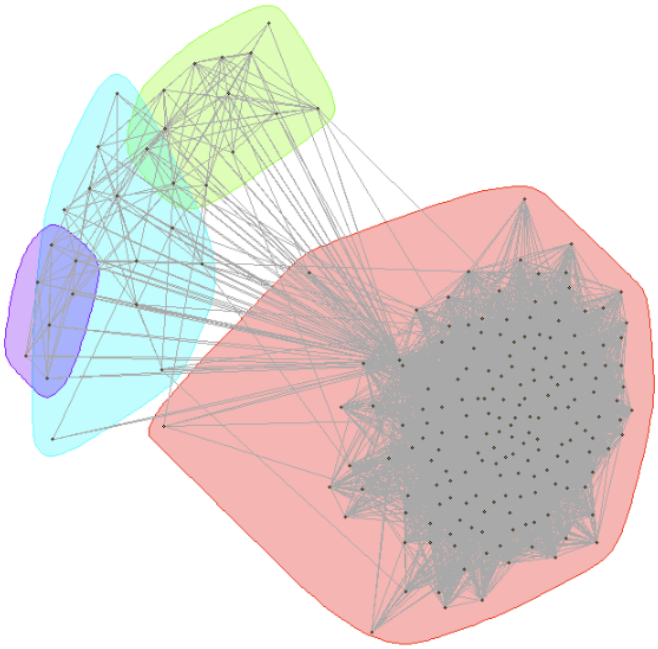
Community Structure of Fast-Greedy for Node ID=1087



Community Structure of Edge-Betweenness for Node ID=1087

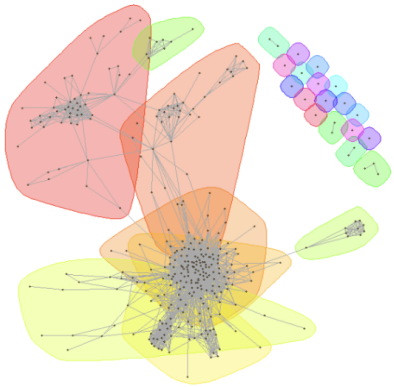


Community Structure of Infomap for Node ID=1087

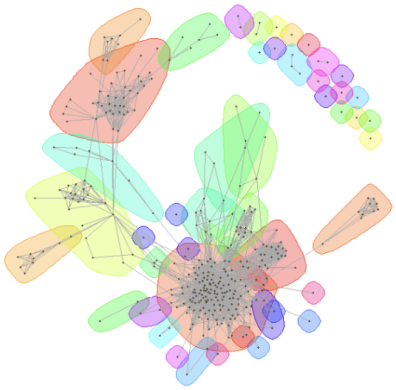


Question 10:
Node ID 1:
Modularity for Fast-Greedy:0.441853
Modularity for Edge-Betweenness: 0.416146
Modularity for Infomap community: 0.418008

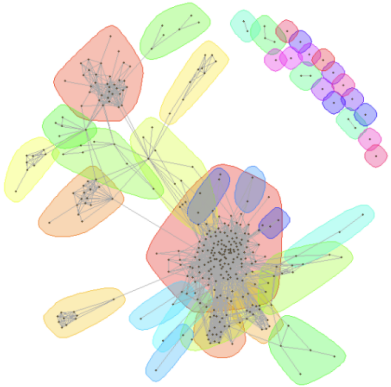
Community Structure of Fast-Greedy for Node ID=1



Community Structure of Edge-Betweenness for Node ID=1

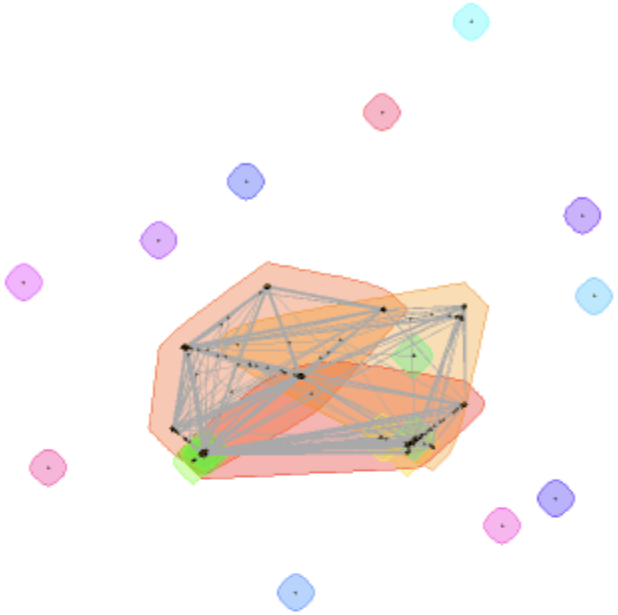


Community Structure of Infomap for Node ID=1

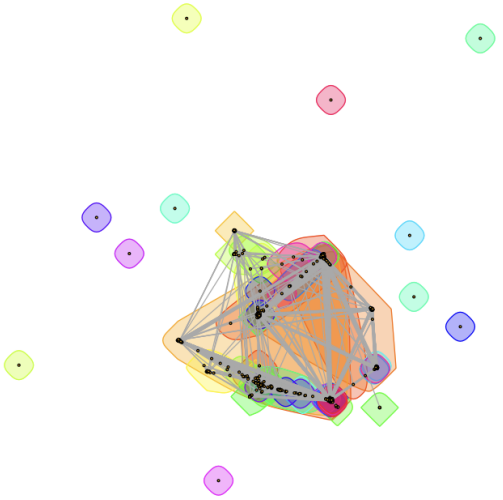


Node ID 108:
Modularity for Fast-Greedy:0.458127
Modularity for Edge-Betweenness:0.521322
Modularity for Infomap community: 0.521369

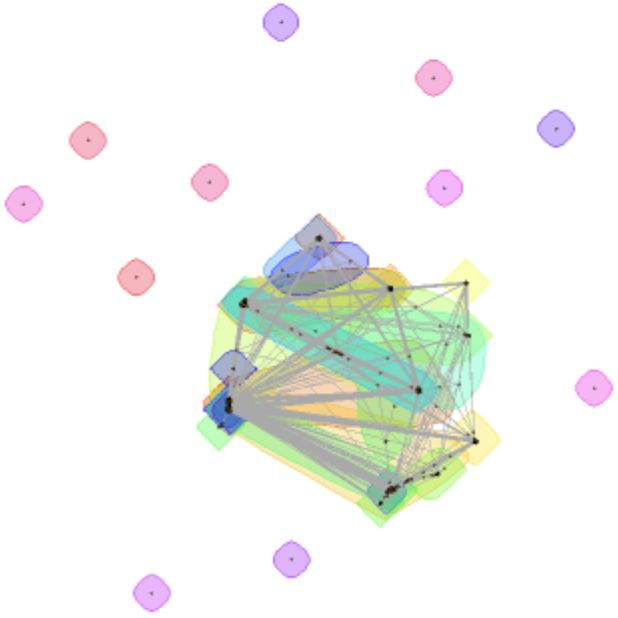
Community Structure of Fast-Greedy for Node ID=108



Community Structure of Edge-Betweenness for Node ID=108

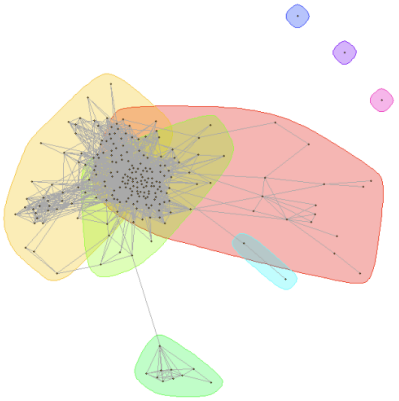


Community Structure of Infomap for Node ID=108

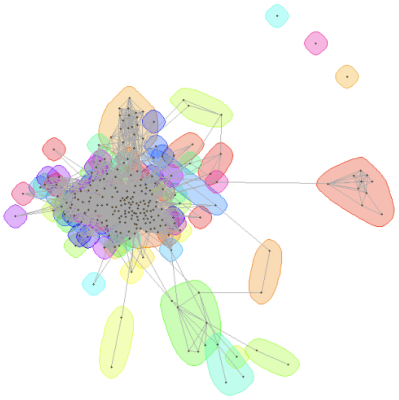


Node ID 349:
Modularity for Fast-Greedy:0.245692
Modularity for Edge-Betweenness: 0.150566
Modularity for Infomap community: 0.246578

