Problem Set 5

Issued: Wednesday, November 1, 2017 **Due:** Sunday, November 19, 2017

The data for this problem set was generously provided to us by Carlo Morselli (University of Montreal). This data set has not been published before 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.

Problem 5.1: [10pts] Gaining an overview of the data set

The data set may be found in `Cooffending.csv`. Additional information on the fields of the data set may be found in `DataDescription.txt`. The first step consists of getting familiar with the data set.

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

(b) How many different offenders are there?

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

(d) 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.

Problem 5.2: [30pts] The co-offender network

After this warmed-up, 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.

(a) 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?

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

(c) How many connected components does the network have?

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

Problem 5.3: [30pts] Studying the largest connected component

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

(a) 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. List the five nodes with the highest degree.

(b) Compute and plot the betweenness centrality of the nodes. List the five highest ranked nodes. Do you see a contrast in the plot? Comment.

(c) Compute and plot the eigenvector centrality of the nodes. List the five highest ranked nodes. Do you see a contrast in the plot? Comment.

(d) Compare the three lists obtained in (a), (b) and (c). Are they similar? Explain.

(e) Describe the general *shape* of the largest connected component. The computed centralities already give you some understanding of it. Compute other features or characteristics of the network to obtain a better overview of the characteristics of this network. You may want to consider the edge density, clustering, diameter, etc. Comment on the results.

Problem 5.4: [15pts] Free form investigation (optional for undergraduates)

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

(b) 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. Ask your own questions, 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.

Remark: Additional thoughts on research directions

An interesting direction of further investigation (along the lines of a class project) consists of constructing a network dual to the cooffending network. In the dual network, the nodes are the crime events and two crime events are connected by an undirected edge whenever the events share an offender. The edges can be weighted by the number of common offenders. The coffending network and its dual are correlated. One may expect clusters in the dual graph to reflect crimes of the same nature. The offenders confined to such clusters may then be thought to be *specialized* in certain crimes. Some offenders may also be seen to be weak links among different clusters, connecting different *communities* of crimes. Geographic data (e.g., municipalities) can be directly attached to the dual graph, and may lend itself to some insightful analysis. The details and outcomes of this direction are of course not clear, and ought to be refined and explored along the investigation.