**Written Report - MITx 6.419x – Module 3**

**Name: Matt7Salomon**

**2. Problem 1: Suggesting Similar Papers**

Part (c): (2 points) Include your answer to this question in your written report. (100 word limit.)

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

**Solution:** Matrix multiplication algorithms generally have a time complexity between O(n^2) and O(n^3). However, usually multiplication algorithms in practice tend to have a time complexity of O(n^3). The more efficient algorithms usually cant work on day-to-day computers because they either need a high memory requirement or they have large constant multipliers which makes them slower then O(n^3). Therefore, in practice, the time complexity of the multiplication algorithm is not better than the friend’s algorithm even when using optimizations for matrix sparsity.

Part (d): (3 points) Include your answer to this question in your written report. (200 word limit.)

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?

**Solution**: Bibliographic coupling is a superior metric. It gives a convincing argument for either metric, but does not recognize that papers include a wide variety of references.

### Part (a) Preliminary Analysis

First, we build a network using the given adjacency matrices using the network package.

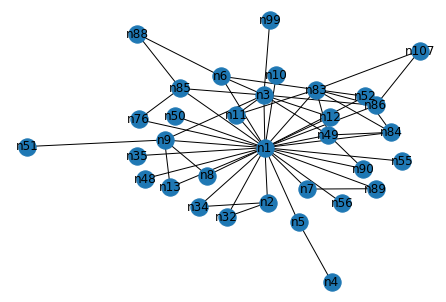
**Question 1**

Chart, line chart

Description automatically generated

**Question 2**

Visualizing the graph at each phase. For networkx we can use nx.draw(g,pos=nx.drawing.nx\_agraph.graphviz\_layout(g),with\_labels=True



Solution: This corresponds to choice (d)

# Part (b) Centrality Measures

We now compute degree centrality, betweenness centrality and eigenvector centrality for the 21 actors across all phases

**Part(b) Question 1**

**Graphical user interface, application, table, Excel

Description automatically generated**

**Part(c) Question 2**

**Graphical user interface, application, table, Excel

Description automatically generated**

**Part(b) Question 3**

**Graphical user interface, application

Description automatically generated**

**Part(b) Question 4**

**Through this experience, we conclude that the fastest algorithm is the degree centrality.**

The time complexity of this algorithm is O(n^2). On the other hand, the time complexities of betweenness and eigenvector centralities are O(n^3)

**Part(b) Question 5**

**Now, we calculate the mean and variance of every column. Based on the network convention, the values are already normalized. For inactive phases of all players, we replace the null values with zero. This helps with comparing the players more efficient. To help understand the data,** we also calculate statistics like standard deviation, variance and a player's activity.

Highest Mean Centrality :

Degree -

index Mean Std. dev Variance Activity

0 n1 0.601485 0.240848 0.058008 11.0

4 n3 0.223505 0.150110 0.022533 11.0

25 n12 0.170893 0.125645 0.015787 10.0

7 n85 0.118010 0.049330 0.002433 11.0

22 n76 0.112235 0.073632 0.005422 10.0

3 n83 0.095836 0.076965 0.005924 11.0

Betweenness -

index Mean Std. dev Variance Activity

0 n1 0.655051 0.238535 0.056899 11.0

25 n12 0.167562 0.187327 0.035092 10.0

4 n3 0.129403 0.179745 0.032308 11.0

22 n76 0.083791 0.099159 0.009833 10.0

54 n87 0.061327 0.092958 0.008641 6.0

84 n41 0.050369 0.167055 0.027907 3.0

Eigenvector -

index Mean Std. dev Variance Activity

0 n1 0.546391 0.117044 0.013699 11.0

4 n3 0.298095 0.124629 0.015533 11.0

7 n85 0.190612 0.065860 0.004338 11.0

22 n76 0.165877 0.080282 0.006445 10.0

3 n83 0.153522 0.083390 0.006954 11.0

14 n8 0.152394 0.067531 0.004560 10.0

Lowest Mean Centrality :

Degree -

index Mean Std. dev Variance Activity

70 n39 0.002217 0.007354 0.000054 1.0

67 n67 0.002217 0.007354 0.000054 1.0

73 n23 0.002217 0.007354 0.000054 1.0

Betweenness -

index Mean Std. dev Variance Activity

53 n77 0.0 0.0 0.0 3.0

73 n23 0.0 0.0 0.0 1.0

71 n59 0.0 0.0 0.0 3.0

Eigenvector -

index Mean Std. dev Variance Activity

88 n103 4.639981e-15 1.538908e-14 2.368237e-28 1.0

89 n104 4.639981e-15 1.538908e-14 2.368237e-28 1.0

64 n75 4.356416e-09 1.444860e-08 2.087619e-16 1.0

# Part (c)

See previously generated plot in Part (a) Question 1 and the provided written solutions.

