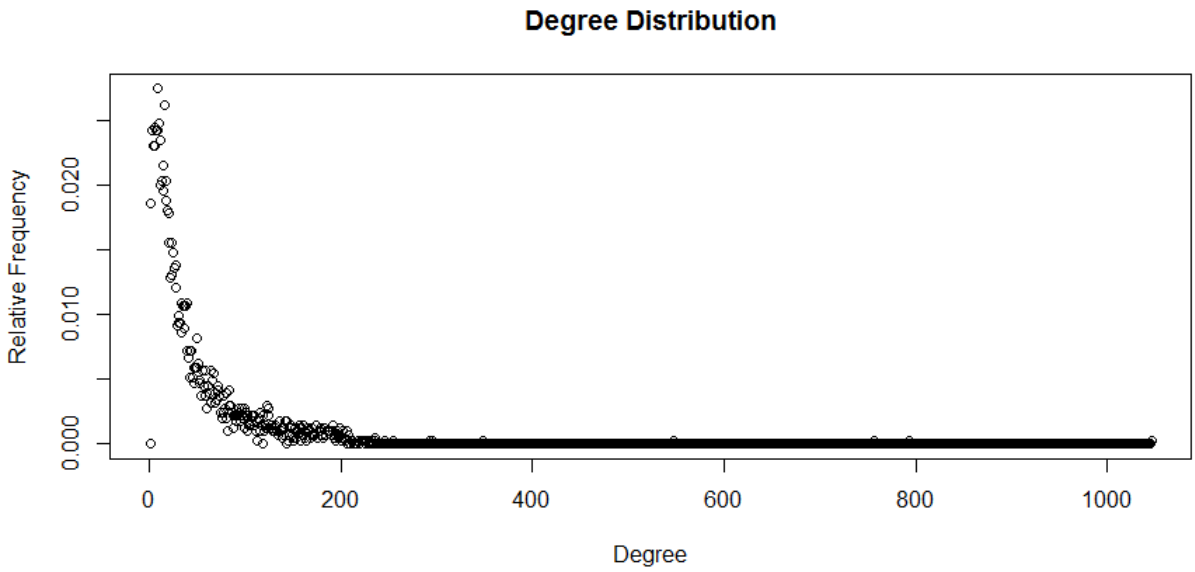
Kevin Wang (604036846)

EE 232E Project 1

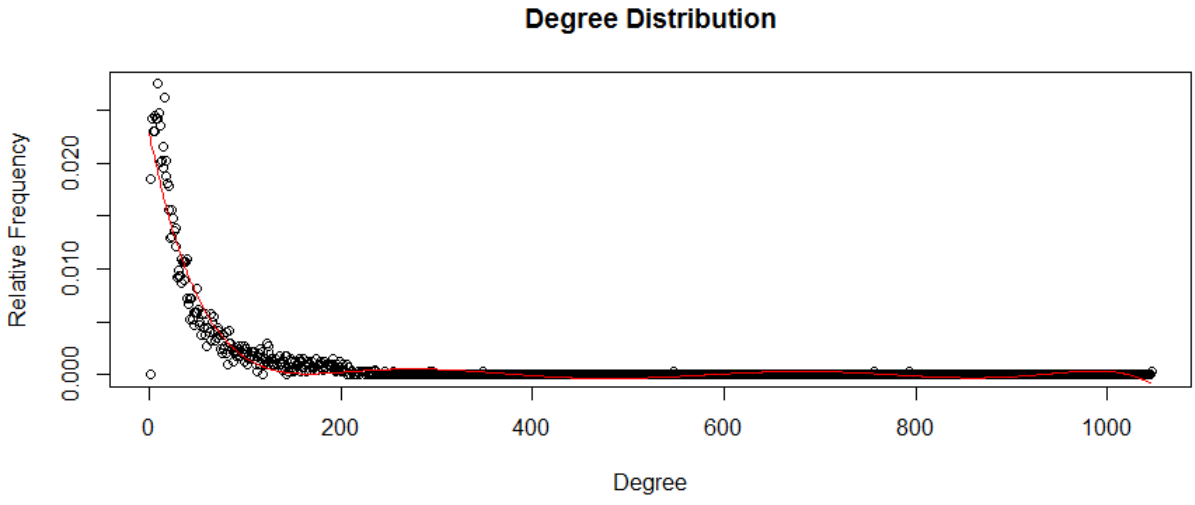
The coding for this assignment was done in R.

Problem 1

The graph/network is connected according to *is.connected*. The diameter of the undirected graph is 8. The average degree is found to be 43.69101.



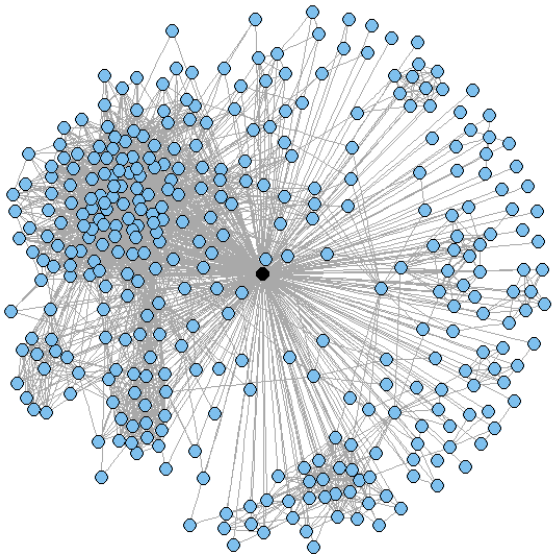
To fit a decent curve, a degree 7 polynomial fit is used:



The total mean square error is 0.01215487, and the actual MSE is 1.162034e-05. The adjusted R^2 value is 0.9002.

Problem 2

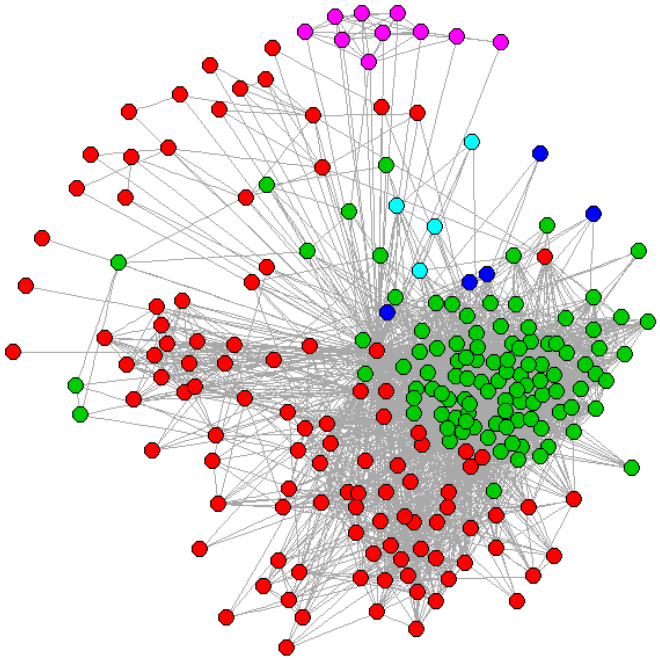
The function *graph.neighborhood* is used to create the personal networks. The personal network of node 1 is found to have 348 nodes and 2866 edges.



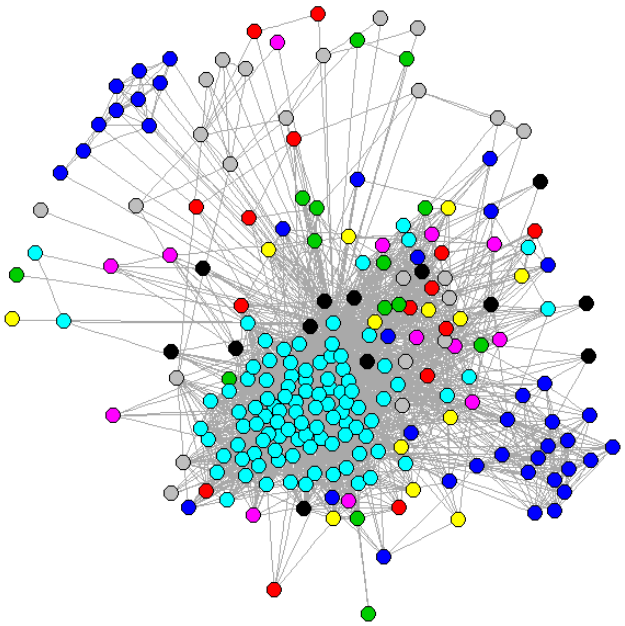
Problem 3

It is found that there are 41 nodes with more than 200 neighbors, so there are 41 core nodes. The average degree of these nodes is 277.439.

Node index 288 was selected for plotting (the actual node name was “348”). There are 230 nodes and 3441 edges.

