HW3: Network Analysis

Problem 1: Suggesting Similar Papers

Part (c):

How does the time complexity of your solution involving matrix multiplication in part (a) compare to your friend's algorithm?

The time complexity of my solution is O(n^x), where 2<x<3. Although theoretically, this algorithm should be faster than O(n^3), In real life practice, Things ain't that promising. It might not be practical due to hardware limitation or it might have huge constant multipliers that dramatically slow down the algorithm. Therefore, an algorithm with matrix multiplication will not be faster than my friend's algorithm.

Part (d):

Bibliographic coupling and cocitation can both be taken as an indicator that papers deal with related material. However, they can in practice give noticeably different results. Why? Which measure is more appropriate as an indicator for similarity between papers?

It is unlikely that papers will only have reference from one related field. It is normal that papers will cite a wide variety of references from different fields to support their ideas. Under such circumstance, Cocitation metric will act poorly as an indicator of paper similarity. Also, if two papers of different field cite the same reference but have no similar content, Cocitation metric will also act poorly as an indicator of paper similarity. Thus, bibliographic coupling is more appropriate as an indicator for similarity between papers.

Problem 2: Investigating a time-varying criminal network

Part (c):

Observe the plot you made in Part (a) Question 1. The number of nodes increases sharply over the first few phases then levels out. Comment on what you think may be causing this effect. Based on your answer, should you adjust your conclusions in Part (b) Question 5?

The reason that the number of nodes increases sharply is because Seizures did not start happening until phase 4. It is possible that police were using wiretaps to grow their list of suspects during early phases. Therefore, by including the early phases in Part (b) Question 5 could bias the result because the police wiretap has no effect on the criminal enterprise yet. The early phases should be removed from the calculation in Part (b) Question 5.

Part (d):

In the context of criminal networks, what would each of these metrics teach you about the importance of an actor's role in the traffic? In your own words, could you explain the limitations of degree centrality? In your opinion, which one would be most relevant to identify who is running the illegal activities of the group? Please justify.

Degree centrality measures how many players a person is connected to. Within this criminal network which is a hierarchical criminal organization, One person is more likely receiving orders only from one person. Therefore, the degree is more likely to measure the number of people one gives orders to. However, using this method to identify a leader has a problem. If the true leader has a proxy such as his second-hand, The true leader will only need to give order to one person and would rank lower on this centrality.

Betweenness centrality measures the importance of a person as a link to others in the network. It helps to find out whose removal would bring the network the greatest impact(fall apart). The key link in the criminal network would be those people with high scores on these metrics. By removing those people, the criminal organization will fall apart.

Eigenvector centrality measures the importance of a person based on the players that person is connected to. If a person is connected to important criminal leaders, that person would score highly under these metrics.

I would choose eigenvector centrality as the most relevant metric to identify who is running the illegal activities of the group. We can use eigenvector centrality to identify important advisors. Betweenness centrality will be used as a secondary metric. It can be used to identify important role players of the network.

Part (e):

In real life, the police need to effectively use all the information they have gathered, to identify who is responsible for running the illegal activities of the group. Armed with a qualitative understanding of the centrality metrics from Part (d) and the quantitative analysis from part Part (b) Question 5, integrate and interpret the information you have to identify which players were most central (or important) to the operation.

From Part (b) Question 5, We can observe that N1 and N3 rank highest on both the betweenness centrality and the eigenvector centrality. They have a high possibility of being members of the central leadership of the criminal group.
We also observed that n85 and n12 rank highly on eigenvector centrality and betweenness centrality, suggesting n85 and n12 acting important and distinct roles in

the criminal organization. With high score on betweenness centrality and relatively lower eigenvector centrality score, n12 is more likely acting as a member of sending orders, but has less direct contact with important members. On the other hand,n85 has a high score on eigenvector centrality and a low score on betweenness centrality. We can presume that n85 connects directly with important members in the criminal group, but does not have to give orders to many members. N85 is possibly an advisor.

Part (f) Question 2:

The change in the network from Phase X to X+1 coincides with a major event that took place during the actual investigation. Identify the event and explain how the change in centrality rankings and visual patterns, observed in the network plots above, relates to said event.

The first police seizure occurred in phase 4. It is reasonable to presume that during this time, the criminal enterprise learned about possible ongoing police investigation for the first time. As a reaction to such possible police investigation, The criminal organization will make changes on its structure.

As we can observe from phase 5, n12 appeared for the first time. N12 had its own group that was distinct from the main group. This is suggesting that after the marijuana seizure in phase 4, the enterprise decided to enter another field-cocaine.
Between those phases, the betweenness centrality went down while the eigenvector centrality went up. It is showing a sign that the criminal organization became less hierarchical. One observation is that in phase5, N1 directly connects with more members now.

