Large-Scale Data Mining: Models and Algorithms
ECE 232E Spring 2018

Project 2
Social Network Mining

Qinyi Tang (204888348)
Shuo Bai (505032786)
Jinxi Zou (605036454)
Xuan Hu (505031796)

5/7/2018

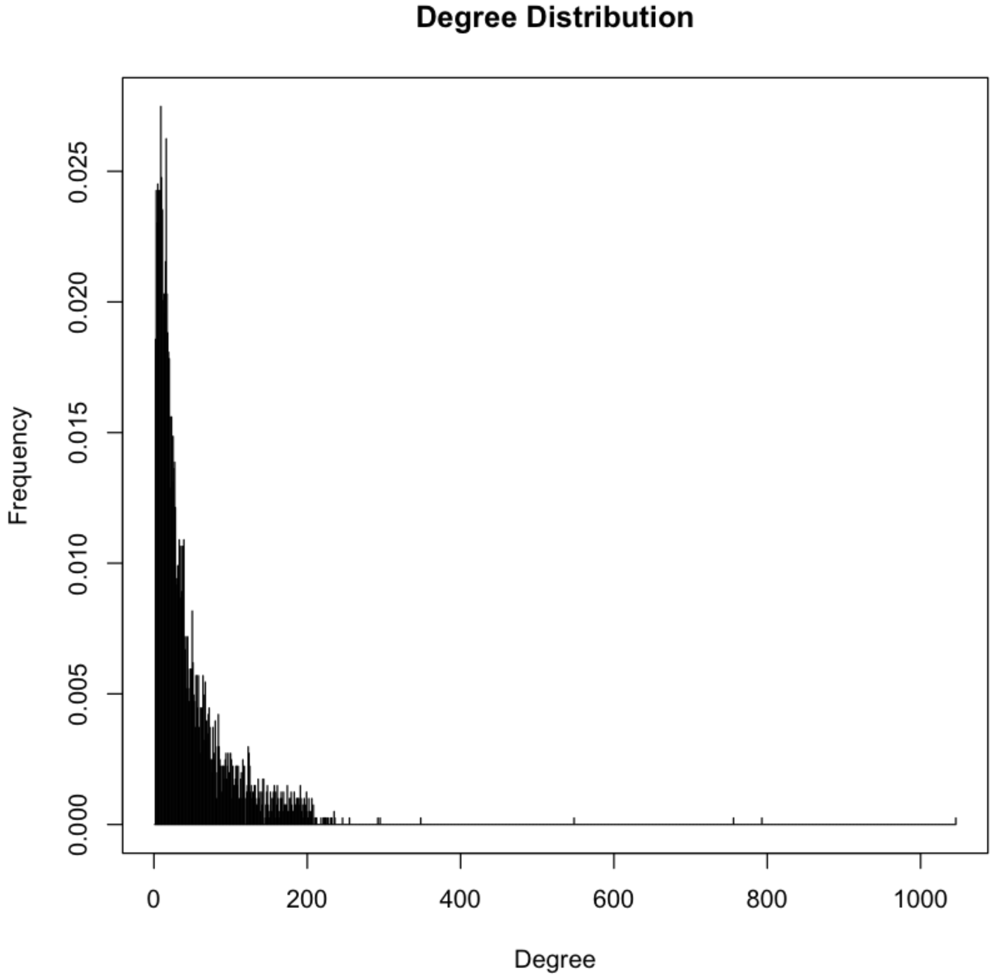
1 Facebook network

1.1 Structural properties of the facebook network

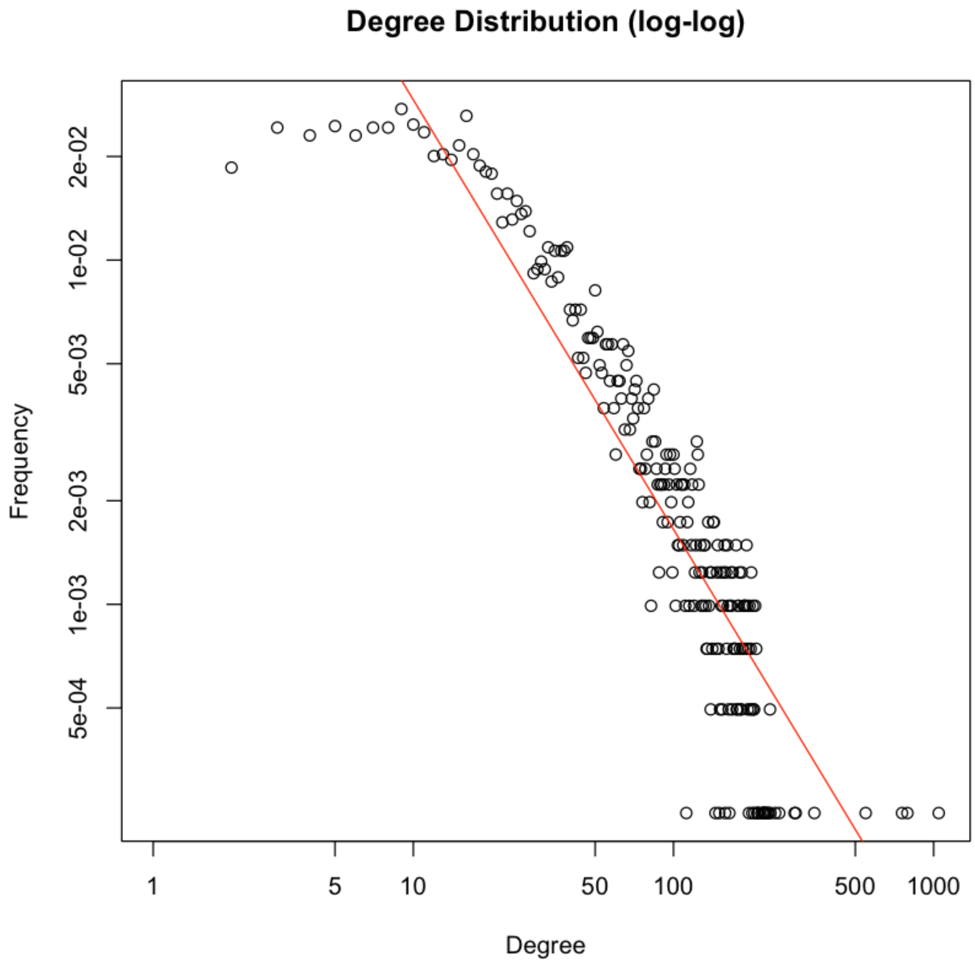
Answer for Q1: The facebook network is connected. It has 4039 nodes and 88234 edges.

Answer for Q2: The diameter of the network is 8.

Answer for Q3: The degree distribution of the facebook network is shown as the following. The average degree is 43.6910126268878.



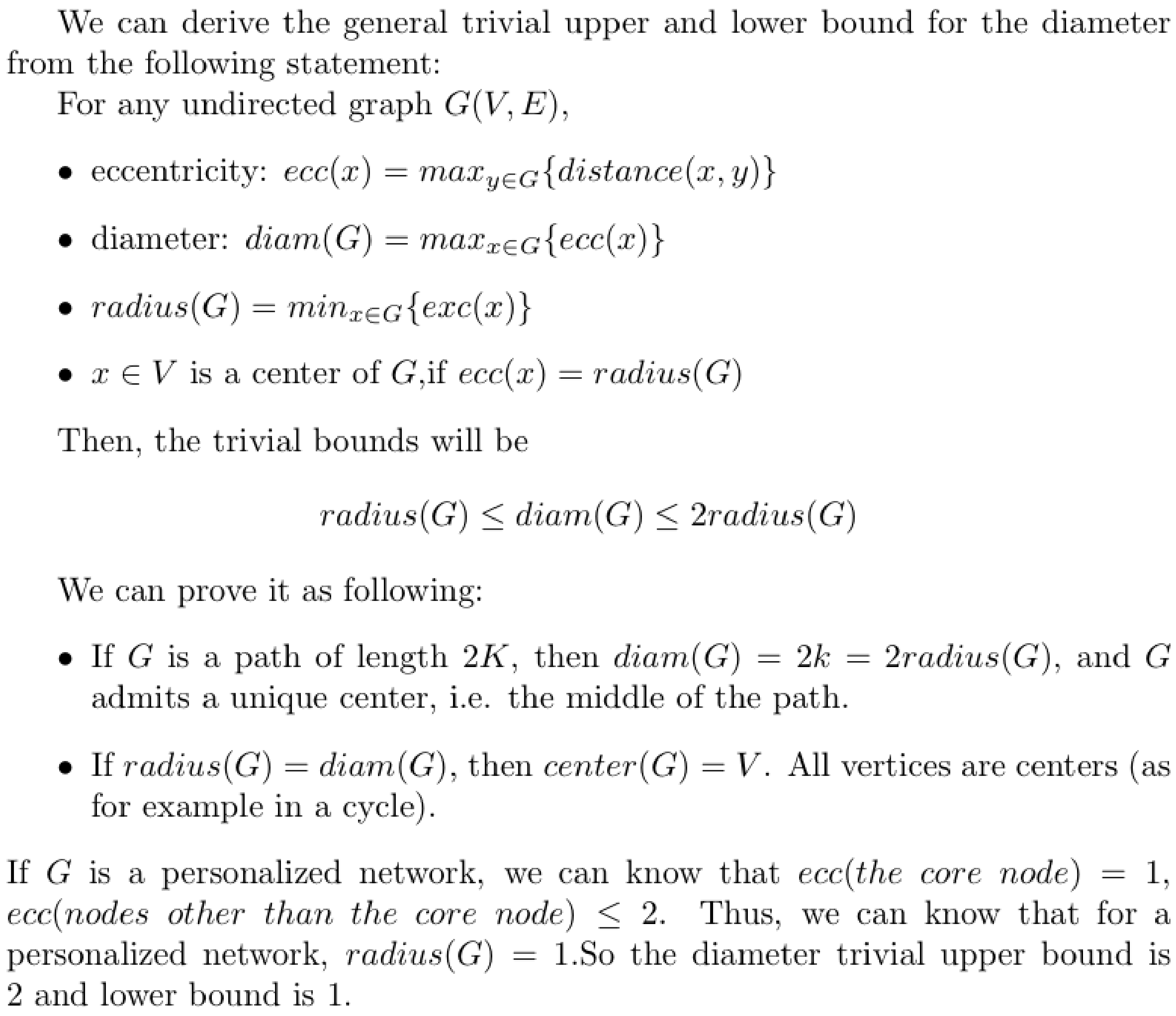
Answer for Q4: The degree of the distribution in log-log scale is shown below, and the line in red is the fitted line. We implemented the fitted line using the function “lm” and omitted 819 zero values when doing log computation. The estimated slope of the fitted line is -1.24752626790776.



