UCLA ECE 232E S2021 Project #2

Social Network Mining

Lin Fan - 505627503

Sunay Bhat - 905629072

Yi-chun Hung - 705428593

Tu Yu-Hsien - 405627283

Table of Contents

Facebook Network

1. Structure properties of the Facebook network

Q1.1)

After building the network from “facebook_combined.txt”, the **number of nodes in the Facebook network is 4039**. And the **number of edges in the network is 88234**.

Q1.2)

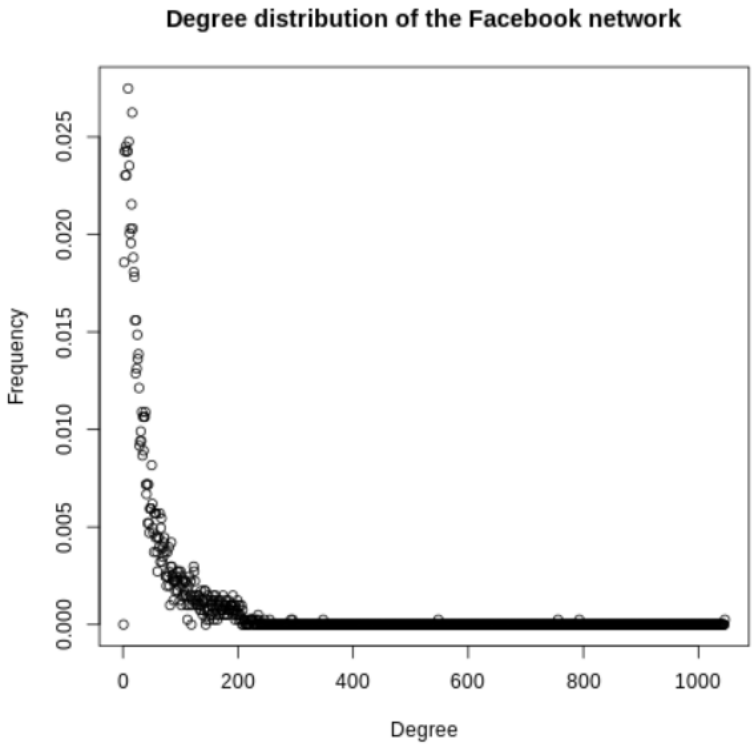
This network is **connected**.

Q2)

The diameter of this network is **8**.

Q3)

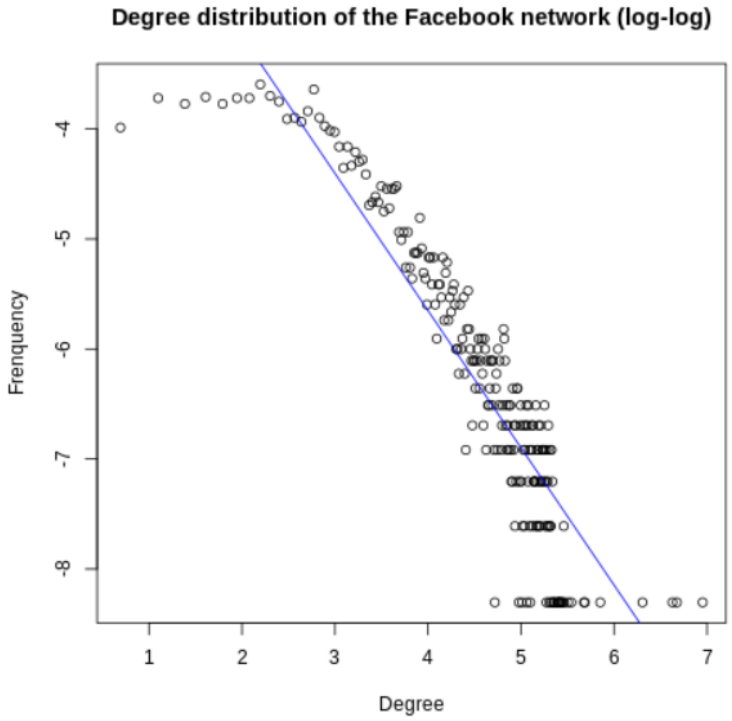
The degree distribution of the network is shown below:



We can see that most of the nodes are with low degrees, and the network follows the power law distribution. The **average degree is 43.69**.

Q4)

The degree distribution is plotted in the **log-log scale** and the slope of the line after estimation is approximately **-1.2475**. The graph and the line fitting the curve are shown below:



2. Personalized network

Q5)

We create a personalized network that describes a node with all 1 distance neighbors on the graph. For the user whose user ID is 1, **the number of nodes we get is 348, and the number of edges we get is 2866.**

Q6)

For this network, **the diameter is 2**. The **trivial upper bound is 2, and lower bound is 1** for the diameter of the personalized network.

Q7)

In the context of the personalized network, our upper bound is 2. It means that there are at least 2 nodes v1 and v2, their distance between each other, except for the central node 1, is 2. In this case, we can conclude that user v1 and user v2 aren't friends with each other, but both of them have connectivity with the central user.

On the other side, our lower bound is 1. This means that any pair of nodes in the subgraph are mutually connected with each other. Distances between any pair of nodes are all equal to 1. In this case, we can conclude that every user within the network is friends with each other.

