41/42 Excellent report!

Written Report – 6.419x Module 3

Name: (Inés da Rosa_inesdarosa)

Problem 1: Suggesting Similar Papers

Part (c): *(2 points)* ***Include your answer to this question in your written report.*** *(100 word limit.)* +2

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

Both solutions have a time complexity = $O(n^3)$, so it does not seem to have a huge difference between them.

However, the algorithm of my friend says 'if the row sum is strictly greater than 1, then do: for each pair ((r, a),(r, b)) in row r that are non-zero (meaning that there is an existing relationship), add 1 to C at the location (a, b).' If this condition is not accomplished, you go to the next row. Honestly, I am not sure, but at first glance the friend's algorithm could be a little faster.

Part (d): (3 points) Include your answer to this question in your written report. (200 word limit.) +3

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?

Why?

In the co-citation, the matrix is built by the papers that are being cited. For example, the numbers outside the diagonal indicate the number of times that both papers were cited by a third paper. The numbers on the diagonal show the number of times that a paper was cited by other papers.

In the bibliographic coupling the matrix is built by the citation of the papers. For example, the numbers outside the diagonal indicate the number of times that both papers cited the same third paper. The numbers on the diagonal show the number of times that such paper cited other papers.

Which measure is more appropriate as an indicator for similarity between papers?

The bibliographic coupling is a static measure, it compares the list of references. This measure can be computed right after the publication.

Co-citation is a dynamic measure, it will change with time. It also gives some indications of the similarity but more important is to know what happens with the diagonal with time. As the number of the diagonal of the paper i indicates the number of papers that are citing to i, with time such number could increase indicating that such paper is gaining popularity. Bibliographic coupling doesn't capture this thing.

For all of the reasons mentioned, bibliographic coupling is more appropriate as an indicator for similarity between papers.

Part (c): ***(2 points)*** ***Include your answer to this question in your written report.*** *(100 words, 200 word limit.)*

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?

I think that such increase could be the outcome of the process of the information accumulation. It is like an accumulation curve; probably it is only a sampling effect or something like that.

Maybe I would analyze the network after the number of nodes levels out, avoiding the lack of nodes because of sampling problems. For example, at first glance I wouldn't consider the first two phases for the analysis. However, re-doing the computation of the measures (between and eigenvector centralities) the three high mean values corresponded to the same individuals.

Part (d): ***(5 points)*** ***Include your answer to this question in your written report.*** *(300 words, 400 word limit.)*

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.

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?

The degree-centrality shows me the number of connections a member of the network has. This can be important because it shows me the high level of relationship a member could have, so, the potential knowledge of the organization that such individual could have.

The between-centrality in a way informs me that the information passes through an individual. So, nothing happens without his knowledge. The person with a high value of that measure shows me the central importance that such a person has. At the same, this measure also says that if that node disappeared the network could be damaged, broken.

The eigenvector-centrality gives information that is very particular because it says that a node is important if it is connected with an important node. Maybe a node doesn't have a lot of connections, but it is connected with a few really important nodes, so, the node will be important too. So, such a node could do some activities essential for the organization. This occurs because this measures also taking into account how well connected a node is, and how many links their connections have, and so on through the network (cascade effect).

In your own words, could you explain the limitations of degree centrality?

The degree-centrality considers the importance of a node based on the number of edges connected to the node, but it does not take into account the cascade effect. This means that it does not consider the fact that a node is important because it is connected with a very important node.

In your opinion, which one would be most relevant to identify who is running the illegal activities of the group? Please justify.

If I had to choose one of them, I would choose between-centrality because it measures the extent to which a node lies on paths between other nodes. At the same time this measure indicates to me that if I removed this node, I could break the network. So, I would think that this node corresponds to someone very important in that criminal network, and for that his illegal activities should be very important for the organization.

Part (e): ***(3 points)*** ***Include your answer to this question in your written report.*** *(100 words, 200 word limit)*

+3

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.

Considering those measures, it is possible to see that n1, n3, and n12 are very important for the organization, they are connected with a lot of nodes and also, they are in the line of the several connections. They are in the middle of the network, probably they know everything about the organization and are fundamentals for its functioning.
The node n85 is an important node because it is connected to important nodes. If it is observed what is the profession of that node it is possible to understand why it is so

important. He is the accountant and in this kind of organization the management of the money is very important, I can imagine that it is necessary to make a lot of magic to hide the money and things like that.

Part (f) Question 2: ***(3 points)*** ***Include your answer to this question in your written report.*** *(200 words, 300 word limit.)*

+3

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.