Part (g):

While centrality helps explain the evolution of every player's role individually, we need to explore the global trends and incidents in the story in order to understand the behavior of the criminal enterprise.

Describe the coarse pattern(s) you observe as the network evolves through the phases. Does the network evolution reflect the background story?

As we mentioned in Part(f), After phase 4, the structure of the criminal network had a significant change. At phase 5, n12 had its own distinct group of members who had no other connections other than n12. It is reasonable to guess that at this point, n1 is giving orders to n12 to operate its group. When it comes to phase 7 and phase 10, we can observe that the connection between the main enterprise and n12 is broken. A reasonable guess would be that they are hiding and protecting themselves. Every time before n12’s disconnection, We can see that there were large seizures that occurred such as phase 6 and phase 9.

From phase 6 to phase 11, due to police seizure, The criminal group was transiting its power and changing its structure. In phase 6, n3 is taking over the power and connects to more actors. In Phase 7, n76 is taking over more power by connecting to three more actors. In phase 8, n87 is taking over more power by connecting to 4 more actors. In phase 10, n37 is taking over more power by connecting to 6 more actors. In phase 11, We can observe that n37 is no longer inside the criminal network. In the beginning, all members of the criminal groups are taking orders passing from n1. N1 is connected to almost every actor. When it came to phase 11, n1 only maintained close connections to n 79,n76,n87 and n12.

Part (h):

Are there other actors that play an important role but are not on the list of investigation **(i.e., actors who are not among the 23 listed above)** ? List them, and explain why they are important.

In the last two phases, n37 and n41 took over the key roles. However, n37 dropped out of the criminal network in the last phase. Their late appearance might be the reason for them to escape from police investigation.

Part (i):

What are the advantages of looking at the directed version vs. undirected version of the criminal network?

When using an undirected graph, It is not hard to find who is acting as the center of the criminal network. It is obvious that actors in the center of a criminal network are acting as a key role, but we are unable to identify what this central actor is doing. We do not know from the undirected graph that if this actor is someone giving orders to others criminals.

The directed graph can give us more information. By studying the difference between incoming edges and outgoing edges (in-degree and out-degree), we can filter out unimportant actors and identify important actors. In this criminal network, if an actor has high in-degree and low out-degree, that actor is more likely someone who is receiving information. On the contrary, that actor is more likely someone giving out information and orders.

Part (j):

*Recall the definition of hubs and authorities. Compute the hub and authority score of each actor, and for each phase. (**Remember** to load the adjacency data again this time using* `create_using = nx.DiGraph()`*.)*

*With **networkx** you can use the* `nx.algorithms.link_analysis.hits` *function, set* `max_iter=1000000` *for best results.*

*Using this, what relevant observations can you make on how the relationship between **n1** and **n3** evolves over the phases. Can you make comparisons to your results in Part (g)?*

N1 has a very high hub score in the first few phases. It is reasonable to presume that he is the major commander of the enterprise during this period of time. When it came to phase 6 and phase 7, the hub score for n1 decreased by two orders of magnitude while the hub score for n3 increased by a significant margin. This is suggesting that there was an ongoing authority transition during this period of time. During phase 8 and 9, we observed a sudden increase in the hud score of n1.In Phase 10, the hud score for n1 dropped by a significant margin. This allies with what we observed in part(f) that n1 is stepping back from the central administration and the enterprise starting shifting to a more compartmentalized model.

Thanks to more information directed graphs provided. We can observe that, from phase1 to phase5, n1 has a lower authority score than n3, meaning the n1 has a weaker strength of incoming links compared to n3. It is reasonable to guess that n3 was taking over more coordination roles during this period of time.

In phase6 and phase 7, we observed another change in authority score, meaning another transition in authority roles. N1 had a higher authority score than n3 during this time.The largest police seizure todate happened in phase7. It is interesting to see that n1 and n3 immediately switch back their roles during phase 8 and phase 9. When it comes to phase 10 and phase 11. N1 and n3 came to the back stage. This allies with what we observed in part(f) that n1 is stepping back from the central administration and the enterprise starting shifting to a more compartmentalized model.

Additional, optional, observations:

N76, N12's hub score and authority score have opposite patterns. N76 has a higher authority score than hubscore. N12 has the opposite pattern as that of N76. All their hub scores are lower than that of n1 and n3. It is suggesting that they are more likely mid-rank managers, but N12, with higher authority score, has more independence than N76.

Project:

Social circles: Facebook

Goal:

This is a homework assignment for the online course "MITx 6.419: Data Analysis: Statistical Modeling and Computation in Applications" that focuses on the application of network analysis techniques. The assignment is an open-ended project that requires me to choose a sociologically relevant research question about either the CAVIAR network or the Facebook or Twitter network from the recitation notebook section.The data used for this assignment can be obtained from publicly available sources such as the Stanford Large Network Dataset Collection (https://snap.stanford.edu/data/index.html) where the Facebook and Twitter ego-networks are available.

