## Problem 1: Suggesting Similar Papers

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

*ANSWER*: The time complexity is also similar to N^3 but because the practical implementation is more cache aware of the naïve implementation, it is actually faster than the naïve implementation. It does not approach N^2. The algorithm, Lapack’s DGEMM, has a polynomial between 2 and 3 and thus has better performance than the runtime complexity of the naïve matrix multiplication. Although there is a constant also attached to the runtime complexity.

**2 (d):** Bibliographic coupling and co-citation 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?

*ANSWER*: They give noticeably different results because they are completely different measures. One is measuring the papers that are listed in the same bibliography (bibliographic) and the other is measuring which papers have the same cited the same papers (co-citation).

Co-citation could be considered a more up-to-date measurement because future works can cite a paper in the future and the co-citation score (also called the co-citation proximity index) for two documents could continue to grow, while the bibliographic coupling would stay the same as those are fixed.

**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?

*ANSWER*: It could be because the operation was not fully baked yet, there was not all the characters involved yet. I think the answer in Part (b) Question 5 should be adjusted somehow because there could be players that have high centrality that are just not introduced yet during the introductory stages. Perhaps the default value should not be zero when a player is not in a phase.

**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.

*ANSWER*: Degree centrality might not measure the fact that an important person actually limits contact with other plyers in the drug game, thus limiting their degrees and their centrality. A middleman might have the highest centrality but might not be nearly as important as the kingpin. If a kingpin is disciplined about interactions, then they would have a much smaller number of connections. Similarly, if a key player in the logistics or money-washing arm of an operation is disciplined, they might limit their in-degrees as well. They do not want many people to connect to them and have many opportunities for law enforcement to have them on wiretaps.

Betweenness centrality might be a better metric because people will be linking other sub components of the graph, yet they won’t be directly linked to every node because that would be too much exposure for someone with w a lot of power and risk.

Eigenvector centrality might be the best measure because it takes into account the centrality of the neighbors as well. In real networks this is often the case that high centrality nodes connect with high centrality nodes.

Also, degree centrality does not take into consideration the outflow of communication, if it is just all degrees and not specifically out-degrees. So making the graph directed would add another dimension of information onto the undirected graph.

**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.

*ANSWER*: I think the most important people could have a high degree centrality, especially for degree centrality would be in the earlier phases of the operation. In the later phases of the operation, there would be betweenness centrality or eigen vector centrality that would be a better judge of who is important. Because in a real network like this, the actors actually limit their degrees purposefully. There is not a need for direct contact with all of the other nodes.

N76 and N41 look like they might be most important other than N1, because they actually connect all of the other nodes to each other. Meaning without them, there would be no large network. This can be seen with multiple centrality measures in later stages.

**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.

*ANSWER*: The major event was the seizure of the marijuana after the Phase 4. The seizure was $2,500,000 and 300 kg of marijuana.

**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?

*ANSWER*: There seems to be a pattern of where some networks will go a particular phase without communicating with the main network (the network with the main actor, Daniel Serero). However, they will do internal communication for that phase. For example, in phase 7, Ernesto Morales was doing a lot of communicating with other players in his own star-shaped network. There was link from that sub-network to the main network though. In the next phase though, phase 8, there was a link from the main network to the sub-network with Ernesto Morales.

Ernesto Morales is the principal organizer of the cocaine imports, so it makes sense that there would be more communication between him and Serero during a cocaine import with a lot of cocaine. After Phase 8, there actually was a seizure of 9 kg of cocaine. This pattern repeated itself a few times over the course fo the investigation. Every time there was a phase which led to a cocaine seizure, that graph for that phase would have connections between sub-graphs of Serero and Morales.

Also, during a phase that led to a large cocaine seizure, there was a much higher out-degree coming from Morales, even without the connection to Serero. This makes sense because Morales probably had to do a lot more coordination and synchronization of distribution, before all the cocaine was seized by the government.