The graph of this personalized network is shown below:

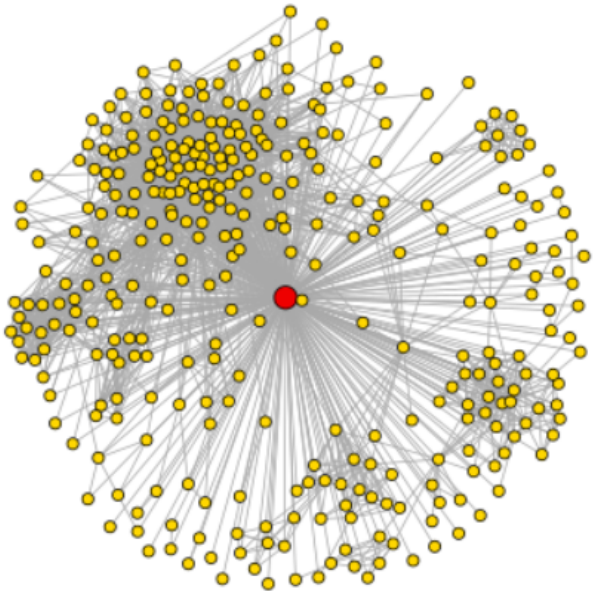


Figure: Personalized network for User ID=1

3. Core node's personalized network

Q8)

A core node is the nodes that have more than 200 neighbors. The **number of core nodes** in the Facebook network is **40**. The **average degree** of the core nodes is **279.375**.

3.1. Community structure of core node's personalized network

Q9)

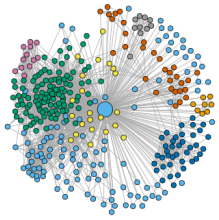
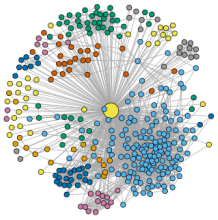
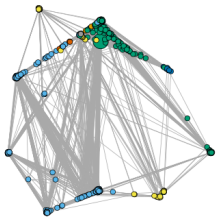
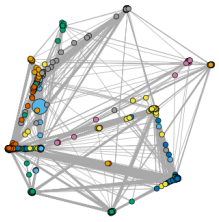
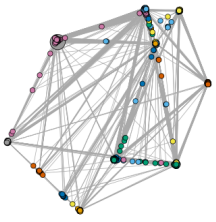
The following table shows the community structure of the core node's personalized network using Fast-Greedy, Edge-Betweenness, and Infomap community detection algorithms. We found community structure for core node 1, 108, 349, 383, and 1087. The modularity of a graph measures how separated are the different vertex types from each other.

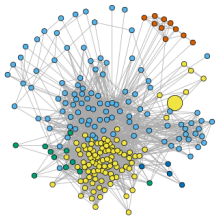
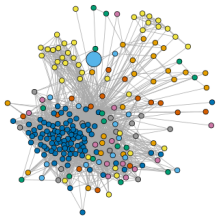
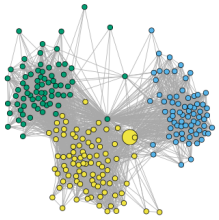
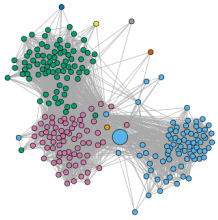
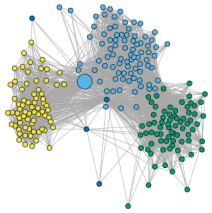
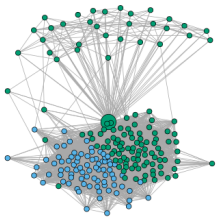
The fast greedy algorithm starts with a non-clustered initial assignment, computes the expected improvement of modularity for each pair of communities, chooses the pair with max modularity improvement and merges into the new community. **The modularity score is not very accurate**

since this algorithm cannot guarantee finding the global maximum and can also find a bad local optimal. The modularity score, especially core node 1087, is very different from the modularity score of the other two algorithms.

The modularity score of the edge-betweenness algorithm is more accurate compared to the fast greedy algorithm. However, the edge-betweenness algorithm is the slowest out of the three algorithms since it goes through the whole graph to compute edge betweenness every iteration. Especially for core node 108, it takes about an hour using this algorithm to find community structure.

Infomap first encodes the network in a way that maximizes the amount of information about the original network and then decodes the information. **The Infomap algorithm generates satisfied modularity scores (scores are very close to the modularity score of edge-betweenness algorithm).** This algorithm is also less time consuming compared to the edge-betweenness algorithm.