I have chosen to investigate the network structure of Facebook friendship connections by utilizing network graph visualization techniques. The primary objective of this analysis is to verify the strength of weak ties theory through empirical analysis. According to this theory, having many weak connections can be more beneficial than having a few strong ones. This project is particularly interesting and useful to me as it allows me to gain insights into the social dynamics of online networks.

In addition, I am aware that network analysis techniques are often utilized in criminology to investigate time-varying criminal networks. I plan to apply similar techniques to study the network structure of connected clients with substance-use disorders. By analyzing the patterns of connections between individuals, I hope to gain a better understanding of the factors that contribute to substance-use disorders and identify potential intervention points for prevention and treatment.

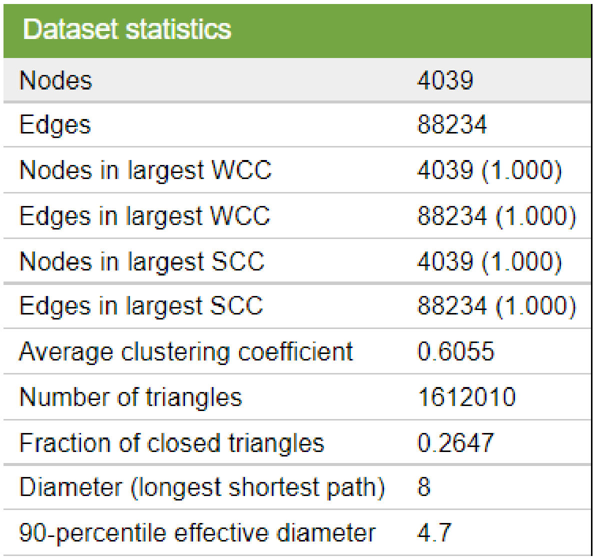
Methodology:

methodology for network analysis

1. Betweenness Centrality
2. Degree Centrality

3. Closeness Centrality
4. Community Detection

Datasets:

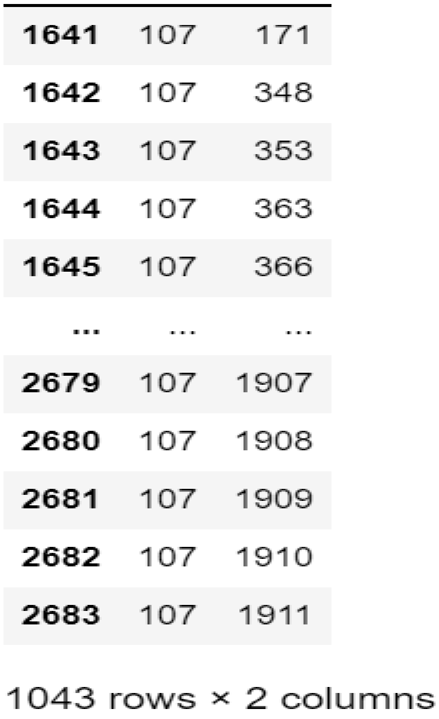


Dataset statistics	
Nodes	4039
Edges	88234
Nodes in largest WCC	4039 (1.000)
Edges in largest WCC	88234 (1.000)
Nodes in largest SCC	4039 (1.000)
Edges in largest SCC	88234 (1.000)
Average clustering coefficient	0.6055
Number of triangles	1612010
Fraction of closed triangles	0.2647
Diameter (longest shortest path)	8
90-percentile effective diameter	4.7

1. Data Cleaning and Wrangling

Step1: Find out the user that has the most connections(most friends)

User 107 has the most connection with others. User 107 has 1043 unique friends. We will then use user 107 as the central user.



1641	107	171
1642	107	348
1643	107	353
1644	107	363
1645	107	366
...	...	...
2679	107	1907
2680	107	1908
2681	107	1909
2682	107	1910
2683	107	1911

1043 rows × 2 columns

Step2: Creating a dataframe that contains information that stores information of how many mutual friends that user 107's friends have with user 107. We will use this dataframe to create our network graph.

2. Building Network Graph

In this question, I will only test two scenarios: At least 2 mutual friends and at least 30 mutual friends.

