Problem Set 4

Issued: Monday, October 29, 2018 **Due:** Sunday, November 11, 2018

Problem 4.1: [15pts] Suggesting Similar Papers

The 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 to find similar papers. In the following, we analyze two approaches for measuring similarity between papers.

(a) *Co-citation network:* 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. How do you compute the (weighted) adjacency matrix of the cocitation network from the adjacency matrix of the citation network?

(b) *Bibliographic coupling:* 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 from the adjacency matrix of the citation network?

(c) 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 4.2: [40pts] 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 `CAVIAR.zip`.

Here is some information on the CAVIAR project and the role of certain individuals arrested following the investigation. This 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 (11 matrices match these phases). This case is rather unique, because unlike other investigative strategies, the mandate of the CAVIAR project was to seize the drugs without arresting villains. 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.** Here is what they represent in terms of the amount of money:

Phase 4	1 seizure	2'500'000$	300 kg of marijuana
Phase 6	3 seizures	1'300'000$	2 x 15 kg of marijuana + 1 x 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 x 500 kg marijuana
Phase 10	1 seizure	18'700'000$	2200 kg of marijuana
Phase 11	2 seizures	1'300'000$	12 kg of cocaine + 11 kg of cocaine

As you can see, this case offers a rare opportunity to study a criminal network in upheaval by 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, attempting 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 the actors of the "Serero organization" under investigation are the following:

- Serero, Daniel (n1) : Mastermind of the network.
- Pierre Perlini (n3) : Principal lieutenant of Serero, he executes his 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.

You will be analyzing the time-varying network, giving a rough sketch of its shape, its evolution and the role of the actors in it.

Questions:

(a) For each of the 11 phases, compute and list the:

- degree,
- betweenness centrality, and
- eigenvector centrality

of the actors under investigation.

(b) Describe which actors are central and which actors are only peripheral. Explain and validate your reasoning. Feel free to compute other graph parameters, in addition to the centrality measures in (a), to aid you in validating your answer. Who seem to be the three principal traffickers?

(c) Are there other actors that play an important role but are not on the list of investigation? List them, and explain why they are important.

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

(e) Describe and interpret the evolution of the role of the central actors found in (b). At which phases are they active? When do they withdraw? Find indices in the network evolution that reflect the description given to them.

(f) Examine the frequency and the directions of the communications of (n1) as the network evolves. Any contrast or pattern(s) you observe? Describe, explain and interpret.

(g) (*optional for undergraduates*) Would you consider that the particular strategy adopted by the police had an impact on the criminal network throughout the different phases of the investigation? What kind of impact? Explain.

Problem 4.3: [45pts] Co-offending Network

The data for this problem set was generously provided to us by Carlo Morselli (University of Montreal). This data set is not publicly available and is only for *in class use*. Do not share it with *anyone* outside this class. If you would like to study this data set further, for example for your final project, Carlo is interested in collaborations and possible research findings could be published with Carlo as co-author.

The data for this problem set consists of individuals who were arrested in Quebec between 2003 and 2010. Some of the individuals have always acted solo, and have been arrested alone throughout their 'career'. Others co-offended with other individuals, and have been arrested in groups. The goal of this problem set is to construct and analyze the co-offender network. The nodes in the network are the offenders, and two offenders share a (possibly weighted) edge whenever they are arrested for the same crime event.

The questions are not fully independent. We recommend reading through all the questions first before attempting to solve the problem set. It may be helpful to create a mental plan first of how to go about solving and implementing. This may save you time and allow you to reuse your code more effectively.

The data set may be found in `Cooffending.csv`. Additional information on the fields of the data set may be found in `DataDescription.txt`. First step consists of getting familiar with the data set. The following questions are optional in that while you are encouraged to compute the summary statistics, there's no need to describe them in report-form.

(a) (optional) How many data points, or cases, does this data set have?

(b) (optional) How many different offenders are there?

(c) (optional) How many different crime events are there? How many different crime events are there, per each year, from 2003 till 2010?

(d) (optional) Which crime(s) involved the greatest number of offenders? List the crime(s), the number of offenders involved, and in which municipality(ies) it/they happened.

After this warm-up data exploration, build the whole co-offender network. Discard the isolated nodes, thus every node will have degree ≥ 2. Given the size of the network, be careful regarding computational and memory constraints. Be sure to use sparse representations of the data whenever possible.

(e) How many nodes does the network have? How many solo offenders are there in the data set? How many (unweighted) edges does the graph contain?

(f) Plot the degree distribution (or an approximation of it if needed) of the network.

g) How many connected components does the network have?

We will now isolate the largest connected component and focus on it. This brings us down to a more manageable size.

(h) How many nodes does the largest connected component have?

(i) Compute the degree of the nodes, and plot the degree distribution (or an approximation of it if needed) for the largest connected component. Comment on the shape of the distribution.

(j) Describe the general *shape* of the largest connected component. Use the degree distribution from above, and compute statistics of the network to obtain an overview of its characteristics. You may want to consider the edge density, clustering, diameter, etc. Comment on the results.

This final section involves some free form investigation. The following parts are *optional for undergraduates.*

(k) How many crime events are executed only by young offenders?

(l) Investigate the relationship between young offenders and adult offenders. Study the structure of the crimes that include both, young and adult offenders. Discuss any patterns you observe.

(m) Ask your own question, build new separate networks if needed, and get as much insight as you like. Feel free to focus on either the whole network, or the largest connected component.