MITx - 6.419x

Data Analysis: Statistical Modeling and Computation in Applications: Spring 2021

Homework 3: Written Report by J. Andrew Seidel

Problem 1: Suggesting Similar Papers

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

Answer: The use of matrix multiplication **greatly reduces the complexity** of the solution, when compared with the algorthim described in earlier parts of the problem.

Part (d): (3 points) 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?

Answer: Bibliographic coupling and cocitation produce noticeably different network results because their network edges represent different relationships between papers:

- Bibliographic coupling network edges occur when two papers cite the same paper, and the edge weight reflects the frequency of this occurence.
- Cocitation network edges occur when two paper are cited by the same paper, and the edge weight reflect the frequency of this occurence.

To highlight the distinction, consider an extreme case: Two paper each cite zero other papers, but both are highly cited by the same set of other papers. In this case, the two papers would not be connected within a Bibliographic network since neither paper cites any other papers. However, within a Cocitation network, the two papers would be highly connected, since many other papers cite both papers. Also note that the Bibliographic network edge weight is a function of only the two papers' citations, while the Cociation network edge weights are a function of the citations of every paper within the network.

- For the reasons above, the Cocitation network is the more appropriate indicator of similarity between papers

Problem 2: Investigating a time-varying criminal network

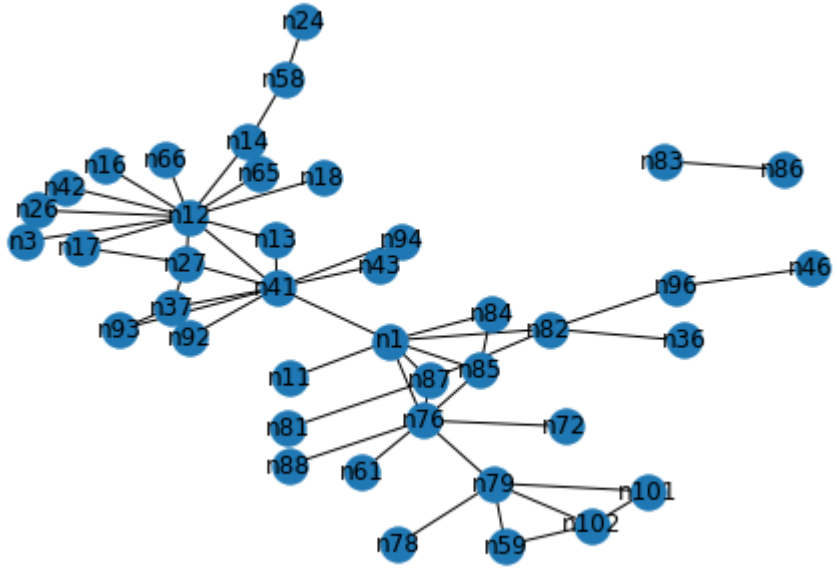
Part (c): (2 points) Observe the plot you made in Part (a) Question 1. The number of nodes increases sharply over the first few phases then levels out. Comment on what you think may be causing this effect. Based on your answer, should you adjust your conclusions in Part (b) Question 5?

Answer: The number of network nodes grew across the first three phases from 15 to 24 to 33, but then stayed the same and then slightly declined from to 33 and 32 for phases 4 and 5. We might expect the growth of the network in the first three phases to be a result of the initial success of drug trafficing without any seizures, and the subsequent stabilization and slight decline of the network in phases 4 and 5 to be a result of the 1st seizure, which occured at phase 4. Following this seizure, we also note that trafficing began to reorient to cocaine import from Colombia, transiting through the United States, using a different network from the initial maruajana import from Morocco through Spain. This shift in the network orientation should also adjust conclusions about centrality of nodes prior to phase 4 and after phase 4.

Part (d): (5 points) 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: Each metric is a measure of how central (or important) an individual is within the drug trafficing network, but each metric is calculated differently and yields different values. Generally, the higher the value, the more central (imporant) the individual is within the trafficing network.

- Degree Centrality simply represents the normalized number of connections an individual has within the trafficing network. A limitation of the Degree Centrality measure is that assigns each connection the same value. In reality, certain connections are more important than others - such as connections that would split a network into two, if the connection was removed. Below in phase 11, we see the connection between individuals n1 and n41 is an important connection joining two seemingly distinct network components.



Phase: 11

- Betweenness Centrality represents the occurence of an individual lying on the shortest path between the combinations of other individual pairs. It addresses the limitation of Degree Centrality by identifying higher value connections and would be the most relevant measure to identify the most important individuals within the network. As evidence, we see the that two highest Betweenness Centraliy measure value is n41 (0.55) and n1 (0.52), which is also consistent with the network graph visualization of two large components connected through the single connection between individuals n41 and n1

Note: we consider importance with the following question in mind: Which individuals, if removed, would cause the trafficing network to break apart? While we take degree centrality and eigenvector centrality into consideration, we place the greatest emphasis on

Part (e): (3 points) 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.

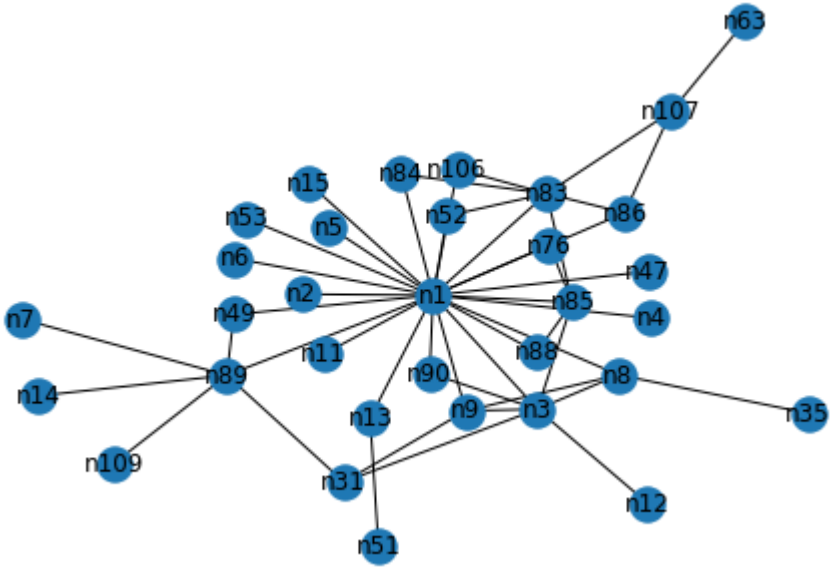
Answer: Given our earlier defintion of 'importance' and looking at the sum of the betweenness centrality measure of the network over the 11 phase, the highest values and greatest importance are associated with the following individuals:

Top 3 Highest Betweenness Centrality Scores (Summed Across All 11 Phases)