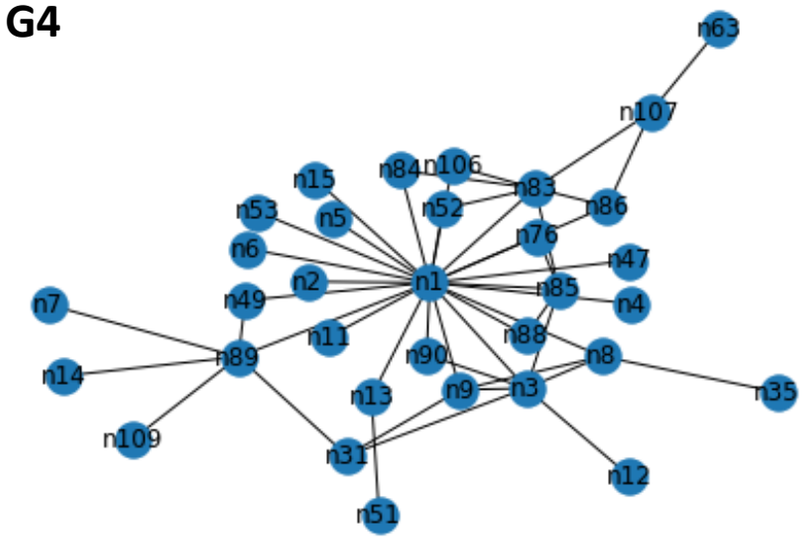
X corresponds to phase number 4, in that time was the first police intervention to take the drugs from the organization, and the traffickers reoriented to cocaine import from Colombia, transiting through the United States. Before, the cocaine came from Morocco, transiting through Spain. The most dramatic change is the position that begins to occupy the node 12 in detriment of n83, n89, n3. It is showed by the change of the centrality measures and the positions in the graphic. I think that the node 12 represents the change of the cocaine provider.

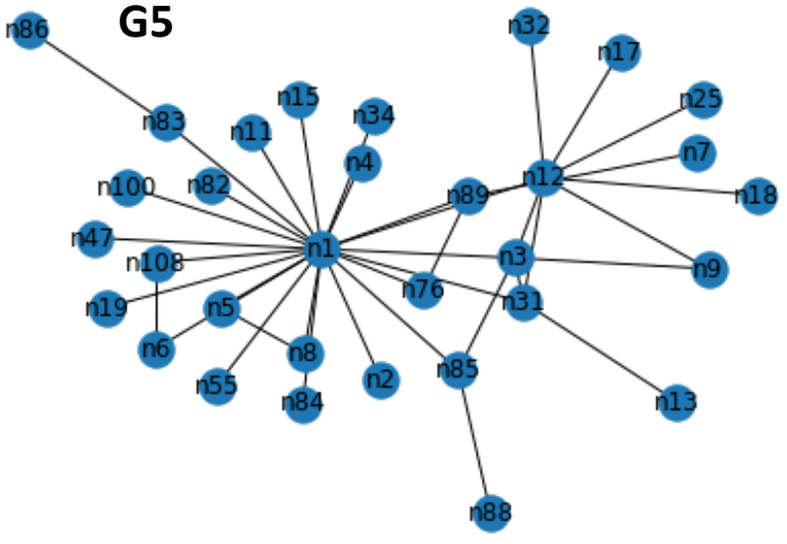
Phase 4:
- Degree-centrality: n1, n83, and n3
- Between-centrality: n1, n89, n3
- Eigenvector-centrality: n1, n3, n83

Phase 5:
- Degree-centrality: n1, n12, and n3
- Between-centrality: n1, n12, and n89
- Eigenvector-centrality: n1, n12, and n3

n12: Ernesto Morales: Principal or ganizer of the cocaine import, intermediary between the Colombians and the Serero organization.
n89: Antonio Iannacci: Investor.
n83: Alain: Investors and transporters of money.
n3: Pierre Perlini: Principal lieutenant of Serero, he executes Serero's instructions.

Basically, there was a re-organization, the investor and transporters money changed their importance, even some of them disappeared, for example n106 (Beverly Ashton Spouse of Lino (n6), transports money and documents). Those changes probably involved the less importance of nodes from Europe.





