Networks

Analysis 3 for MIT 6.419x Data Analysis: Modeling and Applications

Evan Rushton

April 21, 2021

1 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?

The naive time complexity for matrix multiplication is also on the order of $O(n^3)$, which is the same as the proposed algorithm of adding one to each symmetric element of C_{ab} for each pair $((r,a),(r,b))$ per row r of A. There may be optimizations for either of these algorithms that give $O(n^{2.8})$ but these may increase size complexity and are beyond the scope of the question. (66 words)

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?

Bibliographic coupling indicates papers that cite the same references and can be computed by A^TA. Co-citation indicates references cited by the same paper and can be computed by AA^T. From the short lit review I performed, there is no clear-cut choice that will reliably identify more similar papers. Results will vary based on the span of time chosen, the research field, and the size of the network. A high bibliographic coupling will indicate papers that have many references in common because they deal with a similar topic within a field and build upon the same historical record. A high co-citation strength will indicate papers that are often referenced together. The works that cite them may span a broad set of subtopics and research frontiers that all build upon or depend on the co-cited seminal works. One interesting feature of co-citation networks is that they change over time, while bibliographic coupling is static. Figure 1 demonstrates the two similarity measures. (159 words)

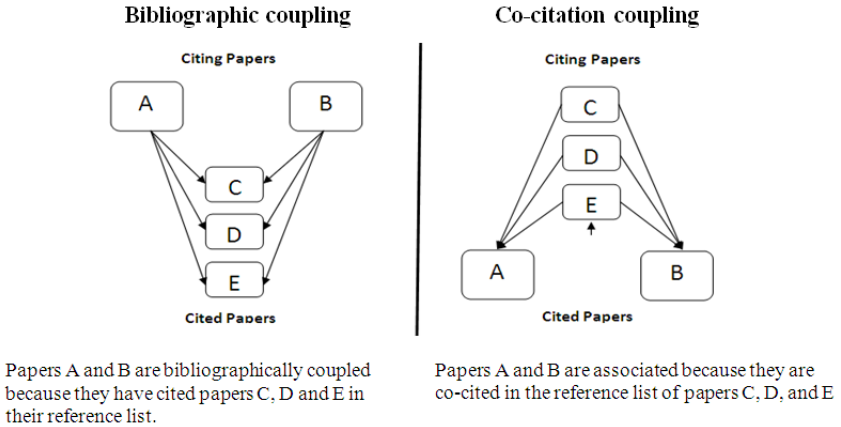

Figure 1: Bibliographic coupling and co-citation (Surwase et al. 2011, Co-citation Analysis: An Overview.)

2 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?

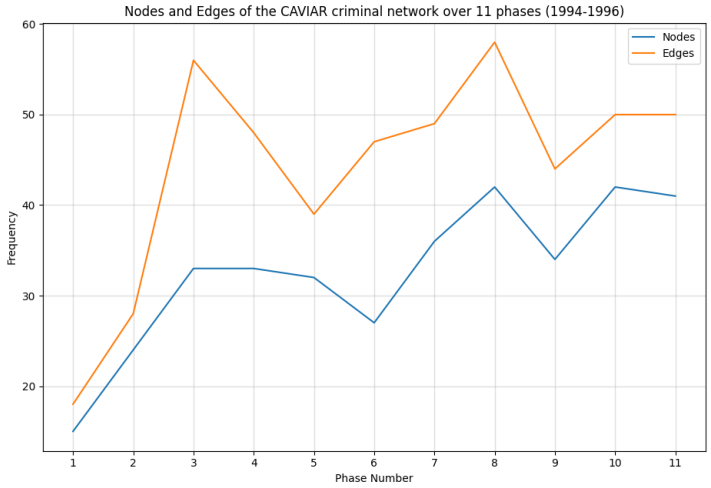

Figure 2: Number of nodes and edges in the CAVIAR network over time (each Phase is 2 months)

The plot of nodes and edges of the CAVIAR network over Phase number is shown in Figure 2. The number of (nodes, edges) rise 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 eigenvector. The top 5-10 players still seem relatively stable regardless of the initial lack of complete network knowledge. (181 words)

Part (d) (5 points) In the context of criminal networks, what would each of these metrics teach you about the importance of an actor's role in the traffic? In your own words, could you explain the limitations of degree centrality? In your opinion, which one would be most relevant to identify who is running the illegal activities of the group? Please justify.