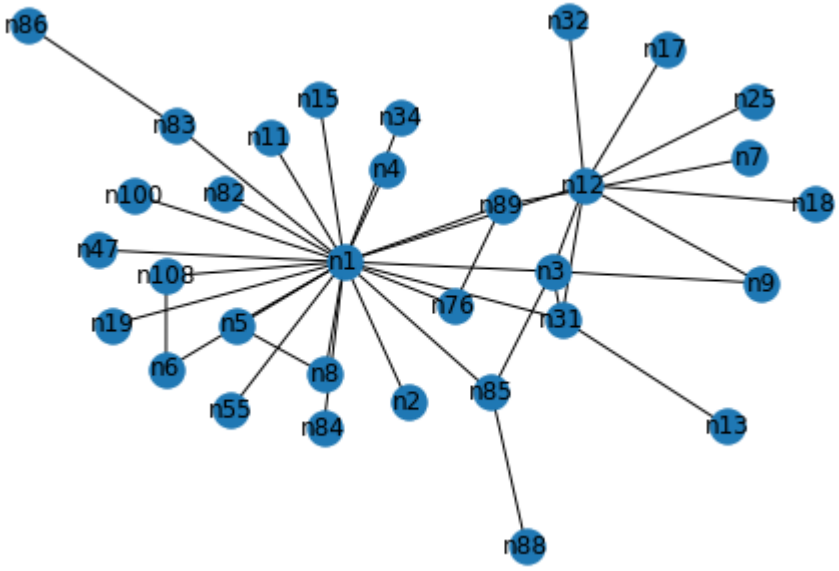
- n1 (7.20): Daniel Serero - Mastermind of the network.
- n12 (1.84): Ernesto Morales - Principal organizer of the cocaine import, intermediary between the Colombians and the Serero organization.
- n3 (1.42): Pierre Perlini - Principal lieutenant of Serero, he executes Serero's instructions.

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.



Phase: 4



Phase: 5

Answer: The networks shown are for phases 4 and 5. The first drug seizure (marijuana) coincided with phase 4 and subsequently, the drug trafficing shifted from marijuana imports from Morocco through Spain, to new cocaine imports from Columbia through the United States. Betweenness Centrality for individual

n12 jumped from 0.00 in phase 4 to 0.26 in phase 5 (the second highest Betweenness Centrality value within the network behind n1, for phase 5), and visually from the two graphs, we see the increased centrality of n12 in phase 5. As noted above, n12 represents Ernesto Morales - Principal organizer of the cocaine import, intermediary between the Colombians and the Serero organization.

Part (g): (4 points) 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.

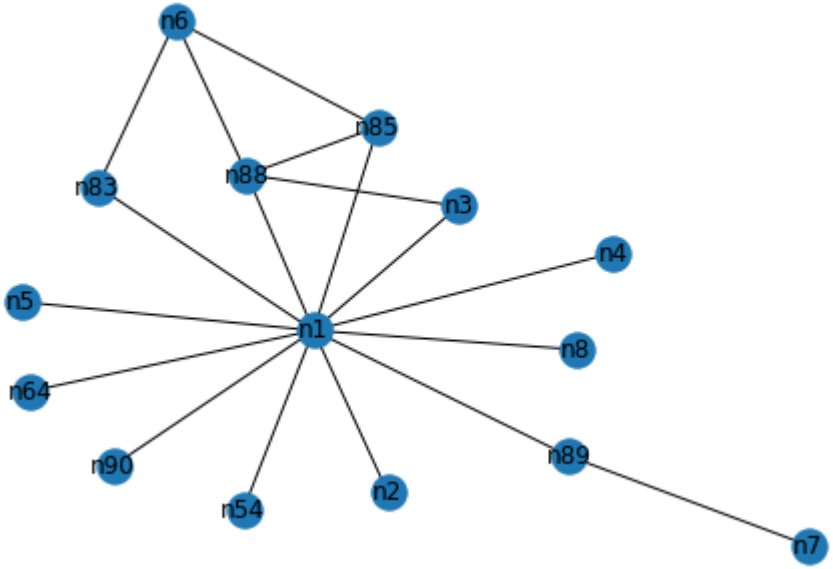
Describe the coarse pattern(s) you observe as the network evolves through the phases. Does the network evolution reflect the background story?

Hint: Look at the set of actors involved at each phase, and describe how the composition of the graph is changing. Investigate when important actors seem to change roles by their movement within the hierarchy. Correlate your observations with the information that the police provided in the setup to this homework problem.

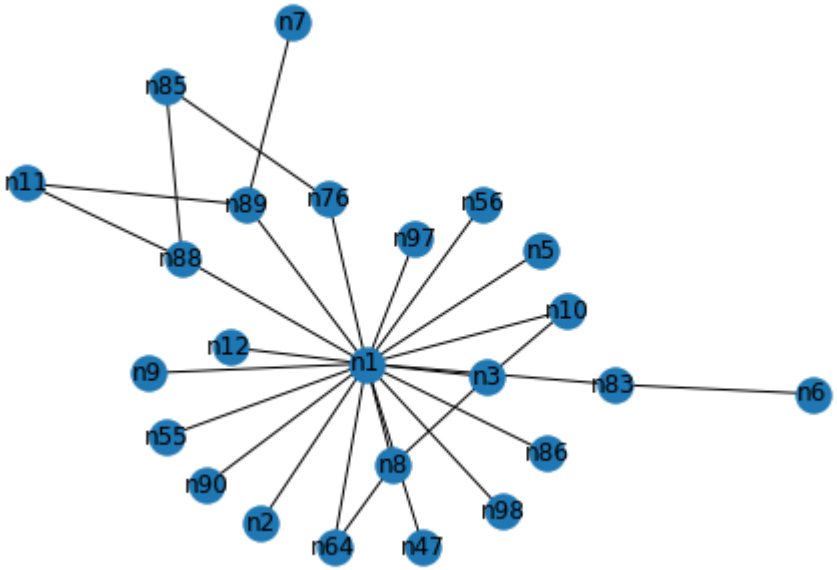
Answer:

Global network trends and the backstory:

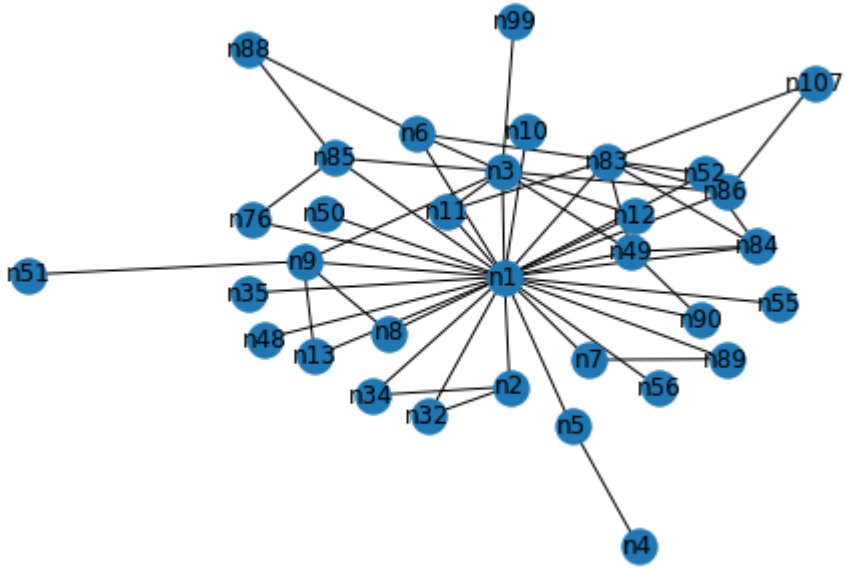
- In phase 1 through phase 3, the network grows in terms of the total number of individuals involved, and remains centered around individual n1. This is conssitent with the backstory that n1 was Daniel Serero - Mastermind of the network, and that there weren't any seizures until phase 4, and drug trafficing was limited to marijuana from Morocco through Spain.
- After the seizure in phase 4, trafficing refocuses to cocaine and a new network from Columbia through the United States. In phases 5 through 11, we see a significant increase in centraliy in individual n12, Ernesto Morales - Principal organizer of the cocaine import, intermediary between the Colombians and the Serero organization.
- And in phase 7 and phase 10, after the first cocaine seizure in phase 6, and additional cocaine seizures in phases 8 and 9 we see the network between individuals n1 and n12 fracture into two distinct component groups, the first centered again around n1, and the second centered around n12.