+4

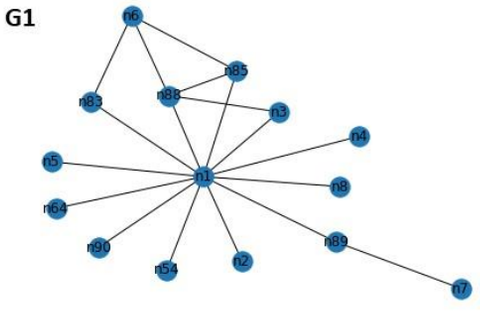
Part (g): ***(4 points)*** ***Include your answer to this question in your written report.*** *(200 words, 300 word limit.)*

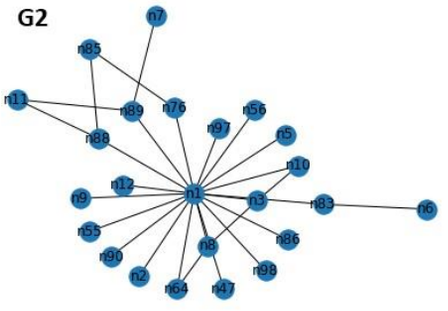
While centrality helps explain the evolution of every player's role individually, we need to explore the global trends and incidents in the story in order to understand the behavior of the criminal enterprise.

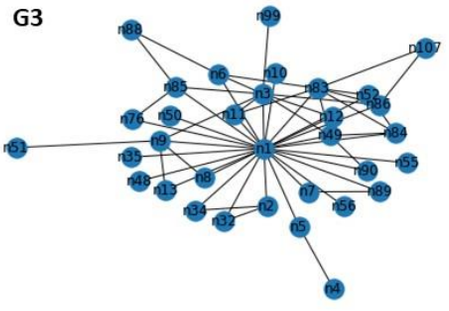
Describe the coarse pattern(s) you observe as the network evolves through the phases. Does the network evolution reflect the background story?

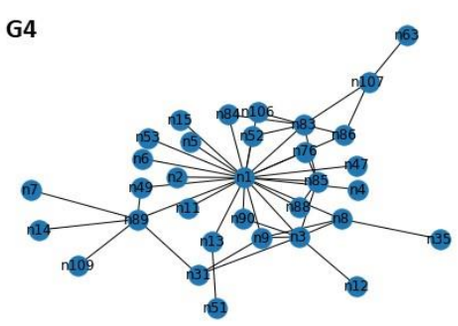
Through the centrality measures and the graphs per phase we can observe some changes. For example, n1 and n3 that at the beginning are the most important and at the end they share the importance with other nodes, and in the phase 10 and 11, n3 lost its importance. At the same time n12 goes increasing in importance with phases. If we observe the last two phases (i.e., 9 and 10) and considering the three higher values of the centrality measures we can observe nodes that at the beginning didn't appear, for example, n37 (trafficker), n41 (trafficker), and n87 (investor).

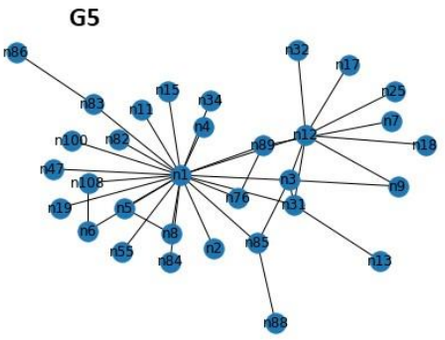
These changes indicate that through the different phases the importance of investors change too. Also, it is possible to see that after police intervention a group separated from the original group is formed, and the center is n12. But then, this new group joins with the original. This separation happened in phases 7 and 10, in the next phases, the group joined with the principal group. Through the graph and the centrality measures the importance of the node n12 is obvious. In the last phase it is observed that the organization has two principal groups, one of them centered in n1 and the second one centered in n12. Here the n41 plays an important role because it appears to link both subgroups.
In general, it appears that the network changes according to the story. After seizures the organization, nodes change its importance, some disappear, and others increase or decrease their relevance.

