Homework Report #3: Network Analysis

Author: Isaac Han

Email: cogitoergosum01001@gmail.com

Class: Data Analysis: Statistical Modeling and Computation in applications

Professors: Prof. Uhler, Prof. Jegelka

Table of Contents

Problem #1

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

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

A, B are n by n matrices whose matrix multiplication is defined entrywise as following;

$$(\mathrm{AB})_{\mathrm{ij}} = \sum_{k=1}^{n} A_{\mathrm{ik}} * B_{\mathrm{kj}}$$

For a naive algorithm, it can be seen that we need to calculate n^2 elements. And each element requires n numbers of additions and n numbers of multiplications. Therefore, the complexity of algorithm for a naive matrix multiplication is $O(2n^3)$ when total numbers of operators were concerned. Following conventions, weighing multiplication more and disregarding coefficient, it would be $O(n^3)$, which was found in Q1.

However, in excahnge of element-wise stability, one can achieve time-wise efficient algorithm. One algorithm is Strassen's algorithm which has complexity of $O(n^{\log_2 7})$, with norm-wise stability (https://www.osti.gov/pages/servlets/purl/1356986).

From matrix partitioning,

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \; A * B = \begin{bmatrix} A_{1,1}B_{1,1} + A_{1,2}B_{2,1} & A_{1,1}B_{1,2} + A_{1,2}B_{2,2} \\ A_{2,1}B_{1,1} + A_{2,2}B_{2,1} & A_{2,1}B_{1,2} + A_{2,2}B_{2,2} \end{bmatrix}$$

A matrix multiplication can be subdivided and combinations of time-wise efficient algorithms may be applied, for desired time-wise efficiency.

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

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?

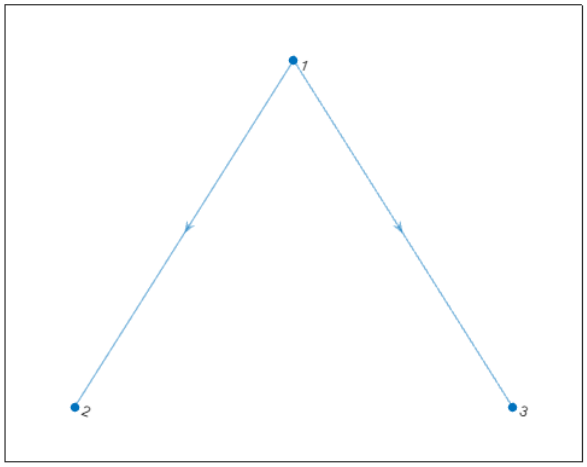
Bibliographic coupling is defeind as two papers citing the same other papers whereas co-citation between two papers is when a third paper cites them. They share common features as they both measures the relateness, but they are mathematically distinct objects.

We solved that Bibliographic coupling is represented as $\mathrm{A}\mathrm{A}^T$, whereas co-citation is A^TA in Q2. To examine how they can be different,

imagine a very simple matrix $A = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$,

$\mathrm{A}\mathrm{A}^T = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ whereas $A^TA = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$.

```
% graph = [0 1 1; 0 0 0; 0 0 0];
% plot(digraph(graph))
```



Notice that the first matrix is effectively a null matrix, considering the diagonal entries are omittable in our application,whereas the second one is a rank 2 matrix. Only when $A^TA = \mathrm{A}\mathrm{A}^T$, two methods are equivalent and different representation of the same object (different basis), as they are similar matrices.

Throughout the illustration above, it was shown that they are mathematically distinct objects. One extreme examples would be papers by a Catholic bishiop and muslim prophets cite manuscripts as common, but there is no other paper referring them in common.

To answer which measurement is more appropriate for measuring similarities(relatedness) between papers, both has their cons and pros. The bibliographic coupling is more suitable measurement for time-independent similarities, whereas co-citation would be proper for time-dependent similarities. For example, imagine there are papers concerning humans and chimps. We consider them very different since they do not have common citation. However, a surprising paper came in and it stated they share 98.8 percent of their DNA. Then a sudden surge of co-citation of papers happened. Are the original papers related? Even after a wave of co-citation, one can argue that the papers are not similar at all, whereas the other side would argue that new information brought similarities between two papers. What methods to use and what criterion to judge is dependent on applications.

