MITx 6.419x Data Analysis
Statistical Modeling and Computation in Applications

Analysis 3
Network Analysis

Jonathan Chang (JonathanChang6d41)

April 21, 2021

Contents

Problem 1
Suggesting Similar Papers

Part (c)

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

First, let's take a closer look at each algorithm.

```python
def friends_algorithm(A):
  """
  Friend's method of creating co-citation
  matrix from adjacency matrix.

  :param      A:    adjacency matrix
  :type       A:    2d numpy array
  :return :         co-citation matrix
  :rtype  :         2d numpy array
  """
  n = A.shape[0]
  C = np.zeros(shape=A.shape, dtype=int)

  # Go through the rows of A one by one
  for i in range(n):
    row = A[i, :]

    # Find pairs in each row
    for j in range(n):
      if row[j] > 0:
        for k in range(j + 1, n):
          if row[k] > 0:
            C[j, k] += 1
            C[k, j] += 1

  return C
```

Listing 1: Co-citation matrix, friend's algorithm.

```python
def my_solution(A):
  """
  Creating co-citation matrix from
  adjacency matrix using matrix
  multiplication. Unlike the lecture
  exercise, we zero out the diagonal so
  the two methods have the same output.

  :param      A:    adjacency matrix
  :type       A:    2d numpy array
  :return :         co-citation matrix
  :rtype  :         2d numpy array
  """
  n = A.shape[0]
  C = A.T @ A

  # Zero out diagonal.
  for i in range(n):
    C[i, i] = 0

  return C
```

Listing 2: Co-citation matrix, matrix multiplication.

In the worst case, if a paper cites every other paper, then there would be $\binom{n}{2} = n(n-1)/2$ possible pairs. For an arbitrarily large n, this amounts to $\mathcal{O}(n^2)$. The friend's algorithm finds on the order of n^2 co-citation pairs for each of n citing papers; therefore, it is $\mathcal{O}(n^3)$ overall.

Matrix multiplication calculates $\vec{a_j} \cdot \vec{a_j} = \sum_{i=1}^{n} a_{ij}^2$ (since the first matrix is transposed) for each cell of the resulting matrix. This operation is $\mathcal{O}(n)$, and there are n^2 cells. It is $\mathcal{O}(n^3)$ overall. However, due to vector optimizations in the hardware instruction set, and the interpreter possibly being able to look ahead and predict repetitive instructions to hand off to multiple threads, matrix multiplication *may* be faster.

Part (d)

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?

Bibliographic coupling and co-citations can be analogous to the complement of authorities and hubs, respectively. If many papers all cite one, then the cited paper might be considered an authority. If one paper cites many others, then it might be considered a hub. However, the coupled and co-cited papers would not be the authorities or hubs, but byproducts of them.

This analogy demonstrates that they are, in some respects, opposites. Papers might cite multiple sources to lend more credence to a single point; for example, in discussion. However, it is likely that many more citations support different points, such that they are not all redundant. Therefore, co-cited papers might be within some distance of each other in overall content, but do not make the same point.

A paper itself, alternatively, generally makes statements about one topic. So if multiple sources cite this same paper, then at least sections of those papers refer to the same topic. Therefore, bibliographically coupled paper have at least one specific section making similar points. The more couplings they make, the more sections may be similar.

Bibliographic coupling should be more appropriate as an indicator for similarity.

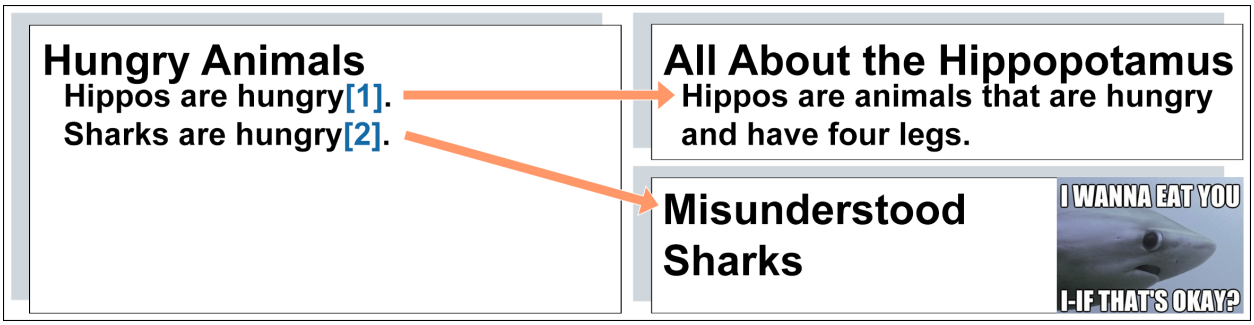


Figure 1.1: Co-cited papers are topically related in general, within some distance.

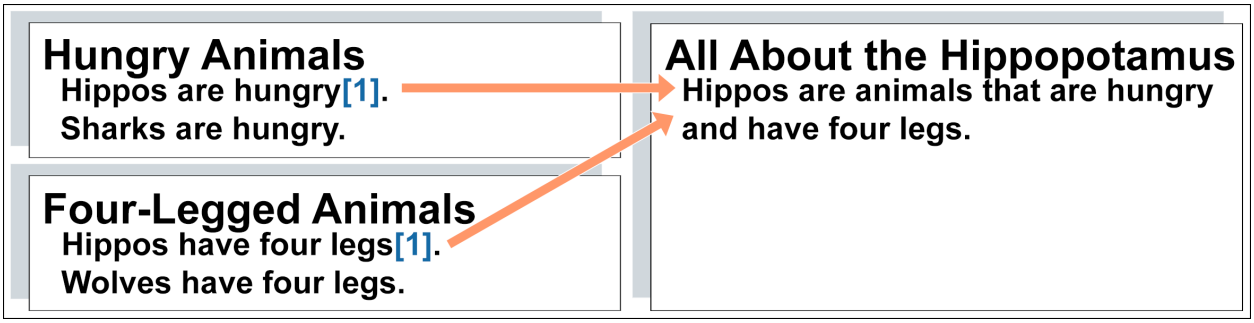
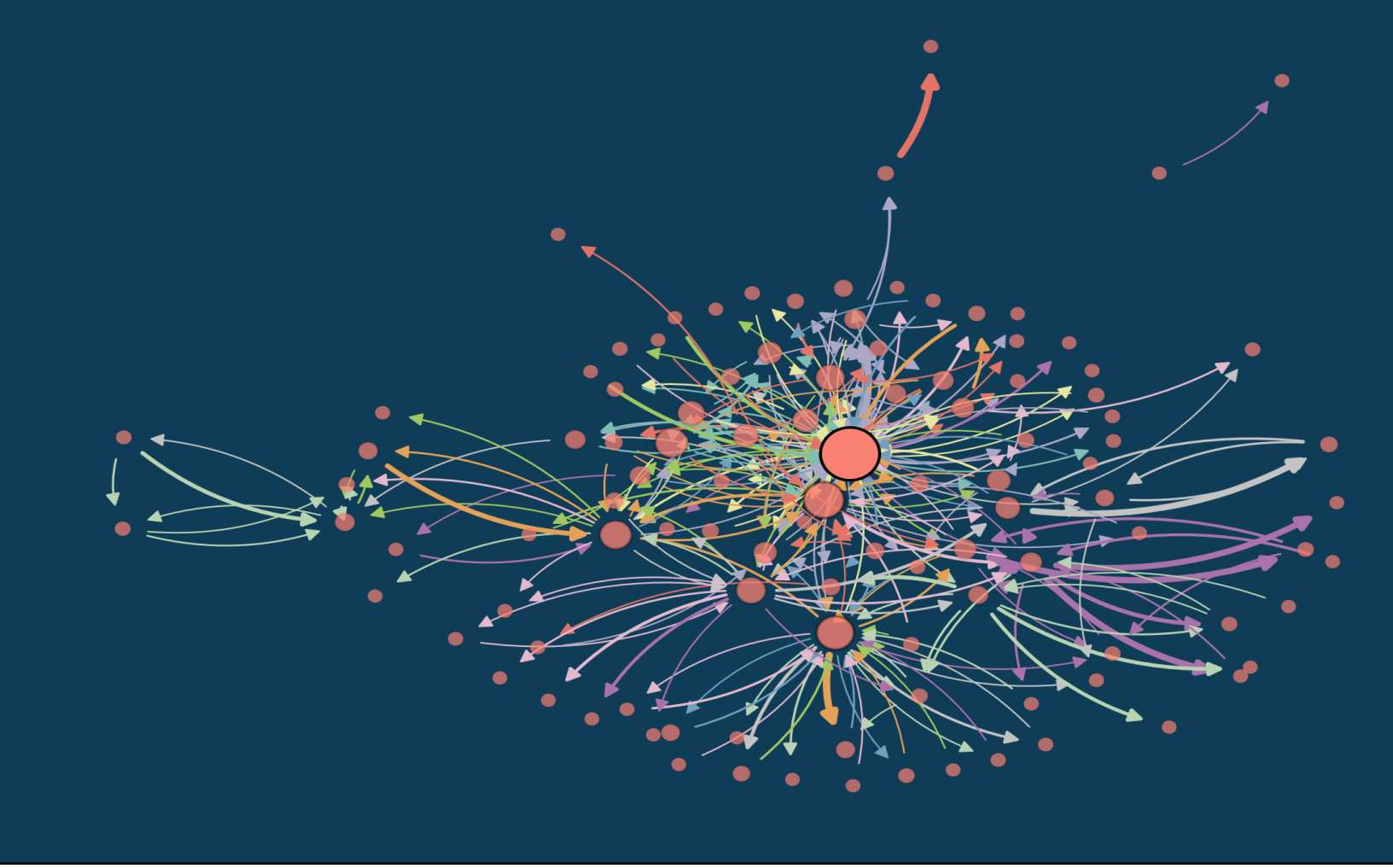


Figure 1.2: Bibliographically coupled papers have sections that are very similar.