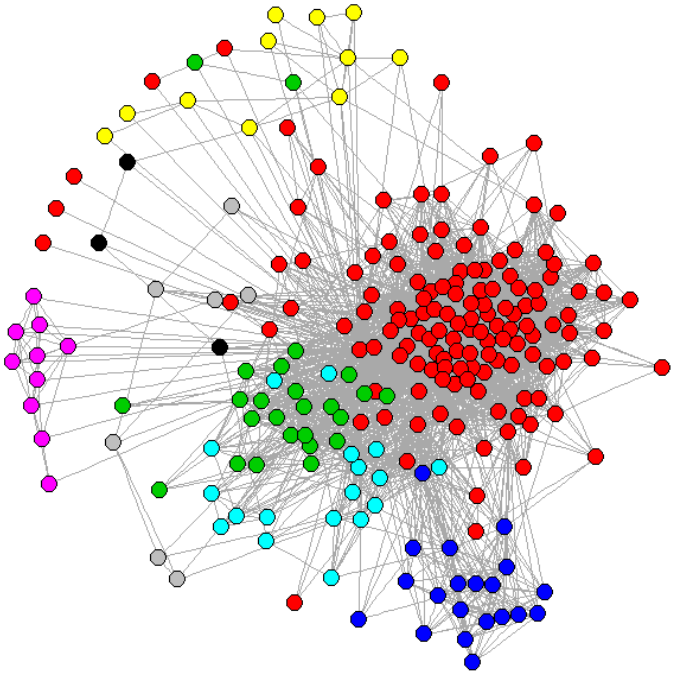


Fast Greedy Method



Edge-Betweenness Method

For this algorithm, there were lots of size 1 communities, and iGraph seems to have run out of unique colors for the nodes.



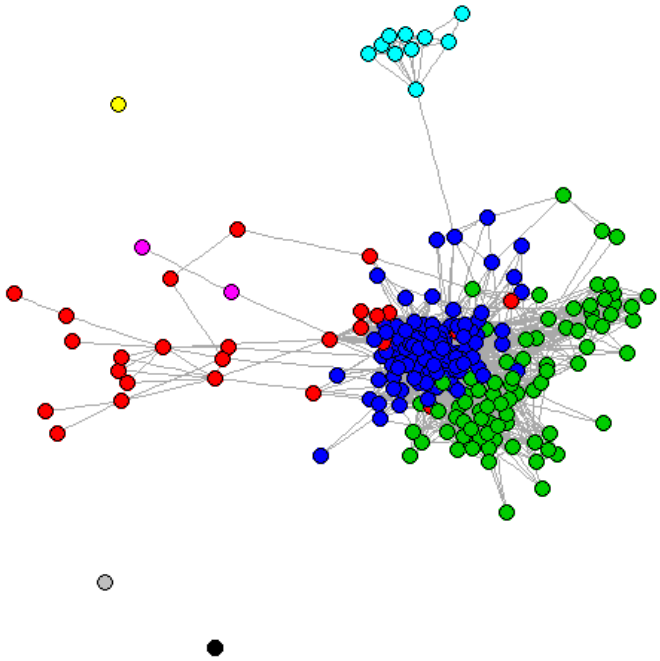
Infomap Method

	Fast Greedy	Edge-Betweenness	Infomap
Modularity	0.2502104	0.133528	0.203753
Number of Communities	5	104	10
Biggest Community Size	107	86	134
Average Community Size	46	2.211538	23

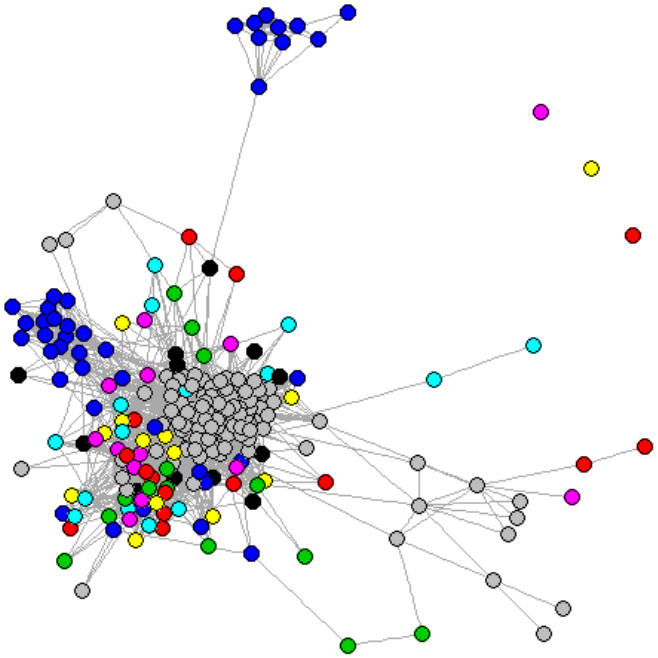
Fast greedy method provided the fewest amount of communities, while edge-betweenness provided lots of size 1 communitites. Thus, for this network at least, the fast greedy and infomap methods seem to be better at producing reasonable community structures.

Problem 4

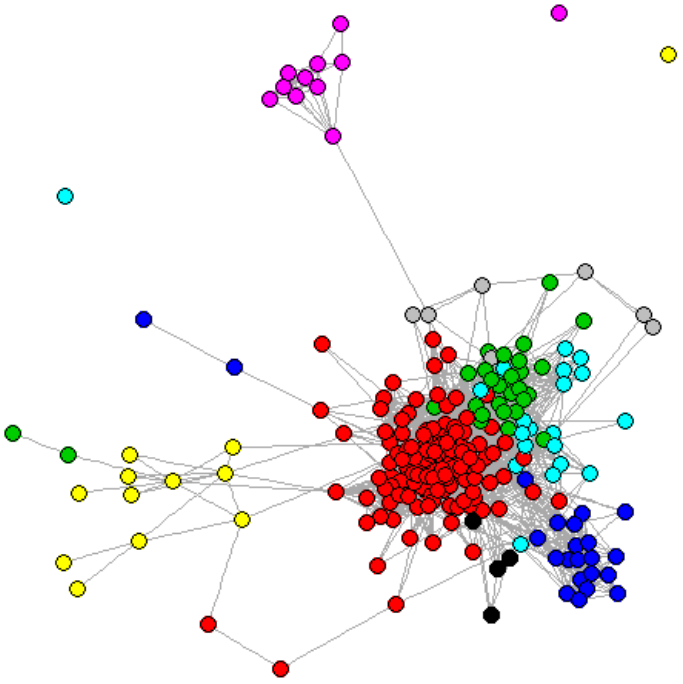
The core node is deleted, and the 3 community finding algorithms are run again to produce the following plots:



Fast Greedy Method



Edge Betweenness Method

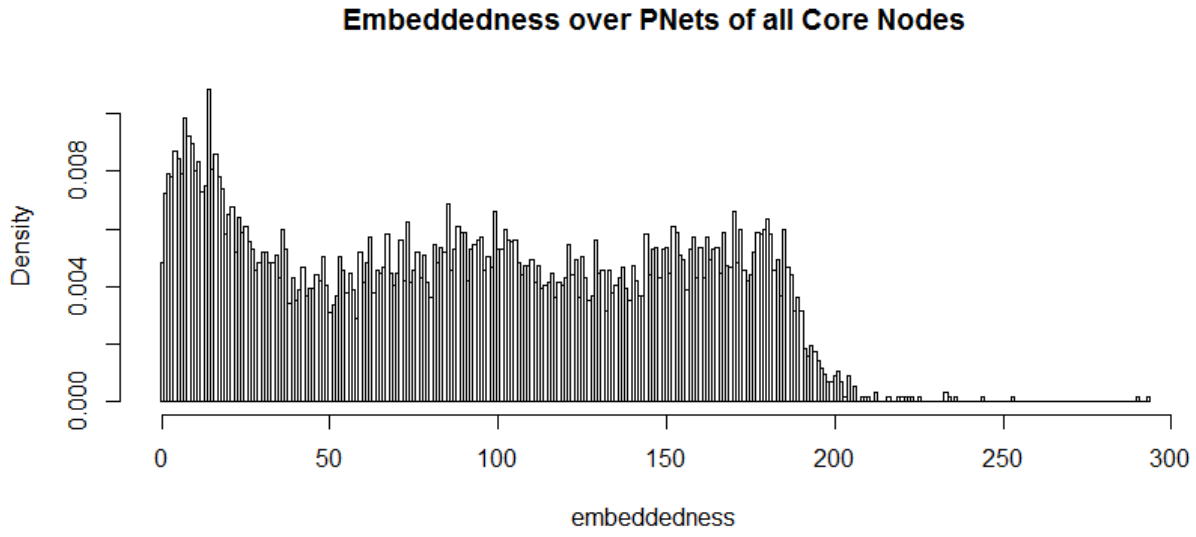


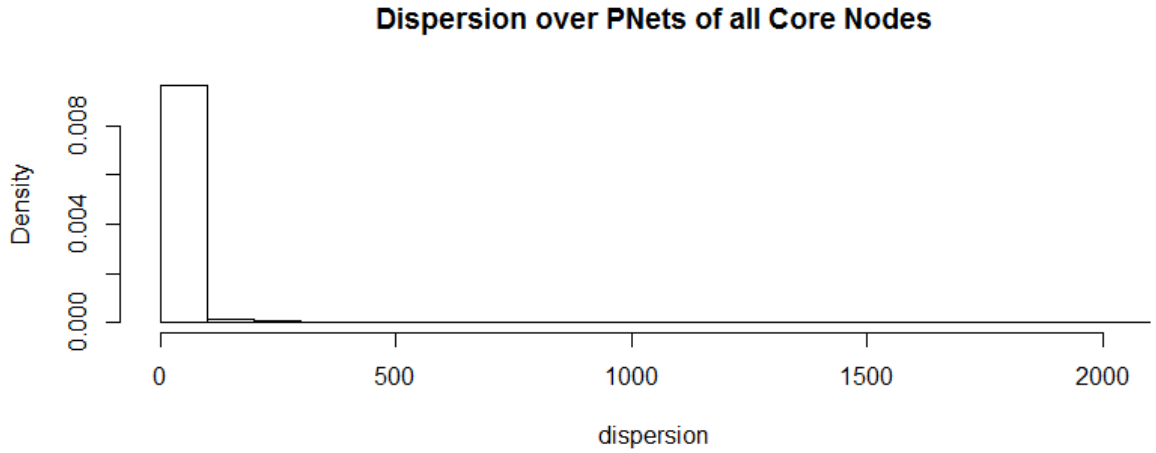
Infomap Method

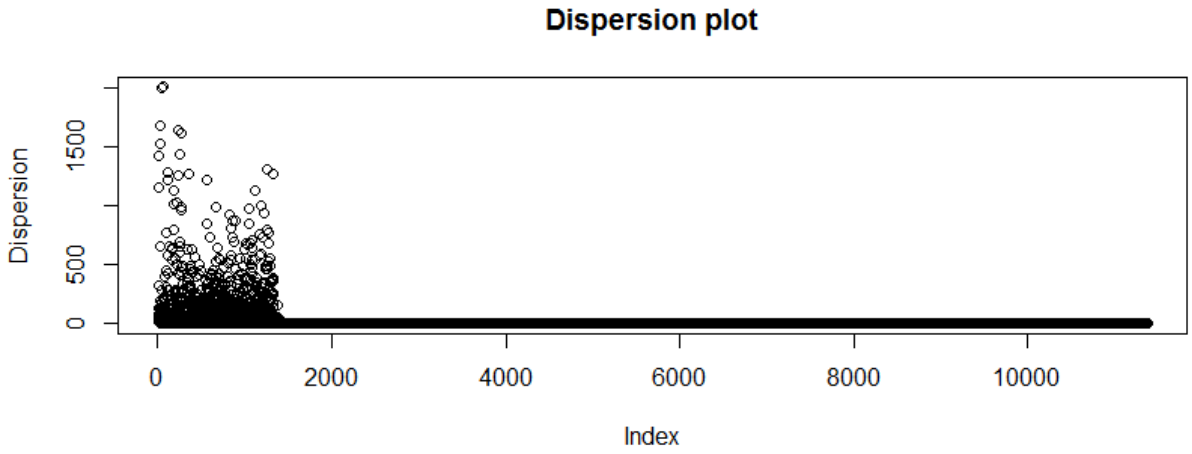
	Fast Greedy	Edge-Betweenness	Infomap
Modularity	0.2456918	0.1505663	0.2337732
Number of Communities	8	103	14
Biggest Community Size	107	85	122
Average Community Size	28.625	2.223301	16.35714

The graph with the core removed was no longer connected. The results show that modularity and biggest community size were relatively unaffected, but the number of communities went up. The new communities were generally small, reducing the average community size. This might be due to the fact that there were now disconnected nodes that were likely in their own communities. Once again, the Edge-Betweenness method produced a lot of communities, much more than there are colors to plot with in igraph.

Problem 5



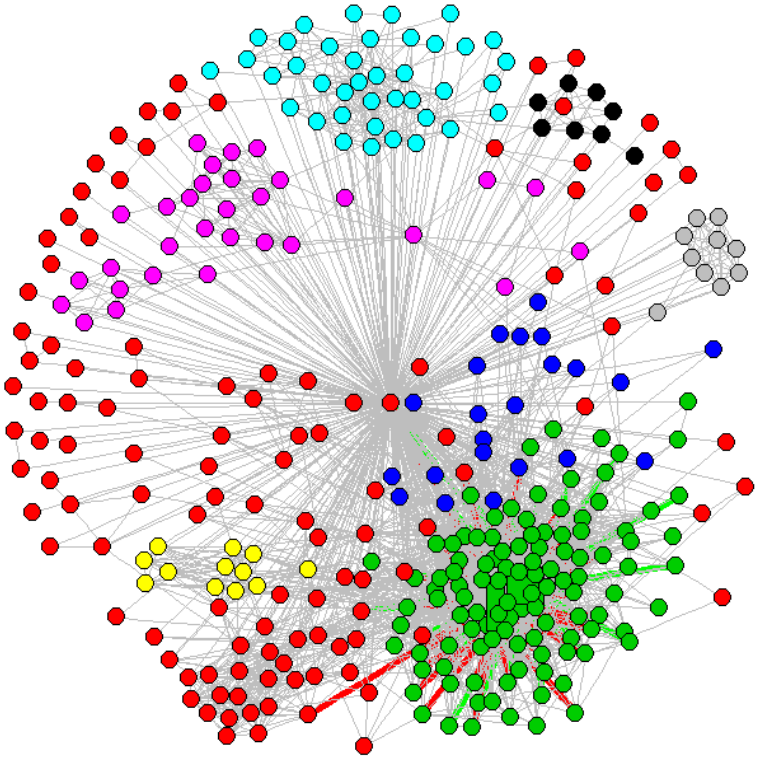




The algorithm used to calculate dispersion had a very long run time, and would often cause R to crash. Thus, the dispersion values for nodes within core node personal networks were calculated for a few personal networks at a time. Apparently, after the first few core nodes, the rest of the core nodes' personal networks all had 0 dispersion. This might be because the core nodes were not hubs but had lots of neighbors, or there were computer memory issues due to the frequency of RStudio encountering fatal errors.

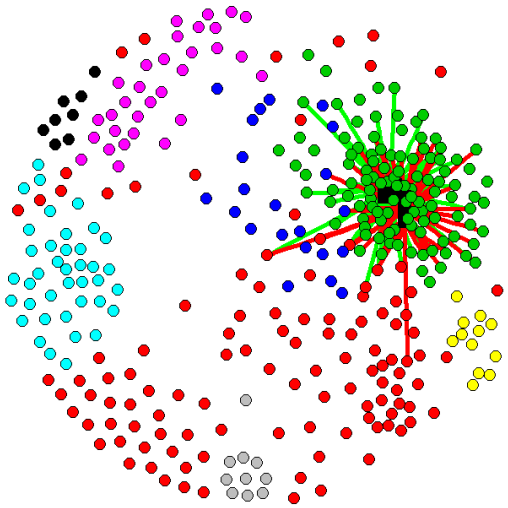
In the following plots, the highest dispersion node is a rectangle with red edges. The highest embeddedness node is a square with green edges. The node with the highest $\frac{dispersion}{embeddedness}$ is a large circle with blue edges.

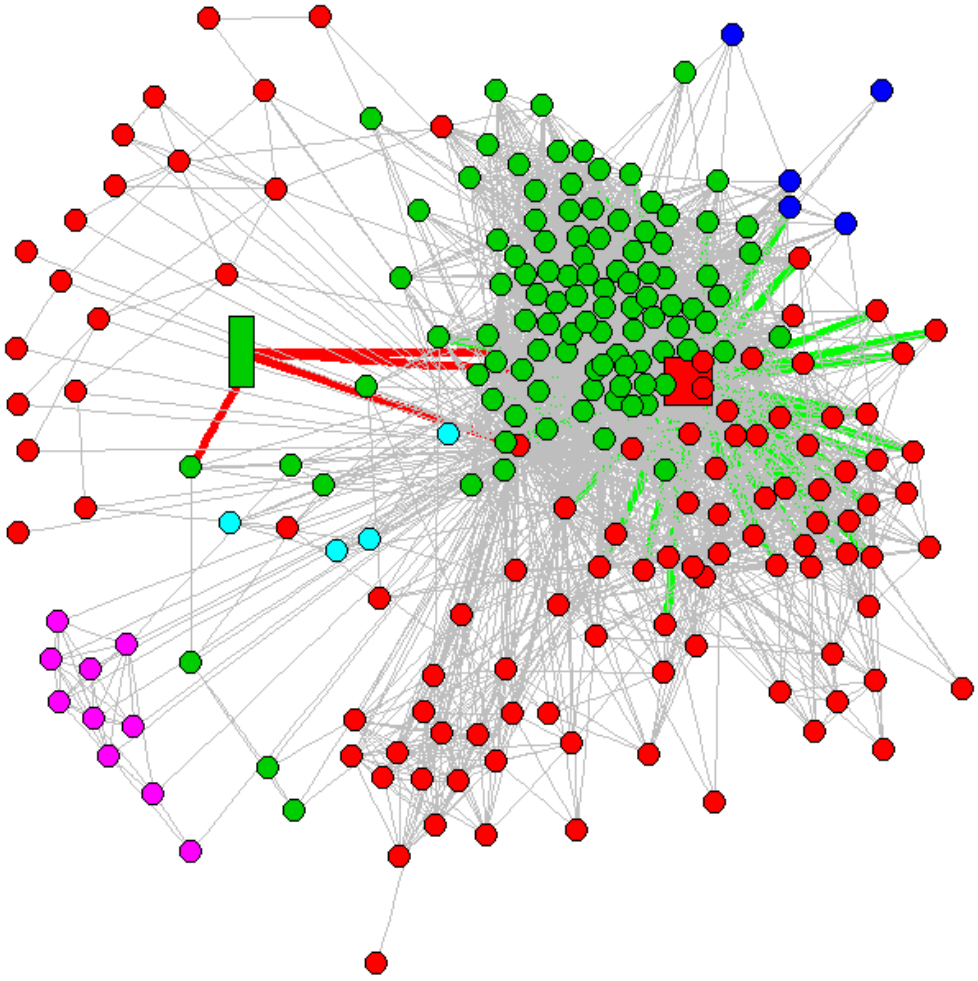
Plots of 3 different personal networks follow. The communities are colored according to the results of the fast greedy community finding method.



