Written Report of Module 3 - Network Analysis

Tien Minh Dam {edX: @damminhtien}

12 July 2024

Problem 1: Citation Network - Suggesting Similar Papers

A citation network is a directed network where the vertices are academic papers and there is a directed edge from paper A to paper B if paper A cites paper B in its bibliography. Google Scholar performs automated citation indexing and has a useful feature that allows users to find similar papers. In the following, we analyze two approaches for measuring similarity between papers.

Part (a): Co-citation network

Two papers are said to be **cocited** if they are both cited by the same third paper. The edge weights in the cocitation network correspond to the number of cocitations. In this part, we will discover how to compute the (weighted) adjacency matrix of the cocitation network from the adjacency matrix of the citation network.

- Problem setup: In order to derive the cocitation matrix, we need to derive it as a function of the original adjacency matrix.
- Problem notation: If there is an edge from paper i to paper j, it means that paper i cites paper j. We will denote by A the corresponding adjacency matrix, such that $A_{ij} = 1$ means there is a directed edge from i to j. Let us denote by C the cocitation network matrix.

Part (b): Bibliographic coupling

Two papers are said to be bibliographically coupled if they cite the same other papers. The edge weights in a bibliographic coupling correspond to the number of common citations between two papers.

Part (c): Algorithm complexity of Co-citation matrix construction

1. *(2 points) How does the time complexity of your solution involving matrix multiplication in part (a) compare to your friend's algorithm? As above, for a brief introduction to the big-O notation, refer to the optional problem 1.7 in Module 1.*

   Answer. To answer this question, firstly, I recall My Friend Algorithm in the Question 1 as Algorithm 1:

Algorithm 1 Mah friend algo pseudocode

Require: Adjacency matrix A of size $n \times n$
Ensure: Co-citation matrix C of size $n \times n$

Initialize an empty matrix C of size $n \times n$ with zeros
for each row r in A **do**
  if sum of row r in A is strictly greater than 1 **then**
    for each a such that $A(r, a) \neq 0$ **do**
      for each b such that $A(r, b) \neq 0$ and $a \neq b$ **do**
        $C(a, b) \leftarrow C(a, b) + 1$
        $C(b, a) \leftarrow C(b, a) + 1$
      end for
    end for
  end if
end for
return C

My friend algorithm, while returns extractly what we want is the co-citation matrix, is quite ineffective due to the high big-O complexity. As we can see in the 1, my friend has 3 nested loop in the code, so the algorithm is $\mathcal{O}(n^3)$.

Now, let's analysis My Algorithm as Algorithm 2:

Algorithm 2 Mah algo pseudocode

Require: Adjacency matrix A of size $n \times n$
Ensure: Co-citation matrix C of size $n \times n$

Compute $C = A^T * A$
return C

The natural complexity of multiplying two $n \times n$ matrices is still $\mathcal{O}(n^3)$ without any additional optimizations. It implies that My Friend Algorithm and My Algorithm to construct the co-citation matrix is in the same level of big-O complexity theoretically.

However, in practical, when we run the code on Numpy or Cuda library (with GPU), the runtime is sightly different between two approaches. The loopness approach is quite slow with large and sparse matrix, however, the approach based on matrix is lighting-fast ϟ. It looks like these library implements several highly-specialized tricks (SIMD (Single Instruction Multiple Data), cuBLAS (CUDA Basic Linear Algebra Subroutines),...) for matrix operations under-the-hood.

Besides, I came across Strassen's algorithm, an algorithm for matrix multiplication that has a time complexity of approximately $\mathcal{O}(n^{2.81})$. This algorithm reduces the number of multiplicative operations at the expense of increased additive operations and complexity. □

Part (d): Co-citation vs. Bibliographic Coupling

1. *(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. To answer this question, we firstly analysis different aspects of Bibliographic coupling and Cocitation in Table 1.

Table 1: Comparison of Bibliographic Coupling and Cocitation method for identifying paper similarity

	Bibliographic coupling	Cocitation
Definition	Two papers are bibliographically coupled if they cite the same third paper(s).	Two papers are cocited if they are cited together by other papers.
Mechanism	The number of shared references between two papers determines the strength of their relationship.	The number of times two papers are cited together by subsequent papers determines the strength of their relationship.
Initial vs. Accumulated Data	Initial Data - relies on the initial set of references a paper cites upon its publication. Accumulated Data - does not change over time; the relationships are fixed once the paper is published. Limited adaptability to new trends or emerging fields.	Initial Data - starts with no information and builds as new papers are published. Accumulated Data - continuously accumulates data, leading to potentially richer and more dynamic insights. High adaptability to emerging trends and evolving research landscapes.
Scope of analysis	Limited to the scope of the references cited by the papers. Scope is narrower, as it only considers direct references.	Broader scope as it considers all future papers that cite the original papers. Scope is broader, as it includes the influence of subsequent citations.
Use case suitability	Best suited for analyzing the initial context in which papers are published. Ideal for understanding the immediate academic environment and historical foundations. Provides a snapshot of research connections at the time of publication.	Best suited for analyzing ongoing impact and current relevance. Ideal for understanding how research is perceived and connected in the broader academic community over time. Provides a dynamic view of research relationships as they evolve.
Strengths vs. Weaknesses	Strengths are simplicity, immediate relevance upon publication. Weaknesses are static nature, less reflective of ongoing research dynamics.	Strengths are dynamic, captures evolving relevance and connections. Weaknesses are requires extensive and ongoing citation data, can be complex to analyze.

In conclusion, to directly answer the question which is better, I think they are both good in their context. Cocitation is generally more appropriate for indicating similarity between papers, especially when the goal is to understand current relevance and evolving research connections. It dynamically reflects how papers are perceived and related in the academic community over time. Bibliographic Coupling is useful for understanding the initial academic context and historical foundations of research but lacks the dynamic perspective that cocitation provides.

□

Problem 2: CAVIAR Network - Investigating a time-varying criminal network

In this problem, you will study a time-varying criminal network that is repeatedly disturbed by police forces. The data for this problem can be found in the CAVIAR directory of the data archive.

The CAVIAR investigation lasted two years and ran from 1994 to 1996. The operation brought together investigation units of the Montréal police and the Royal Canadian Mounted Police of Canada. During this two year period, 11 wiretap warrants, valid for a period of about two months each, were obtained (the 11 matrices contained in phase1.csv, phase2.csv, ... correspond to these eleven, two month wiretap phases).

This case is interesting because, unlike other investigative strategies, the mandate of the CAVIAR project was to seize the drugs without arresting the perpetrators. During this period, imports of the trafficking network were hit by the police on eleven occasions. The arrests took place only at the end of the investigation. Monetary losses for traffickers were estimated at 32 million dollars. Eleven seizures took place throughout the investigation. Some phases included no seizures, and others included multiple. The following summarizes the 11 seizures:

Table 2: CAVIAR Phase at the glance

Phase ID	No. Seizure	Amount of money	
Phase 4	1	$2,500,000	300 kg of marijuana
Phase 6	3	$1,300,000	2 x 15 kg of marijuana + 1 x 2 kg of cocaine
Phase 7	1	$3,500,000	401 kg of marijuana
Phase 8	1	$360,000	9 kg of cocaine
Phase 9	2	$4,300,000	2 kg of cocaine + 1 x 500 kg marijuana
Phase 10	1	$18,700,000	2200 kg of marijuana
Phase 11	2	$1,300,000	12 kg of cocaine + 11 kg of cocaine

This case offers a rare opportunity to study a criminal network in upheaval from police forces. This allows us to analyze changes in the network structure and to survey the reaction and adaptation of the participants while they were subjected to an increasing number of distressing constraints.

The network consists of 110 (numbered) players. Players 1-82 are the traffickers. Players 83-110 are the non-traffickers (financial investors; accountants; owners of various importation businesses, etc.). Initially, the investigation targeted Daniel Serero, the alleged mastermind of a drug network in downtown Montréal, who attempted to import marijuana to Canada from Morocco, transiting through Spain. After the first seizure, happening in Phase 4, traffickers reoriented to cocaine import from Colombia, transiting through the United States.

According to the police, the role of 23 of the players in the "Serero organization" are the following, listed by name (unique id):

- Daniel Serero (n1): Mastermind of the network.
- Pierre Perlini (n3): Principal lieutenant of Serero, he executes Serero's instructions.
- Alain (n83) and Gérard (n86) Levy: Investors and transporters of money.
- Wallace Lee (n85): Takes care of financial affairs (accountant).
- Gaspard Lino (n6): Broker in Spain.
- Samir Rabbat (n11): Provider in Morocco.

- Lee Gilbert (n88): Trusted man of Wallace Lee (became an informer after the arrest).
- Beverly Ashton (n106): Spouse of Lino, transports money and documents.
- Antonio Iannacci (n89): Investor.
- Mohammed Echouafni (n84): Moroccan investor.
- Richard Gleeson (n5), Bruno de Quinzio (n8) and Gabrielle Casale (n76): Charged with recuperating the marijuana.
- Roderik Janouska (n77): Individual with airport contacts.
- Patrick Lee (n87): Investor.
- Salvatore Panetta (n82): Transport arrangements manager.
- Steve Cunha (n96): Transport manager, owner of a legitimate import company (became an informer after the arrest).
- Ernesto Morales (n12): Principal organizer of the cocaine import, intermediary between the Colombians and the Serero organization.
- Oscar Nieri (n17): The handyman of Morales.
- Richard Brebner (n80): Was transporting the cocaine from the US to Montréal.
- Ricardo Negrinotti (n33): Was taking possession of the cocaine in the US to hand it to Brebner.
- Johnny Pacheco (n16): Cocaine provider.