Problem 2
Investigating a time-varying criminal network

Part (c)

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?*

The investigation started out with wiretaps on known main players of the network. As new information came in, it is reasonable to assume that the investigation expanded with more wiretaps to the peripheral players, sharply expanding until each of the central players have been covered. *Part (b) Question 5* finds the mean with regards to temporal consistency; however, if this assumption is correct, then many of the players may have consistent involvement, but were yet to be discovered at the beginning of the investigation; in that case, it wouldn't make sense to take the mean over every phase. On the other hand, we couldn't accurately extrapolate a player's importance from one phase to others, since the network isn't so stable.

Part (d)

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.

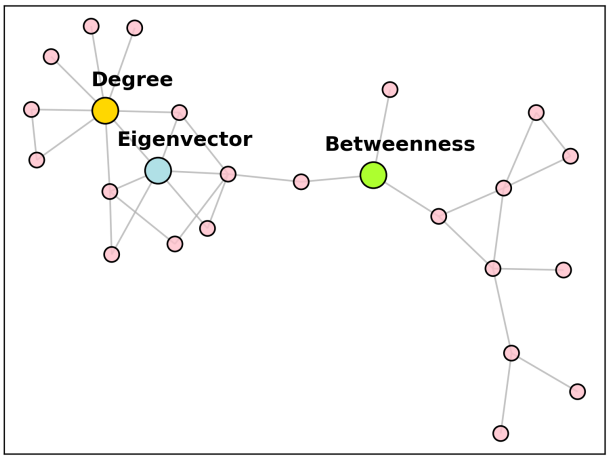


Figure 2.1: Difference between the centralities.

While degree centrality is easy to visualize, betweenness and eigenvector centralities take a bit more involvement. I devised a graph using my intuition to highlight these differences. One should immediately notice that a node could have a high degree, but not appear to be central or intimately connected in the network; for example, all its nodes are leaves. In a criminal network, hypothetically the manager of a small-time drug-running operation could deal with more people than a kingpin who acts through intermediaries.

Betweenness centrality, which measures the ratio of shortest paths passing through a node, is surely more influenced by all the clusters in a graph compared to the other two. The number of shortest paths on any side must be similar, so the node of highest betweenness centrality must be rather central to the graph. Consider that from *Part (b) Question 5*, the second highest node was that of a cocaine import intermediary between the Colombians and the Serero organization. That's what the betweenness identifies; not necessarily the top player in a tight network, but perhaps the people who bridge two tightly connected networks.

Eigenvector centrality extends degree centrality by measuring how relatively connected a node is by propagating the degrees of all its neighbors, their neighbors, and so on... It's a bit hard to interpret from the math once the iterations cause a node's weight to circle back and influence itself, but we could see from the graph above that it favors nodes that are situated amidst other nodes with high degrees. This makes the eigenvector centrality most suitable for identifying, if not the kingpin, the head of operations – the player who is coordinating with many other groups of people.

Part (e)

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.

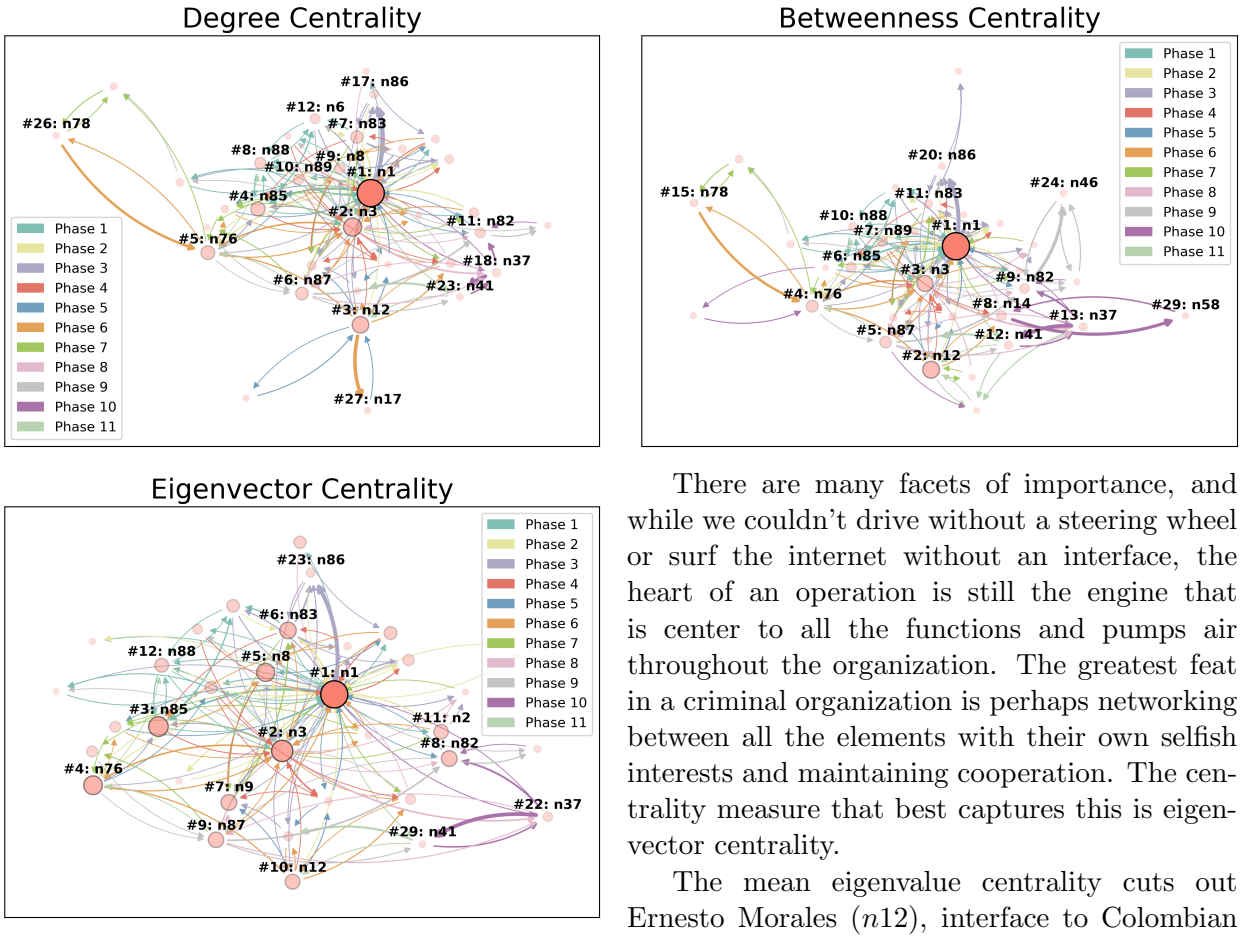


Figure 2.2: Phase at first detected communication and centrality rankings.

There are many facets of importance, and while we couldn't drive without a steering wheel or surf the internet without an interface, the heart of an operation is still the engine that is center to all the functions and pumps air throughout the organization. The greatest feat in a criminal organization is perhaps networking between all the elements with their own selfish interests and maintaining cooperation. The centrality measure that best captures this is eigenvector centrality.

The mean eigenvalue centrality cuts out Ernesto Morales ($n12$), interface to Colombian cartels, in favor of accountant Wallace Lee ($n85$). Tasked to recuperate the marijuana, Gabrielle Casale ($n76$) and Bruno de Quinzio ($n8$), are similarly favored by all three measures. Investor and money transporter Alain Levy ($n83$) is named as a top important player by eigenvector centrality, but not betweenness. Although, it's interesting to note that while his brother Gérard ($n86$) is more private, he communicates with the boss much more frequently. Edge weight isn't captured by any of the three measures of centrality. On top of this list is West End Gang leader[1] Daniel Serero ($n1$) and his right-hand man Pierre Perlini ($n3$). What's interesting is half this list consists of non-traffickers, including the VP of a financial firm[1]. Of course, money is just as important as the drugs.

Part (f)

The change in the network from Phase 4 to 5 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.

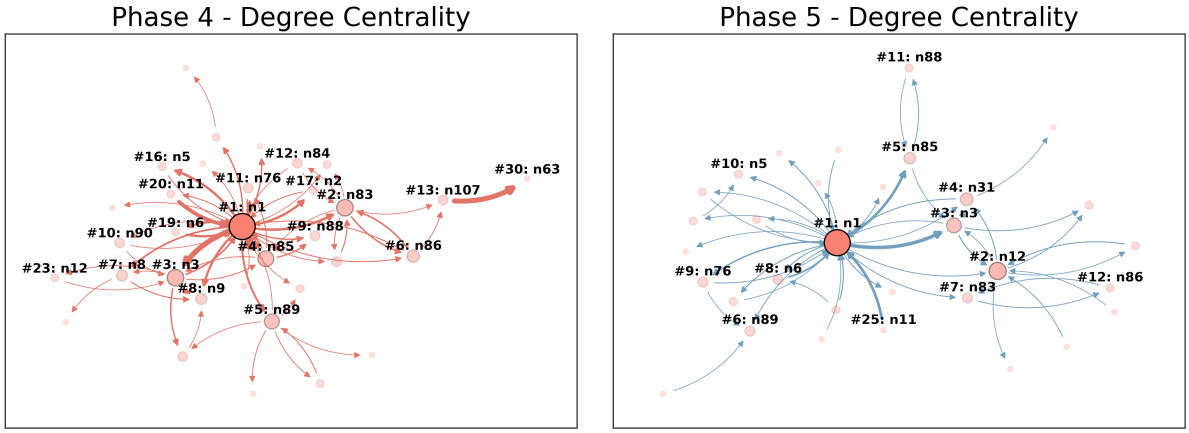


Figure 2.3: Degree centrality before and after phase 5. Top 12 ranks labeled, along with players with high levels of communication.