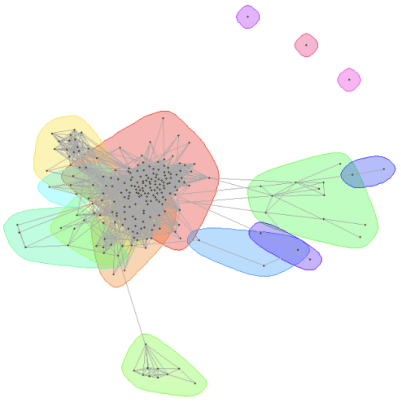
Community Structure of Fast-Greedy for Node ID=349



Community Structure of Edge-Betweenness for Node ID=349

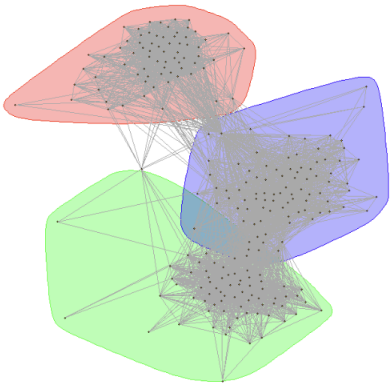


Community Structure of Infomap for Node ID=349

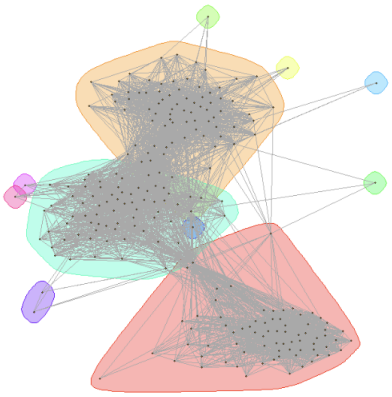


Node ID 484:
Modularity for Fast-Greedy:0.534214
Modularity for Edge-Betweenness: 0.515441
Modularity for Infomap community: 0.543444

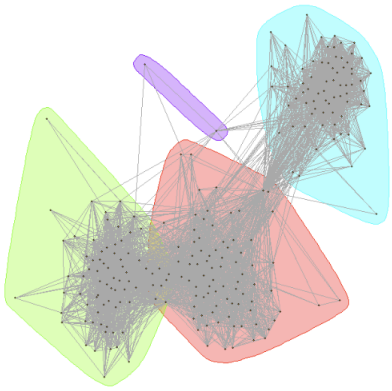
Community Structure of Fast-Greedy for Node ID=484



Community Structure of Edge-Betweenness for Node ID=484

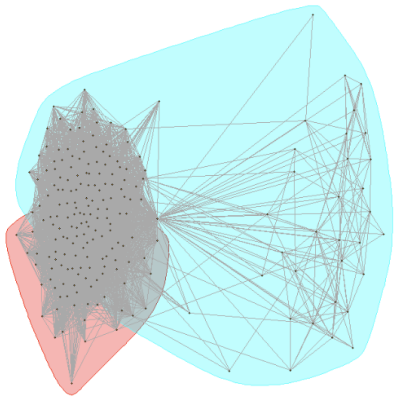


Community Structure of Infomap for Node ID=484

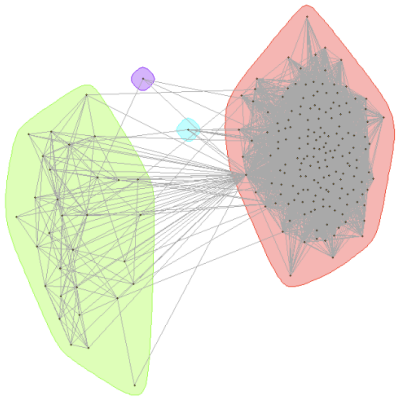


Node ID 1087:
Modularity for Fast-Greedy:0.148196
Modularity for Edge-Betweenness: 0.032495
Modularity for Infomap community: 0.027372

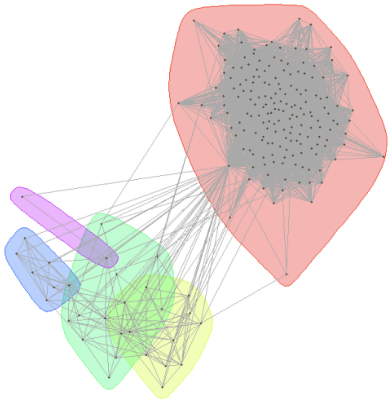
Community Structure of Fast-Greedy for Node ID=1087



Community Structure of Edge-Betweenness for Node ID=1087



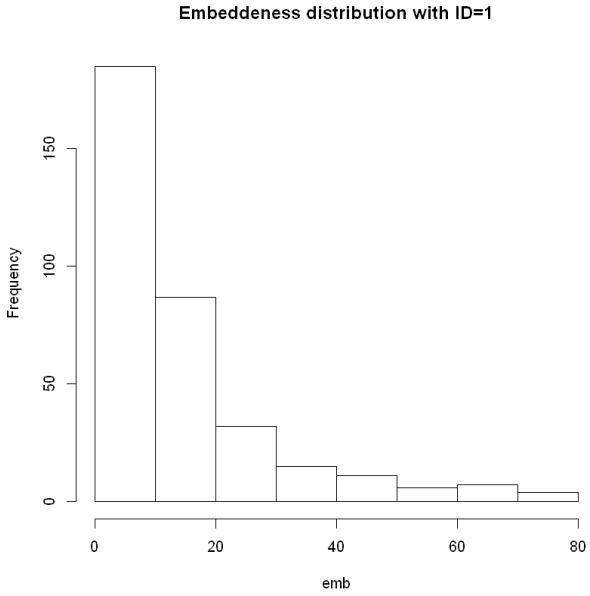
Community Structure of Infomap for Node ID=1087

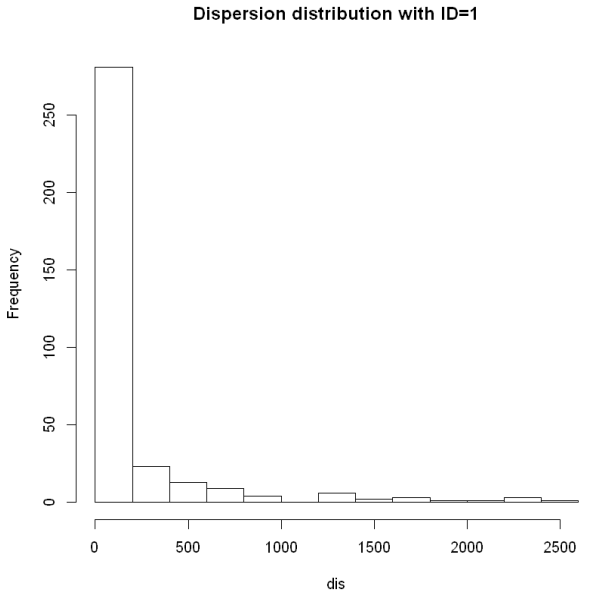


Compare the modularity score of the community structure of personalized networks in question 9 & 10, we can suggest that the modularity score will increase in all 5 personalized networks after the core node is removed, which means modified personalized networks with higher modularity score indicates a better community structure.

