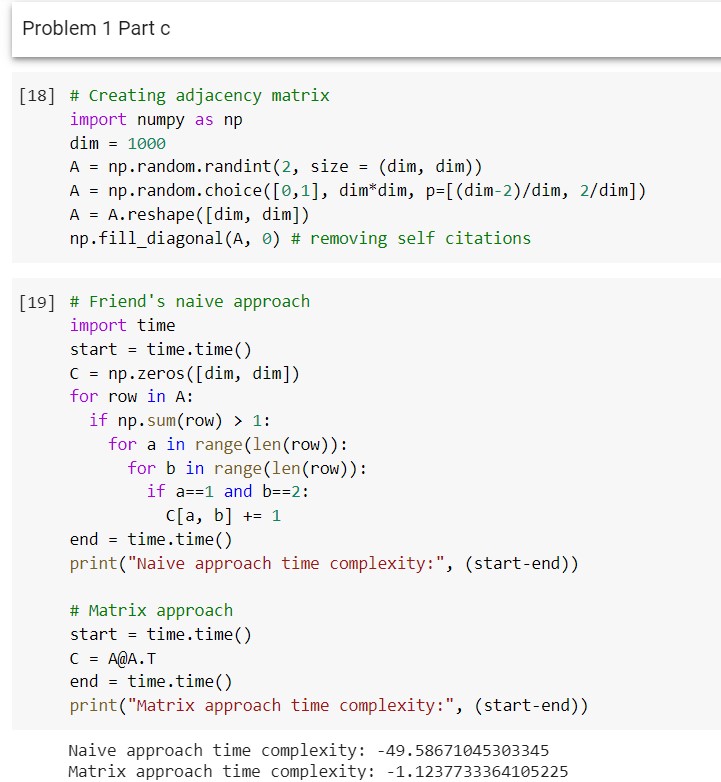
Problem 1

Part C

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



We see that for an adjacency matrix having 1000 papers, the time complexity of the naïve approach is almost 50 times greater compared to matrix approach.

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?*

We could come up with the example of an overview paper, describing the problem from different perspectives. Different sections of this paper will probably cite different papers, but those should not necessarily be content-related, as initial section topics could differ significantly. On the other hand, if two papers have the same bibliography, one could quite confidently argue that their content is similar. Therefore, bibliographic coupling is a better indicator for similarity.

Problem 2

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 increase of criminal network is natural process, as each new criminal have possibility to introduce more connections. But the leveling-out of the network is caused by the influence of police activity. In Part B Q5 the two centrality measures were compared: betweenness centrality and eigenvector centrality. The first one is highly dependent on number of nodes, as more nodes in the network tend to create more paths through a particular node. This can be accounted for by normalizing the centrality metrics according to the number of nodes at each stage.

Regarding eigenvector centrality need not to be adjusted, as it does not depend on the size of the network.

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.*

The degree centrality could simply give us the information on how many people this person is interacting with. It already gives some information but completely ignores the importance of each other person. The simple drug dealer acting in a group of 5 other dealers would have the same degree of centrality as the chief of the mafia, interacting with his top 5 consiglieres. But overall, it will tend to mark local group leaders, for example those responsible for cocaine production in specific location.

The betweenness centrality measure would favoritize the “middle” criminal executers, or “connectors” – those how actually connect different parties or interacts with several groups. They would have the maximum number of shortest paths in the network. This metric is best suited for identifying the “deliverers” and “managers” of the criminal network.

The eigenvector centrality highlights those who have important connections. In case of directed network, it would help to identify the boss of the Mafia and other most important criminals in its head. For undirected network, it will tend to identify those group leaders, that are closer to “deliverers” and “working horses”, as it will receive the “importance” not only from its own group but also from the different part of the network.

Regarding limitations, it is for example hard to identify the boss of the mafia if the network data is undirected – the head of mafia usually interacts with a very few people, which make him undistinguishable in the undirected graph.