Although Ernesto Morales ($n12$) was first found to contact Daniel Serero ($n1$) in phase 2, it was primarily Serero that dealt with marijuana broker Gaspar Lino ($n6$) and Samir Rabbat ($n11$) throughout the investigation, through the accountant Wallace Lee ($n85$) and investor Alain Levy ($n83$). From phase 4, Pierre Perlini ($n3$) had established a line of contact with Morales, and then became the main cocaine contact between him and Serero. Throughout the remaining phases, Serero still handled the marijuana side of the business, and maintained more frequent contact with recuperators like Richard Gleeson ($n5$) and Gabrielle Casale ($n76$). Phase 4 marked the first seizure of 300kg of cannabis, which is probably why the organization sought to diversify or reorient its business, which brought Perlini to more prominence.

Despite so, comparing centrality rankings in the assignment is misleading since the trend is not so direct (phase 9 is somewhat of an outlier). Hub and authority scores show a more complex story from phase 6. Looking at betweenness, Alain and Gérard Levy ($n86$), who introduced Lino in the first 2 phases, distanced from communication, and both the investigation wiretapping and the business network expanded through Morales. In general, centrality rankings reflected the shift at #2 from Lee as major marijuana contact, to Morales as principle organizer of the cocaine import. In response, a there was a responsibility shift from Serero to Perlini in terms of who took initiative to communicate.

Betweenness Centrality

Phase	n1	n3	n6	n12	n83
1	0.91	0.00	0.01	N/A	0.04
2	0.94	0.002	0.00	0.00	0.09
3	**0.83**	**0.10**	0.03	0.00	0.05
4	0.84	0.09	0.00	0.00	0.08
5	0.88	0.04	0.00	0.27	0.06
6	0.54	0.23	0.00	0.38	0.00
7	0.59	0.07	0.00	0.02	0.00
8	0.55	0.31	0.00	0.36	0.00
9	**0.25**	**0.58**	0.00	0.36	0.00
10	0.34	0.00	0.00	0.03	0.03
11	0.53	0.00	0.00	0.43	0.00

Figure 2.4: Five important nodes at every phase.

Part (g)

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?

Investors and money transporters Alain ($n83$) and Gérard Levy ($n86$) were the main correspondents between the European marijuana trade, particularly broker Gaspar Lino ($n6$), and they often kept contact with VP of prestigious accountant firm Wallace Lee[1] ($n85$). Along with the drug runners ($n5$, $n8$ and $n76$), they formed the core marijuana operation of the Serero ($n1$) organization. It was Ernesto Morales ($n12$) (see phase 2) that first contacted Serero, and after the first seizure at the end of phase 4, Serero probably assigned his right-hand man Pierre Perlini ($n3$) to the task.

Thus, by phase 6, the organization had largely pivoted to the cocaine trade. Where Serero was making the calls in the beginning, it was then Perlini (see the hub score in phase 6), with Serero handling remnants of the marijuana operation. Phase 6 saw three seizures, so the organization cut contact with Morales in phase 7 to focus on marijuana, but with another seizure, that didn't last either. A more balanced organization reemerged, and by phase 8, the organization reached out to unnamed intermediary $n14$ to restore contact with the cocaine importer Morales.

A major reorganization occurred during phase 9: Patrick Lee ($n87$) heavily communicated with unnamed players $n41$ and $n37$, whom would later serve as intermediaries between Morales, replacing Perlini's role in the final phases. Perlini was in heavy contact with the investors before exiting.

As a point of interest, it has been reported that Daniel Serero only served 3 years in prison, and accumulated $55 million *despite* these drug seizures![2]

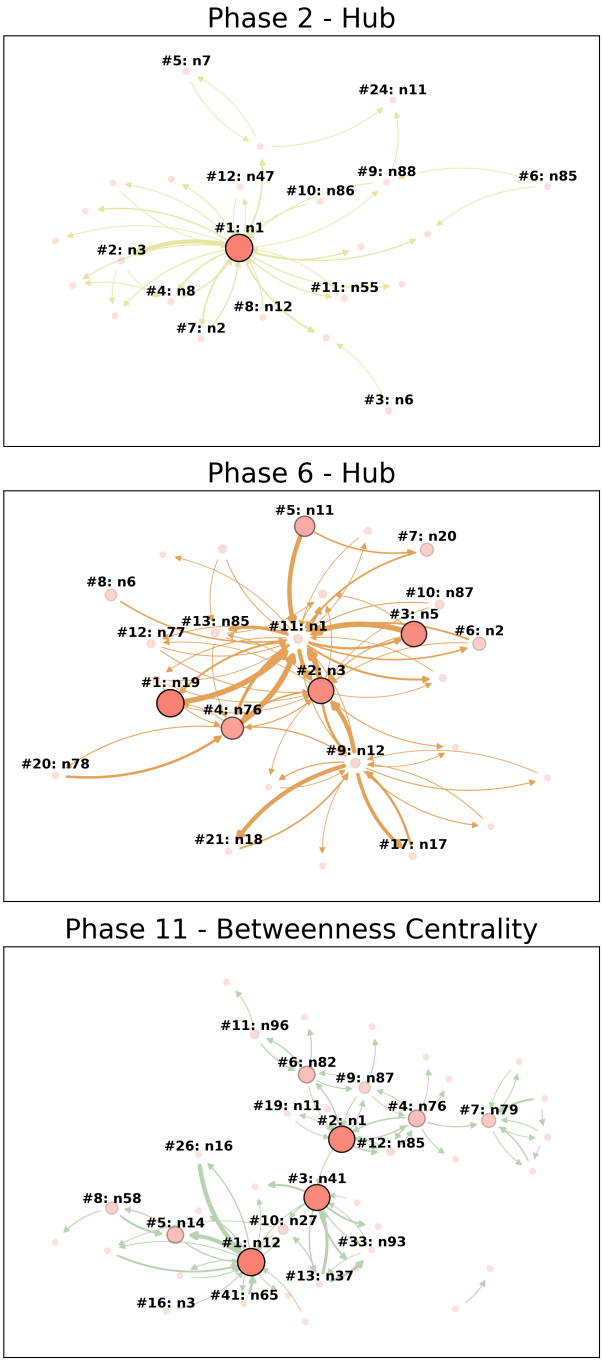


Figure 2.5: Some important phases.

Part (h)

Are there other actors that play an important role but are not on the list of investigation (i.e., actors who are not among the 23 listed above)? List them, and explain why they are important.

There are numerous players with prominent centrality rankings or were in frequent contact with important players that are not among the 23 named. There were $n9$ and $n19$, who were in frequent contact with Serero ($n1$) and Perlini ($n3$), as well as marijuana recuperators ($n5$, $n8$, $n76$) and cocaine importer Morales ($n12$) throughout nearly every phase. There was $n14$, who helped reestablish contact with Morales in phase 8. There was $n41$, who served as sole intermediary between Serero and Morales in phase 11, and $n37$, whom in phase 10 connected them to the transportation arrangements manager Salvatore Panetta ($n82$).

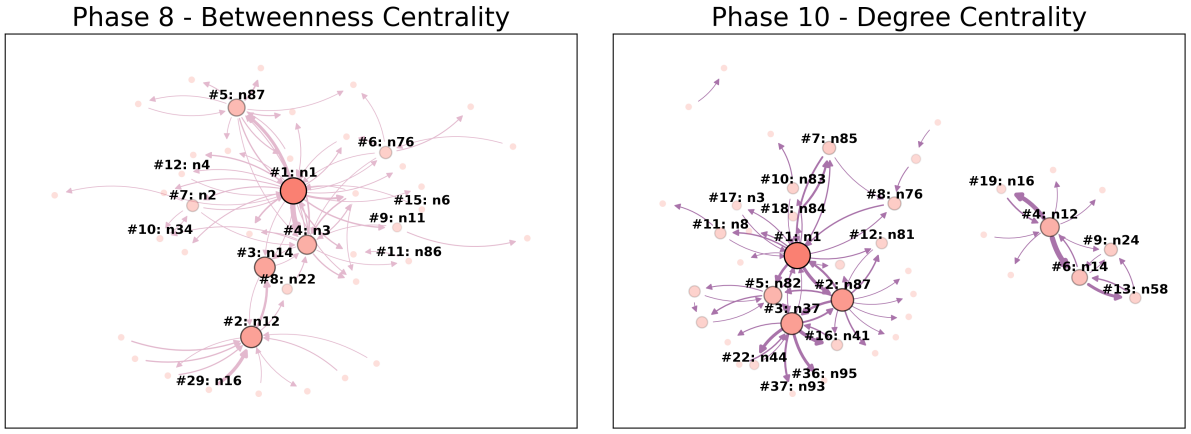


Figure 2.6: Highlighting $n14$ and $n41$.

Part (i)

What are the advantages of looking at the directed version vs. undirected version of the criminal network?

Unlike study citations, it's a bit hard to interpret in-degrees and out-degrees in a hierarchical social network, like a criminal network. At the very least, we could see who were initiating the lines of communication, although the consequences of that are up to interpretation. For example, it probably doesn't mean much to give orders to many peripheral players, whereas it would indicate importance to initiate many calls to other important players. It would be unclear how the latter compares to receiving many calls from other important players in terms of an importance measure. Further, this may depend on the stage of a relationship, whether at the beginning of a drug deal or the end, in which the buyer or seller might initiate respectively.

Regardless, with directed graphs, the left eigenvector centrality would only measure in-degrees, while the right would measure out-degrees. Such centrality measures, along with hubs and authorities, could be used to determine initiative as a precursor to network expansion or, perhaps, power. It could be argued that increasing contacts shifts the center of the network, and it's more likely to make that happen than to expect calls without effort. In that case, we could certainly use this to inspect network dynamics, such as the rising importance of Pierre Perlini ($n3$) or Ernesto Morales ($n12$), or the falling importance of Alain Levy ($n83$).

Part (j)

Recall the definition of hubs and authorities. Compute the hub and authority score of each actor, and for each phase. Using this, what relevant observations can you make on how the relationship between $n1$ and $n3$ evolves over the phases. Can you make comparisons to your results in *Part (g)*?