In the data files (phase1.csv, phase2.csv, ...), you will find matrices that report the number of wiretapped correspondences between the above players in the network, where players are identified by their unique id. You will be analyzing this time-varying network, giving a rough sketch of its shape, its evolution and the role of the actors in it.
Paper: `https://www.researchgate.net/publication/292304919_Modeling_Verdict_Outcomes_Using_Social_Network_Measures_The_Watergate_and_Caviar_Network_Cases`

Part (a): Basics visualization

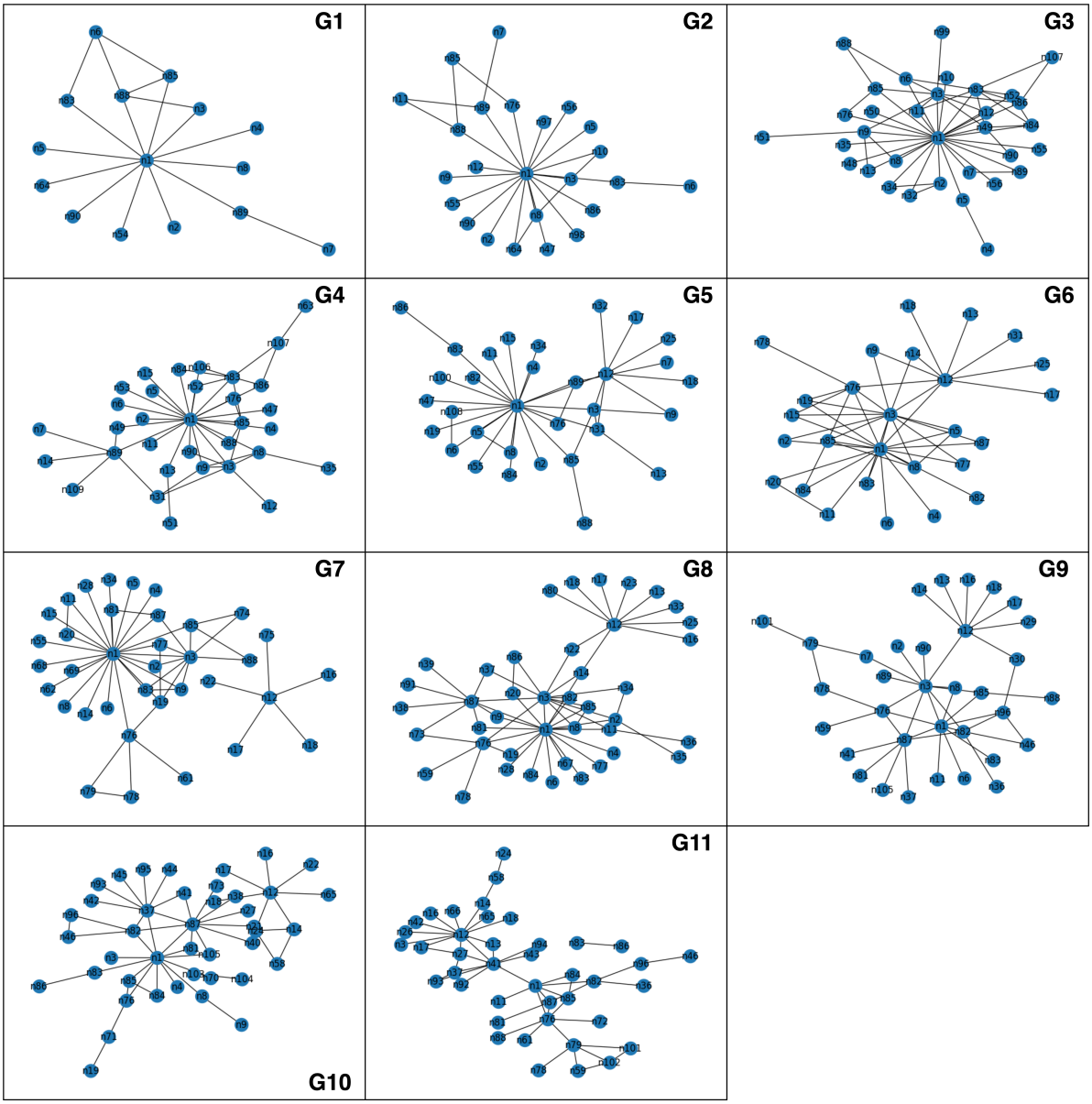

Figure 1: CAVIAR networks in one picture

Part (b): Centrality

(b1) Degree centrality

(b2) Betweenness centrality

(b3) Eigenvector centrality

Part (c): Changes overtime in the network

1. *(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?*

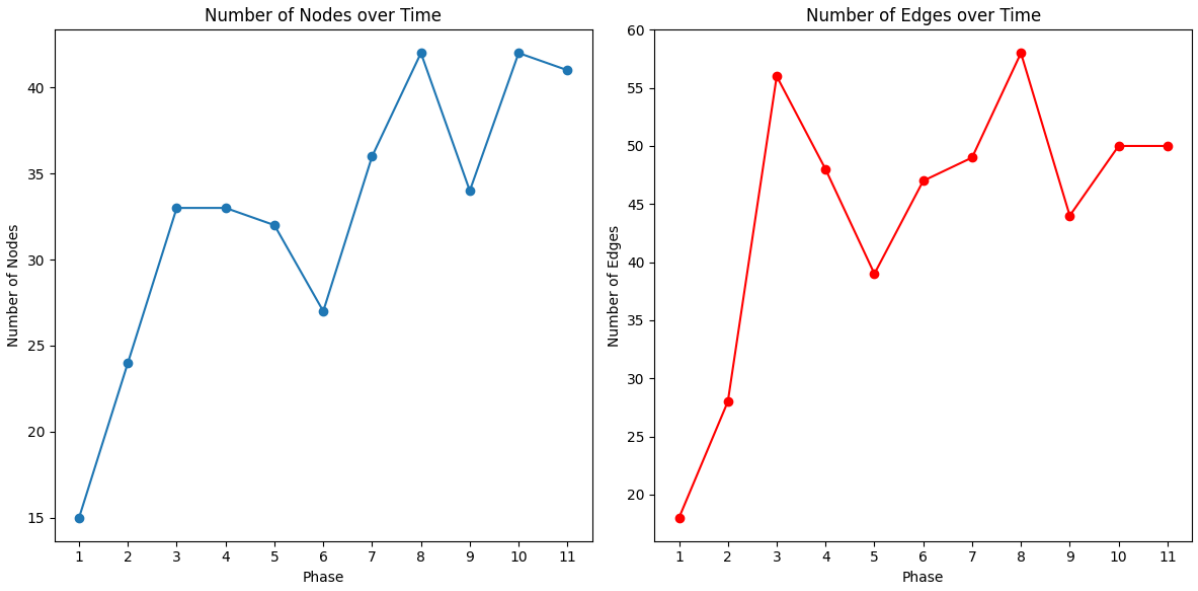

Figure 2: Number of node and edge CAVIAR network over time

Answer. As we can see in Figure 2, the number of nodes increases sharply over the first few phases then levels out. The initial sharp increase in the number of nodes indicates the period of rapid network expansion and identification of new actors. The number of nodes falls down later because the initial identification and inclusion of key players in the network by the police forces. As the investigation progresses, the network stabilizes as fewer new actors are identified. Once the network stabilizes, the centrality measures become more reliable as indicators of consistent player involvement.

In the question 5 in Part (b), I believe that adjustments to conclusions should consider the dynamic nature of the early phases. Temporal consistency in centrality is more meaningful during stable phases. The mean centrality values should be interpreted with an understanding of the network's growth pattern, ensuring comparisons are fair and contextualized within the investigation's timeline. □

Part (d): Meaning of centrality metrics

1. *(5 points) 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.*

   Answer. To answer this question, firstly, I recall the definition of 3 centrality metrics:

   1. Degree centrality measures the number of direct connections an actor has.
   2. Betweenness centrality measures how often an actor appears on the shortest paths between other actors.
   3. Eigenvector centrality measures an actor's influence based on the number and quality of connections, giving higher scores to nodes connected to other high-scoring nodes.

   These terms leads to the insights in criminal network as follow:

1. Actors with high degree centrality are highly connected and can interact with many others directly. In a criminal network, these individuals might be brokers or coordinators who manage several relationships, facilitating communication and resource flow.
2. Actors with high betweenness centrality control information flow and can influence communication and coordination between different parts of the network. In a criminal network, these individuals might act as gatekeepers or intermediaries, controlling access to resources and information.
3. Actors with high eigenvector centrality are connected to other highly influential actors. In a criminal network, these individuals might be leaders or highly trusted members who have access to other influential network members.

So, We come up with the main issues of degree centrality:

- Local view - degree centrality only considers immediate connections and does not capture the broader network structure.
- Neglects influence - an actor with many low-influence connections may appear more central than an actor with fewer, but more influential, connections.
- Vulnerability - focusing on degree centrality may miss key players who maintain low visibility but hold critical strategic positions.

Choosing the Most Relevant Metric: In my opinion, betweenness centrality is the most relevant metric for identifying who is running the illegal activities of the group because of its position in information flow, strategic importance, and network vulnerability. My justify as follow:

- Control over information flow -leaders or coordinators in a criminal network often control the flow of information and resources. Betweenness centrality identifies these strategic positions, highlighting those who act as intermediaries and gatekeepers.
- Network vulnerability - disrupting actors with high betweenness centrality can fragment the network, making it an effective strategy for law enforcement. These individuals are crucial for the network's cohesion and operations.
- Strategic importance - actors with high betweenness centrality can exert significant influence over the network, even if they do not have the most connections (degree centrality) or are not connected to other highly influential actors (eigenvector centrality).

□

Part (e): Meet the bosses

1. *(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.*

   Hint: Note that the definition of a player's "importance" (i.e. how central they are) can vary based on the question you are trying to answer. Begin by defining what makes a player important to the group (in your opinion) ; use your answers from Part (d) to identify which metric(s) are relevant based on your definition and then, use your quantitative analysis to identify the central and peripheral traffickers. You may also perform a different quantitative analysis, if your definition of importance requires it.

