Written Report – 6.419x Module 3

Name: Sandipan Dey

2. Problem 1: Suggesting Similar Papers

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

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

Going by the rows of matrix O(n) and considering O(n^2) pairs for each row takes O(n^3) time.

Matrix multiplication A^TA takes O(n^3) with naive implementation but using Strassen's Algorithm it takes O(n^{lg7})=O($n^{2.8074}$) time, so it's faster in the worst case.

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

Bibliographic coupling and co-citation can both be taken as an indicator that papers deal with related material. However, they can in practice give noticeably different results. Why? Which measure is more appropriate as an indicator for similarity between papers?

Co-citation is a better indicator of similarity between papers, since

- Similarity of the papers measured by bibliographic coupling is questionable sometimes, since two works may reference completely unrelated subject matter in the third.

- Outgoing citation counts are fixed, hence the similarity relationship between documents lies in the past and is static, i.e. bibliographic coupling strength cannot change over time.

- Co-citation addresses this shortcoming of bibliographic coupling by considering a document's incoming citations to assess similarity, a measure that can change over time.

- Additionally, the co-citation measure reflects the opinion of many authors and thus represents a better indicator of subject similarity.

3. Problem 2: Investigating a time-varying criminal network

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?

With the increasing number of phases the police got more informed about among which of the players the communications are likely to happen (and the roles of players in the Serero organization) - that's why number of players chosen for wiretapping gradually levelled out.

The conclusion obtained from part (b) Q5 using centrality measures is along the same lines with the roles of the players identified with wiretapping, e.g., n1 being the mastermind has the highest centrality measure. So we need not adjust the conclusion.

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.

Degree centrality: proportional to the number of players a given player is communicating with. A player with higher value communicates to more players (hence has more influence), limitations: does not capture the notion of communication to important players (may fail to capture cascading effect)

Betweenness Centrality: Measures the extent to which a player lies on the communication channels between other players.

Eigenvector Centrality: Depends both on the number of neighboring players a given player communicated with and the importance of the neighbors.

Since the Eigenvector Centrality measures the cascading effect of importance, it should be the best measure to capture the mastermind running the illegal activities (the player with the highest eigenvector centrality).

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

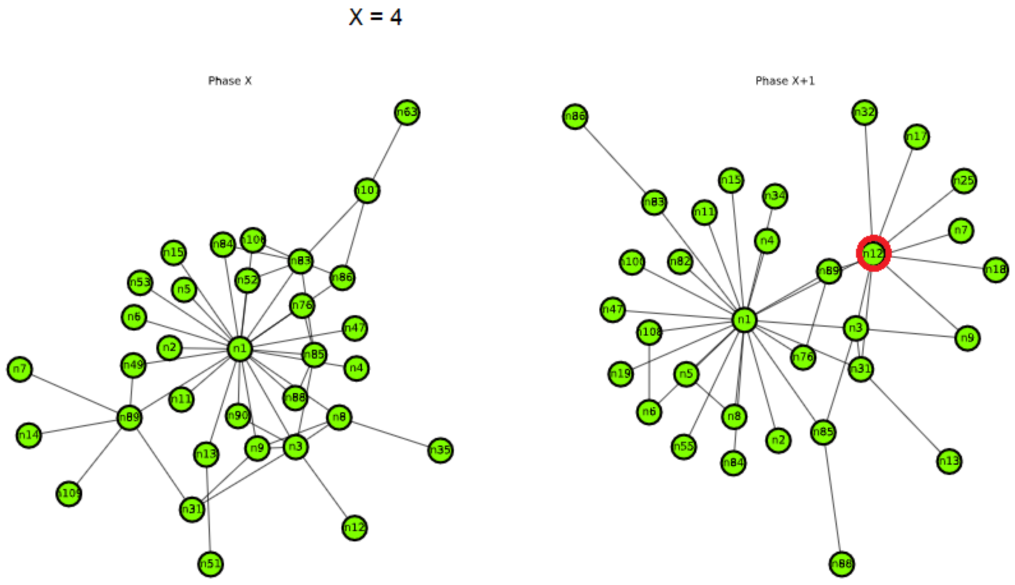
In real life, the police need to effectively use all the information they have gathered, to identify who is responsible for running the illegal activities of the group. Armed with a qualitative understanding of the centrality metrics from Part (d) and the quantitative analysis from part Part (b) Question 5, integrate and interpret the information you have to identify which players were most central (or important) to the operation.

A player is most important if the player communicates with many and / or important players. That way eigenvector centrality is the best measure to identify the most central players and the top 3 most important players are n1, n3 and n85.