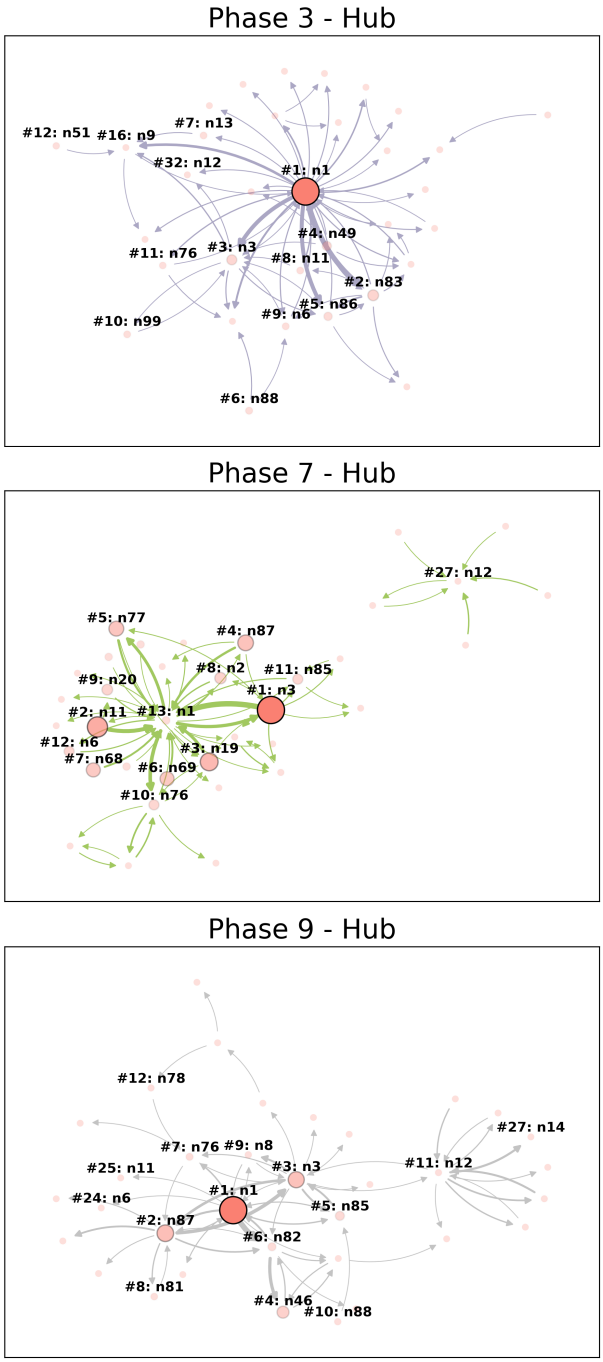


Figure 2.7: Shift of roles between Daniel Serero ($n1$) and Pierre Perlini ($n3$).

Whether it's due to a limited early investigation, Pierre Perlini ($n3$) didn't play a prominent role in the early phases. Daniel Serero ($n1$) coordinated with a team of investors and money transporters on one hand, and the marijuana drug runners on the other. Perlini contacted them here and there. Ernesto Morales ($n12$), the cocaine importer, first contacted Serero in phase 2, and Serero reciprocated in the subsequent phase. By then (see phase 3), Perlini's role increased somewhat, but it was still Serero making calls and Perlini largely on the receiving end. Serero being the top hub reflects that. By phase 4, Perlini became the primary contact for Morales, and after the first marijuana seizure, the organization pivoted towards cocaine. But not completely, since Serero maintained the marijuana trade while Perlini tread new ground. From phase 5 onward, we could already see the network fragmenting into two clusters.

Perhaps the reversal in rankings, however, is due to Perlini's departure as much as his rise in power. As he worked to gear the drug runners and investors toward the cocaine deal, Serero comfortable cashed in on preexisting connections; hence Serero received more calls, and by phase 6, he became the top authority, and Perlini top hub. It's a role reversal that persisted even after the second round of seizures after phase 6; despite temporarily losing contact with Morales, Perlini maintained with the rest of his network.

In phase 8, Ernesto reestablished contact through unnamed intermediaries $n12$ and $n14$, and investor Patrick Lee ($n87$) made his way onto the scene. Serero added Steve Cunhra ($n96$) and Salvatore Panetta ($n82$) to the roster to establish a direct line into Morales's circle,

circumventing Perlini. He regains top hub for these efforts. Perlini, still in between Serero and Morales's efforts, regained top authority. Perlini was in heavy talks with Lee, and it would be Lee that later brought unnamed players $n37$ and $n41$ to replace Perlini's role as primary coordinator between Morales. With that, Perlini was relegated to a minor player by phase 10, with the investor Lee and transport Panetta maintaining high centrality ranking until the end.

It's hard to tell exactly what happened to Perlini near the end, whether he willingly ceded his position, or he fell from favor. Regardless, I'm sure you, dear reader, could agree that this is largely similar to the analysis in *parts (f)* and *(g)*, since I constructed the narrative first before I answered any of these questions, after all... ☺

Phase	Hub		Authority	
	Serero	**Perlini**	**Serero**	**Perlini**
1	**0.7031**	0.0144	0.0118	**0.1357**
2	**0.9730**	0.0076	0.0003	**0.3367**
3	**0.7931**	0.0463	0.0032	**0.1496**
4	**0.8598**	0.0240	0.0022	**0.2755**
5	**0.9065**	0.0105	0.0006	**0.3236**
6	0.0080	**0.1953**	**0.8054**	0.0321
7	0.0068	**0.3433**	**0.7274**	0.0069
8	**0.8259**	0.0174	0.0020	**0.4672**
9	**0.5879**	0.1395	0.0162	**0.0675**
10	0.0000	0.0000	0.0000	0.0000
11	0.0000	**0.0379**	0.0000	0.0000

Figure 2.8: Hub and authority scores between Daniel Serero ($n1$) and Pierre Perlini ($n3$). **Bold** is highest.

Problem 3
Co-offending Network

Part (g)

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?

At first glance, the graph has a mean degree of 2.945, between 2 and 3, and one large component of 19924, which is consistent with the behavior of Erdos-Renyi when $np = c < \log(n)$. But a closer examination shows that the other components are larger than they're supposed to be under the model, and the tail distribution is too wide to be Poisson, which the Erdos-Renyi approaches for a large n. The graphs show this comparison, and the log-log power law test, in which a straight line shows this might be a power law distribution – definitely *not* Erdos-Renyi.

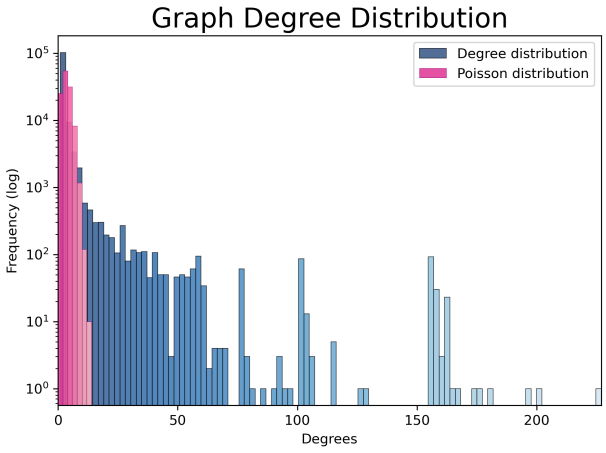


Figure 3.1: Degree distribution has wider tails than Poisson, indicating it's not Erdos-Renyi.

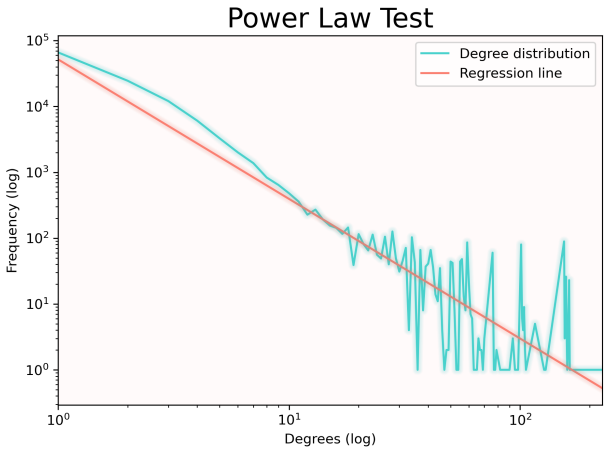


Figure 3.2: Relatively straight log-log plot indicates it might be a power law distribution.

Part (m)

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?

A single crime event is represented as a fully-connected induced subgraph; therefore, offenders that only co-offend with members within the same group would have a clustering coefficient of 1.0. Otherwise, interlinked co-offending groups where each member has co-offended with other members in different crimes, or cliques, would also give a coefficient of 1.0. Within repeat offenders, however, 88% of them are members of components sized 10 or less.

Since isolates have been removed from these graphs, offenders with one co-offender or many unconnected co-offenders in a line or star network would have a clustering coefficient of 0.0. 72% of them are in components sized 2.

Offenders that co-offend with multiple groups where *some* of each group co-offends with other groups might have coefficients in between. Within the top cluster (note that the y-axis is log-scale), between 35-40% of offenders belong to line or star subgraphs with 0.0 clustering coefficient. Around 25% belong to clique subgraphs with 1.0. This is true for both non-repeat and repeat co-offenders. The most common degree among repeat offenders is 4, as opposed to 1 for non-repeats; however, the mean degree is higher for non-repeats. Offenders with clustering coefficients between 0.25-0.75 have degrees on average twice as large as offenders with coefficients on both ends, with 8.7 degrees as opposed to 4.3-4.4. Many more small groups are cliques.