Core Node 1

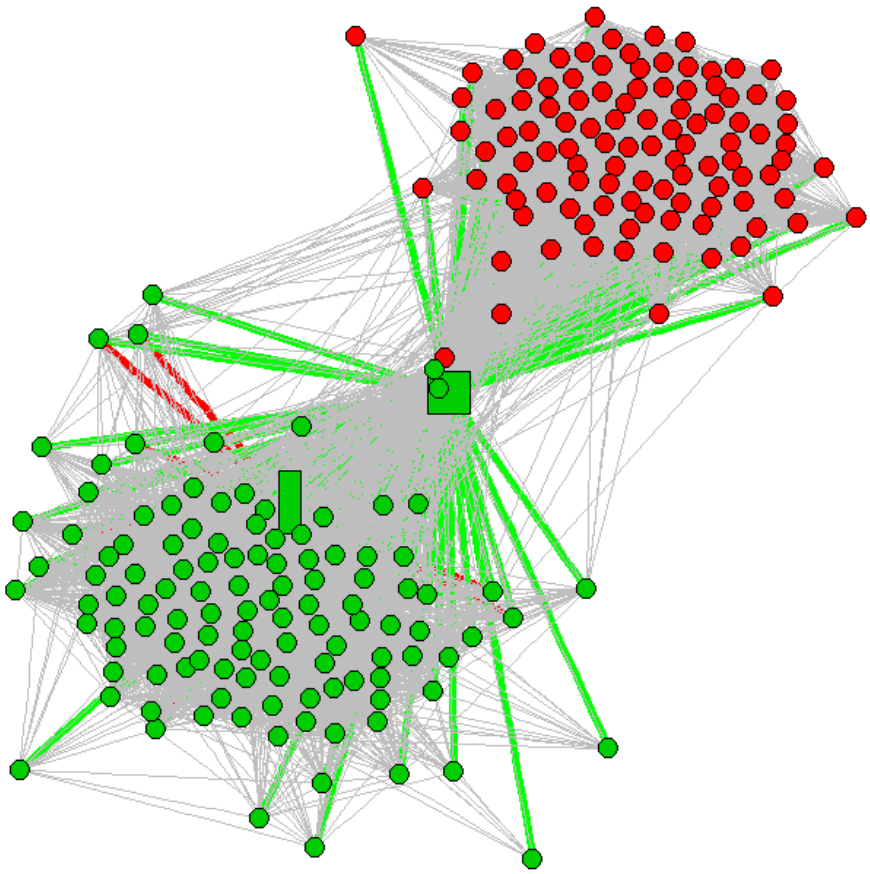
In this network, the node with the highest dispersion and highest $\frac{dispersion}{embeddedness}$ was the same node so the highest $\frac{dispersion}{embeddedness}$ markings were not used. A plot with the 2 max nodes colored black and other non-incident edges removed follows:





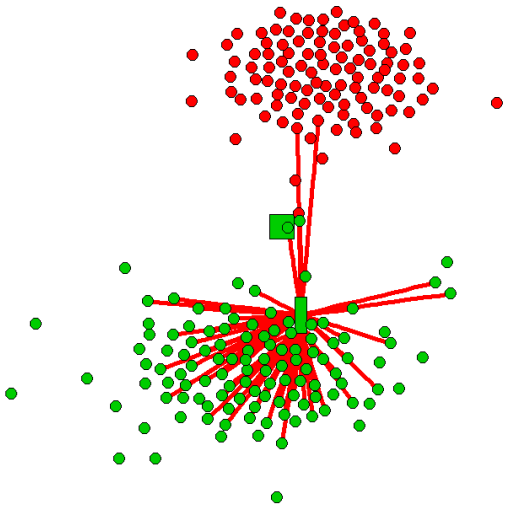
Core node 3

The node with the highest dispersion and highest $\frac{dispersion}{embeddedness}$ was the same node so the highest $\frac{dispersion}{embeddedness}$ markings were not used. It is possible that this is usually the case.



Core Node 20

Once again, the node with the highest dispersion and highest $\frac{dispersion}{embeddedness}$ was the same node so the highest $\frac{dispersion}{embeddedness}$ markings were not used. It appears most of the red edges (edges of max dispersion) were early edges so they appear behind others and is somewhat difficult to see. A more clear graph with other edges removed:

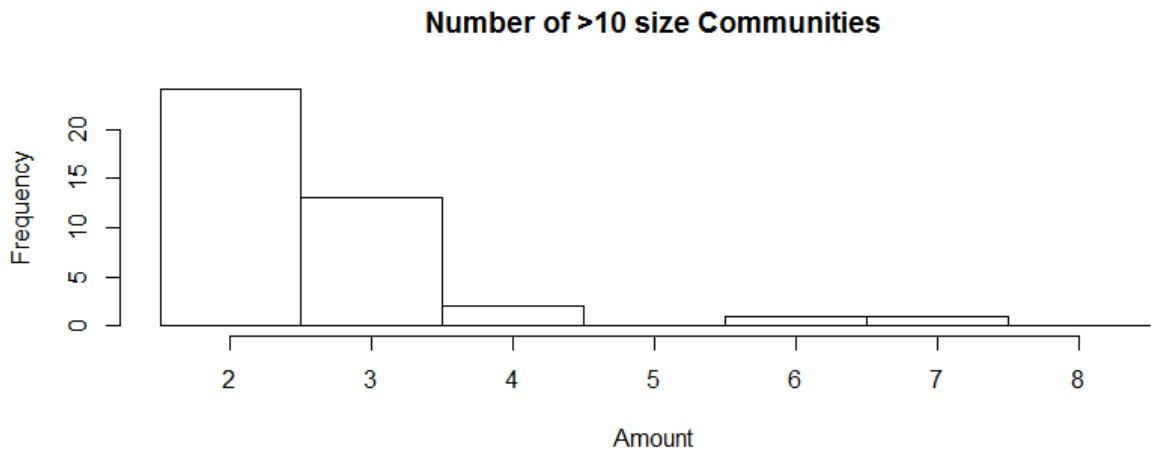


Dispersion seems to measure how much of a "hub" the node is in addition to the core node. The node of interest's mutual neighbors with the core node may only be connected to each other through the core and node of interest for the dispersion to be high.
Embeddedness measures how much a node's mutual neighbors overlap with that of the core node. This measure may show how similar a node's group of friends is to the core node.
Theoretically, $\frac{dispersion}{embeddedness}$ should give a ratio of how much a node is a secondary hub in a personal network to how similar it is to the core of the persona network. However, at least in the 3 cases that were tested, the node that had maximum value was merely the node with the highest dispersion.

Problem 6

This part dealt with looking for similar community structures between personal networks. The personal networks of the core nodes from the previous sections were used in the analysis, with the fast greedy community finding method to determine community structure.

For each personal network, the fast greedy algorithm is used to find the communities. Then, for communities with more than 10 nodes, subgraphs that only include the nodes in that community are created for further analysis. The modularity and clustering coefficient are calculated for each community. The percentage of nodes in the overall personal network that the community envelops is also found. The statistics calculated can then be examined to find similarities and potential same type communities.



As shown in the plot, many of the personal networks only had 2 or 3 communities that were large enough to qualify. Thus, to simplify the process of finding 2 types of communities, these networks' statistics are examined first for similarities.