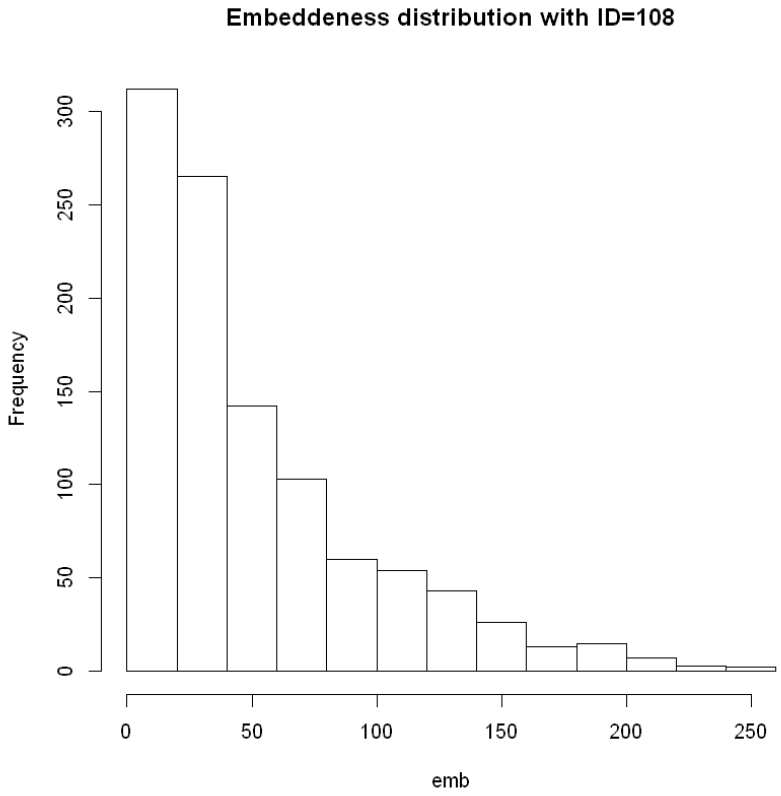
Question 11:
Since the direct connection between core node (C) and non-core node (N) in the personalized network is 1, then the expression of Embeddedness (E) is E = Degree(N) - 1, where Degree(N) is the degree if the non-core node in the personalized network of the core node.

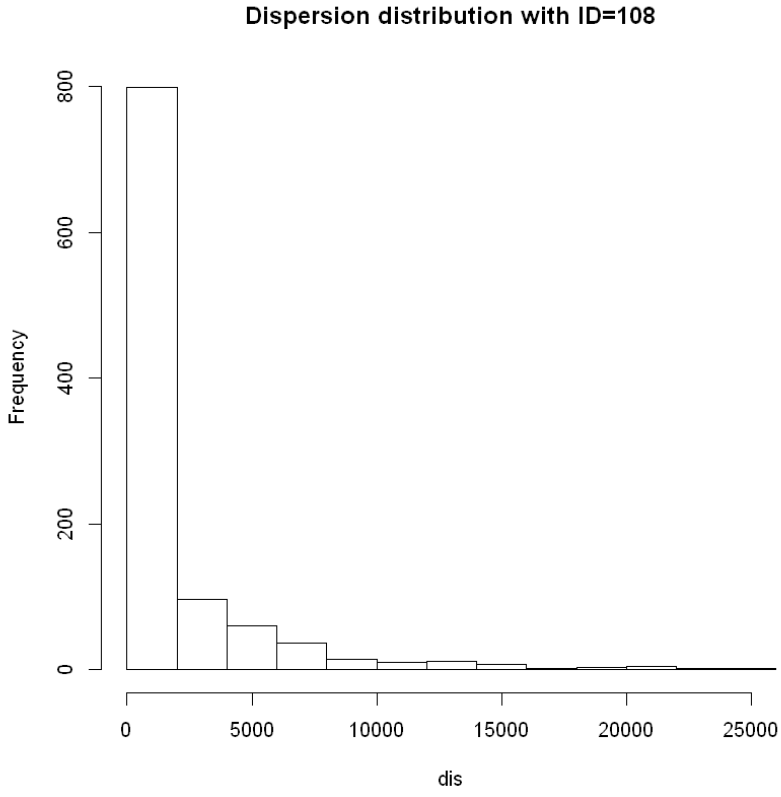
Question 12:
Node ID 1:



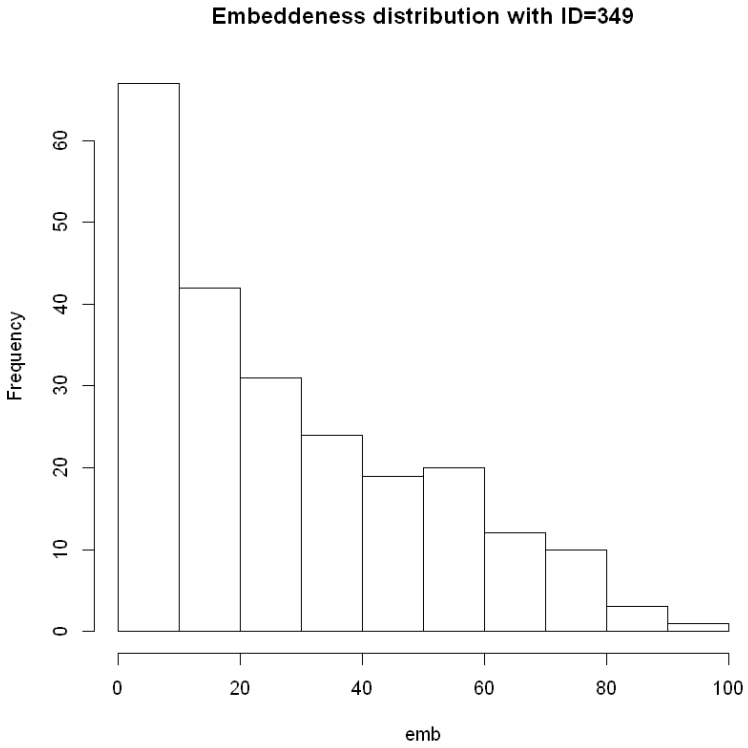


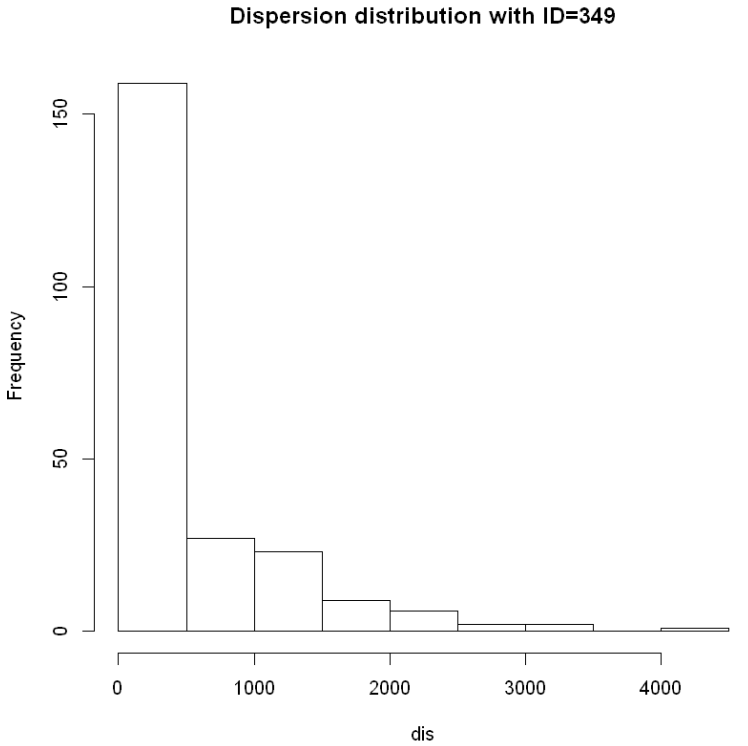
Node ID 108:


Embeddeness distribution with ID=108
Frequency
0
50
100
150
200
250
300
0
50
100
150
200
250
emb

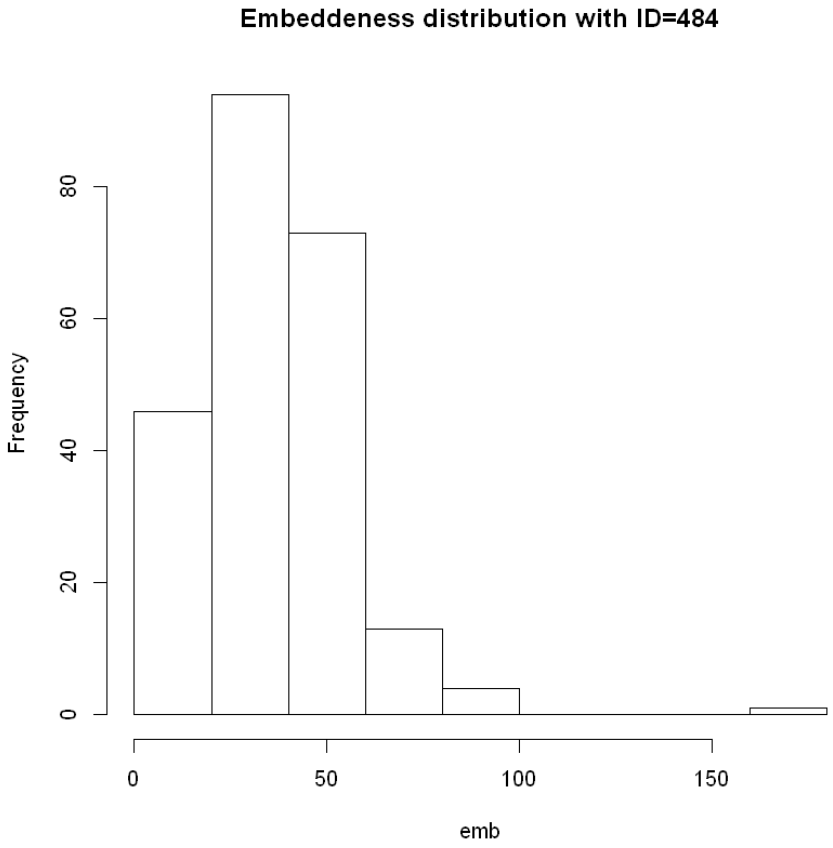

Dispersion distribution with ID=108
Frequency
0
200
400
600
800
0
5000
10000
15000
20000
25000
dis