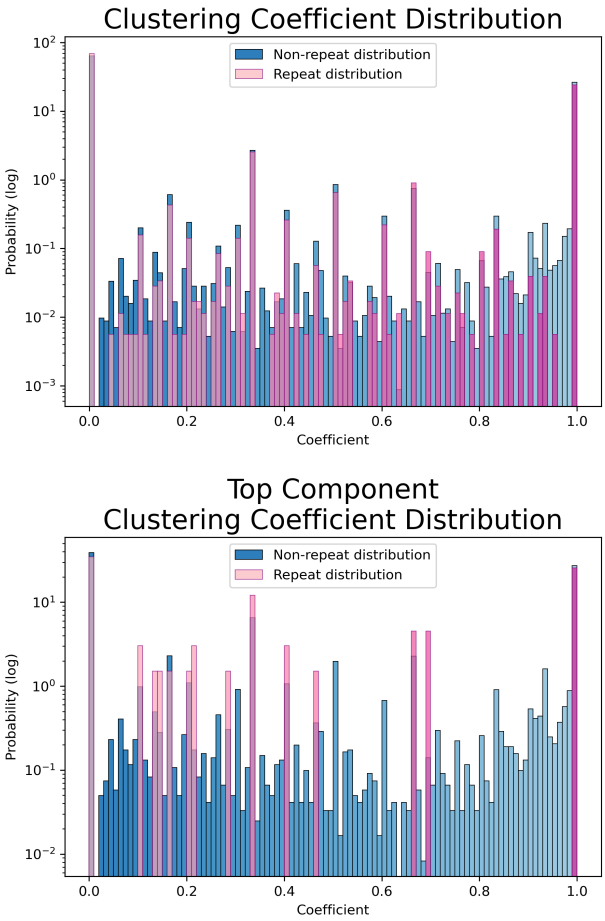


Figure 3.3: Clustering coefficients for repeat and non-repeat co-offenders.

The results are largely within expectations. I suspect that finding trustworthy partners should be among the greatest barriers to criminal enterprising. It would require people who are vigilant about self-protection, but also possess a degree of loyalty and intelligence. And if one such group amasses, it might have to be broken up by the FBI. As such, I'd imagine that most crimes are committed, if not solo, in small groups. That's why over 96% of components have 5 or less members. With the CAVIAR network as reference, I would expect that the centers of star networks are probably gang leadership, whereas cliques are part of operations.

Part (n)

Pick a centrality measure (degree, eigenvector, betweenness, etc) and compute the scores for the top 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?

The distributions are shown in log-log plots, due to the number of extremely small values. It's no surprise that repeat co-offenders might also commit crimes with others, hence they have higher centrality rankings, both in general and absolutely. I will primarily be studying the **eigenvector centrality** (degree is shown for curiosity). Repeat co-offender 596946 has the highest eigenvector centrality of 0.4056, with 36 crimes committed.

Since this is far below the most crimes by a single offender, I tried several means of justification. The average group size per crime is 1.917, which is in line with the top 10 offenders, who have committed between 196-456 crimes, all with average group sizes between 1-2. The offender has 13 neighbors (i.e. degrees), 8 are repeat co-offenders, above average but nowhere near the most. Of the neighbors, mean crime is 27.69, mean group size is typical at 1.67, and mean degree is 7.69. They're all adult males.

The neighbors or neighbors, and their neighbors as well, all have similar stats. And they all specialize in breaking and entering, robbery, with the occasional automobile theft, assault and drug charges.

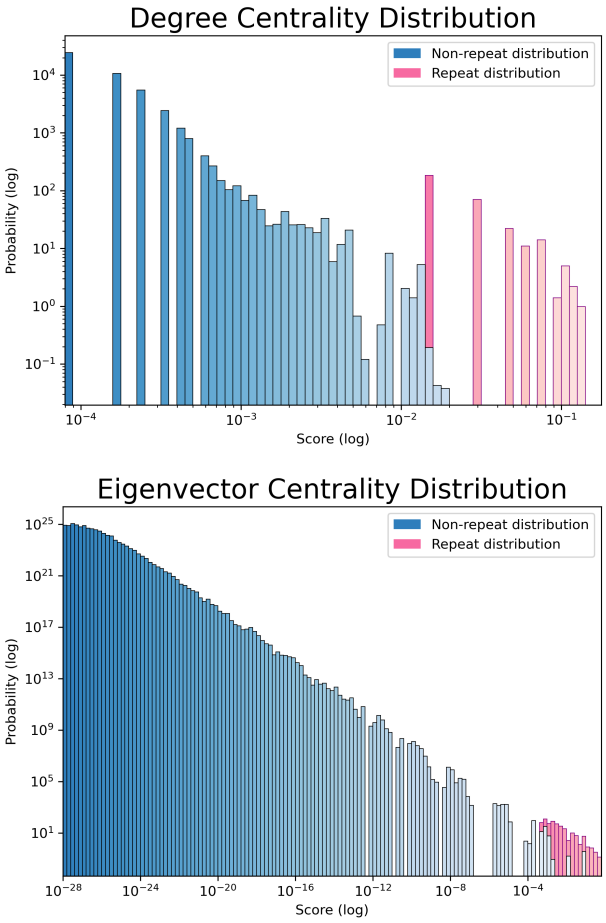


Figure 3.4: Centrality scores for repeat and non-repeat co-offenders.

While I wasn't sure what to expect at first, this certainly makes sense. Egregious crimes like murder have long sentences, preventing a broad network. Crimes with long sentences compel a smaller network, and sex crimes are usually personal. The offender is clearly gang-related with such a large network of co-offending criminals all focused around one thing. Of interest, the most central actor of degree centrality, with 15 degrees, is also a robber with a similar network.

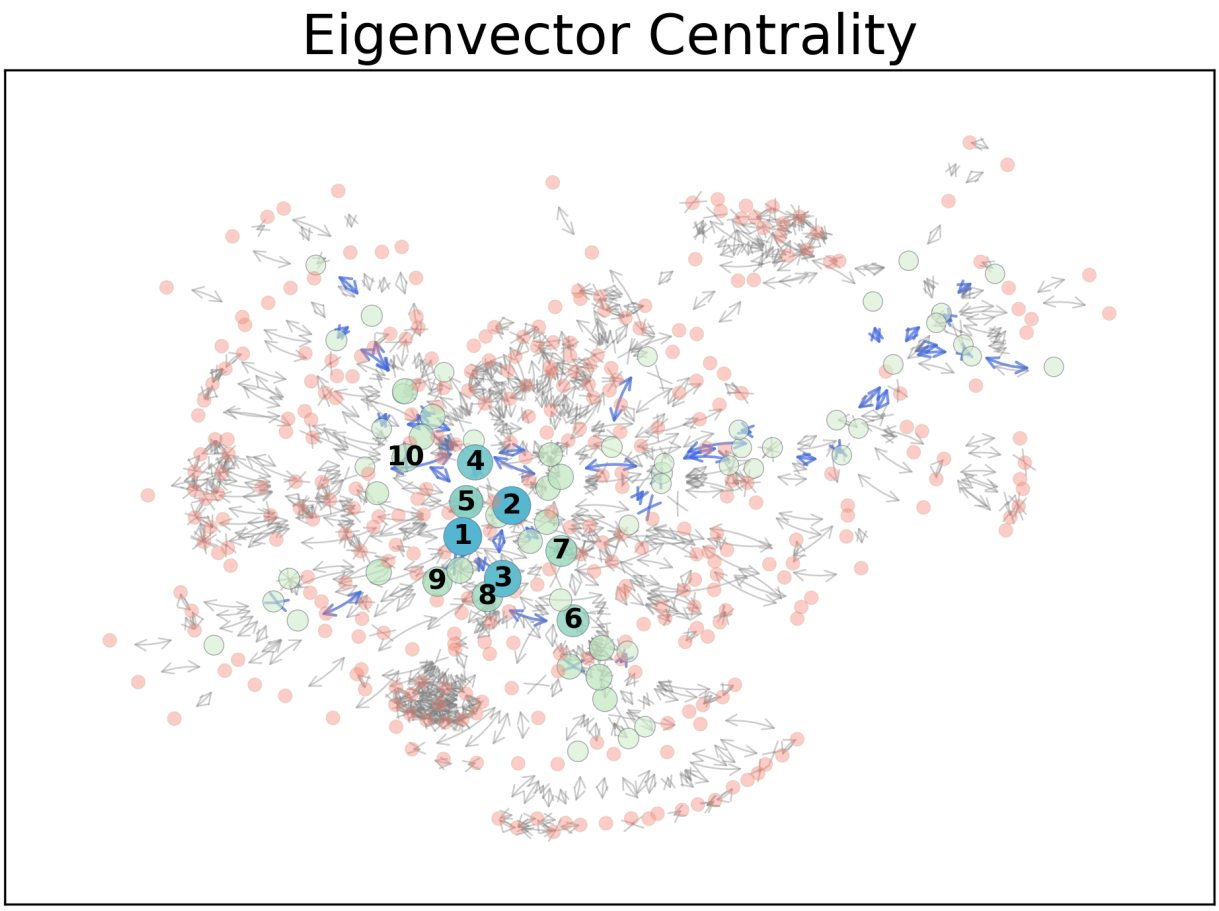


Figure 3.5: Eigenvalue centrality rank. Blue nodes are the repeat co-offender subgraph. Pink nodes are their neighbors, for context. They are mostly car and car property thieves.

Project
Impact of Crime On A Network

I.
Methodology

As we've previously seen, different clustering coefficients[m] and centralities[n] have sociological explanations. It's also conceivable that different kinds of crime could be conducted differently. For example, it was previously observed that robbers have high centrality probably due to that, by itself, misdemeanors have lighter sentences, that it's difficult to prosecute, or that there's a gang element that causes many such criminals to be associated. In contrast, personal crimes such as sexual offenses are usually committed in smaller groups.

I'll be exploring further whether the type of crime could impact co-offense network, and whether certain crimes have higher incidence of local structures such as cliques or star graphs. I'll look at some examples of the stories told by local structures with respect to crime type.