**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.

*ANSWER*: There are quite a few actors who have high importance that are not listed in the above chosen actors. Actors n10, n105, n13, n14, n15, n,19, n2, n27, n31, n37, n41, n46, n49, n52, n64, n81, n9, n90, n93, n98 all had eigenvector centrality measurements in the top 10 for at least one of the phases. Actors n10, n13, n19, n27, n31, n37, n41, n49, n81, and n9 all had eigenvector centrality measurements in the top 5 for at least one of the phases. And n41 actually was the second highest eigenvector centrality for the 11th phase.

**Part (i):** What are the advantages of looking at the directed version vs. undirected version of the criminal network?

*ANSWER*: The directed will give you who is in contacting the other nodes/actors. So by getting the outdegree of the node in question you can better predict the verdict classifications (as was said in the study). The out-degree centrality was formally shown to be a very good indicator of whether or not a person was found guilty in their trial.

Having the graph be directed vs undirected just adds another dimension of information to the overall analysis. Without the graph being directed, there is nothing really taken away from the graph, there is no information lost. And the out-degree has shown to be a good feature to have an a predictive model, so it is valuable information gained at that. There is lot of information, like for example finding cliques, that is a little different, and that can be still gained with undirected graph as well, so using an undirected graph is also still valuable.

**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().)

*ANSWER*: The HITS algorithm has two components – the hubs and authorities. Hubs are nodes that have a lot of outgoing links and authorities are nodes that have a lot of incoming links.

Based on the chart below, we can see that n1 and n3 do have an inverse relationship with respect to hub and authority. The hub and authority for a directed graph would usually have an inverse relationship within itself. But even though n1 hub and n1 authority would have inverse relationship, it is interesting that n1 hub and n3 hub would have inverse relationship. Every time n1 authority goes down, so does the n3 authority go up. Similarly, every time the n3 hub goes down, the n1 hub score goes up. So in a nutshell, every time an actor has a lot of incoming links, they have fewer outgoing links, and every time n1 has either a lot or little of a type a link, n3 has the opposite. This seems to correlate with the phases and the direction of the drugs that they are moving, and the type of drugs. But even that is not an extremely clear correlation with the type of drugs because during phases with both marijuana and cocaine seizures, the hub and authority scores for both these actors can go either way a little bit.**Chart, line chart

Description automatically generated**

Comparing against the centralities below, there does seem to be a slight correlation between the centralities of the undirected graph and the hub and authority of the directed graph. When the centralities of the undirected graph go up, the hub of the directed graph seem to go down. Looking at n1 for phase 4 specifically, there is clearly a huge drop in hub score and then on the centrality chart there is also a large drop in both eigenvector and betweenness centrality, and even a slight drop in degree centrality. **Chart, line chart

Description automatically generated**

**Co-offending Network**

**Part (g):** Plot the degree distribution (or an approximation of it if needed) of 𝐺. Comment on the shape of the distribution. Could this graph have come from an Erdos-Renyi model? Why might the degree distribution have this shape?

*ANSWER*:

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

The log scale shows that this is a power law distribution, a real-life network. This type of network would not be generated by the Erdos-Renyi model. But can be by the configuration model. Only a few of the actors have a large number of degrees while the majority of the actors don’t.

This shape is basically like this because there are some people that know a lot of other people and are very well connected to everyone. The majority of the people are just one-offs and serve only a small function or have a relatively small piece of the pie and are not the most influential and well-connected people.

**Part (m):**Plot the distribution of clustering coefficients for each node for 𝐺𝑟 and 𝐺𝑛𝑟. What shape do the plots make? What does this tell you about the behavior of the actors? Hint: What does it mean for an actor to have a clustering coefficient of 0.5? Are there as many actors with intermediate clustering coefficients (say, between 0.25 and 0.75) as you expect for each graph?

*ANSWER*:

A picture containing histogram

Description automatically generated

A picture containing histogram

Description automatically generated

The clustering coefficients for the nodes are mostly likely to be 0, but second most likely to be 1. In both cases for Gr and Gnr. This means the actors are either in a clique, or more likely they are in a star. Being in a star means that the neighbors are not connected at all, and that there are fewer edges between the neighbor nodes.