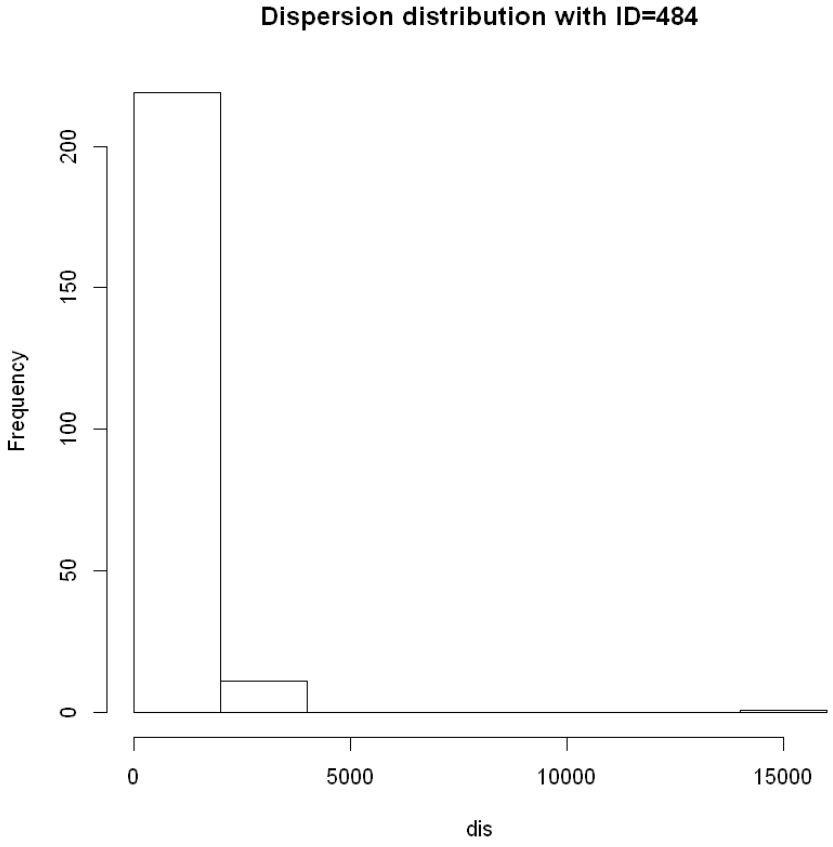
Node ID 349:



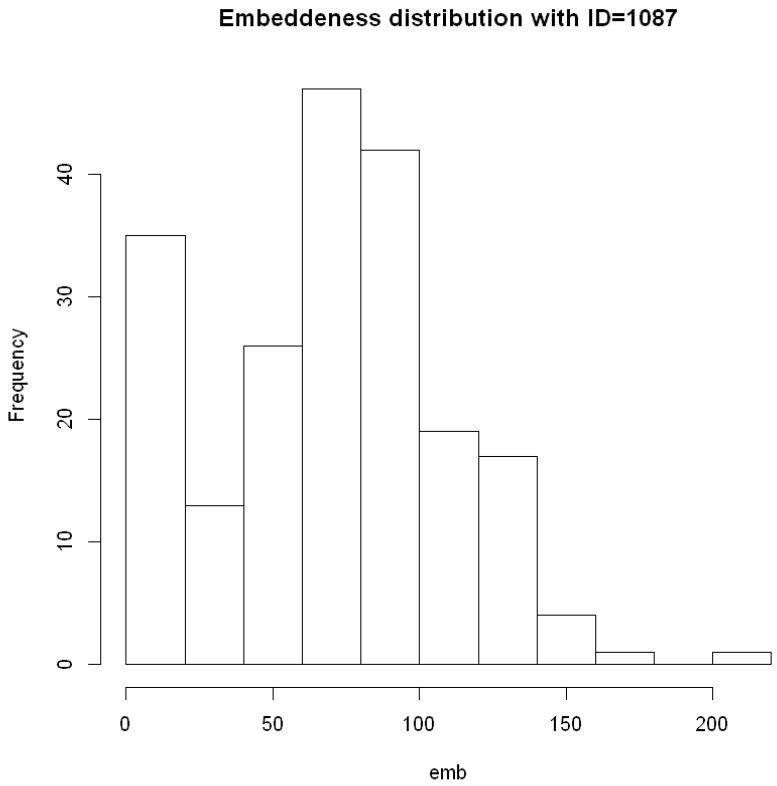


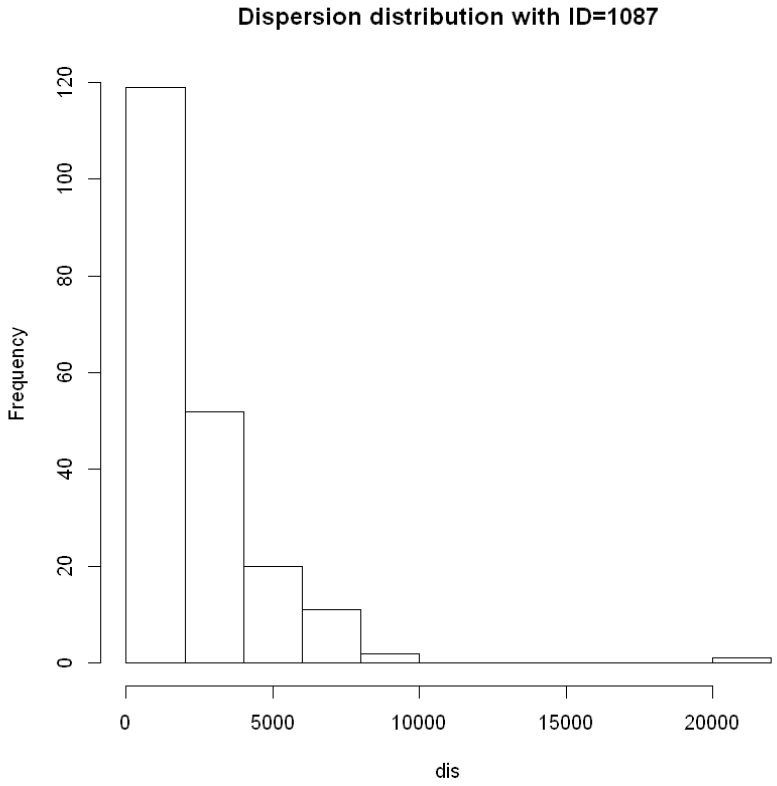
Node ID 484:


Embeddeness distribution with ID=484
Frequency
0
20
40
60
80
0
50
100
150
emb


Dispersion distribution with ID=484
Frequency
0
50
100
150
200
0
5000
10000
15000
dis

Node ID 1087:


Embeddeness distribution with ID=1087
Frequency
0
10
20
30
40
0
50
100
150
200
emb


Dispersion distribution with ID=1087
Frequency
0
20
40
60
80
100
120
0
5000
10000
15000
20000
dis

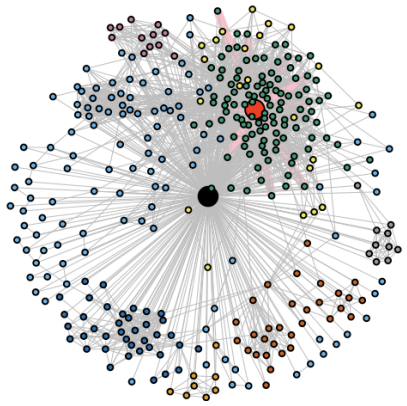
Question 13 & 14
Maximum dispersion for each of core node's personalized network:
Node ID 1: 57
Node ID 108: 1023
Node ID 349: 33
Node ID 484: 1
Node ID 1087:1