Answer. As mentioned in the last question, I believe that the betweeness centrality is the most important metric to disrupt this criminal network. In this question, I also combine the eigenvector centrality to identify the bosses of criminal. Using the results from Part (b) Question 5, we can integrate and interpret the information to identify central and peripheral traffickers:

- Player n1 (Daniel Serero) - highest betweenness centrality, indicating he controls critical pathways within the network. Highest eigenvector centrality, indicating he is highly influential and connected to other key players. He is likely the mastermind and central coordinator of the network, managing both direct interactions and broader influence.
- Player n12 (Ernesto Morales) - second highest betweenness centrality, indicating a strategic role in bridging different parts of the network. He is likely a key intermediary, possibly responsible for coordinating specific operations or resources.
- Player n3 (Pierre Perlini) - third highest betweenness centrality, indicating significant influence over communication pathways. Second highest eigenvector centrality, showing strong connections with other influential members. He is likely a principal lieutenant, executing instructions and maintaining network cohesion.
- Player n2 (Unknown, Hypothetical) - third highest eigenvector centrality, indicating significant overall influence. He is likely a trusted associate or advisor, connected to other central figures in the network.

Understanding these roles allows law enforcement to target key individuals effectively to disrupt the network. In this case, Daniel Serero (n1) emerges as the most crucial player, followed by Ernesto Morales (n12) and Pierre Perlini (n3). These players' central roles suggest they are pivotal in running the illegal activities of the group. □

Part (f): Overall evolution of the network and Correlate the patterns

Now, we will attempt to analyze the overall evolution of the network and correlate the patterns we observe to events that happened during the investigation.

1. *(3 points) 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. In Part f, Question 1, We visually identify the correct $X = 4$. So, in this question, I first analyze the eigenvector centrality and betweeness centrality of the criminal in phase 4 and phase 5. Note that: from the introduction of this problem, we all know in the phase 4, monetary losses for traffickers were estimated at 2.500.000 USD, 300 kg of marijuana is taken by the law enforcement. It implies the possibility of a significant refactor in criminal flow to patch the problem.

Table 3: Eigenvector centrality and betweenness centrality for the top players in both phases

Player	Phase X Eigenvector	Phase X+1 Eigenvector	Phase X Betweenness	Phase X+1 Betweenness
n1	0.6104	0.6402	0.8393	0.8839
n3	0.2726	0.2757	0.0904	0.0441
n89	0.1629	0.1568	0.1962	0.0645
n83	0.2710	0.1267	0.0796	0.0645
n85	0.2517	0.1813	0.0165	0.0645
n12	0.0472	0.2810	0.0000	0.2699

I summarize the centrality for the top players in both phases in Table 3, then visualize them in bar chart plotted in Figure

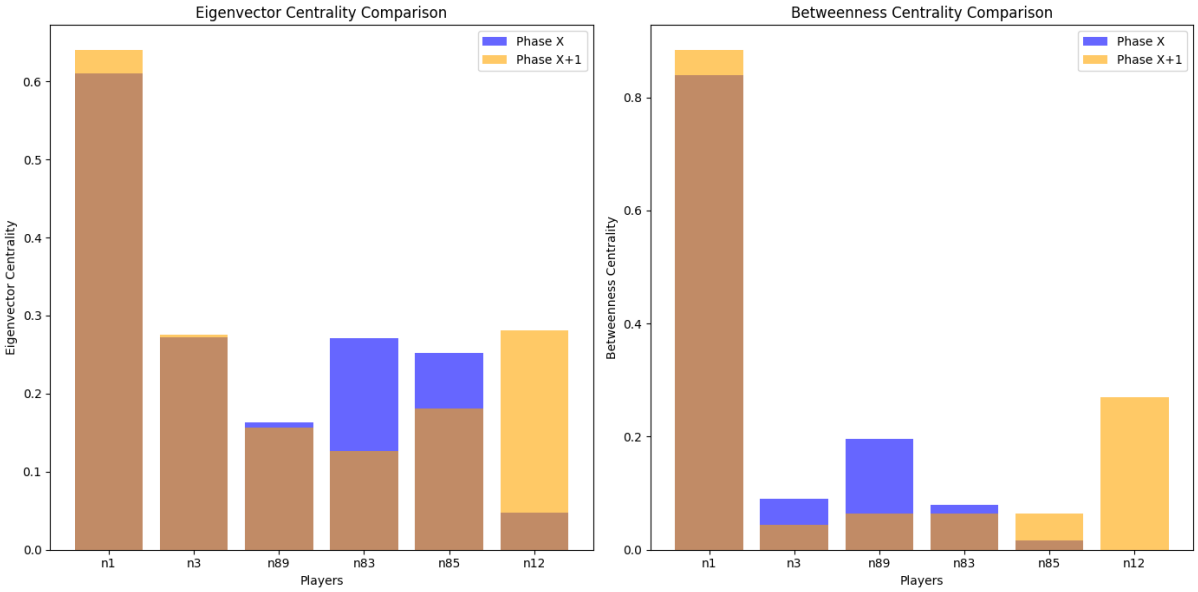

Figure 3: Bar chart: Eigenvector centrality and betweenness centrality for the top players in both phases X and X+1

The change of eigenvector centrality as follow:

- n1 (Daniel Serero) maintains the highest eigenvector centrality, indicating his continued influence and central role in the network.
- n3 (Pierre Perlini) slight increases in eigenvector centrality, indicating a stable influence.
- n89 and n83 slight decreases in eigenvector centrality, indicating a reduction in their relative influence.
- n85 decreases in eigenvector centrality, reflecting a decline in influence.
- n12 significant increases in eigenvector centrality, suggesting a rise in importance and connections with influential actors.

The change of betweeness centrality as follow:

- n1 (Daniel Serero) maintains the highest betweenness centrality, indicating his continued control over key pathways in the network.
- n3 (Pierre Perlini) decreases in betweenness centrality, suggesting a reduced role in controlling communication pathways.
- n89, n83, and n85 reduced betweenness centrality, indicating diminished roles as intermediaries.
- n12 significant increases in betweenness centrality, indicating a rise in strategic importance and control over key communication pathways.

Act as a real police, I guess what happen:

- Increased centrality of n12 implies the rise in importance of n12 suggests that Ernesto Morales or a similar figure stepped up to fill the gap created by the disruption. This change could be due to the network's reorganization in response to the seizure.
- Despite the disruption, Daniel Serero (n1) remains central, indicating his strong control and ability to adapt to changes.

- The decrease in centrality for players like n83, n89, and n85 suggests a shift in their roles, possibly due to the network's restructuring to mitigate the impact of the seizure.

The rise in centrality of certain players like n12 and the stability of n1 highlight the network's adaptability and the central figures' resilience. The analysis shows how the network responded to a major event, such as a police seizure, by reorganizing and adjusting the roles and importance of its members. □

Part (g): Global trends and incidents

1. *(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. Remind Table 2 ,the data of police shows the incident of criminal network in phase 4, 6, 7, 8, 9, 10, 11 where multiple seizures occurs. The biggest one is in phase 6 (3 seizures), follow by phase 9 and 10 after (2 seizures). To answer this question, firstly, I recall the Figure 2 which plot the change of nodes and edges in CAVIAR network:

   (g1) Number of nodes over time

   - There is a sharp increase in the number of nodes from Phase 1-3, 6-8, 9-10 indicating the identification and inclusion of new actors in the network.
   - The number of nodes stabilizes between Phases 3 and 5, suggesting a period of consolidation where the core network is established.
   - A noticeable drop in nodes in Phase 6 and 9 may indicate a disruption, such as a police seizure or internal conflict, leading to the removal of certain actors.
   - Subsequent phases show fluctuations but generally reflect a stable network with minor adjustments.

   (g2) Number of edges over time

   - The number of edges shows a similar initial increase, peaking around Phase 3 and 8, which correlates with the highest activity and connectivity within the network.
   - After Phase 3 and 8, there are a notable decreases, likely reflecting the impact of law enforcement actions disrupting the network's connectivity.
   - The network attempts to recover in later phases, showing fluctuations in connectivity.

