Large Scale Social Network : Models and Algorithms
Project #2

Due on May 1st, 2020 at 11:59pm

Professor Roychowdhury, Vwani

Wang, Yin, He

Question 1

According to the properties of Facebook, the graph created should be a undirected graph where each vertex represents a single user and the edge between two nodes denotes the friendship of two individuals. Since this, we should carefully set the directed parameter as false when using *read_graph*. Thanks to the undirected attribute of our graph, we do not need to worry about the different definition of connectivity and giant connected component(GCC). Simply using functions *gorder*, *gsize* and *is_connected*, we can get the result as follows: for the graph we study, we denote it as $G = (E, V, W_E, W_V)$, where $|V| = 4039$, $|E| = 88234$. Moreover, the graph itself is connected.

Question 2

Since the graph is connected, we just need to check the diameter of the graph rather than the giant connected component. According to the result of the funciton *diameter*, the diameter of our Facebook network is 8. The results listed in Question 1 and here are both consistent with the statistics shown in the SNAP website.

Question 3

The degree distribution of the Facebook network can refer to Figure 1. Notice that the curve plotted shares high resemblance with the power-law distribution, making it reasonable for us to plot the degree distribution in log-log scale in the following questions. The average degree is easy to derive after getting the degree distribution. We just need to use degrees times the their corresponding frequency and finally sum them up. The average degree obtained by this method is 43.6910.

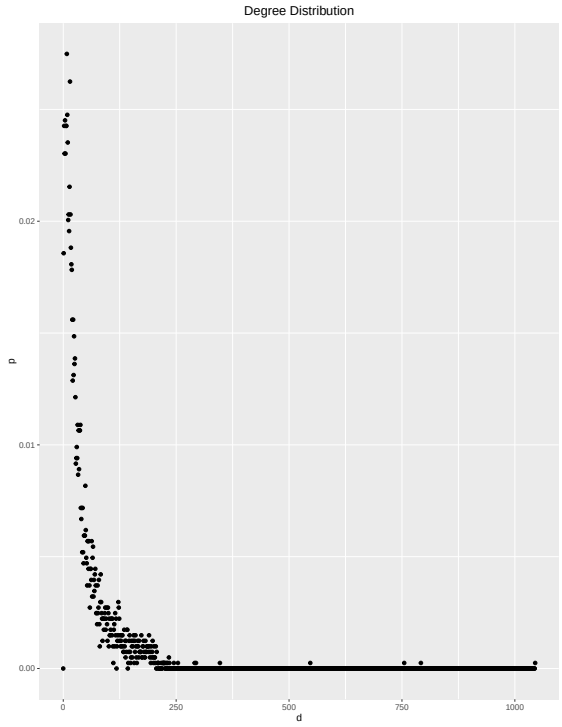

Figure 1: Degree distribution of the Facebook network.

Question 4

Similar in project 1, we are going to plot the degree distribution in log-log scale and try to derive the parameter γ for power-law distribution by linear regression. In order to avoid the disturbance of degrees with 0 frequency, we will delete them first before estimating and plotting. The result is shown as Figure 2. According to the annotation of the figure, the slope of the line is -1.180.

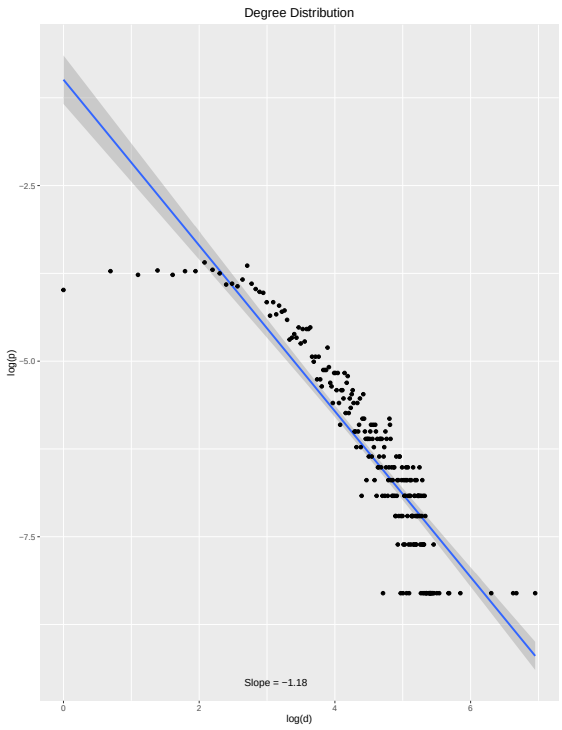

Figure 2: Degree distribution of the Facebook network in log-log scale.

Question 5

Using the function provided in *igraph* package, *make_ego_graph*, we can easily get the personalized graph for any specific node in the Facebook network. There are 348 nodes and 2866 edges in the personalized network for node 1. To get a more clear impression of the personalized network, we plot it as Figure 3.

Question 6

The diameter of this personalized network is 2, which is a very trivial result actually. Notice that for any pair of vertices in this network, they are both directly connected with the center node. Formulating formally, let us denote the personalized network as G_C which is generated according to the node V_C and its neighbors. According to the definition, for any vertex V_i in this network, there always exists edge $(V_i, V_C) \in E_C$. Hence, for any pair of nodes (V_i, V_j), there always exists a path $[(V_i, V_C), (V_C, V_j)]$ that connects these two nodes. In other words, the distance of every two nodes must be less than or equal to 2. Hence, the diameter for the personalized network is always less than or equal to 2 as well. The lower bound of the diameter can be obtained by noticing that it must be a positive integer. In all, $1 \leq diameter(G_C) \leq 2$

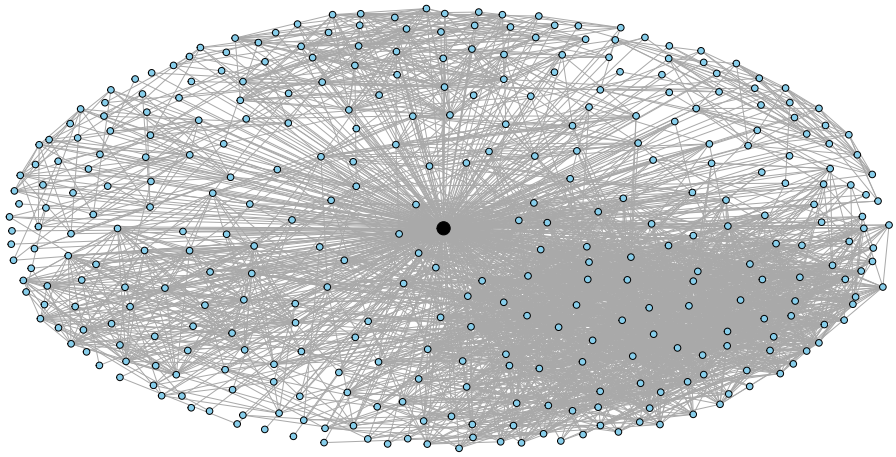

Figure 3: Personalized network of node 1 in Facebook network.