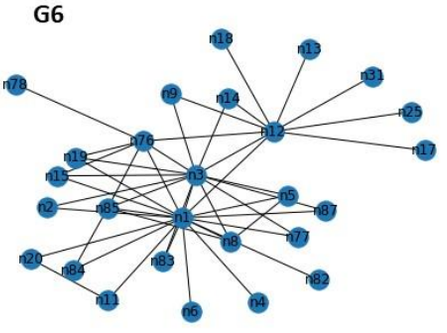


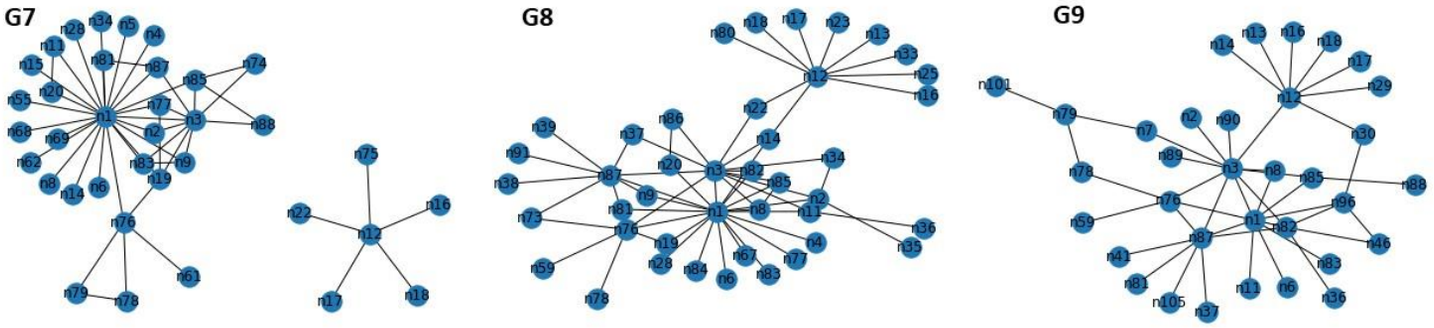


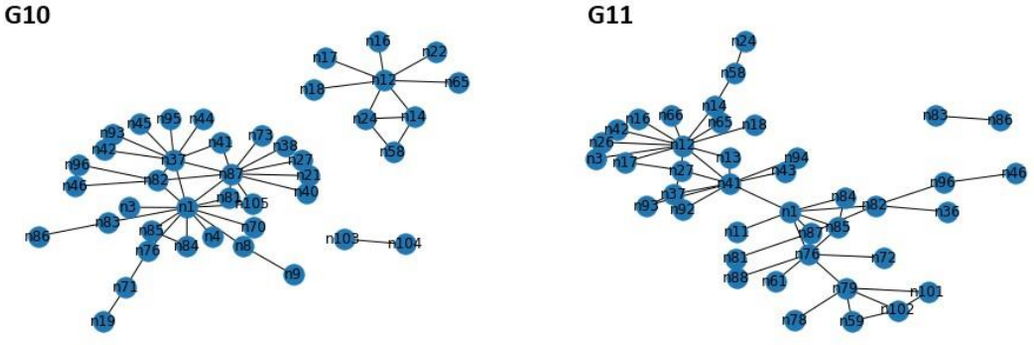












+2

Part (h): (2 points) Include your answer to this question in your written report. (50 words, 100 word limit.)

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.

n27(trafficker), n41(trafficker), and n76(trafficker) seem to be important.

n76 appears as an important trafficker after phase 5, very well connected, in the line of the important relationships and related to the important nodes.

n41 is particularly important in the last two phases; it seems fundamental in the architecture of the network in phase 11.

n27 is observed in the last two phases, I don't know, but it is probably that n27 appeared as the outcome of the police activities. It appears with high values of eigenvector-centrality, suggesting good relationships with important nodes.

The remaining two questions will concern the directed graphs derived from the CAVIAR data.

Part (i): (2 points) Include your answer to this question in your written report. (150 words, 250 word limit.)

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

Hint: If we were to study the directed version of the graph, instead of the undirected, what would you learn from comparing the in-degree and out-degree centralities of each actor? Similarly, what would you learn from the left- and right-eigenvector centralities, respectively?

Out-degree centrality measures explain the propagation of messages in a telephone communication network. And an in-degree centrality measure could represent, like in social media, the popularity of the node, and maybe the importance of receiving the message. Considering a directed graph, it could be possible to know the flow of information, and as the police have the hours of the day that communications happen, this flow could be better understood.

In a general way, the right eigenvector indicates that the importance comes from nodes i point to j. So, the importance of i is the result of pointing to an important node j; you are important if you are pointing to an important node. The importance in the left eigenvector comes from nodes pointing to i. So, if you are pointed at from an important node, then you are important. Direct-graph would permit finding nodes with a low number of connections but with a high level of importance for the criminal network. You can distinguish hierarchies using the eigenvector-centrality with a direct-graph.