Degree Centrality identifies nodes in the network with the most edges (highest degree). In the context of criminal networks, players with a high degree centrality would be in contact with many other players. This measure is limited because it isn't always the case that the players who are in contact with the most other players are in fact the most important players. Sometimes bosses try to lay low and not communicate with many individuals in the network. It is more likely that the second and third in command will give instructions to the individuals moving the product, and there will be regional leaders for the various activities that go on like financial affairs, transportation logistics, providing drugs, and buying drugs. (120 words)

Eigenvector Centrality identifies nodes in the network that have the most important neighbors. In the context of criminal networks, these players are in communication with many important players as defined as having a high degree. In the CAVIAR network this identifies Daniel Serero the mastermind, Pierre Perlini the principal lieutenant, and players in direct contact with them. (57 words)

Betweenness Centrality identifies nodes in the network that are contained in the most shortest paths between all nodes. In the context of criminal networks, players with high betweenness reduce the number of contacts it takes for information to travel across the network. One way to think about this measure is if a node with high betweenness were removed from the network, then communication would take longer across the network. It may even be that a node with high betweeness is the only node connecting different clusters of a network. I think this is the most relevant to identify who is running the illegal activities of the group because these nodes are critical to maintain communication across the network. In the CAVIAR network this identifies Ernesto Morales the principal cocaine organizer before some of the lesser players in direct contact with Serero and Perlini. (143 words)