Only keep friends with at least 2 common friends

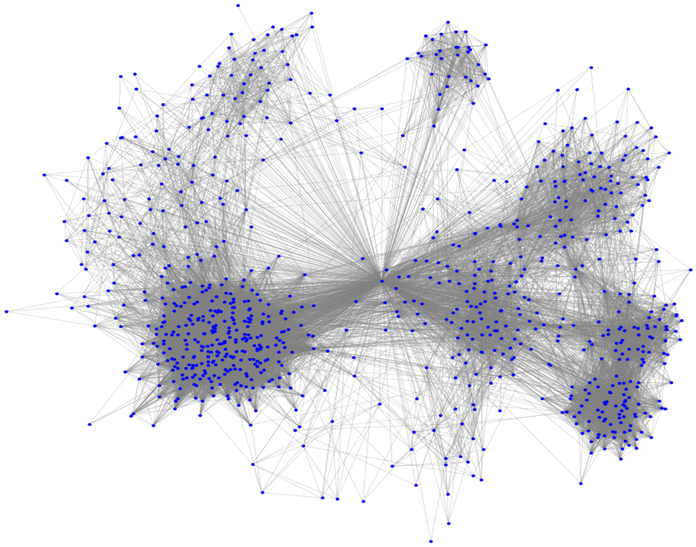
There are 877 friends with at least 2 common friends.

network graph:

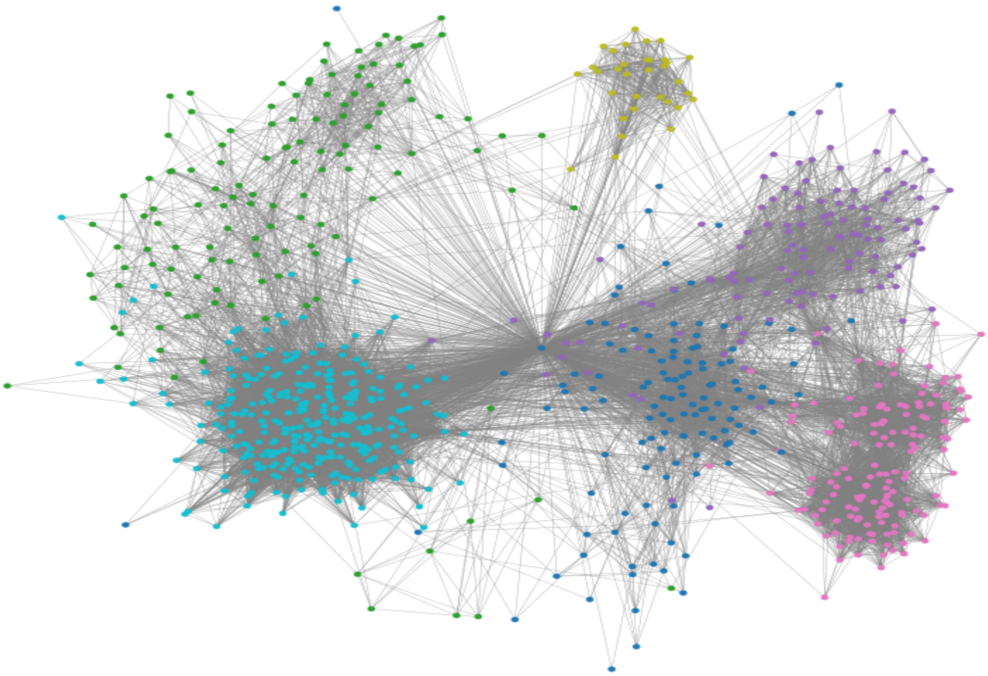
The blue nodes below are friends of the central user(user 107), and those gray lines are ties of friendship.

The visualization below used a very cool funcion called force-directed layout function to calculate the position of each node. Generally, The more mutual connections nodes share, The closer those nodes will cluster. Also different clusters of nodes will be pulled away from each other.

From this step, we can observe that there are approximately 7 communities of friends. However, we need further study to find out.



Next, what we do is straightforward, we will color different communities of friends.