Part (c): (2 points) Include your answer to this question in your written report. (100 words, 200 word limit.)

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?

**Solution**: Until phase 4, we did not see any seizures. Therefore, we can conclude that the police were probably using wiretaps to grow the suspect list in an early phase. Hence, if a player is not available in an early phase, we can conclude that the police has not started wiretapping them. Although we already know that the police wiretap does not have any effect on criminal enterprise which means their mean score has been biased down. Therefore, we can remove the early phases from the calculation in part(b) question 5.

# Part (d)

Part (d): (5 points) Include your answer to this question in your written report. (300 words, 400 word limit.)

In the context of criminal networks, what would each of these metrics (including degree betweenness, and eigenvector centrality) 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.

**Solution**:

We know that degree centrality measures the number of connected players to a person. Since this is a hierarchical organization, we can conclude that a person is usually receiving orders from just one other person. Therefore, the degree measures the number of people they give order to. However, this is not a good method to recognize a leader. For example, the main leader may only give orders to one person and lead through him which means they would rank low on this type of centrality. Also, a key courier may transmit information only between two people which causes them not to rank high using this type of metric. On the other hand, betweenness centrality shows if a person is a key network link by showing if a person’s removal from the criminal network might break the network. Therefore, if a person scores high on betweenness centrality, we can conclude that their presence is critical to the information distribution in the criminal network. In addition, eigenvector centrality shows the importance of a person based on their connected players. As an example, an important advisor may not have a high score on degree centrality and betweenness centrality even though they are connected to an important leader. However, they will have a high score in eigenvector centrality as seen in Part(b) Question 5. N12 has a high rank which means they are an important advisor. Also, N12 is a leader and serves as an critical information distributor. The eigenvector centrality is the best algorithm to show the important players in the network since it identifies the key advisors and important leaders that are hidden behind a lieutenant. Also, the betweenness centrality may be used as a secondary metric, since it shows important information. For example, if the betweenness centrality targets an unimportant player as an information carrier, we can target them for surveillance.

# Part (e)

See previously generated plot in Part (b) Question 5 and Part (d) and the provided written solutions.

Part (e): (3 points) Include your answer to this question in your written report. (100 words,~ 200 word limit)

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.

**Solution**: The central leaders are n1 and n3 since these players have high betweenness centrality and eigenvector centrality scores. Therefore, we can consider them as belonging to the central leadership in the criminal operation. Additionally, n12, n85, n76, n12 have a high rank of betweenness centrality and n85 has a high rank of eigenvector centrality which means they are key players as well. We can conclude that n12 is likely distributing orders and coordinating subordinates in the operation. Also, since n12 has a lower eigenvector centrality which means that they do not deal with key actors. The high eigenvector centrality and low betweenness suggest that this player acts like n1 and n3. Also, according to the police records, n85 is not a trafficker, however the police would likely investigate this player due to high degree of centrality.

# Part (f)

We can now attempt to analyze the evolution of the network and correlate the patterns we observe to events that happened during the investigation.

Identify X using your visualization in Part (a) Question 2.

**Diagram

Description automatically generated**

Answer: We see that X=4

**Question 2**

Part (f) Question 2: (3 points) Include your answer to this question in your written report. (~200 words, 300 word limit.)

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.

**Solution**: The first police seizure occurred in phase 4. Therefore, we assume that phase 4 was the first time that the enterprise learned there will be a police investigation on their case, thus, potentially initiating large changes in the organization. For example, we see that n12 appeared in phase 5 for the first time maintaining a separate group from the rest of the enterprise. The police identified n12 as a main player for a cocaine operation which means the operation was diversified after phase 4 in at least one more market. If we check the betweenness centrality, we can see that the metrics went down for most players in these phases while the eigenvector centrality metric went up. Therefore, we can conclude that the organization became less hierarchical. The visualization shows that except for n12, most players are connected to n1 meaning that n1 has trusted the rest of the players less and less and decided to be in charge of the operation more and more after this change.

# Part (g)

Part (g): (4 points) Include your answer to this question in your written report. (200 words, 300 word limit.)

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?