Part (j): (4 points) Include your answer to this question in your written report. (300 words, 400 word limit)

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.*

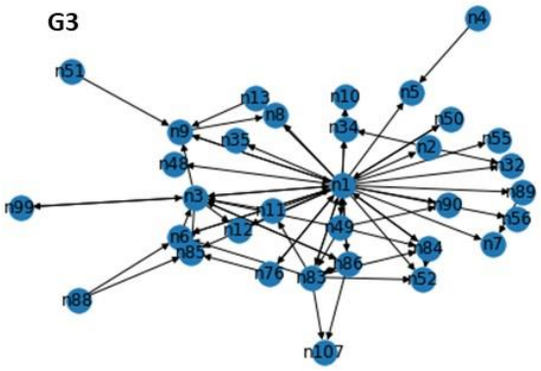
Using this, what relevant observations can you make on how the relationship between n1 and n3 evolves over the phases. Can you make comparisons to your results in Part (g)?

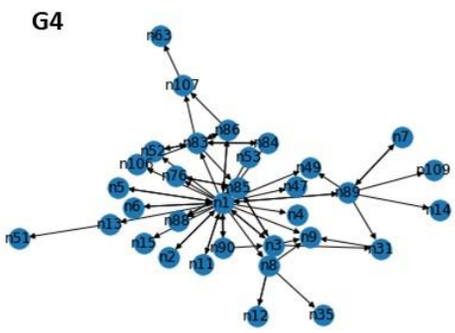
- Hub: A node is high-quality if it links to many high-quality nodes
- Authority: A node is high-quality if many high-quality nodes link to it

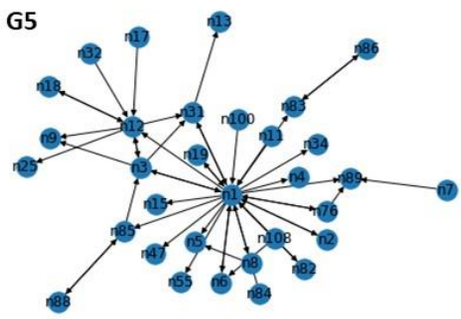
Observing the data of hub and Authority (at the end of this part there's an example of the node scores that belong to 8 and 11 phases) we can see that in general n1 presents higher values in a Hub perspective than in an Authority perspective. Conversely, n3 presents higher values in an Authority perspective than in a Hub perspective. The values tend to be higher in n1 than in n3. But in the last two phases the importance of such nodes decreases; this is detected when we observe the values of the other nodes. For example, the value of n12 in phase 11 is bigger than n1 in the hub perspective, but lower than n3, suggesting some change of the node behavior. On the other hand, n12 has a higher value than n1 and n3 in the Authority perspective.

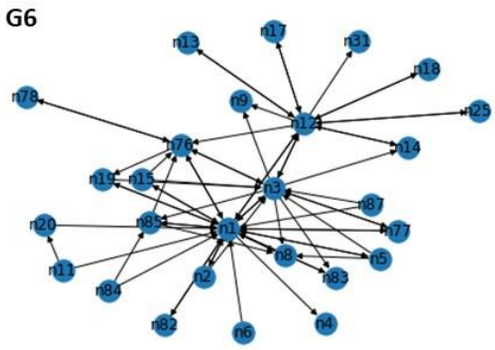
Considering the mentioned above, n1 tends to have importance because it links to high-quality nodes. On the other hand, n3 has its importance based on the important nodes linked to it. But in the last phases the importance of both nodes decreases. At the same time, n12 increases its importance because it links to many high-quality nodes, but its importance is higher through the important nodes linked to it.

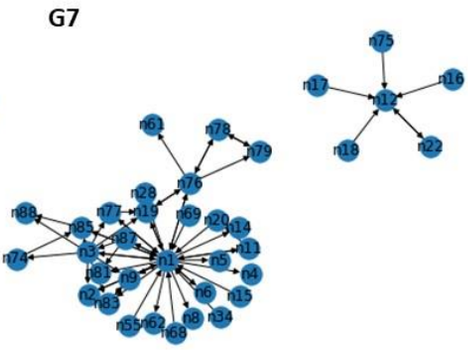
In the comparison with part g, I think that these results are consistent with the results in part g. The changes that suffered n1, n3, and n12 are well identified, and the relevance of n12 in the organization too. The analysis shows that n12 should be a target by the police, besides of n1 and n3.