Data Explanation

Properties

- A time-varying criminal network that is repeatedly disturbed by police forces, whose data set is given in the `CAVIAR` directory.
- During criminal investigation, 11 wiretap warrants, valid for a period of about two months each, were obtained. (the 11 matrices contained in `phase1.csv`, `phase2.csv`, correspond to these eleven, two month wiretap phases).

- The project mandated to seize drugs without arresting the perpetrators, which allows to observe perturbations on criminal networks.
- The network is consisted of 110 (numbered) players, where 1-82 are the traffickers and 83-110 are the non-traffickers (financial investors; accountants; owners of various importation businesses, etc.).
- Initially, the investigation targeted Daniel Serero, the alleged mastermind of a drug network in downtown Montréal, who attempted to import marijuana to Canada from Morocco, transiting through Spain. After the first seizure, happening in Phase 4, traffickers reoriented to cocaine import from Colombia, transiting through the United States.

Notable Players(Vertices)

- Daniel Serero (n1) : Mastermind of the network.
- Pierre Perlini (n3) : Principal lieutenant of Serero, he executes Serero's instructions.
- Alain (n83) and Gérard (n86) Levy : Investors and transporters of money.
- Wallace Lee (n85) : Takes care of financial affairs (accountant).
- Gaspard Lino (n6): Broker in Spain.
- Samir Rabbat (n11): Provider in Morocco.
- Lee Gilbert (n88): Trusted man of Wallace Lee (became an informer after the arrest).
- Beverly Ashton (n106): Spouse of Lino, transports money and documents.
- Antonio Iannacci (n89): Investor.
- Mohammed Echouafni (n84): Moroccan investor.
- Richard Gleeson (n5), Bruno de Quinzio (n8) and Gabrielle Casale (n76) : Charged with recuperating the marijuana.
- Roderik Janouska (n77): Individual with airport contacts.
- Patrick Lee (n87): Investor.
- Salvatore Panetta (n82): Transport arrangements manager.
- Steve Cunha (n96): Transport manager, owner of a legitimate import company (became an informer after the arrest).
- Ernesto Morales (n12): Principal organizer of the cocaine import, intermediary between the Colombians and the Serero organization.
- Oscar Nieri (n17): The handyman of Morales.
- Richard Brebner (n80): Was transporting the cocaine from the US to Montréal.
- Ricardo Negrinotti (n33): Was taking possession of the cocaine in the US to hand it to Brebner.
- Johnny Pacheco (n16): Cocaine provider.

