Statistics Review: Homework/Analysis

MITx 6.419x Data Analysis: Statistical Modeling and Computation in Applications

Ellick Hou - 4/17/21

table of contents

Problem 1: Suggesting Similar Papers....2
- Part (c) 2
- Part (d) 3

Problem 2: Investigating a time-varying criminal network 4
- Part (c) 5
- Part (d) 6
- Part (e) 7
- Part (f) Question 2 8
- Part (g) 8
- Part (h) 9
- Part (j) 10

Problem 3: Co-offending Network....11
- Part (g) 11

Problem 1: Suggesting Similar Papers

✦ Part (c)

(2 points) How does the time complexity of your solution involving matrix multiplication in part (a) compare to your friend's algorithm? (100 word limit.)

➡ Ans:

(1) Adjacency matrix($A \in \mathbb{R}^{N \times N}$) is a Dense matrix.

- Big-O of friend's algorithm = $O(N^3)$
- Big-O of dense matrix multiplication = $O(N^3)$
- The time complexity of these two solutions is the same, but we can access accelerated linear algebra libraries such as OpenBLAS(by Numpy) to accelerate the actual matrix multiplication execution time.

(2) Adjacency matrix($A \in \mathbb{R}^{N \times N}$) is a Sparse matrix.

- If we assume every paper has a maximum of citations, c, then the adjacency matrix defined in Part(a) would become a sparse matrix when N increase, with the number of non-zero entries $< cN$.
- Big-O of friend's algorithm = $O(N^2)$
- Big-O of dense matrix multiplication = $O(N^3)$
- However, if we store A in a sparse matrix format, CSR, for example, could reduce the time complexity to $O(N)$.

friend's algorithm	Dense	Sparse
for each row of A, do	N	N
if sum(i_row) > 1, do	N	N
for edge in comb([where(i_row>0)], 2), do	$\frac{N(N-1)}{2}$	$\frac{c(c-1)}{2}$
C[edge[0],edge[1]] +=1	1	1
C[edge[1],edge[0]] +=1	1	1
	$O(N^3)$	$O(N^2)$

generating the co-citation matrix by A in CSR format	
Let	
col_ind = [$j_{nonzero_1}$, $j_{nonzero_2}$, $\cdots$]	
ind_pointer = [0, p_{row_0}, p_{row_1}, $\cdots$, $p_{row_{N-1}}$]	
for i in [0, 1, …, N-1], do	N
col_in_irow = col_ind[ind_pointer[i] : ind_pointer[I+1]]	c
for edge in permutations(col_in_i_row, 2), do	c(c-1)
C[edge[0],edge[1]] +=1	1
	$O(N)$

Part (d)

(3 points) 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? (200 word limit.)

➡ Ans:

- Bibliographic coupling captures the relationships depend on papers' out-degree, co-citation captures relationships depend on papers' in-degree.
- Assume we add new paper into the two different networks relies on its first-ever publication date.
  - Once the paper has been added, the bibliographic coupling network decides the similarity between itself and other papers. This relationship is fixed and will not change by adding more new papers. It depends on the properties of the paper itself.
  - The co-citation network won't build any relationship until the paper has been cited. The relationship depends on other papers' properties and allows to gain more weight during the network expands.
- The appropriateness of these two measures depends on what kind of similarity we are looking for.
  - If we look for the similarity between research papers, it is more appropriate to describe the similarity by co-citation. The greater the weight an edge has, the more authors think these two papers relate to each other in their work.
  - If we look for the similarity between review papers, the bibliographic coupling works better. Review papers might cite some classical paper together when they are reviewing similar subjects.
  - Additionally, if we want to find papers that are similar to a very new paper, co-citation might give us a few or even no neighbors. Finding the bibliographic couples of the very new paper might give us much information if this is the case.

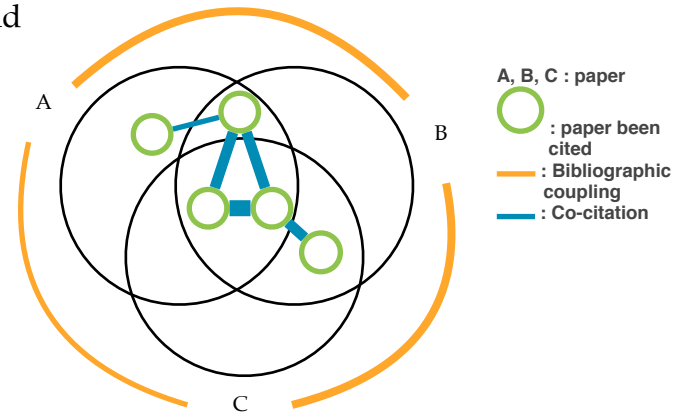

Figure P1.1: Illustration of Bibliographic coupling and co-citation relationships between papers.