The next page shows statistics for networks with 2 qualified communities:

[1] 0.4054964 0.1436934
[1] 0.07161101 0.18483743
[1] 0.11017552 0.08674278
[1] 0.11240558 0.07157481
[1] 0.07661574 0.15278500
[1] 0.1113902 0.1123970
[1] 0.1375443 0.1274657
[1] 0.11354499 0.07095212
[1] 0.07948774 0.10877946
[1] 0.10065166 0.07866595
[1] 0.02372436 0.07387341
[1] 0.02765862 0.03366824
[1] 0.07414583 0.29926783
[1] 0.02590140 0.03456357
[1] 0.09502005 0.02972739
[1] 0.03104839 0.04453442
[1] 0.02180758 0.04441916
[1] 0.06271727 0.02479201
[1] 0.12122242 0.02814943
[1] 0.03230029 0.06974776
[1] 0.05003996 0.02678854
[1] 0.03644368 0.02708271
[1] 0.02837366 0.07530173
[1] 0.07808408 0.02722120

Modularity

[1] 0.4429170 0.5710029
[1] 0.7493293 0.6897602
[1] 0.6792788 0.7311728
[1] 0.6800228 0.7600346
[1] 0.7552254 0.6870958
[1] 0.6578281 0.7128246
[1] 0.6675509 0.6622667
[1] 0.7120287 0.7247414
[1] 0.7016317 0.6574025
[1] 0.6966114 0.7347162
[1] 0.8864055 0.7533701
[1] 0.8613593 0.8312328
[1] 0.7453361 0.5680869
[1] 0.8479354 0.8485176
[1] 0.7378685 0.8739549
[1] 0.8628080 0.8131676
[1] 0.8738207 0.7967090
[1] 0.8274198 0.8289931
[1] 0.6994635 0.8661631
[1] 0.834086 0.777089
[1] 0.7950728 0.8649022
[1] 0.8429489 0.8640182
[1] 0.8464978 0.7311993
[1] 0.7613767 0.8789593

Clustering Coefficient

[1] 0.4652174 0.4521739
[1] 0.3932039 0.6067961
[1] 0.5 0.5
[1] 0.60181 0.39819
[1] 0.3915094 0.6084906
[1] 0.5889831 0.4110169
[1] 0.4405286 0.5198238
[1] 0.5714286 0.4285714
[1] 0.5142857 0.4666667
[1] 0.5990099 0.3762376
[1] 0.4732143 0.5267857
[1] 0.5343137 0.4656863
[1] 0.5825243 0.3980583
[1] 0.4803922 0.4705882
[1] 0.4369369 0.5630631
[1] 0.6116505 0.3883495
[1] 0.5432692 0.4567308
[1] 0.4798206 0.5201794
[1] 0.4212766 0.5446809
[1] 0.609589 0.390411
[1] 0.4375 0.5625
[1] 0.480198 0.519802
[1] 0.5322034 0.4677966
[1] 0.4423077 0.5576923

Percentage of Personal Network

The same stats can be found for networks with 3 qualified communities:

```
[1] 0.5175140 0.1747091 0.2698966
[1] 0.1018167 0.1095664 0.1763953
[1] 0.08163756 0.08207306 0.10474288
[1] 0.04380058 0.07588627 0.07621932
[1] 0.06955986 0.12415661 0.07565824
[1] 0.06898019 0.06530244 0.12496075
[1] 0.29170031 0.02386163 0.02385570
[1] 0.009995835 0.021764986 0.037218560
[1] 0.038928200 0.018863166 0.005770636
[1] 0.02151777 0.02555884 0.03210150
[1] -7.806256e-17  2.573744e-02  3.011234e-02
[1] 3.209238e-17 2.958358e-02 2.218583e-02
[1] 3.469447e-18 4.201208e-02 2.836578e-02
```

Modularity

```
[1] 0.4853627 0.5074544 0.8119763
[1] 0.7102949 0.6293934 0.5433162
[1] 0.7740693 0.7545116 0.6239783
[1] 0.8393372 0.7548901 0.7020220
[1] 0.8113208 0.6697541 0.6972503
[1] 0.7596897 0.7612374 0.6630535
[1] 0.6638692 0.8854427 0.8781265
[1] 0.9021717 0.8827116 0.8257142
[1] 0.8164923 0.8845085 0.9006686
[1] 0.8671457 0.8743243 0.8538044
[1] 0.9375000 0.8461883 0.8718708
[1] 0.9245902 0.8366947 0.8733532
[1] 0.9416810 0.8506187 0.8573642
```

Clustering Coefficient

```
[1] 0.44359465 0.46271511 0.06692161
[1] 0.3060345 0.3103448 0.3836207
[1] 0.06382979 0.41702128 0.51914894
[1] 0.1213592 0.3932039 0.4854369
[1] 0.05284553 0.55691057 0.39024390
[1] 0.2901961 0.2784314 0.4313725
[1] 0.3066667 0.1022222 0.5911111
[1] 0.1170732 0.4341463 0.4487805
[1] 0.48815166 0.41232227 0.09952607
[1] 0.4455446 0.1188119 0.4356436
[1] 0.06467662 0.52736318 0.40796020
[1] 0.06896552 0.49753695 0.43349754
[1] 0.05940594 0.47029703 0.47029703
```

Percentage of Personal Network

Now we attempt to find some communities with similar statistics. We will denote the stats in the format (Modularity, CC, % of PN).

It appears most of the modularity values are relatively low, however, some are somewhat high. One community has the stats (0.405, 0.443, 0.465). Another has (0.518, 0.485, 0.443); thus, perhaps communities with relatively high modularity scores are of the same type.

There are also multiple communities with very low modularity (~0.02) and high clustering coefficient (>0.8). This might be another type of community.

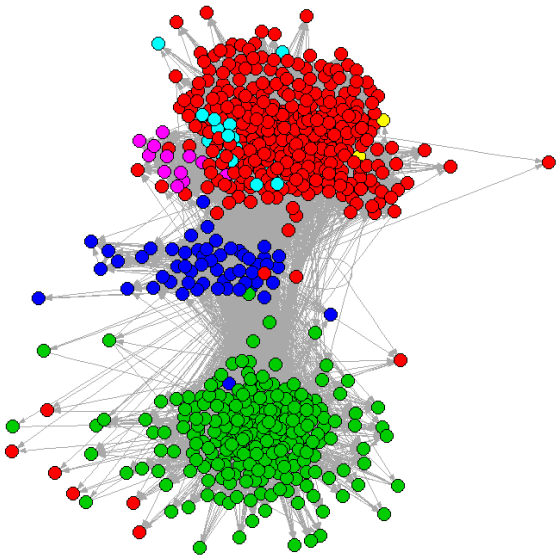
Finally, there are also communities with moderate modularity (~0.1) and slightly lower clustering coefficient (0.6-0.7). These communities might all be of the same type as well.

Judging from the percentage of personal network statistic, it appears most of the personal networks had 2 main clusters with various other smaller ones.

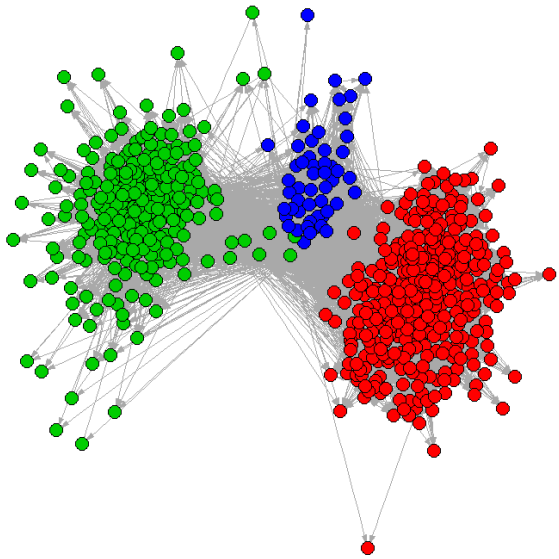
Problem 7

The data files are formatted in such a way that merely importing as dataset did not work. Thus, the data needed to be parsed manually. First, the .circles file was parsed and the number of circles determined. All further operations were only conducted if there were more than 2 circles. It turns out 57/132 of the users qualified.