1.2 Personalized network

Answer for Q5: We created the personalized network of the user whose ID is 1 by creating a subgraph with the node 1 and all its neighbors. It has 348 nodes and 2866 edges.

Answer for Q6: The diameter of the personalized network is 2. The trivial upper and lower bound of the diameter is 2 and 1, and the reason is given below.



We can derive the general trivial upper and lower bound for the diameter from the following statement:

For any undirected graph $G(V, E)$,

- eccentricity: $ecc(x) = max_{y \in G}\{distance(x, y)\}$
- diameter: $diam(G) = max_{x \in G}\{ecc(x)\}$
- $radius(G) = min_{x \in G}\{exc(x)\}$
- $x \in V$ is a center of G,if $ecc(x) = radius(G)$

Then, the trivial bounds will be

$$radius(G) \leq diam(G) \leq 2radius(G)$$

We can prove it as following:

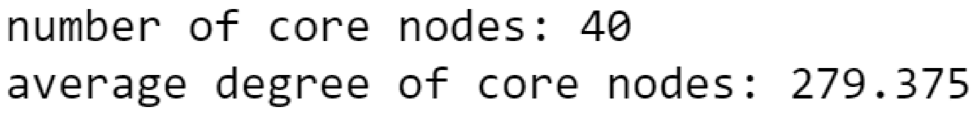
- If G is a path of length $2K$, then $diam(G) = 2k = 2radius(G)$, and G admits a unique center, i.e. the middle of the path.
- If $radius(G) = diam(G)$, then $center(G) = V$. All vertices are centers (as for example in a cycle).

If G is a personalized network, we can know that $ecc(the\ core\ node) = 1$, $ecc(nodes\ other\ than\ the\ core\ node) \leq 2$. Thus, we can know that for a personalized network, $radius(G) = 1$.So the diameter trivial upper bound is 2 and lower bound is 1.

Answer for Q7: When the diameter equals the trivial upper bound, it means that there exists two neighbors nodes (two nodes connected to the core node) has no connection. The path between these two nodes is 2 because they both connected to the core node from the definition of the personalized network. When the diameter equals the trivial lower bound, it means that any two neighbors nodes are connected, which indicates any two nodes in the graph is connected to each other.

1.3 Core node's personalized network

Answer for Q8:



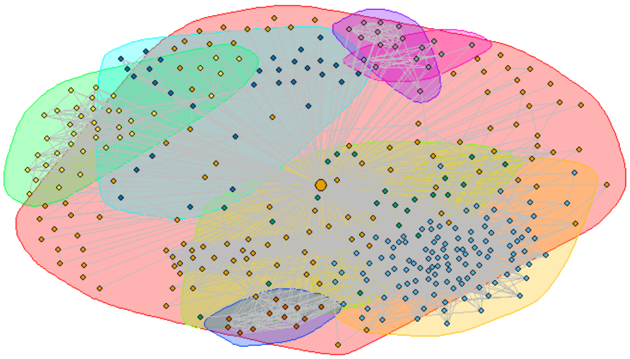
```
number of core nodes: 40
average degree of core nodes: 279.375
```

Answer for Q9:

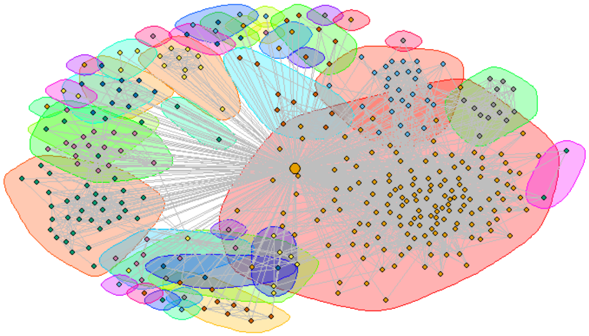
Node\Detection Algorithms	Fast-Greedy	Edge-Betweenness	Infomap
1	0.4131014	0.3533022	0.3891185
108	0.4359294	0.5067549	0.5082492
349	0.2517149	0.133528	0.203753
484	0.5070016	0.4890952	0.5152788
1087	0.1455315	0.02762377	0.02690662

