Written Report – 6.419x Module 3

Name: (vatsalya243001)

- **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?

Solution:

The naive time complexity for matrix multiplication is O(n^3), which matches the time complexity of the proposed algorithm that involves adding one to each symmetric element of C_{ab}Cab for every pair (r,a),(r,b) in each row r of A. While there are potential optimizations for both algorithms that can achieve O(n^{2.8}), these optimizations may increase space complexity and are beyond the scope of this question.

Part (d) (3 points) Bibliographic coupling and co citation can both be taken as an indicator that papers deal with related material. However, they can in practice give noticeably different results. Why? Which measure is more appropriate as an indicator for similarity between papers?

Solution:

Bibliographic coupling indicates that papers cite the same references and can be computed by A^T A. Co-citation indicates references cited by the same paper and can be computed by AA^T. From the literature review I conducted, there is no definitive choice that will reliably identify more similar papers. Results vary based on the time span chosen, the research field, and the size of the network. High bibliographic coupling indicates papers that share many references, suggesting they deal with similar topics within a field and build upon the same historical record. High co-citation strength indicates papers frequently referenced together. The works citing them may span various subtopics and research frontiers that all rely on the co-cited seminal works. Notably, co-citation networks evolve over time, while bibliographic coupling remains static. Figure 1 illustrates the two similarity measures.

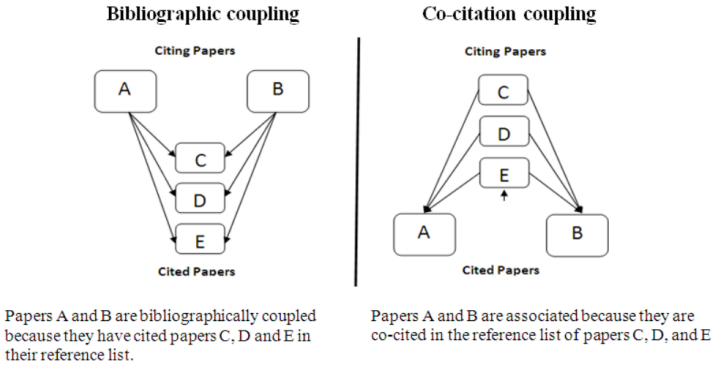

Figure 1: Bibliographic coupling and co-citation (Surwase et al. 2011, Co-citation Analysis: An Overview.)

- **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?

Solution:

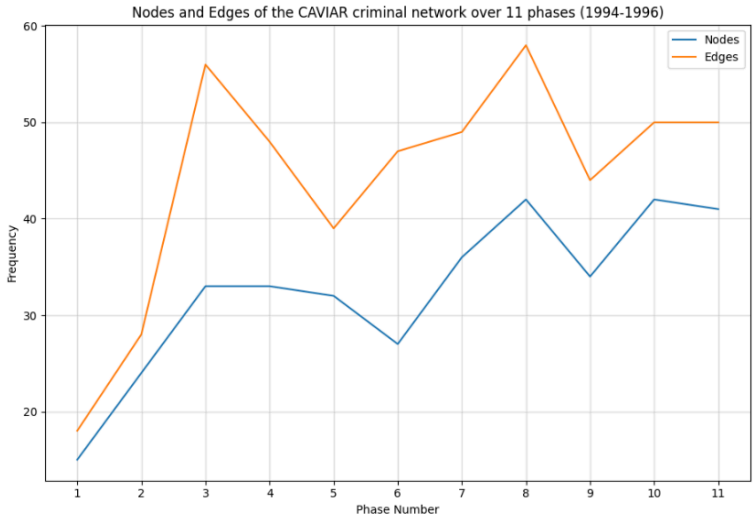

Figure 2: Number of nodes and edges in the CAVIAR network over time (each Phase is 2 months)

The plot of nodes and edges of the CAVIAR network over Phase number is shown in Figure 2. The number of (nodes,edges) rises sharply between Phases 1 and 3 from (15, 18) to (33, 56). This rise is likely due to the police force adding new players to the network as they listened to conversations and learned of new contacts. The number of nodes plateaus from Phase 3 to 5 and there is a sharp drop in the number of edges. They likely discovered the majority of the nodes at this point, and might have calibrated which nodes were actually connected. Since not all players were known from the start of Phase 1, our assumption of imputing zero for actors in Phases they aren't present may not be justified (especially for Phases 1 and 2). If we don't impute any zeros and compute mean centrality measures then n1, n21, and n3 are still top betweenness, but n1, n3 and n87 are top eigenvectors. The top 5-10 players still seem relatively stable regardless of the initial lack of complete network knowledge.