For each qualified network, the Infomap and Walktrap community algorithms are run to find community structure. The following plots are examples of the results for a particular network:



Infomap Method



Walktrap Method

The circles data is then accessed and the membership with each circle is compared with the community memberships. The Jaccard index, $\frac{A \cap B}{A \cup B}$ is found for each pair of community/circle memberships. Thus, for each network, a $n * m$ matrix is found of Jaccard indices for n circles and m communities. This process is repeated for both the Infomap and Walktrap methods, so each network would have 2 matrices of data. An example of such matrices follows:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	0.268571429	0	0.000000000	0.08181818	0	0.00000000
[2,]	0.041055718	0	0.000000000	0.00000000	0	0.00000000
[3,]	0.026392962	0	0.000000000	0.00000000	0	0.00000000
[4,]	0.116959064	0	0.000000000	0.01785714	0	0.00000000
[5,]	0.040935673	0	0.000000000	0.03333333	0	0.00000000
[6,]	0.317280453	0	0.000000000	0.09375000	0	0.00000000
[7,]	0.371104816	0	0.000000000	0.08163265	0	0.00000000
[8,]	0.142441860	0	0.009433962	0.03030303	0	0.00000000
[9,]	0.260000000	0	0.000000000	0.08411215	0	0.00000000
[10,]	0.383098592	0	0.000000000	0.09210526	0	0.00000000
[11,]	0.008797654	0	0.000000000	0.00000000	0	0.00000000
[12,]	0.204081633	0	0.000000000	0.00000000	0	0.02777778

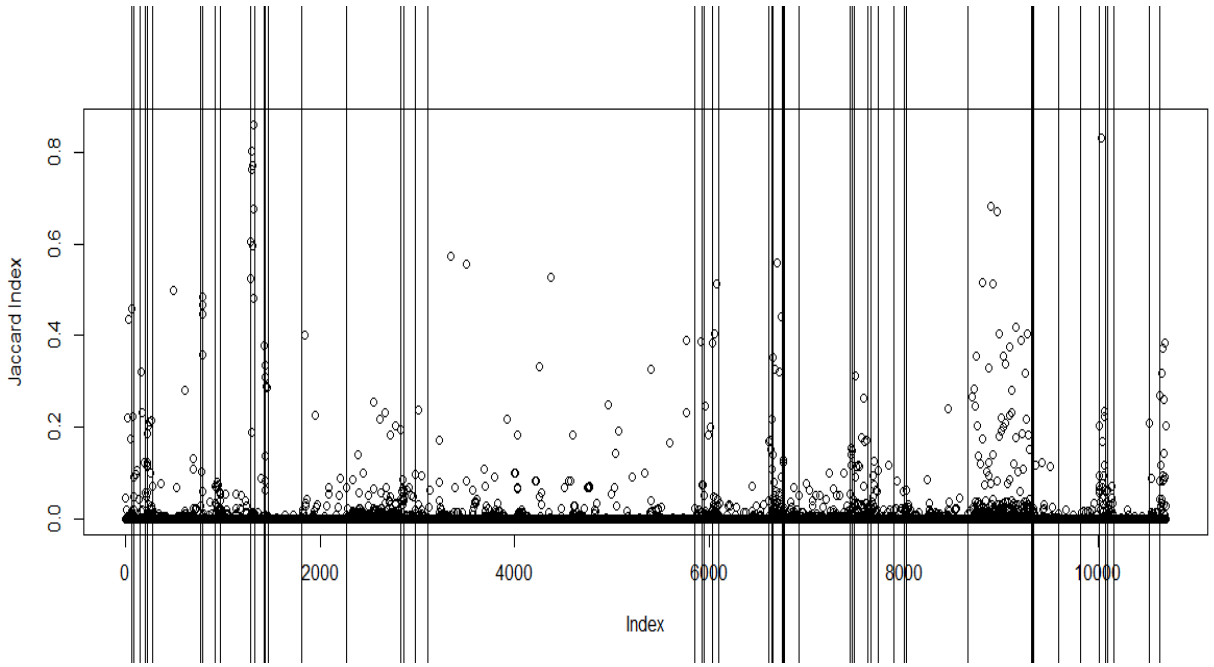
Infomap Method

```
             [,1] [,2] [,3]
 [1,] 0.283746556    0    0
 [2,] 0.038567493    0    0
 [3,] 0.024793388    0    0
 [4,] 0.112947658    0    0
 [5,] 0.041322314    0    0
 [6,] 0.341597796    0    0
 [7,] 0.393939394    0    0
 [8,] 0.143250689    0    0
 [9,] 0.275482094    0    0
[10,] 0.413223140    0    0
[11,] 0.008264463    0    0
[12,] 0.198347107    0    0
```

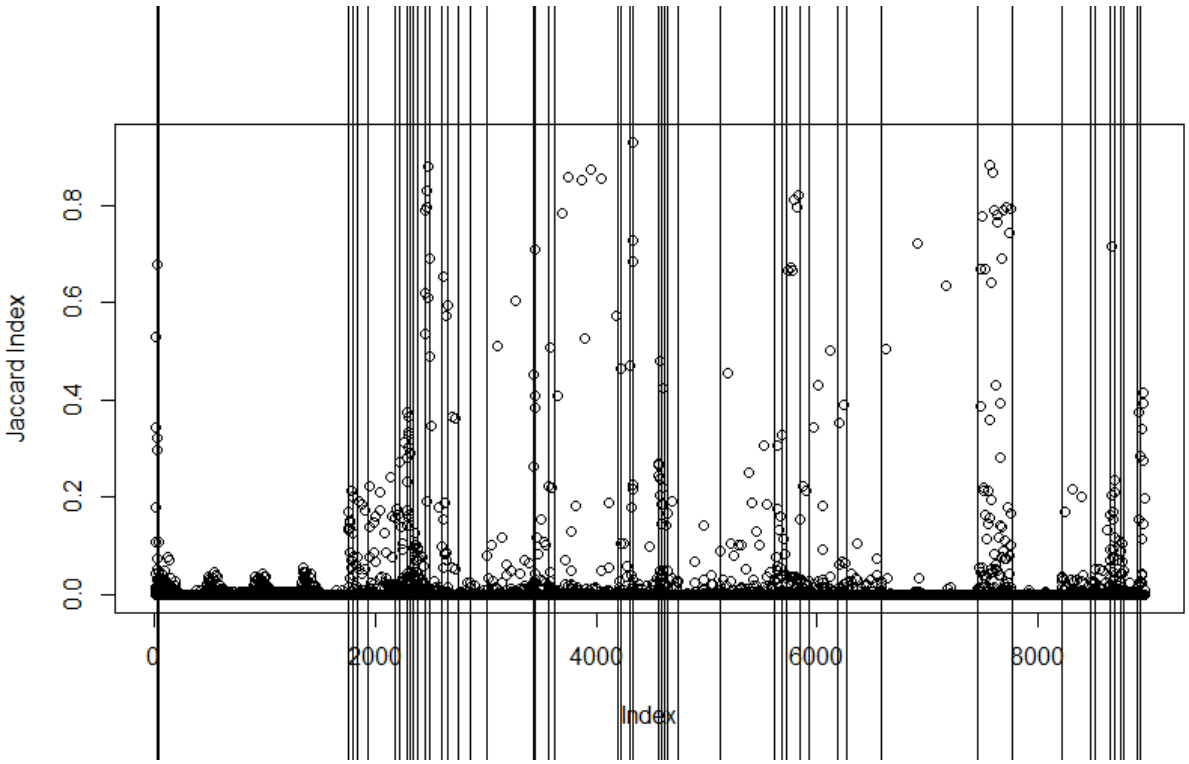
Walktrap Method

Evidently, at least for this network, some communities had no overlap with the user's circles. However, some circles had a moderately significant amount of overlap.

To compare the differences in Jaccard Indices between users, each particular Jaccard value can be plotted for both types of community methods. Vertical lines are drawn to show separation between values of different networks.