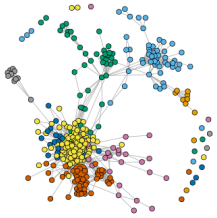
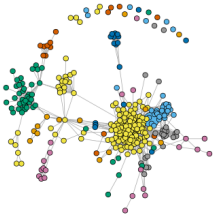
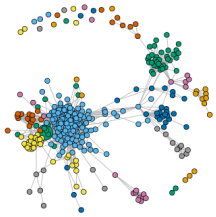
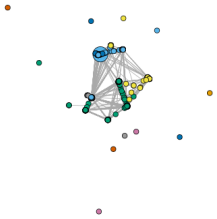
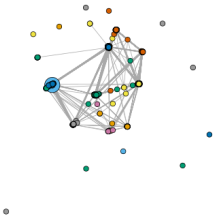
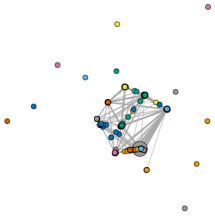
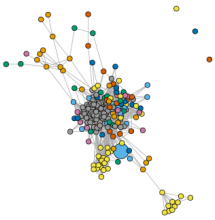
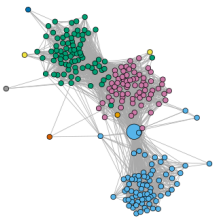
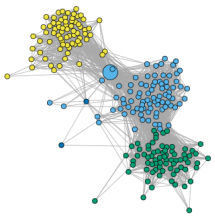
	Fast-Greedy	Edge-Betweenness	Infomap
Node 1	 Modularity: 0.4131014	 Modularity: 0.3533022	 Modularity: 0.3891185
Node 108	 Modularity: 0.4359294	 Modularity: 0.5067549	 Modularity: 0.5082492

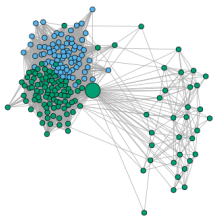
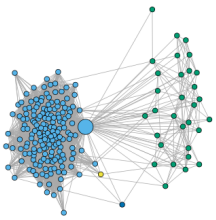
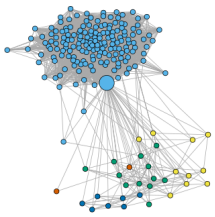
Node 349	 Modularity: 0.2517149	 Modularity: 0.133528	 Modularity: 0.203753
Node 484	 Modularity: 0.5070016	 Modularity: 0.4890952	 Modularity: 0.5152788
Node 1087	 Modularity: 0.1455315	 Modularity: 0.02762377	 Modularity: 0.02690662

3.2. Community structure with the core node removed

Q10)

The following table shows the community structure of the core node's personalized network using Fast-Greedy, Edge-Betweenness, and Infomap community detection algorithms but with core nodes removed. We found the modified community structure for core node 1, 108, 349, 383, and 1087. **It is clear that the modularity score increased for all three algorithms after the core nodes are being removed from the personalized network.** This makes sense because the core node shares edges with all communities and thus making it less separated between communities. Deleting the core nodes make connections more sparse.

	Fast-Greedy	Edge-Betweenness	Infomap
Node 1	 Modularity: 0.4418533	 Modularity: 0.4161461	 Modularity: 0.4180077
Node 108	 Modularity: 0.4581271	 Modularity: 0.5213216	 Modularity: 0.5185931
Node 349	 Modularity: 0.2456918	 Modularity: 0.1505663	 Modularity: 0.2448156
Node 484	 Modularity: 0.5342142	 Modularity: 0.5154413	 Modularity: 0.5434437

Node 1087	 Modularity: 0.1481956	 Modularity: 0.0324953	 Modularity: 0.02737159

3.3. Characteristic of nodes in the personalized network

Q11)

Embeddedness of a node is defined as the number of mutual friends a node shares with the core node. **Degree of a node is defined as the number of edges that are incident to the node.**