Data Import

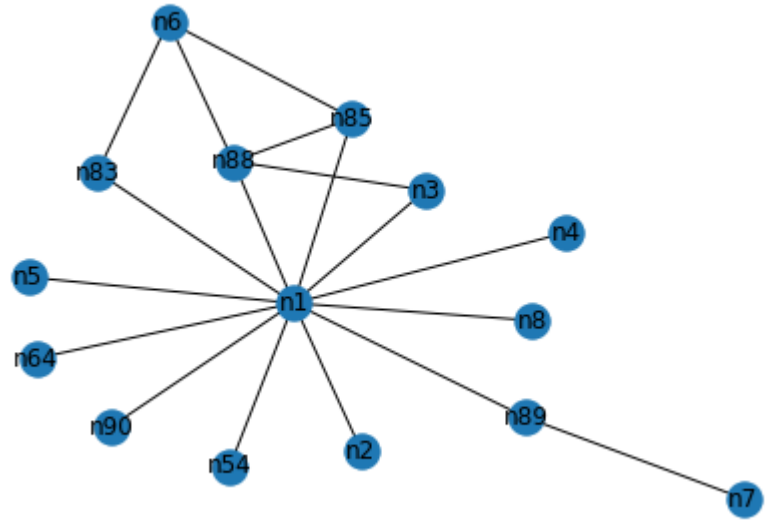
From Internet

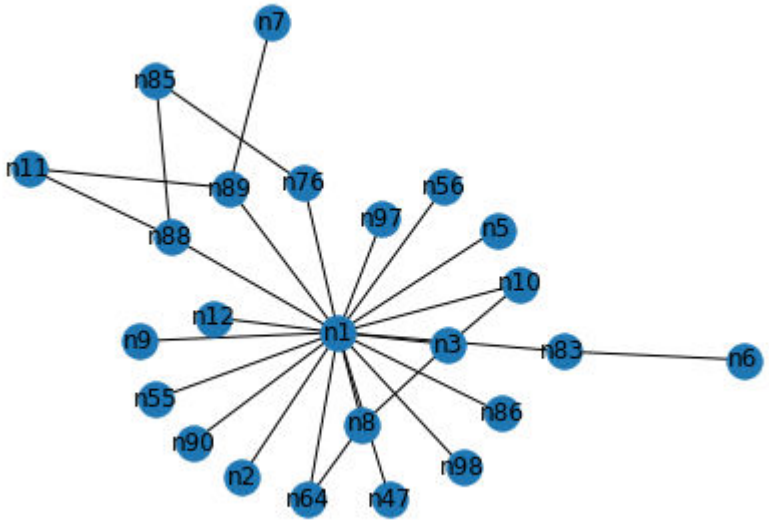
```
import pandas as pd
import networkx as nx
phases = {} G = {}
for i in range(1,12):
var_name = "phase" + str(i)
```

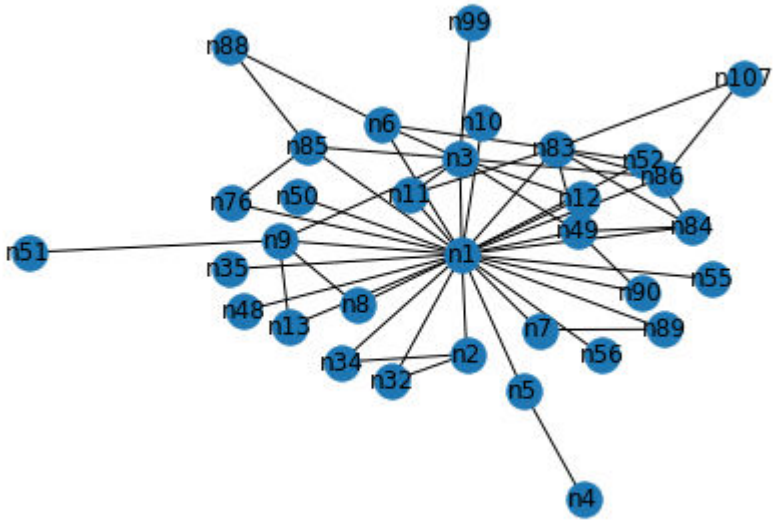
```
    file_name = "https://raw.githubusercontent.com/ragini30/Networks-Homework/main/" + var_name + ".csv"
    phases[i] = pd.read_csv(file_name, index_col = ["players"])
    phases[i].columns = "n" + phases[i].columns
    phases[i].index = phases[i].columns phases[i][phases[i] > 0] = 1
    G[i] = nx.from_pandas_adjacency(phases[i])
    G[i].name = var_name
```

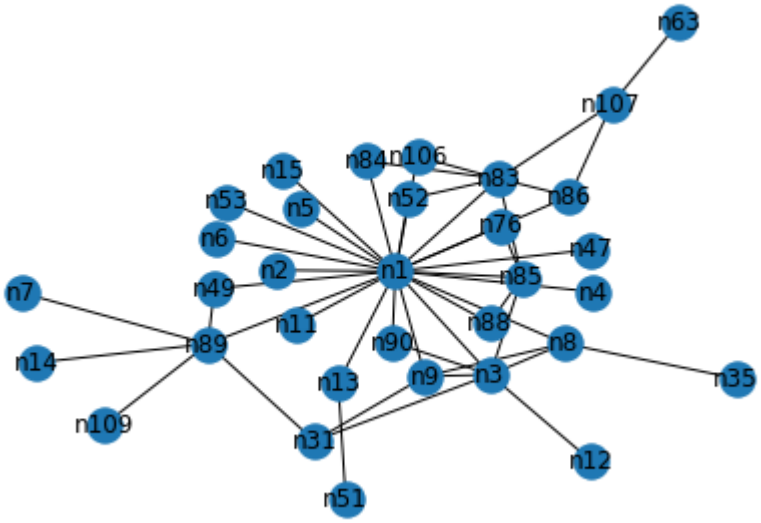
One may obtain dataset using given Python code.