Maximum embeddedness for each of core node's personalized network:
Node ID 1: 57
Node ID 108: 1023
Node ID 349: 33
Node ID 484: 1
Node ID 1087:1

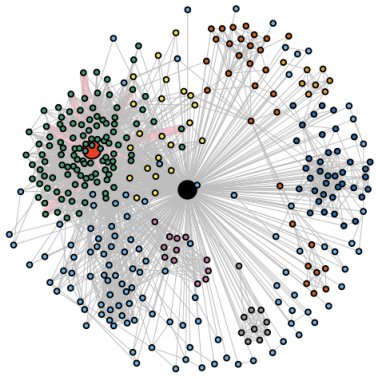
Maximum dispersion/embeddedness for each of core node's personalized network:
Node ID 1: 26
Node ID 108: 1023
Node ID 349: 33
Node ID 484: 1
Node ID 1087:1

The following graph shows what the cord node is. Highlight red node is the core node with the pink highlight edges.

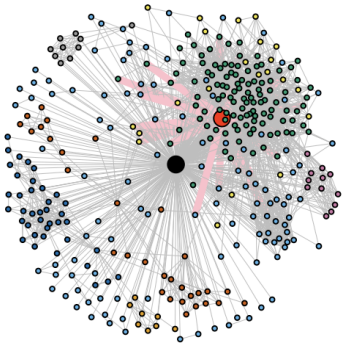
Node ID: 1 , Max Dispersion Node: 57



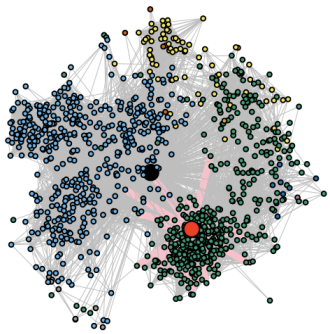
Node ID: 1 , Max Embeddedness Node: 57



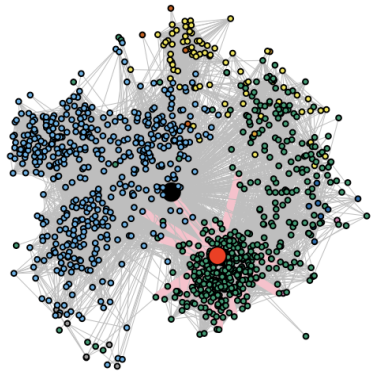
Node ID: 1 , Max Dispersion/Embeddedness Node: 26



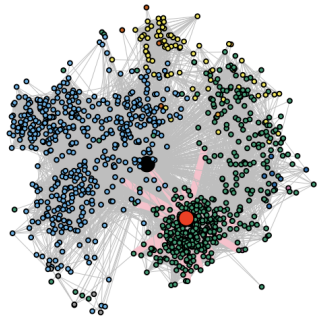
Node ID: 108 , Max Dispersion Node: 1023



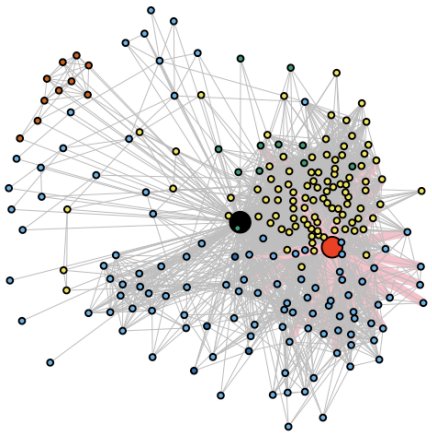
Node ID: 108 , Max Embeddedness Node: 1023



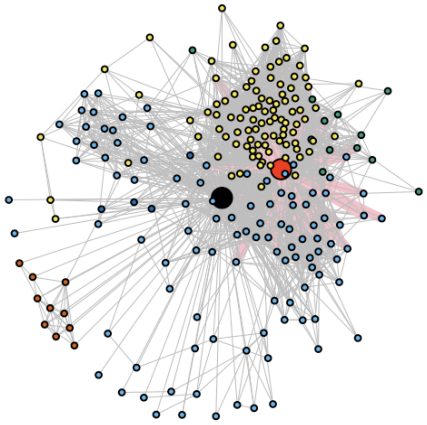
Node ID: 108 , Max Dispersion/Embeddedness Node: 1023



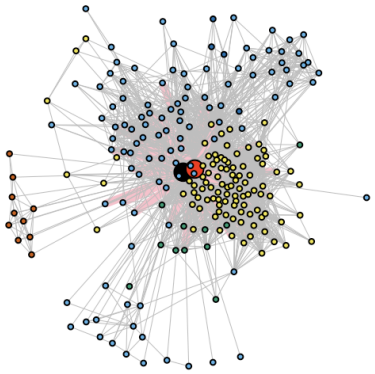
Node ID: 349 , Max Dispersion Node: 33



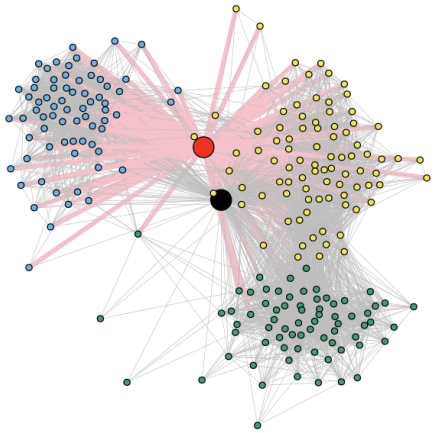
Node ID: 349 , Max Embeddedness Node: 33



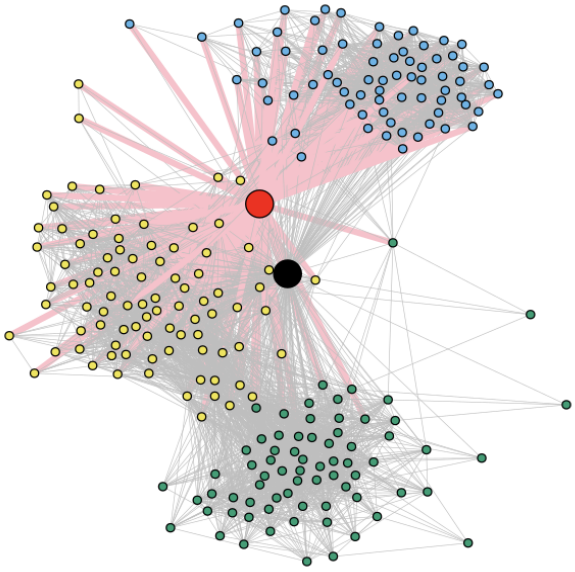
Node ID: 349 , Max Dispersion/Embeddedness Node: 33



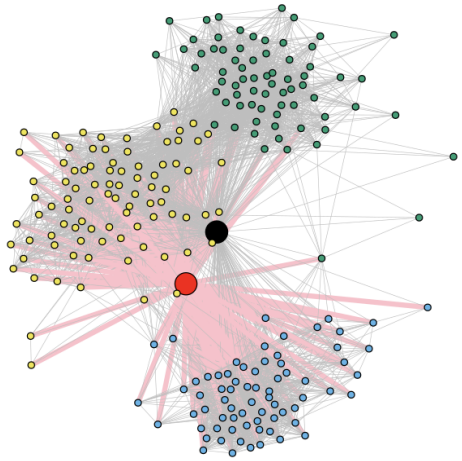
Node ID= 484 , Max Embeddedness Node : 1



