**Written Report – 6.419x Module 3**

**Name:** 4sowmya

**Part (c):** (**2 points)**

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

The time complexity of the solution with matrix multiplication is O(n\*n) whereas the time complexity for the friend’s algorithm is O(n\*n\*n). This is because, the iteration is going through each row twice to get the non-zero pair. The matrix multiplication method is faster than the iteration method. I have tried to confirm this with the python script (Appendix-1). The following script runs both the methods and prints the time taken to complete each task. From the output, I could see that friend’s iterative method takes longer time to execute than the matrix method. I tried with n=10, n=100, n=1000. As the value of n increase, the time difference increases significantly.

**Part (d):** (**3 points)** (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?**

Citation counts are interpreted as measures of the impact or influence of a publication. Both bibliographic coupling and co-citation are semantic similarity measures for documents in citation analysis. Co-citation is the frequency in which two documents are cited together by other documents. Two documents are bibliographically coupled if they both cite the same third document.

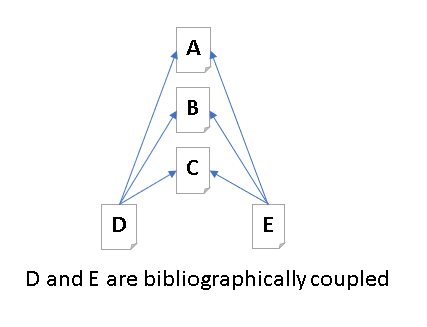
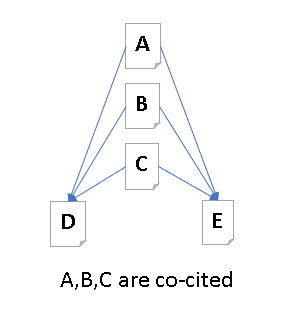
 

Figure 1 Figure 2

Considering the figure 1 and figure 2, we can infer that the bibliographic coupling is a mirror image of the co-citation semantically. They both can have different result as bibliographic coupling can become static over time while co-citation is dynamic.

To analyze the similarity between two papers will need domain knowledge to understand the reasoning/ use of the cited resources. But as an indicator for similarity analysis, I would consider bibliographic coupling. The reason can be explained using reference to figure1,2 above.

Publication A, B, C can be co-cited in multiple other papers. It does not mean that the content in each of these papers are similar topics. In case of, bibliographic coupling, document D, E can have similar topics as they are using the same citations. Hence bibliographic coupling will provide semantic similarity between the papers.

# **Problem 2: Investigating a time-varying criminal network**

**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?**

At every phase of the wiretapping, more connections of the current network can be found. Hence the nodes are increases sharply over the first few phases. Due to the continuous police activity, the network connections can be interrupted, or the periphery of the network is reached at each cluster and hence the level out.

Part(b)question 5 finds the mean with reference to the temporal consistency using two of the centrality metrics (betweenness centrality and eigenvector centrality). Since the network is not stable, it will not make sense to get the mean of the centrality to find the most influential actors. If a connection is identified in the later part of the stage, then the mean value may not represent the actual strength of the connection. I would prefer to get median of the scores to understand the strength of each actor and identify the most influential links. Since this is a criminal network analysis, eigen centrality is optimal to identify the highly influential actor so that when this person is stopped, it will have high impact in reducing the crime.

**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.**

In the context of criminal networks, degree centrality shows connection between individuals.

Degree centrality assigns the score based on the number of links in each node. Hence, if there are 3 smaller agents who has 5 connections each will have the same score. It does not differentiate if the 5 connections are smaller connections or high influential connections.

Betweenness centrality calculates the score based on the sum of short paths between nodes. This can help identify the high influential connections. A high betweenness count could indicate someone holds authority over disparate clusters in a network or just a periphery to both clusters. The central node may have the high betweenness value. I could not infer if that person is the main leader or just a contractor with lots of connection.

Eigen centrality measures the influence of a node based on the number of links it has to other nodes in the network. Eigen centrality also considers how we a node is connected and how many links their connections have and so on through the network. This helps in identifying the nodes with influence over the whole network, not just those directly connected to it. This will be able more appropriate measure to find the top master mind.

In terms of limitations, degree centrality might not have high score for the mastermind if the master mind communicates only with the top middleman. The betweenness centrality can also not highlight the master mind if one of the middlemen has more smaller connections.

Hence the best option will be Eigen centrality as this will consider the middleman’s connection too.