Question 7

If the diameter of the personalized network achieves the lower bound derived in Question 6, it means that there exists a edge between any pair of nodes in the personalized network. In other words, such personalized network is a complete graph. Since there are only two possibilities for the diameter of personalized graph, if the graph itself is not a complete graph, then according to the deduction above, it will achieve the upper bound of the diameter.

Question 8

In the undirected graph, the number of neighbors for a node can be considered as the degree of it. Hence, to figure out the number of the core code in Facebook network, we just need to iterate the whole graph to find the nodes with degree larger than 200. In all, we find out 40 nodes having more than 200 neighbors and their average degree is 279.375.

Question 9

According to the question, we are going to compare the three different clustering methods based on the personalized network generated according to 5 different core nodes. The clustering results are shown as figures where the color of each vertex denotes the membership of different communities. Specifically, Figure 4 represents the clustering results of three different methods for the personalized network of Node 1 while Figure 5 illustrates the results for the personalized network of Node 108; Separated communities according to three different methods for the personalized network of node 389 are shown as Figure 6; Clustering results for the personalized network generated according to node 484 are shown as Figure 7; For the memberships of vertices in the personalized network of node 1087, one may refer to Figure 8.

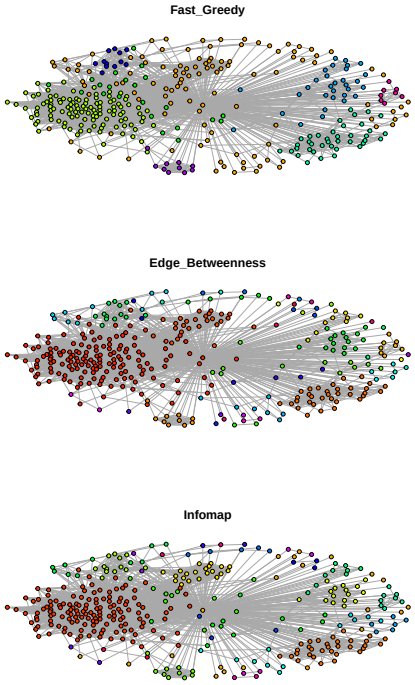

Figure 4: Clustering results for the personalized network of node 1 based on different methods.

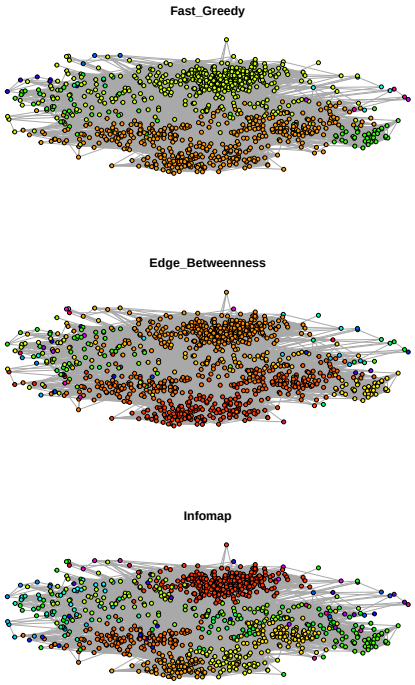

Figure 5: Clustering results for the personalized network of node 108 based on three different methods.

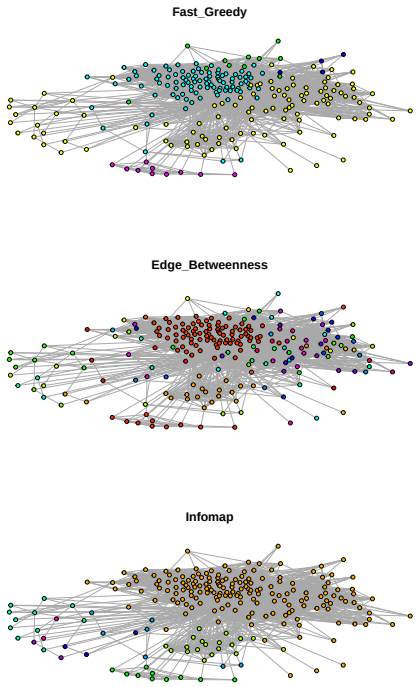

Figure 6: Clustering results for the personalized network of node 349 based on different methods.

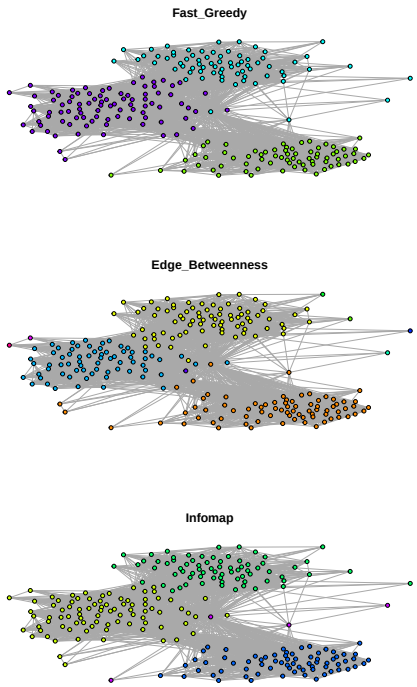

Figure 7: Clustering results for the personalized network of node 484 based on different methods.

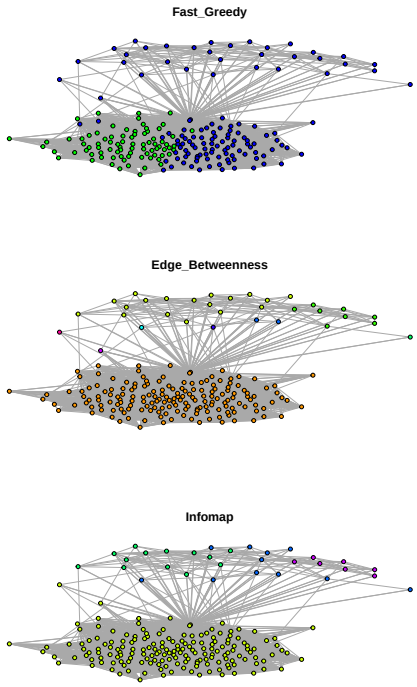

Figure 8: Clustering results for the personalized network of node 1087 based on different methods.

