Written Report – 6.419x Module 3

Name: ptoche

Long calculations and code snippets are in the appendix at the end of this report.

Problem 1: Suggesting Similar Papers

*A citation network is a directed network where the vertices are academic papers and there is a directed edge from paper **A** to paper **B** if paper **A** cites paper **B** in its bibliography. **Google Scholar** performs automated citation indexing and has a useful feature that allows users to find similar papers. In the following, we analyze two approaches for measuring similarity between papers.*

Part (a): Co-citation network (auto-graded)

Two papers are said to be cocited if they are both cited by the same third paper. The edge weights in the cocitation network correspond to the number of cocitations. In this part, we will discover how to compute the (weighted) adjacency matrix of the cocitation network from the adjacency matrix of the citation network.

***Problem setup:** In order to derive the cocitation matrix, we need to derive it as a function of the original adjacency matrix.*

***Problem notation:** If there is an edge from paper i to paper j, it means that paper i cites paper j. We will denote by A the corresponding adjacency matrix, such that $A_{ij} = 1$ means there is a directed edge from i to j. Let us denote by C the cocitation network matrix.*

Question 1 (auto-graded)

In attempting to derive the cocitation matrix, your friend came up with the following algorithm:

Assuming the row indices of the matrix mean that the paper is citing others, and the column indices that the paper is being cited, then the algorithm's steps would be:

- *Construct an empty matrix for C.*
- *Go through the rows of A one by one.*
- *For each row r of A, 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). Note that by following this rule, you will naturally also add 1 to C at location (b, a) as the pair $((r, b), (r, a))$ must also be present.*

After reading carefully through the proposed steps, please answer the following:

Does this generate the cocitation weighted adjacency matrix?

What is the big-O complexity, $\mathcal{O}$, of the proposed algorithm, in terms of n, the number of nodes in the graph?

Question 2 (auto-graded)

Write the cocitation weighted adjacency matrix, C, in terms of A using matrix operations. The diagonals in your answer need not match the diagonals generated by the definition in Question 1; the off-diagonals should match Question 1.

Hint: How can you use A_{ki} and A_{kj} to represent a logical AND for edge (k, i) and (k, j) being present in the graph? Use this to write down C in terms of C_{ij} then find the corresponding matrix operations.

Part (b): Bibliographic coupling

(5 points) *Two papers are said to be bibliographically coupled if they cite the same other papers. The edge weights in a bibliographic coupling correspond to the number of common citations between two papers.*

How do you compute the (weighted) adjacency matrix of the bibliographic coupling, B, from the adjacency matrix of the citation network, A? Write your answer in terms of matrix operations.

Part (c)

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

Part (d)

(3 points) *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?*

Problem 2: Investigating a time-varying criminal network

In this problem, you will study a time-varying criminal network that is repeatedly disturbed by police forces. The data for this problem can be found in the CAVIAR directory of the data archive.

The CAVIAR investigation lasted two years and ran from 1994 to 1996. The operation brought together investigation units of the Montréal police and the Royal Canadian Mounted Police of Canada. During this two year period, 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).

This case is interesting because, unlike other investigative strategies, the mandate of the CAVIAR project was to seize the drugs without arresting the perpetrators. During this period, imports of the trafficking network were hit by the police on eleven occasions. ***The arrests took place only at the end of the investigation.*** *Monetary losses for traffickers were estimated at* 32 *million dollars. Eleven seizures took place throughout the investigation.* ***Some phases included no seizures, and others included multiple.*** *The following summarizes the* 11 *seizures:*

Phase 4	1 seizure	$2,500,000	300 kg of marijuana
Phase 6	3 seizures	$1,300,000	2×15 kg of marijuana $+1 \times 2$ kg of cocaine
Phase 7	1 seizure	$3,500,000	401 kg of marijuana
Phase 8	1 seizure	$360,000	9 kg of cocaine
Phase 9	2 seizures	$4,300,000	2 kg of cocaine $+1 \times 500$ kg marijuana
Phase 10	1 seizure	$18,700,000	2,200 kg of marijuana
Phase 11	2 seizures	$1,300,000	12 kg of cocaine +11 kg of cocaine

This case offers a rare opportunity to study a criminal network in upheaval from police forces. This allows us to analyze changes in the network structure and to survey the reaction and adaptation of the participants while they were subjected to an increasing number of distressing constraints.

The network consists of 110 *(numbered) players. Players* 1-82 *are the traffickers. Players* 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.*

According to the police, the role of 23 *of the players in the "Serero organization" are the following, listed by name (unique id):*

- *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.*

In the data files, you will find matrices that report the number of wiretapped correspondences between the above players in the network, where players are identified by their unique id. You will be analyzing this time-varying network, giving a rough sketch of its shape, its evolution and the role of the actors in it.