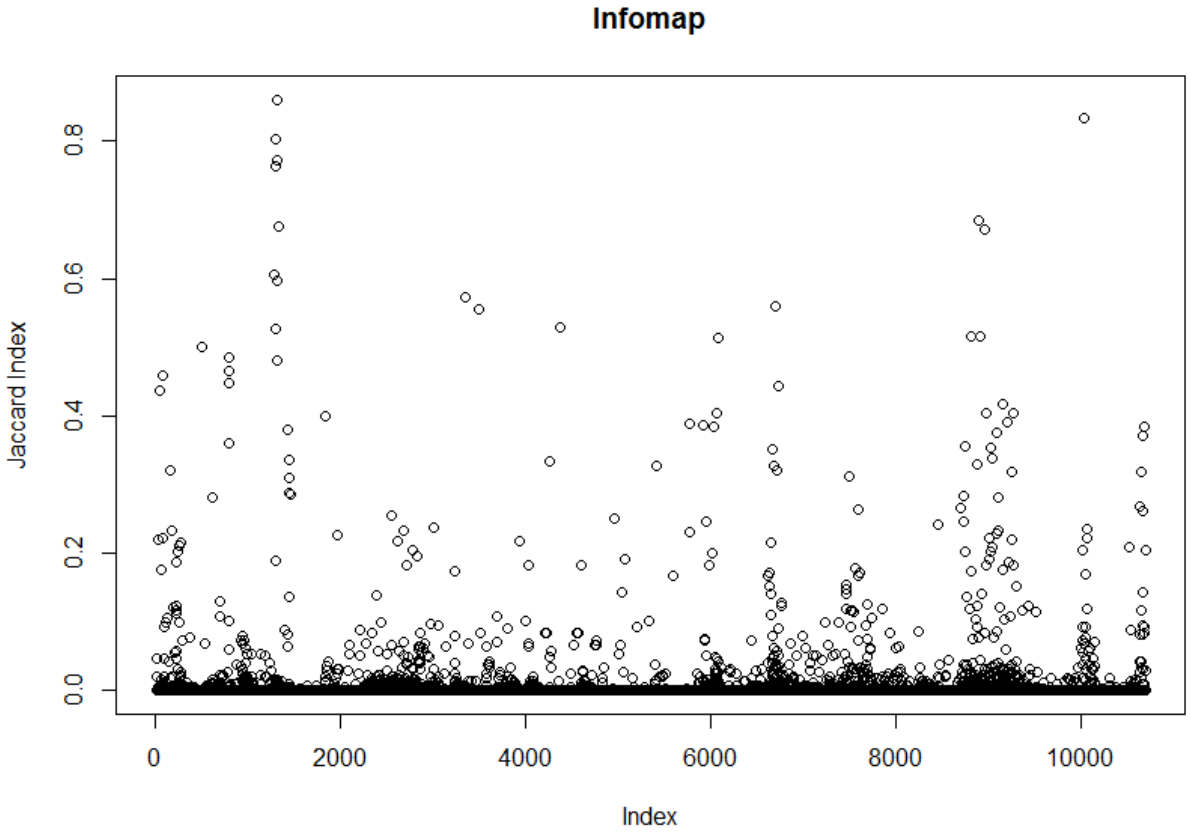


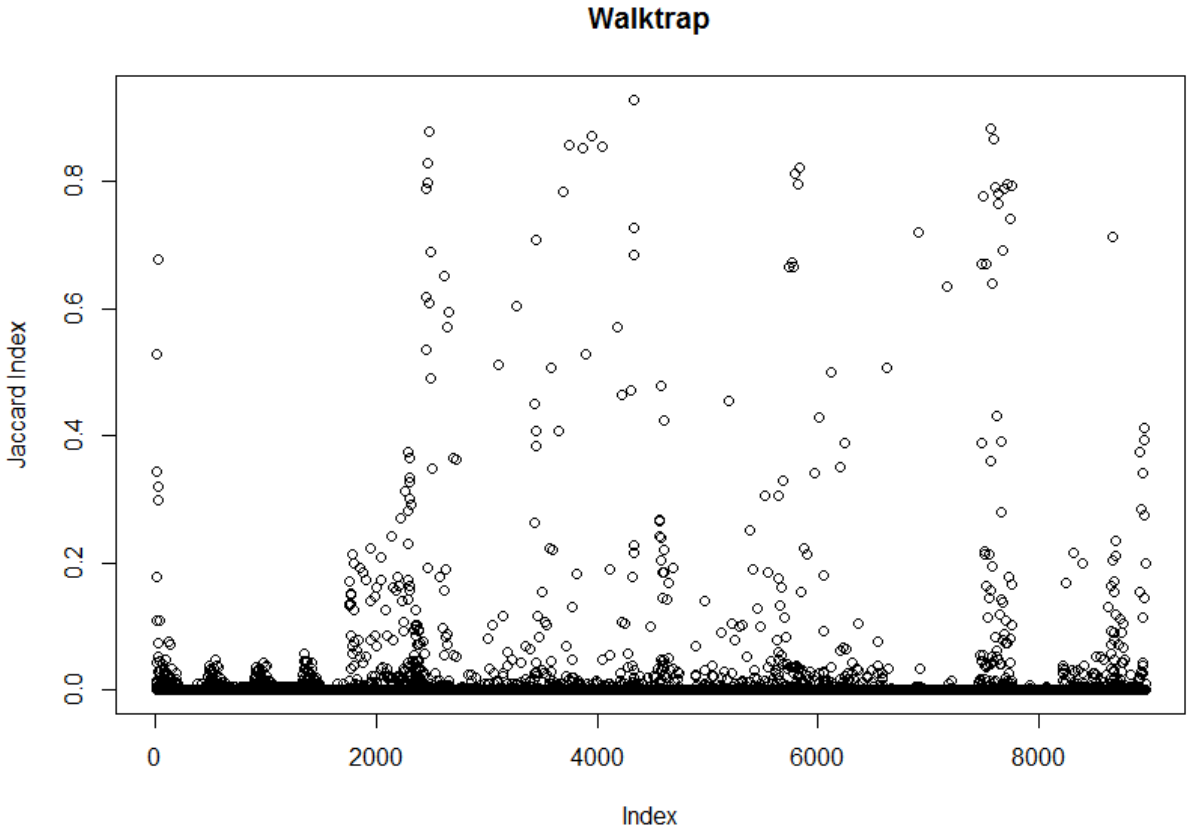
Infomap Method



Walktrap Method

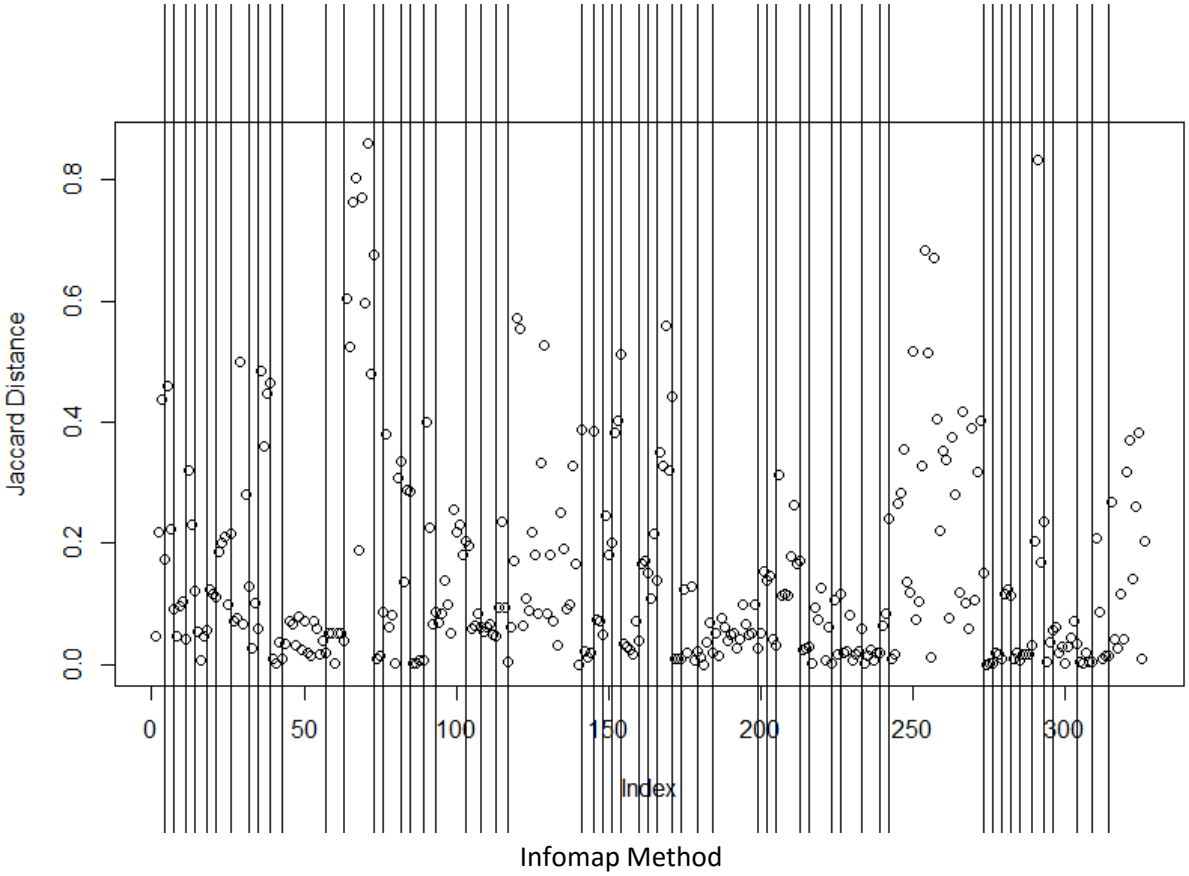
The following plots show the data without lines denoting different networks/users.