Question 10

Similarly, we can derive the clustering results for the networks removed core nodes and show their corresponding results as figures by inheriting the denotation utilized in Question 9. Specifically, the results for the node 1, 108, 389, 484 and 1087 can be found as Figure 9, 10, 11, 12, 13 respectively. Notice that in Figure 9, we can clearly see that the personalized network removed the core node is no-longer connected, which is a possible situation when paths between two nodes have to pass the core node. The modularity for the original networks and the modified networks based on different clustering methods are shown as Table 1. According to the table, the modularity increases for every clustering method no matter which personalized network is considered after the core node is removed. This can be explained by the definition of modularity. In project 1, we claim that modularity reflects the concentration of edges within modules compared with random distribution of links between all nodes regardless of modules. For the core node, it always have edges connection it with nodes within other communities no matter which community it belongs to, which impairs the modularity of the personalized network. Hence, if excluding the core node, we can improve the modularity of our network.

Question 11

Notice that the embeddedness of a node for a specific core node is defined as the number of friends sharing with the core node. Moreover, the degree of a non-core node in the personalized network of the core node can be considered as two parts. One part is the degree generated by the mutual friends. The other part is the degree caused by the connection of the non-core node and the core node. Hence, the relationship between the degree of a non-core node and the embeddedness of it can be written as

$$Embeddedness(V_i) = deg(V_i) - 1 \ \forall V_i \neq V_C \tag{11.1}$$

Question 11 continued on next page...

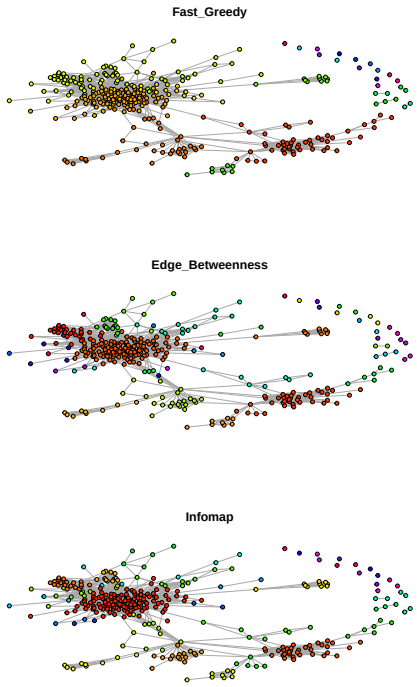

Figure 9: Clustering results for the modified personalized network of node 1 based on different methods.

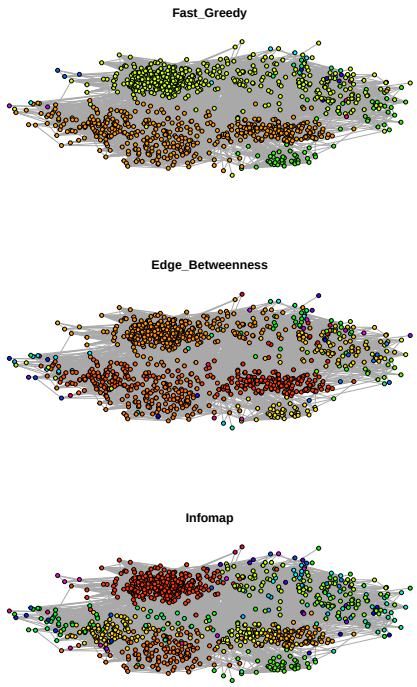

Figure 10: Clustering results for the modified personalized network of node 108 based on different methods.

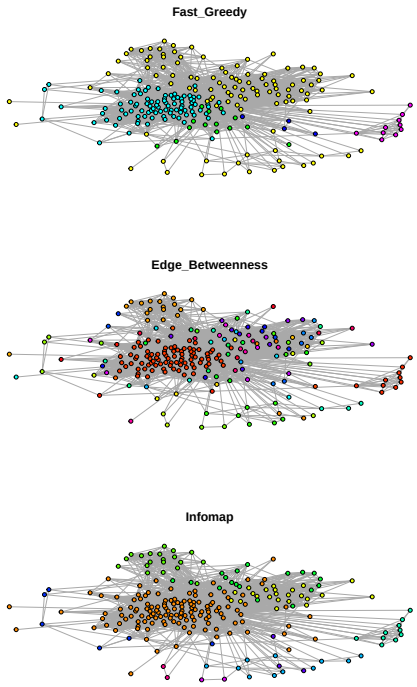

Figure 11: Clustering results for the modified personalized network of node 349 based on different methods.

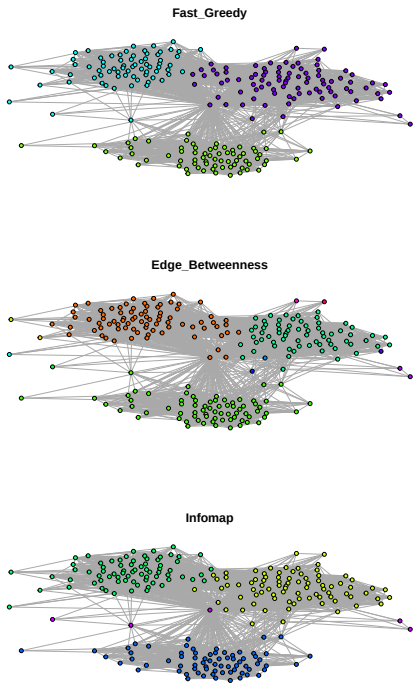

Figure 12: Clustering results for the modified personalized network of node 484 based on different methods.

Question 11 continued on next page...

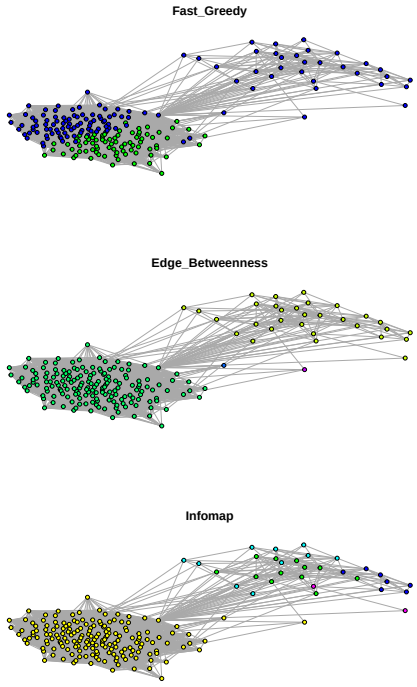

Figure 13: Clustering results for the modified personalized network of node 1087 based on different methods.