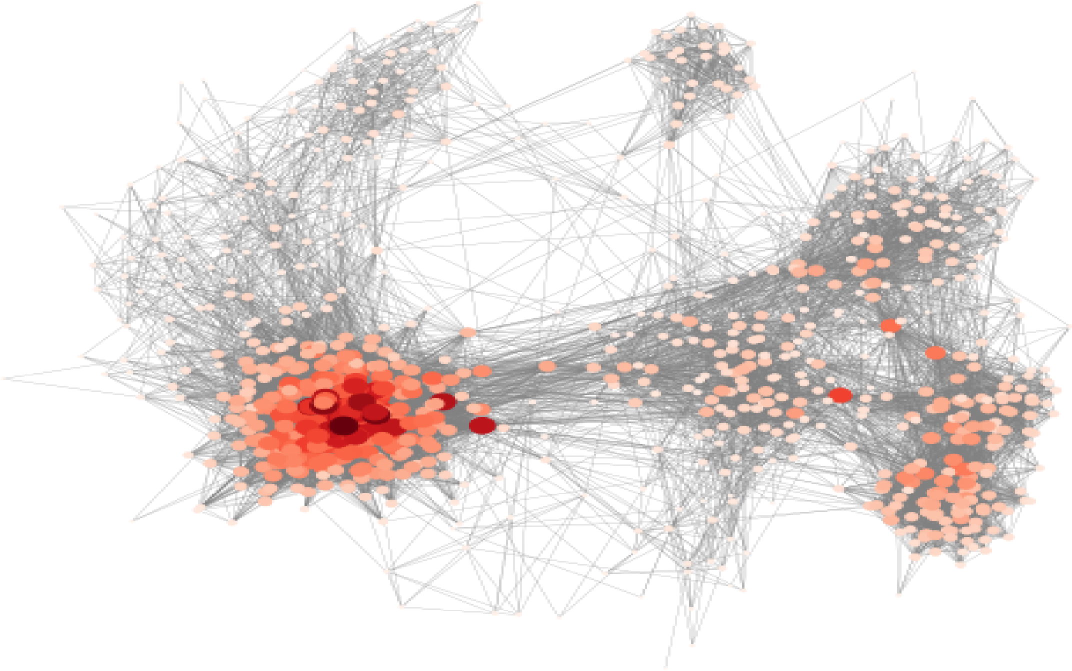
What we observed is that there are actually 6 communities of friends. We can observe that some communities are tangled together(next to each other), while there are some communities that are alienated with each other. One reasonable guess would be that communities of left and right represent friends user 107 knows during different periods of time.

3. Centrality:

Here we will use more mathematical ways we are taught in the class to do a more precise and professional study.

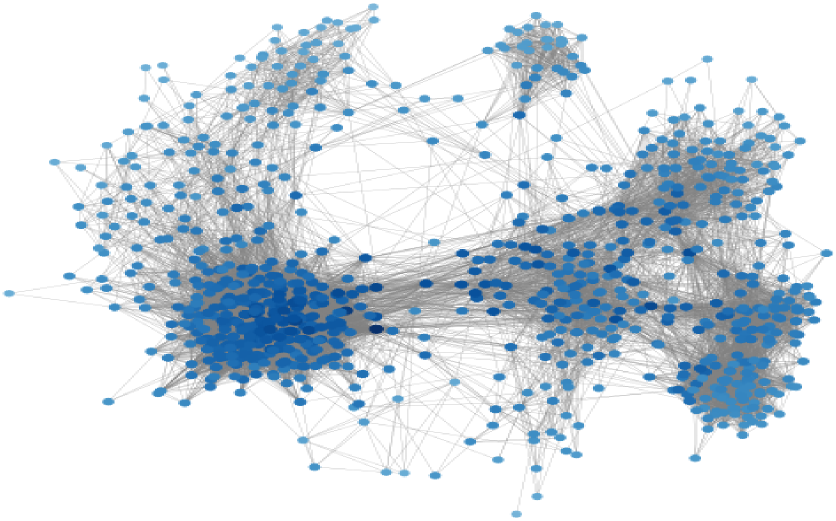
Degree Centrality:

The first method we will use is the one degree centrality. We moved the central user 107 from the graph. This visualization is highlighting friends who have the most mutual friends with user 107.



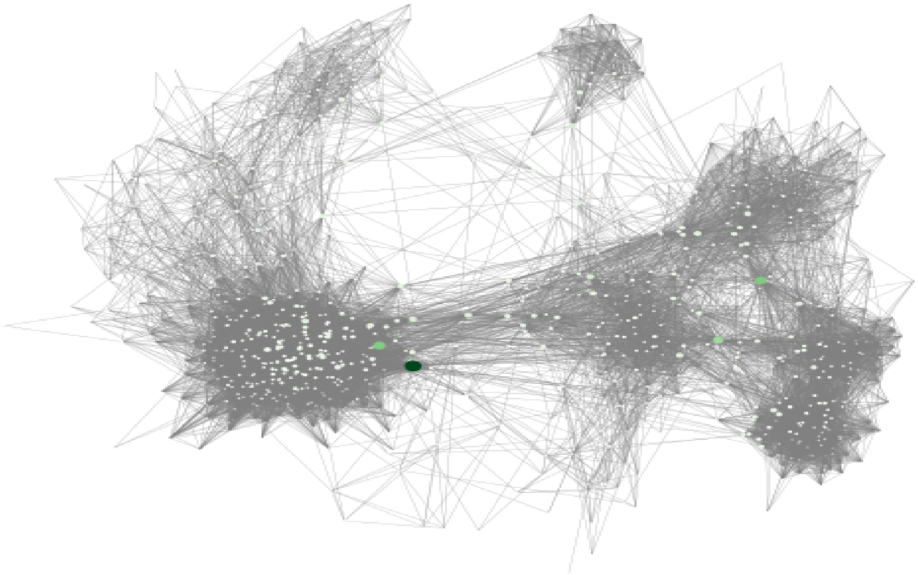
Closeness Centrality:

The visualization below used closeness centrality. Being an inverse metric of degree centrality, closeness centrality measures the sum of shortest path to other nodes.