Table for modularity score

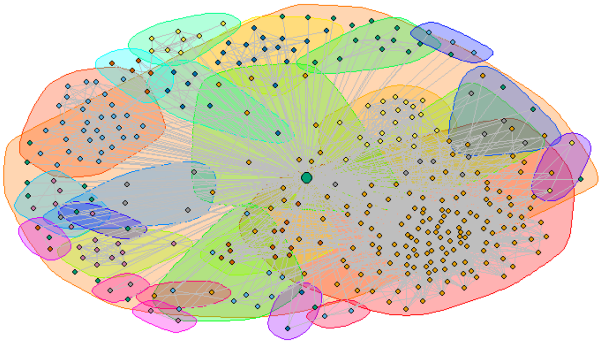
fast_greedy for Node 1



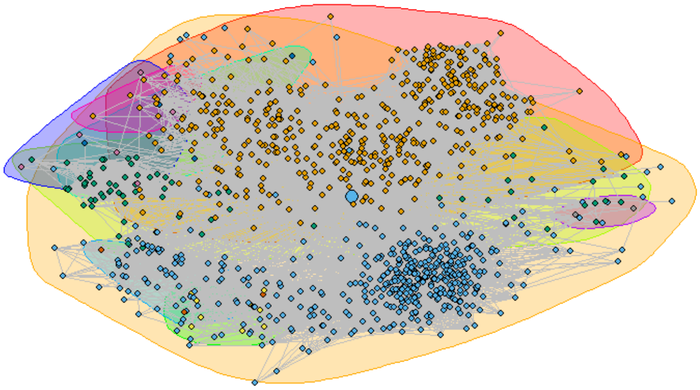
edge_betweenness for Node 1



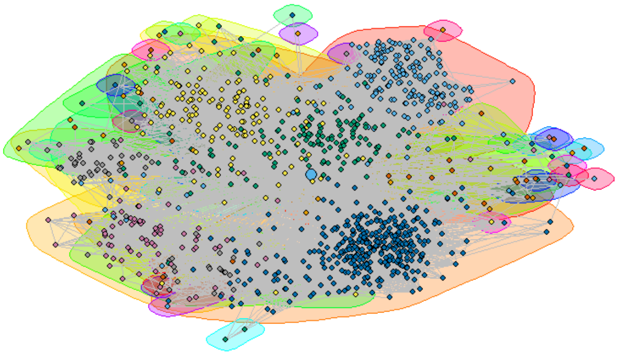
infomap for Node 1



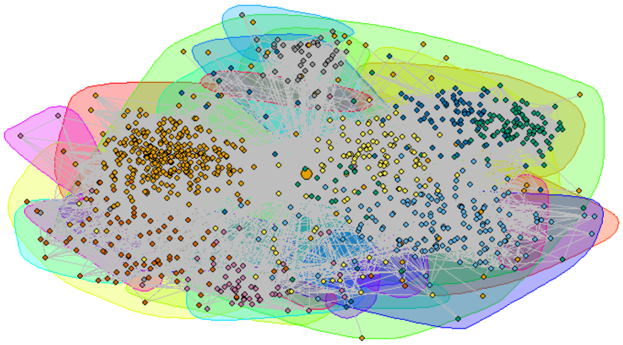
fast_greedy for Node 108



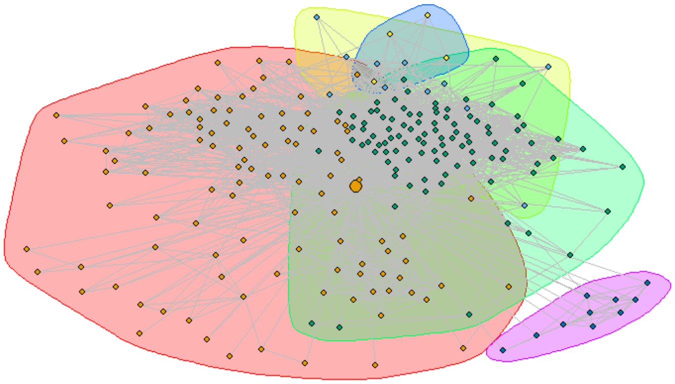
edge_betweenness for Node 108



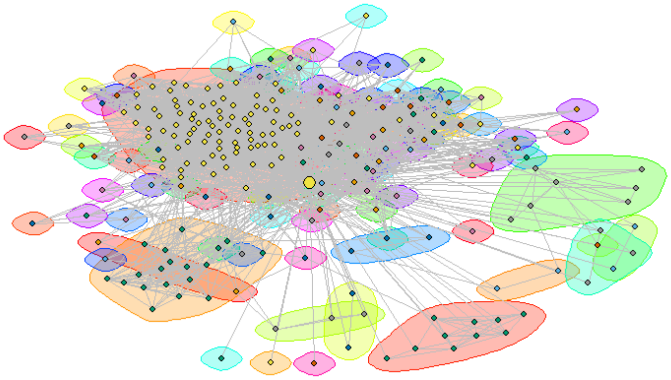
infomap for Node 108



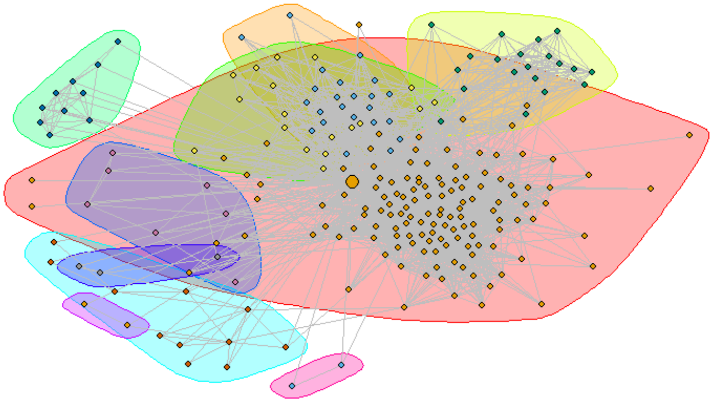
fast_greedy for Node 349



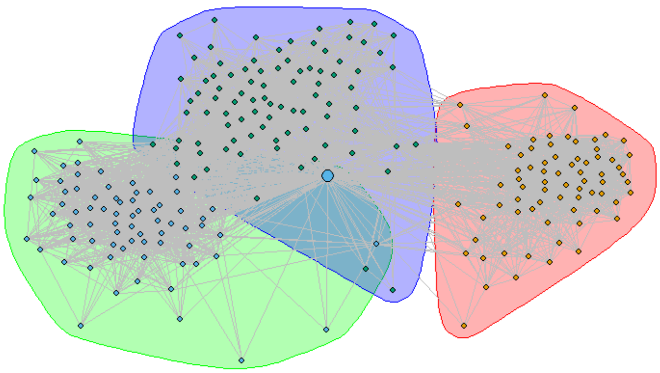
edge_betweenness for Node 349



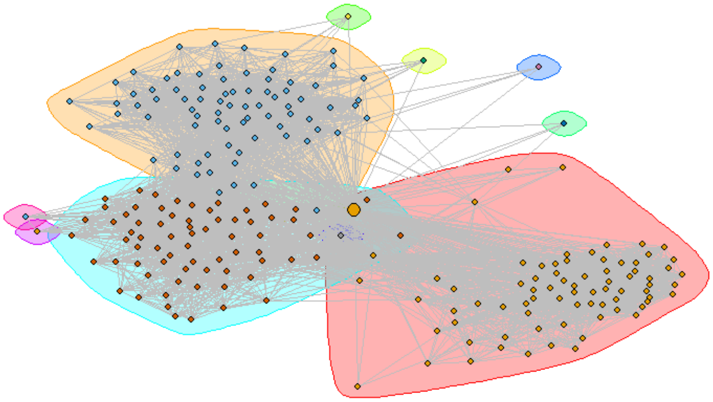
infomap for Node 349



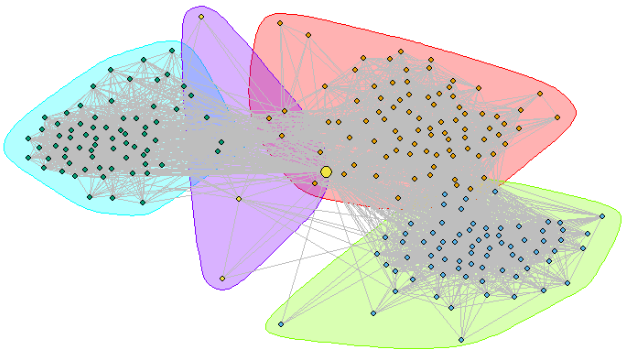
fast_greedy for Node 484



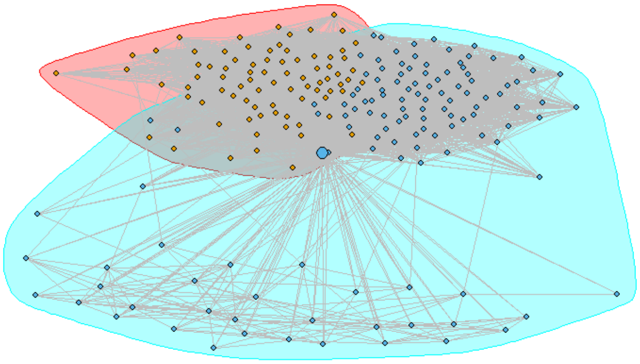
edge_betweenness for Node 484



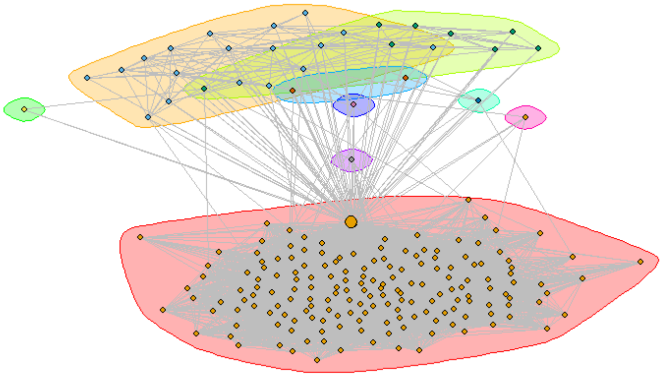
infomap for Node 484



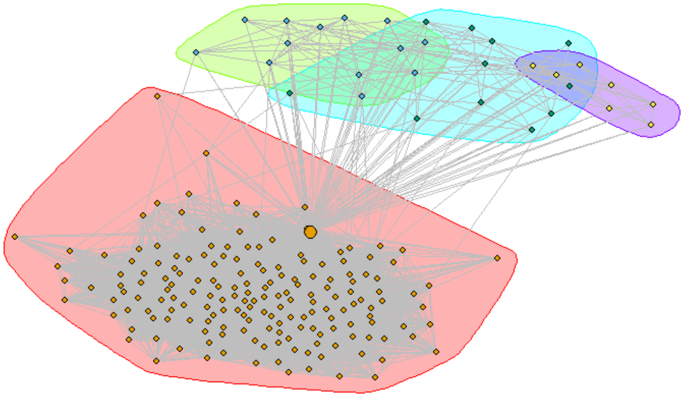
fast_greedy for Node 1087



edge_betweenness for Node 1087



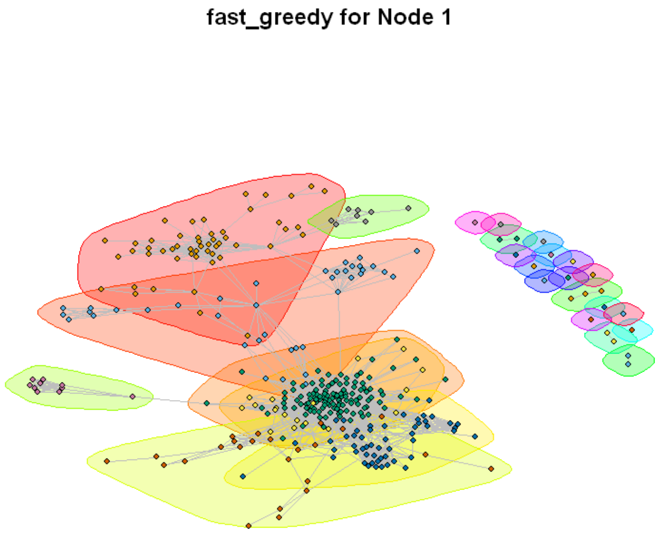
infomap for Node 1087



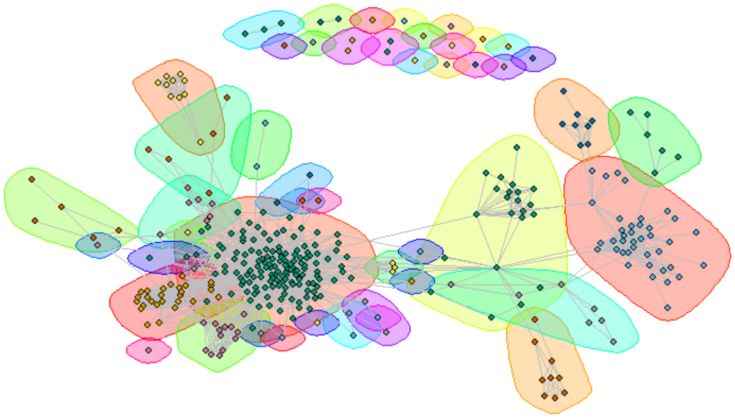
Answer for Q10: In contrast with question 9, if we remove the core node from network, we get less community of different algorithm. Besides that, although we removed core nodes, there isn't a big influence on modularity score, which means even without core node, we still get similar distribution of nodes.

Node\Detection Algorithms	Fast-Greedy	Edge-Betweenness	Infomap
1	0.4418533	0.4161461	0.4180077
108	0.4581271	0.5213216	0.5201497
349	0.2456918	0.1505663	0.2448156
484	0.5342142	0.5154413	0.5434437
1087	0.1481956	0.0324953	0.02737159