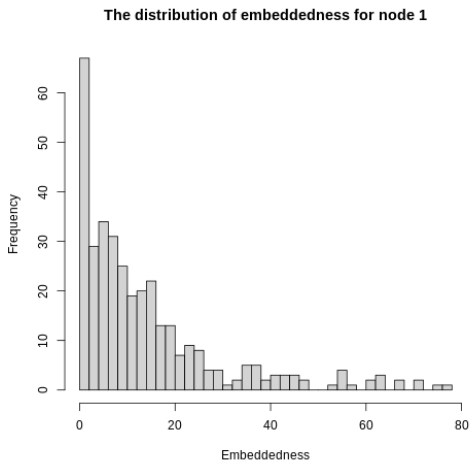
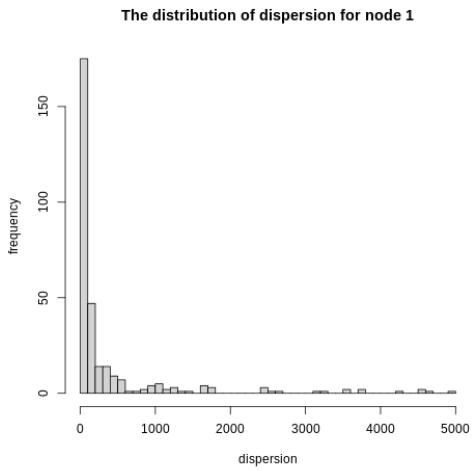
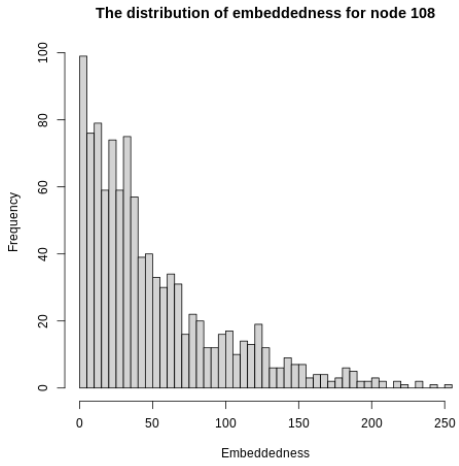
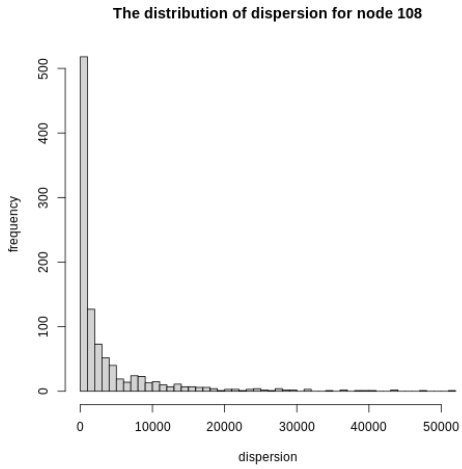
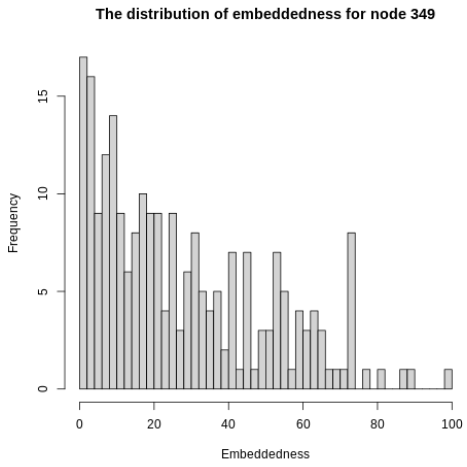
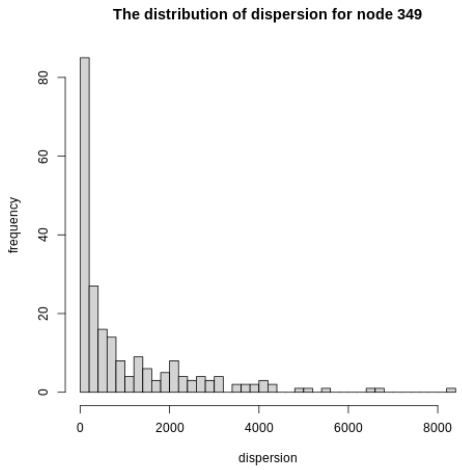
The expression relating the Embeddedness between the core node and a non-core node to the degree of the non-core node in the personalized network of the core node is given below:

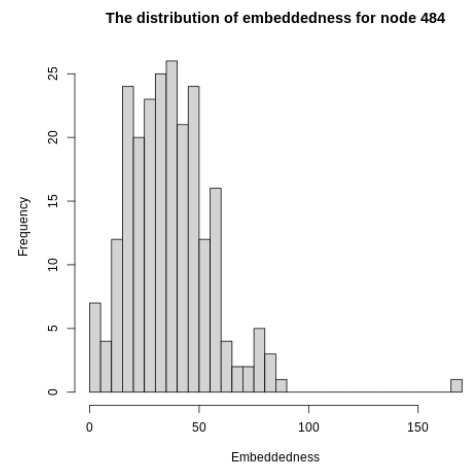
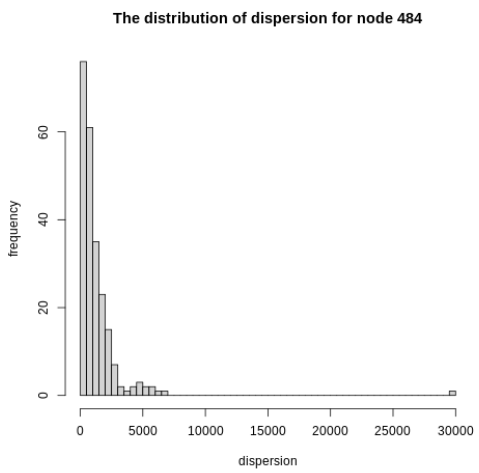
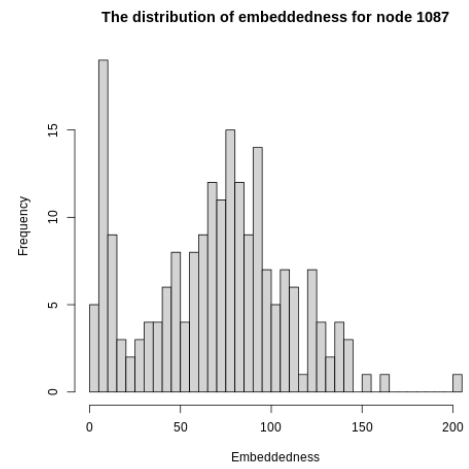
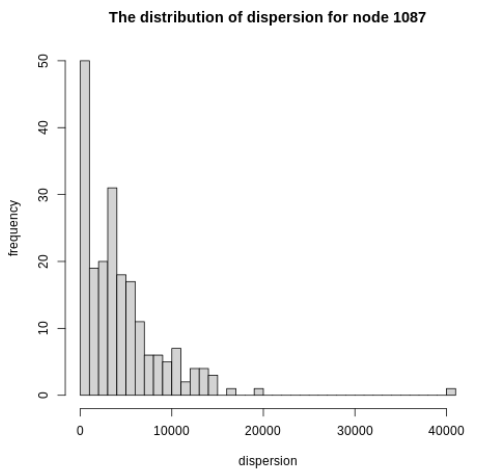
$$Embeddedness\,(i) \;=\; Degree\,(i) \;-\; 1$$

This makes sense because in the core node network, **the core node is connected to all other non-core nodes**. **Thus the embeddedness becomes the number of nodes connected to that non-core node, which is the degree of that non-core node minus 1** (since the non-core node is also connected to the core-node and is counted in the node's degree).