Part (d) (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.

Solution:

Degree Centrality measures the number of connections (edges) a node has in the network. In criminal networks, individuals with high degree centrality are in contact with many others. However, this measure is limited because the most connected individuals are not necessarily the most important. Key figures often avoid direct communication with many network members to remain inconspicuous. Instead, second and third-in-command typically relay instructions to those involved in various activities, such as financial operations, transportation logistics, drug provision, and purchases.

Eigenvector Centrality assesses nodes based on the importance of their neighbors. In criminal networks, individuals with high eigenvector centrality are connected to other influential figures. In the CAVIAR network, this identifies key players like Daniel Serero, the mastermind, and Pierre Perlini, the principal lieutenant, along with their direct contacts.

Betweenness Centrality identifies nodes that frequently lie on the shortest paths between other nodes. In criminal networks, individuals with high betweenness centrality are crucial for information flow, as they bridge different parts of the network. Removing a node with high betweenness would significantly disrupt communication. This measure is most relevant for identifying those running illegal activities, as these nodes maintain essential network connectivity. In the CAVIAR network, this highlights Ernesto Morales, the principal cocaine organizer, as a critical figure, even more so than some direct contacts of Serero and Perlini.

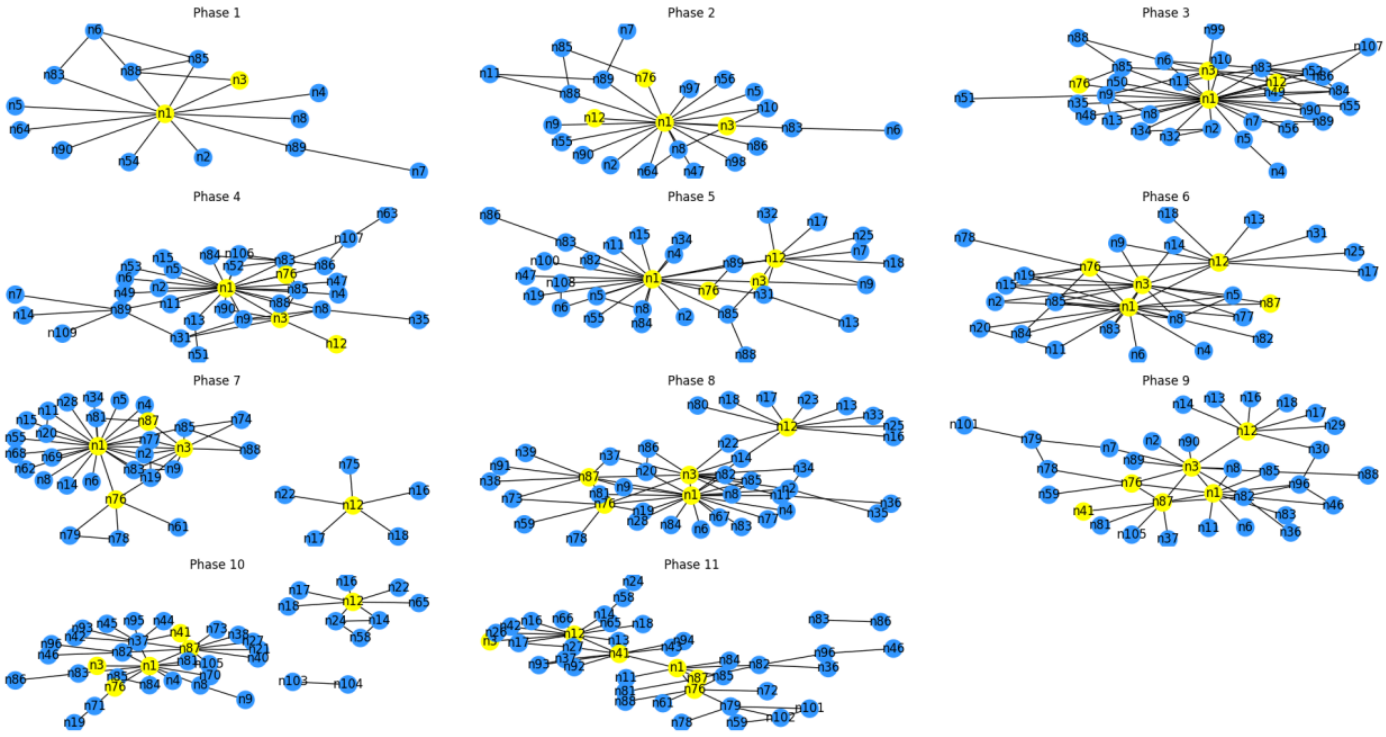

Figure 3: CAVIAR network over time (central players colored yellow)