The following questions will use undirected graphs derived from this data.

Part (a): Question 1

(auto-graded) *What is the size of the network at each phase? Plot the evolution of the number of node and number of edges over time, from phase 1 to 11.*

Provide the number of nodes and edges for the three phases listed below:

Part (a): Question 2

(auto-graded) *Try visualizing the graph at each phase. The graphviz algorithm is recommended for these complex graphs, and you will need it to answer some of these questions. Visualize the graph for Phase* 3*. Which of the following plots below correspond to Phase* 3*?*

Part (b): Question 1

(auto-graded) *For each of the* 11 *phases and for each of the players under investigation (i.e., the* 23 *listed above), compute and list the normalized degree centrality of the player.*

The normalized degree centrality of node i is defined as:

$$\tilde{k}_i = \frac{k_i}{n-1}$$

where k_i is the degree of node i and n is the number of nodes in the graph. Provide the degree centrality for the following four players, at the specified phases:

Part (b): Question 2

(auto-graded) *For each of the* 11 *phases and for each of the players under investigation, compute and list the normalized betweenness centrality of the player.*

For undirected graphs, the normalized betweenness centrality for node i is defined as:

$$\tilde{B}_i = \frac{2}{(n-1)(n-2)} \sum_{s \neq i \neq t} \frac{n_{st}^i}{g_{st}}$$

where n_{st}^i is the number of shortest paths between s and t that pass through i, and g_{st} is the total number of shortest paths between s and t. Note that this considers both orderings of each pair of nodes, so for undirected graphs, a path counts twice (as it counts both for n_{st}^i and for n_{ts}^i).

Provide the normalized betweenness centrality for the following four players, at the specified phases:

Part (b): Question 3

(auto-graded) *For each of the* 11 *phases and for each of the players under investigation, compute and list the eigenvector centrality of the player.*

Ensure your eigenvector centrality is normalized as

$$\sqrt{\sum_i v_i^2} = 1$$

Provide the eigenvector centrality for the following four players, at the specified phases:

Part (b): Question 4

(auto-graded) *Recall the mathematical definition of each of these metrics, along with the algorithm that is best suited to compute it and the corresponding time complexity.*

Which algorithm is the fastest for this data set?

1. *Degree centrality*
2. *Betweenness centrality*
3. *Eigenvector centrality*

Part (b): Question 5

(auto-graded) *The data from questions* 1 *to* 3 *can be used to perform different types of quantitative analyses. In this question we will look at performing one such analysis — we will determine the temporal consistency of a player's centrality , i.e. which players consistently remained active and central throughout most of the phases and which didn't?*

To answer this question, look at the temporal evolution of the networks and calculate the mean centrality for each of the centrality metrics, across all phases, for every player.

Note: As every actor might not be present in every phase, attach a centrality of zero to an actor for the phases in which they are not present, before calculating these statistics, so that you take a mean over all 11 *phases for all actors.*

Food for thought (not graded): *What are the implications of this step? What else could you do to ensure that your numbers are comparable with each other?*

For the betweenness centrality, which three players have the highest mean?

For the eigenvector centrality, which three players have the highest mean?

Part (c)

(2 points)(100-200 ***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?*

Part (d)

(5 points)(300-400 ***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.*

Part (e)

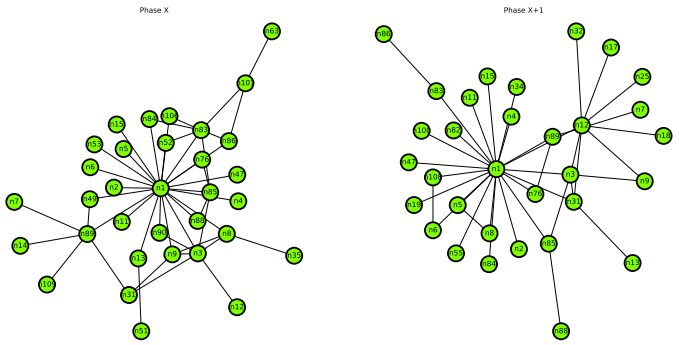
(3 points) (100-200 ***word limit)*** *In real life, the police need to effectively use all the information they have gathered, to identify who is responsible for running the illegal activities of the group. Armed with a qualitative understanding of the centrality metrics from Part (d) and the quantitative 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.*

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 1

(auto-graded) *Now, we will attempt to analyze the overall evolution of the network and correlate the patterns we observe to events that happened during the investigation.*

The plots below visualizes the criminal network for 2 *consecutive phases Phase* X *and* $X + 1$*. Identify* X *using your visualization in Part (a) Question 2.*



$$X = 4$$

The plot represents Phase 4*.*

Part (f): Question 2

(3 points) (200-300 ***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.*

Part (g)

(4 points) (200-300 ***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.