Prediction from local structures requires that the only information available be relative to a node. That is realistically, we could not obtain population data over the entire data set. Therefore, we should make use of summary statistics for just the node itself, an average of its immediate neighbors, and perhaps an average of the connect component it's in. Since co-offense is undirected, hub and authority scores are somewhat irrelevant, so we'll direct attention on clustering coefficient, degree, and betweenness and eigenvector centrality. Small components $n \leq 3$ are first removed because they are too common and there would be too little structural data to make an observation. For nodes in such components, we might as well make MLE predictions of the most common crime type.

Crime types are aggregated using the first 3 digits of the crime type number (e.g. `df['CrimeType'] % 100`). These crimes are grouped into related categories, such as (11) representing different types of murder, (12) conspiracy to murder, (13) sexual offenses, (14) assault, (21) theft, (31) prostitution, (41) drug possession, (42) drug trafficking, (212) robbery, $(213 - 214)$ automobile theft, and so on.

To determine whether any local structures exist, we could observe different components for each crime type graphically to gain some intuition. The graphs will focus on a particular crime color-coded on the blue-green spectrum depending on centrality ranking, along with a pink layer of their neighboring nodes for context. Cliques and star graphs of the crime in question will be colored as well. Every node will be labeled by the most common crime committed by the offender, which is often different from the crime type being examined. Graph will center on largest betweenness centrality. Eigenvector centrality based graphs will also be generated for larger components $n \geq 20$. We should be able to visually ballpark the clustering coefficient.

Finally, we will look at summary statistics for holistic perspective. They will be grouped by crime type for each of the statistics described above.

II.

Results

There are many examples of local graph structures that reveal something about the nature of the time. For some crimes, there are high occurrences of star graphs in larger components, such as running afoul of transportation laws such as the Taxi Transport Act (73007), Lotteries or Races Act (73003), or going beyond truck weight limits (73005). Together, these are grouped under the (730) label. Drivers might usually be with few people, and the long distance ensures unconnected co-offenders.

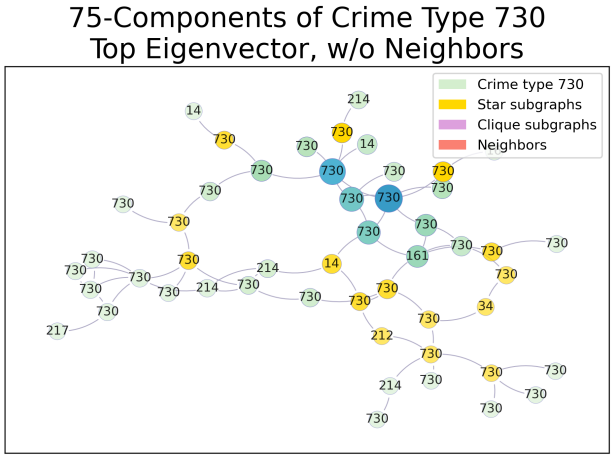


Figure 4.1: Long distance relationships.

Prostitutes (31) who are probably independent, *not* affiliated with a syndicate, tend to form small component star graphs.

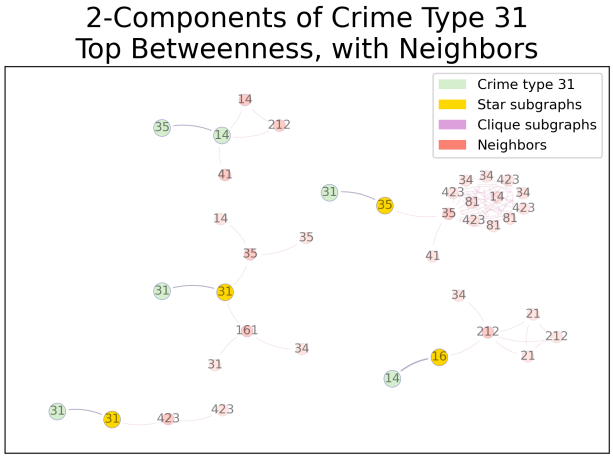


Figure 4.2: Traveling with friends?

Through these structures, we may be able to speculate on groups even *within* a crime type. For example, prostitutes who probably are connected to criminal organization tend to form large cliques bridged by few central actors.

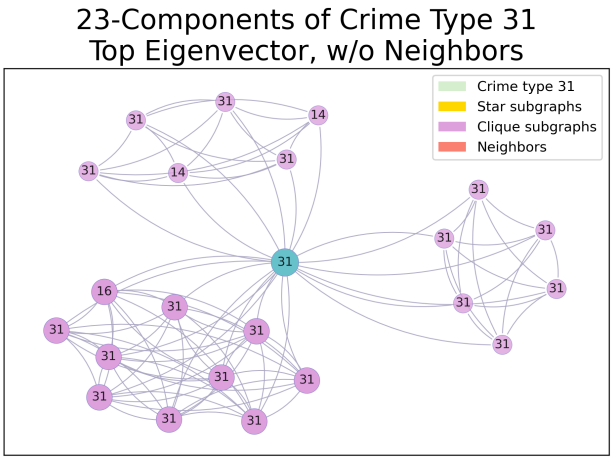


Figure 4.3: Could it be a pimp?

Another example of a large clique would be heroin addicts (41), whom seem to have no trouble finding buddies to use with.

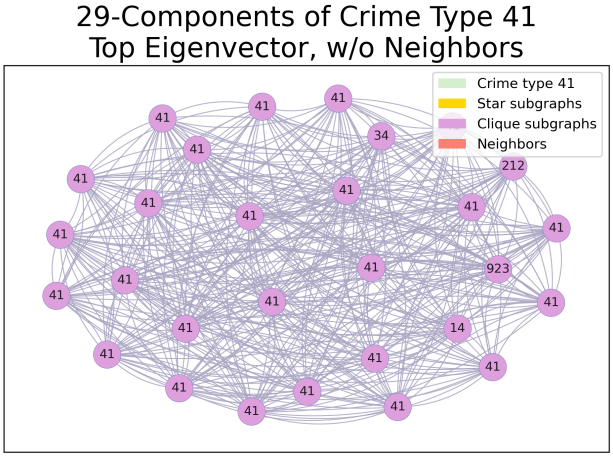


Figure 4.4: I'm sure it was a party.

But local structures extend beyond just whether they form clique or star graph patterns. Many networks are not quite either but close, and what might it tell us that just some people might not know each other? I think we should keep an open mind and look for general patterns as opposed to strictly clustering coefficients of 0 or 1. There could be less connected, large subgraphs of the same crime.

For example, we have this network of robbery (212), and weapons offenses and smuggling (33), respectively.

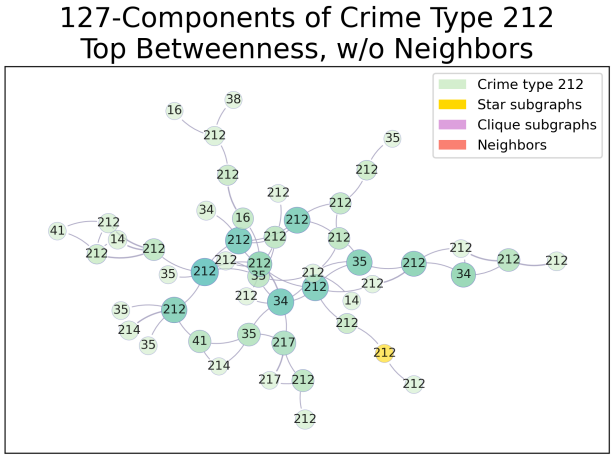


Figure 4.5: It looks street-gang related.

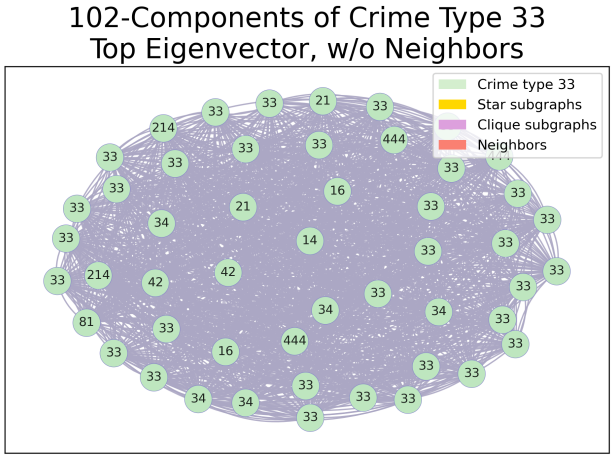


Figure 4.6: More sophisticated gang related.

Where it becomes interesting is when we look at their neighbors and find potential relationships between crime types. Some crimes are often committed in conjunction with other crime, and the graphs reveal that.

We might be able to classify two types of elucidations here:

1. Structures that tell us something about how certain crimes work in tandem.

2. Structures that demonstrate select central actors.

For the first, here's a central group of prostitutes that also deal in counterfeit money and law evasion (34), with key members connected to other prostitution groups. We could speculate that the central actors are probably management related.

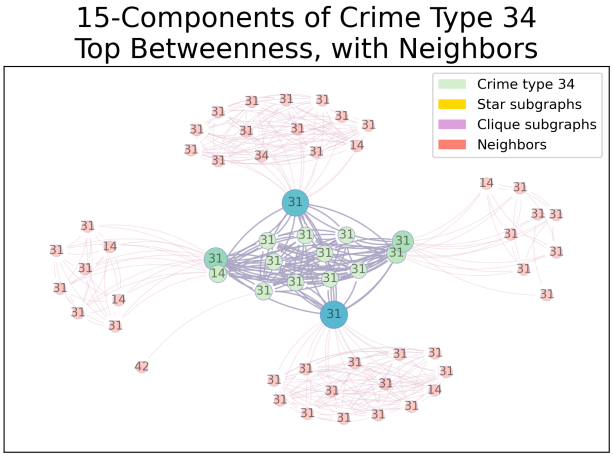


Figure 4.7: Intersection of multiple prostitution groups.

In other news, we could clearly see the structural link between drug importers (43) on one side and traffickers (42) on the other side of this network.