**Solution**:

As mentioned in the previous section, starting from phase 4, the criminal network had a major change in direction. For example, N12 gains a major number of players which for these players, N12 remains the only point of contact. This means that n1’s influence gradually moved to N12. Also, in phase 7 and phase 10, N12 is getting disconnected from N1 creating a new network and disconnecting from the main enterprise. When we consider the police activity, we find out that some major seizures occurred at phase 6 and phase 9 which were before N12 disconnection. At phase 6 and afterwards, n3 starts gaining authority in the enterprise and at phase 7, n76 gains authority of six players. In phase 8, n87 gains authority over four players. In phase 10, n37 gains authority over six players. In phase 3 n1 was connected to almost every player but by phase 11, n1 has delegated most of his managed players to other actors due to the police activity. In phase 11, n1 is only connected to a small number of players. N37 apparently has left the enterprise in the later phases and an axis of control has emerged around n79, n76, n41, n87, and n12.

# Part (h)

Part (h): (2 points) Include your answer to this question in your written report. (50 words, 100 word limit.)

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.

**Solution:** Those actors are n37 and n41 which take critical roles in the last two phases of the act. These two players evaded the police activity since they entered the operation late and after most police activity has already occurred. It should be mentioned that n37 leaves the operation in the last phase.

# Part (i)

Part (i): (2 points) Include your answer to this question in your written report. (150 words, 250 word limit.)

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

**Solution:** The directed graph gives more information about the studied problem. For example, we can’t tell the relationship between actors when an actor gains authority over the other actors by simply looking at an undirected graph. Further, directed graphs reveal the flow of information which helps us identify the unimportant players who only receive the information. We can also identify the important actors who distribute the information and control the enterprise. The important actors usually have a more balance in-degree and out-degree and the unimportant actors have a high in-degree and low out-degree in the network. Similarly, we can use the eigenvector centralities to gain the same information. The eigenvector centrality also measures actor importance.

# Part (j)

Part (j): (4 points) Include your answer to this question in your written report. (300 words, 400 word limit)

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().)

**Solution:** In the early phases, n1 appears to be the major commander in the enterprise with a high score. However, the hub score of n1 decreases in phase 6 and phase 7 by two orders of magnitude. In the same period, the hub score of n3 increases by a high margin showing that n1 delegates commandership to n3 in these phases. We can also see that for a short time, n1 regains control of the enterprise with a high hub score in phase 8 and phase 9. The hub score of n1 drops by many orders of magnitude in phase 10 though which confirms the observations in part(f) that n1 withdrew from commandership in phase 10 and the leadership was delegated to other actors in the enterprise. We know that the authority score is based on the strength of incoming links. In phase 1 through 5, n3 has a higher authority than n1 meaning that n3 had the main coordination role in this time although its not explicitly noticeable in the undirected graph’s network structure. However, in phase 6 and phase 7, n1 achieves a higher authority than n3 due to a trade in authority in this phase between the two actors. Shortly after in phase 7, the largest police seizure occurs which causes n1 and n3 to conclude that the trade in role was not optimal, thus, they return back to their previous roles in phase 8 and phase 9. In phase 10 and phase 11, n1 and n3 delegate their power to other users in the enterprise. Some notable information from the network is that n12 has a higher hub score than authority score. N76 has a higher authority score than hub score showing that they take orders from another important player. Therefore, although in part(f) we concluded that this player is highly connected, they likely are a middle manager. According to the police reports, n76 is managing the transportation of marijuana. N85 has a higher authority score than hub score which means they receive information from others and coordinate and advise. Police records show that n85 is an accountant.

### Additional Analysis

Here is a scatter plot of mean v/s *activity* for each of the centrality measures.

**Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated**

**Visualizing centrality:**

**Diagram

Description automatically generatedDiagram, radar chart

Description automatically generatedDiagram

Description automatically generated**

### Other Characteristics

In order to interpret the data, we can also plot other relevant statistics such as a centrality histogram plot, a centrality rank plot etc.

**Chart, histogram

Description automatically generated**

### Visualizing Directed Graphs

A sample visualization is shown below for the directed criminal network in phase 1.

**A picture containing accessory

Description automatically generated**

### Other Characteristics[¶](http://localhost:8888/notebooks/solution_notebooks_caviar_network_solutions_allquestions.ipynb#Other-Characteristics)

In order to interpret the data, we can also plot other relevant statistics such as a centrality histogram plot, a centrality rank plot etc.