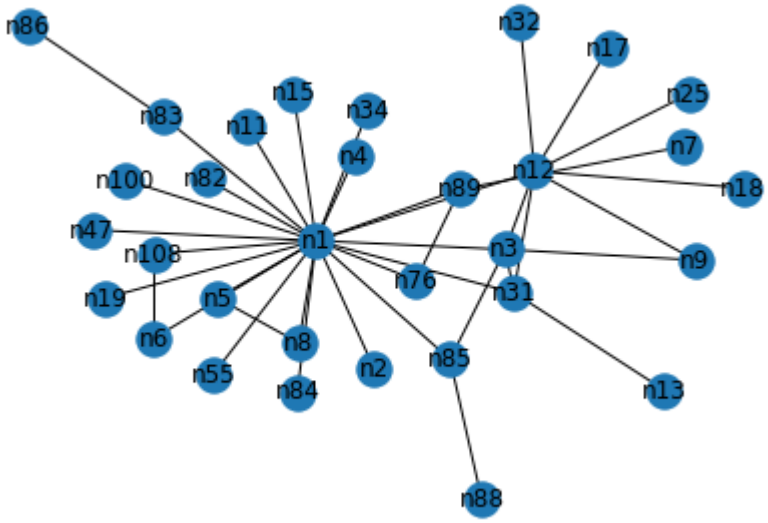
Graph from Phase1 to Phase11

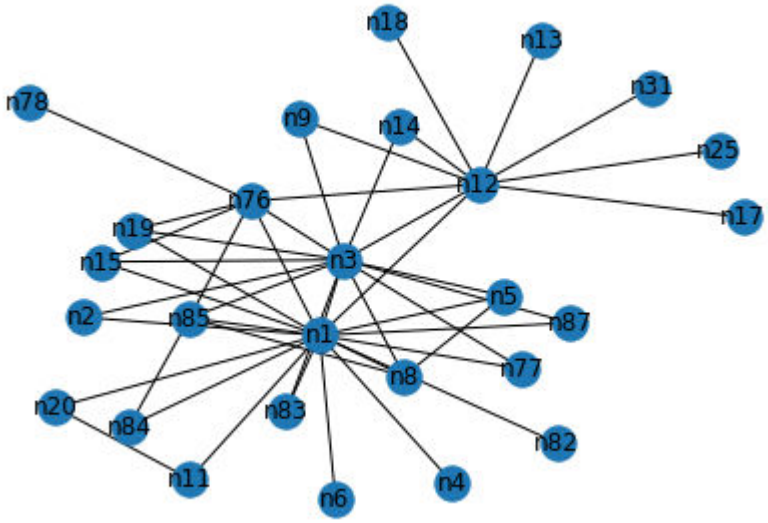


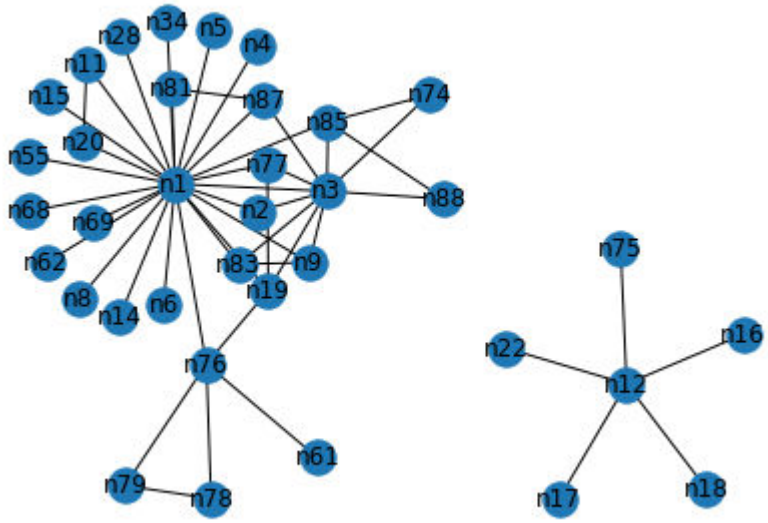


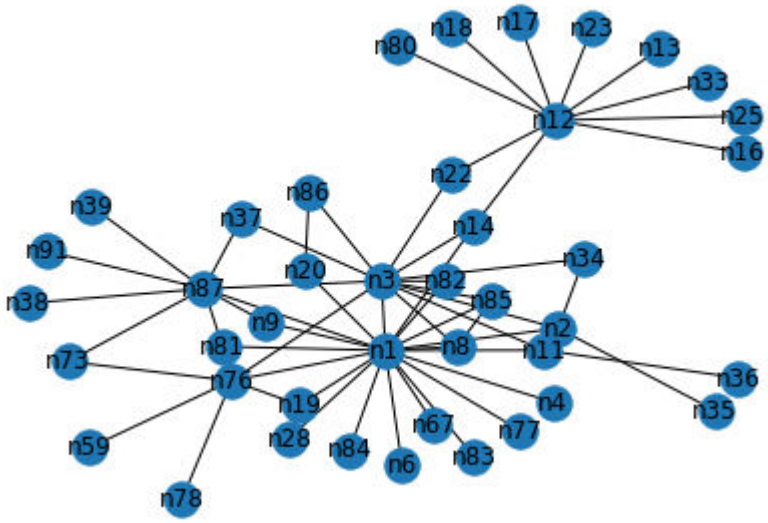

n99
n88
n107
n6
n10
n85
n3
n83
n52
n86
n76
n50
n11
n12
n49
n84
n51
n9
n1
n35
n55
n90
n48
n8
n13
n7
n89
n34
n2
n56
n32
n5
n4