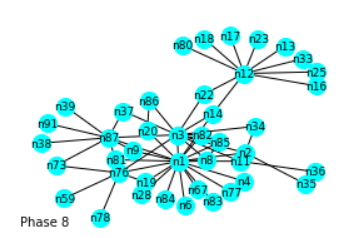
The best choice of centrality measure is the betweenness centrality, as it helps to identify those who connects different parts of the traffic and actually makes it possible.

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.*

Node 1 is pointed out by all the centrality metrics, so we definitely choose this one. Nodes 3 and 12 have shown the highest temporal consistency of betweenness degree measures. Looking on the network graph, those are probably the “managers” – working horses of the traffic who interacts with a lot of people from different groups, executing control or lead functions.

Phase 8, for example, gives quite general representation of the network:

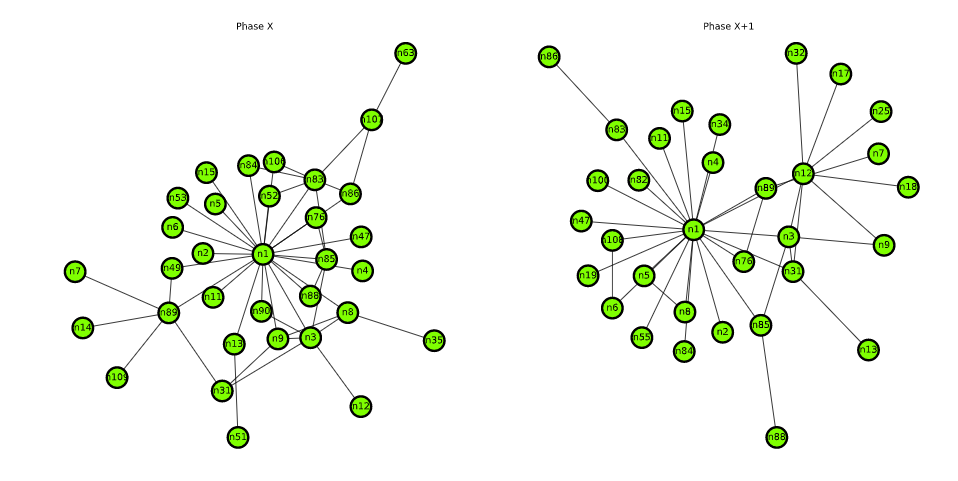


Nodes 1, 3 and 12 are “top managers”, controlling different groups, and node 85 is just quite close to nodes 1 and 3 and therefore marked as “important” by eigenvector centrality measure.