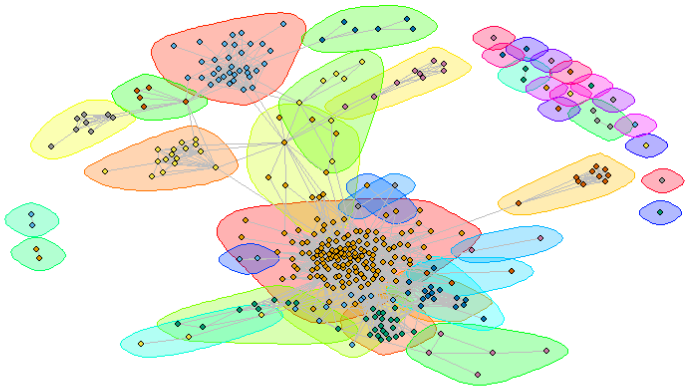
Table for modularity score



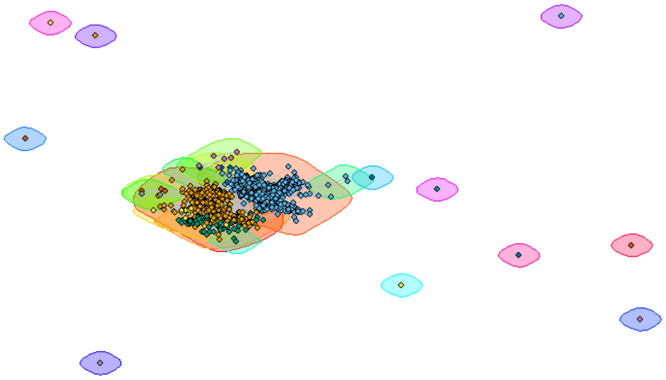
edge_betweenness for Node 1



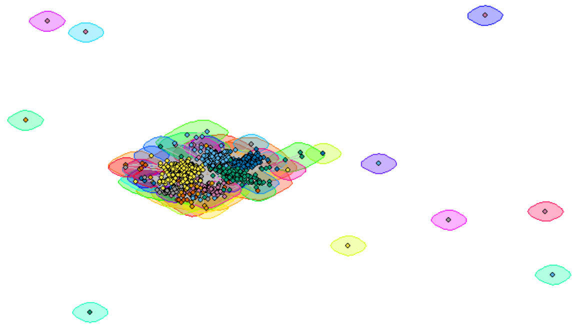
infomap for Node 1



fast_greedy for Node 108

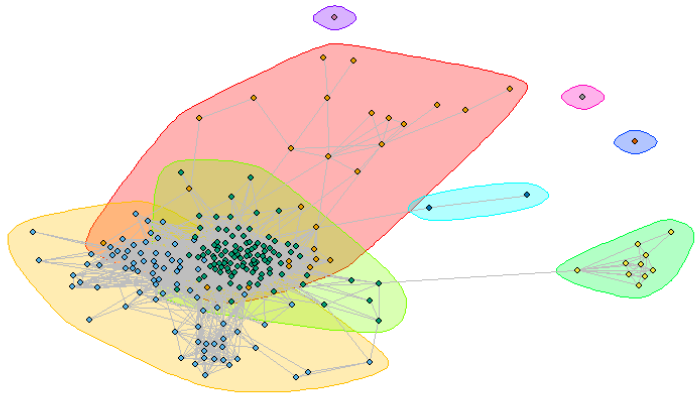


edge_betweenness for Node 108

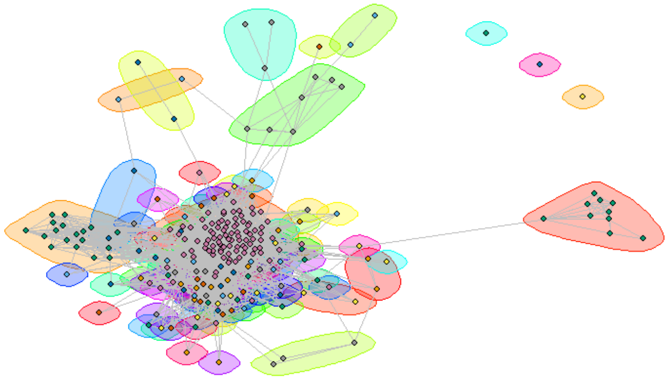


infomap for Node 108

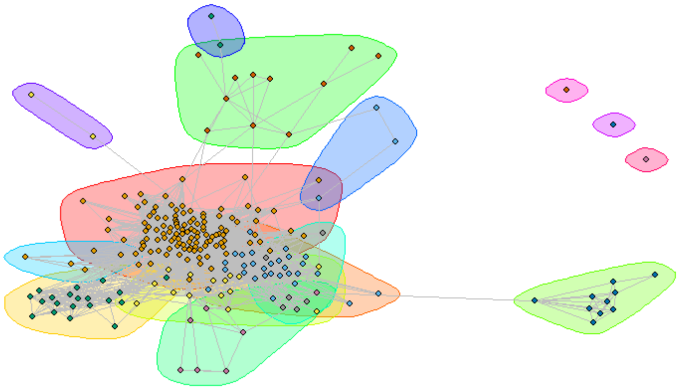
fast_greedy for Node 349



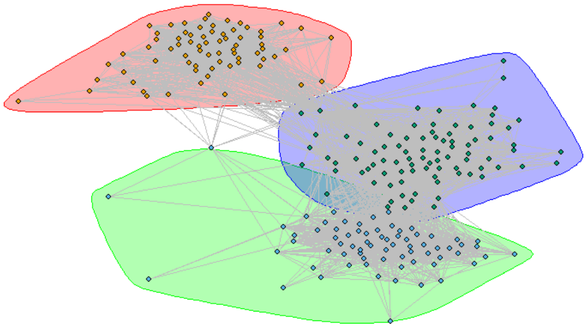
edge_betweenness for Node 349



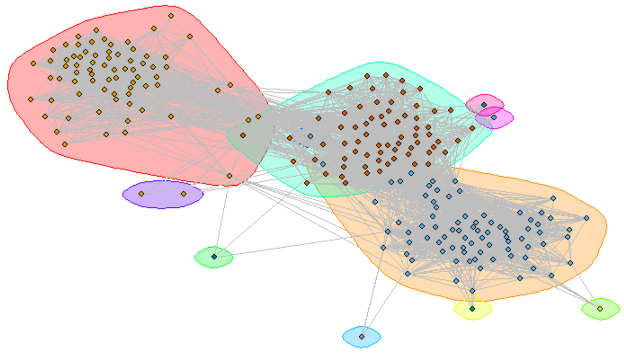
infomap for Node 349



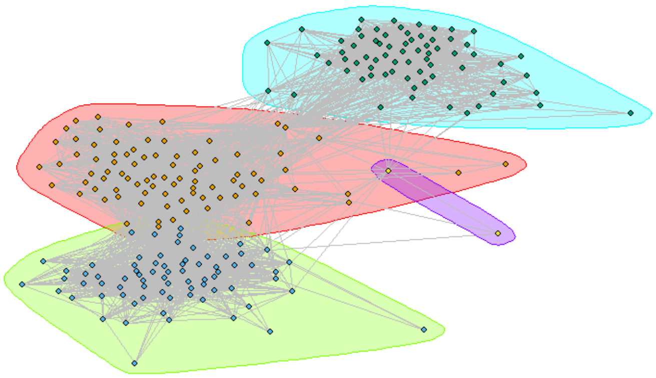
fast_greedy for Node 484



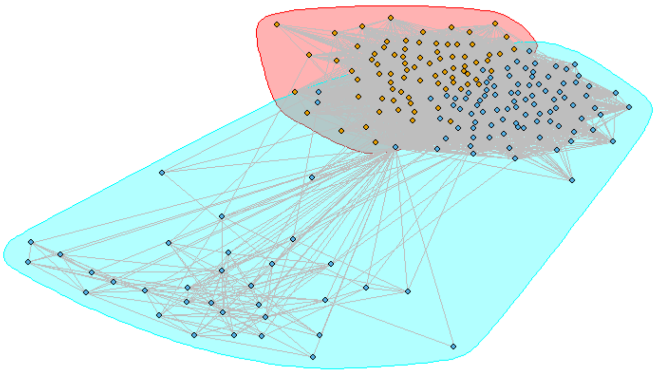