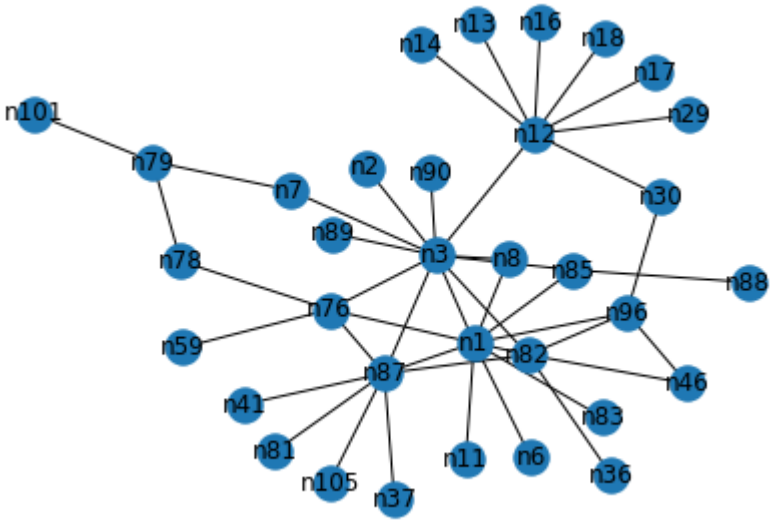

n63
n107
n84
n106
n15
n83
n53
n5
n52
n86
n6
n76
n47
n2
n1
n85
n7
n49
n4
n11
n88
n89
n90
n8
n14
n13
n9
n3
n35
n109
n31
n12
n51

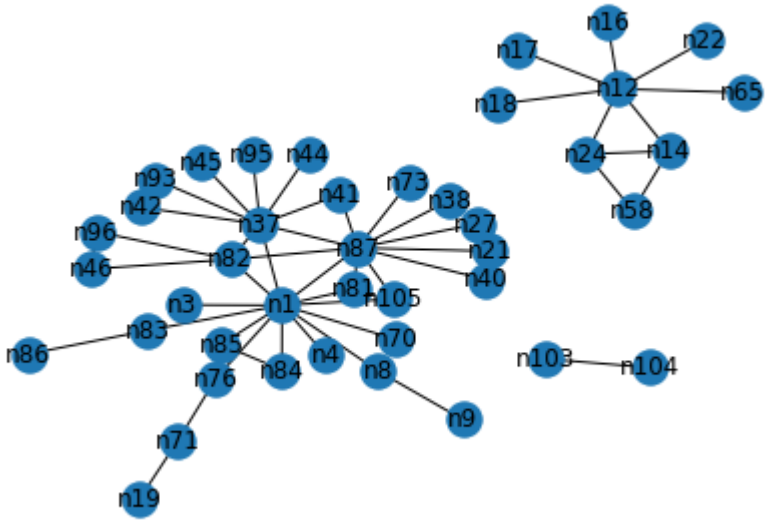

n86
n32
n17
n15
n25
n83
n34
n11
n7
n4
n12
n100
n82
n89
n18
n47
n1
n3
n108
n9
n76
n31
n19
n5
n6
n8
n85
n2
n55
n84
n13
n88