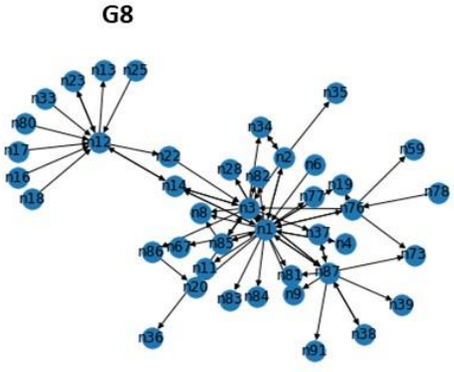


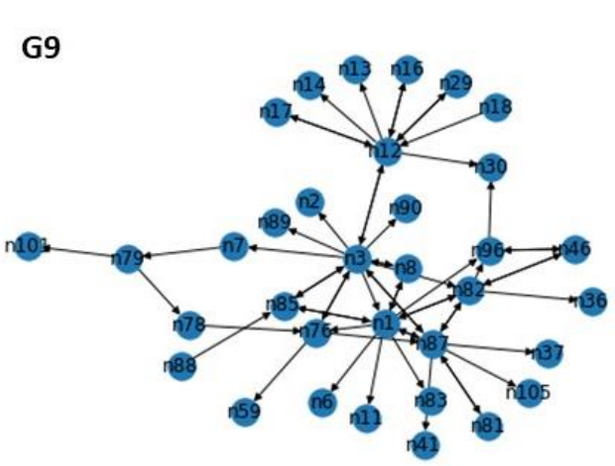


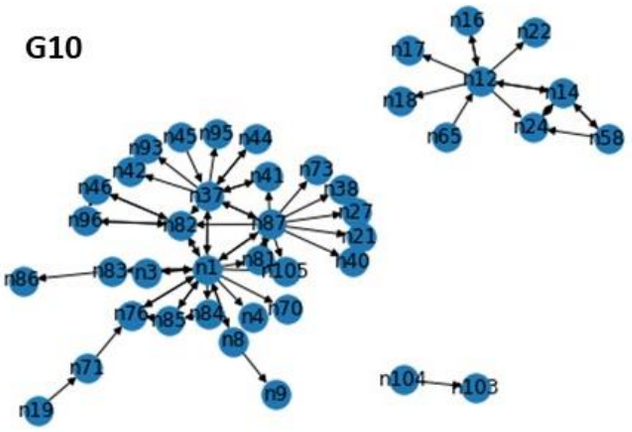


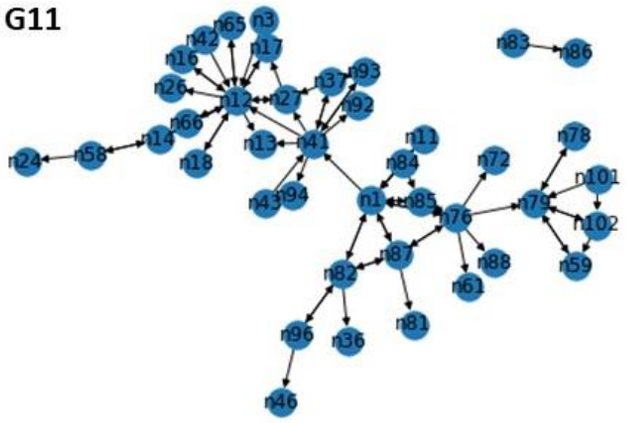












This summary is hard to read - consider plots to highlight patterns more clearly.

```
nx.algorithms.link_analysis.hits(G[8], max_iter=1000000, tol=1e-
08, nstart=None, normalized=True)
```

```
({'n1': 0.8258784632754057,
  'n11': 0.009455097861075304,
  'n12': 0.0020937948516502173,
  'n13': -0.0,
  'n14': 0.00948519817300598,
  'n16': 3.9668008278192666e-05,
  'n17': 9.52032198676624e-06,
  'n18': 9.52032198676624e-06,
  'n19': 4.1193908489932734e-05,
  'n2': 0.009532825916519714,
  'n20': -0.0,
  'n22': 0.028327252632554756,
  'n23': 7.933601655638534e-06,
  'n25': 4.76016099338312e-06,
  'n28': 4.1193908489932734e-05,
  'n3': 0.01737960174318472,
  'n33': 1.5867203311277067e-06,
  'n34': 0.0004923369395606741,
```

```
 'n35': -0.0,
 'n36': -0.0,
 'n37': 0.005434486978175028,
 'n38': 0.002717243489087514,
 'n39': -0.0,
 'n4': 4.1193908489932734e-05,
 'n59': -0.0,
 'n6': 4.1193908489932734e-05,
 'n67': -0.0,
 'n73': -0.0,
 'n76': 0.010622930042372726,
 'n77': -0.0,
 'n78': 0.00012308423489016852,
 'n8': -0.0,
 'n80': 1.5867203311277067e-06,
 'n81': -0.0,
 'n82': -0.0,
 'n83': -0.0,
 'n84': -0.0,
 'n85': 0.06640968297952266,
 'n86': 0.000738505409341011,
 'n87': 0.011070143984131024,
 'n9': -0.0,
 'n91': -0.0},
{'n1': 0.002038101507186539,
 'n11': 0.04262148338215955,
 'n12': 7.85042550440492e-05,
 'n13': 1.5436488269945664e-05,
 'n14': 0.00646389814799882,
 'n16': -0.0,
 'n17': -0.0,
 'n18': -0.0,
 'n19': 0.01850130245888137,
 'n2': 0.02435876311196457,
 'n20': 0.03653814466794686,
 'n22': 4.6309464809836996e-05,
 'n23': 1.5436488269945664e-05,
 'n25': -0.0,
 'n28': 0.01826635002092552,
 'n3': 0.46717114577731894,
 'n33': -0.0,
 'n34': 0.0003265426848246005,
 'n35': 7.028069408223628e-05,
 'n36': 0.0002091229335700841,
 'n37': 0.0014796372025625425,
 'n38': 8.161456105547586e-05,
 'n39': 0.0006529164884438069,
 'n4': 0.012177566680617013,
 'n59': 0.00023495243795584772,
 'n6': -0.0,
 'n67': 0.01826635002092552,
 'n73': 0.00015993204037409176,
 'n76': 0.006089690777991144,
 'n77': 0.03653270004185104,
 'n78': -0.0,
 'n8': 0.013435957587881683,
 'n80': -0.0,
 'n81': 0.012504024924838918,
```

```
  'n82': 0.04883839771783924,
  'n83': 0.012177566680617013,
  'n84': 0.006088783340308507,
  'n85': 0.07357792406518682,