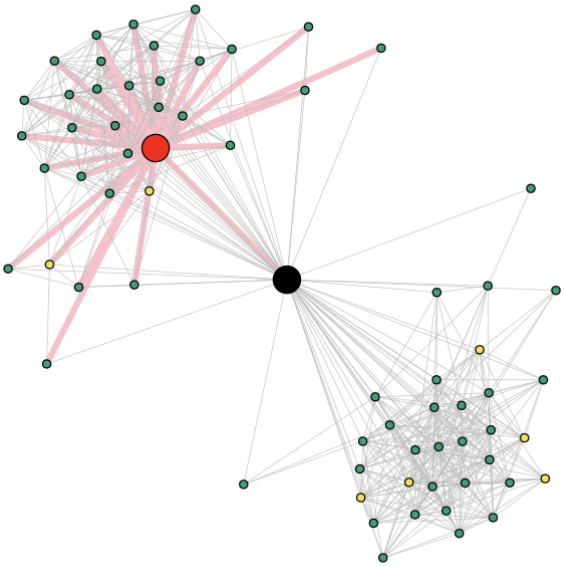
Node ID= 484 , Max Dispersion Node : 1



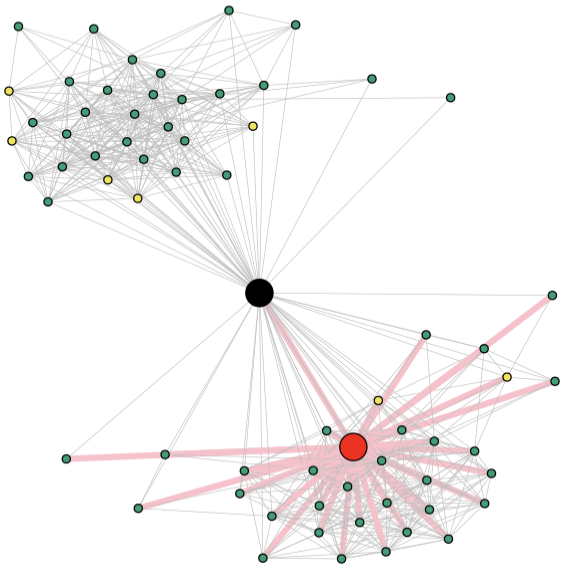
Node ID= 484 , Max Dispersion/Embeddedness Node : 1



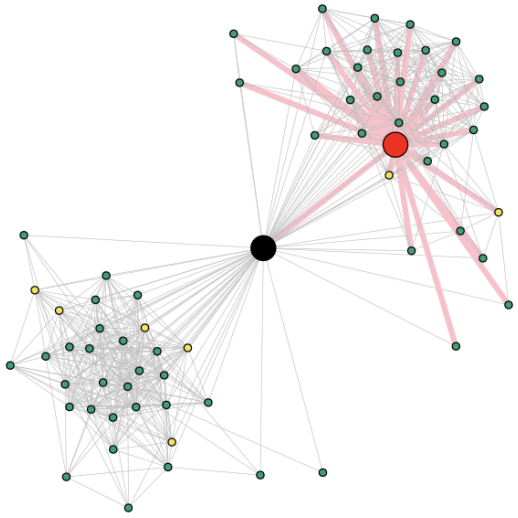
Node ID= 1086 , Max Embeddedness Node : 1



Node ID= 1086 , Max Dispersion Node : 1



Node ID= 1086 , Max Dispersion/Embeddedness Node : 1



Question 15:
For the maximum dispersion, we can observe that the nodes distribute sparsely, which means the mutual friends between core nodes and nodes with high dispersion are not closely connected to each other. For the maximum embeddedness, we can observe that the nodes distribute densely, which means the node with high embeddedness has many mutual friends with the core nodes. This suggests that the members in the community have strong connections with each other. For the dispersion/embeddedness, we can use this to know how close connections are between the non-core node and the core node. If the dispersion/embeddedness is high, we can infer that the node has high dispersion and low embeddedness which means this node has mutual friends with the core node but with less close connection. Also, the node and core node may come from different communities.

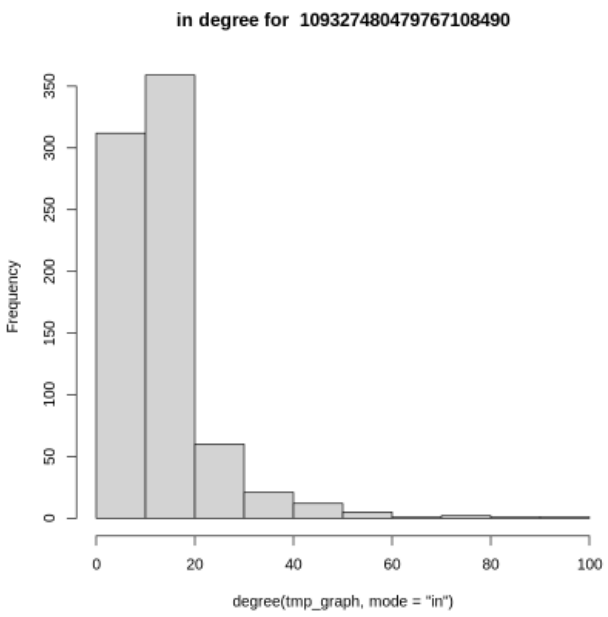
Question 16:
The length of the list Nr is 11.

Question 17
The average accuracy of the friend recommendation algorithm:
Common Neighbors Measure:0.842022825659189
Jaccard Measure: 0.832441951987407
Adamic Adar measure: 0.857924701561065
There is no significant difference among these accuracy, but we can observe that the accuracy of the Adamic Adar measure is highest. Hence the average accuracy of the Adamic Adar Measure is the best.

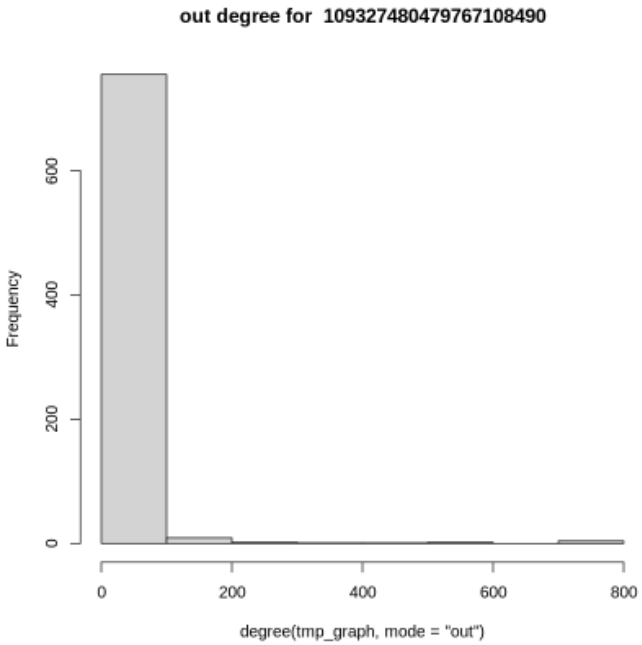
Question 18:
There are 132 nodes and there are 57 personal networks