(g3) Centrality rank over time

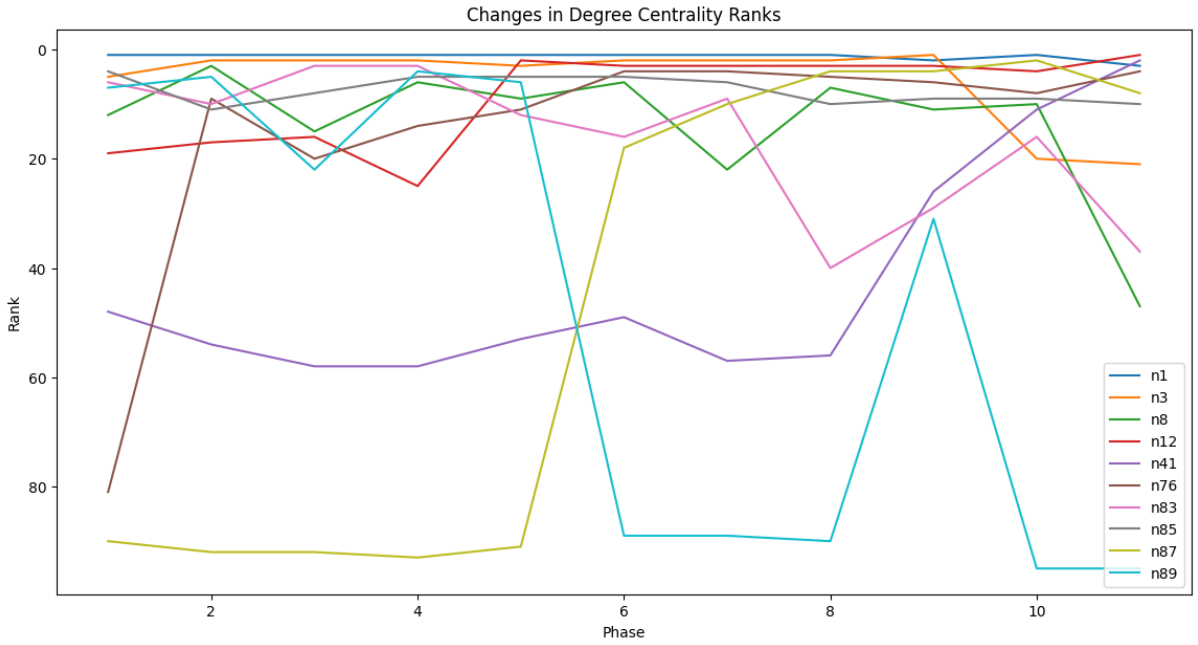

Figure 4: Changes in Degree Centrality Ranks Over Time

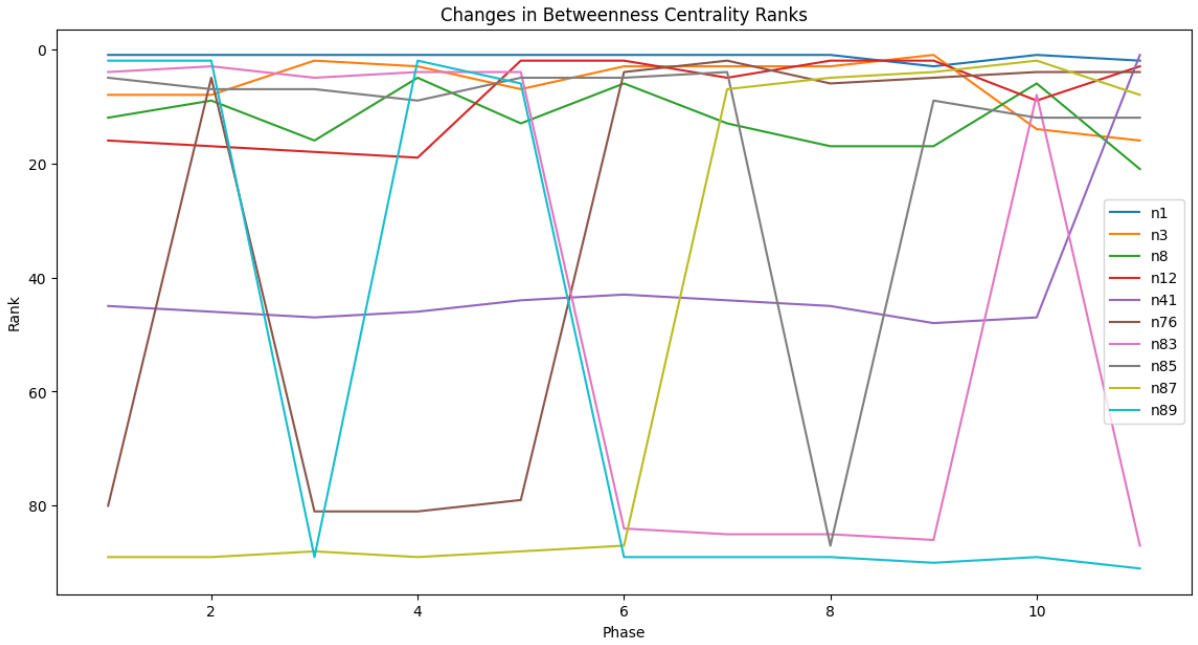

Figure 5: Changes in Betweeness Centrality Ranks Over Time

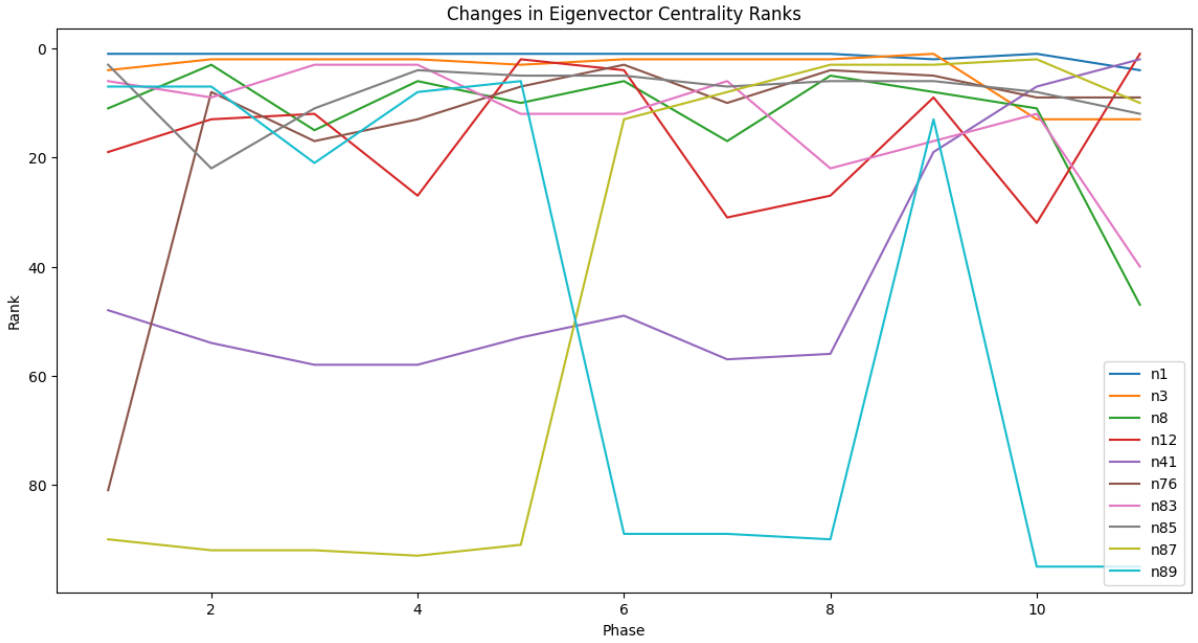

Figure 6: Changes in Eigenvector Centrality Ranks Over Time

As shown in Figure 4, 5, 6, We analyze the changes in Centrality Ranks Over Time of top 10 highest ranked players:

1. Player n1 (Daniel Sereno) - blue line: truly the mastermind of the network, the rank always in the top consistently.
2. Player n3 (Pierre Perlini) - orange line: principal lieutenant of Serero, he executes Serero's instructions, the rank of him remain stable, drop a little bit after phase 9.
3. Player n8 (Bruno de Quinzio) - green line: charged with recuperating the marijuana, the rank of him remain stable in top 5, only drop slightly after phase 10.
4. Player n12 (Ernesto Morales) - red line: principal organizer of the cocaine import, intermediary between the Colombians and the Serero organization, he is grow very fast after phase 4.
5. Player n41 (A danger person - We dont have him in police database): his rank is above 40th in the first phases but after phase 8 he climb very very fast on the criminal status ladder, in last phase, his rank is beyond the rank of player n1 Sereno. This phenomenon indicates him as a new important player in CAVIAR network. The police should take care him carefully later.
6. Player n76 (Gabrielle Casale) - brown line: same mission as player n8, he retire in phase 3,4,5 but recover in phase 6.
7. Player n83 (Alain) - pink line: investor, his rank go down very fast after phase 5, he retire and only work in phase 10 after.
8. Player n85 (Wallace Lee) - olive line: accountant of CAVIAR, he only retire in phase 8.
9. Player n87 (Patrick Lee) - gray line: investor, only highly-participate in CAVIAR network after phase 6.
10. Player n89 (Antonio Iannacci) - cyan line: investor, he retire after seizures of phase 6.

(g4) Correlation with known events

- Phase 4 Seizure: Initial disruption leading to network reorganization, with n12 starting to rise in importance.
- Phase 6 Multiple Seizures: Significant reorganization, with n87, n76, n87 showing a marked increase in centrality, likely taking on a crucial role to stabilize the network. Player n89 is likely dropped out after this phase.
- Phase 7-8 Major Seizure: Continued adaptation with key players like n87 and n85 maintaining high centrality.
- Phase 9 Seizures: not thing down but the come back of n85.
- Phase 10-13 Seizures: the rise of n41, n12, the drop of n83.

In conclusion, the network evolution reflects the background story. □

Part (h): The Invisible-man

1. *(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, as mention before, the player n41 is not on the list of investigation, we dont know him. He claim the highest centrality in the last final phase. As can be seen from Figure 5, his rank is above 40th in the first phases but after phase 8 he climb very very fast on the criminal status ladder, in last phase, his rank is beyond the rank of player n1 Sereno. This phenomenon indicates him as a new important player in CAVIAR network. The police should take care him carefully later. □

Part (i): Advantages of directed graph in criminal network

1. *(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. Analyzing a directed version of the criminal network offers significant advantages over an undirected one. Directed graphs capture the directionality of relationships, crucial for understanding who commands or influences whom within the network. By comparing in-degree and out-degree centralities, we can differentiate between coordinators (high in-degree) and leaders (high out-degree). Moreover, left-eigenvector centrality reveals actors central to influential members, highlighting key coordinators, while right-eigenvector centrality identifies those who influence other key actors, indicating leaders or controllers of operations. This nuanced understanding of roles and influences is vital for law enforcement to disrupt the network effectively, targeting not only prominent figures but also those pivotal in maintaining the network's cohesion and functionality.