Problem 2: Investigating a time-varying criminal network

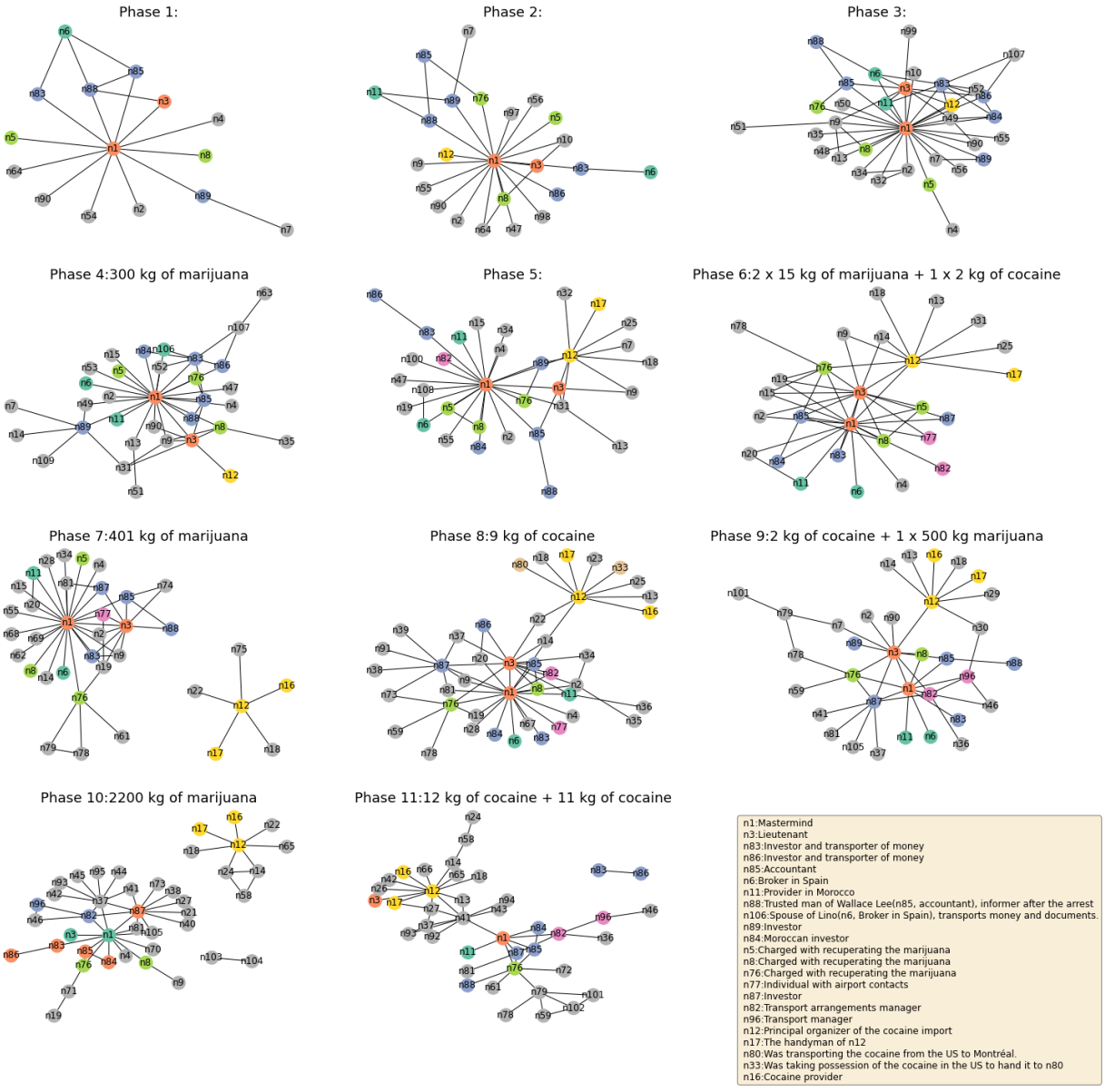

Figure P2.1: Undirected graph of CAVIAR Phase 1 to Phase 11

✦ Part (c)

(2 points) 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?

➡ Ans

❖ Perhaps the police are finding who they should wiretap and focus on in the first few phases, and maybe they need more evidence to wiretap more suspects. After they target down the key player, they need to figure out who is really involved in the criminal activities and filter out the suspects, so the number of nodes won't change that sharply.