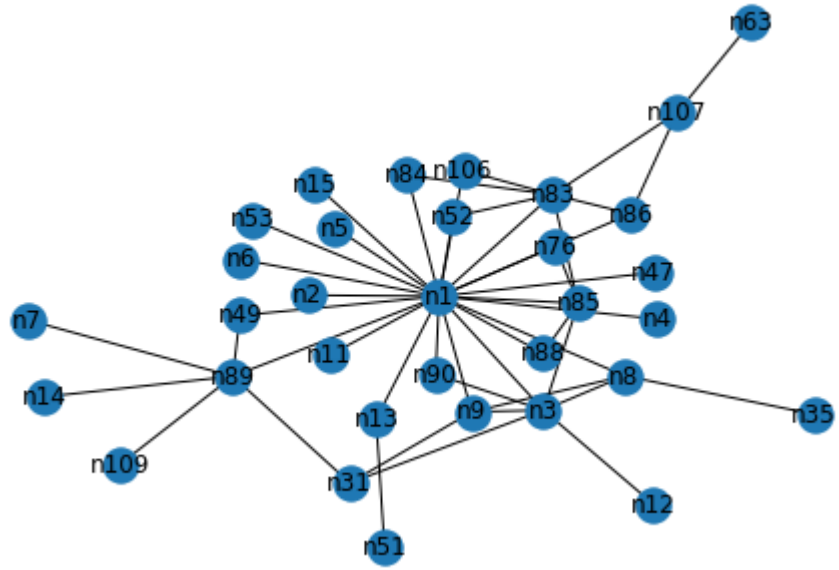
Phase: 1



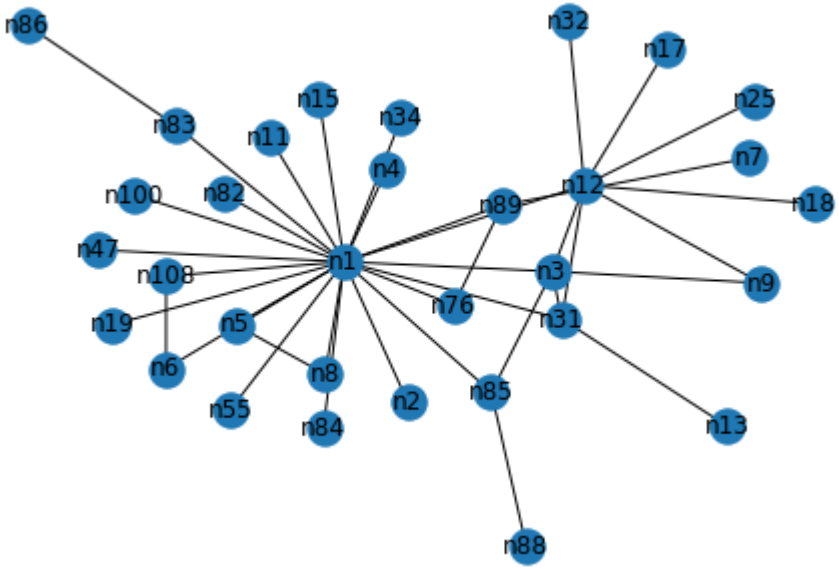
Phase: 2



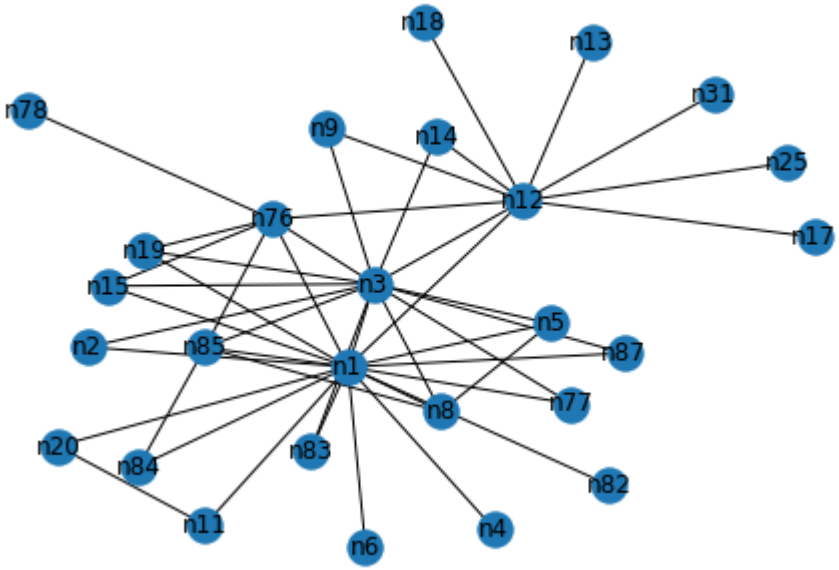
Phase: 3



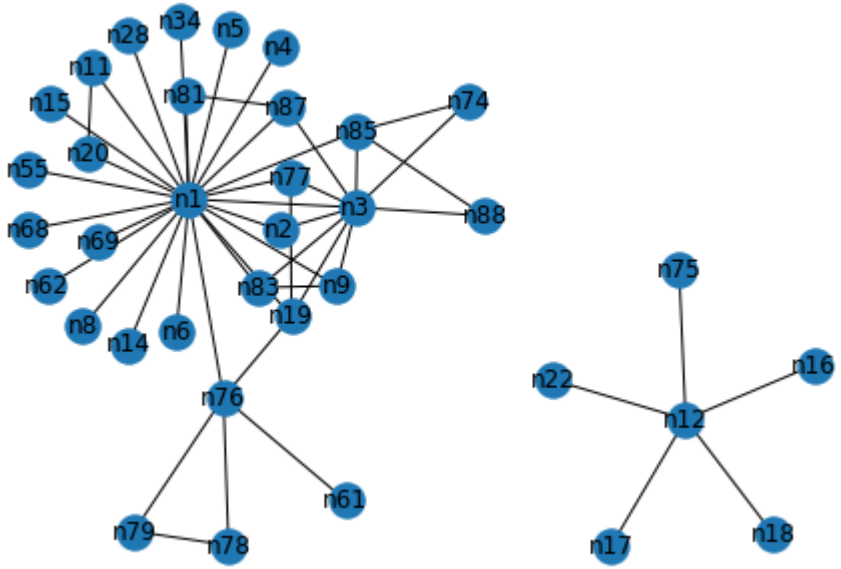
Phase: 4



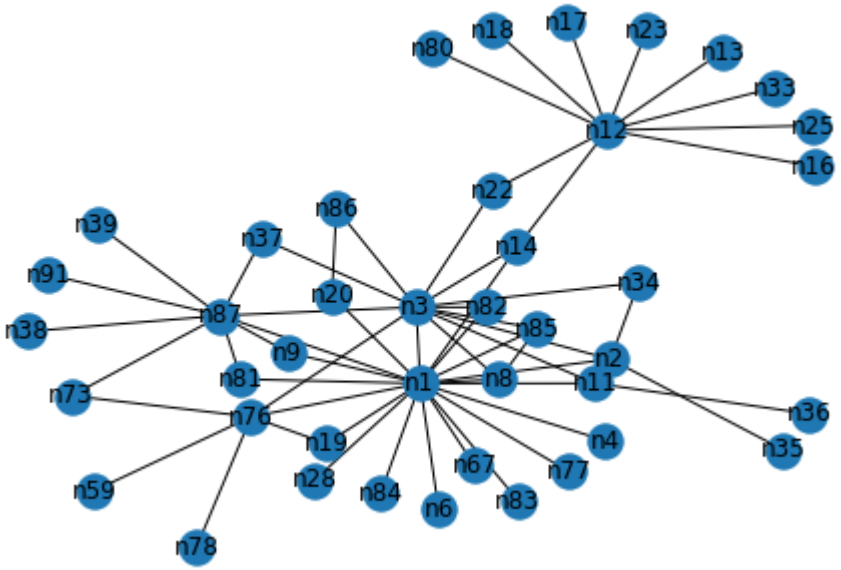
Phase: 5



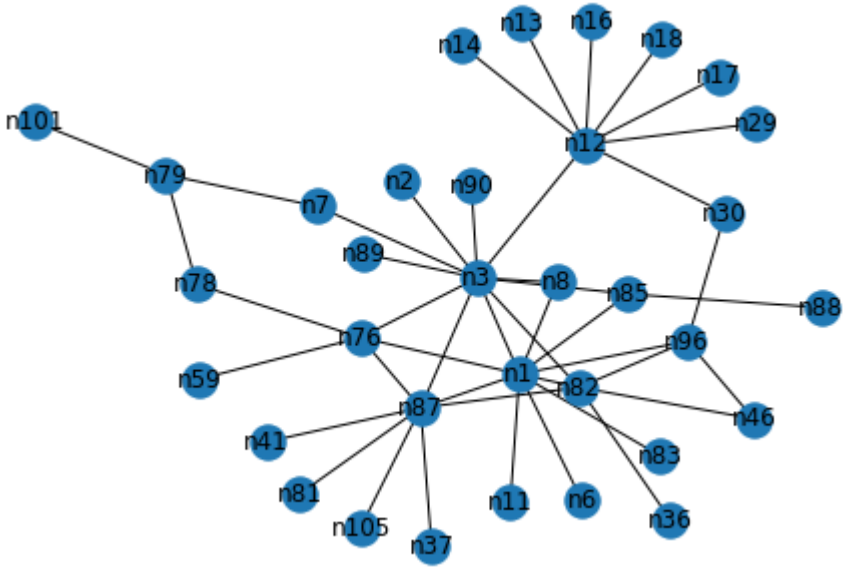
Phase: 6 ./Network_Img/N_6.png



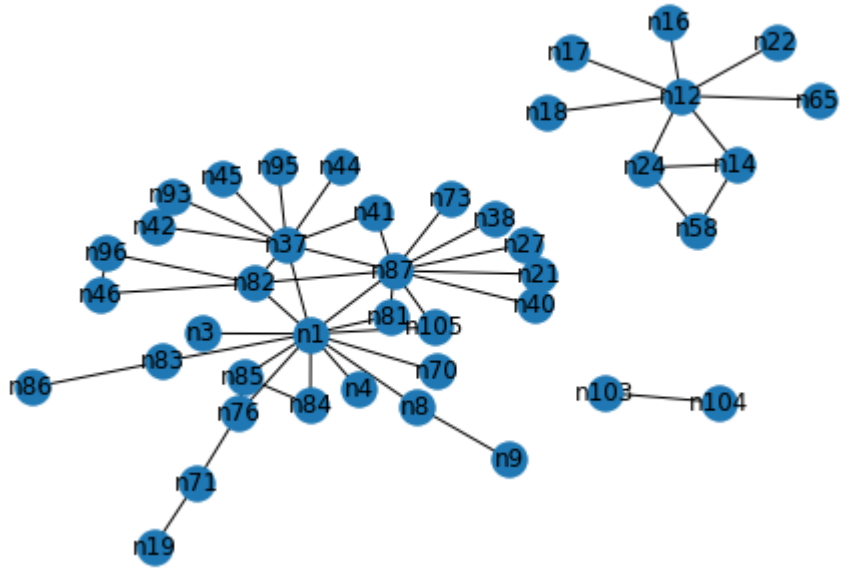
Phase: 7



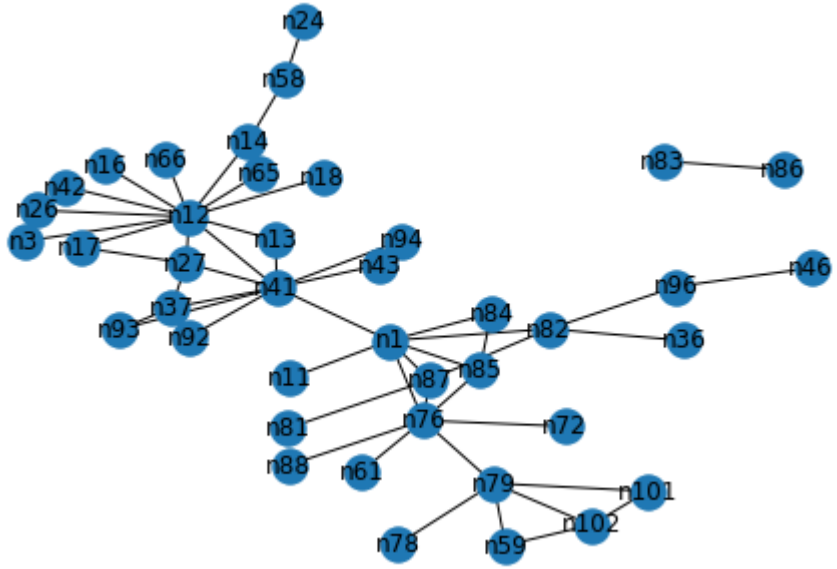
Phase: 8



Phase: 9



Phase: 10



Phase: 11

Part (h): (2 points) 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: Yes, from the various centrality measure and by visual inspection of the graphs above, we see that the following individuals were important parts of the network but not listed among the 23 provided in the investigation:

- Individual n41 played a central role in phase 11, bewtween n12 and n1 (betweenness sum across phase of 0.55, 6th higest)
- Similarly, individual n79 in phase 11 was important in connecting 4 other individuals into the networt, but isn't listed in the investigation.