Part f

*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.*

The phase X is the Phase 4 where the seizure of 300kg of marijuana took place



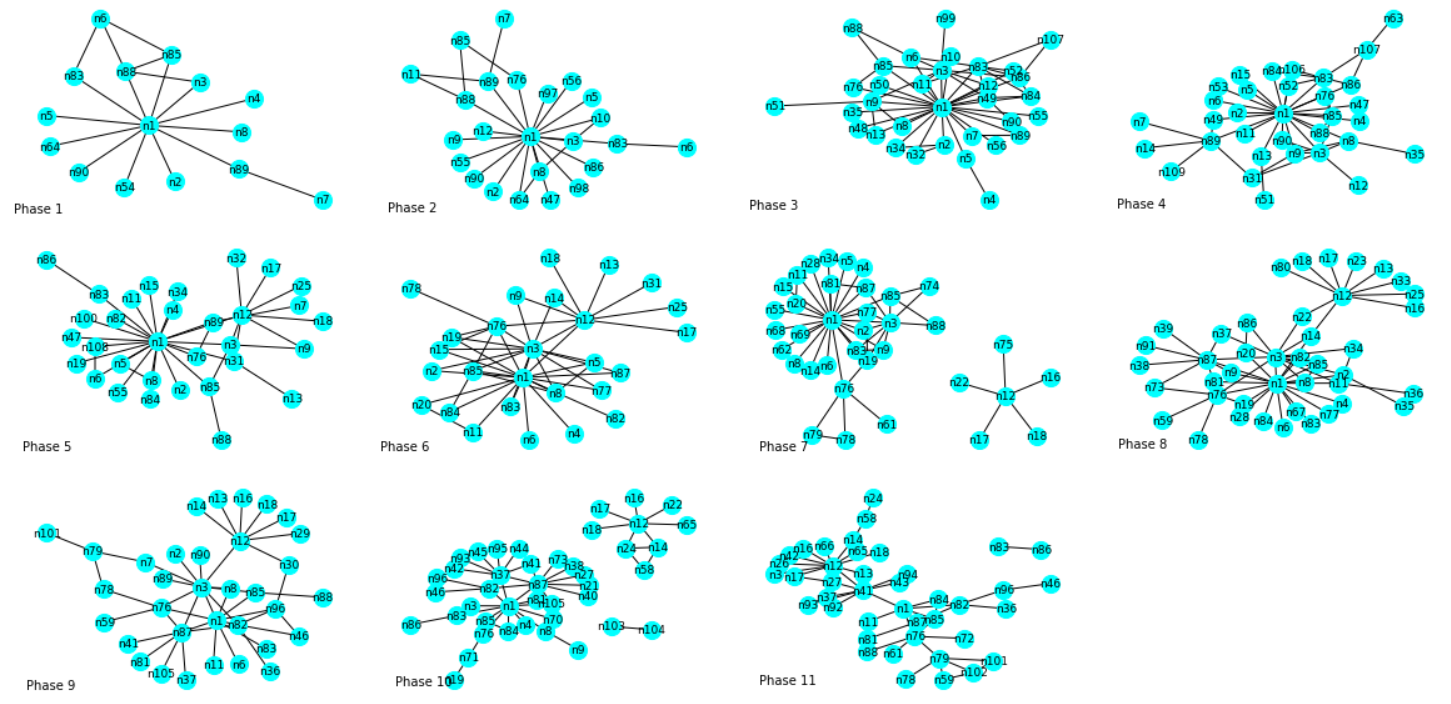
The network before the seizure was denser, more centralized towards node 1. After the seizure, the criminals needed to find new drug traffic opportunities. They were lucky to have Ernesto Morales (n12) which had his contacts in Colombia and helped organize the new cocaine traffic. Wee see node n12 getting edges with previously unseen nodes.

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?*

The all 11 phases are depicted in the figure below



We see that it has started from quite small network, focusing on marijuana drug traffic. Up to phase 4 it has quite grown but the seizure of 300kg of marijuana, after which the criminal net started to explore the cocaine traffic from Colombia, without renouncing to marijuana traffic. We then the cocaine seizures at phases 6 and 9 resulted in temporary loss of connection between Ernesto Morales and the network of Daniel Serero, as wee see node n12 disconnected from the main network at phases 7 and 10.

We also observe that after cocaine seizures, the network around Daniel Serero becomes denser, pointing that he tries to focus more on marijuana traffic.

After the huge seizure of marijuana in Phase 10, the network focused on cocaine traffic, as in Phase 11 we see the highest centrality of the node 12 (Ernesto Morales).

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.

It is definitely node n41, which plays the most important role in cocaine traffic at Phase 11. This node has the highest betweenness centrality: 0.55 and second highest eigenvector centrality: 0.49

Also, the node 14 which had the directed connections to Pierre Perlini (n3) and Ernesto Morales (n12) at phases 6 and 8.