Node	Original Network			Network removed core node		
	Fast_Greedy	Edge_Betweenness	Infomap	Fast_Greedy	Edge_Betweenness	Infomap
1	0.4131	0.3533	0.3891	0.4419	0.4161	0.4180
108	0.4359	0.5068	0.5082	0.4360	0.5056	0.5082
349	0.2517	0.1335	0.0960	0.2575	0.1337	0.2037
484	0.5070	0.4891	0.5153	0.5211	0.4956	0.5291
1087	0.1455	0.0276	0.0269	0.1482	0.0325	0.0274

Table 1: Modularity for different clustering methods of the original networks and networks removed the core node.

Question 11 continued on next page...

where V_C represents the core node by whom the personalized network is generated.

Question 12

In this question, we are going to plot the distribution histogram of embeddedness and dispersion for each of the core node's personalized network. As the same in Question 9, the core nodes are 1, 108, 349, 484, and 1087. The resulting plots are shown as figures below. Figure 14, 16,18,20,22 illustrate the distribution of embeddedness of nodes in personalized networks for core node 1, 108, 349, 484 and 1087 while Figure 15,17,19,21,23 demonstrate the distribution of dispersion for these networks. From the plots, we can see

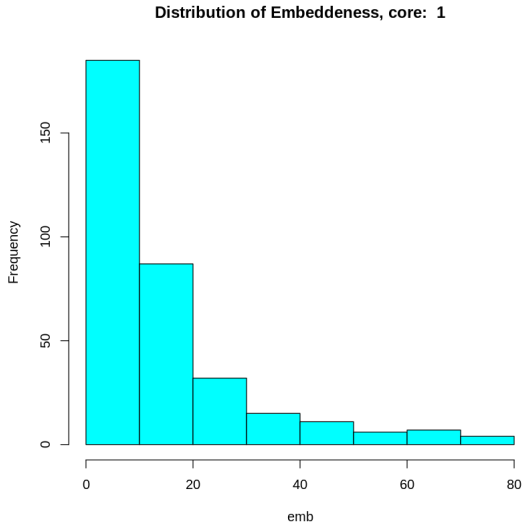

Figure 14: Histogram of embeddedness for the personalized network generated by core node 1

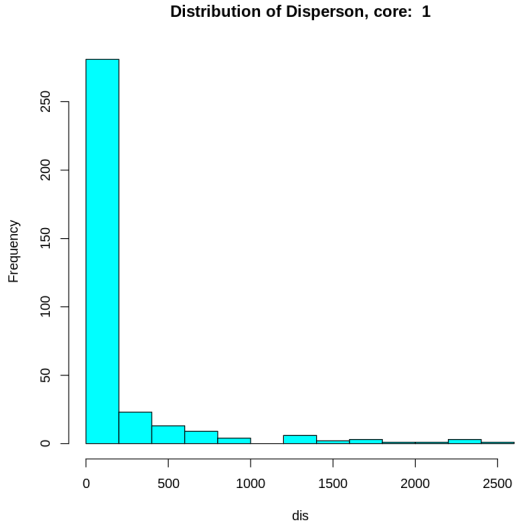

Figure 15: Histogram of dispersion for the personalized network generated by core node 1

that in general, the distribution of embeddedness and the distribution of dispersion follow the same trend,

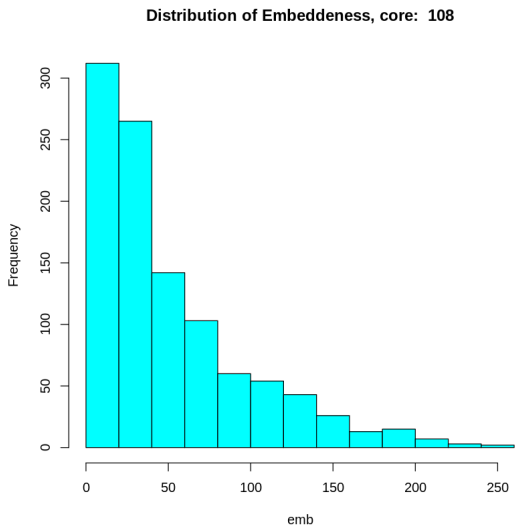

Figure 16: Histogram of embeddedness for the personalized network generated by core node 108

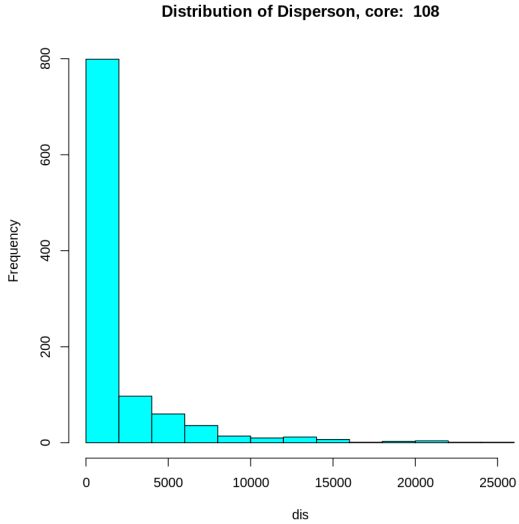

Figure 17: Histogram of dispersion for the personalized network generated by core node 108

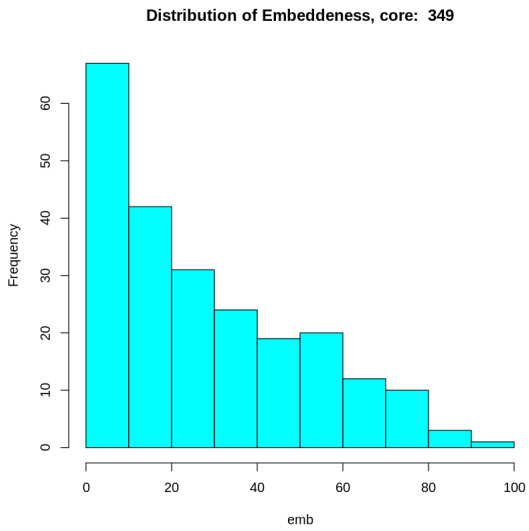

Figure 18: Histogram of embeddedness for the personalized network generated by core node 349