  'n86': 0.0001281309953711821,
  'n87': 0.13443779076839704,
  'n9': 0.006170397901363982,
  'n91': 0.00024484368316642763})
```

```
nx.algorithms.link_analysis.hits(G[11], max_iter=1000000, tol=1e-
08, nstart=None, normalized=True)
```

```
({'n1': 7.931250529069014e-05,
  'n101': 7.946126964705447e-08,
  'n102': 1.3176849181666843e-08,
  'n11': 1.3687921998480012e-08,
  'n12': 0.0012870619604367637,
  'n13': 0.0,
  'n14': 0.0637970505809652,
  'n16': 0.4168121070476854,
  'n17': 0.13893736901589512,
  'n18': 0.02526133982107184,
  'n24': 0.0,
  'n26': 0.0,
  'n27': 0.025278904331072084,
  'n3': 0.03789200973160776,
  'n36': 0.0,
  'n37': 0.0027918864440531514,
  'n41': 0.03472465793379243,
  'n42': 0.02526133982107184,
  'n43': 3.720253085455867e-05,
  'n46': 0.0,
  'n58': 0.0003126125747054468,
  'n59': 3.9504983746763516e-08,
  'n61': 0.0,
  'n65': 0.22735205838964656,
  'n66': 0.0,
  'n72': 0.0,
  'n76': 2.304343731365323e-06,
  'n78': 6.58416395779392e-09,
  'n79': 1.6790010395526544e-10,
  'n81': 0.0,
  'n82': 4.879318614913173e-07,
  'n83': 2.4055856970142477e-20,
  'n84': 3.227037970850096e-07,
  'n85': 1.0560232148232396e-06,
  'n86': 0.0,
  'n87': 7.914529871075482e-07,
  'n88': 0.0,
  'n92': 0.0,
  'n93': 0.00016869604888451945,
  'n94': 0.0,
  'n96': 1.2762242862412885e-06},
 {'n1': 9.824171226409951e-07,
  'n101': 0.0,
  'n102': 1.0797036081867121e-08,
```

```
'n11': 3.589745741626812e-06,
'n12': 0.9065354399239576,
'n13': 0.01891823407998588,
'n14': 0.0018591384287317736,
'n16': 0.00029126713213478364,
'n17': 0.0012606497448592552,
'n18': 0.0001747602792808702,
'n24': 4.244726558750716e-05,
'n26': 0.00011650685285391347,
'n27': 0.006908618887512601,
'n3': 0.0,
'n36': 8.833670380540481e-08,
'n37': 0.009437625641297558,
'n41': 0.002670120659743453,
'n42': 0.0,
'n43': 0.0,
'n46': 1.1552580970929934e-07,
'n58': 0.011550016724956836,
'n59': 6.115930309599817e-10,
'n61': 3.128891742816089e-07,
'n65': 5.825342642695673e-05,
'n66': 0.00023301370570782693,
'n72': 2.0859278285440592e-07,
'n76': 1.821155799663058e-05,
'n78': 1.5198578365007743e-11,
'n79': 4.725622640990165e-07,
'n81': 7.164355683152889e-08,
'n82': 1.827336501082307e-05,
'n83': 0.0,
'n84': 0.0,
'n85': 2.217885862343526e-05,
'n86': 5.755167655566476e-19,
'n87': 1.1044082535588896e-05,
'n88': 1.0429639142720296e-07,
'n92': 0.01728831559909568,
'n93': 0.00686316538481401,
'n94': 0.015716650544632435,
'n96': 1.1042087975675601e-07})
```