Part i

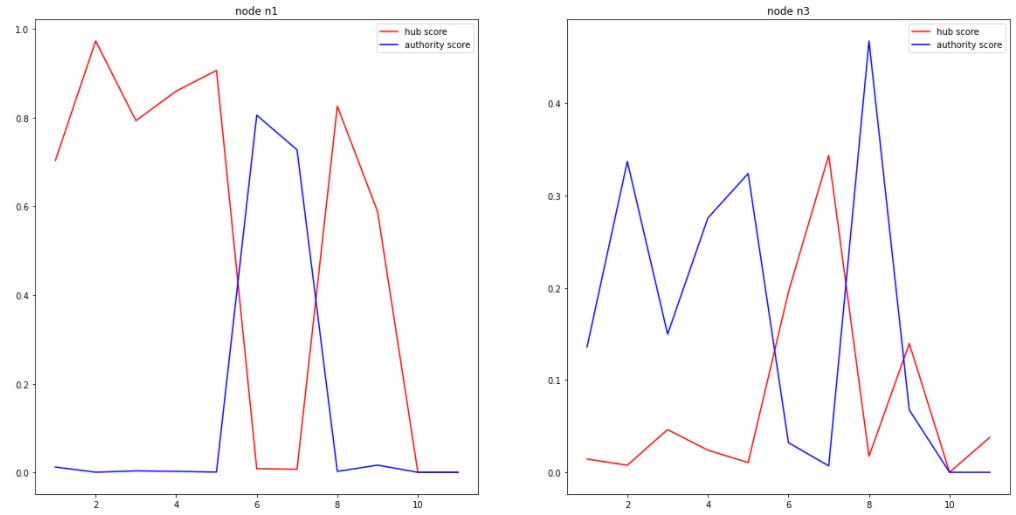
*What are the advantages of looking at the directed version vs. undirected version of the criminal network?*

We can observe the hierarchy in the network – who issue orders and who receives them. It would also be easier indetify some hidden big-bosses who interacts only with Daniel Serero, for example, and therefore their importance is unidentifiable at the undirected graph. It will be interesting to see from whom Daniel Serero is dependent, for example. And the overall importance of each node could be determined by the ratio of in- and out-degrees, assuming that the directed edge mainly means the order from higher criminal or some useful information provided to that criminal.

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)?*



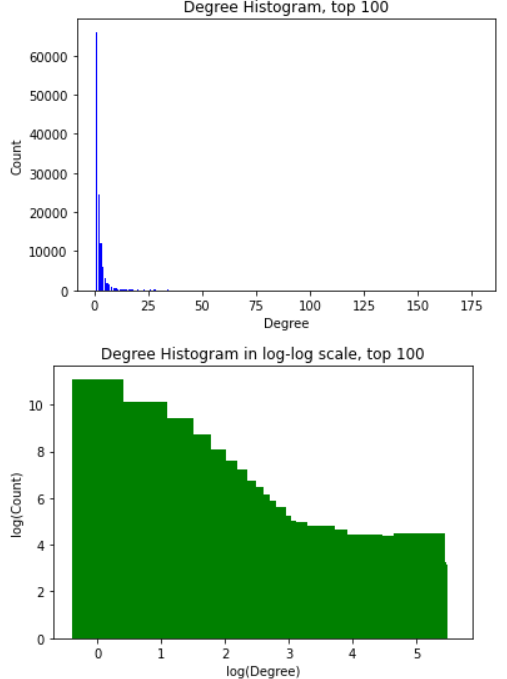
In the context of criminal network, high hub score means that this person most probably gives the instructions to a lot of other criminals, while high authority score means that he might receive the reports from others, without actually managing the situation.

We see that after the seizure in Phase 6 Daniel Serero stepping back from managing the traffic – hist hub score dramatically drops – he starts to manage through his lieutenant Pierre Perlini whose hub score rises. In other words, in post-crisis times, Perlini executes the orders and Serero receives the reports. But seizures in phases 9 and 10 make them stepping back both, as all their scores become very low.

Problem 3

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?*



Yes, it could have come from Erdos-Renyi Model as the log-log shape looks like binomial with low degree. It is related to the fact that each node in Erdos-Renyi model have binomial number of edges.

This is the end, run out of time, unfortunately…

Thanks for reading :)