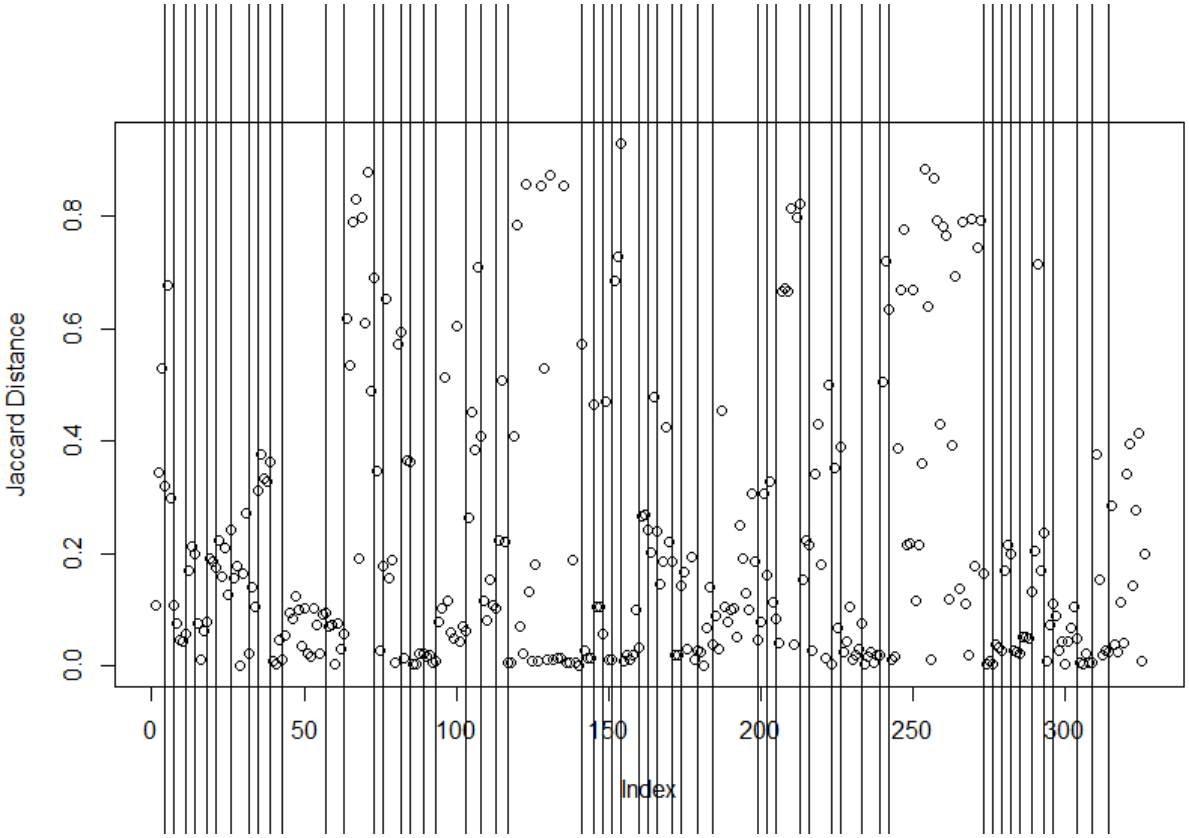


It appears that a good amount of users use circles to tag specific communities that can be detected with the community finding algorithms. However, the algorithms frequently returned more communities than circles, with many of these communities having little to no overlap with a user's circles.

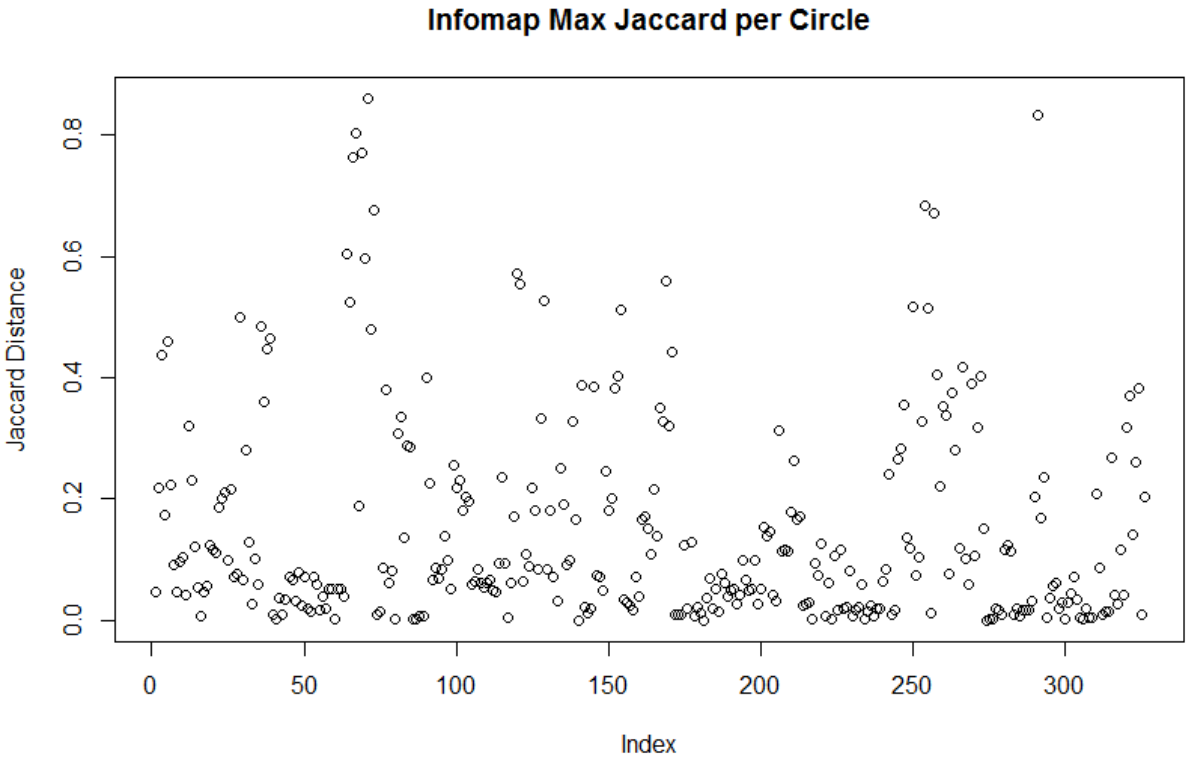
Now, plots are created using only the maximum Jaccard index for each circle. The idea behind these plots is to show circles paired with the community that it is most likely to correspond to. Once again, vertical lines represent separate networks/users.

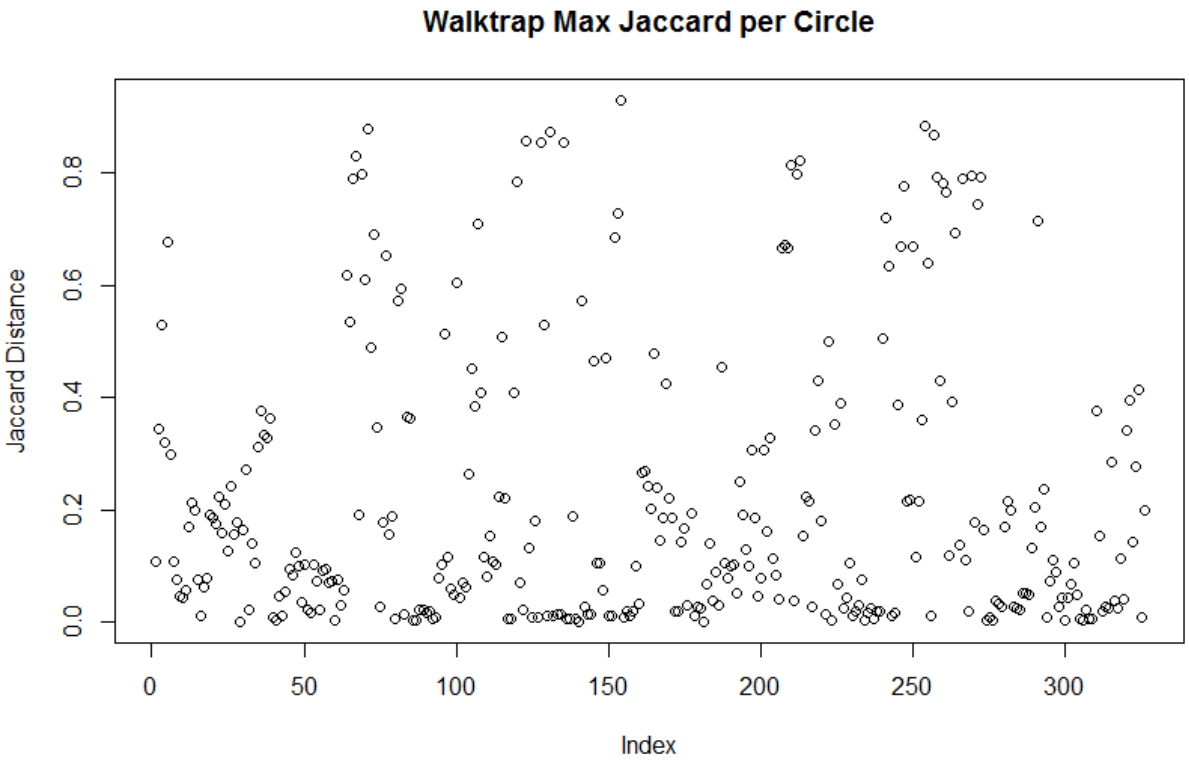


Infomap Method



Walktrap Method


Infomap Max Jaccard per Circle
Jaccard Distance
0.0
0.2
0.4
0.6
0.8
0
50
100
150
200
250
300
Index


Walktrap Max Jaccard per Circle
Jaccard Distance
0.0
0.2
0.4
0.6
0.8
0
50
100
150
200
250
300
Index

Once again, it is apparent that some users are more likely to tag certain community relationships as circles. However, it is still clear that many circles are not represented well in communities found using the algorithms. It is noted that some who have high Jaccard Index circles often have high values for all of their circles, and those with many low Jaccard Index circles frequently have many low Jaccard Index circles. Perhaps this is representative of how users behave and how each of them tag their circles; maybe some tag circles as relationships that make sense according to the notion of community structure in graphs, while others do not.

Codes

There are 4 .R code files. P1-1.R contains code for problems 1-4, while the rest, P1-5, P1-6, and P1-7 contain code for the problems they correspond to.