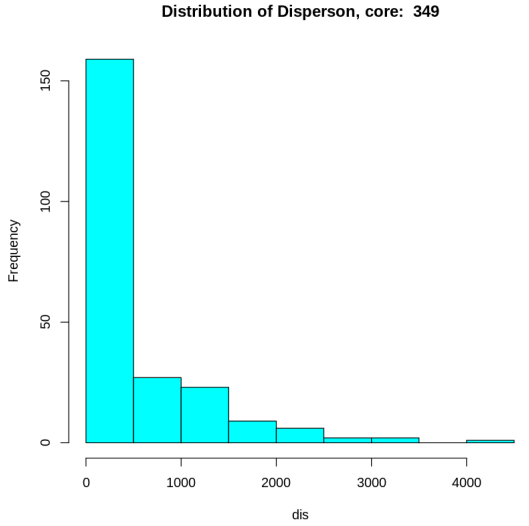

Figure 19: Histogram of dispersion for the personalized network generated by core node 349

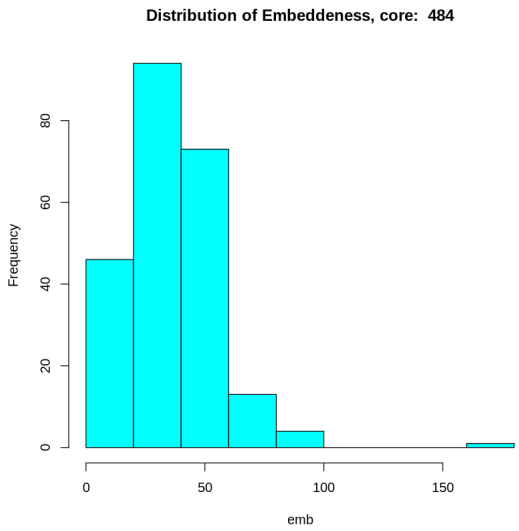

Figure 20: Histogram of embeddedness for the personalized network generated by core node 484

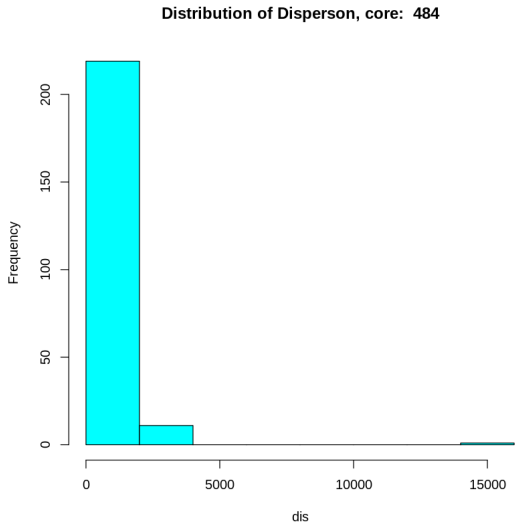

Figure 21: Histogram of dispersion for the personalized network generated by core node 484

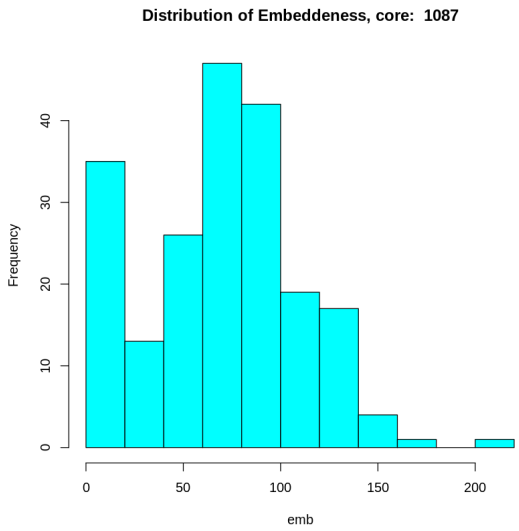

Figure 22: Histogram of embeddedness for the personalized network generated by core node 1087

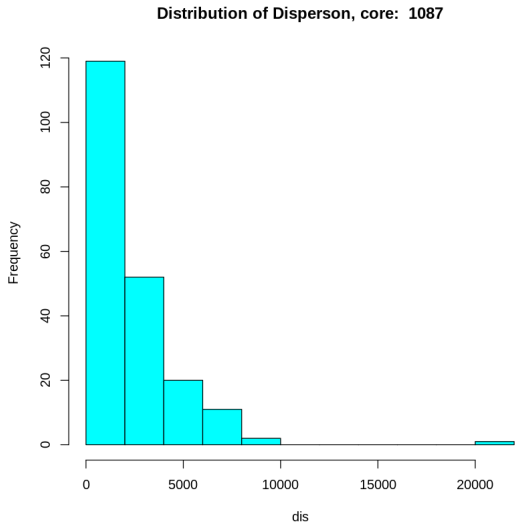

Figure 23: Histogram of dispersion for the personalized network generated by core node 1087

except for code node 1087. And we also can see that the histograms are all right skewed, meaning that most of the nodes are having smaller embeddedness and dispersion numbers.

Question 13

In this part, we are going to plot the community structure for each of the core node's personalized network then identify and highlight the node with maximum dispersion. In particular, to detect the community structure, we will use the Fast-Greedy algorithm. The nodes having maximum dispersion for personalized networks of different core nodes are shown as Table 2. For the visualization of the community structure for different personalized networks, one can refer to Figure 24, 25, 26, 27, 28.

	Core Node 1	Core Node 108	Core Node 349	Core Node 484	Core Node 1087
Max dispersion	57	1889	377	108	108

Table 2: The nodes having maximum dispersion in different personalized network generated according to core node.

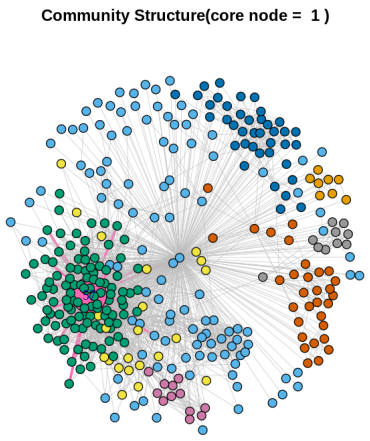

Figure 24: Community structure of the personalized network for core node 1 with node having maximum dispersion highlighted.

Question 14

Again, in this part, we are going to plot the community structure for each of the core node's personalized network, however this time we will highlight the node with maximum embeddedness and the node with maximum dispersion/embeddedness portion. To be clear, the node with maximum embeddedness is colored in cyan, the node with maximum dispersion/embeddedness portion is colored in pink, and the node with both is colored in green. The founded nodes which have maximum dispersion/embeddedness ratio, maximum embeddedness respectively for different core node are shown as Table 3.
We can directly see that besides core node 1, in all other cases maximum embeddedness and maximum portion appears to be the same node in each personalized network, which are shown straightforward in the

Question 14 continued on next page...