Q12)

Dispersion of a node is defined as the sum of distances between every pair of the mutual friends the node shares with the core node. The distances are calculated in a modified graph where the node (whose dispersion is being computed) and the core node are removed.

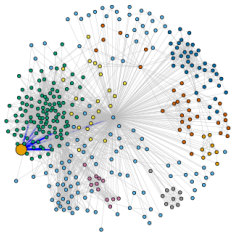
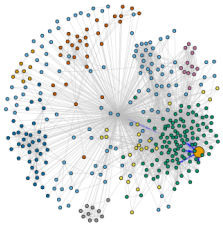
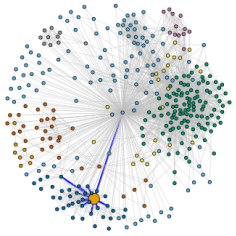
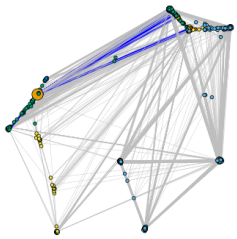
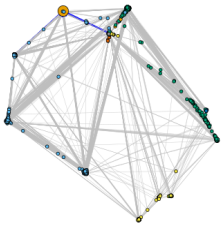
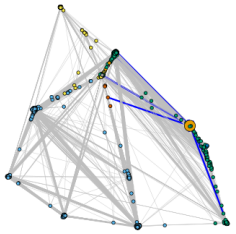
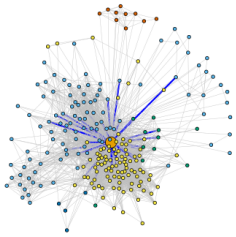
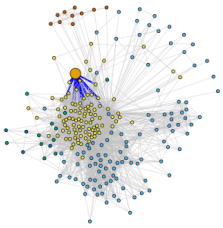
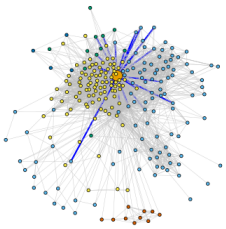
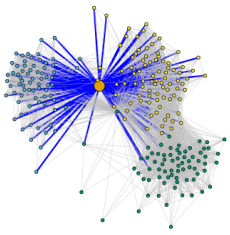
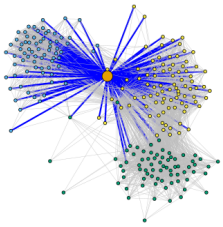
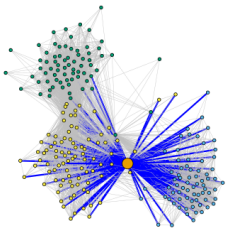
	Distribution histogram of embeddedness	Distribution histogram of dispersion
Node 1		
Node 108		
Node 349		

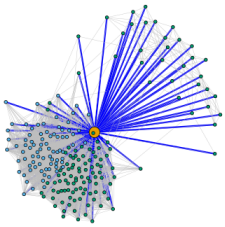
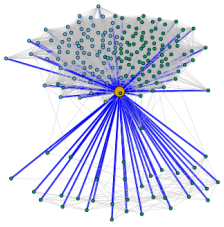
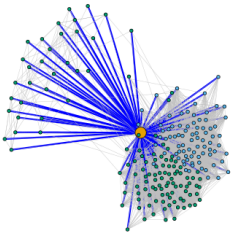
Node 484		
Node 1087		

Q13) & Q14)

The results for questions 13 and 14 are combined in the table and plots below.

Max Value Node IDs			
Central Node ID	**Max Dispersion (Q13)**	**Max Embeddedness (Q14)**	**Max Ratio (Q14)**
1	**46**	**56**	**18**
108	**993**	**1022**	**993**
349	**30**	**32**	**30**
484	**1**	**1**	**1**
1087	**1**	**1**	**1**

	Community structure Max dispersion	Community structure Max embeddedness	Community structure Max $\frac{dispersion}{embeddedness}$
Node 1			
Node 108			
Node 349			
Node 484			

Node 1087			

Q15)

Performance is measured by how many different communities the measurement can find. From this definition, **we can conclude that dispersion/embeddedness is a better measurement than dispersion; and dispersion is a better measurement than embeddedness**. This makes sense because embeddedness is the amount of mutual connections between friends whereas dispersion is the amount of connections a person has between unrelated groups of friends. The neighbors of the node with max dispersion are very likely unconnected and belong to different communities. On the other hand, the node with max embeddedness shares many mutual nodes with the core node. The maximum ratio of dispersion/embeddedness is the best measurement because the dispersion is large whereas the embeddedness is small.