   I also plot some centrality charts as follow:

   □

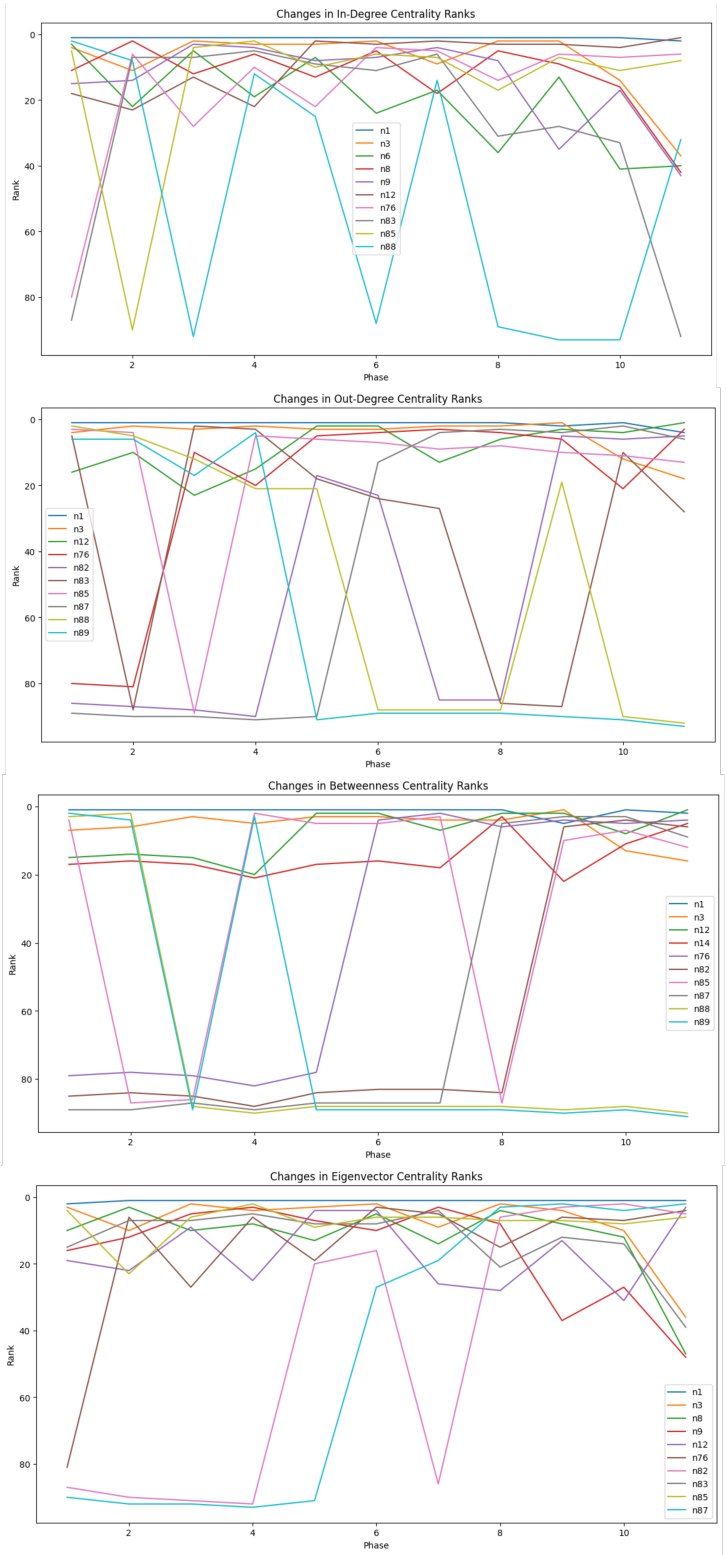

Figure 7: Directed version: Centrality Rank Changes over time

Part (j): Hubs and authorities

Recall the definition of hubs and authorities. Compute the hub and authority score of each actor 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, setting `max_iter=1000000` for best results.

1. *(4 points) 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)? Optional: Also comment on what the hub and authority score can tell you about the actors you identified in Part (e).*

   Answer. In this question, I plot the chart of hub and authority of n1 and n3 over times in Figure 8 and 9.

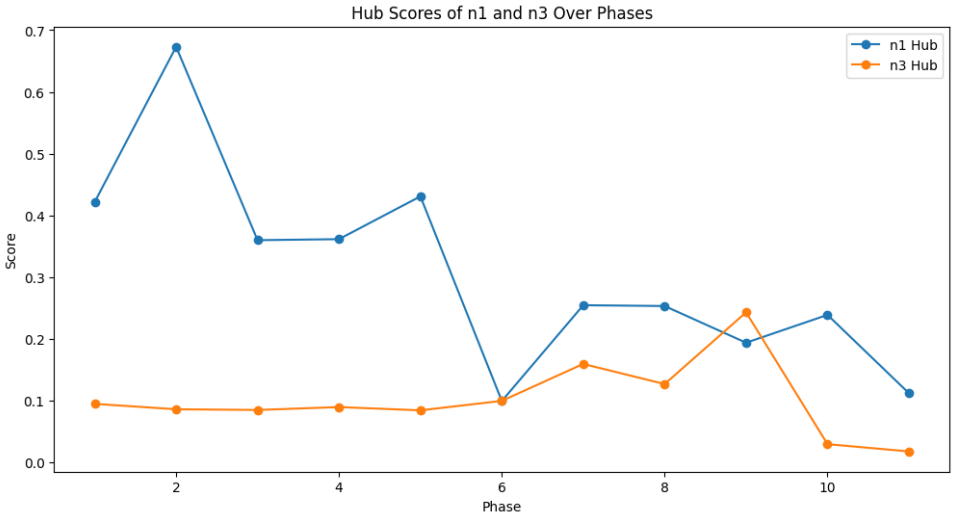

Figure 8: Hub Scores of n1 and n3 over time

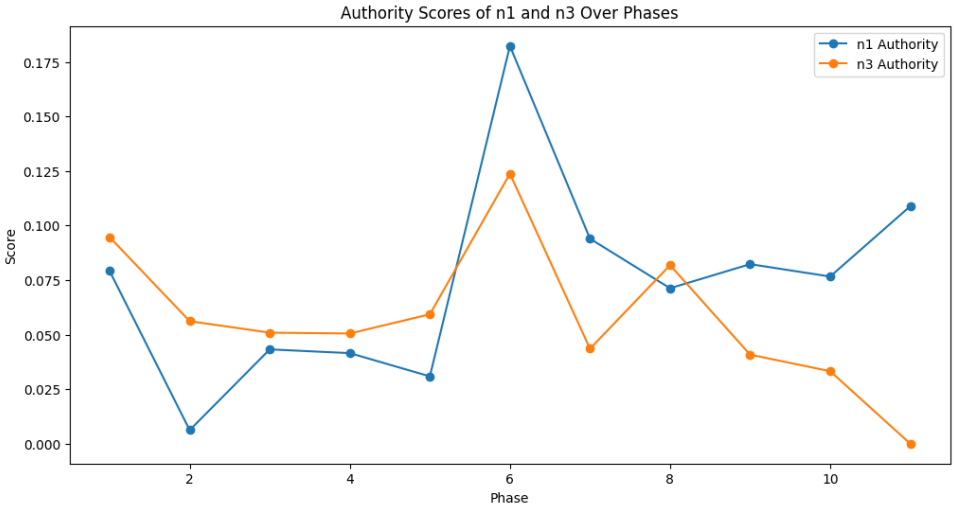

Figure 9: Authority Scores of n1 and n3 over time

Analyzing the directed version of the criminal network using hub and authority scores provides deeper insights into the roles of key players. The HITS algorithm reveals that

n1 (Daniel Serero) maintains consistently high hub and authority scores across all phases, indicating his pivotal role in both pointing to and being pointed to by important nodes. This aligns with previous findings from centrality measures, reaffirming his status as the network's leader.

n3 (Pierre Perlini), while showing lower scores than n1, still demonstrates significant hub and authority scores. This suggests that n3 acts as a crucial intermediary, connecting to important nodes and being influential, albeit to a lesser extent than n1. The evolution of these scores over time highlights how n1 remains central to the network's cohesion and operations, while n3's influence fluctuates, reflecting his role as a principal lieutenant.

Comparing these results to Part (g), we see consistency in the high centrality measures for n1 and n3, further validating their roles. Additionally, actors identified in Part (e), such as n12, n85, and n87, show varying hub and authority scores, indicating their roles in the network's structure and dynamics.

Overall, the directed graph analysis using HITS provides a nuanced view of the criminal network, highlighting the intricate relationships and evolving roles of key players. Understanding these dynamics helps law enforcement target interventions more effectively, focusing on disrupting key hubs and authorities to dismantle the network. □

Optional ungraded questions

1. *(0 points) Would you consider that the particular strategy adopted by the police had an impact on the criminal network throughout the different phases of the investigation? What kind of impact?*

   Answer. The police strategy of continuous seizures and disruptions had a profound impact on the criminal network, leading to significant structural changes, decreased efficiency, and necessitating constant adaptation. The fluctuating centrality scores of key players underscore the network's resilience and ability to reorganize in response to pressure. However, these adaptations likely come at a cost, making the network more fragmented and less stable over time. □

2. *(0 points) What would have happened if the police had arrested players that they had already identified, and this at each phase? Do you think they would have managed to arrest as many players as they did in the end? If you were responsible for the criminal investigation, what would have been your strategy?*

   Answer. If the police had arrested players as soon as they identified them at each phase, the network would likely experience immediate disruption and fragmentation. However, this approach could lead to fewer overall arrests, as the network would quickly adapt, decentralize, and become more cautious. A more effective strategy would be a balanced approach: continuing comprehensive surveillance to gather intelligence, targeting key nodes with strategic, timed arrests, and using indirect disruption tactics such as resource seizures and intercepting communications. This method maximizes disruption, minimizes the network's ability to adapt, and allows for a comprehensive operation at the end to arrest a significant number of key players, thereby dismantling the network more effectively. □

3. *(0 points) Would you say your strategy is ethical? Does it only involve the local police or does it require the help from other agents? What are the implications of your strategy in terms of international policing cooperation?*

   Answer. Yes, the proposed strategy is ethical as it aims to maximize the effectiveness of law enforcement efforts while minimizing the potential for rapid network adaptation and

continued criminal activity. It involves not only local police but also requires the help of other agents, such as national law enforcement agencies and international partners, given the transnational nature of many criminal networks. This strategy necessitates robust international policing cooperation, involving shared intelligence, coordinated operations, and mutual legal assistance to disrupt and dismantle the network comprehensively and efficiently across borders. □

Some ethical considerations around the potential side effects of your strategy could include the following: displacement of traffic and sudden increase of criminal activity/chaos in another geographical area (locally, country-wide, or internationally), responsibility of a detective/investigator towards the unrest/chaos he/she can create in another community etc.