**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.**

While referring to the quantitative and centrality metrics, we could see that the nodes “n1” and “n3” are high score nodes in all the metrics. Hence it will be good to look at these two nodes in specific to understand the connections. Removing these two nodes will have a big impact in the network. The connections can be cut. Phase 3 and Phase 9 as shown below as samples. The node size represent the score based score from the centrality calculation.

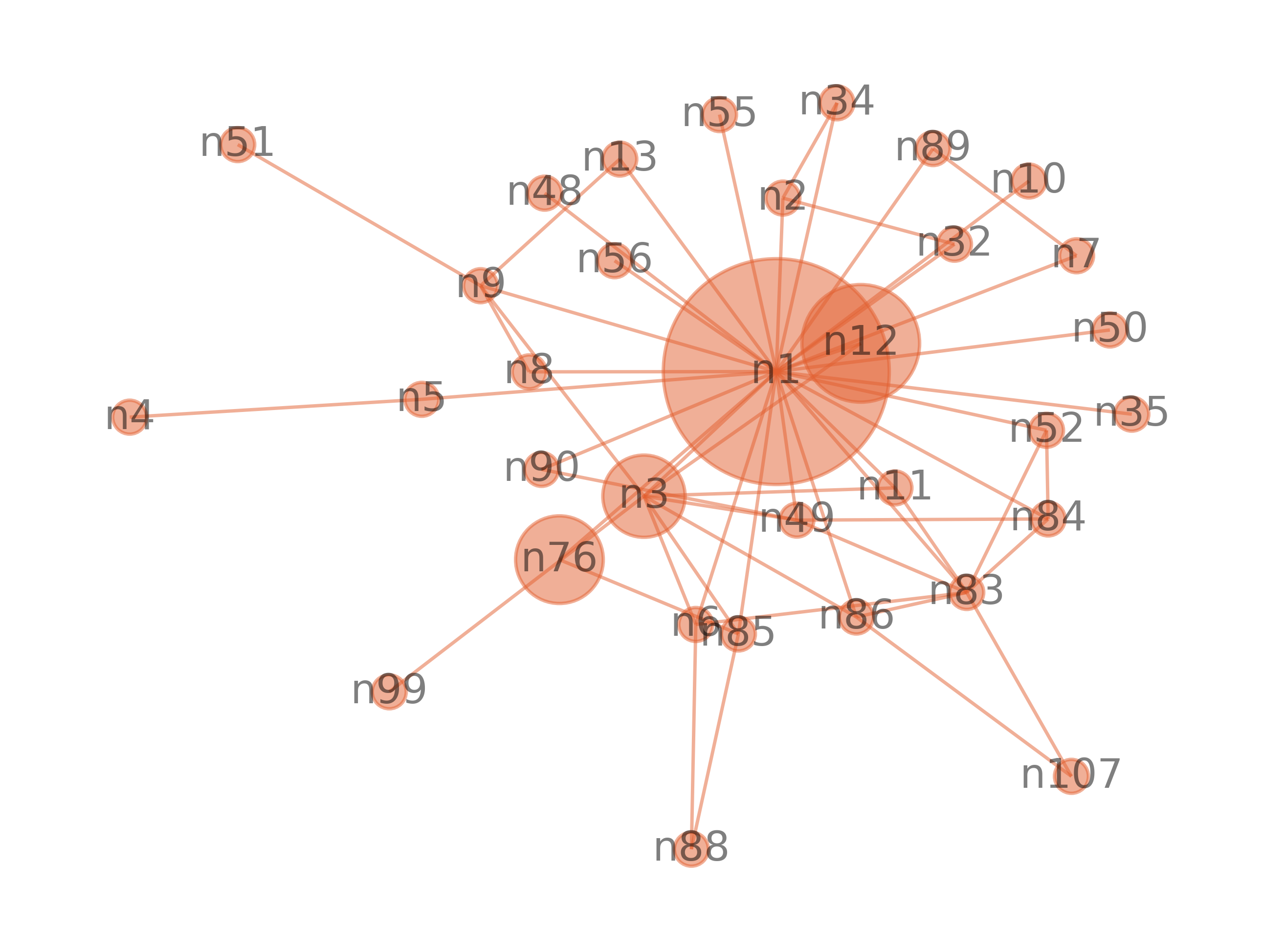
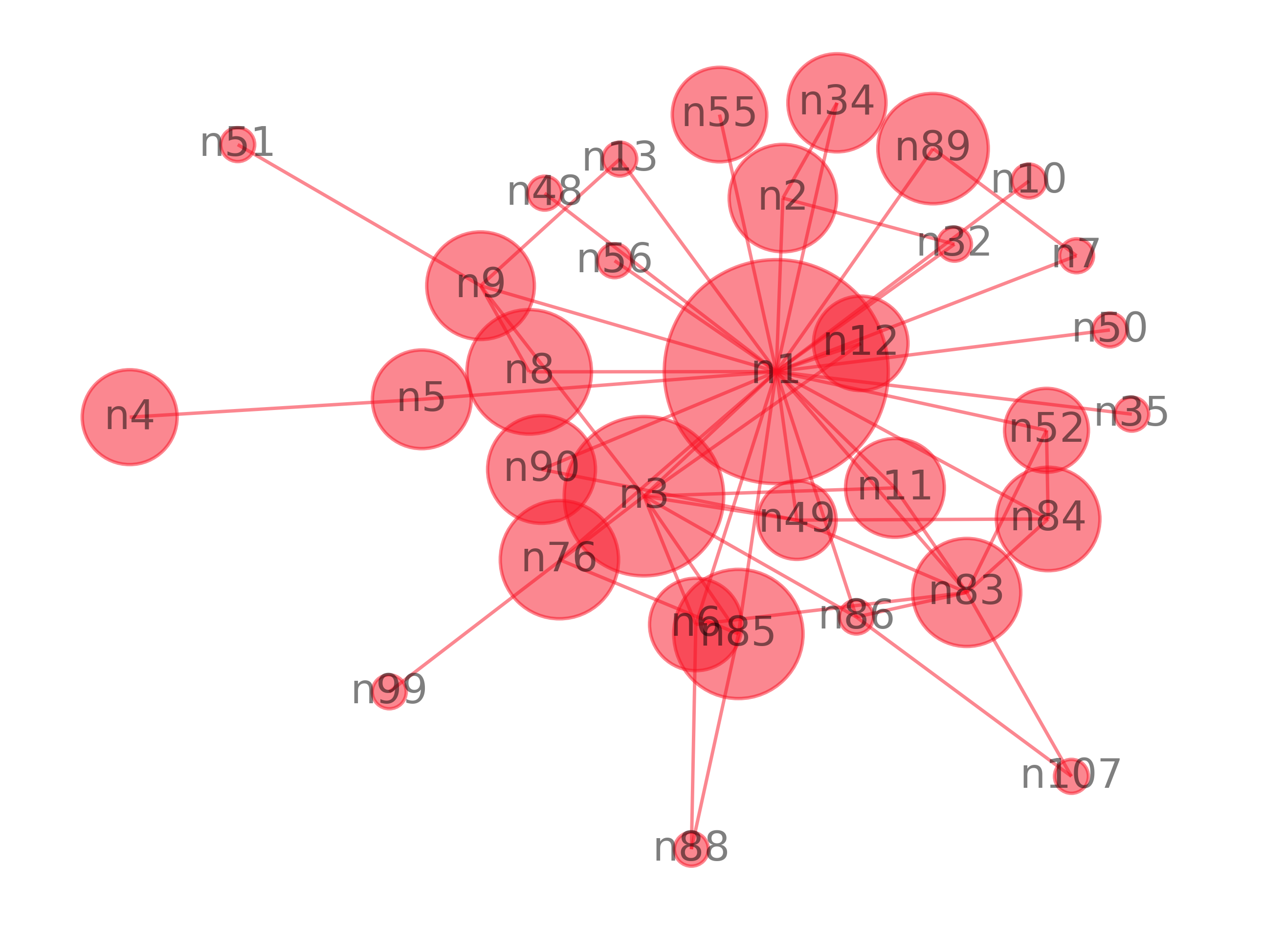
 