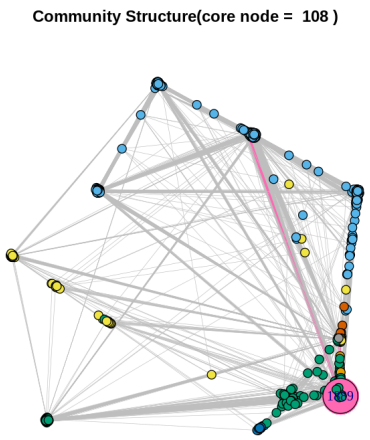

Figure 25: Community structure of the personalized network for core node 108 with node having maximum dispersion highlighted.

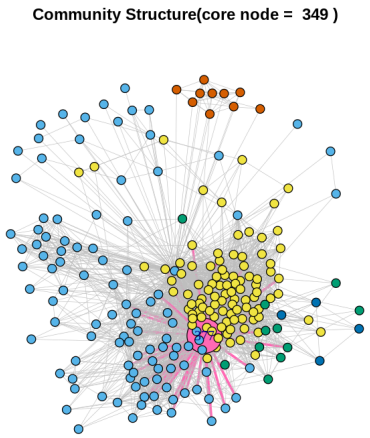

Figure 26: Community structure of the personalized network for core node 349 with node having maximum dispersion highlighted.

	Core Node 1	Core Node 108	Core Node 349	Core Node 484	Core Node 1087
Max embeddedness	57	1889	377	108	108
Max $\frac{\text{dispersion}}{\text{embeddedness}}$	323	1889	377	108	108

Table 3: Nodes with maximum characteristic in different personalized networks generated according to core nodes.

Question 14 continued on next page...

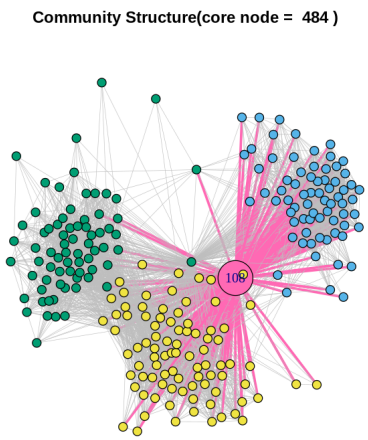

Figure 27: Community structure of the personalized network for core node 484 with node having maximum dispersion highlighted.

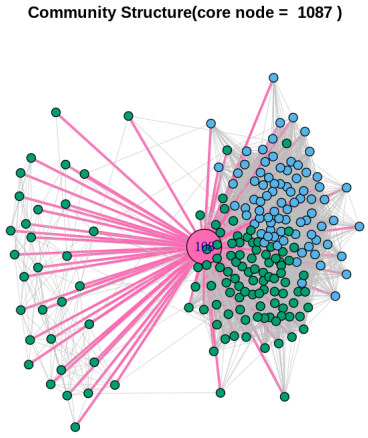

Figure 28: Community structure of the personalized network for core node 1087 with node having maximum dispersion highlighted.

resulting figures. Figure 29, 30, 31, 32 and 33 represent the results for the personalized network of core node 1, 108, 349, 484 and 1087 respectively.

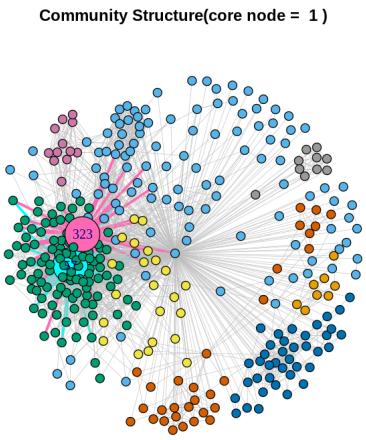

Figure 29: Community structure of the personalized network for core node 1 with nodes having maximum embeddedness and $\frac{\text{dispersion}}{\text{embeddedness}}$ highlighted.

Question 15

Recall the definition of *dispersion* and the definition of *embeddedness*, which is defined as the sum of distances between every pair of the mutual friends the node shares with the core node and is defined as the number of mutual friends a node shares with the core node. Considering embeddedness, we can see that nodes with maximum embeddedness share the same largest community with the core node, and also we can notice that the node is in the center of that particular community, which makes sense because the node with maximum embeddedness will share the maximum number of friends with the core node, thus will be placed somewhere near the core node, which is placed in the center of the communities. Considering dispersion, the node and the core node may not in the same community, which is because a farther distance between these two nodes will make the dispersion value larger. The maximum portion implies very large dispersion and very small embeddedness, thus this basically means that this particular node and the core node?s mutual friends are not strongly connected when removing this node and the core node, thus we can say that there must be some strong relationship between these two nodes.

Question 16

In this question, we are going to first create a personalized network for node 415, and then create the list of users who we want to recommend new friends to. To create the list is simple, we just pick all the nodes with degree 24, and we denote this list as N_r, and according to our calculation, the length of this list is 11.

Question 17

In this question, we will compare the performance of three different friend recommendation algorithms. These

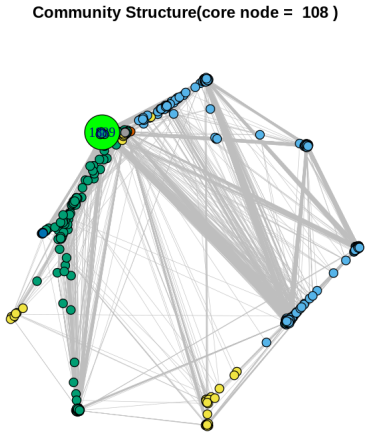

Figure 30: Community structure of the personalized network for core node 108 with nodes having maximum embeddedness and $\frac{\text{dispersion}}{\text{embeddedness}}$ highlighted.

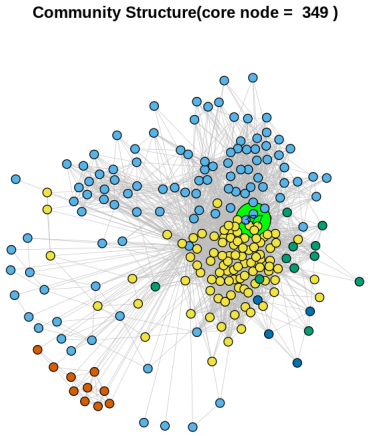

Figure 31: Community structure of the personalized network of core node 349 with nodes having maximum embeddedness and $\frac{\text{dispersion}}{\text{embeddedness}}$ highlighted.

