Problem Set 4

Issued: Monday, November 1st, 2021

Part 1 Suggested Due Date: Friday, November 5th, 11:59 PM EST
Part 1 & 2 Due Date: Monday, November 15th, 11:59 PM EST

Part 1
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 of weight 1 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 co-cited if they are both cited by the same third paper. The edge weights in the co-citation network correspond to the number of co-citations (e.g., an edge of weight 3 relates paper A and paper B if there are 3 distinct papers that cited both of them). How do you compute the (weighted) adjacency matrix of the co-citation network from the adjacency matrix of the citation network?

(b) *Bibliographic coupling network:* Two papers are said to be bibliographically coupled if they cite the same other papers. The edge weights in a bibliographic coupling network correspond to the number of common citations between two papers (e.g., an edge of weight 3 relates paper A and paper B if their reference lists have 3 papers in common). How do you compute the (weighted) adjacency matrix of the bibliographic coupling network from the adjacency matrix of the citation network?

(c) Bibliographic coupling and co-citation can both serve as an indicator that papers deal with related material. However, they can in practice give noticeably different results. Why? Feel free to use examples to illustrate your argumentation. Which measure is more appropriate as an indicator of similarity between papers and why?

Part 2
Problem 4.2: [45pts] Co-offending Network

The data for this problem set was generously provided to us by Carlo Morselli (University of Montreal, Canada). 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, collaborations with the School of Criminology at the University of Montreal, Canada are possible and potential research findings could be published with members of their labs as co-authors.

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-offending 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(s). The weight of the edge will represent the number of crimes for which any two individuals were arrested together.

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 first create a mental plan 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`. The first part of the exercise consists of getting familiar with the data set (questions a, b, c, and d). The following questions are optional: while you are strongly encouraged to compute the summary statistics to diagnose any potential issues with data loading, there is no need to describe them in your report.

(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 per year?

(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-offending network. Discard the isolated nodes, thus every node will have degree ≥ 2 (note: by construction, solo offenders have a degree of 1, in that they only co-offend with themselves). Given the size of the network, be careful regarding computational and memory constraints. Be sure to use sparse representations of the data whenever possible. In particular, we encourage you to look into Python's dedicated *scipy.sparse* package for sparse matrices (`https://docs.scipy.org/doc/scipy/reference/sparse.html`).

(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 of the network (or an approximation of it if needed). Use a log scale for the x-axis. Does this plot exhibit a power law degree distribution?

(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 graph size.

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

(i) Compute the degree of the nodes, and plot the degree distribution for the largest connected component (or an approximation of it if needed). Again, use a log scale for the x-axis. Comment on the differences between this distribution and the degree distribution of the overall network derived in (f).

(j) Describe the general *shape* of the largest connected component. For this, you can use the degree distribution obtained in (i). In addition, you can compute relevant metrics that describe the network to obtain an overview of its characteristics. You may want to consider the following metrics: edge density, clustering, diameter, etc. Comment on the results.

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

(k) For the largest connected component, plot a homophily matrix by municipality. That is, plot the modularity between each pair of municipalities (i.e., the fraction of edges that run between the same type of nodes minus the fraction of such edges if the edges were placed at random). Comment on the patterns you observe.

(l) Produce a homophily matrix with another variable in the data set (or an interaction of multiple variables), and again comment on any patterns you observe.

(m) Ask your own question. If needed, build new separate networks. Derive as many insights as you would like. Feel free to focus on either the whole network or the largest connected component.