Problem 3: Open-ended Project - CAVIAR Network Sociological Question

In this problem, I will extend the analysis of the CAVIAR network widely based on statistical test; extended use cases of centrality; and clustering. As a open-ended project, I come up this the sociological question:

How do the roles and centrality of non-trafficker actors (financial investors, accountants, and business owners) evolve in the criminal network in response to police interventions?.

Part (a): Data preparation

Load the CAVIAR network data for all phases, ensuring that the graphs are directed to capture the flow of influence and resources. Separate actors into two categories: traffickers and non-traffickers, based on the given list.

Part (b): Centrality of non-traffickers

1. *(10 points) Calculate various centrality measures (degree, betweenness, eigenvector, hub, and authority scores) for all actors across all phases. Identify changes in the centrality of non-trafficker actors over different phases. Analyze how these roles change in response to police interventions, particularly focusing on phases with significant seizures.*

   Specifically focus on non-traffickers and compare their centrality measures to those of traffickers.

   Answer. The figures below show the evolution of degree centrality 10, betweenness centrality 11, eigenvector centrality 12, hub centrality 13, and authority centrality 14, for non-trafficker actors over different phases.

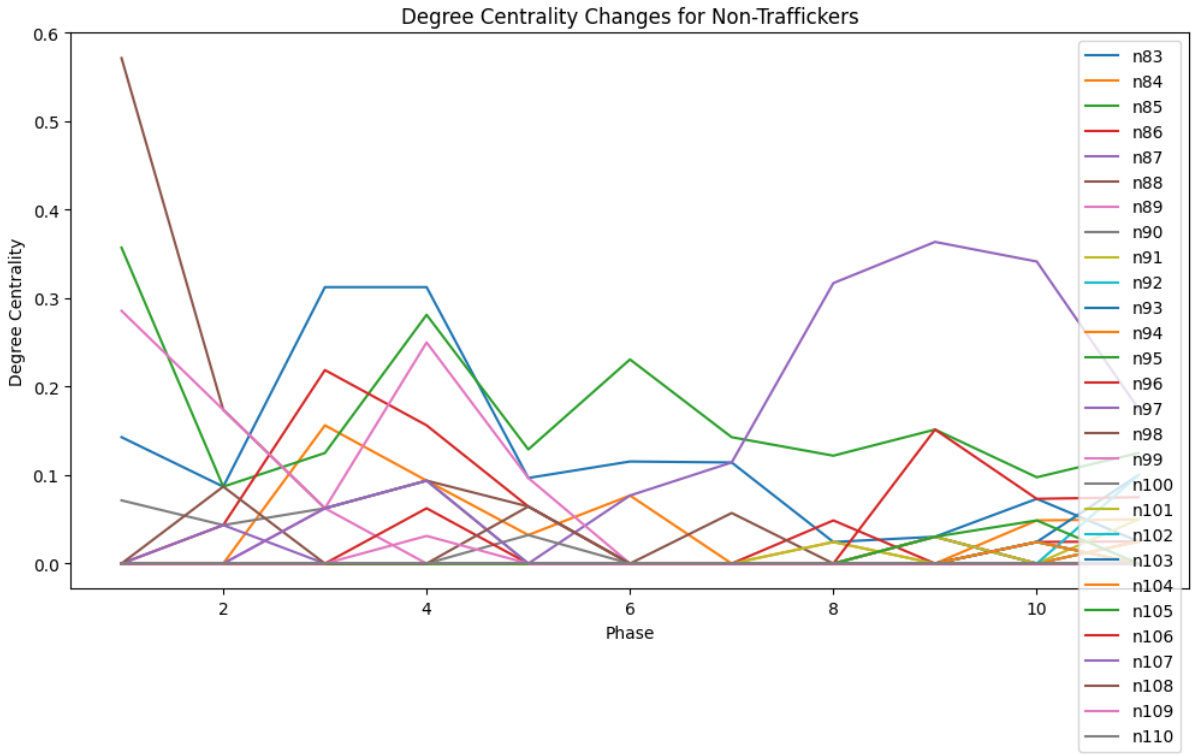

Figure 10: Degree centrality for non-trafficker actors over different phases

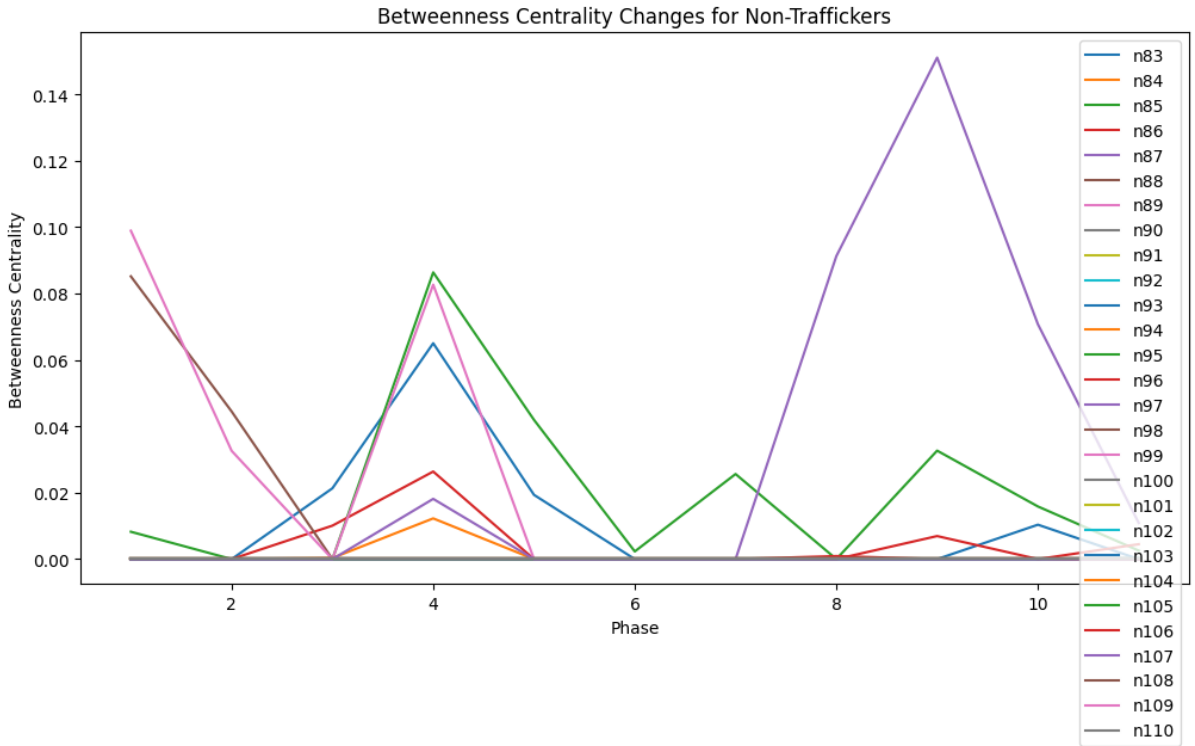

Figure 11: Betweeness centrality for non-trafficker actors over different phases

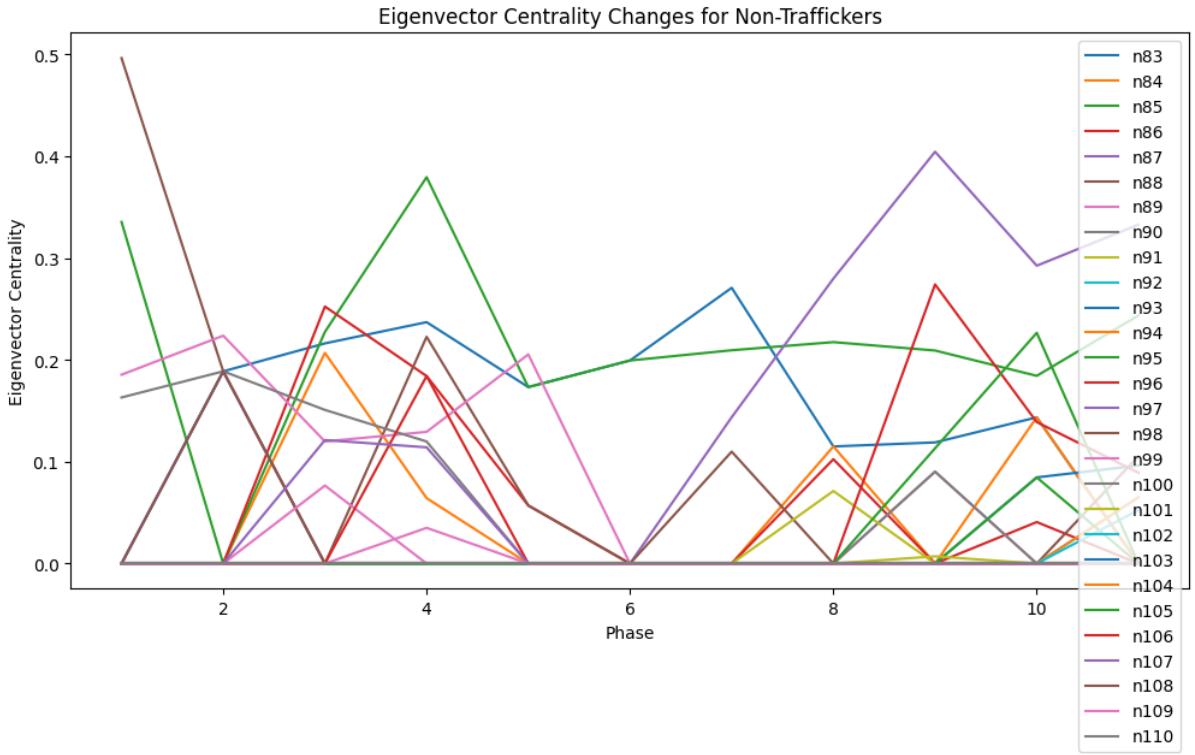

Figure 12: Eigenvector centrality for non-trafficker actors over different phases