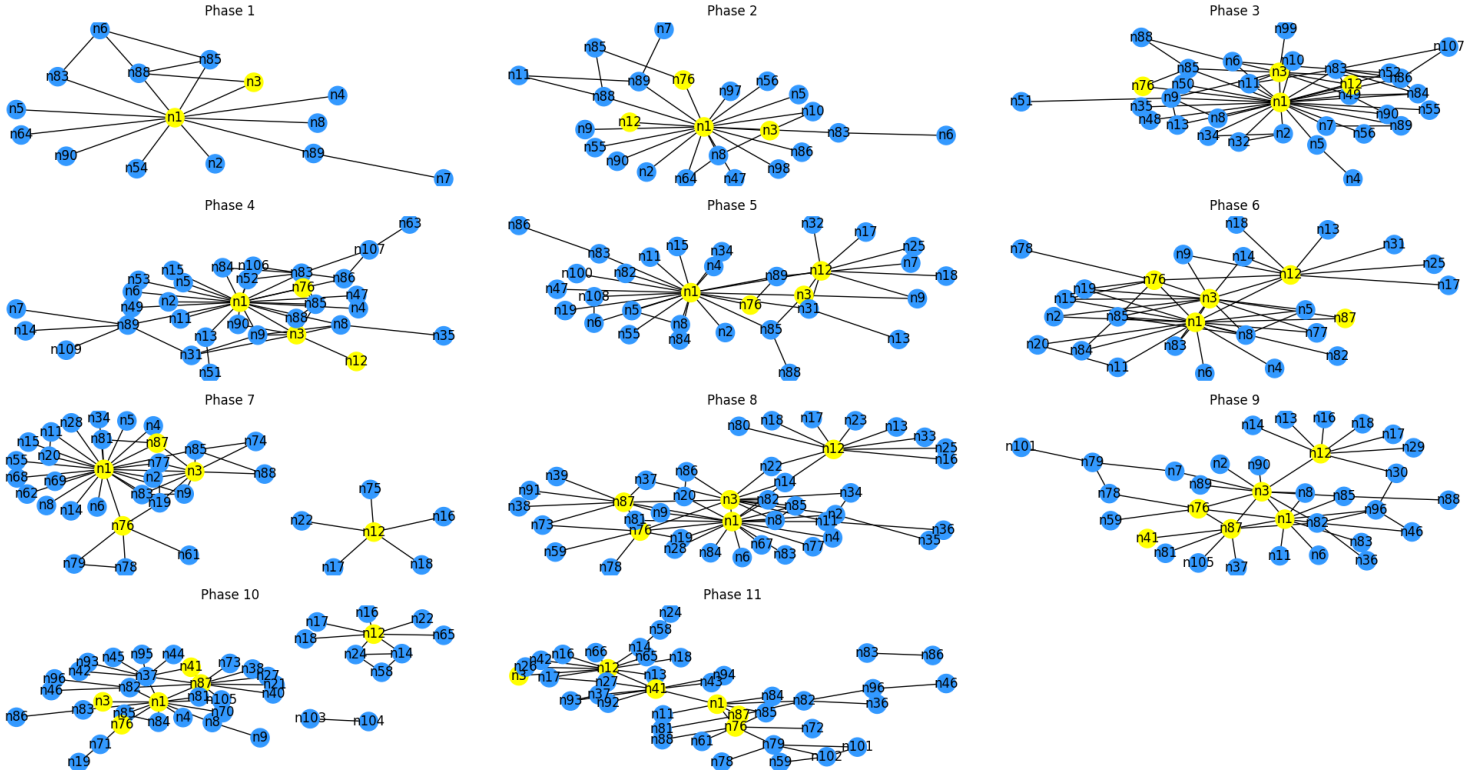

Figure 3: CAVIAR network over time (central players colored yellow)

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.

Table 1 and Table 2 show centrality measures for the top 10 betweenness centralities ordered by descending betweenness for all players in the network with and without imputed zeros. Looking at all centrality measures together gives 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 colored yellow in the network. Notice that they seem to be central even as the network reshapes. These central traffickers are (Daniel Serero, Ernesto Morales, Pierre Perlini, Gabrielle Casale, Patrick Lee, and n41). The cut-off of top 5 is arbitrary, and I tuned this parameter based on the location of the nodes in Figure 3. Nodes that appear to stay central are likely central players. Other players with high centrality like Antonio Iannacci (n89), Alain Levy (n83) and Salvatore Panetta (n82) are connected to the important players, but still appear peripheral in the network. (156 words)

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.

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. (63 words)

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?

Referring back to Figure 2 and 3 we see the initial growth of the network from Phases 1-3 (as the police built a complete picture) that centers primarily around Serero (n1) and Perlini (n3). The number of edges decreases between Phases 3-5 (from 56 to 39) and the network restructured to expand the cocaine imports centered around Morales (n12) as mentioned earlier. From Phases 6-8 the number of (nodes, edges) increased from (27, 47) to (42, 58). There were three smaller seizures of both marijuana and cocaine during Phase 6 followed by a large seizure of marijuana in Phase 7. Lieutenant Perlini takes a more central role in the network during Phase 6. This may be a sign that Serero was getting worried about contacting too many players. The structure of the network is decentralizing Serero and handing more authority to n3, n76, n82, n85, and n86 from Phases 6-8. Morales (n12) is the center of his own connected component during Phase 7 and 10 possibly indicating a communication freeze between the marijuana and cocaine sides of the operation. It is also worth nothing that both Phase 7 and 10 correspond to large marijuana seizures. From Phases 9-11 Serero begins to reduce the number of edges directly tied to him and his lieutenant Perlini is also seen cutting ties during this time. In these later Phases the network looks less centralized and Patrick Lee (n87), Gabrielle Casale (n76), and n41 take on more central roles in the network. Given the huge seizure in Phase 10 of 2200kg marijuana ($18,700,000) it is not surprising that the network began to break ties and decentralize as can be seen in the Phase 11 graph. (281 words)

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.

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. (65 words)

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

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. (154 words)

Part (j) (4 points) Recall the definition of hubs and authorities. Compute the hub and authority score of