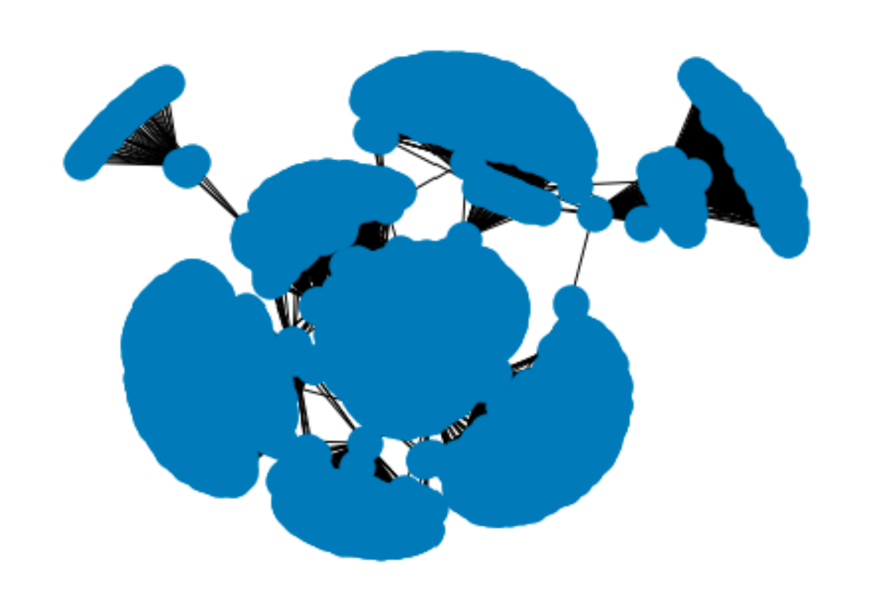
**Chart, bar chart

Description automatically generated**

**Project**

The goal of this project is to evaluate connecting community’s role of a central social network. A central node is generally an important person, a lead, an influencer, or head of a cult. We can study the question of if the relationship is one-sided or if the connections directly communicate with each other. Also, additional roles of a central node can be investigated based on the connecting nodes. For example, if a center node is a criminal, the surrounding nodes are generally criminals as well and they are connected around crime. In this project, we will study the Facebook dataset and how the nodes are connected to each other. The Facebook dataset forms a large network; therefore, it is computationally heavy to study. We will. Minimize the network size to be able to investigate it while maintaining the characteristics of the network. We will do this by visualizing the network and choosing a fraction of the network. Then, we will select the significant influencers of the network by using centrality metrics. We will apply clustering to find smaller networks as well. Further, we will run a graphical model to determine if we can create a community without influencer’s impact.

The Facebook dataset corresponds to a large network with 3663 nodes and 53498 edges. We can see from the following figure that the network forms various communities. These communities could be formed based on different features such as gender, nationality, interests, location, nationality, etc. However, we cannot determine based on the given dataset, which features are included in the study.



We can see from the figure that the network is very large and might be too large for a computer analysis even though we can visualize it. Therefore, we need to sample the dataset for the analysis. If we don’t sample, we may not be able to analyze the dataset using a simple computer.

One method of choosing a smaller subset is connected component. However, the results show that the network is only one connected component. Another method is to select one community to study. We will pick a dense cluster to get the most out of the experiment using a centrality metric: degree and betweenness. The results which are shown in a table demonstrate that 107 is the most connected node. Therefore, we will choose a network incorporating 107. The smaller subset has 333 nodes and 5038 edges.

The following table shows the top 10 highest scoring node based on degree centrality.