Hint: Note that the definition of a player's “importance" (i.e. how central they are) can vary based on the question you are trying to answer. Begin by defining what makes a player important to the group (in your opinion) ; use your answers from Part (d) to identify which metric(s) are relevant based on your definition and *then*, use your quantitative analysis to identify the central and peripheral traffickers. You may also perform a different quantitative analysis, if your definition of importance requires it.

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.



After the first seizure, happening in Phase 4, traffickers reoriented to cocaine import from Colombia, transiting through the United States. As can be seen from the above figure, it supports the above change, where ‘n12’, the Principal organizer of the cocaine import, intermediary between the Colombians and the Serero organization, became a important node with high degree centrality.

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?

After the first seizure, happening in Phase 4, traffickers reoriented to cocaine import from Colombia, transiting through the United States. The coarse pattern is that after phase 4, the player ‘n12’ (the Principal organizer of the cocaine import, intermediary between the Colombians and the Serero organization) became an important player and started playing a key role in trafficking along with ‘n1’, where many of the players started communicating with him directly.

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.

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.

- 'n8' (being the player with 6th highest eigenvector centrality ~0.1524)
- 'n2' (being the player with 9th highest eigenvector centrality ~0.1143)
- 'n9' (being the player with 10th highest eigenvector centrality ~0.1143)

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?

In-degree centrality / Left eigenvector: importance of a given player comes from the (importance of the) players starting communication with the player.

Out-degree centrality / Right eigenvector: importance of a given player comes from (the importance of the) other players he starts communication with.

But we shall miss the bidirectional importance accumulation.

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?

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

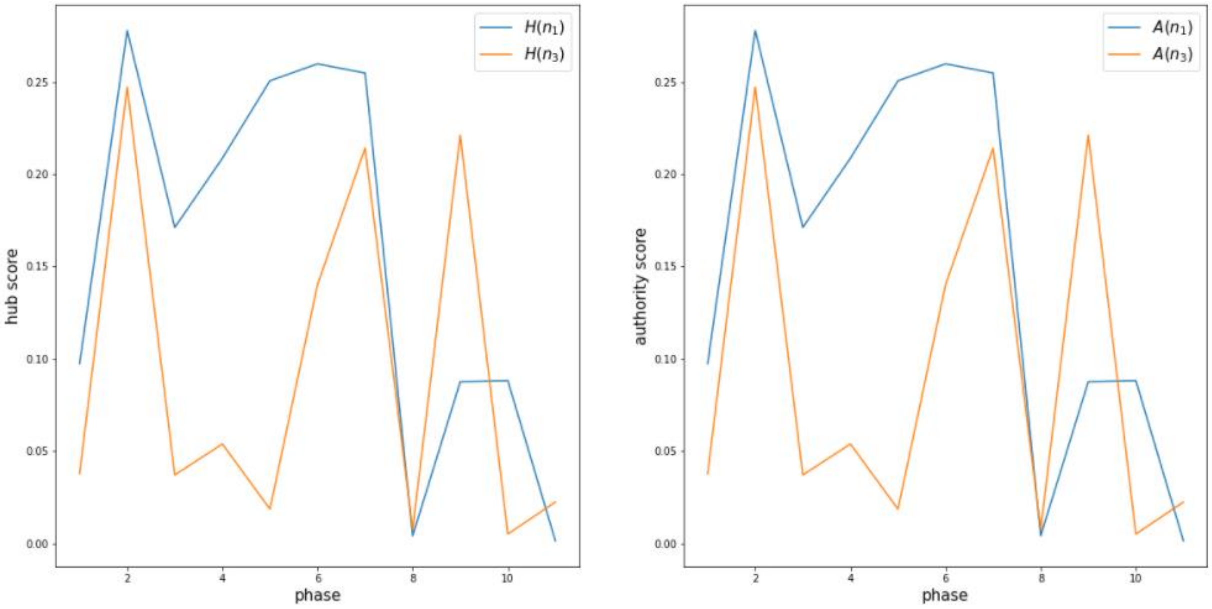
Recall the definition of hubs and authorities. Compute the hub and authority score of each actor, and for each phase. (**Remember** to load the adjacency data again this time using `create_using = nx.DiGraph()`.)

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)?

With the following code, computing the hub and authority scores for **n1** and **n3** for different phases,

```
for i in G.keys():
    h, a = nx.hits(G[i], max_iter=1000000)
```



We can see that the hub / authority score for both **n1** and **n3** decreases over time in general, since a reorientation happens after phase 4, after which the player **n12** gains more importance.