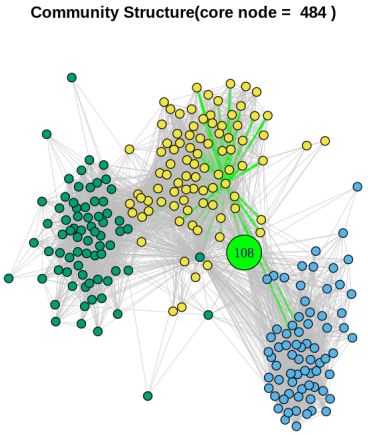

Figure 32: Community structure of the personalized network of core node 484 with nodes having maximum embeddedness and $\frac{\text{dispersion}}{\text{embeddedness}}$ highlighted.

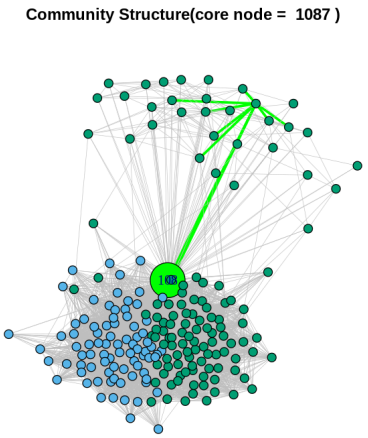

Figure 33: Community structure of the personalized network of core node 1087 with nodes having maximum embeddedness and $\frac{\text{dispersion}}{\text{embeddedness}}$ highlighted.

Question 17 continued on next page...

three algorithms are Common Neighbor measure, Jaccard measure, and Adamic-Adar measure. And we will use the average accuracy to compare the performance. The detailed descriptions of these three algorithms and steps to compute the average accuracy are discussed in the project handout, thus we will not discuss them in this report. The measurements are shown in the following table as Table 4.

Algorithm Name	Common Neighbor	Jaccard	Adamic-Adar
Average Accuracy	0.84644	0.83270	0.85279

Table 4: The Performance of three different friend recommendation algorithms.

Overall, we can see that there is no big difference between these three algorithms in terms of average accuracy, and we can see that the Adamic-Adar method has the highest average accuracy, thus in this case, we say that Adamic-Adar is the best recommendation algorithm.

Question 18

In all 792 data sets, there are 132 personal networks, and 57 of them are more than 2 circles. The detailed distribution is showed as Figure 34.

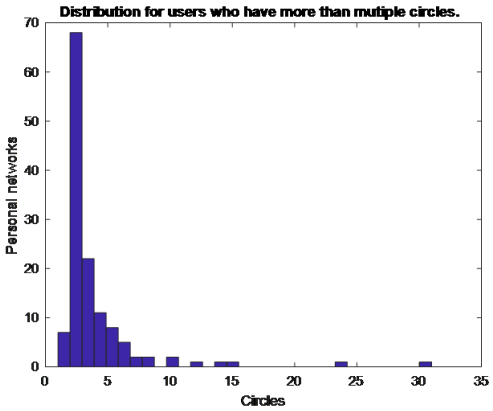

Figure 34: Distribution of personal networks having different number of circles.

Question 19

The in-degree and out degree distributions are show as follow. It?s clear that even they have the similar tendency (the density decreases as the degree increases), but their distributions are not the same. The visualization for these results can refer to Figure 35, 36, 37, 38,39, 40. Specifically, Figure 35 and 36 demonstrates the in-degree and out-degree distribution of the personal network of node 109327480479767108490 respectively; The in-degree and out-degree distribution of the personal network of node 115625564993990145546 are shown as Figure 37 and 38;For the results for the personal network of node 101373961279443806744, one can refer to Figure 39 and 40.

As for in-degree, nodes 109327480479767108490 and node 101373961279443806744 are quite similar. Seems that they have personalized networks, and these personalized networks seem to decrease exponentially as the degree increases. In these two networks, only a few members have the highest degree. Node

Question 19 continued on next page...

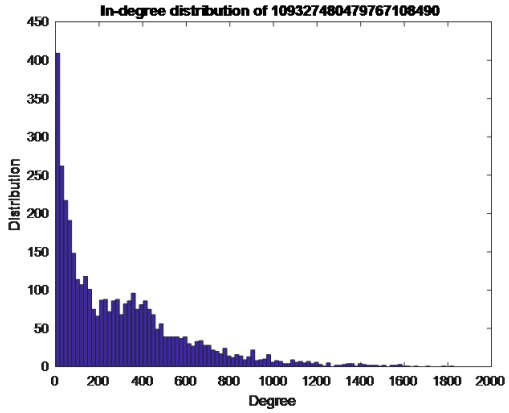

Figure 35: In-degree distribution of the personal network for node 10932748047976710849O.

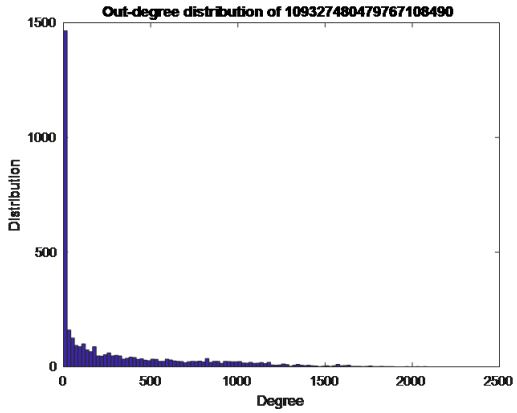

Figure 36: Out-degree distribution of the personal network for node 10932748047976710849O.

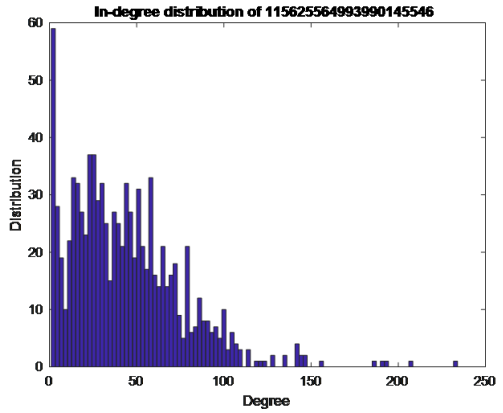

Figure 37: In-degree distribution of the personal network for node 11562556499399014554O.

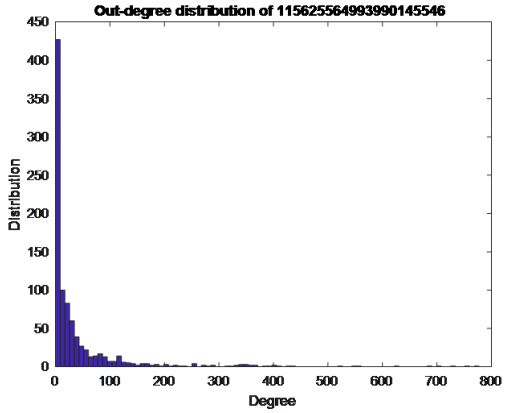

Figure 38: Out-degree distribution of the personal network for node 11562556499399014554 6.