Part (i): (2 points) What are the advantages of looking at the directed version vs. undirected version of the criminal network?

Hint: If we were to study the directed version of the graph, instead of the undirected, what would you learn from comparing the in-degree and out-degree centralities of each actor? Similarly, what would you learn from the left- and right-eigenvector centralities, respectively?

Answer: As this data reflects the wire-tapping (listening to phone calls between individuals) the directed version of the graph would capture information regarding who initiated the call and who received it. In the most general form, it represents the flow of communication - but we also recongize that information could be shared bi-laterally even a single party is consistently the initiator of the call.

- The Left Eigenvector Centrality represents the importance of an individual with respect to the number of other individual making calls to that individual (edges pointing to the individual). Individuals with the highest left eigenvector centrality would receive the greatest number of phone calls
- The Right Eigenvector Centrality represent the importance of an individual with respect to the number of other individuals that individual calls (edges pointing from the individual). Individuals with the highest right eigenvector centrality would initiate the greatest number of phone calls.

Part (j): (4 points) Recall the definition of hubs and authorities. Compute the hub and authority score of each actor, and for each phase.

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)?

Answer:

An individual has a high "hub score" if he or she points to many important authorities, and inversely, an individual has a high "authority score" if he or she is pointed to by many important hubs.

We see that in the phases 1 through 5:

- Individual n1 has a high hub score (0.70 to 0.90 across phases), and an authority score near zero, and that conversely,
- Individual n3 has a low hub score and a moderatly high authority score (0.14 to 0.34).

These relationships switch in phases 6 and 7:

- Individual n1's high hub score falls to nearly zero, and his authority score rises from nearly zero to a relatively high 0.81 and 0.73,
- Individual n3's low hub score increase to a moderately high 0.20 and 0.34

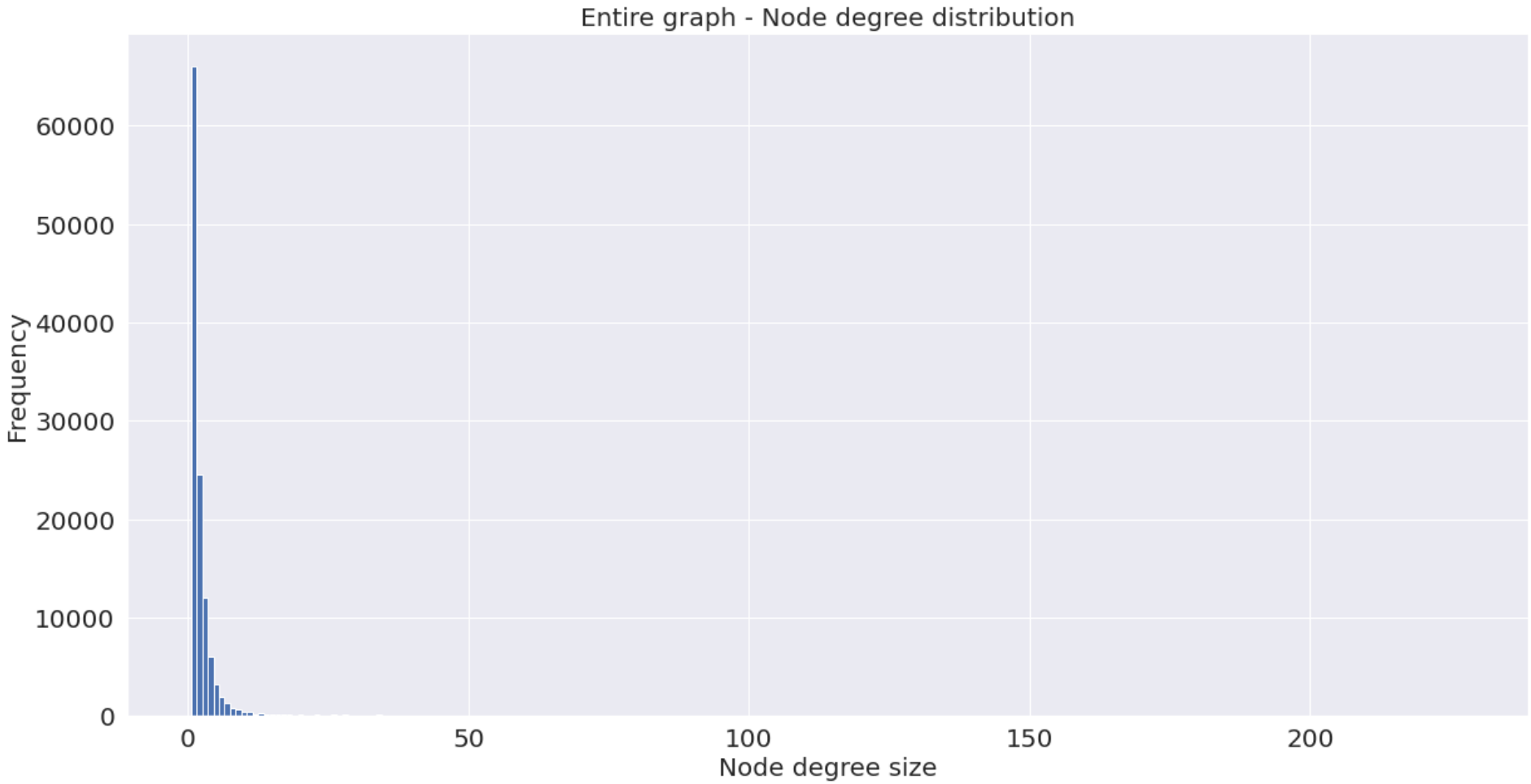
As noted above in part G, we cite the cocaine seizure in phase 6 of the fracturing of the cocaine network in phase 7 as support for the shifting behavior of individual n1 - Daniel Serero - Mastermind of the network from a hub to an authority, and the associated shift from authority to hub for individual n3 - Pierre Perlini - Principal lieutenant of Serero, who executes Serero's instructions.