Question 19:
Node ID: 10932748047976710849O

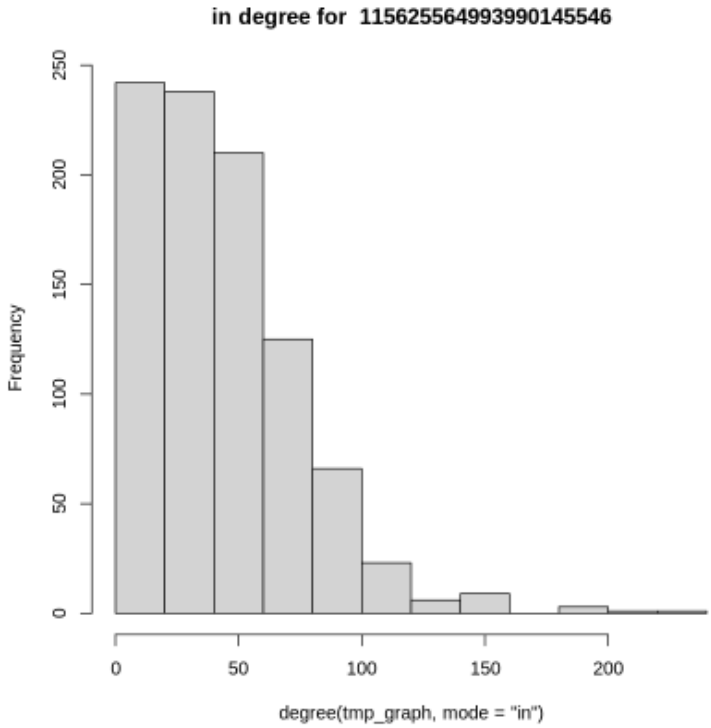
```
[1] 69
```



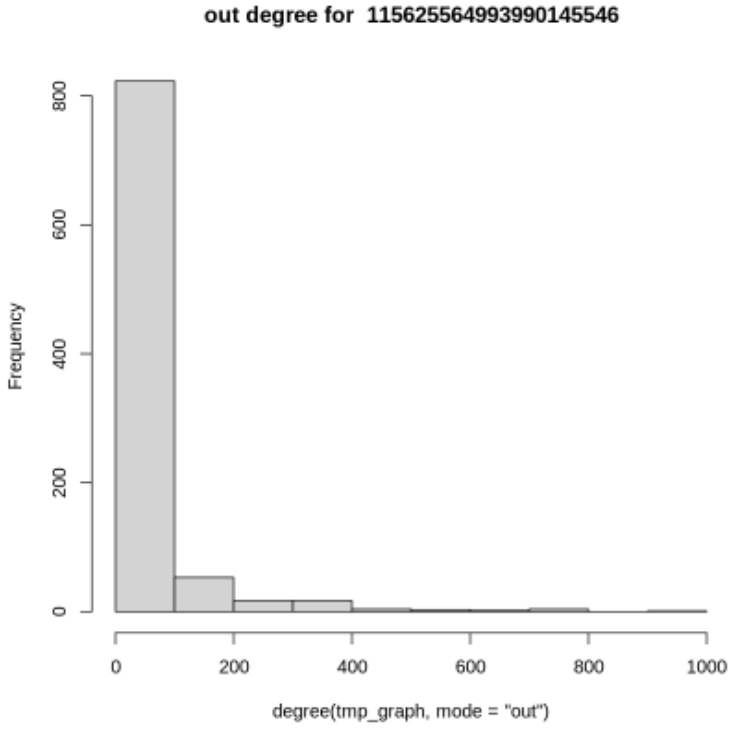

```
[1] 115
```



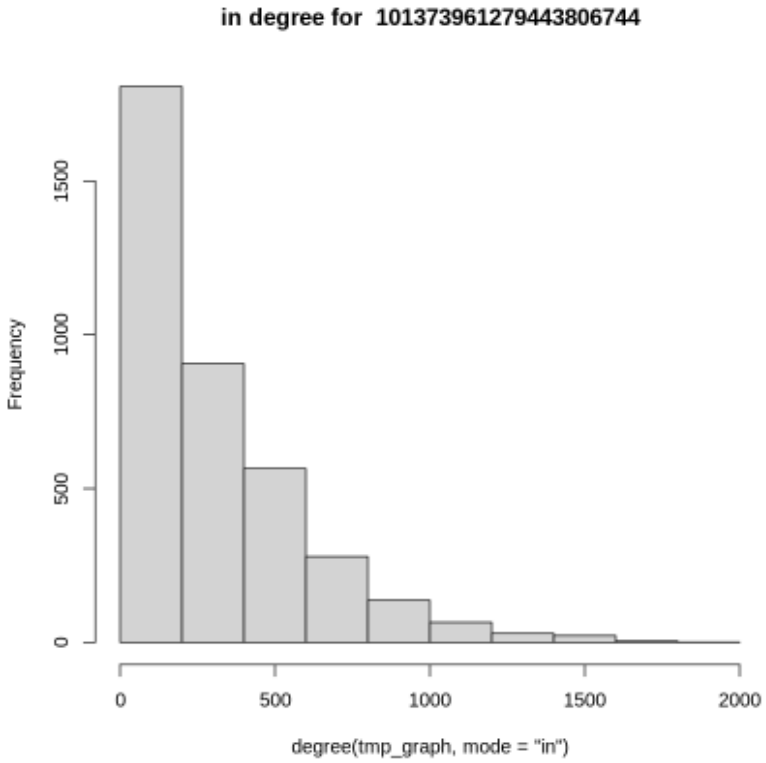
Node ID: 115625564993990145546

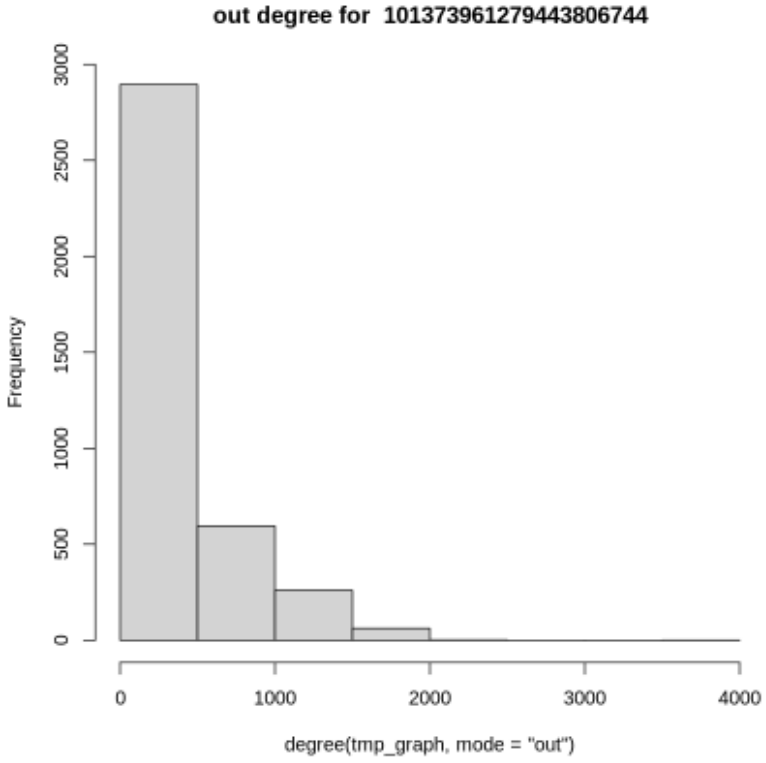



```
[1] 17
```



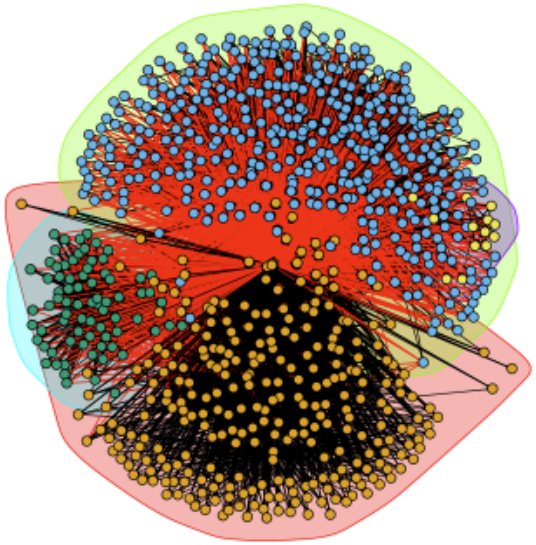
Node ID: 101373961279443806744


in degree for 10137396127944380674
Frequency
0
500
1000
1500
0
500
1000
1500
2000
degree(tmp_graph, mode = "in")

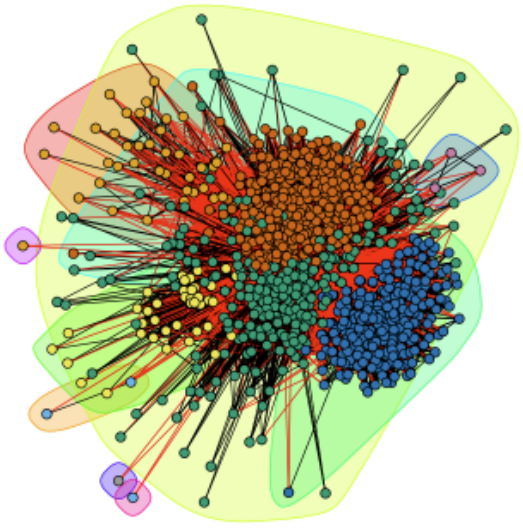

out degree for 10137396127944380674
Frequency
0
500
1000
1500
2000
2500
3000
0
1000
2000
3000
4000
degree(tmp_graph, mode = "out")

The in-degree and out-degree distribution of 3 personal networks are plotted above. We can observe that both in-degree and out-degree's plots have the same tendency, but their distributions are different.