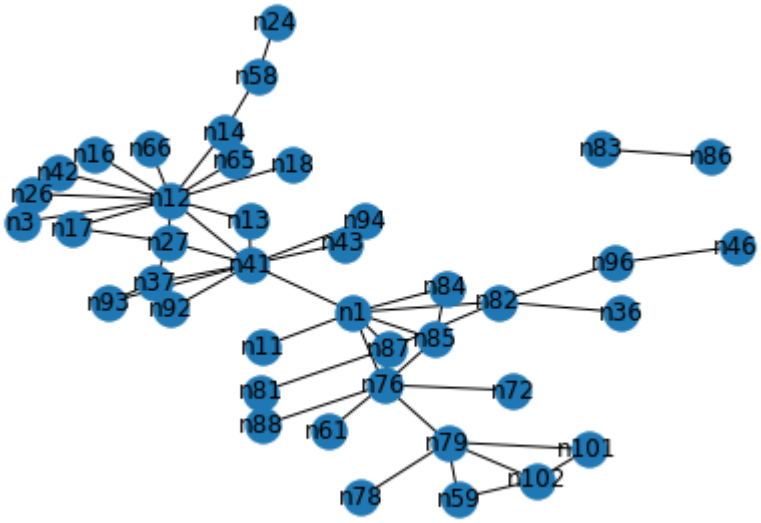

n18
n13
n78
n31
n9
n14
n25
n12
n76
n17
n19
n3
n15
n5
n2
n85
n87
n1
n77
n8
n20
n83
n84
n82
n11
n4
n6


n28
n34
n5
n4
n11
n81
n15
n87
n74
n85
n20
n55
n77
n1
n3
n88
n68
n69
n2
n62
n83
n9
n19
n8
n14
n6
n76
n61
n79
n78
n75
n16
n22
n12
n17
n18


n80
n18
n17
n23
n13
n33
n12
n25
n16
n22
n86
n39
n37
n14
n91
n20
n3
n82
n34
n87
n85
n38
n9
n2
n81
n1
n8
n11
n73
n36
n76
n19
n4
n35
n67
n77
n59
n28
n84
n83
n6
n78


n14
n13
n16
n18
n17
n29
n101
n12
n79
n2
n90
n7
n30
n89
n3
n8
n85
n88
n78
n76
n96
n59
n1
n82
n87
n46
n41
n83
n81
n11
n6
n36
n105
n37


n16
n17
n22
n12
n18
n65
n24
n14
n58
n45
n95
n44
n93
n73
n41
n38
n42
n37
n27
n96
n87
n21
n46
n82
n40
n81
n105
n3
n1
n83
n70
n86
n85
n4
n8
n84
n76
n9
n71
n19
n103
n104



Problem #2

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?

First, the criminal network was investigated from phase1 and it can be assumed that it is already well established. The investigation mandated that no one to be arrested until the end of investigation and the first seizure took at phase 4. Furthermore, it is hard to imagine that a criminal network grows so rapidly in 2-month interval. From this, it can be inferred that increased number of nodes simply corresponds to the discovery of criminal network that was once hidden from investigators.

Part (b) Q5 concerns the temporal consistency of a player's centrality. Since some players might have hidden, as it was not discovered yet in early phases, the centrality measures we performed might not accurate represent reality, as it underestimate individuals discovered at later stage.

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 (including degree, betweenness, and eigenvector centrality) teach you about the importance of an actor's role in the traffic? In your own words, could you explain the limitations of degree centrality? In your opinion, which one would be most relevant to identify who is running the illegal activities of the group? Please justify.

Definition:

The research paper posted by the course has the meaning of measurement in page 9 and relevant cases in page 4. According to the paper;

1. Degree centrality is the number of neighbours of an agent. In-degree and out-degree can be devised if the direction of graph is believed to be important.
2. Betweenness centrality measures across all agent pairs that have the shortest path containing the player, the percentage that pass through the player
3. Eigenvector Centrality measures the degree of connected agents who are themselves connected to many players.