The clustering coefficients that are intermediate are not as many because a lot of times people are completely unknown to other groups of people. In other words, a group of people will commit a crime together, and they will all know each other, but those people will not know a lot of other people. So the components are contained within themselves and not highly connected to other components. Since a group of people will not know another group of people, there isn’t likely to have a lot intermediate clustering coefficients because to have an intermediate clustering coefficient, a person A would need to commit an offense with a person B, and then a person C would offend with a person B as well, and also there is a person D that committed an offense with person B. For an intermediate clustering coefficient, person A and person C would not commit an offense together, but person A and person D would. This scenario is less likely than for person A, person C and person D to have all committed offenses together or none of them commit offenses together, which would leave person B as the hub in a star shape graph (thus a clustering coefficient of zero).

There seems to be not much difference though between the shapes of the distribution of Gnr and Gr. The only difference would be that repeat offenders have an ever higher amount of people with clustering coefficient of zero.

**Part (n):** Pick a centrality measure (degree, eigenvector, betweenness, etc) and compute the scores for the top (largest) component of 𝐺𝑟 and 𝐺𝑛𝑟. Compare the distribution of the centrality across nodes (for example, with summary statistics and/or a histogram). Examine the number of crimes committed by the most central actor in the repeat offender graph, does this support your conclusions?.

*ANSWER*:

Chart, histogram

Description automatically generated



In the largest component of Gr, the node with the highest eigenvector centrality was offender 596946. This person also committed 36 crimes. This offender committed crimes with as many as 3 co-offenders and also this person committed crimes by themselves. Most of the time though, this offender committed crimes with at least one other co-offender.

The repeat offender graph is different in that a lot offenders, almost half of the offenders actually, have eigenvector centralities that are above zero. It makes sense that this graph would represent a higher amount of offenders with respect to eigenvector centrality because there are more offenders repeating offenses with other offenders. So offenders that commit an extreme amount of offenses will all be in this distribution and therefore they will on average have a higher centrality. Those offenders that commit offenses with different groups of people, multiple times and have offenses spanning years and years will all be in this category as long as they repeat with the same co-offender more than once. For offender 596946, and offenders with similar amount of crimes, they will all have higher centrality.

**Project:**

“Are offenders more apt to commit crimes alone as they get older?”

I will investigate whether or not people tend to get in trouble with more people when they are younger vs when they are older. We will take the top 10 people that are repeat offenders over the years and investigate whether the amount of co-offenders these repeat offenders have actually goes up or down over the years.

The hypothesis is that over the years, repeat offenders would need fewer people to do crimes with, as they are older and more confident criminals. This could be called the “crime confidence index”. This could also give insight into the mind of the mind of early offenders before they become career criminals.

The methodology will be that we will take the top ten offenders. The top ten offenders will be measured by the number of offenses they are involved in cumulatively. Over time, we will also look at whether the number of co-offenders goes up or down

We will just look at the averages for each year for each offender. All of these offenders have hundreds of offenses so there will be plenty of samples for each year. So for each year we will take the averages and see if there is a correlation between the year and the average amount of offenders.

|  |  |
| --- | --- |
| **Offender Identifier** | **Offense Count** |
| 300728 | 456 |
| 440431 | 358 |
| 253577 | 357 |
| 120976 | 350 |
| 373749 | 275 |
| 396148 | 247 |
| 189402 | 220 |
| 614546 | 209 |
| 623487 | 206 |
| 7540 | 196 |

Here below is the line chart of the top 10 offenders and we can clearly see there is no positive correlation between year and average amount of offenders. There is a slight negative correlation though.

Line chart

Description automatically generated with low confidence

In this study we found that as time goes on, offenders do not get more confident and commit more crimes alone. They have as many crimes with more offenders or with themselves when they are young as when they are getting older. We could look for different demographics in the future and test the same things and find if there are any certain demographics that lean one way or the other.