Figure 3 Betweenness centrality (Phase 3) Figure 4: Eigen centrality (Phase 3)

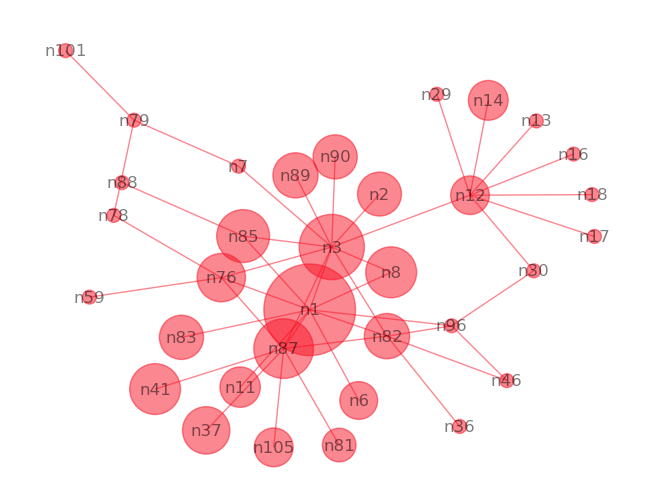
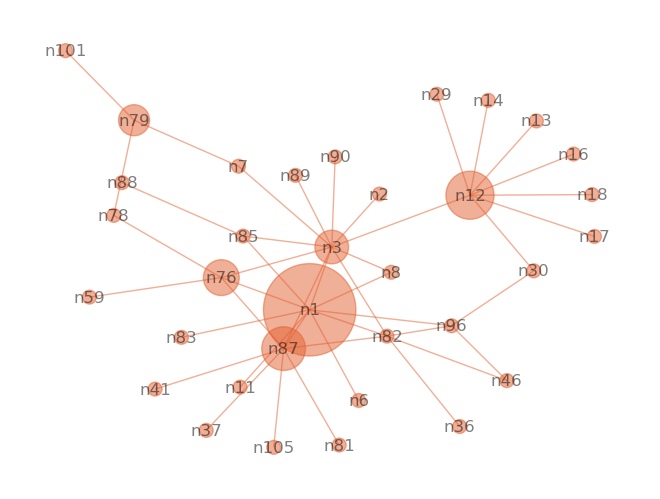


Figure 5 Betweenness Centrality (Phase 9) Figure 6 Eigen centrality (Phase 9)

While considering the connections and the sub-connections in the network through eigen centrality, we can infer that the nodes “n1”,”n3”, “n85” are the high influencing nodes in eigen centrality. We also see that “n12” also gains some importance in the betweenness centrality along with other nodes specified. From the documentation about the study, those represent

* Daniel Serero (n1) : Mastermind of the network.
* Pierre Perlini (n3) : Principal lieutenant of Serero, he executes Serero's instructions.
* Wallace Lee (n85) : Takes care of financial affairs (accountant).
* Ernesto Morales (n12): Principal organizer of the cocaine import, intermediary between the Colombians and the Serero organization.

**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 X = 4 and Phase X+1 = 5

At Phase 4, the first seizure of 300kg Marijuana happened. After this there is a change in the clustering happened in the graph. Also, the connection started changing after the phase 4. The degree of node “n89” is reduced after this phase meaning the connections in the network is broken. The degree of node “n12” started to increase in the further phases. Antonio Iannacci (n89) is not able to traffic Marijuana. The seizure could have impacted him financially. Since the seizure is started, criminals had to start a different route to traffic drugs. Ernesto Morales (n12): Principal organizer of the cocaine import, intermediary between the Colombians and the Serero organization was able to get more importance in the upcoming phases because of that.