Betweenness Centrality:

Betweenness centrality measures frequency of a node being located on the shortest path between two other nodes.

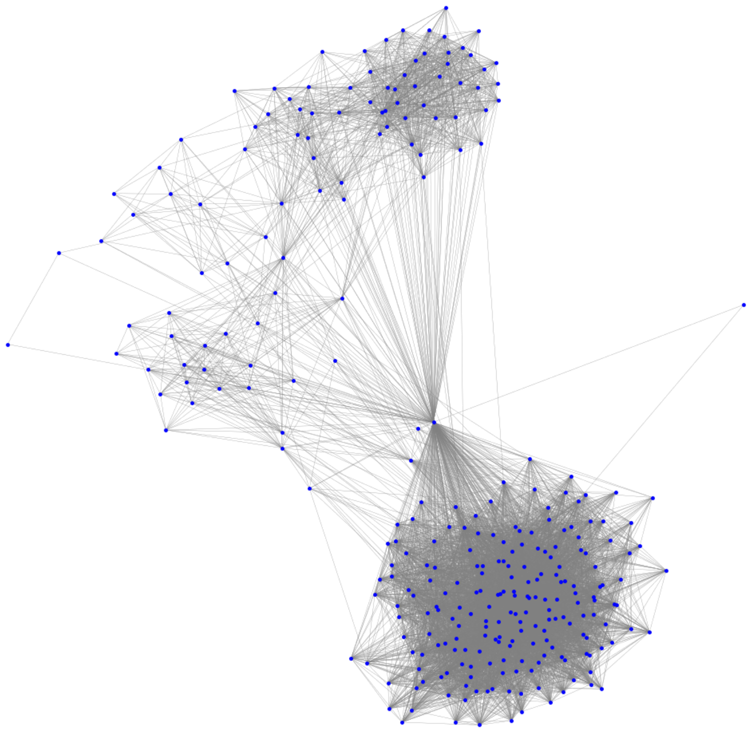


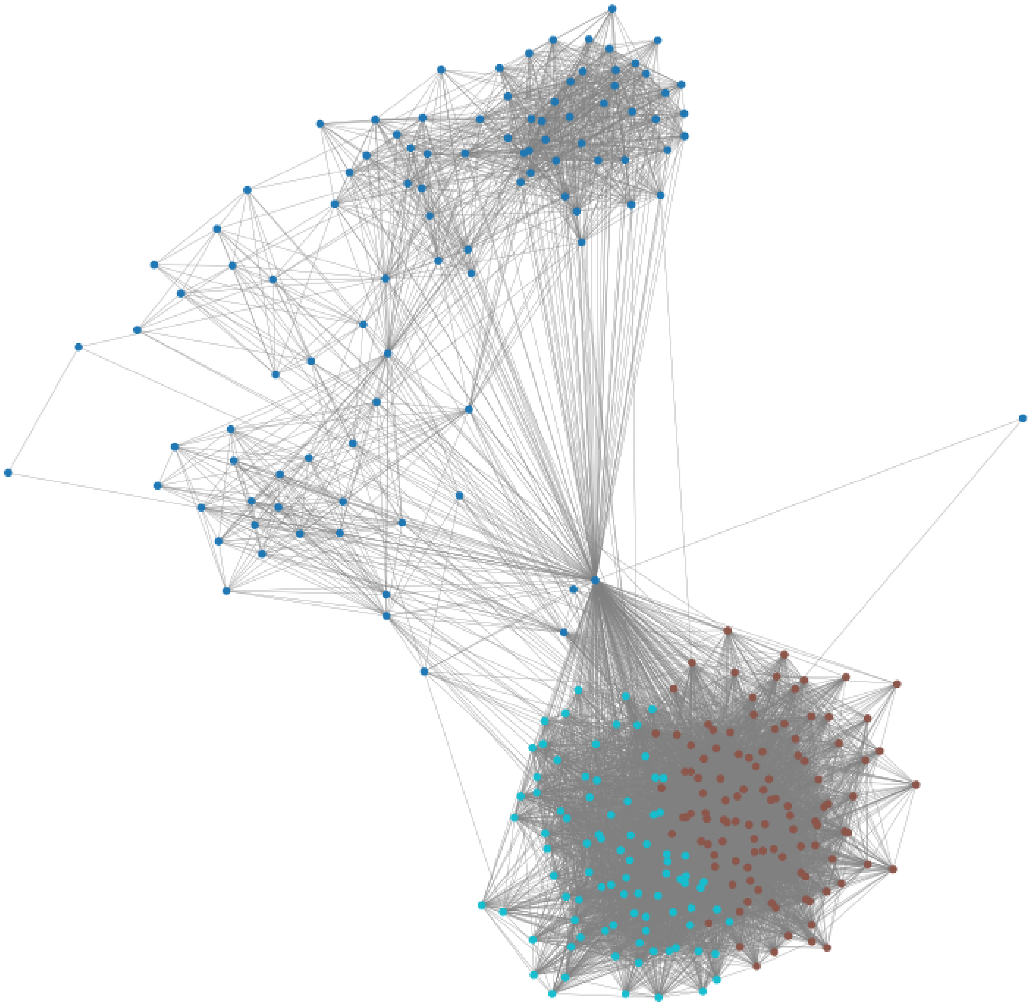
Only keep friends with at least 30 common friends