Importance:

Degree centrality: The one giving out or receiving orders from many others would have high degree centrality. Below are list of some important actors according to the measurement.

- n1: Mastermind of the network
- n12: Principal Organizer of the cocaine import, intermediary between the colombians and the Serero organization.
- n41: no info
- n76: Charged with recuperating the marijuana
- n79: no info

N1, the mastermind of the network, is highly connected as expected. As police investigated around him, the observational bias would also contribute to the higher degree of centrality. Another interesting aspect is n12, an international drug trafficker. As diversifying traffic routes are important, to reduce the risk. Another one is n76, who took drugs from numerous drug sellers.

Betweenness centrality: It is essentially a measurement where if the player is removed, what would be the stress given to an optimized network. It does not give a relevant information in a criminal network. Criminal organizations operate in a somewhat decentralized manner. However it would be an effective measurement in a Watergate Conspiracy, where highly optimized government affairs were concerned.

- n1, n12, n41, n76, n79: Mentioned avove
- n82: Transport arrangements manager

Since only highly relevant nodes are listed, both centrality measurements referred to the same nodes. N82 was newly revealed to be extremely significant since transporting drugs from a nation to another had to go through certain individuals.

Eigenvector centrality: Measures individuals whose neighors are the most connected to the graphs.

- n1, n12, n76: Mentioned above
- n87: Investor

Eigenvector analysis brought investor as important since n1 is important.

Limitation and which methods to use:

The limitation of degree centrality in criminal investigation case, along with all other measurement, would be missing data problem. In investigation, the full graph is not given but it is constructed by investigating a link at a time. Therefore, degree centrality would hardly reveal any surprising information as the one investigated would have highest centrality.

Following the statistical significance of the research paper, I would state the out-degree centrality(degree centrality) measurement would be the most proper. The nature of criminal organization is one-directional and top-down. If individuals giving out most orders are removed, the network would be most disrupted.

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

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.

I will define important individuals based on propagating powers. I would also look how the network responded when a stress is introduced, which are introduced during phases 4, 6, 7, 8, 9, 10, and 11. Find nodes with high importance. In this case, I used degree centrality for consistancy. Find connections that are newly found and connections that got lost. Newly found connections are responses to the stress and lost connections means the edge strength was not strong enough to endure a given stress level.

With emphasis on ideas, rather than coding for days, I only compared phases4 and phases5, using degree centrality. N12 and n83 had significant absolute value changes, where N12 went from 1 edge connection to 9 edge connections and n83 went from 7 connections to only 2 connections.

On the other hand, n11, n15, and n47 had very samll changes, where n11 is negligible as it was relatively not significant. It might not be perturbed due to lack of significance.

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

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.

Actually I have used phase 4 as a way to distinguish important player, in the previous problem. Therefore, any redundant explanation would be skipped.

In the class webpage, it stated that there were 11 seizures and the first seizure was during phase4, where $2.5M worth of marijuana was confisticated. Assuming all variables are properly understood, I used the following code to analyze the behavior.

```
def diff_val(key,raid):
  return -(data[raid-1].get(key,0)-data[raid].get(key,0))

raided = [4, 6, 7, 8, 9, 10]
data = [nx.degree_centrality(G[i]) for i in range(1,12)]

result = [{} for i in range(6)]
for i in range(len(raided)):
  result[i] = {key: diff_val(key,raided[i]) for key in key_list}
i = 0
{k:v for k,v in sorted(result[i].items(), key=lambda item: item[1], reverse=True) if v !=0}
```

```
{'n12': 0.22681451612903225,   'n52': -0.0625,
 'n108': 0.06451612903225806,  'n90': -0.0625,
 'n31': 0.035282258064516125,  'n106': -0.0625,
 'n5': 0.033266129032258 06,   'n89': -0.09072580645161 29,
 'n6': 0.033266129032258 06,   'n107': -0.09375,
 'n17': 0.032258064516129 03,  'n83': -0.15423387096774194}
```

The positive means edges are added and negative means edges are lost. After searching a few players, it was found that confisticating drugs resulted in more activity in international drug import, through n12 and n6 (rest had no info). Meaning, the drug import network got bigger and dense. And confisticating resulted in reduction in investment, as investors(n83, n89, and n90) became dormant. Meaning, investment network got smaller and sparse.