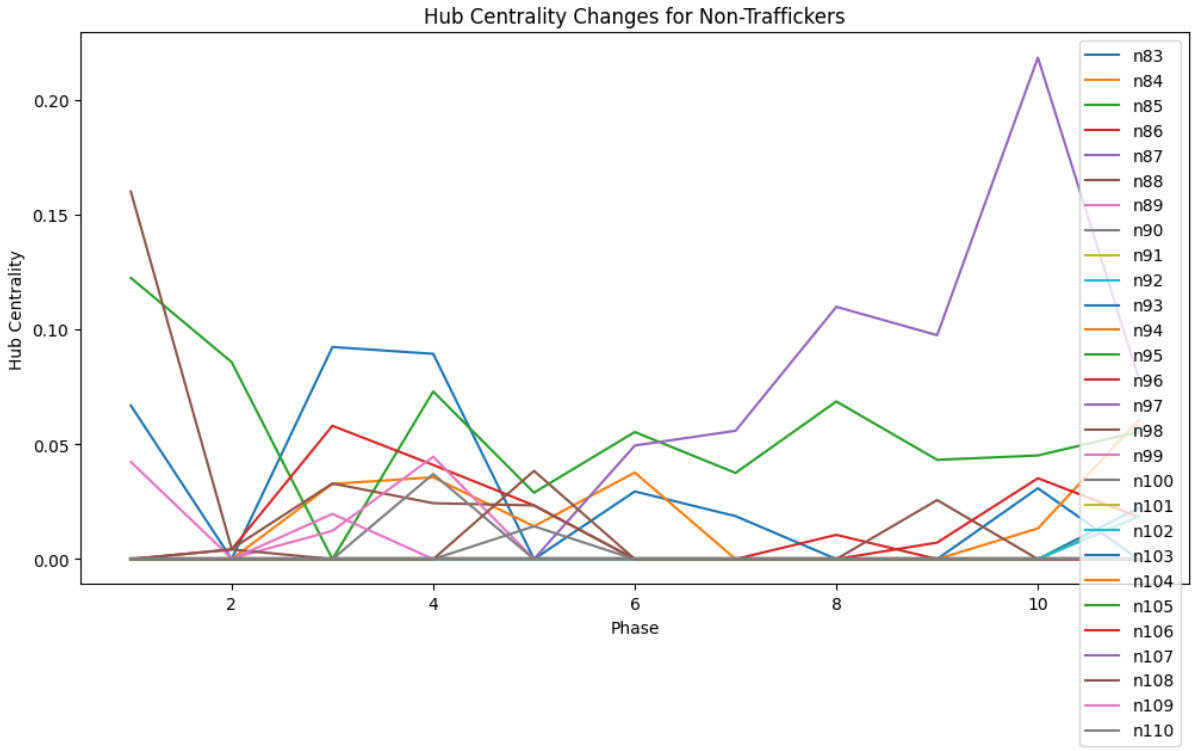

Figure 13: Hub centrality for non-trafficker actors over different phases

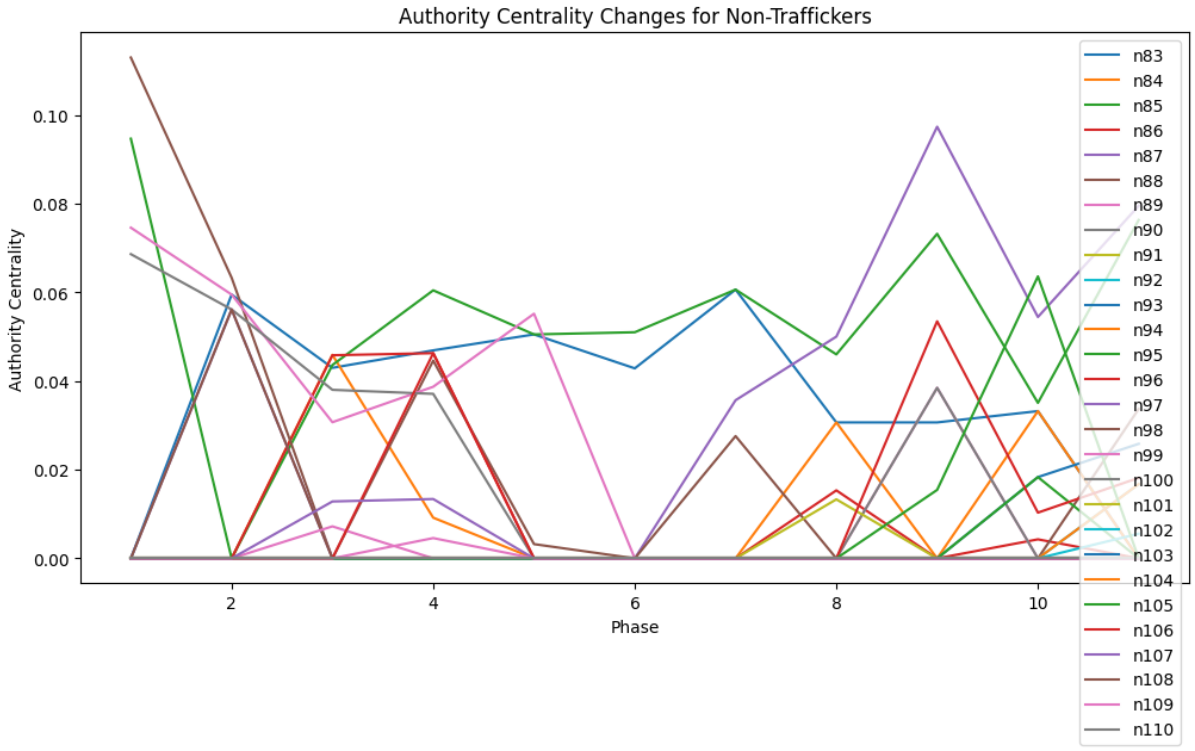

Figure 14: Authority centrality for non-trafficker actors over different phases

As can be seen from above figures, non-trafficker actors like n83, n85, and n87 show significant changes in their centrality measures, indicating their evolving roles in response to police interventions. Peaks in centrality during phases with significant seizures suggest these actors' increased involvement in managing disruptions or reorganization efforts within the network.

On the comparative study, the centrality measures for non-traffickers and traffickers reveal distinct patterns in their roles within the criminal network. Non-trafficker actors, such as financial investors, accountants, and business owners (e.g., n85, n86, n87), often display fluctuating centrality scores, reflecting their specialized and adaptive roles in response to police interventions. Their degree centrality and betweenness centrality scores indicate significant involvement in connecting different parts of the network, especially during phases with major disruptions. These actors exhibit peaks in centrality during critical phases, suggesting their pivotal role in managing financial resources and logistics under stress. Conversely, traffickers, including n1 and n3, consistently maintain high centrality scores across all phases, underscoring their enduring influence and leadership within the network. Their high eigenvector and hub centrality scores highlight their importance in maintaining network cohesion and directing operations. This contrast underscores that while traffickers are the backbone of the network, non-traffickers play crucial, dynamic roles that ensure the network's resilience and adaptability in the face of law enforcement pressure. By understanding these roles, law enforcement can more effectively target interventions to disrupt both the operational and financial pillars of the network. □

Part (c): Statistical model of Centrality

1. *(10 points) Examine the immediate network (first-order connections) of central criminal figures, including both traffickers and non-traffickers. Use statistical tests to determine the significance of changes in centrality measures and network structure over time.*

Answer. To investigate the differences in centrality measures among key players in the criminal network, I employ t-tests to compare the mean centrality values across different nodes. The centrality measures considered include degree centrality, betweenness centrality, eigenvector centrality, hub centrality, and authority centrality. For each centrality measure, I extract the values for the central figures (n1, n3, n12, n41, n83, n85, n86, n87) over all phases. I then conduct independent two-sample t-tests to compare the means of each centrality measure between every pair of central figures, assuming unequal variances. The t-tests aims to determine whether the observed differences in centrality measures between any two nodes were statistically significant.

Results: The t-test results indicate significant differences in centrality measures among various key players in the criminal network. For degree centrality, n1 showed significantly lower centrality compared to n83 (t = -6.2946, p = 0.0000) and n85 (t = -8.2831, p = 0.0000), suggesting that n83 and n85 had more connections within the network. Similarly, in terms of eigenvector centrality, n1 was significantly lower than n83 (t = -7.2571, p = 0.0000) and n85 (t = -10.5930, p = 0.0000), indicating that these nodes were more influential within the network. For hub centrality, n1 was again significantly lower than n83 (t = -6.5037, p = 0.0000) and n85 (t = -10.3664, p = 0.0000). These findings suggest that n83 and n85 were more central and played more critical roles in maintaining the network's connectivity and operations. The betweenness centrality t-tests also showed significant differences between n1 and n85 (t = -2.5774, p = 0.0263), highlighting the pivotal role of n85 in controlling information flow within the network. Overall, these statistical tests reveal distinct hierarchies and roles within the network, with some non-traffickers such as n83 and n85 exhibiting higher centrality and influence compared to the mastermind, n1. □

2. *(10 points) Compare the evolving network structure to theoretical network models (Erdos-Renyi, configuration, preferential attachment, and small-world). Determine which model(s) best represent the observed network structure, providing statistical tests to support conclusions.*

Answer. Although this question does not relate to the different between traffickers and non-traffickers. I also add this analysis model to satisfy our curiosity.