Table 1: Top 10 Betweenness Centrality Measures of all players in CAVIAR network (missing values imputed as 0)

Node	Name	Betweenness	Eigenvector	Degree	Role
n1	Daniel Serero	0.655051	0.546391	0.601485	Mastermind of the network
n12	Ernesto Morales	0.167562	0.141893	0.170893	Principal organizer of the cocaine import
n3	Pierre Perlini	0.129403	0.298095	0.223505	Principal lieutenant of Serero
n76	Gabrielle Casale	0.083791	0.165877	0.112235	Charged with recuperating the marijuana
n87	Patrick Lee	0.061327	0.141080	0.090261	Investor
n41	NaN	0.050369	0.063869	0.027644	NaN
n89	Antonio Iannacci	0.047948	0.078354	0.059124	Investor
n14	NaN	0.032671	0.051697	0.033035	NaN
n83	Alain Levy	0.031785	0.153522	0.095836	Investor and transporter of money
n82	Salvatore Panetta	0.029196	0.100067	0.047570	Transport arrangements manager

Table 2: Top 10 Betweenness Centrality Measures of all players in CAVIAR network (missing values not imputed)

Node	Name	Betweenness	Eigenvector	Degree	Role
n1	Daniel Serero	0.655051	0.546391	0.601485	Mastermind of the network
n41	NaN	0.184687	0.234188	0.101361	NaN
n12	Ernesto Morales	0.184318	0.156083	0.187982	Principal organizer of the cocaine import
n3	Pierre Perlini	0.129403	0.298095	0.223505	Principal lieutenant of Serero
n87	Patrick Lee	0.112433	0.258647	0.165478	Investor
n76	Gabrielle Casale	0.092170	0.182465	0.123459	Charged with recuperating the marijuana
n89	Antonio Iannacci	0.087905	0.143649	0.108395	Investor
n79	NaN	0.080449	0.038775	0.091017	NaN
n82	Salvatore Panetta	0.053527	0.183455	0.087212	Transport arrangements manager
n14	NaN	0.051340	0.081239	0.051913	NaN

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.

Solution:

Tables 1 and 2 show centrality measures for the top 10 betweenness centralities ordered by descending betweenness for all players in the network, both with and without imputed zeros. Considering all centrality measures together provides a more complete picture of these players. The evolution of the network across the phases is displayed in Figure 3, and the top 5 players from Tables 1 and 2 are highlighted in yellow. Notice that these players remain central even as the network reshapes. The primary traffickers are Daniel Serero, Ernesto Morales, Pierre Perlini, Gabrielle Casale, Patrick Lee, and n41. The cut-off for the top 5 is arbitrary and was adjusted based on the nodes' positions in Figure 3. Nodes that consistently appear central are likely key players. Other players with high centrality, such as Antonio Iannacci (n89), Alain Levy (n83), and Salvatore Panetta (n82), are connected to important players but still appear peripheral in the network.

Part (f) Question 2 (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.

Solution:

This restructuring happened between Phase 4 and 5 (X = 4) and corresponds with the first seizure by the police of 300kg of marijuana ($2,500,000). This represents a shift in the operation toward cocaine as indicated in the expanding network around n12 (Ernesto Morales), the principal organizer of the cocaine import, whose degree centrality increased from 0.03125 to 0.258065 from Phase 4 to 5.

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?

Solution:

Referring back to Figures 2 and 3, we observe the initial growth of the network from Phases 1-3, centered primarily around Serero (n1) and Perlini (n3), as the police built a complete picture. The number of edges decreases between Phases 3-5 (from 56 to 39) as the network restructures to expand the cocaine imports centered around Morales (n12). From Phases 6-8, the number of nodes and edges increases from (27, 47) to (42, 58). During Phase 6, there were three smaller seizures of both marijuana and cocaine, followed by a large seizure of marijuana in Phase 7.

Lieutenant Perlini becomes more central in the network during Phase 6, possibly indicating Serero's concern about maintaining too many direct contacts. The network's structure begins to decentralize Serero's authority, handing more power to n3, n76, n82, n85, and n86 from Phases 6-8. Morales (n12) forms his own connected component during Phases 7 and 10, possibly indicating a communication freeze between the marijuana and cocaine operations. Notably, both Phases 7 and 10 correspond to significant marijuana seizures. From Phases 9-11, Serero and his lieutenant Perlini reduce their direct connections, leading to a less centralized network. During these later phases, Patrick Lee (n87), Gabrielle Casale (n76), and n41 take on more central roles. The huge seizure in Phase 10 of 2200kg of marijuana ($18,700,000) likely prompted the network to break ties and further decentralize, as reflected in the Phase 11 graph.

