Assignment 1

Getting started with network analysis

Notes:

Try to answer the questions as precisely as possible. The page limit for this assignment is **3 Pages**. This page limit includes your textual answers and plots/figures (*if any*). You can submit your code/tables/appendices etc. separately or as a single markdown file. **Please upload everything as a single zip file to Canvas latest by Feb 9th.**

Q1> Constructing the network *(Total: 5 marks)*

Download any 1 network from the link below. Alternatively, you can also use your own network dataset, as long as it is a real-world one and not simulated/synthetic.

http://networkrepository.com/soc.php

You are free to download any network of your choice from this page. You can refer to the network descriptives (e.g., number of nodes, edges, average degree etc.) on the same page before choosing your network. Feel free to work with a mid-sized or large network depending on your preference.

Most of the network datasets on the page are in the form of edge lists, where each row contains a pair of node IDs representing an edge.

Q1(a): Load the downloaded edge list using your preferred tool (e.g., R) and module/package (e.g., igraph). Let's call this "*Network A*" *(1 mark)*

Q1(b): If the above network has k nodes, try to take a random sample of k/2 nodes from the list of k nodes, if k is even, or a random sample of (k+1)/2 nodes if k is odd. Let's call this set of nodes $Vsample$. Now create a network ($Vsample$, $Esample$) using this sample of nodes and ALL edges e_{ij} that have both endpoints *i* and *j* in the sampled set of nodes, $Vsample$. This is often referred to as an 'induced subgraph'. Let's call this "$Network\ A_{sample}$"*(1 mark)*

Q1(c): Compare "$Network\ A_{sample}$" with "*Network A*" in terms of any 3 graph-level measures of your choice, including but not limited to the ones we discussed in class. One **example** of a relevant graph-level measure would Graph Density. *(1 mark)*

Q1(d): Do you find any differences in the measures across the two networks? If yes, why do you observe these differences? What does this tell you about random sampling in graphs/networks? *(2 marks)*

Q2> Computing centralities *(Total: 8 marks)*

Q2(a): Very briefly define/explain the following 5 node-level measures. Also, mention one business application that might benefit from the computation of each of these measures *(be as specific as you can in explaining the how the measure can be used, use examples of actual organizations/business models if required). (5 marks)*

1. Degree
2. Closeness
3. Clustering coefficient
4. Page Rank

5. Eccentricity

Q2(b): Generate the above-mentioned measures for all nodes in "Network A" that you created in Q1. *(1 mark)*

Q2(c): Compute the Pearson's correlation among the 5 node-level measures (i.e., you should create a 5X5 correlational matrix). What do you observe from the correlation analyses in terms of the strengths of correlation between the measures? *(2 marks)*

Q3> Benchmarking your measures *(Total: 7 marks)*

A key problem in network analysis is understanding whether a particular measure that you computed is sufficiently high or low. One way to address this is to use a random graph as a comparison group.

Q3(a): Create a random graph (in R or Python) that has the same number of nodes and edges as the Network A in Q1. Feel free to make any other assumptions (e.g., directed/undirected, whether or not loops exist etc.), but describe any assumptions that you make in clear detail. We will call this $Network_{random}$. *(1 mark)*

Q3(b): Re-compute the 3 graph-level measures from Q1(c) but now using $Network_{random}$ as your network. Are the measures significantly different from the measures originally computed in Q1(c)? *(Perform a statistical test if needed, to test for significance of the difference between two sets of measures) (2 marks)*

Q3(c): Re-compute the 5 node-level measures from Q2(b) but now using $Network_{random}$ as your network. Are the measures significantly different from the measures originally computed in Q2(b)? *(Perform a statistical test if needed, to test for statistical significance of the difference between two sets of measures) (2 marks)*

Q3(d): What can you comment about the graph-level and node-level measures in Network A, based on what you found in Q3(b) and Q3(c)? ***(e.g.,** do you now think that the Network A is a highly dense network?). (2 marks)*