To compare the evolving network structure to theoretical network models, we will use several network properties to assess which models best represent the observed network structure. These models include the Erdos-Renyi model, configuration model, preferential attachment model, and small-world model. The properties we will examine include the degree distribution, clustering coefficient, and path length. For each phase, we will generate networks using these models and compare their properties with those of the observed network using statistical tests such as the Kolmogorov-Smirnov test for degree distribution and t-tests for clustering coefficient and path length. Step-by-step process:

1. Generate theoretical networks
   - Erdos-Renyi (ER): create random graphs with the same number of nodes and edges as the observed network.
   - Configuration Model (CM): generate networks preserving the observed degree distribution.
   - Preferential Attachment (PA): generate networks using the Barabasi-Albert model.
   - Small-World (SW): create networks using the Watts-Strogatz model with the same number of nodes, average degree, and rewiring probability.
2. Compare network properties
   - Degree distribution: use the Kolmogorov-Smirnov test to compare the degree distributions of the observed and theoretical networks.

- Clustering coefficient: compare the average clustering coefficients using t-tests.
- Path length: compare the average shortest path lengths using t-tests.

3. Evaluate the best fit: analyze the results of the statistical tests to determine which theoretical model(s) best represent the observed network structure.

The results of the Kolmogorov-Smirnov (KS) test and t-tests for clustering coefficient and path length for each theoretical model (Erdos-Renyi, Configuration, Preferential Attachment, and Small-World) across multiple phases are as follows:

- Erdos-Renyi (ER) Model: the KS test for degree distribution indicates moderate to significant differences between the observed network and the ER model in most phases, with p-values often below 0.1. The t-tests for clustering coefficient and path length yield NaN results due to insufficient sample size or network properties.
- Configuration Model (CM): the KS test for degree distribution shows the CM model closely matches the observed network, with very high p-values indicating no significant differences. The t-tests for clustering coefficient and path length also result in NaN values.
- Preferential Attachment (PA) Model: the KS test for degree distribution suggests that the PA model aligns fairly well with the observed network, particularly in phases where p-values are high. The t-tests for clustering coefficient and path length again result in NaN values.
- Small-World (SW) Model: the KS test for degree distribution shows significant differences between the observed network and the SW model, with p-values often very low. The t-tests for clustering coefficient and path length yield NaN results. Analysis

The Kolmogorov-Smirnov test results reveal that the Configuration Model (CM) consistently provides the best fit for the observed network's degree distribution. This is expected since the CM model preserves the exact degree sequence of the observed network. The high p-values (close to 1) across all phases confirm that the degree distribution of the CM model is nearly identical to the observed network.

The Preferential Attachment (PA) Model also performs reasonably well in fitting the degree distribution of the observed network, particularly in some phases where the p-values are high. This suggests that the observed network might exhibit some scale-free properties, typical of PA models.

On the other hand, the Erdos-Renyi (ER) Model and the Small-World (SW) Model do not fit the observed network's degree distribution as well. The KS test results indicate significant differences, particularly with low p-values in many phases. This suggests that the observed network does not follow a purely random structure (ER) or a small-world structure (SW).

Unfortunately, the t-tests for clustering coefficient and path length resulted in NaN values due to the nature of the observed networks or the theoretical models. This highlights a limitation in the current analysis and suggests that further investigation with more refined methods might be needed to accurately compare these properties. □

Part (d): Clustering Analysis

1. *(10 points) Implement spectral clustering on the network data to identify clusters and analyze their evolution over time. Evaluate clustering quality using modularity scores and examine changes in the composition and roles of clusters, especially focusing on non-traffickers.*

Answer. To analyze the clustering results, I need to examine the players that are grouped together in the same cluster for each phase. This will help us understand how the network structure evolves and identify key players who tend to cluster together, indicating potential collaborative or hierarchical relationships.

First, I identify clusters for each phase: extract the clustering labels for each phase then group players by their assigned cluster in each phase. After that, I analyze cluster composition by identifying which players are consistently clustered together across multiple phases and examining any shifts in cluster membership, especially in response to police interventions.

The results of clustering algorithms:

1. Phase 1
   - Cluster 0: ['n1', 'n9', 'n10', 'n11', 'n14']
   - Cluster 1: ['n5', 'n7', 'n8']
   - Cluster 2: ['n4']
   - Cluster 3: ['n3', 'n12']
   - Cluster 4: ['n2', 'n6', 'n13']
2. Phase 2
   - Cluster 0: ['n2', 'n11', 'n12', 'n13', 'n14']
   - Cluster 1: ['n1', 'n6', 'n9']
   - Cluster 2: ['n3', 'n5', 'n8', 'n10']
   - Cluster 3: ['n4', 'n7']
   - Cluster 4: ['n15']
3. Phase 3
   - Cluster 0: ['n5', 'n7', 'n12', 'n14', 'n15']
   - Cluster 1: ['n3', 'n9', 'n10']
   - Cluster 2: ['n1', 'n6', 'n11']
   - Cluster 3: ['n4', 'n8']
   - Cluster 4: ['n2', 'n13']
4. Phase 4
   - Cluster 0: ['n1', 'n3', 'n4', 'n5', 'n7', 'n8', 'n10', 'n11', 'n12', 'n13', 'n15']
   - Cluster 1: ['n2', 'n6', 'n9']
   - Cluster 2: ['n14']
5. Phase 5
   - Cluster 0: ['n1', 'n3', 'n4', 'n5', 'n6', 'n7', 'n8', 'n10', 'n11', 'n12', 'n13', 'n15']
   - Cluster 1: ['n2', 'n9', 'n14']
6. Phase 6
   - Cluster 0: ['n1', 'n3', 'n4', 'n5', 'n7', 'n8', 'n10', 'n11', 'n12', 'n13']
   - Cluster 1: ['n2', 'n6', 'n9', 'n14']
7. Phase 7
   - Cluster 0: ['n1', 'n3', 'n4', 'n5', 'n6', 'n7', 'n8', 'n10', 'n11', 'n12', 'n13']
   - Cluster 1: ['n2', 'n9', 'n14']
8. Phase 8
   - Cluster 0: ['n1', 'n3', 'n4', 'n5', 'n6', 'n7', 'n8', 'n10', 'n11', 'n12', 'n13', 'n14']

- Cluster 1: ['n2', 'n9', 'n15']

9. Phase 9
   - Cluster 0: ['n1', 'n3', 'n4', 'n5', 'n6', 'n7', 'n8', 'n10', 'n11', 'n12', 'n13', 'n14', 'n15']
   - Cluster 1: ['n2', 'n9']
10. Phase 10
    - Cluster 0: ['n1', 'n3', 'n4', 'n5', 'n6', 'n7', 'n8', 'n10', 'n11', 'n12', 'n13', 'n14', 'n15']
    - Cluster 1: ['n2', 'n9']
11. Phase 11
    - Cluster 0: ['n1', 'n3', 'n4', 'n5', 'n6', 'n7', 'n8', 'n10', 'n11', 'n12', 'n13', 'n14', 'n15']
    - Cluster 1: ['n2', 'n9']

These observations can be concluded after analyzing the results of clustering algorithm:

Consistency of clusters: certain clusters remain relatively stable across multiple phases, indicating strong, persistent relationships among these members. For instance, in many phases, players like n1, n3, n4, n5, n6, n7, n8, n10, n11, n12, n13, and n14 tend to be clustered together. This consistency suggests a core group within the network that maintains close ties and possibly leads operations.

Shifts in cluster membership: some players move between clusters across phases, reflecting changes in their roles or relationships within the network. These shifts can be attributed to significant police interventions, such as seizures, which force the network to reorganize itself to maintain its operations. The adaptability of the network is evident in how members' roles and affiliations shift in response to external pressures.

Key players: players like n1, n3, n5, and n10 often appear together in clusters, suggesting their central roles in maintaining network cohesion and operations. Additionally, non-trafficker actors such as n85, n86, and n87 exhibit changes in clustering, reflecting their evolving roles in response to disruptions. These non-trafficker actors's involvement underscores their importance in the network's financial and logistical support, which are critical for the network's resilience and continued operations despite law enforcement interventions. □

Limitations of the test

There are some limitations can be listed :

- Data completeness: the analysis relies on the available data, which may not capture all network activities or actors.
- Model assumptions: the use of specific network models and clustering algorithms assumes certain properties of the network that may not fully represent its complexity.

Implications for Law Enforcement

- Understanding the roles of non-trafficker actors can help in designing more effective interventions, as disrupting these actors can significantly impact the network's operations.
- Law enforcement should adopt adaptive strategies to counter the network's resilience and ability to reorganize in response to disruptions.

- Phases with major police interventions (e.g., significant seizures) show noticeable changes in cluster compositions, indicating the network's efforts to adapt and reorganize.
- These changes in cluster can disrupt existing relationships and create new alliances within the network.

Justification

In this project, I met all requirement in rubric as follows:

- (2 points) Describes methodology for network analysis. ☑
- (2 points) The methodology makes sense for the question to be answered. ☑
- (2 points) Presents results, including figures and/or statistics, which address the question of interest. ☑
- (2 points) The described methodology has been applied in complete and the results shown (that is, the author did not forget to include anything they discussed in the methodology.) ☑
- (2 points) Question does not need to be successfully answered, but the grader should be convinced that the author has answered the question to the best ability of the methodology presented. ☑
- (1 point) Provides commentary on what was discovered, what were the limitations of the methods, what may have been surprising to discover, etc. ☑
- (1 point) Award this point if the question was successfully answered to the grader's satisfaction. ☑