❖ To eliminate the doubt that the high centrality in the first few phases is due to the investigation just beginning. The mean centrality rank across different ranges of phases has calculated in Table P2.1. Combine the information in Figure P2.1 and Table P2.1, we can have the inference below.

- The The accountant(n85) has been wiretapped from the beginning of the investigation. However, as the investigation proceeds, one of the investors(n87) exposed. From the first time n87 exposed(phase 6), n87 plays as the center of the "Serero organization" till the end. Although the accountant(n85) been squeezed out of the top3 rank, n85 still directly contacts the mastermind and plays an important role.
- The principal organizer of cocaine import (n12) never shows on the top5 rank of eigenvector centrality. But, we can tell in Figure P2.1 that the cocaine criminal group seems to be divided out from the main group and cause the low eigenvector centrality of n12.
- If we rank the top5 key players in the "Serero organization", maybe [n1, n3, n12, n76, n87] is more appropriate.

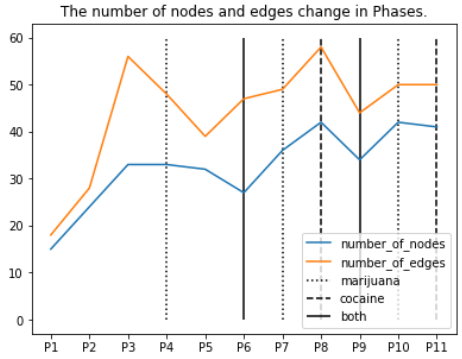

Figure P2.2: The number of nodes and edges change in Phases.

rank	All phases			4 to 11			6 to 11		
	Eigen vector	Between ness	Degree	Eigen vector	Between ness	Degree	Eigen vector	Between ness	Degree
1	n1	n1	n1	n1	n1	n1	n1	n1	n1
2	n3	n12	n3	n3	n12	n3	n3	n12	n3
3	n85	n3	n12	n87	n3	n12	n87	n3	n12
4	n76	n76	n85	n85	n76	n76	n76	n76	n87
5	n83	n87	n76	n76	n87	n87	n85	n87	n76

Table P2.1: Centralities rank across different phases range

Id dict: {n1: Mastermind,
n3: Mastermind's lieutenant,
n12: Principal organizer of the cocaine import,
n85: Accountant,
n87: Investor,
n76: Charged with recuperating the marijuana}

Part (d)

(5 points) 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.

➡ Ans

- The eigenvector centrality tells who directly contacts the mastermind or other key players and might imply who is the head of the different functional groups(investor, finance, sells, etc.).
- The betweenness centrality tells who connects the whole organization.
- The degree centrality tells who contacts the most people been wiretapped in the investigation.
  - The degree centrality can not inform us who has the power to contact other "C-levels" in the organization.
  - For example, in the [n84, n79, n31, n89] set from Phase 4 to Phase 11
    - degree centrality = [0.037 , 0.034, 0.033, 0.039]
    - eigenvector centrality = [0.088, 0.015, 0.049, 0.052]
    - If we only look at the degree centrality, these four guys show no difference. However, if we check the eigenvector centrality, n84 is higher than others. Turns out n84 is one of the investors, and others play no roles in the organization.
- In comprehensive, I think eigenvector centrality identifies the key player the most for 2 reasons.
  - If we check the top10 of each metrics, the accuracy of players is in "Serero organization" or not is [degree: 8/10, eigenvector: 8/10, betweenness: 6/10]. The degree centrality ties the eigenvector centrality, but for the limitation mention above, The eigenvector centrality has a better chance to identify organization members when the degree centrality is low.
  - As mentioned in Part (c), although n12 not on the top5 leaderboard of the eigenvector centrality, eigenvector centrality still rank n12 in 6th.

Part (e)

(3 points) 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.

➡ Ans

- n1, n3 are the most critical players in the organization because they have the highest score in all three metrics.
- n12 is also important because he has the 2nd highest score in betweenness centrality and 6th in eigenvector centrality. Also, he connects the cocaine criminal group and the main group, as shows in Figure P2.1.
- n87 might be the main investor because he has the highest score among all investors, which shows he contacts other organization members the most.
- n85 is important because although n87 squeezed him out of the top3 leaderboard of eigenvector centrality, he is still on the top5 leaderboard. Moreover, he is one of the players who continuously contacts with n1 from the beginning to the end.
- n76 might be the head of the marijuana sellers because he is the only one who can contact the accountant and the investors, among the three players who "Charged with recuperating the marijuana," and 41.5% of his total degrees across all phases is not the organization members.