11/12

Problem 3 Project - I continue with CAVIAR

- Introduction

The intervention by the police on the criminal network produced several changes through the time. Those changes were detected through different centrality measures. It was possible to detect that important people changed their importance through the phases in the network. However, it was very difficult for me to evaluate how the composition of the criminal network, which interacted by telephone communication, changed through the time. It is true that by the visualization with the graphs it was allowed to see that one or two nodes appeared or disappeared. But we don't manage measures that globally informs about such modification.

In this sense, I want to know how the whole network change through time, considering the important nodes and unimportant nodes. Since that the intervention of the police

produce some changes, I work the data with the hypothesis that the composition of the network will be different when comparing the firsts phases with the last phases. For do that I calculated a similarity index with the aim to compare the composition of the nodes, so considering the participants in each phase.

- Method

I calculated the Jaccard coefficient index that measures similarity between finite sample sets, and it is defined as the size of intersection divided by the size of the union of the sample sets:

$J(A,B)=A\cap B/A\cup B$

Jaccard only counts as matches when the attribute is present in both sets. In this case each set corresponds to each phase, so, it is compared the composition similarity between phases. The index takes values from 0 to 1. When the value is 0, the phases are totally different, so, the nodes are not shared between the phases. If the value is 1, the nodes are the same in both phases.

- Results and Discussion

The results are showed in the table, it is a symmetric table, each column, and row, correspond to the phases, and the number under the name of the column is the number of nodes in the phase. The diagonal has value 1 in every case. In general, it is possible to see that the values that compare phase x with x + 1 present higher values (pink cells), but when the phases are more distant in the time it is obvious the fall in similarity (cells in the red box).

	PHASE 1 (n = 15)	PHASE 2 (n = 24)	PHASE 3 (n = 33)	PHASE 4 (n = 33)	PHASE 5 (n = 32)	PHASE 6 (n = 27)	PHASE 7 (n = 36)	PHASE 8 (n = 42)	PHASE 9 (n = 34)	PHASE 10 (n = 42)	PHASE 11 (n = 41)
PHASE 1	1										
PHASE 2	0,5	1									
PHASE 3	0,37	0,54	1								
PHASE 4	0,37	0,46	0,61	1							
PHASE 5	0,34	0,47	0,51	0,51	1						
PHASE 6	0,27	0,31	0,33	0,43	0,59	1					
PHASE 7	0,24	0,3	0,3	0,3	0,42	0,54	1				
PHASE 8	0,16	0,22	0,29	0,29	0,4	0,53	0,47	1			
PHASE 9	0,29	0,32	0,29	0,31	0,35	0,39	0,37	0,41	1		
PHASE 10	0,12	0,16	0,17	0,19	0,25	0,3	0,28	0,38	0,33	1	
PHASE 11	0,1	0,16	0,17	0,19	0,24	0,28	0,28	0,34	0,5	0,43	1

If we focus on the phase 3, because the number of nodes is already stabilized, and we observe the values comparing with phase 4, we can see that the similarity is relatively high (0.61). When phase 4 is compared with phase 5 the value in similarity decreases to 0.51. Remember that the first seizure happened in phase 4, and traffickers reoriented to cocaine import from Colombia, transiting through the United States. The decrease in similarity from phase 3 onwards appears to be a process of similarity decline. It is likely

that the intervention of the police affects the whole organization, and not only on the important people. Also, the result suggests that there is some facility to replace people by other people, even, to increase the network.

I would like to make a final mention. Honestly, I was surprised with those similarity index values. Before the analysis, I expected for more similarity between the phases. This network is very variable in the composition through time. I expected that this activity would be more closed, with the same people working in this business.

→ Might reflect that so many agents enter/leave the network? It is somewhat surprising.