There are 877 friends with at least 30 common friends.

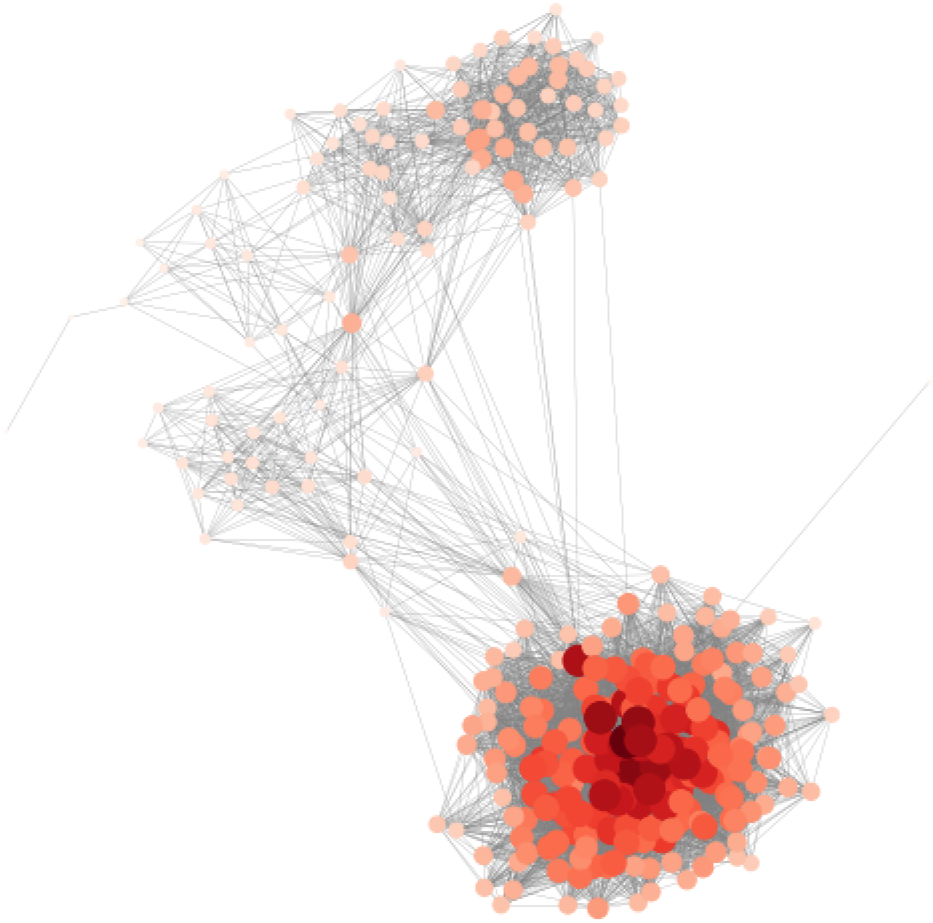
We will use the same procedure for this section.

Network Graph:

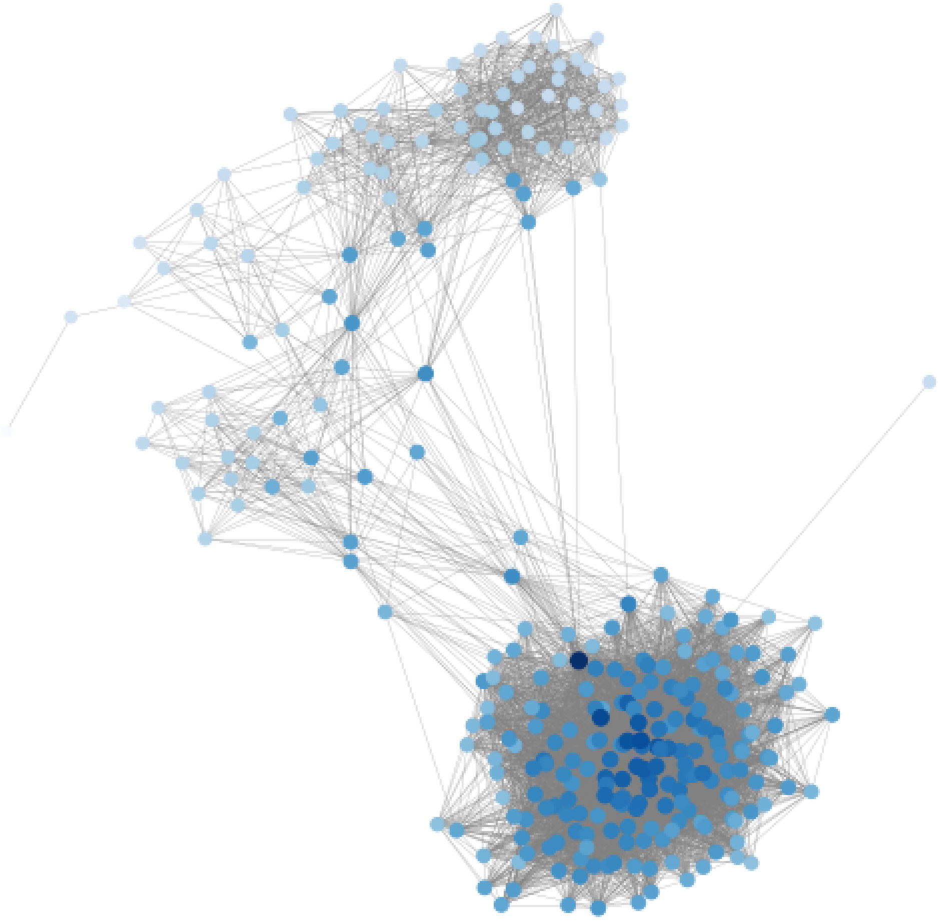




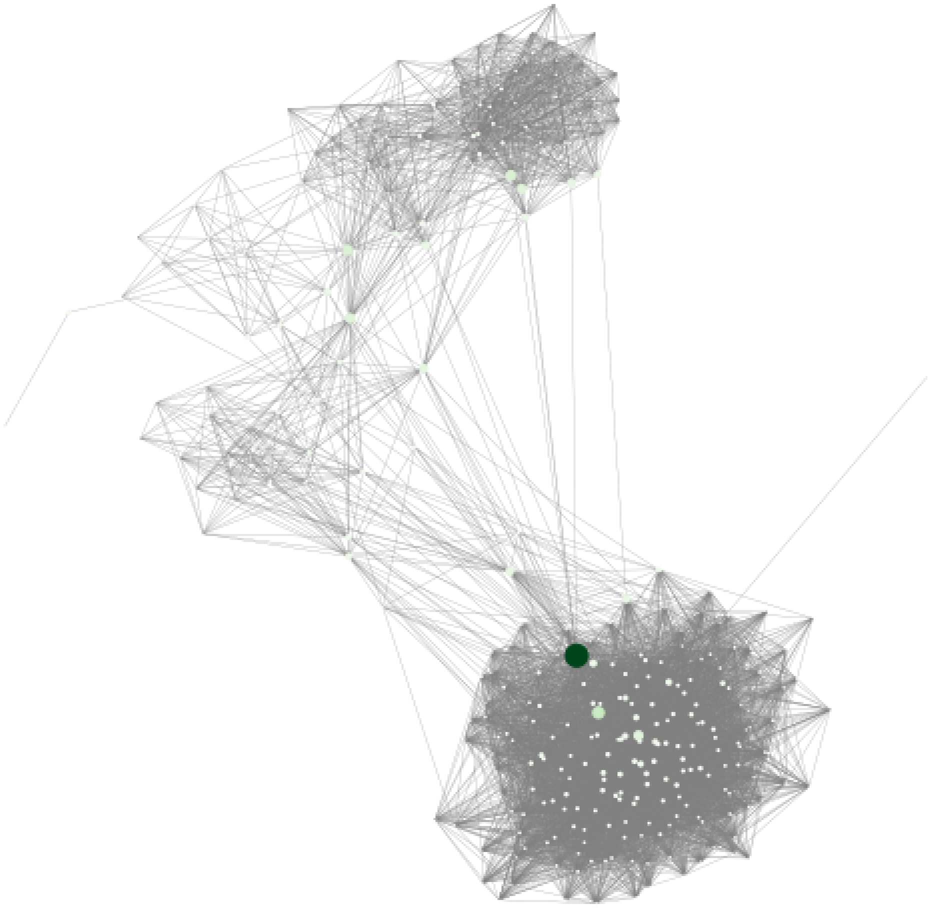
Degree Centrality:



Closeness Centrality



Betweenness Centrality:



Conclusion: By using weak ties, We can maintain more communities of friends. Enlarge the scope of friends.