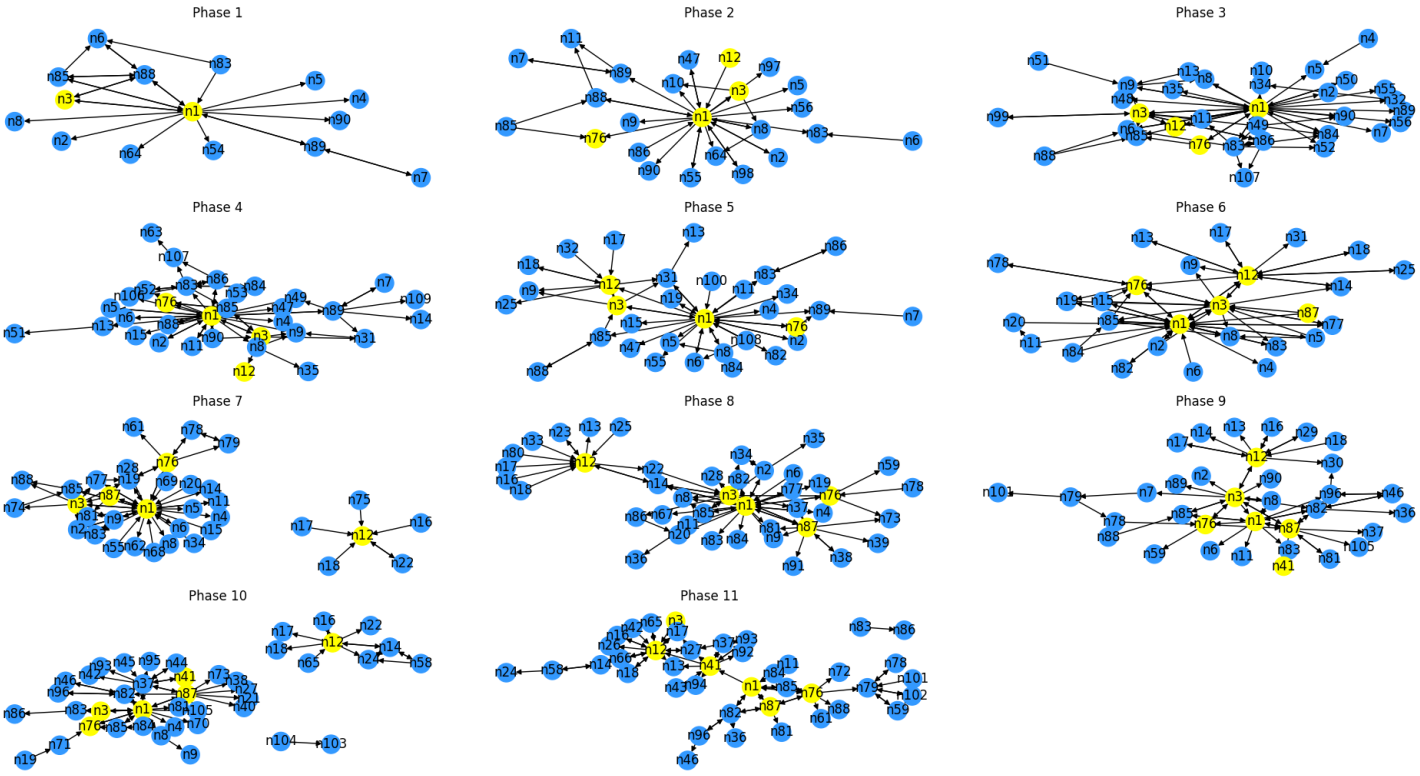

Figure 4: CAVIAR network over time (central players colored yellow)

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)?

Hub score is defined as the sum of the authority score of each node it points to and authority score is the sum of the hub score of each node that points to it. A player with a high authority score is contacted by players that are recognized as hubs of information and a player with a high hub score contacts players that are recognized as authorities. In the criminal network context 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 is shown in Table 3 and Table 4. In comparing n1 (Serero) and n3 (Perlini), notice that Serero's hub score is high for Phases 1-5 and 8-9, while Perlini's hub score is higher for Phases 6-7. A reverse pattern is true for authority score where Perlini is higher for Phases 1-5 and 8-9, while Serero's authority score is higher for Phases 6-7. This adds an additional layer to the earlier analysis. While it is true that Perlini has more edges in Phase 6 and seems to take a more central role, it wasn't clear that his role switched from collecting information to sharing information and vice versa for Serero. This evidence supports the claim that Serero was preparing to remove himself from the network. First by distributing other players as new hubs while maintaining his authority in Phases 6-7, and then acting as a hub with some final instructions in Phases 8-9 before removing himself. The fact that Serero and Perlini removed themselves from the network in Phases 10-11 is also evident by the drop in their hub and authority scores. (287 words)

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

3 Co-offending Network

Three subgraphs of the co-offender network are considered in this section. G (co-offenders) is the unweighted complete co-offender network with isolates removed. G_r (repeating co-offenders) is the unweighted co-offender network only containing edges for pairs of offenders who co-offended at least twice with isolates removed. G_{nr} (non-repeating co-offenders) is the unweighted co-offender network only containing edges for pairs of offenders who co-offended exactly once with isolates removed.