4. Friend recommendation in personalized networks

4.3. Creating the list of users

Q16)

The number of nodes with degree 24 is 11. And the node indices of these nodes are as follows:
`[578 600 615 618 627 643 658 659 661 662 496].`

4.4. Average accuracy of friend recommendation algorithm

Q17)

	Common Neighbors	Jaccard	Adamic Adar
Average Accuracy of the Personalized Network	**0.832**	**0.807**	**0.828**

	Common Neighbors	Jaccard	Adamic Adar
Average Accuracy of the Whole Network	**0.457**	**0.403**	**0.478**

As shown in the tables, the method based on the **common neighbors measure performs slightly better than the one based on the adamic adar measure in the personalized network**; but, **in the whole network, the one based on the common neighbors measure does not perform better**. We think this is due to the structure of the network. Additionally, we have also found that the variance of the accuracy is relatively large in all three methods. It also implies that the performance of the methods varies among nodes and the structure of the network.

Google+ Network

Q18)

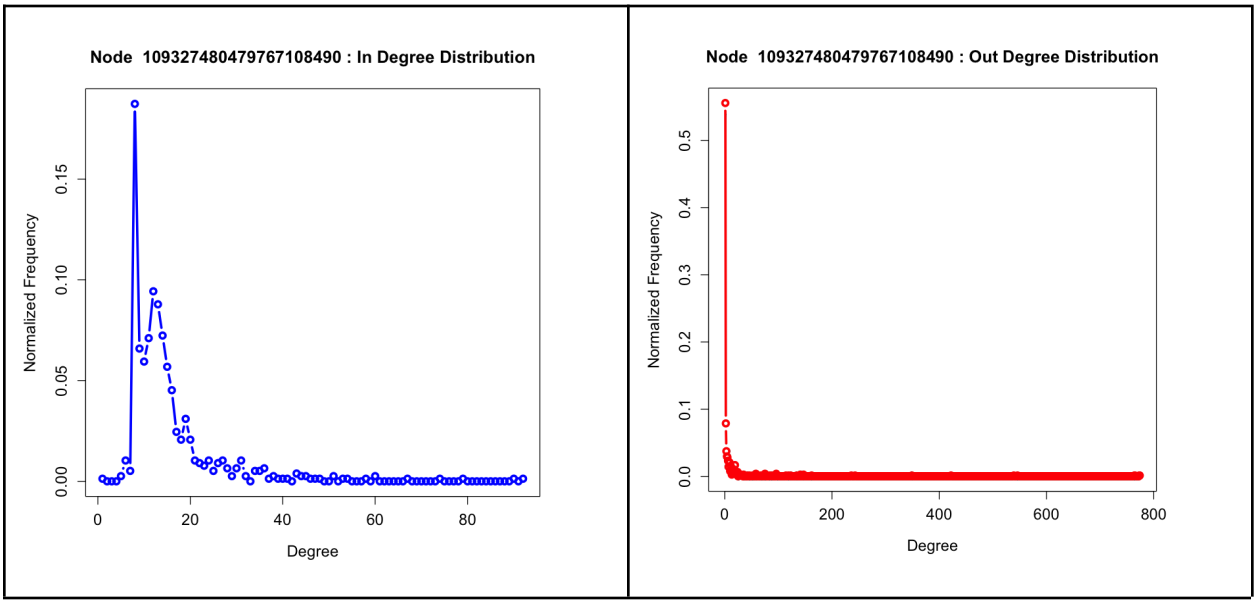
We are asked to create personal networks for all users in more than 2 circles. By analyzing the circles files, we find there are **57 users** with membership in more than 2 circles.

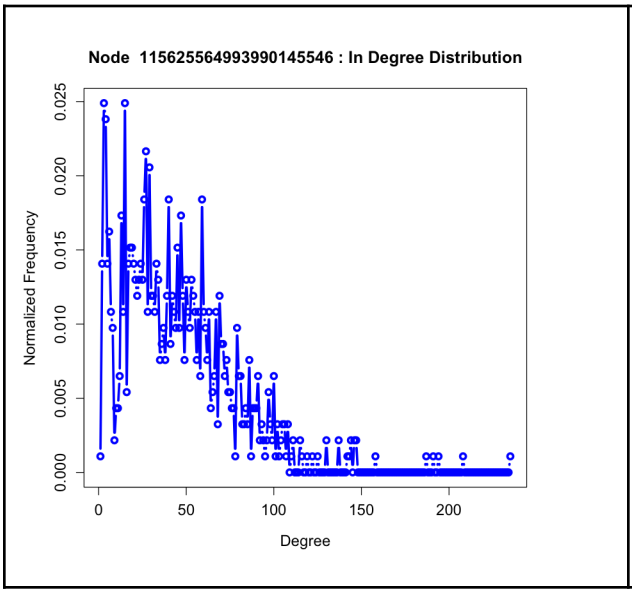
Q19)