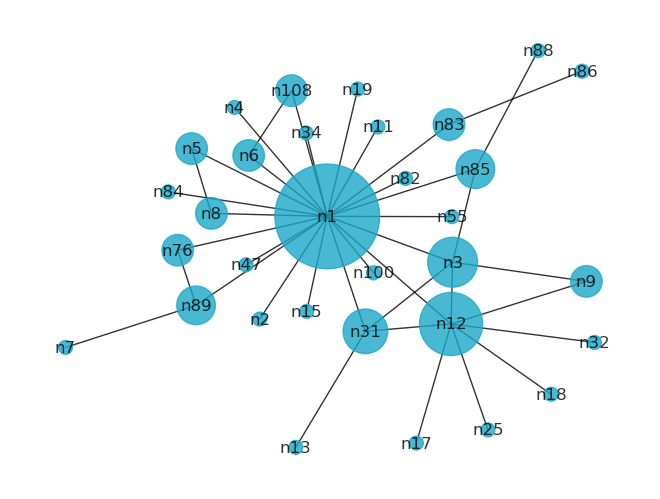
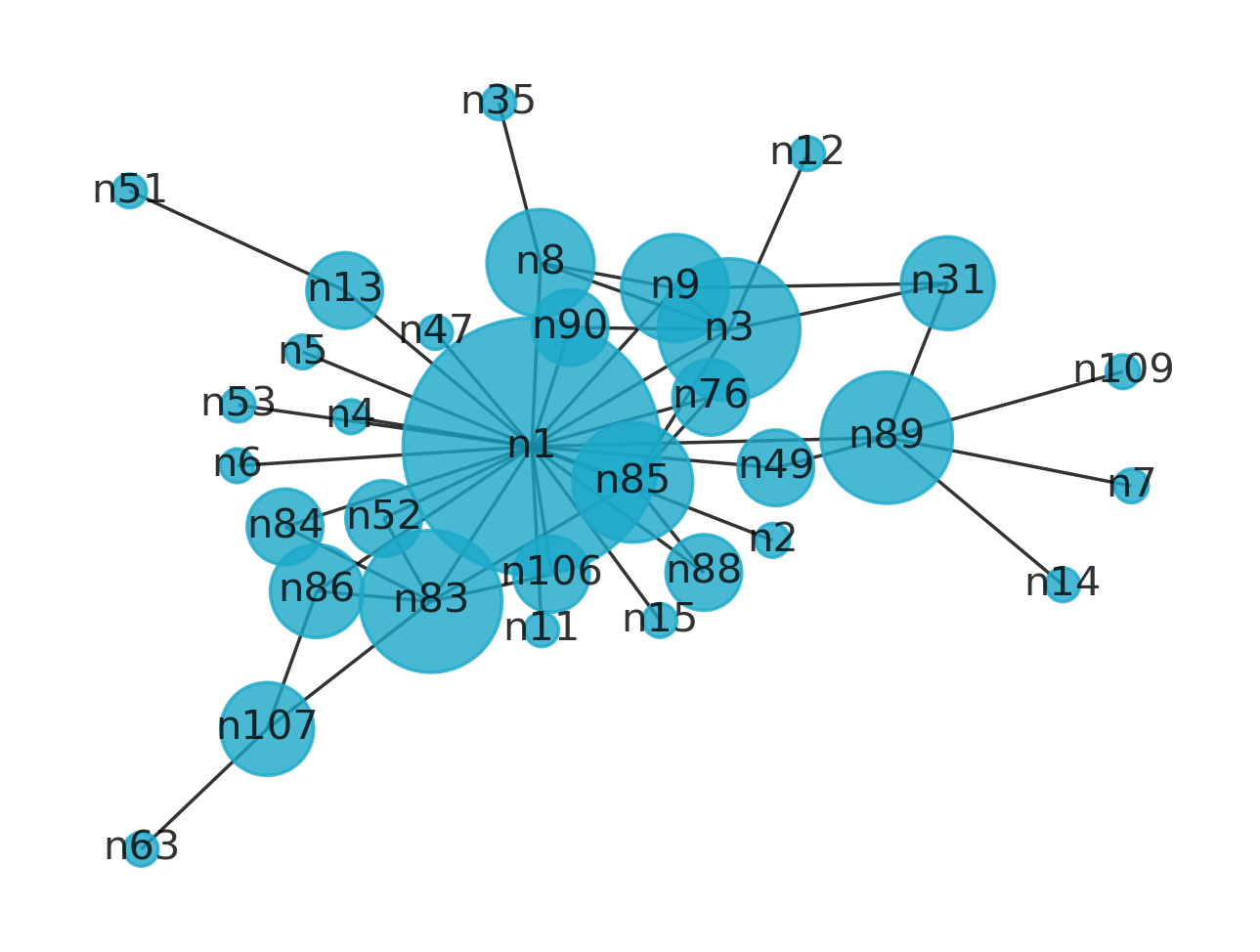


Figure 7 Degree centrality (phase 4) Figure 8 Degree centrality (phase 5)

**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?**

* Table 1 shows the graph at each phase. The graph started with fewer nodes meaning at the phase 1 the main dealer in suspicion is wiretapped and his connections are established as nodes. The connection and nodes increased until phase 4 where the first seizure happened. Until phase 4 Daniel Serero (n1 - Mastermind of the network) maintained the connections. After phase 4, there is a change in the cluster. Antonio Iannacci (n89) lost connections. Probably, the seizure would have affected him. The criminals had to change paths. Cocaine became next possible drug for trafficking. Ernesto Morales (n12 -Principal organizer of the cocaine import, intermediary between the Colombians and the Serero organization), started gaining connections from phase 5. His connections increased in further phase. Ernesto seems to have direct connection to Pierre Perlini (n3 - Principal lieutenant of Serero, he executes Serero's instructions) in many of the phases from phase 5 and maintained a short path between him and Ernesto. Pierrie (n3). Ernesto acted as a main communication between Daniel Serero (n1) and other dealers from phase 5. Perlini also had his own network disconnected from the main network in phase 7 and 10.