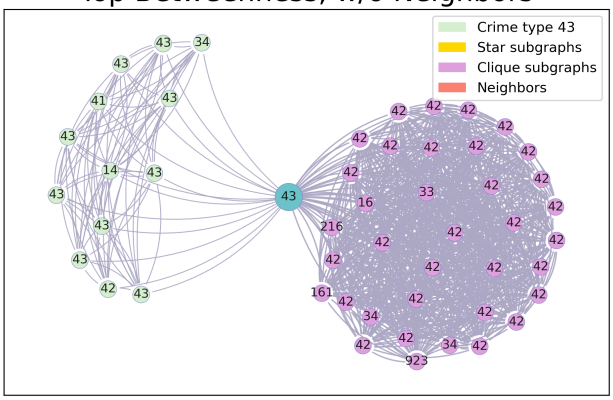


Figure 4.8: The two cultures?

The next example shows stolen cars (213) being potentially used to transport drugs (42). The entire network shown deals in automobile theft, but the highlighted subgraph with high eigenvector centrality coincides with trafficking and possession.

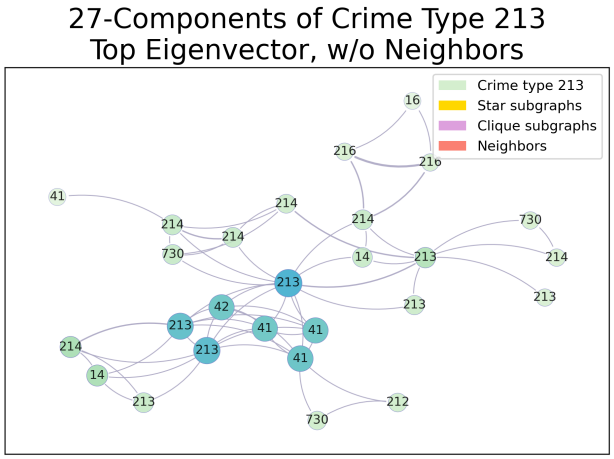


Figure 4.9: The two cultures?

The prostitutes form a clique, as well as traffickers, interfaced by one central actor that connects both.

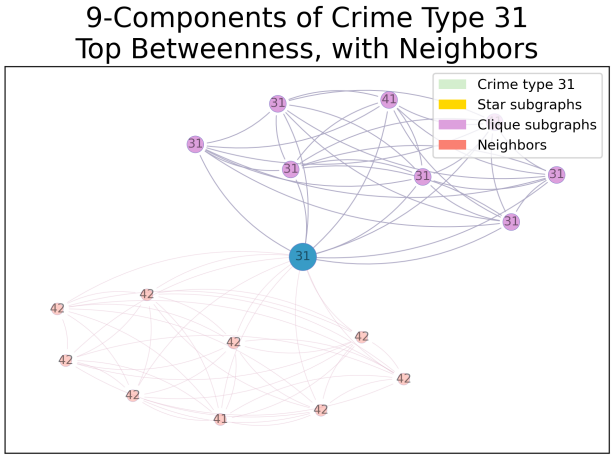


Figure 4.10: Sex, drugs and rock n' roll.

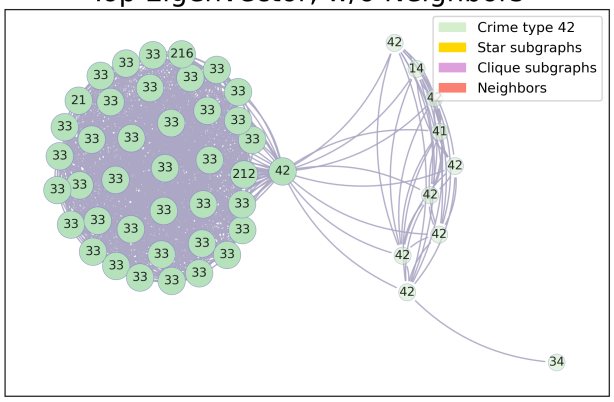


Figure 4.11: Drug kingpin.

As we round out the graphs, let's look at some examples of possible criminal enterprise leadership. The first one (above) has an army of armed (33) criminals behind his drug trafficking (42) network.

This next person is the single actor that connects two tightly connected group of car thieves (213).

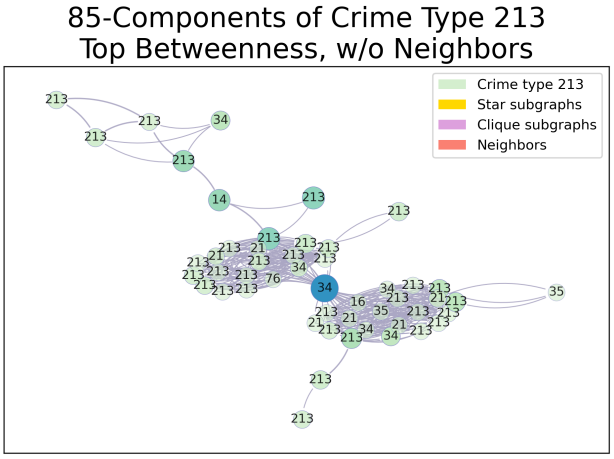


Figure 4.12: He's scoping out your BMW.

Finally, here is a cannabis grower (444) and trafficker (42) that mediates other traffickers, armed criminals (33), people who commit violent assault (14), and members of criminal organizations (399).

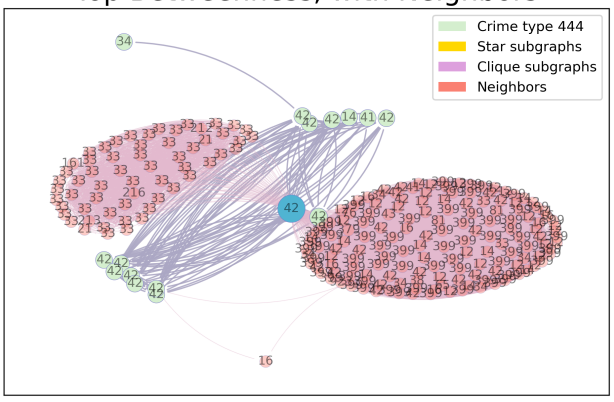


Figure 4.13: He owns all the hydroponic drug houses.

Here are the summary statistics per crime type.

Crime Type	Node				Neighbor				Component			
	Clst	**Degr**	**Eigv**	**Btwn**	**Clst**	**Degr**	**Eigv**	**Btwn**	**Clst**	**Degr**	**Eigv**	**Btwn**
11	1e-03	7e-05	2e-18	3e-05	1e-03	9e-05	9e-17	6e-05	1e-03	9e-05	2e-04	7e-05
12	2e-03	6e-05	4e-12	5e-05	1e-03	1e-04	3e-11	2e-04	1e-03	1e-04	3e-04	9e-05
13	1e-03	5e-05	1e-11	1e-04	9e-04	9e-05	9e-09	3e-04	1e-03	9e-05	2e-04	7e-05
14	1e-03	6e-05	1e-05	8e-05	1e-03	1e-04	3e-05	2e-04	1e-03	9e-05	2e-04	7e-05
15	1e-03	5e-05	2e-17	9e-06	2e-03	8e-05	2e-17	2e-04	2e-03	1e-04	4e-04	1e-04
16	1e-03	1e-04	2e-06	7e-05	1e-03	1e-04	5e-05	2e-04	1e-03	9e-05	3e-04	8e-05
21	1e-03	1e-04	2e-07	8e-05	1e-03	1e-04	7e-06	2e-04	1e-03	1e-04	3e-04	8e-05
31	2e-03	1e-04	5e-14	4e-05	2e-03	1e-04	1e-13	1e-04	2e-03	1e-04	2e-04	6e-05
32	2e-03	2e-04	6e-17	2e-05	2e-03	2e-04	1e-14	4e-05	2e-03	2e-04	1e-04	3e-05
33	3e-03	1e-04	2e-03	7e-05	3e-03	2e-04	2e-03	2e-04	2e-03	1e-04	3e-04	8e-05
34	1e-03	1e-04	3e-07	8e-05	1e-03	1e-04	4e-05	2e-04	1e-03	1e-04	3e-04	8e-05
35	1e-03	6e-05	2e-06	8e-05	1e-03	1e-04	5e-05	2e-04	1e-03	1e-04	3e-04	8e-05
37	2e-03	5e-05	7e-06	8e-05	2e-03	1e-04	7e-05	2e-04	2e-03	1e-04	3e-04	9e-05
38	3e-03	9e-05	6e-05	6e-05	3e-03	1e-04	9e-05	1e-04	2e-03	1e-04	3e-04	8e-05
41	1e-03	1e-04	3e-07	8e-05	1e-03	1e-04	4e-05	2e-04	1e-03	9e-05	2e-04	7e-05
42	2e-03	2e-04	2e-03	6e-05	2e-03	3e-04	2e-03	2e-04	2e-03	1e-04	3e-04	8e-05
43	3e-03	2e-04	1e-06	2e-05	3e-03	2e-04	2e-06	6e-05	2e-03	1e-04	2e-04	7e-05
44	1e-03	6e-05	2e-18	2e-05	1e-03	9e-05	6e-16	1e-04	1e-03	8e-05	2e-04	5e-05
45	1e-03	4e-05	5e-20	5e-05	8e-04	9e-05	3e-18	1e-04	1e-03	1e-04	3e-04	9e-05
46	2e-03	2e-04	8e-20	5e-05	1e-03	2e-04	2e-18	1e-04	2e-03	1e-04	4e-04	1e-04
49	2e-03	4e-05	1e-22	1e-04	2e-03	6e-05	1e-20	2e-04	3e-03	1e-04	3e-04	8e-05
51	1e-03	9e-05	2e-11	9e-05	1e-03	1e-04	9e-09	3e-04	1e-03	1e-04	2e-04	7e-05
52	2e-03	1e-04	1e-13	9e-05	2e-03	2e-04	2e-12	2e-04	2e-03	1e-04	3e-04	8e-05
53	7e-04	5e-05	3e-16	2e-04	5e-04	8e-05	1e-14	6e-04	1e-03	8e-05	2e-04	6e-05
54	3e-03	7e-05	8e-22	1e-05	3e-03	8e-05	2e-20	9e-06	3e-03	8e-05	1e-04	3e-05
61	3e-03	5e-05	3e-10	1e-05	3e-03	1e-04	4e-09	2e-05	1e-03	9e-05	3e-04	8e-05
62	3e-03	8e-05	2e-27	0e+00	3e-03	8e-05	2e-27	7e-10	2e-03	7e-05	2e-27	5e-10
64	1e-03	8e-05	6e-12	1e-04	1e-03	1e-04	2e-10	3e-04	1e-03	1e-04	3e-04	9e-05
65	3e-03	5e-05	1e-12	3e-05	2e-03	9e-05	6e-11	5e-05	2e-03	8e-05	2e-04	6e-05
76	1e-03	5e-05	7e-07	5e-05	1e-03	1e-04	2e-04	1e-04	1e-03	9e-05	2e-04	7e-05
81	1e-03	8e-05	1e-05	5e-05	1e-03	1e-04	1e-05	1e-04	1e-03	1e-04	2e-04	6e-05
91	4e-04	1e-04	5e-23	5e-04	4e-04	1e-04	2e-21	7e-04	1e-03	1e-04	4e-04	1e-04
92	1e-03	5e-05	7e-14	4e-05	9e-04	9e-05	1e-11	1e-04	1e-03	9e-05	2e-04	7e-05
93	1e-03	6e-05	4e-07	7e-05	1e-03	1e-04	7e-05	2e-04	1e-03	1e-04	2e-04	7e-05
94	4e-04	9e-05	8e-19	1e-04	5e-04	1e-04	1e-16	3e-04	1e-03	1e-04	4e-04	1e-04
145	2e-03	5e-05	3e-11	2e-05	1e-03	9e-05	4e-10	1e-04	1e-03	9e-05	3e-04	8e-05
146	1e-03	6e-05	1e-11	8e-05	1e-03	1e-04	2e-10	3e-04	1e-03	1e-04	3e-04	8e-05
151	2e-03	8e-05	9e-11	1e-04	2e-03	1e-04	4e-08	3e-04	2e-03	1e-04	3e-04	8e-05
161	1e-03	6e-05	2e-09	1e-04	1e-03	1e-04	3e-09	3e-04	1e-03	1e-04	3e-04	8e-05
162	3e-03	1e-04	3e-04	5e-05	3e-03	2e-04	3e-04	3e-04	2e-03	1e-04	6e-04	9e-05
167	1e-03	8e-05	4e-11	2e-04	1e-03	1e-04	2e-10	2e-04	1e-03	1e-04	2e-04	6e-05
211	1e-03	7e-05	7e-09	1e-04	1e-03	1e-04	8e-09	2e-04	1e-03	9e-05	3e-04	8e-05
212	1e-03	6e-05	9e-08	8e-05	1e-03	1e-04	1e-05	2e-04	1e-03	1e-04	2e-04	7e-05