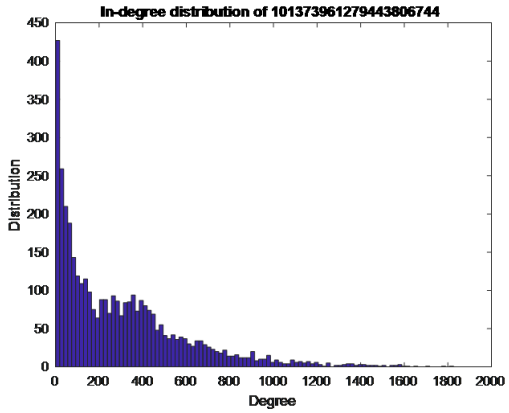

Figure 39: In-degree distribution of the personal network for node 10137396127944380674 4.

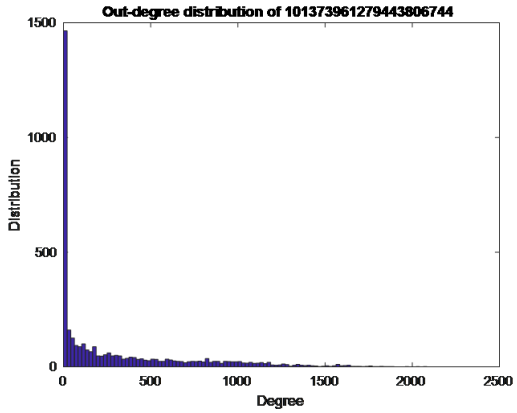

Figure 40: Out-degree distribution of the personal network for node 10137396127944380674 4.

11562556493993990145546's personalized network seems to have more linear attenuation, so compared with the other two networks, a large number of users have higher degrees.

Question 20

Table 5 demonstrates the modularity scores for nodes mentioned above. The communities for these nodes are shown as figures. Specifically, Figure 41 represents the community structure for the personal network of node 109327480479767108490; The community structure detected for the personal network of node 11562556493990145546 is shown as Figure 42; The result of Walktrap algorithm on the personal network of node 101373961279443806744 is plotted as Figure 43. According to the chart, it does seem the community has not generated a high degree of modularity.

Node	109327480479767108490	11562556493990145546	101373961279443806744
Scores	0.2528	0.3195	0.1911

Table 5: Modularity score for the personal network of three different nodes.

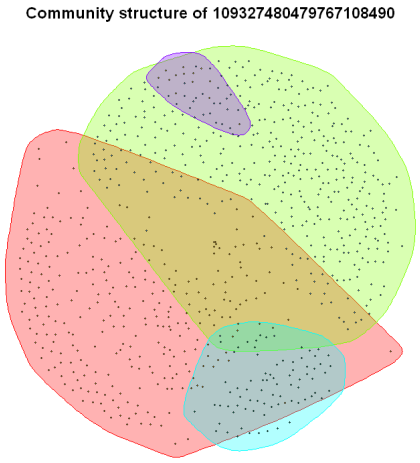

Figure 41: Community structure detected by Walktrap algorithm for the personal network of node 109327480479767108490.

Question 21

Homogeneity and completeness are two criteria for evaluating the clustering results of clustering algorithms. The two often have a certain negative correlation.

Homogeneity means that the data contained in each cluster (cluster result cluster) should belong to a class.

Completeness means that all data belonging to the same class should be grouped into the same cluster.

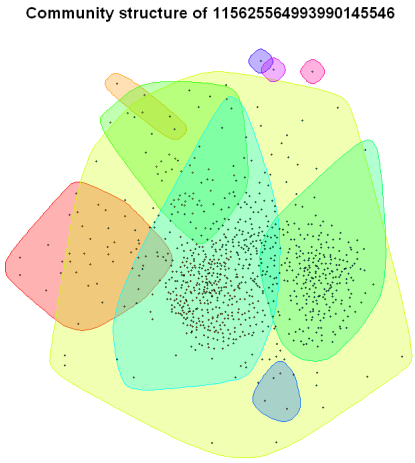

Figure 42: Community structure detected by Walktrap algorithm for the personal network of node 115625564993990145546.

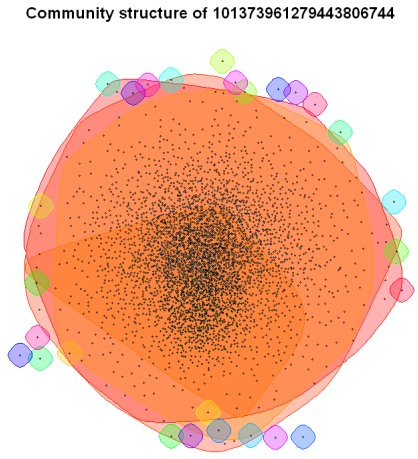

Figure 43: Community structure detected by Walktrap algorithm for the personal network of node 101373961279443806744.

Question 22

The homogeneity and completeness of three node are listed in the table. For a balanced evaluation of the performance, we introduce the V-measure:

$$V = \frac{(1+\beta)hc}{\beta h + c} \tag{22.1}$$

which is a harmonic mean of homogeneity and completeness, and the factor β can be adjusted to favor either the homogeneity or the completeness of the clustering algorithm. The concrete measurements for the node mentioned in Question 19 are shown as Table 6.

Node ID	109327480479767108490	115625564993990145546	101373961279443806744
Homogeneity	0.8936	0.9549	0.7850
Completeness	0.4967	0.6890	0.0398
V-measure	0.6385	0.8005	0.0758

Table 6: Three different measurements for the personal network of three nodes.

In the test, we found that the homogeneity score is very high, especially the second node; while the integrity score is very low, especially the third node, which is almost 0.

The explanation for this is as follows: when the community cluster of node three is visualized, no obvious community structure can be observed. There is a lot of overlap in larger communities. At the same time, since the homogeneity of each community decreases with the amount of conditional entropy. For the node-one network, the homogeneity score is significantly higher than the integrity score. One possible factor is that there are more communities predicted than circles. Another factor may be that the degree distribution is biased towards low-level users, which may harm Walktrap's community discovery capabilities.

Finally, for the second network, the homogeneity and integrity scores are the highest among the three personalized networks. A potential reason is that the degree distribution is hardly biased towards low-degree users and may have fewer "central" attributes. Its V-measure score is also the highest among the three nodes.