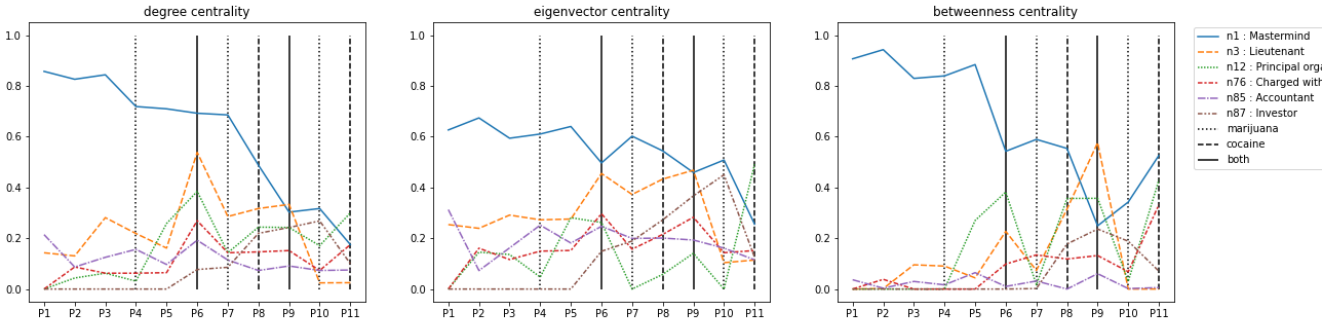

Figure P2.3: The centralities of key players' change in Phases.

Part (f) Question 2

(3 points) 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.

➡ Ans

- From phase 4 to phase 5, after the first seizure, n12 starts to run the cocaine business, eigenvector centrality starts to increase and exceed n76,n85. Betweenness centrality become the second highest in the 6 key players mentioned above. (Figure P2.3)

Part (g)

(4 points) 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?

➡ Ans

- Yes
- Phase 1 to Phase 4, finding suspects and evidence to get more wiretap permit, network expand.
- Phase 5, after the first seizure, n12 starts to run the cocaine business, eigenvector centrality starts to increase.
- Phase 6, n3 join the center of the network, eigenvector centrality increase. At the same time, the eigenvector centrality of n1 decreases, n1 might release power to n3.
- Phase 7, after the first time cocaine been seized, cocaine group disconnect to the main group. n12's eigenvector centrality = 0.
- Phase 8, cocaine group connect to the main group, n12's betweenness back to the original level.
- Phase 9, after the second time cocaine been seized, n96(Transport manager, owner of a legitimate import company), another transport accomplice shows up, n77 disappear in the network.
- Phase 10, another cocaine seizure, new import way didn't help, the cocaine group disconnect to the main group again. n12's eigenvector centrality = 0.
- Phase 11, after the marijuana business undergoes a disastrous attack. Network shift its center to the cocaine group.

Phase	# seizures	monetary loss	drugs
Phase 4	1	2,500,000	300 kg of marijuana
Phase 6	3	1,300,000	2 x 15 kg of marijuana + 1 x 2 kg of cocaine
Phase 7	1	3,500,000	401 kg of marijuana
Phase 8	1	360,000	9 kg of cocaine
Phase 9	2	4,300,000	2 kg of cocaine + 1 x 500 kg marijuana
Phase 10	1	18,700,000	2200 kg of marijuana
Phase 11	2	1,300,000	12 kg of cocaine + 11 kg of cocaine

Table P2.1: Seizures

✦ Part (h)

(2 points) What are the advantages of looking at the directed version vs. undirected version of the criminal network?

➡ Ans

- ❖ The information flow is more clear. We can tell who is more active and who is more passive.