edge_betweenness for Node 484



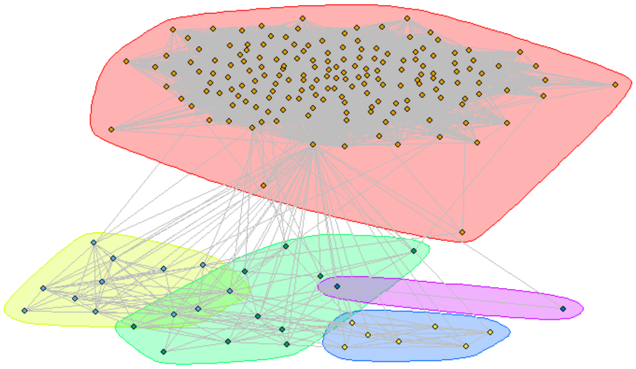
infomap for Node 484

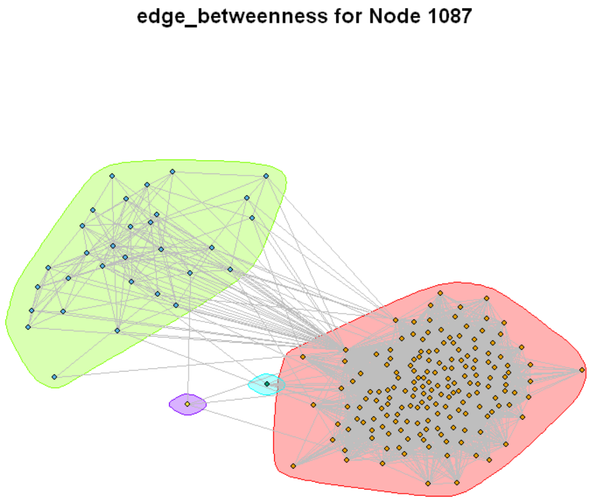


fast_greedy for Node 1087



infomap for Node 1087

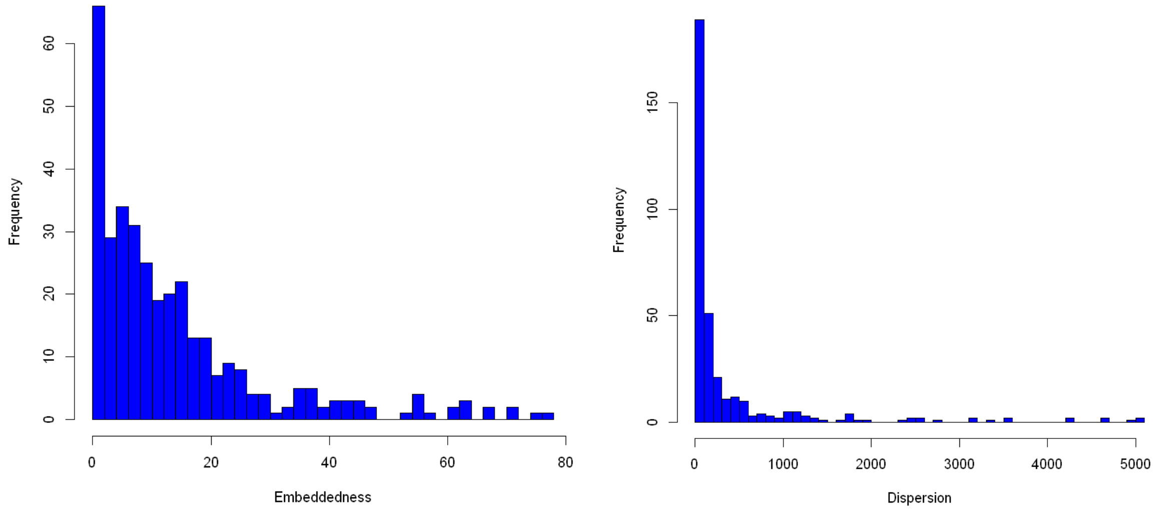




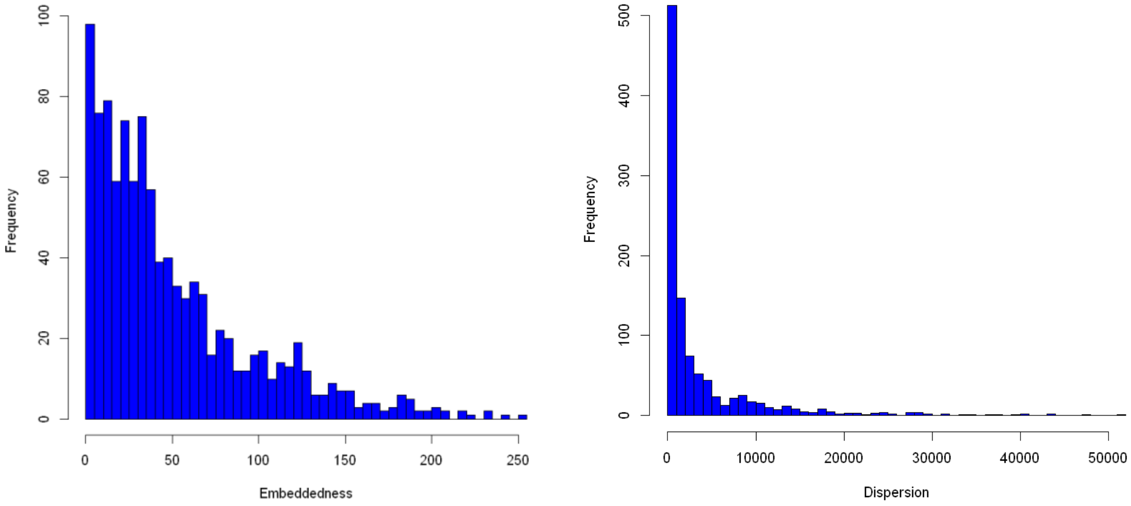
edge_betweenness for Node 1087

Answer for Q11: Embeddedness of a node is the number of mutual friends a node shares with the core node. Since in personalized network, core node is connected with every other node, the friend of selected node i must be the mutual friend of core node except itself since they are already connected. Therefore, for node i, Embeddedness(i) = Deg(i)-1

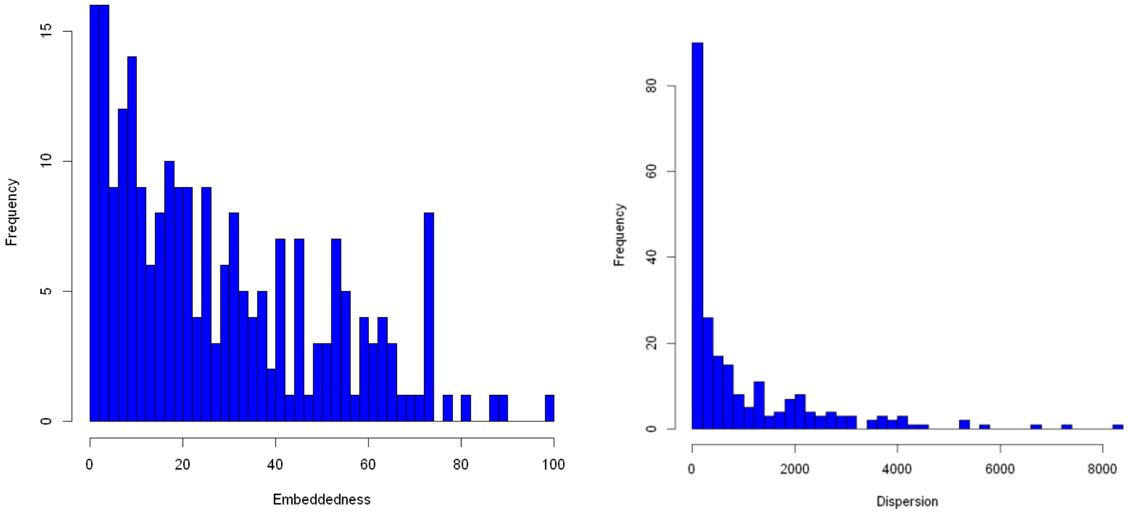
Answer for Q12:



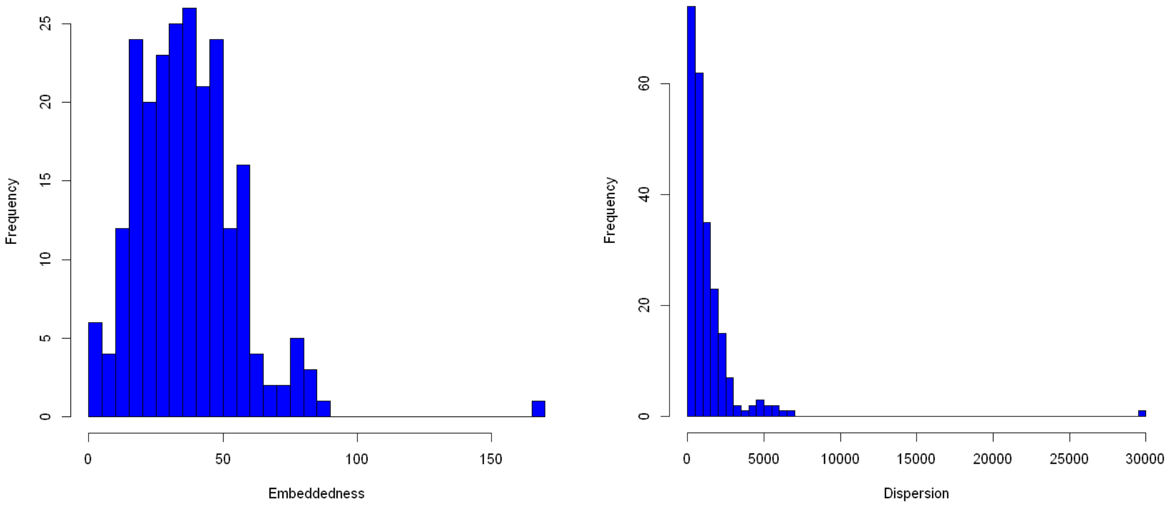
Node ID = 1



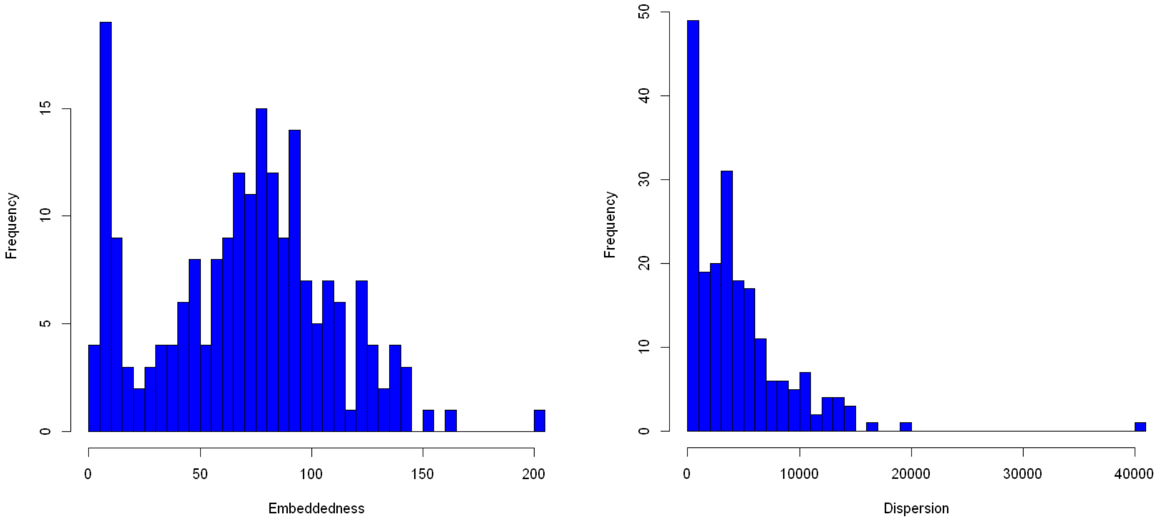
Node ID = 108



Node ID = 349



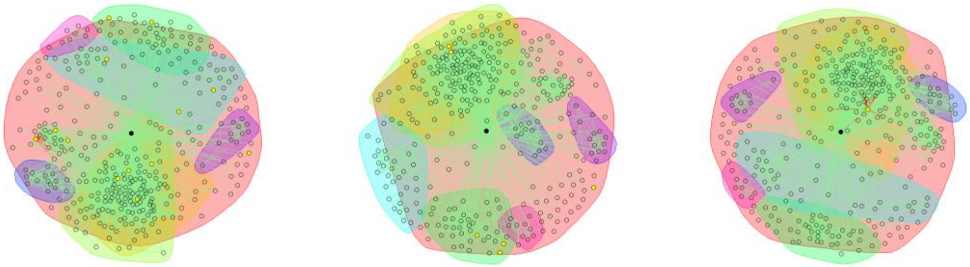
Node ID = 484



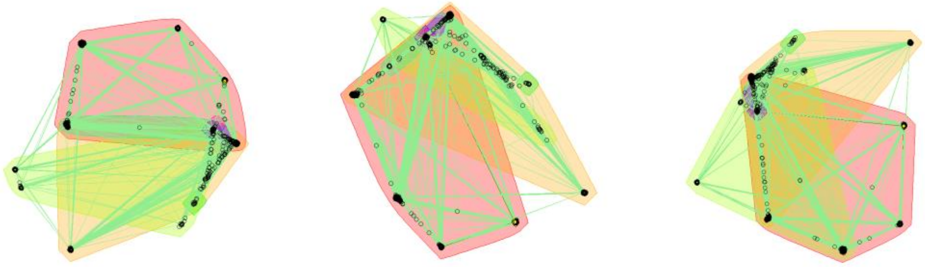
Node ID = 1087

Answer for Q13 & 14:

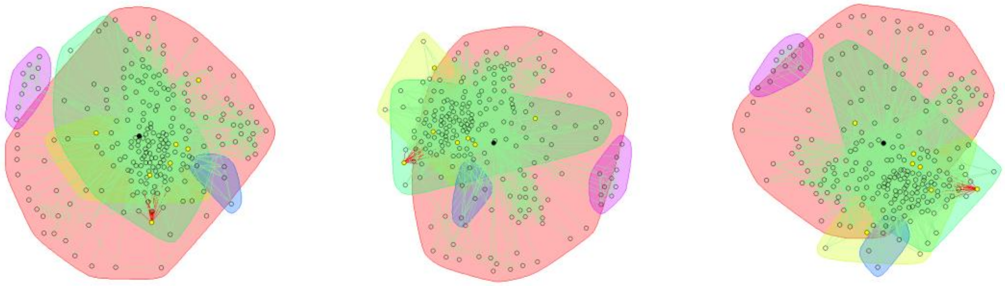
Max_Dispersion Max_Embeddedness Max_Dispersion/Embeddedness