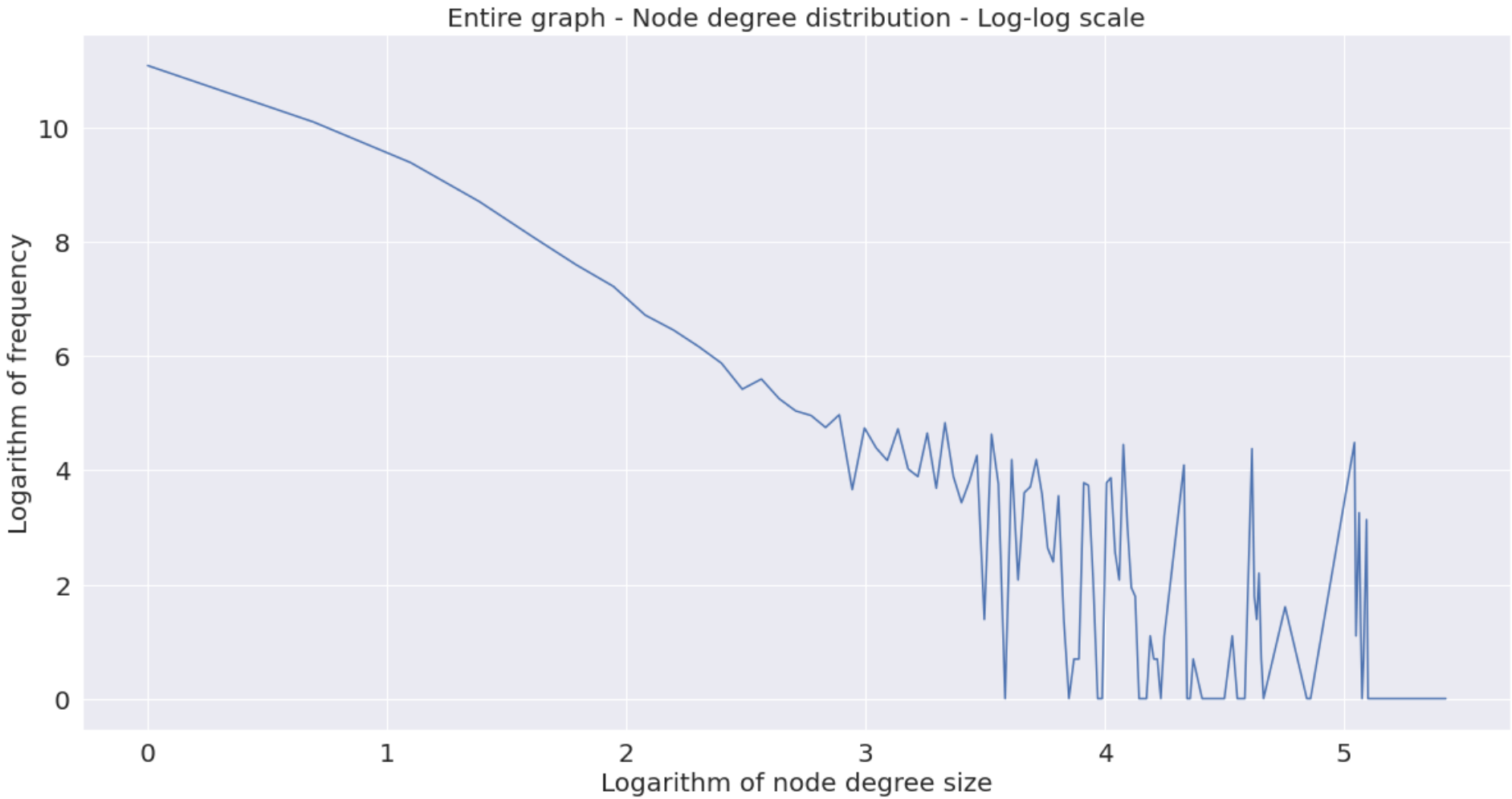
Problem 3: Co-offending Network

Part (g): (3 points) 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: From the two plots below of the degree distribution and the log scale degree distribution, we see that the log scale distribution appears to be approximately linear in shape, such that is in fact a power law distribution. As many natural networks follow power-law distributions, this is inline with our expectations.

We would not expect this model to come from an Erdos-Renyi model, because Erdos-Renyi models follow a binomial degree distribution, which is inconsistent with the power-law distribution we see with the linear log-scale plot.

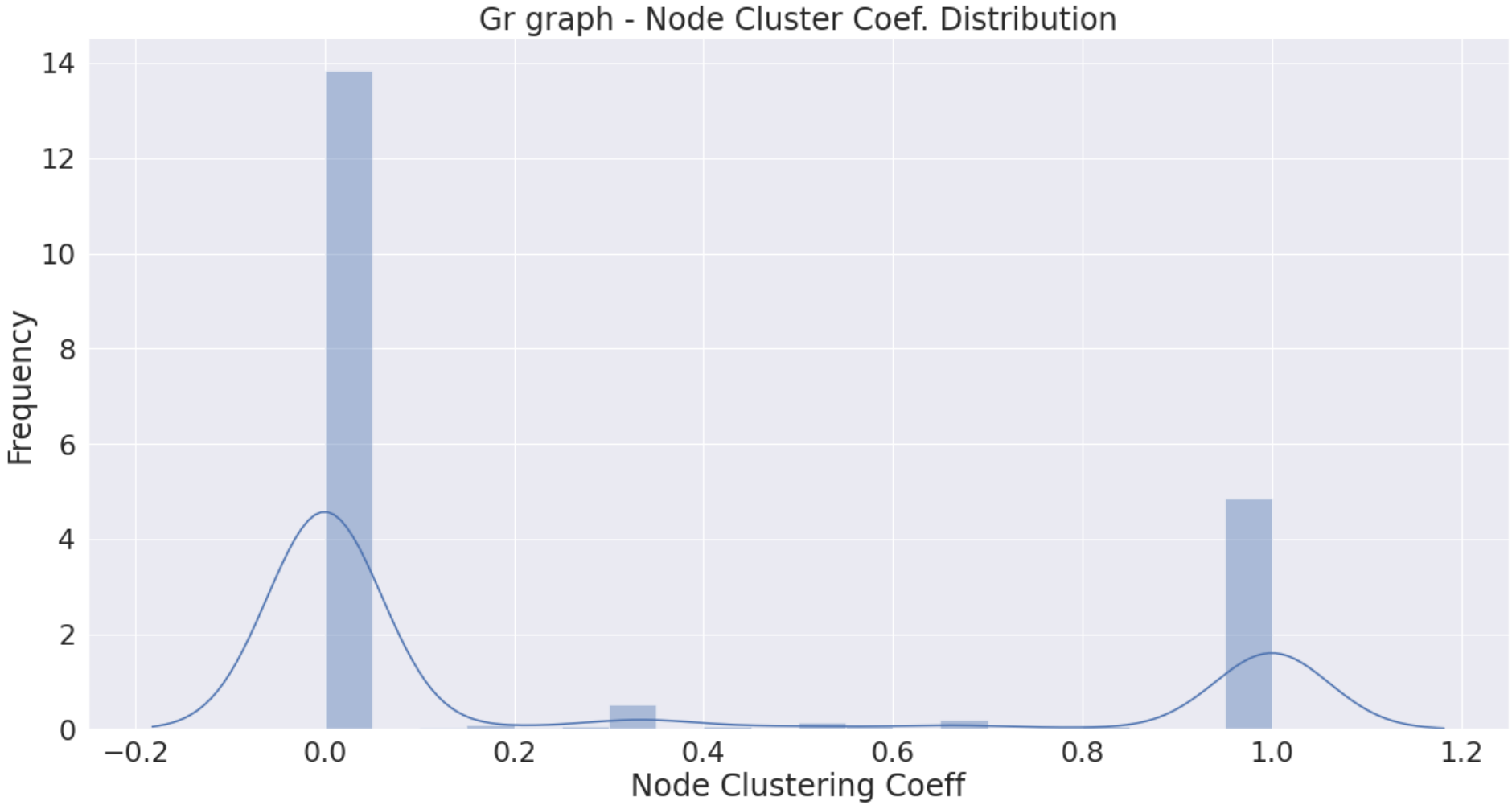

Entire graph - Node degree distribution
Frequency
60000
50000
40000
30000
20000
10000
0
0
50
100
150
200
Node degree size