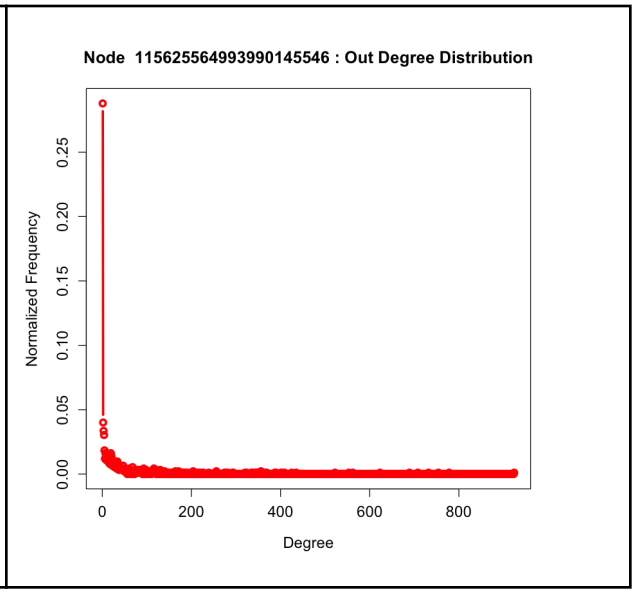
The following 6 plots show the “In” and “Out” degree distributions for each of the three nodes’ ego networks. The distributions and statistics on the nodes show quite different characteristics. The **out-degree distributions for all three nodes are quite similar in shape**, where the vast majority of nodes have very small out degrees and a steady but low percentage of nodes populate the higher degrees. Node 1 (“...8490”) in-degree also has a higher peak at the low

degree, but Node 2 (“...5546”) and 3 (“...6744) **each progressively have a wider in-degree distribution** with more nodes that have medium degree counts. Node 1, 2 and 3 **increase in the total number of edges**, and hence the **degree distributions and statistics scale up as well**. The table at the end summarizes the statistics, and we can see how the **variance increases with the mean and shape of the distribution**.

In practical terms, the **third ego network contains the most interconnected nodes** due to the in-degree distributions, with the second having more interconnections than the first. But the **out-degree indicates a similar behavior across networks which makes intuitive sense for an ego-network** unless there were many very tightly connected clusters in a node’s ego-network.







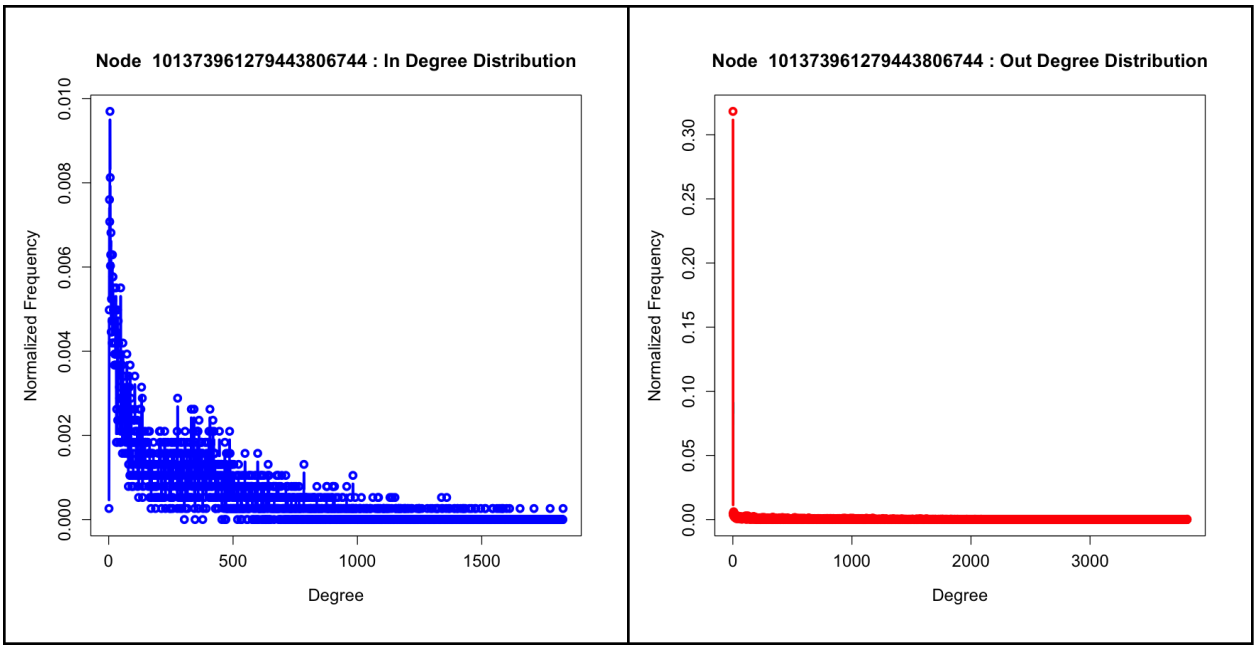


Table of Statistics

Node ID	μ	σ^2 (In-Deg)	σ^2 (Out-Deg)
109327480479767108490	**14.06**	**96.00**	**4588.18**
115625564993990145546	**43.64**	**1020.62**	**9351.30**
101373961279443806744	**298.12**	**86408.77**	**166186.74**

1. Community structure of personal networks

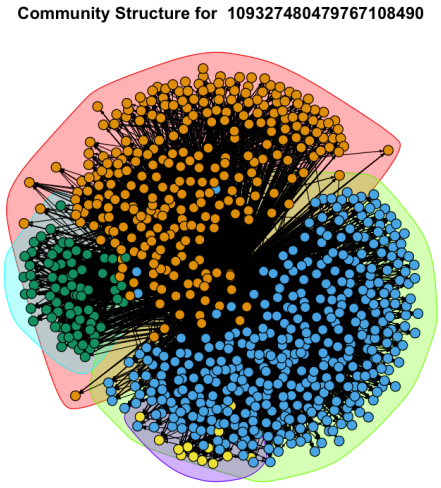
Q20)