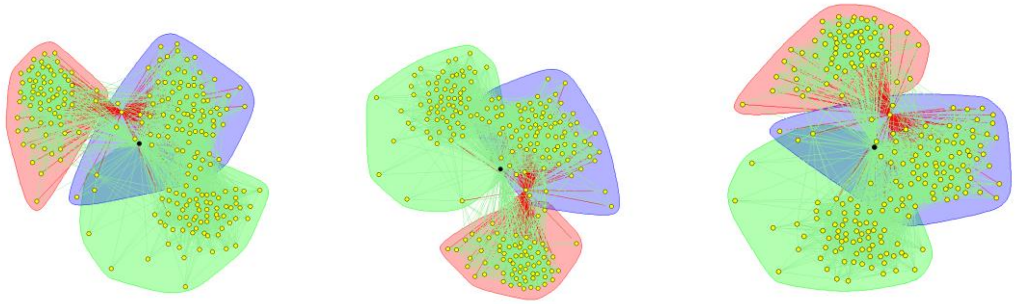
Node ID = 1



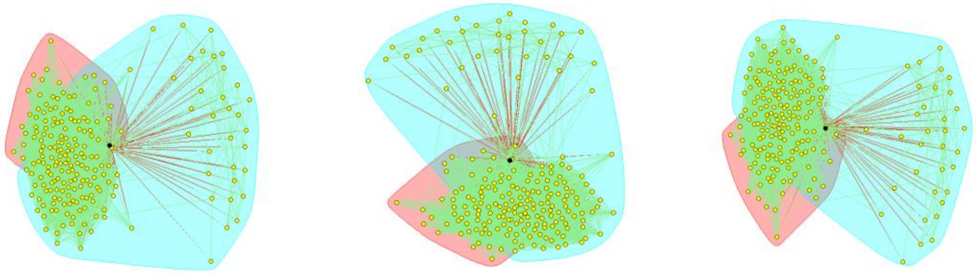
Node ID = 108



Node ID = 349



Node ID = 484



Node ID = 1087

Answer for Q15: Embeddedness is the number of mutual friends a node shares with the core node. If a node has large embeddedness with core node, it means this node has an intersection with core nodes with large probability. Dispersion is the sum of distances between every pair of the mutual friends the node shares with the core node. A low dispersion mean those two nodes are in the same group probably. Just like two workmates in an office, other workmates also know each other in this group. In other words, a high dispersion means those two nodes, or two person, may not be familiar with each other. Obviously, a node with high embeddedness, usually also has large dispersion. So for k=Dispersion/Embeddedness, is like doing a normalization process. A small k means high relativity between two nodes. Otherwise, a large k stands for low correlation.

1.4 Friend recommendation in personalized networks

For this part, we take the read graph as a reading edge list. In this case, we avoid the change of nodes number problem when automatically reconstructing of graph.

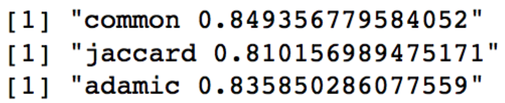
Answer for Q16:

We pick out all ID 415 node’s neighbours and in this personalized network, filter others to keep those nodes with degree 24. The number for the total list is 11.

Answer for Q17:

In this problem, we take three measures:
Common Neighbors measure : length of intersection Si and Sj
Jaccard measure: length of intersection Si and Sj over length of union Si and Sj
Adamic Adar measure: sum of 1 over log(subset of intersection Si and Sj)

After calculating, the accuracy is shown as following：



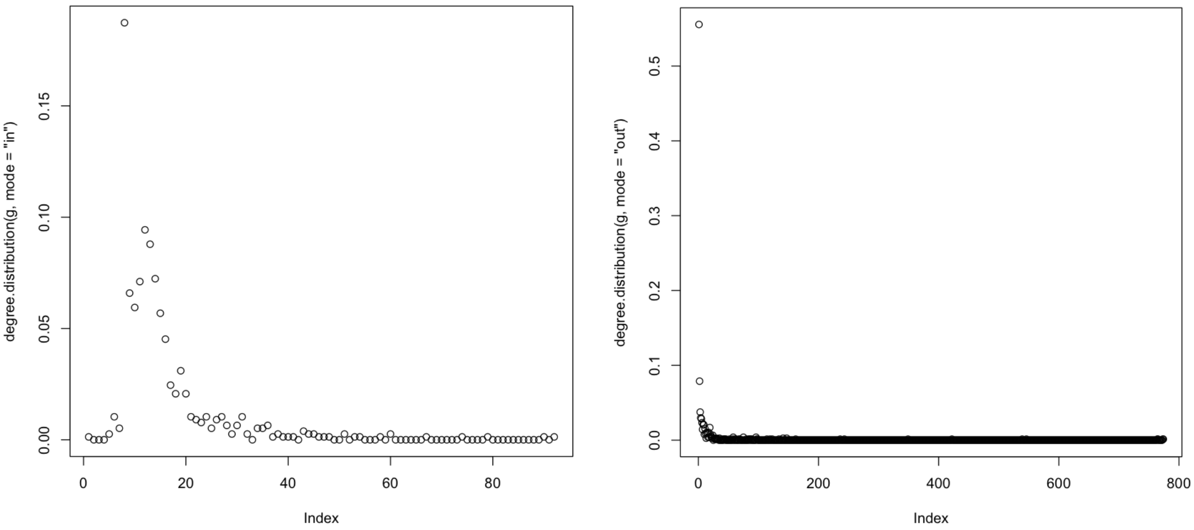
```
[1] "common 0.849356779584052"
[1] "jaccard 0.810156989475171"
[1] "adamic 0.835850286077559"
```

Therefore, from the result we can observe that the first method has the highest accuracy.

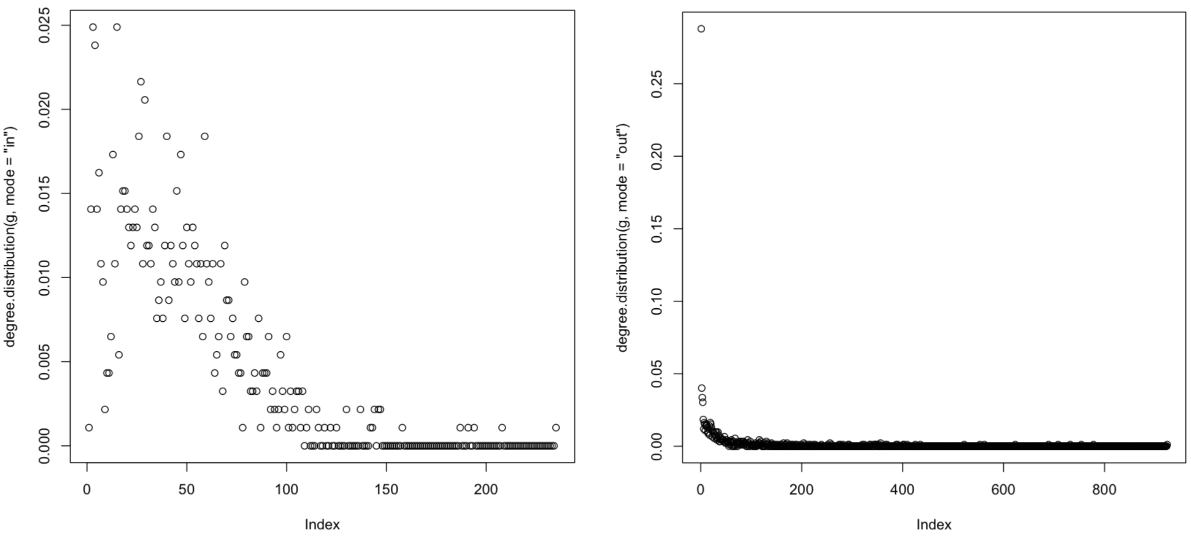
2 Google + network

Answer for Q18: We checked files of circle and calculated the number of rows in each personal network circles file. Finally, we got 57 networks having more than 2 circles.

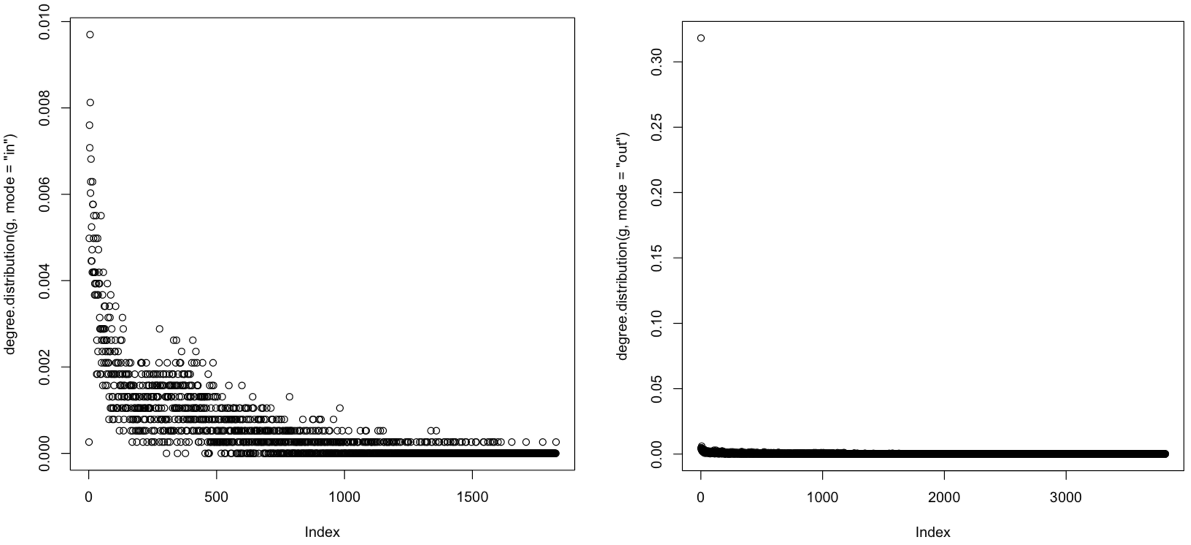
Answer for Q19: We rebuild the personal networks for the selected node with the provided files. First, we create the graph from the “.edges” file. Because the “.edges” file records all the connection of the neighbouring nodes except for the connection between core node between its neighbors, we add the edges between all the existed nodes with the core nodes.
The in-degree and out-degree distribution of the personal network for 109327480479767108490 is shown as below:



The in-degree and out-degree distribution of the personal network for 115625564993990145546 is shown as below:



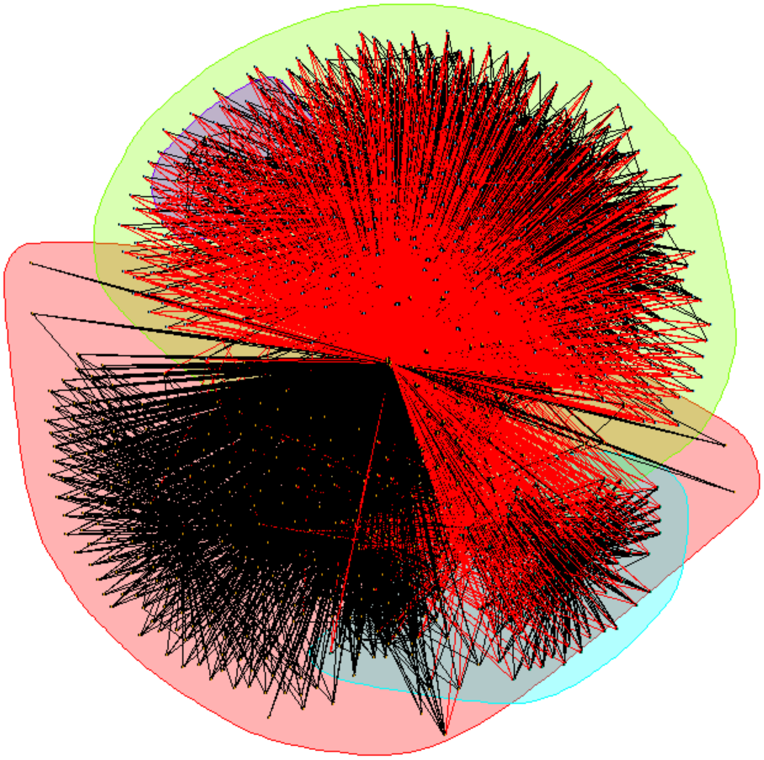
The in-degree and out-degree distribution of the personal network for 101373961279443806744 is shown as below:



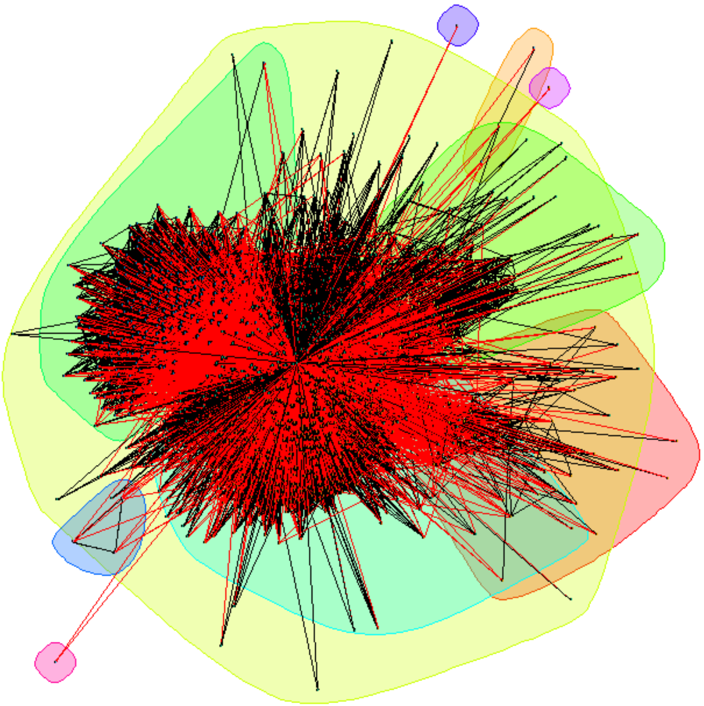
From above diagrams, all in-degree distributions of three nodes have similar trend and all out-degree distributions of three nodes are similar with a large portion of nodes having zero degree and others distributed similarly.

Answer for Q20: We extracted the community structure of each personal network using walktrap.community and got their modularity scores and community structure plots.

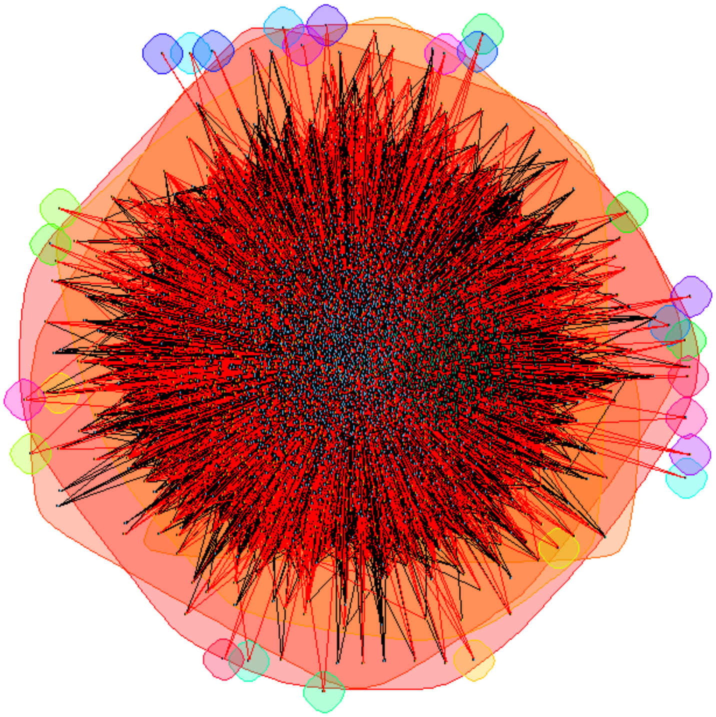
The modularity score of network ‘109327480479767108490’ is 0.25276535939251 and its community structure is shown as below:



The modularity score of network ‘1156255649939901455466’ is 0.319472554647349
and its community structure is shown as below:



The modularity score of network ‘101373961279443806744’ is 0.191090282684037
 and its community structure is shown as below:



All modularity scores of three networks are similar and in a range from 0.1 to 0.4, which indicating that all three personal networks have similar community structure even with different number of nodes inside networks.

Answer for Q21:

From those expressions for h and c, homogeneity represents the purity of circles inside each community, the higher the homogeneity is, the less number of circles information are included in each community, the most ideal situation is that each community just include nodes from the same one circle.

Completeness represents the completeness of circle information in each community. If all nodes of each circle are included in the same one community, the completeness is highest. If every community only includes a small part of nodes from each circle, then the completeness is low.

Answer for Q22:

By calculating the results of expression, we got $H(C)$, $H(V)$, $H(C|K)$, $H(K|C)$results of three networks.

For personal network of node ‘109327480479767108490’:

$H(C)\ = 0.4563477$

$H(V)$ = 0.4365564

$H(C|K)$ = 0.06759188

$H(K|C)$= 0.2925478

Homogeneity h = 0.8518851

Completeness c = 0.3298739

The homogeneity of network ‘109327480479767108490’ is 0.851, which is relatively high, completeness is 0.329, which is relatively low. The results h and c represent each community of network includes limited number of circles’ information but information of each circle inside each community is not completed.

For personal network of node '115625564993990145546':

$H(C)\ = 3.676366$

$H(V)$ = 0.4695553

$H(C|K)$ = 2.015052

$H(K|C)$= 2.077295

Homogeneity h = 0.4518903

Completeness c = -3.423962

The homogeneity of network ‘115625564993990145546’ is 0.451, which is relatively low. The result h represents each community of network includes a relatively large number of circles’ information inside. The completeness is a negative number which is not normal. This may because there are many communities in this network, which only have one node inside resulting too small number of $H(K)$, which caused completeness to be negative.

For personal network of node '101373961279443806744':

$H(C)\ = 0.166908$

$H(V)$ = 0.2142508

$H(C|K)$ = 0.1662627

$H(K|C)$= 0.536535

Homogeneity h = 0.003866707

Completeness c = -1.504238

The homogeneity of network ‘101373961279443806744'’ is 0.0038, which is relatively low. The result h represents some communities of network include a relatively large number of circles’ information inside. The completeness is a negative number which is not normal. This may because there are many communities in this network, which only have one node inside resulting too small number of $H(K)$, which caused completeness to be negative.