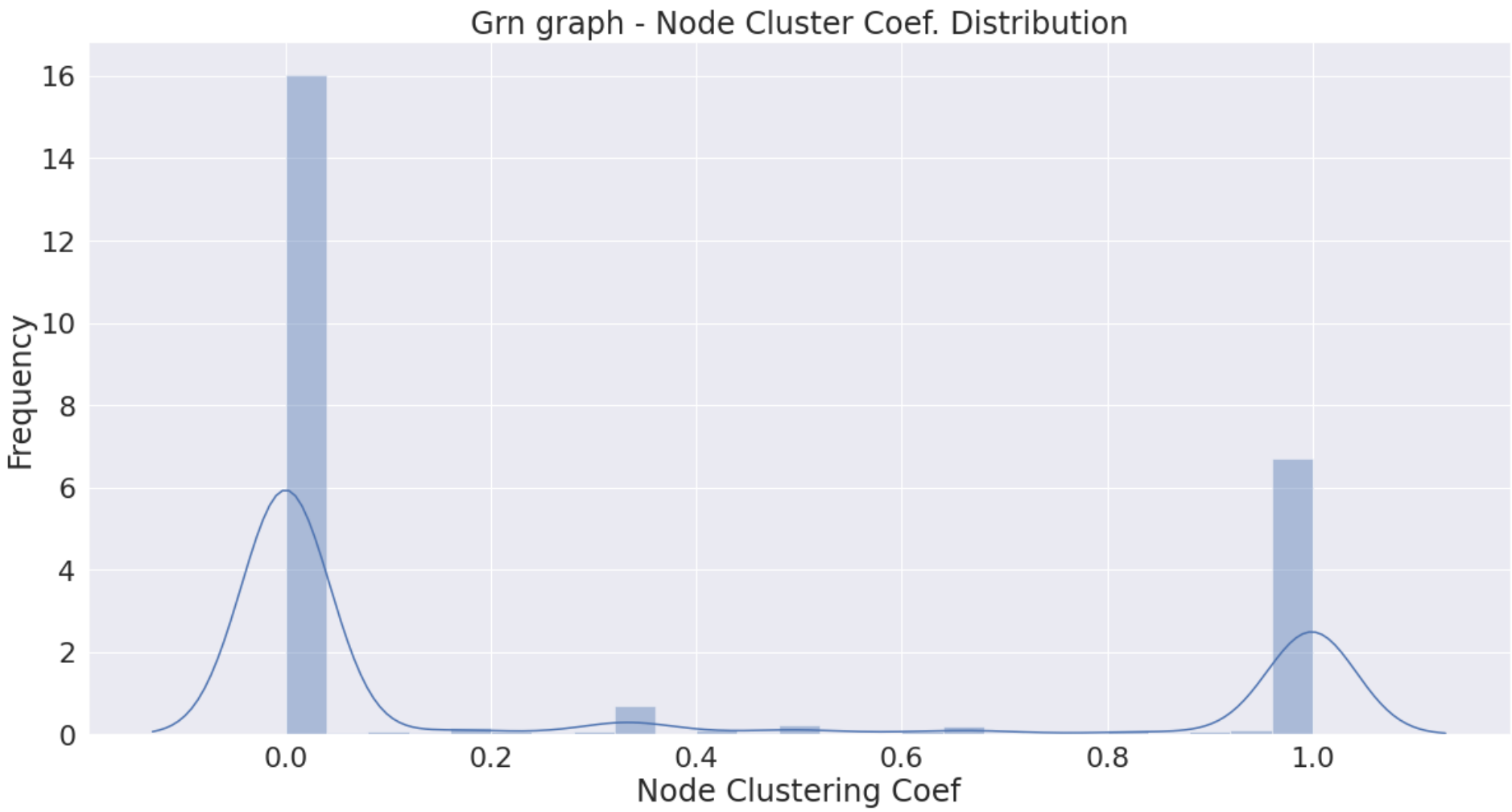


Part (m): (4 points) Plot the distribution of clustering coefficients for each node for Gr and Gnr. What shape do the plots make? What does this tell you about the behavior of the actors? Hint: What does it mean for an actor to have a clustering coefficient of 0.5? Are there as many actors with intermediate clustering coefficients (say, between 0.25 and 0.75) as you expect for each graph?

Answer: The nodes within both the Gr and the Grn networks follow a bi-modal distribution, with a large number of nodes having a clustering coefficient near 0, and smaller number of nodes having a clustering coefficient near 1, and only a few nodes having a clustering coefficent between 0 and 1. This tells us that

the for most offenders, their network of co-offenders are themselves not co-offenders amongst each other (clustering coefficient of zero), but that there is moderately sized number of offenders with a network of co-offenders who are also co-offenders with each other (clustering coefficient of one), and that there are relatively few offenders with a network of co-offenders who are co-offenders with only a subset of each other (clustering coefficient between zero and one).