5. Problem 3 Project

Project:

Let’s consider the clustering problem and algorithm as discussed in the “Spectral Clustering" lecture, and implement this procedure to the **CAVIAR** networks. The question that we shall like to look into is the following:

    - what are some changes that we would notice in the output clusters, over time (phases)

- **(2 points)** Describes methodology for network analysis.

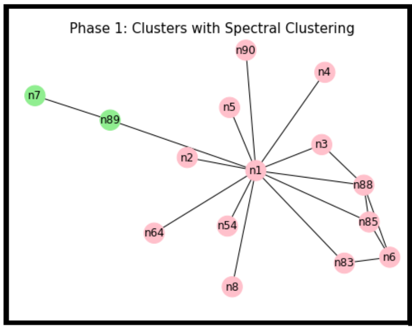
    - We shall apply spectral clustering on the COVIAR network to partition it into 2 primary clusters and try to understand how the output clusters change over phase.
    - We shall use the eigenvector corresponding to the second smallest eigenvalue of the Graph Laplacian corresponding to each component and threshold the values of eigenvector at zero to find those two clusters.
    - Since the graph becomes disconnected for some of the phases, we shall repeat the analysis for each of the components in each phase.

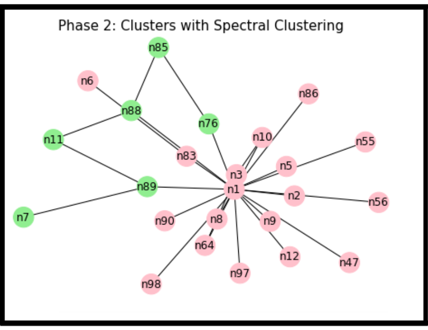
- **(2 points)** Grader is convinced that the methodology makes sense for the question to be answered. Grader is convinced that no additional methodology **within the bounds of techniques taught and discussed in this module** could be applied beyond what was described. The grader should only consider additional methodology that adds meaningfully to the answer for the question: additions that simply repeat or confirm the presented results should not be considered by the grader. If a justification is provided for why a particular method was not used, the grader should be convinced by that argument.

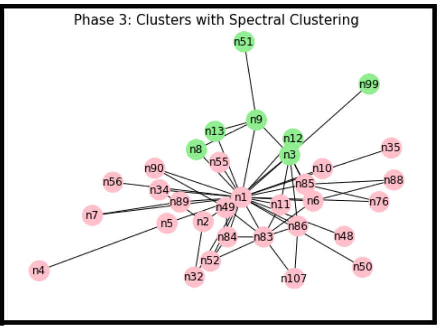
- **(2 points)** Presents results, including figures and/or statistics, which address the question of interest.

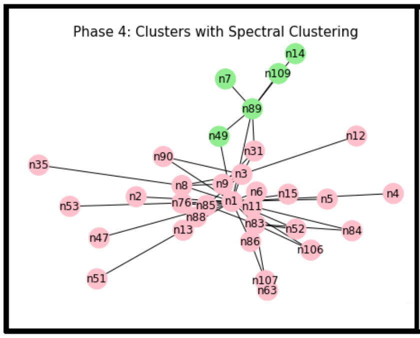
  The following figure shows how the output clusters change as the phase change, output clusters with spectral clustering was generated using the following code.

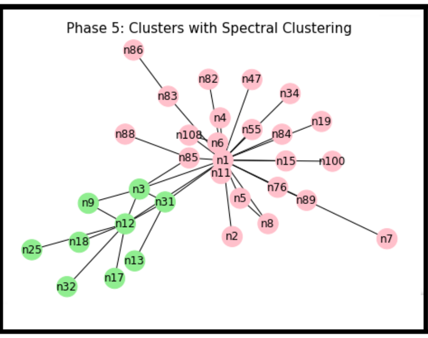
```
for  i in G.keys():
    g = G[i]      # ith phase
    for gcc in sorted(nx.connected_components(g),  key=len,  reverse=True):
        g0 = g.subgraph(gcc)
        ev = nx.linalg.algebraicconnectivity.fiedler_vector(g0) # spectral partitioning
        partition = {node:(1 if val <0 else 2) for (node, val) in zip(g0.nodes, ev)}
```

