Problem Set 3

Issued: Wednesday, March 8th
Due Date: Friday, March 24th, 11:59 PM ET
Note: Problem 3.3 is optional for all students, we'll scale the total points accordingly.

Problem 3.1: [50pts] 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.

The data for this problem set consists of individuals who were arrested in Quebec between 2003 and 2010. Note that all individual data has been anonymized. The "Name" column is a random name chosen from existing databases of first names. 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_Data.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). They will also help diagnose potential issues with data loading.

(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 per year for 2003-2010?

(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.

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. If you are unable to perform step (g) and (h), we can provide a subgraph on which you can perform parts (i),(j), to enable you to do these. Our subgraph is not the largest connected component, and we will *deduct some points for students using our substitute for the largest connected component.*

(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.

Problem 3.2: [25pts] Detecting Disordered Speech

Psychosis due to various causes such as schizophrenia affect over 21 million people worldwide as of 2020 [1]. An important component of diagnosis is subjective assessments based on oral interviews with patients. Methods to quantify and understand speech patterns from such interviews are thus of immense value. Past research has represented the speech transcript of a patient as a directed graph, with words denoted by nodes and edges connecting consecutive words (with direction) [1]. *Mota et al.* [1] summarized various properties of the speech graph of patients.

(a) Read the paper by *Mota et al.* [1]. Beyond what is measured in the paper, suggest two quantitative measures of the speech graph that one might compute.

[1] Source: `https://www.who.int/news-room/fact-sheets/detail/schizophrenia`

(b) Research has shown that patients with speech disorders tend to repeat words (e.g., say the same word *twice* in succession). How would you detect if a patient exhibits this pattern from the adjacency matrix of their speech graph? Specify a property of the adjacency matrix you could assess to test if a given patient exhibits this pattern.

(c) Consider a case where a speech transcriptionist only measures the number of times a given word was spoken by a patient (and hence edge direction information is lost). Can you recover the directed degree (i.e., indegree or outdegree) of the nodes of the speech graph in each case? If yes, describe the procedure to recover the degree of each node. If no, what other information do you consider essential to recovering the directed degree.

Problem 3.3: [15pts] Suggesting Similar Papers (Optional)

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?

References

[1] Mota, N. B., Vasconcelos, N. A., Lemos, N., Pieretti, A. C., Kinouchi, O., Cecchi, G. A., Copelli, M., and Ribeiro, S. Speech graphs provide a quantitative measure of thought disorder in psychosis. *PloS one 7*, 4 (2012), e34928.