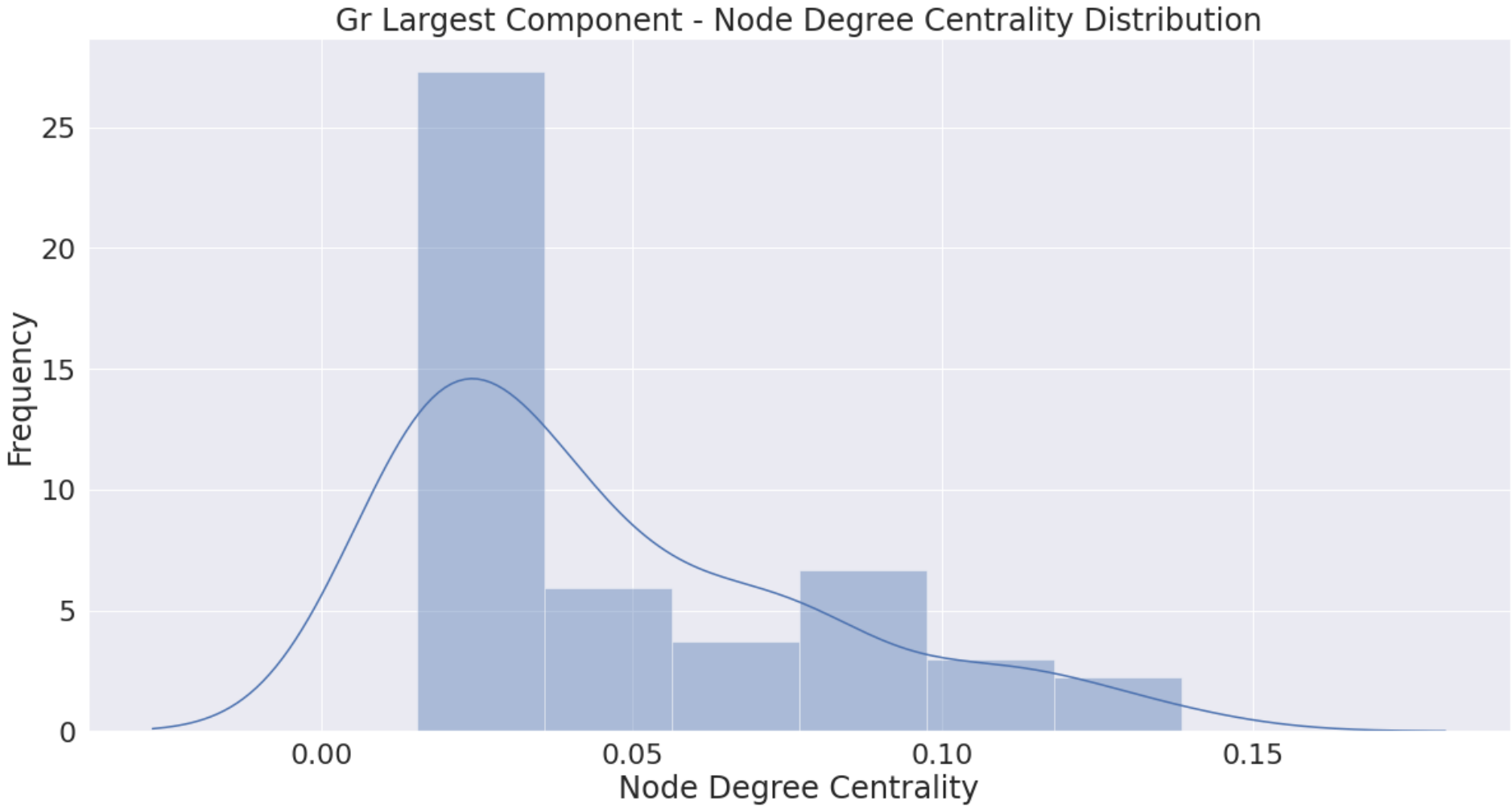


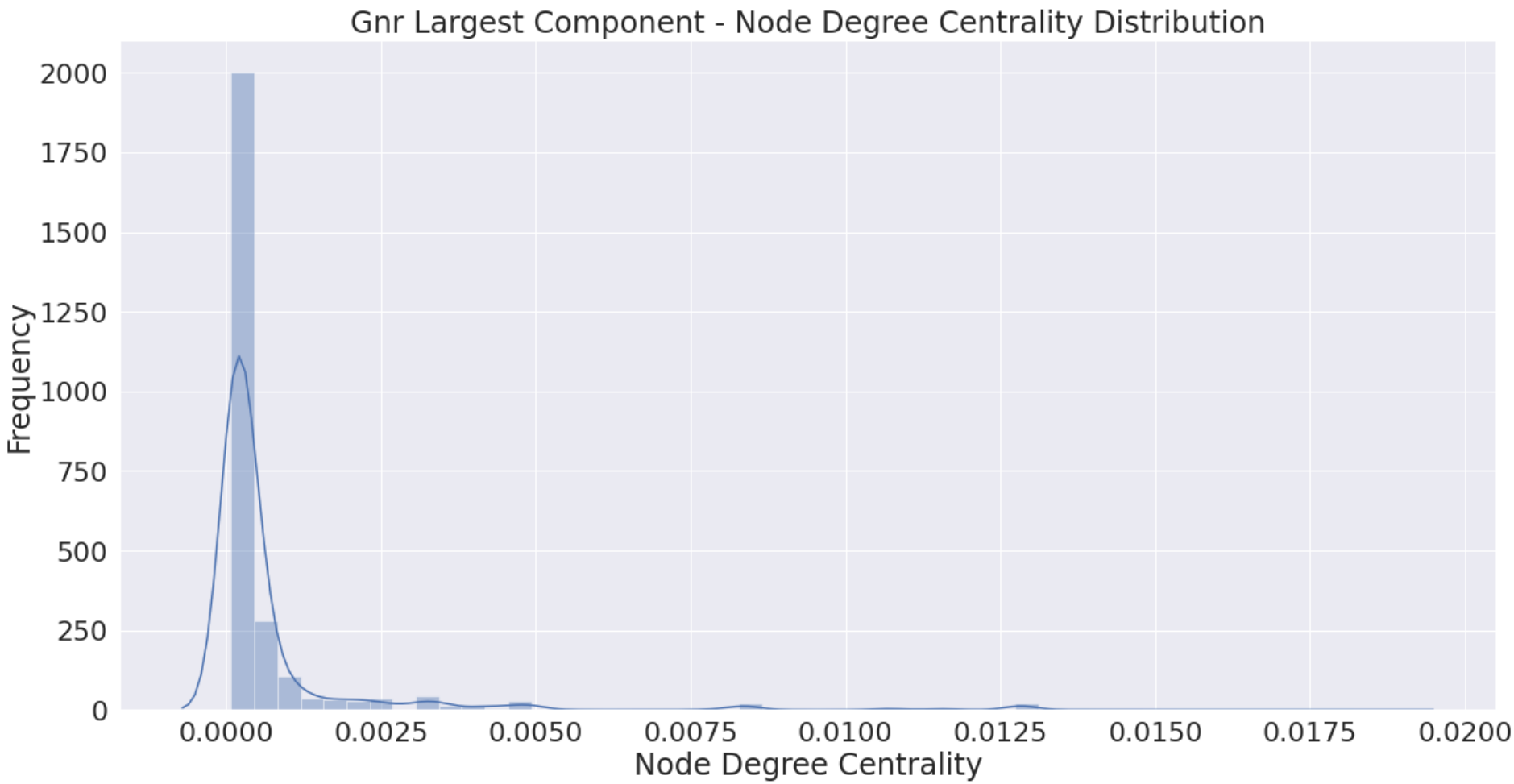
Part (n): (4 points) Pick a centrality measure (degree, eigenvector, betweenness, etc) and compute the scores for the top (largest) component of Gr and Gnr. Compare the distribution of the centrality across nodes (for example, with summary statistics and/or a histogram). Examine the number of crimes committed by the most central actor in the repeat offender graph, does this support your conclusions?.

Answer: For computational efficiency, we choose the degree centrality measure and below plot the distribution of node degree centrality for the largest connected components of the network for repeat co-

offenders (Gr) and non-repeat co-offenders (Gnr). We find that both plots show higher frequencies of lower degree centrality, but we find that within the network of repeat co-offenders, the degree centrality is higher overall.

The offender (OffenderIdentifier 610924) with the highest degree centrality (0.13) was involved with a relatively high 46 different crimes. If we filter all records by these 46 crimes we also find a relatively high number of other co-offenders (17 in total) associated with this one central offender, which is consistent with the high degree centrality found with this one offender.




Gnr Largest Component - Node Degree Centrality Distribution
Frequency
2000
1750
1500
1250
1000
750
500
250
0
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
0.0200
Node Degree Centrality

```
1
2
```