*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.*

Solution:

Player n41 becomes active in Phase 9 through the investor Patrick Lee (n87) and quickly becomes connected to n37 and then becomes a central node for six players by Phase 11. n37 becomes active in

Phase 8 connected to Patrick Lee and Pierre Perlini and quickly becomes a central node for six players in Phase 10, many of whom are then handed off to n41.

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

Solution:

A directed graph contains more information since the adjacency matrix is not necessarily symmetric. In-degree centrality would indicate the amount that other players contacted a particular player. This would show which players are often on the receiving end of information sharing. Out-degree centrality would indicate the amount that a particular player contacted other players. This would show which players are often on the giving end of information sharing. Intuitively it would seem that higher out-degree would indicate players with more network influence and higher in-degree would indicate players with more network knowledge. Left-eigenvector centrality would indicate how important the other players communicating with a player are. This would show which players are receiving the most important information. Right-eigenvector centrality would indicate how important the other players a player is communicating to are. This would show which players are giving the most important information. See Figure 4 to see the directed graphs for each Phase.

Part (j) (4 point) Recall the definition of hubs and authorities. Compute the hub and authority score of each actor, and for each phase. (Remember to load the adjacency data again this time using create_using = nx.DiGraph().) With networkx you can use the nx.algorithms.link_analysis.hits function, set max_iter=1000000 for best results. 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)?

Solution:

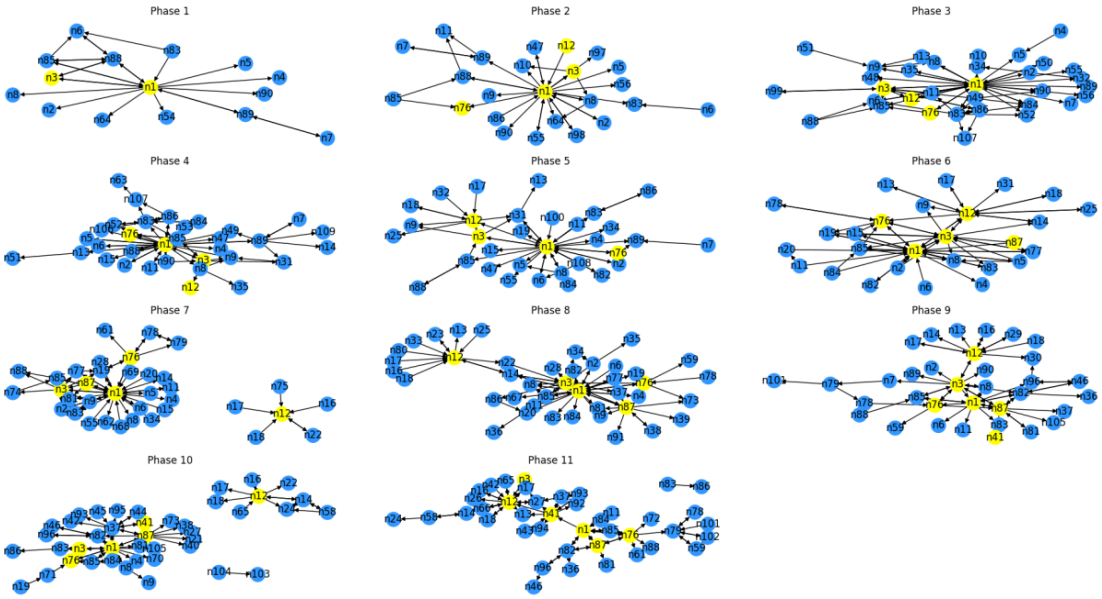

Figure 4: CAVIAR network over time (central players colored yellow)

The hub score is defined as the sum of the authority scores of the nodes it points to, and the authority score is the sum of the hub scores of the nodes that point to it. In a criminal network, a player with a high authority score is contacted by players recognized as hubs of information, while a player with a high hub score contacts players recognized as authorities. Hubs spread important information, and authorities collect important information. It is therefore not surprising that Serero, Morales, and Perlini have high scores. The hub and authority scores of the central players for each phase are shown in Tables 3 and 4. Comparing n1 (Serero) and n3 (Perlini), we see that Serero's hub score is high for Phases 1-5 and 8-9, whereas Perlini's hub score is higher for Phases 6-7. The reverse pattern holds for authority scores: Perlini's authority score is higher for Phases 1-5 and 8-9, while Serero's is higher for Phases 6-7.