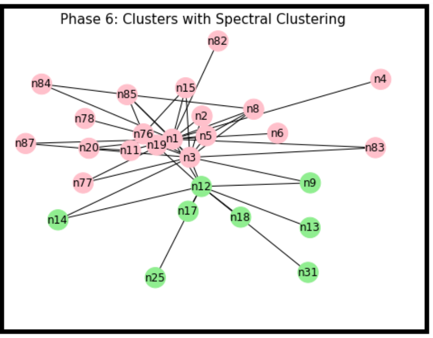

Phase 1: Clusters with Spectral Clustering

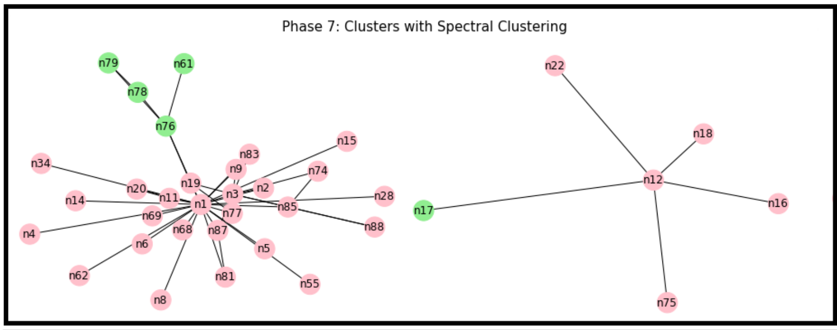

Phase 2: Clusters with Spectral Clustering

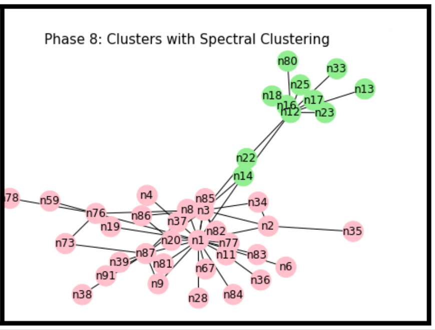

Phase 3: Clusters with Spectral Clustering

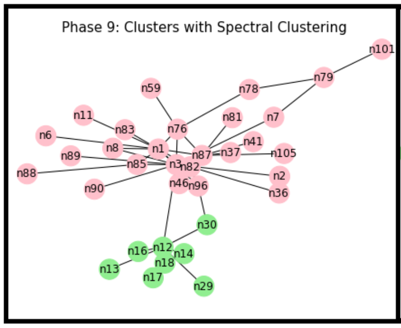

Phase 4: Clusters with Spectral Clustering

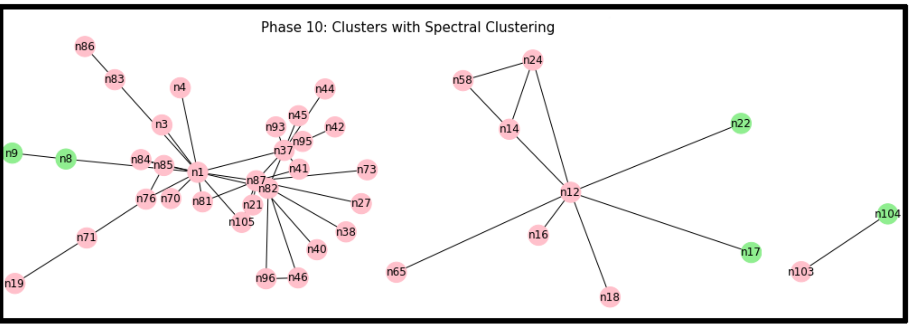

Phase 5: Clusters with Spectral Clustering

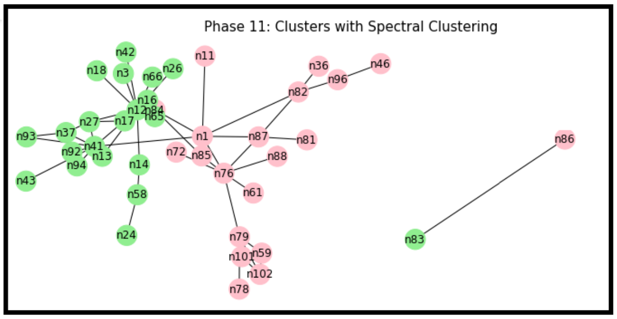

Phase 6: Clusters with Spectral Clustering


Phase 7: Clusters with Spectral Clustering


Phase 8: Clusters with Spectral Clustering

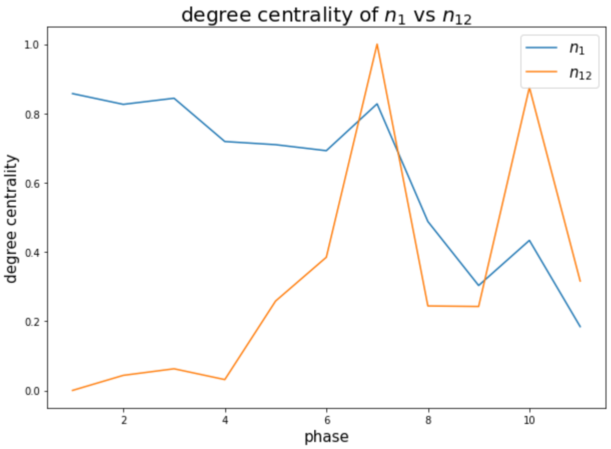

Phase 9: Clusters with Spectral Clustering

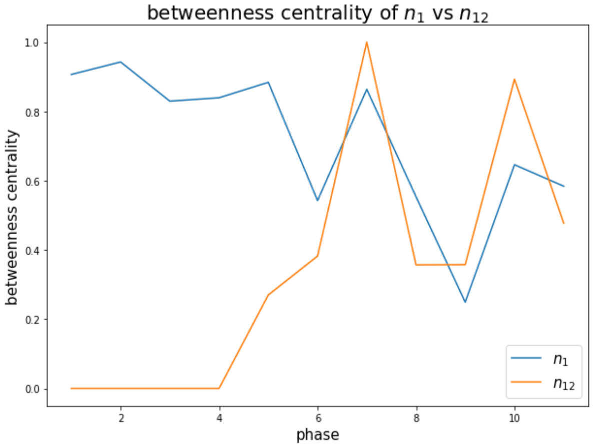

Phase 10: Clusters with Spectral Clustering


Phase 11: Clusters with Spectral Clustering

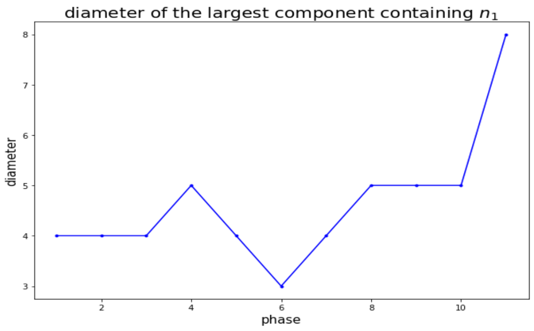
As can be seen from the above figure, the cluster outputs changed gradually as the phase increased. First let's note that for phases 7, 10 and 11 the graph did not remain connected, so there were multiple components in these phases.

From the above figure (showing the results of spectral clustering) and the following 2 plots, we can see that the player n_1 (being the Mastermind of the network) has the highest degree / between-ness centrality, but his centrality score decreases as phase increases, whereas centrality score of the player n_{12} (being the Principal organizer of the cocaine import, intermediary between the Colombians and the Serero organization) increases in general as phase increases (particularly so, after phase 4). At phases 7, 10 and 11 the centrality score of n_{12} appear to be higher than that of n_1, since they are at different components of the graph. This supports the event that after the first seizure, happening in Phase 4, traffickers reoriented to cocaine import from Colombia, transiting through the United States and the player n_{12} became very important player.