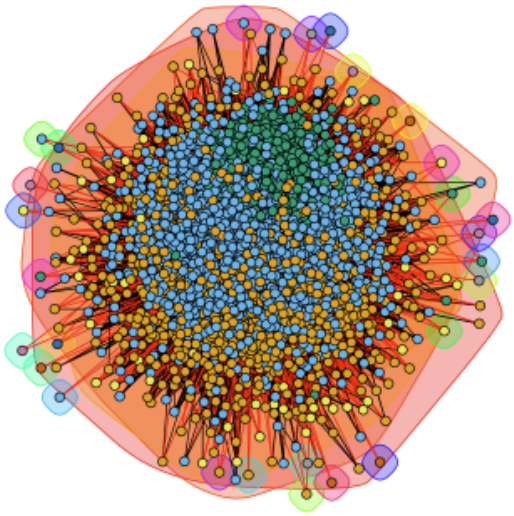
Question 20:
Modularity Scores:
Node ID 109327480479767108490: 0.2527654



Node ID 115625564993990145546: 0.3194726



Node ID 101373961279443806744: 0.1910903



From above plots and modularity scores, we find that these modularity scores are not similar. Higher modularity scores means the network has a clear community structure with dense communities. For the third personal network, it has the worst modularity score which means it has no clear and dense communities in the network.

Question 21:
Homogeneity: It shows whether the nodes in one cluster belong to the same class. The homogeneity will be close to 1 if most of the nodes in the cluster are the same class. It indicates the purity of the cluster.

Completeness: It shows whether all the nodes belong to the same class are assigned to the same cluster. The completeness will be close to 1 if all the clusters have been assigned by different and complete classes.

Question 22:
Node ID 109327480479767108490:
Homogeneity: 0.8518851
Completeness: 0.3298739

Node ID 115625564993990145546:
Homogeneity: 0.4518903
Completeness: -3.423962

Node ID 101373961279443806744:
Homogeneity: 0.003866707
Completeness: -1.504238
For the first personal network, we can find it has the highest value of homogeneity which means each community in the network represents a circle well, but with low completeness. This means

that the distribution of members across circles is uncertain. For the second personal network, it has medium value of homogeneity which means a few communities in the network can represent a circle well, but it has lowest completeness and it is negative. This suggests that H(K|C)/H(K) is greater than 1 from the expression of completeness. The value of H(K|C) will be too large, which means we can not be certain to know which circle the members belong to. For the third personal network, it has lowest homogeneity and negative completeness, which means there are no clear and dense communities in the network. Also, the reason for the negative completeness is the same as the reason in the second personal network. In addition, the plots in the questions can visually show the details above.

Question 23:
We chose channels = 9, dropout = 0.6, learning_rate = 1e-3, and epoch = 144 in the 3 layers Graph Convolutional Network to get the optimal performance with accuracy of 0.53. The number of channels indicates the complexity of the model, and if we choose channels = 9, the model can help each node get more features. The dropout = 0.6 can help to prevent overfitting and improve generalization. The learning rate suggests how fast the training process is. Epoch means how many times the model will iterate, and larger epoch can allow the model to learn mode from the data. Choosing the learning_rate = 1e-3 and epoch =144 will get the best accuracy for the model.

Question 24:
Node2Vec explores the network structure by random walk to find the node features. In this case, we use the SVM classifier, and its performance is the best because compared to other classifiers, the SVM is well-suited for 1433-dimensional text features that are high-dimensional data. Also, the SVM can be faster to train compared to other classifiers. After combining the Node2Vec and text features and training the classifier, we can get the best classification accuracy is 0.6648.

Question 25:
Accuracy for a random walk with different customized properties:
No teleportation, NO TFIDF:

unvisited = 1

	precision	recall	f1-score	support
0	0.65	0.72	0.68	285
1	0.80	0.90	0.85	406
2	0.82	0.53	0.64	726
3	0.76	0.68	0.72	379
4	0.43	0.57	0.49	214
5	0.29	0.77	0.42	131
6	0.62	0.49	0.55	344
accuracy			0.65	2485
macro avg	0.62	0.67	0.62	2485
weighted avg	0.70	0.65	0.65	2485

0.6454728370221328

With teleportation, without TFIDF:

```
tp:  0
unvisited =  0
              precision    recall  f1-score   support

           0       0.29      0.48      0.36       285
           1       0.49      0.57      0.53       406
           2       0.64      0.33      0.43       726
           3       0.67      0.58      0.62       379
           4       0.13      0.17      0.15       214
           5       0.09      0.18      0.12       131
           6       0.33      0.27      0.30       344

    accuracy                           0.39      2485
   macro avg       0.38      0.37      0.36      2485
weighted avg       0.46      0.39      0.41      2485

0.39235412474849096
tp:  0.1
unvisited =  0
              precision    recall  f1-score   support

           0       0.66      0.81      0.73       285
           1       0.81      0.91      0.86       406
           2       0.86      0.53      0.65       726
           3       0.79      0.70      0.74       379
           4       0.50      0.80      0.62       214
           5       0.42      0.79      0.55       131
           6       0.65      0.58      0.62       344

    accuracy                           0.69      2485
   macro avg       0.67      0.73      0.68      2485
weighted avg       0.74      0.69      0.70      2485

0.6945674044265594
tp:  0.2
unvisited =  0
              precision    recall  f1-score   support

           0       0.71      0.81      0.76       285
           1       0.83      0.92      0.87       406
           2       0.85      0.55      0.67       726
           3       0.81      0.71      0.76       379
           4       0.52      0.84      0.64       214
           5       0.43      0.80      0.56       131
           6       0.66      0.60      0.63       344

    accuracy                           0.71      2485
   macro avg       0.69      0.75      0.70      2485
weighted avg       0.75      0.71      0.71      2485

0.7106639839034206
```

Without teleportation, with TFIDF:

```
unvisited =  0
              precision    recall  f1-score   support

           0       0.31      0.51      0.39       285
           1       0.48      0.56      0.52       406
           2       0.50      0.26      0.34       726
           3       0.62      0.51      0.56       379
           4       0.09      0.13      0.11       214
           5       0.12      0.27      0.17       131
           6       0.30      0.23      0.26       344

    accuracy                           0.36      2485
   macro avg       0.35      0.35      0.33      2485
weighted avg       0.41      0.36      0.37      2485

0.3609657947686117
```

With teleportation, with TFIDF:

```
tp:  0
unvisited =  0
              precision    recall  f1-score   support

           0       0.32      0.49      0.39       285
           1       0.49      0.59      0.53       406
           2       0.62      0.33      0.43       726
           3       0.70      0.55      0.62       379
           4       0.13      0.17      0.15       214
           5       0.11      0.25      0.15       131
           6       0.30      0.25      0.27       344

    accuracy                           0.39      2485
   macro avg       0.38      0.38      0.36      2485
weighted avg       0.46      0.39      0.41      2485

0.39436619718309857
tp:  0.1
unvisited =  0
              precision    recall  f1-score   support

           0       0.68      0.81      0.74       285
           1       0.78      0.92      0.84       406
           2       0.87      0.54      0.67       726
           3       0.80      0.70      0.74       379
           4       0.53      0.82      0.64       214
           5       0.42      0.80      0.55       131
           6       0.66      0.60      0.63       344

    accuracy                           0.70      2485
   macro avg       0.68      0.74      0.69      2485
weighted avg       0.74      0.70      0.70      2485

0.7026156941649899
```

```
tp:  0.2
unvisited =  0
              precision    recall  f1-score   support

           0       0.71      0.81      0.76       285
           1       0.80      0.93      0.86       406
           2       0.86      0.56      0.68       726
           3       0.80      0.71      0.75       379
           4       0.53      0.83      0.65       214
           5       0.43      0.79      0.55       131
           6       0.67      0.60      0.64       344

    accuracy                           0.71      2485
   macro avg       0.69      0.75      0.70      2485
weighted avg       0.75      0.71      0.71      2485

0.7130784708249497
```

The purpose of the teleportation used in the pagerank is to prevent the random walk from going to the dead end or trap. Through teleportation, the random walker can reach any page even though there is no link to the current page. TFIDF is to check the relevance of the documents to a query. The frequency of items in TFIDF suggests how important it is in the documents. Through these details of teleportation and TFIDF and above accuracy, we can suggest that the random walk with both teleportation and TFIDF has highest accuracy. In addition, when random walk without teleportation or the teleportation probability equal to 0, the accuracy of mode will be small (around 0.37). Then we can conclude that whether to use teleportation will have a significant difference on the performance of random walk because random walkers are more likely to approach the dead end and trap without teleportation.