Figure 4.14: Mean clustering coefficient and degree, eigenvector and betweenness centralities per node, neighbors and components. Some of the outliers are highlighted.

Crime Type	Node				Neighbor				Component			
	Clst	**Degr**	**Eigv**	**Btwn**	**Clst**	**Degr**	**Eigv**	**Btwn**	**Clst**	**Degr**	**Eigv**	**Btwn**
213	1e-03	7e-05	2e-07	9e-05	1e-03	1e-04	2e-05	2e-04	1e-03	1e-04	3e-04	8e-05
214	1e-03	6e-05	3e-07	8e-05	1e-03	9e-05	2e-05	2e-04	1e-03	9e-05	2e-04	7e-05
216	1e-03	1e-04	1e-05	9e-05	1e-03	1e-04	3e-05	2e-04	1e-03	1e-04	3e-04	8e-05
217	1e-03	7e-05	1e-06	6e-05	1e-03	1e-04	3e-05	2e-04	1e-03	1e-04	2e-04	6e-05
323	3e-02	1e-04	8e-03	7e-06	3e-02	2e-04	9e-03	4e-05	3e-02	2e-04	8e-03	3e-05
345	2e-03	3e-04	2e-15	6e-05	2e-03	3e-04	1e-13	2e-05	2e-03	3e-04	8e-05	2e-05
371	2e-03	1e-04	3e-11	8e-05	2e-03	2e-04	6e-11	2e-04	2e-03	1e-04	2e-04	7e-05
373	2e-03	7e-05	2e-11	7e-05	1e-03	1e-04	1e-09	2e-04	2e-03	9e-05	2e-04	8e-05
379	2e-03	8e-05	5e-14	3e-05	1e-03	1e-04	4e-12	8e-05	1e-03	1e-04	3e-04	8e-05
381	2e-03	5e-05	3e-09	3e-05	2e-03	9e-05	3e-09	2e-04	2e-03	9e-05	2e-04	7e-05
384	7e-03	5e-04	2e-06	5e-05	7e-03	7e-04	2e-06	8e-05	2e-03	2e-04	4e-04	1e-04
399	2e-03	4e-04	2e-04	1e-04	2e-03	4e-04	2e-04	3e-04	2e-03	1e-04	5e-04	1e-04
413	2e-03	8e-05	6e-10	8e-05	2e-03	1e-04	8e-09	3e-04	1e-03	1e-04	3e-04	1e-04
421	3e-03	5e-05	8e-13	3e-05	3e-03	9e-05	2e-11	1e-04	2e-03	9e-05	2e-04	8e-05
422	3e-03	2e-04	4e-03	5e-05	3e-03	3e-04	5e-03	2e-04	2e-03	1e-04	3e-04	8e-05
423	2e-03	2e-04	6e-06	8e-05	2e-03	2e-04	4e-05	2e-04	2e-03	1e-04	3e-04	9e-05
424	2e-03	2e-04	5e-06	5e-05	2e-03	2e-04	7e-05	1e-04	2e-03	1e-04	2e-04	7e-05
425	2e-03	5e-05	2e-04	3e-05	3e-03	9e-05	4e-04	7e-05	3e-03	9e-05	6e-04	7e-05
426	2e-03	2e-03	9e-05	1e-04	2e-03	2e-03	3e-04	2e-04	2e-03	1e-04	6e-04	1e-04
433	4e-04	4e-05	5e-22	1e-05	3e-04	1e-04	3e-20	2e-04	1e-03	1e-04	5e-04	2e-04
444	2e-03	1e-04	3e-03	3e-05	2e-03	2e-04	3e-03	8e-05	2e-03	9e-05	2e-04	5e-05
521	2e-03	5e-05	1e-12	2e-05	2e-03	1e-04	3e-10	2e-04	2e-03	1e-04	2e-04	7e-05
690	1e-03	6e-05	2e-15	2e-05	1e-03	1e-04	6e-14	4e-05	2e-03	1e-04	3e-04	8e-05
710	8e-04	2e-04	2e-14	2e-05	8e-04	3e-04	6e-13	6e-05	1e-03	2e-04	9e-05	3e-05
730	7e-04	6e-05	3e-13	2e-04	7e-04	1e-04	9e-12	4e-04	1e-03	1e-04	3e-04	1e-04
750	1e-03	5e-05	3e-10	4e-05	1e-03	1e-04	6e-08	1e-04	1e-03	9e-05	2e-04	7e-05
911	1e-03	5e-05	7e-21	5e-05	1e-03	9e-05	1e-18	6e-05	1e-03	9e-05	3e-04	8e-05
912	8e-04	9e-05	5e-12	1e-04	9e-04	1e-04	2e-11	2e-04	1e-03	9e-05	3e-04	8e-05
913	1e-03	8e-05	2e-07	9e-05	1e-03	1e-04	6e-05	2e-04	1e-03	1e-04	3e-04	9e-05
923	1e-03	6e-05	3e-07	7e-05	1e-03	9e-05	5e-05	2e-04	1e-03	9e-05	2e-04	6e-05
924	1e-03	6e-05	2e-12	1e-04	1e-03	1e-04	3e-12	2e-04	1e-03	9e-05	2e-04	7e-05
931	9e-04	4e-05	1e-11	3e-05	9e-04	8e-05	8e-10	1e-04	1e-03	8e-05	2e-04	6e-05

Figure 4.15: Results (continued).

III.

Discussion

Results were found that weakly indicates predictability of crimes based on network structure. While numerous individual examples indicate that patterns exist, it was difficult to gauge them from the summary statistics. While clique and star induced subgraphs could be found throughout the network for different crime types, there has not been a clear pattern. Many criminals commit multiple crimes, so different subgraphs coincide. Probably capturing just the node, neighbor and component coefficients and centralities is overly simplistic, and does not take into account many of these dynamics. Analogously, it is closer to a degree centrality than an eigen-

vector.

One thing I noticed is that only by exploring the graph intensively was I able to develop the imagination of how useful it might be. For example, the police might use these networks to discover the modes in which crimes are committed, figure out how criminals connect with each other, and even predict where and what kind of future crimes might exist by gauging the similarity of a person or community to these profiles. In retrospect, the results presented aren't surprising per se, I could imagine the relations between these crimes. But they are illuminating nevertheless, maybe because they solidify my hunches.

Next steps would be to take these statistics by node and train a logistic regression or neural network with the crime types as labels to see if we could predict some of the crime types with greater accuracy. Such a model could then be used to see if a criminal that fits the profile might have other potential connections, and serve as a launching pad for investigation.

List of Figures

List of Listings

References

[1] Martin Stone. "Canada arrests 30 for drug trafficking," *United Press International*, April 18, 1996. [Online]. Available: https://www.upi.com/Archives/1996/04/18/Canada-arrests-30-for-drug-trafficking/4693829800000/. [Accessed April 10, 2021].

[2] "Daniel 'The Arab' Serero," *Oocities*, October, 2009. [Online]. Available: https://www.oocities.org/wiseguywally/DanielSerero.html. [Accessed April 11, 2021].