Table Graphs for each phase

|  |  |  |  |
| --- | --- | --- | --- |
| Figure 8 Phase -1 | Figure 9 Phase 2 | Figure 10 Phase 3 | Figure 11 Phase 4 |
| Figure 12 Phase 5 | Figure 13 Phase 6 | Figure 14 Phase 7 | Figure 15 phase 8 |
| Figure 16 Phase 9 | Figure 17 Phase 10 | Figure 18 phase 11 |  |

**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.**

There were other players not listed in the 23 main player list at different phases who played important role. Player 31 acted as a main communication point between n12 and n1 in phase 5 after the first seizure. n14 and n22 acted as main communication between n12 and n3 in the phase 8. But there was a big seizure of cocaine in phase 8. After that n14 stayed with n12 , but as a peripheral node.

**The remaining two questions will concern the directed graphs derived from the CAVIAR data.**

**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?

**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?**

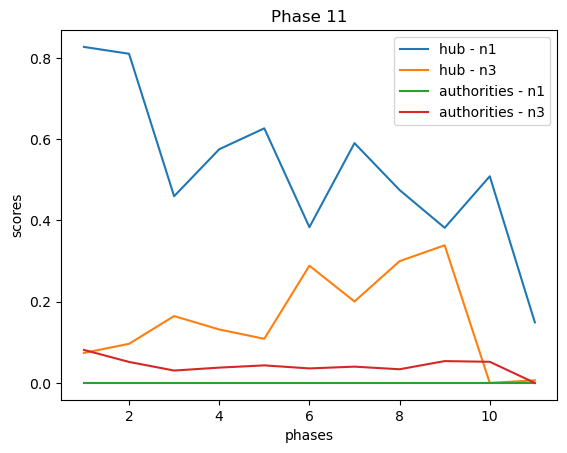
In the criminal network analysis, directed graph will show the hierarchy of the connections. Depends on how the connections are made, we can infer flow of network. Also breaking the connection to the hierarchical branch with many connections and disrupt the trafficking and reduce the crime. Assuming the connections are made from who gives order to who, then this graph will help in identifying the actor with high influence, actor with more authority.

The left eigen vector centrality will measure the in-degree for the directed graph and the right eigen vector centrality will measure the out-degree.

**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.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)?**

Following figure showcases the plot for hubs and authorities score for various stages for n1 and n3. n1 showed the high hub value meaning the main connection source in the phase 1. The connection decreases and fluctuate in the mid-range for the rest of the phases. n3 did not have high impact in the initial phase. The hub score increases after phase 4. n3 became the next high connection in the network after phase 4. Both has a big dip in phase 10 where there was a big seizure and in phase 11, the graph represented a wide branch of network with denser connections to n12.



**Project:**

**Abstract:**

For the project, I have chosen the ego-facebook data. The goal is to run an analysis on the edges and understand the connection patterns and highly active ego.

**Dataset:**

The Facebook ego data contains data for connections and circles for each user. The user is called ego and all the user’s connections are nodes in the graph. There will be one graph per ego. Each user can create a group of connections called circle to share information private to them. In the dataset, the circle for each user is given in file with extension “. circle”. This file contains the l set of circles for the ego node. Each line contains one circle, consisting of a series of node ids. The first entry in each line is the name of the circle.

“. edges” file contains the edges for ego id in the filename.

**Analysis:**

For this project, only analysis is performed on the edges to identify the highly influential edge in the network. There are data for 9 ego in the dataset