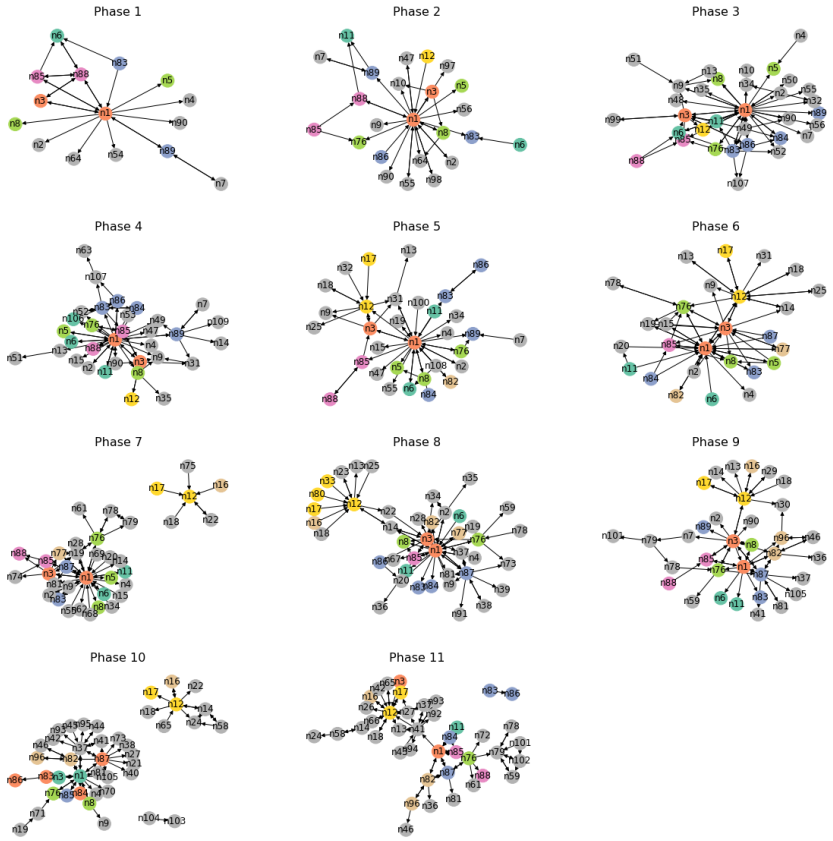

Figure P2.4: Directed graph of CAVIAR Phase 1 to Phase 11

Part (j)

(2 points) 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)?

➡ Ans

- In Phase 6, it's the first time n3 join the center of the network, its own eigenvector centrality and degree centrality sharply increase. At the same time, the eigenvector centrality of n1 decreases. Combine the message in Figure P2.5, by the definition of hub score, function(1), A node has a high hub score because he points to many authorities. From Phase 6 to Phase 7, the hub score of n1 decrease and the hub score of n3 raise. It shows n1 want n3 to pass his order. n1 is from active to passive during the time. After the big monetary loss in Phase 7, n1 want to re-hold his power so the hub and authority score back to before.

$$x^{k+1} = \alpha A y^k \qquad (1)$$
$$(y^{k+1})^T = \beta (x^{k+1})^T A \qquad (2)$$

- In Phase 11, when the center of the network shift to the cocaine group, the n12's authority score sharply increase.

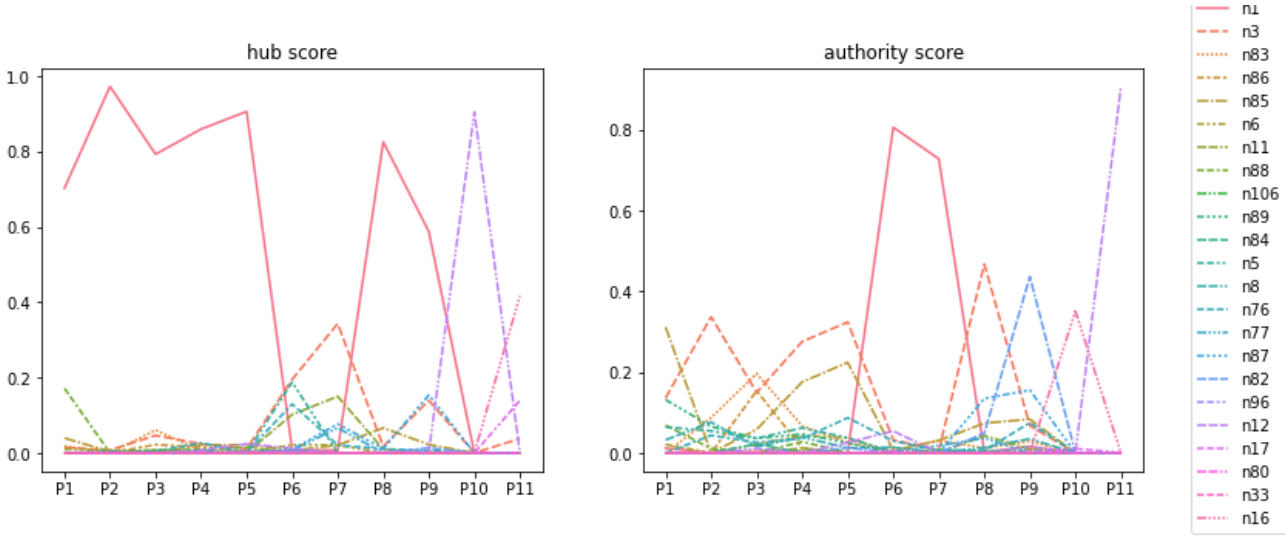

Figure P2.5: The hub and authority scores.

Problem 3: Co-offending Network

✦ Part (g)

(3 points) 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?