This adds an additional layer to the earlier analysis. Although Perlini had more edges in Phase 6 and seemed to take a more central role, it wasn't clear that his role switched from collecting information to disseminating it, and vice versa for Serero. This evidence supports the claim that Serero was preparing to withdraw from the network. He first distributed responsibilities to other players as new hubs while maintaining his authority in Phases 6-7, and then acted as a hub with final instructions in Phases 8-9 before withdrawing. The drop in their hub and authority scores in Phases 10-11 further indicates that Serero and Perlini removed themselves from the network.

Table 3: Hub scores of the central players in the CAVIAR network for each Phase

Phase	Daniel Serero	Pierre Perlini	Ernesto Morales	Gabrielle Casale	Alain Levy	Antonio Iannacci	Salvatore Panetta	Patrick Lee	unknown
	n1	n3	n12	n76	n83	n89	n82	n87	n41
1	0.70306	0.01436	NaN	NaN	0.01808	0.00196	NaN	NaN	NaN
2	0.97296	0.00764	0.00004	0.00000	0.00000	0.00000	NaN	NaN	NaN
3	0.79310	0.04625	0.00000	0.00474	0.06045	0.00020	NaN	NaN	NaN
4	0.85979	0.02397	0.00699	0.02674	0.00737	0.00017	NaN	NaN	NaN
5	0.90650	0.01054	0.02403	0.00170	0.00000	0.00000	0.00004	NaN	NaN
6	0.00805	0.19529	0.01188	0.12969	0.00307	NaN	0.00307	0.00935	NaN
7	0.00681	0.34332	NaN	0.02084	0.00065	NaN	NaN	0.07697	NaN
8	0.82588	0.01738	0.00209	0.01062	0.00000	NaN	0.00000	0.01107	NaN
9	0.58793	0.13947	0.00216	0.00614	0.00000	0.00000	0.01327	0.15385	0.00000
10	0.23082	0.00199	NaN	0.00698	0.00100	NaN	0.00307	0.10496	0.00694
11	0.00008	0.03789	0.00129	0.00000	NaN	NaN	0.00000	0.00000	0.03472

Table 4: Authority scores of the central players in the CAVIAR network for each Phase

Phase	Daniel Serero	Pierre Perlini	Ernesto Morales	Gabrielle Casale	Alain Levy	Antonio Iannacci	Salvatore Panetta	Patrick Lee	unknown
	n1	n3	n12	n76	n83	n89	n82	n87	n41
1	0.01181	0.13571	NaN	NaN	0.00000	0.13164	NaN	NaN	NaN
2	0.00027	0.33670	0.00000	0.04349	0.08696	0.06522	NaN	NaN	NaN
3	0.00315	0.14957	0.01088	0.02571	0.19627	0.03599	NaN	NaN	NaN
4	0.00216	0.27547	0.00022	0.03517	0.06665	0.04656	NaN	NaN	NaN
5	0.00058	0.32359	0.02574	0.08695	0.02493	0.02489	0.01242	NaN	NaN
6	0.80542	0.03209	0.05411	0.02932	0.00465	NaN	0.00013	0.00000	NaN
7	0.72742	0.00689	NaN	0.00637	0.03076	NaN	NaN	0.00030	NaN
8	0.00204	0.46717	0.00008	0.00609	0.01218	NaN	0.04884	0.13444	NaN
9	0.01616	0.06749	0.00250	0.07464	0.02902	0.00229	0.43625	0.15550	0.00760
10	0.02493	0.00765	NaN	0.00781	0.01021	NaN	0.11999	0.13397	0.18187
11	0.00000	0.00000	0.90654	0.00002	NaN	NaN	0.00002	0.00001	0.00267

- **4 Project**

2. (2 points) Describes methodology for network analysis.

Solution:

A co-offender subgraph, G_type, will be generated for each type of crime. Various graph measures will be calculated for each of these networks to determine if any stand out as noticeably different across crime types. We will compare the number of nodes, edges, isolated nodes, mean degree, and connected components. The largest connected component will be analyzed to measure edge density, degree distribution, diameter, average path length, clustering coefficient, homophily, and modularity. Centrality measures will not be included in this analysis as the focus is not on individual nodes. Observations will be made for each measure, and a hypothesis test will be conducted to determine if crime types can be distinguished by the structure of the network.