Part (h)

(2 points) (50-100 ***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.*

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

Part (i)

(2 points) (150-250 ***word limit)*** *What are the advantages of looking at the directed version vs. undirected version of the criminal network?*

Part (j)

(4 points) (300-400 ***word limit)*** *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)?*

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

Project

The last part of this assignment is an open-ended project. Choose a sociologically interesting question about either the CAVIAR network or the facebook or twitter network from the recitation notebook section.

Try to answer your own question using the data. You can subset the data in whichever way you desire as long as it is (sociologically) meaningful. For example, in the case of a cooffending network, you could group nodes by attributes such as sex, group edges such as repeat/non-repeating cooffenses, use the weighted or unweighted co-offending networks, focus on the largest connected component, etc. Think of how you may want to subset the data in the context of the CAVIAR or the social networks, or the publicly available network data set you have chosen.

Project expectations/Rubric:

1. *Clearly states a sociological question which is interesting and relevant to the data. The question must be sociologically motivated: for example, "Compare the network structure in 2003 vs 2009" is not a good question, without further context. If you have some reason to believe that the network structure changes in those years, then you should make that your central question: for example, "Did crimes involving youth offenders become more organized and structured over the years" is a better question, from which comparing the structure in different years becomes part of the methodology to answer the question. More examples of possible questions for cooffending networks are provided below.*
2. ***(2 points)** Describes methodology for network analysis.*
3. ***(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.*
4. ***(2 points)** Presents results, including figures and/or statistics, which address the question of interest.*
5. ***(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.) Adequately discusses the results obtained.*
6. ***(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.*
7. ***(1 point)** Provides commentary on what was discovered, what were the limitations of the methods, what may have been surprising to discover, etc.*
8. ***(1 point)** Award this point if the question was successfully answered to the grader's satisfaction.*

1. Possible Project Suggestions for the CAVIAR network:

1. *Make use of centrality measures to identify concrete and quantifiable aspects of the criminal network that has changed over time, If possible, you could support them with a test of statistical significance. Then, provide a coherent explanation that provides examples using specific datapoints.*

2. *How has the immediate network of central criminal figures evolved in response to the police operations?*

3. *Consider the starting links as well as the new connections that arise throughout phases. Of the four network models discussed (Erdos-Renyi, configuration, preferential attachment, and small-world), which one(s) are the most realistic? Provide statistical tests if/when appropriate.*

4. *Consider the clustering problem and algorithm as discussed in the "Spectral Clustering" lecture, and implement this procedure to the CAVIAR networks. What are some changes that you would notice on either the clustering quality (i.e. modularity) or the output clusters, over time? (You may feel free to perform the algorithm on an appropriate subset of the graph, say by excluding certain nodes.)*

2. Possible Project Suggestions for the Facebook/Twitter network:

1. *Which attributes of the social network are assortative? Are there any disassortative attributes?*

2. *Pick one or both of the social media network graphs, and study select features of the graph such as the degree distribution, clustering, and centrality metrics. To what extent does the power-law distribution hold, and examine how these statistics can be used to select a candidate model among the four network models discussed (Erdos-Renyi, configuration, preferential attachment, and small-world)?*

3. *Formulate several hypotheses on how the Facebook and Twitter network graphs could differ, based on features or quantities discussed throughout the module. Then, devise a methodology to test for whether the difference is statistically significant, and carry out the analysis.*

4. *Consider the nodes that have the most connections (in Facebook) and who are followed by the most other accounts (on Twitter). Then, consider the subgraphs that consist of each of these nodes' followers. Formulate any interesting questions based on these subgraphs or their properties.*

3. Possible Project Suggestions for the Cooffending Network (since you will not be using this data set, this is just to give you further ideas of what is expected.):

1. *A common question in co-offending networks is the stability of relationships over time: do offenders commit crime with the same people over time, or do they choose new people to co-offend with as time passes? You may wish to investigate individual nodes to see how their neighborhoods change over time.*

2. *There are other potential examples of time-based studies. For example, is there a concept of seasonality in the structure of the data? Do crime networks become more clustered during certain months or seasons?*

3. *One can choose groups of nodes and consider homophily/assortativity, that is, how these nodes interact. For example, one can ask how females interact with males in the network: does one of the two sexes tend to have higher centrality in the network? One can also extend this to time-varying studies: for example, does the influence of females increase in more recent years?*

4. *The CAVIAR data involved an organized crime ring. Is there organized crime in the data? How does this network compare with the organized crime in the CAVIAR network?*

5. *How does the type of crime impact the network? Are there certain local structures, such as cliques or star graphs, that are associated with different types of crime? Can you identify different types of crime by the structure of a co-offending relationship alone?*