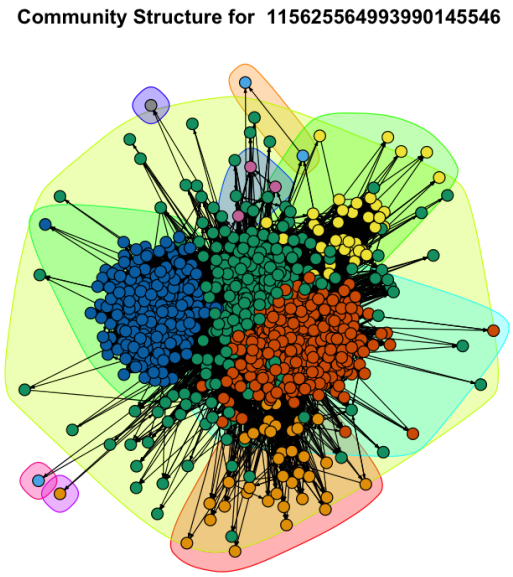
The following table contains the modularity scores and the plots show the community structure highlighted. Community structure was found using the **walk-trap detection** algorithm. This algorithm was developed based on the idea that **random walks tend to get "trapped" in dense communities** and thus one can use the paths of random walks to cluster communities.

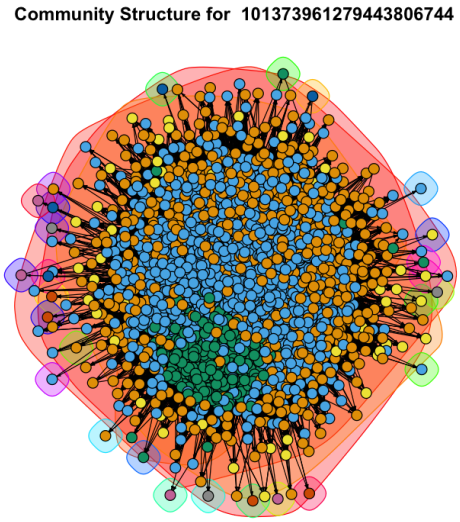
The **modularity is highest for Node 2, then Node 1, and Node 3 last**. This indicates Node 2 has the highest connectedness within communities and lower connections between communities. The balance (essentially ratio), between these two connectedness features is well-illustrated by these three ego nodes. **Node 1 has only 4 communities and too many connections between them to have a high modularity**, **which makes sense as both the in and out degree distributions are narrow (few nodes dominate connections). Node 2 has**

higher modularity due to the 10 communities and relatively weaker links between them as the second plot visually shows and the degree distribution variances might indicate. Node 3 has 31 communities, but the majority appear to be trivially small or single nodes. It also has a high degree of interconnectedness between communities thus giving it the lowest modularity score.

Node ID	Modularity	Communities
109327480479767108490	**0.253**	4
115625564993990145546	**0.319**	10
101373961279443806744	**0.191**	31







Q21)

Both homogeneity and completeness are defined in the information theoretic sense.

Homogeneity is based on **1 minus the ratio of entropies (or uncertainty) of circles (C) given communities (K) over just circles.** When this ratio is near 1, then h=0. This occurs when the community information contains no information about circles, and thus the homogeneity is zero. On the other hand, if the ratio is close to zero, then the community

structure contains a large amount of information about the circles, and thus the uncertainty of circles is very low given communities. This means h=1, or the homogeneity is high. **Homogeneity is essentially a measure of how similar the circle membership is within communities.**

Completeness is a similar measure based on **1 minus the ratio of entropies (or uncertainty) of communities (K) given circles (C) over just communities.** The interesting thing about this measure is that it can be negative since nodes can belong to multiple circles. Thus the uncertainty could be increased by knowing the circles. **Completeness is essentially a measure of how contained within communities a circle membership is, or how many communities are needed to describe a circle.**

Essentially, homogeneity and completeness are two different ways of describing the overlap between community structure and circle structure, with high homogeneity indicating communities have few or one circle membership, and high completeness indicating how well contained circle membership is to a small set of communities.

Q22)

The table below summarizes the homogeneity and completeness statistics for each of the three nodes. **We see node 1 has high homogeneity indicating that most of the communities have largely the same circle membership within them. The second node has about half the value, indicating the communities have a mix of circle membership with some majorities. For node 3, the near zero homogeneity value indicates that almost every member in a community has a different circle membership.**

The medium to low completeness in node 1 indicates that most circles have decent overlap with communities but are probably split over many communities. The negative completeness values for node 2 and 3 come from the fact that nodes can be in many circles, which can result in a H(K|C) which is higher than H(K). Node 2 has more of these nodes with multiple memberships and thus has the larger negative completeness.

These two measures are somewhat related, but do not necessite correlation. Large circles with small communities could result in high homogeneity with low completeness as we see in these three graphs.

Node ID	Homogeneity	Completeness
109327480479767108490	**0.85188512**	**0.32987391**
115625564993990145546	**0.45189030**	**-3.42396235**
101373961279443806744	**0.00386671**	**-1.50423839**