Part (g) (3 points) Plot the degree distribution (or an approximation of it if needed) of G. Comment on the shape of the distribution. Could this graph have come from an Erdos-Renyi model? Why might the degree distribution have this shape?

The degree distribution of the co-offender network in Figure 5 looks to follow a power law distribution. This could not have come from an Erdos-Renyi model since that follows a poisson distribution and does not follow a power law distribution. (40 words)

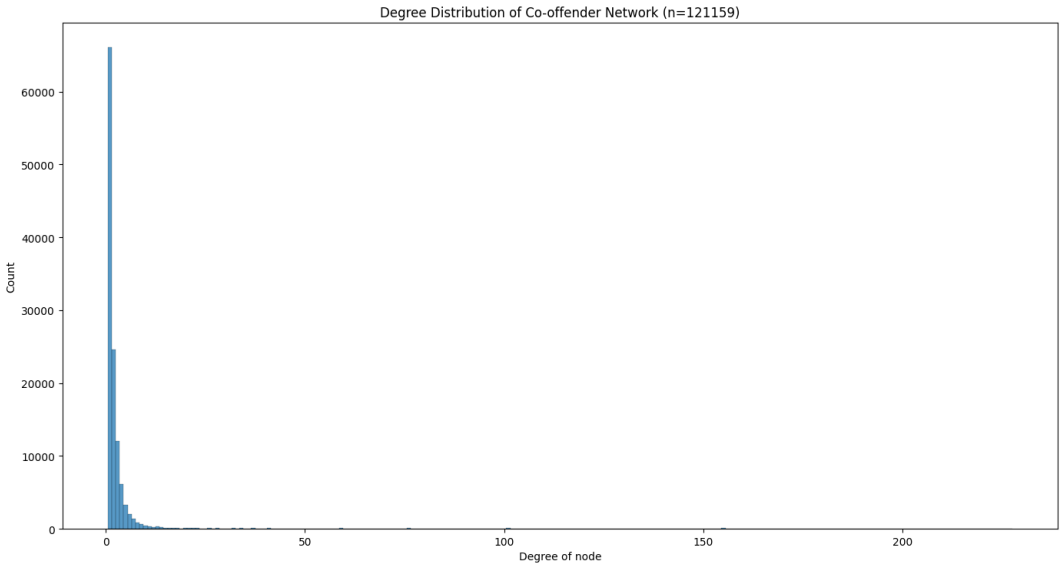

Figure 5: Degree distribution of the co-offender network G with isolates removed

The next questions focus on the largest connected components (LCC) of G, G_r, and G_{nr}. Summary statistics for the thee graphs are displayed in Table 5.

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

The clustering coefficients (cc) for each node in the LCC of G_r and G_{nr} give a measure for how close to being fully

Table 5: Network measures for three subgraphs of the co-offender network

Graph	Nodes	Edges	Mean Degree	Connected Components	Average Component Size	LCC Size	Fraction in LCC
G	121159	178413	2.94511	36098	3.35639	19924	0.164445
G_r	17764	15448	1.73925	6463	2.74857	66	0.00371537
G_{nr}	113519	162965	2.87115	35596	3.18909	12086	0.106467

connected each node is with their neighbors. I expect there to be one or two nodes in each neigborhood that are more connected as the leading actors, and the rest sparsely connected as peripheral players in each crime ring. Their distributions are shown in Figure 6. Values were obtained using $C_i = \frac{1}{k_i(k_i-1)} \sum_{j,k} A_{ij}A_{jk}A_{ki}$ calculated by `networkx.clustering(G)`. The shape of both distributions is bimodal with a majority of nodes with cc = 0 (meaning they form no triangles with their neighbors) and the second majorty of nodes with cc = 1 (meaning they form triangles with all of their neighbors). This is in line with my initial guess that some players act as the leaders in co-offending crimes and the other players are peripheral and only in contact with the leader. The fact that there are very few intermediate clustering coefficients and so many strictly equal to 1 and 0 is suprising and goes to show that there isn't an organized network hierarchy, but rather an individual leader who organizes the rest of the criminals for each coordinated criminal activity. The bimodal split of (cc = 0, cc = 1) is similar for each graph. G_r: (69.25% cc=0, 24.2% cc=1); G_nr: (64.15% cc=0, 25.7% cc=1). For comparison, G: (59.9% cc=0, 30.0% cc=1) indicating that this trend is consistent. (240 words)