**Degree**:

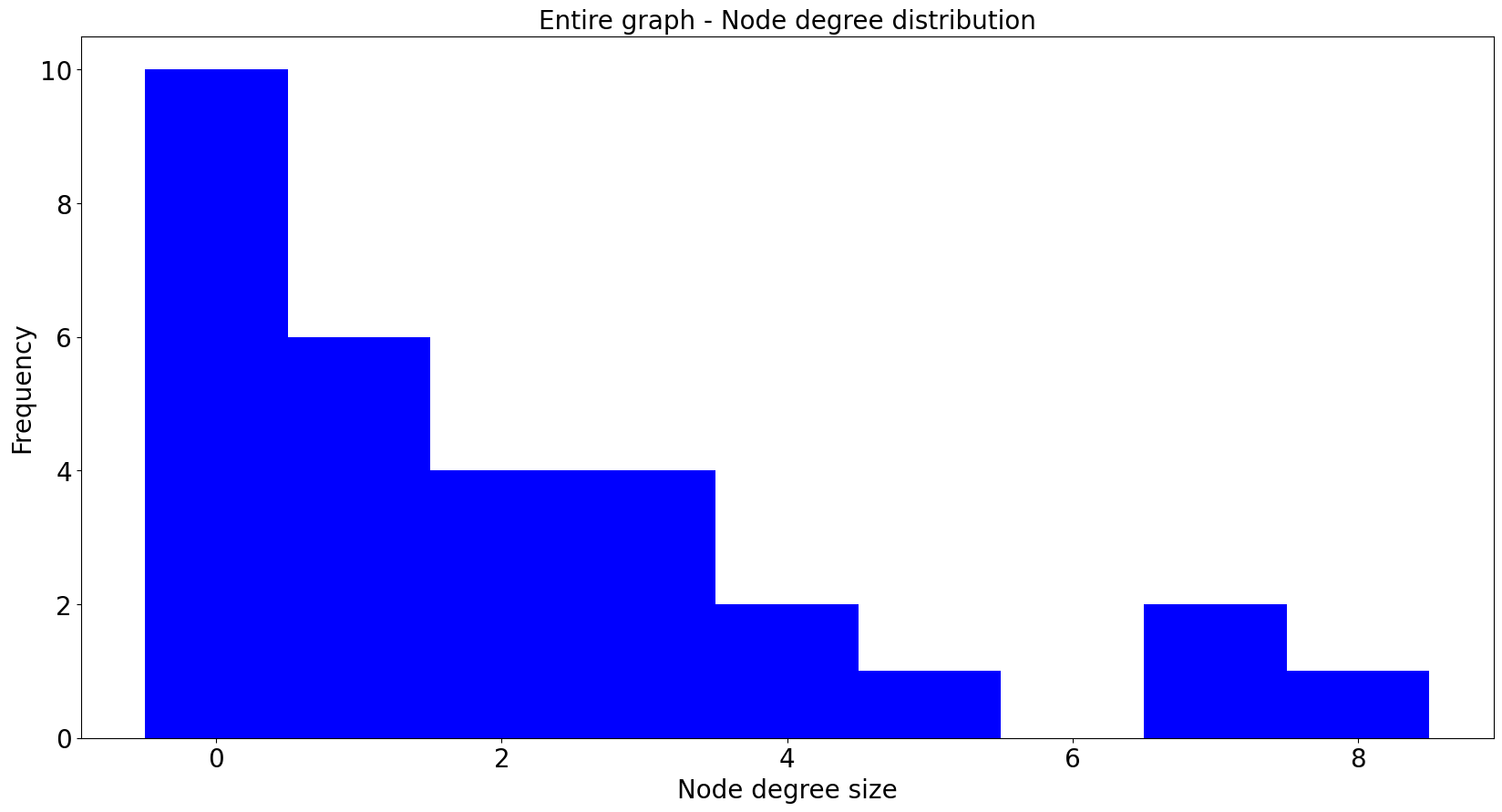
Here is the degree statistics for each ego

Ego0: average = 4.26923, median = 1.50000, standard deviation = 5.98861

Ego107: average = 4.07087, median = 2.00000, standard deviation = 5.18049 Ego1684: average = 5.73723, median = 4.00000, standard deviation = 5.69065 Ego1912: average = 2.54082, median = 2.00000, standard deviation = 2.77685 Ego3437: average = 4.94444, median = 1.00000, standard deviation = 7.21859 Ego348: average = 2.24000, median = 2.00000, standard deviation = 2.20055 Ego3980: average = 2.73684, median = 2.00000, standard deviation = 2.89904 Ego414: average = 2.58621, median = 2.00000, standard deviation = 2.10938