Table

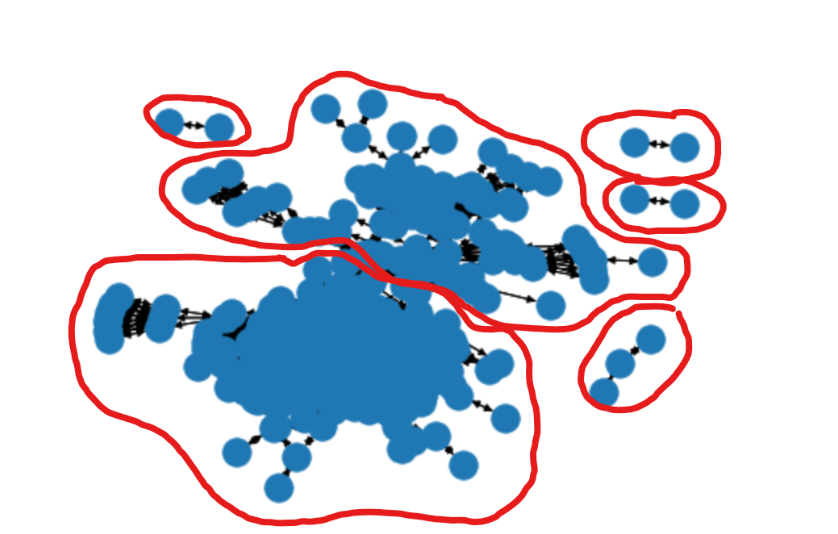
Description automatically generated

The following table shows the top 10 highest scoring node based on degree centrality

Table

Description automatically generated

The following figure shows the potential samples of the dataset. From these potential subnetworks, we will choose one containing 107.



Next, we will find the influencers in this network. Influencers are nodes that have impact on the connected nodes. Therefore, we will choose the betweenness centrality as a metric. We cannot choose metrics that are based on directed graph since we have an undirected graph in this study. Further, we cannot choose degree centrality due to its simplicity and closeness. Influencers are not necessarily close to their followers. In addition, influencers help create a community. Thus, we choose betweenness.

We choose top 1% betweenness centrality weight since their weights are greater than other nodes. These nodes have weights of 0.266 and 0.246 while most of the other nodes have a weight of less than 0.05. These nodes include: 277,175,19, and 23.

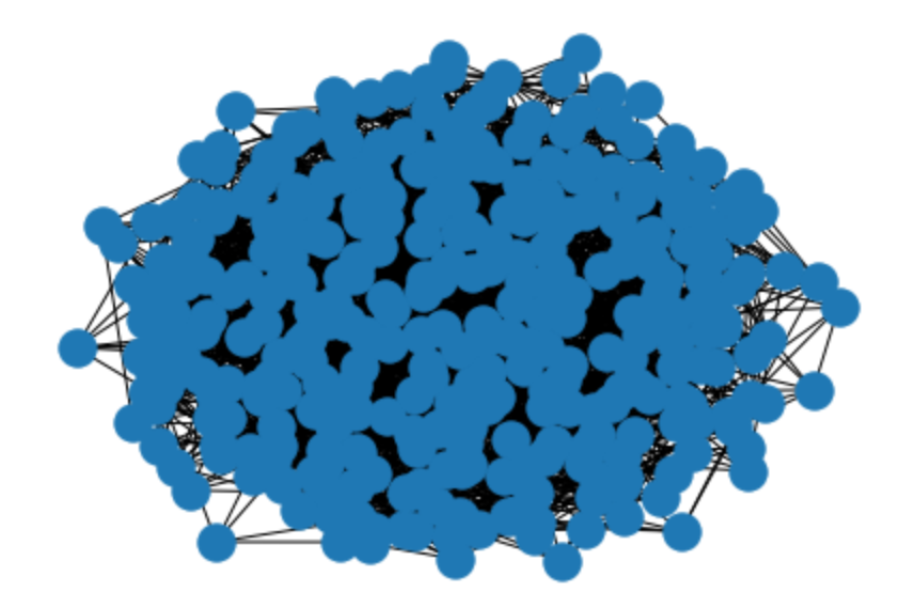
Further, we run a community detection algorithm with this new graph. The community detection algorithm detects subsets of nodes that are densely connected. The results are visualized in the following figure.

Chart, scatter chart

Description automatically generated

We can see from the plot that influencers had a major effect in engaging people in the Facebook network. Visually, a network without influencers is much sparser than the network with influencers. Although a visualization is simply not a great proof and a quantitative analytics method on a more complete dataset needs to be performed to sufficiently prove this conjecture. We cannot draw a better conclusion with the type of the data provided.

The studied sample has 333 nodes and 5038 edges which means that the number of possible edges are 110889. Therefore, the probability for a graph to occur is 0.045. Hence, if we use an Erdos-Renyi model, we can get a connected component graph as follows.



We can see from our study that we have some influencers in our network which was shown with the centrality weights. Without the influencers, we might have a major disconnection within the network. The information provided from the dataset makes it impossible to investigate further. However, further analysis is required to study the effect of influencers in our network. By comparing the two figures, we can conclude that the random generated model is connected but seem to be sparser than the actual network. This suggests that the users can be very connected inside a group. Also, there are some society parts who are separated from the rest of the society. Therefore, Erdos-Renyi is not generalizable to a human connection network.