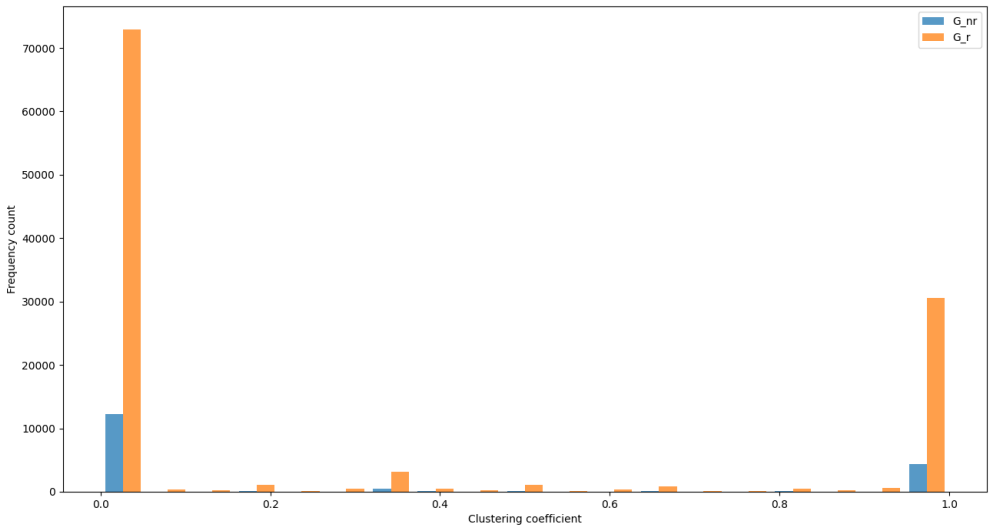

Figure 6: Distribution of clustering coefficients for G_r and G_{nr}

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

Degree and eigenvector centrality were computed for nodes in the LCC of G_r and G_nr. The descriptive stats for these centralities are shown in Table 6. The most central actors in the repeat offender graph as measured by these centralities had a high numbner of repeat crimes, which is not a surprise.

Table 6: Summary statistics of centrality measures for G_r and G_{nr}

Graph	DC Min	DC Median	DC Max	DC Mean	DC SD	EC Min	EC Median	EC Max	EC Mean	EC SD
G_r	0.015385	0.030769	0.138462	0.045688	0.032827	4.939356e-04	2.429702e-02	0.405597	0.070578	0.101620
G_{nr}	0.000083	0.000248	0.018701	0.000723	0.001721	2.276354e-29	1.596741e-21	0.084271	0.001058	0.009035

4 Project

Clearly states a sociological question which is interesting and relevant to the data:

How does the type of crime impact the network? Are there certain local structures, such as cliques or star graphs, that are associated with different types of crime? Can you identify different types of crime by the structure of a co-offending relationship alone?

(2 points) Describes methodology for network analysis.

A co-offender subgraph, G_{type} will be generated for each type of crime. A number of graph measures will be generated for each of these networks to determine if any stand out as being noticeably different across crime types. The number of nodes, edges, isolated, mean degree, and connected components will be compared. The largest connected component will be used to measure edge density, degree distribution, diameter, average path length, clustering coefficient, homophily and modularity. Centrality measures will not be used for this analysis as individual nodes are not the focus of the analysis. Remarks will be made for each measure and a hypothesis test will be performed to decide if crime types can be determined by the structure of the network.

(2 points) Presents results, including figures and/or statistics, which address the question of interest.

There are (number) crime types in the co-offender network and a subgraph G_{type} was made for each. The number of nodes, edges, isolates, mean degree, number of connected components, and size of the largest connected component is shown in Table

The largest connected component of each graph was further analyzed

(3 points) Provides commentary on what was discovered, what were the limitations of the methods, what may have been surprising to discover, etc.

Sadly, I ran out of time again. I would like to perform a hypothesis test to see if these subnetworks are much different from the parent netweeork and if the differences are significant.