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?

Hint: Look at the set of actors involved at each phase, and describe how the composition of the graph is changing. Investigate when important actors seem to change roles by their movement within the hierarchy. Correlate your observations with the information that the police provided in the setup to this homework problem.

One particular actor that came to me is n12. In the first few graphs, n12 was appears to be not important. It was simply connected to n1 or n3. But after the stress was introduced, n3 and n12 together forms a very centralized role. However, as time pass by, n3 became outcast starting from phase 10 while n12 became one of the most important nodes.

Presumably, n12 was not significant in the early phases, but it became an integral part of drug import after phase4. After continued failed operations, it sems n1 abandoned n3 and it became almost isolated in the last phase of investigation. As indicated in phase 10 graphs, the networks concerning supplying drugs seems to be evolving as they face stresses(confistications).

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.

- n41: no info
- n31: no info
- n108: no info

N41 was insignificant but became integral(connects n1 and n12) part in phase11, I suspect it might be an undercover agent.

N31 and n108: they are networks response to drug confistication, but not explanined among the 23 listed.

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?

The only advantage of undirected graphs would be the efficient data representation, which possibly reduce time and memory complexity of algorithms. Besides, anything that can be done to undirected graphs can be done

to directed graphs. Criminal networks are a special kind of networks famous for a top-down structures and many durg sellers would only know their immediate supervisor. Therefore, one would be interested in finding ones giving out orders rather than receiving from others, finding n1 who has many going out edges would be preferrable over finding n76 who bought drugs from multiple agents.

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)?

Optional: Also comment on what the hub and authority score can tell you about the actors you identified in Part (e).

Following codes are used to analyze the given task;

```
for i in range(1,12):
  hubs, authorities = nx.algorithms.link_analysis.hits(G[i], max_iter=1000000)
  hubs_sorted = {k:v for k,v in sorted(hubs.items(), key=lambda item: item[1], reverse=True)}
  authorities_sorted = {k:v for k,v in sorted(authorities.items(), key=lambda item: item[1],
reverse=True)}
  print(hubs_sorted['n1'], ' : ',hubs_sorted['n3'],'|||', authorities_sorted['n1'], ' :
',hubs_sorted['n3'] )
```

```
0.18725288268260323  :  0.07587619182703625 ||| 0.1872528826826032  :  0.07587619182703625
0.16770629581129273  :  0.05949794063168375 ||| 0.16770629581129323  :  0.05949794063168375
0.12713765557988813  :  0.062362445093895644 ||| 0.12713765557988815  :  0.062362445093895644
0.13501191764079312  :  0.06030410851802328 ||| 0.13501191764079304  :  0.06030410851802328
0.14382183658602773  :  0.061930903030158364 ||| 0.1438218365860275  :  0.061930903030158364
0.11969374265773357  :  0.10962640792059754 ||| 0.11969374265773357  :  0.10962640792059754
0.13621252284871307  :  0.08427894765259066 ||| 0.13621252284871327  :  0.08427894765259066
0.11745086762077393  :  0.09346435879800764 ||| 0.11745086762077403  :  0.09346435879800764
0.10992849617321215  :  0.11209054721813097 ||| 0.1099284961732121  :  0.11209054721813097
0.11996366356435789  :  0.024532558539870002 ||| 0.1199636635643579  :  0.024532558539870002
0.05646752030535488  :  0.02523448798976474 ||| 0.05646752030535488  :  0.02523448798976474
```

I mentioned that the importance of n3 decreases rapidly while the importance of n12 increases sharply. It can be seen in each iteration, hub scores decreases for both n1 and n3, which indicated that under the stress, the criminal network died out.

The comments in part(g) would be identical as what I have written down.

'''

One particular actor that came to me is n12. In the first few graphs, n12 was appears to be not important. It was simply connected to n1 or n3. But after the stress was introduced, n3 and n12 together forms a very centralized role. However, as time pass by, n3 became outcast starting from phase 10 while n12 became one of the most important nodes.

Presumably, n12 was not significant in the early phases, but it became an integral part of drug import after phase4. After continued failed operations, it sems n1 abandoned n3 and it became almost isolated in the last phase of investigation. As indicated in phase 10 graphs, the networks concerning supplying drugs seems to be evolving as they face stresses(confistications).

'''