Now, let's plot the diameter of the (largest) component containing n_1 at each phase and observe from the following figure that it also decreases, right after phase 4, when the reorientation took place.



Now, let’s try to explain the result of the spectral clustering from the visualization of the network graph. For the first 4 phases the graph contains almost all the players in pink cluster (the cluster containing n_1), there are very few players in the green cluster (mostly containing n_{89} and n_7, Investor). But from phase 5 onward, the player n_{12} started to gain more importance after reorientation (containing n_{12}) of the network and the size of the green cluster kept on increasing (for phases 7 and 10, the players communicating with n_{12} formed a different component altogether).

- **(2 points)** The described methodology has been applied in complete and the results shown (that is, the author did not forget to include anything they discussed in the methodology.)

4. Adequately discusses the results obtained.

- **(2 points)** Question does not need to be successfully answered, but the grader should be convinced that the author has answered the question to the best ability of the methodology presented.

- **(1 point)** Provides commentary on what was discovered, what were the limitations of the methods, what may have been surprising to discover, etc.

- **(1 point)** Award this point if the question **was** successfully answered to the grader's satisfaction.

Discovered

- The change in the output clusters that corroborated the fact that there was a reorientation in the network after phase 4 (seizure) and the player n_{12} became more important in the following phases, with more players communicating to him directly.

Limitations

- This does not show the quality of the clusters (e.g., with modularity etc.) that can be done by community detection algorithms such as Louvain method and then successive computation of modularity.