Ego686: average = 2.15385, median = 1.00000, standard deviation = 2.51190

Ego698: average = 2.03333, median = 1.00000, standard deviation = 2.25807



**Centrality measures:**

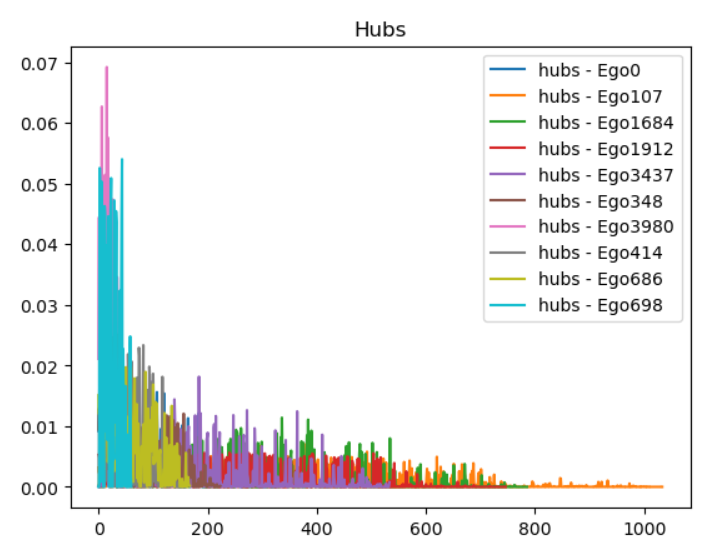
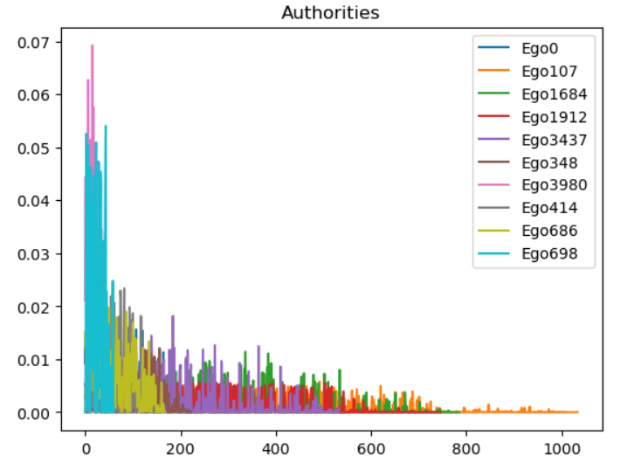
We then perform a quantitative analysis on the connections to find the most influential node across all ego connects. Comparing all the centrality measures, we see that user 376 and 856 are more influential. 376 and 856 are on the top 10 list for all the metrics.

Table Top 10 nodes by centralities score

|  |  |  |
| --- | --- | --- |
| Degree Centrality | Betweenness Centrality | Eigen Centrality |
|  |  |  |

**Hubs and authorities:**

In the hubs and authorities measure, we can see that Ego 3980 has high connections and is the most connected user. Ego id 698 takes the next role.



Conclusion:

Based on the analysis so far, we could infer that edgo id 3980 is the most connected user based on the hub score even though 3980’s average degree is not the top. Among all the networks 376 and 856 are connected across network. Hence they both are identified as most influential connections. We could use the nodes 376, 856 as a main communication point for spreading the information faster.

References:

1. <https://www.researchgate.net/publication/292304919_Modeling_Verdict_Outcomes_Using_Social_Network_Measures_The_Watergate_and_Caviar_Network_Case>
2. <https://en.wikipedia.org/wiki/Co-citation>
3. <https://en.wikipedia.org/wiki/Bibliographic_coupling>
4. <https://en.wikipedia.org/wiki/Citation_impact>
5. <https://snap.stanford.edu/data/ego-Facebook.html>

# Appendix:

1. **Part(c) example code:**

import numpy as np

import time

def getC(A):

    dim = A.shape

    C = np.zeros(dim)

    for r in A:

        if np.sum(r)>1:

            for a in range(len(r)):

                for b in range(len(r)):

                    if a==1 and b==1:

                        C[a][b]=1

    return C

for n in [10,100,1000]:

    print("n={}".format(n))

    A = np.random.choice([0,1],n\*n, p = [0.8,0.2])

    A = np.reshape(A,[n,n])

A = np.random.choice([0,1],n\*n, p = [0.8,0.2])

A = np.reshape(A,[n,n])

start = time.time()

getC(A)

end = time.time()

print("Time to complete the Friend's algorithm = ", (start-end))

start = time.time()

C = A@A.T

end = time.time()

print("Time to complete the matrix algorithm = ", (start-end))

OUTPUT:

n=10

Time to complete the Friend's algorithm = -0.0012998580932617188

Time to complete the matrix algorithm = 0.0

n=100

Time to complete the Friend's algorithm = -0.03846120834350586

Time to complete the matrix algorithm = -0.00800466537475586

n=1000

Time to complete the Friend's algorithm = -39.36552095413208

Time to complete the matrix algorithm = -1.2720036506652832