Final exam

Fall 2021

Analysis of Network Data Wedensday 12/07

Please do not turn the page until instructed to do so.

This exam has 4 questions. You are expected to pick 3 to work on and submit.

Every question you pick will be equally weighted.
The exam is **a take-home exam**.

It is due by end-of-day on Friday 12/16.

You are **not** allowed to collaborate with other students in the class.

You are **though** allowed to ask me questions and request hints.

Answer every question you pick to the best of your knowledge.

My office hours for the exam this week will be:

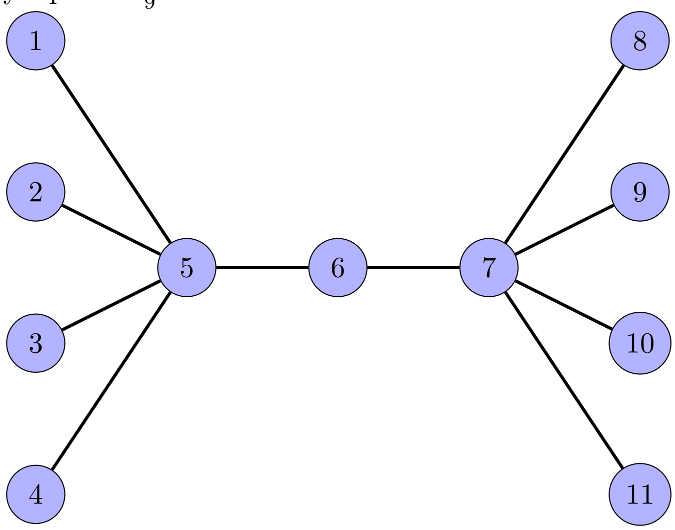
- Online office hours by request on Zoom.
- Or during the following days on Zoom:
  - Friday 12/09, Monday 12/12, and Friday 12/16 between 1pm and 3pm.

Make sure to show your work for **partial credit**.

Question 1: A version of betweenness centrality

In class, we saw the definitions of three of the most standard centrality metrics as in degree, closeness, betweenness, PageRank (eigenvector). While betweenness centrality is often very important as a global metric of centrality, it is also limited in the sense that it favors nodes that are the "center of stars". As an example, consider Figure 1, where nodes $5, 7$ have the highest betweenness centrality, but node 6 appears to be the true "central" node given its integral position for allowing communications between the two star communities.

Figure 1: An example where betweenness centrality would fail to find the arguably most critical node in the network. Nodes $5, 7$ have betweenness centrality equal to $\frac{2}{3}$, while node 6 has betweenness centrality equal to $\frac{5}{9}$.



A possible solution to this is to propose a new centrality metric which decomposes betweenness centrality for a node i in two components: (i) $b_i^{(1)}$: shortest paths between two nodes k, ℓ that are both not neighbors of i (i.e., $k, \ell \notin N(i)$); (ii) $b_i^{(2)}$: shortest paths between nodes k, ℓ where at least one endpoint is a neighbor of i (i.e., $k \in N(i)$ or $\ell \in N(i)$, or both $k, \ell \in N(i)$). For every node in the network, we may then report the betweenness centrality of a node i, b_i, as the summation of the two terms: $b(i) = b_i^{(1)} + b_i^{(2)}$.

(a) Write a code in networkx that calculates $b_i^{(1)}, b_i^{(2)}$ for every node in the Les Misérables graph.

(b) For the same graph, calculate also the betweenness centrality b_i for every node. Then, rank b_i and $b_i^{(1)}$ in decreasing order (i.e., rank the nodes from highest to lowest value). Using this rankings, calculate Kendall's τ correlation (also known as rank correlation).

Question 2: A sensor network protocol

A network $G(V, E)$ consists of sensors (V) that can serve as data collectors. Some of the sensors can communicate with other sensors in their vicinity (E). Every time a sensor collects data it passes the data on to the other sensors in the network. All sensors can send data they have collected to at least one neighbor. However, to minimize battery consumption, only a subset of the sensors is allowed to relay data that other sensors have collected. Let this set be called D: it has the following two properties:

1. any sensor in D can send data to another sensor in D using only intermediary nodes in D; and
2. any sensor in the network is either in D or adjacent to a sensor in D.

Answer the following questions.

(a) Formulate the problem of finding the set of sensors D with minimum cardinality $|D|$.

(b) A related problem can be stated as follows. Find a spanning tree T with the maximum number of leafs. We define a leaf as a node that only has one edge incident to it in the tree. Formulate this spanning tree extension as an integer program.

(c) Let $|D|$ the cardinality of the optimal set of part (a) and $|L|$ be the number of leafs from the optimal solution of part (b). Show that $|D| + |L| = |V|$.

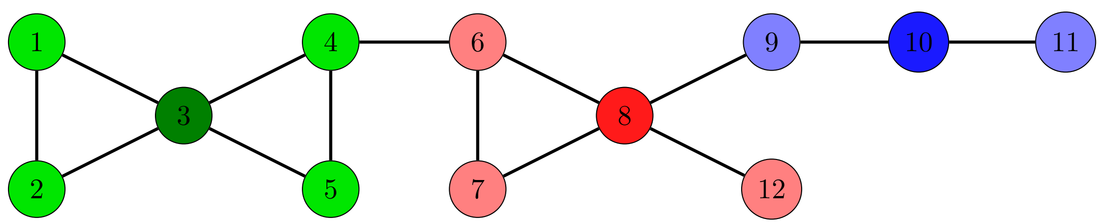
Question 3: Spectral clustering and special structures

(a) In a social network, we would like to identify "leaders". Here, we define a leader as an entity with the highest betweenness among its peers. More specifically, let b_i be the highest betweenness in the network (corresponding to node i) and let b_j be the second highest betweenness (corresponding to some node j): then, if $b_i - b_j \geq 0.4$, we say that node i is the leader of the network.

Consider the following approach. First, check whether a leader exists in the whole network. If not, then use spectral clustering to obtain a bipartition, and check whether a leader exists in the two induced subgraphs of the bipartition. If there exists a leader, report it; otherwise perform spectral clustering on each of the partitions and continue. We terminate when all leaders have been found or whether an induced subgraph has only 2 nodes.

If a leader exists in the whole network, then report that node as the leader; otherwise, if there is a leader in each of the two partitions obtained by spectral clustering, report these two nodes as the leaders. However, these are not the only possibilities. If a leader does not exist in either partition, you take that partition and you use spectral clustering again, leading to more leaders. Hence, a network could potentially have multiple leaders. As an example consider the network of Figure 2.

Figure 2: The leaders of the network below would be 3, 8, and 10. First, we partition, the network using spectral clustering clustering in two sets: $\{1,2,3,4,5\}, \{6,7,8,9,10,11,12\}$. The first partition has a leader in node 3; the second does not, so it is further partitioned into $\{6,7,8,12\}, \{9,10,11\}$. The first of these partitions has a leader in node 8 and the second in node 10.



Implement the algorithm described above using `networkx`. What are the leaders of the les misérables network and what are the leaders for the karate club network?

(b) In class, we discussed numerous clique and star relaxations, like the quasi-clique (where a fraction $\gamma \in [0,1]$ of all edges in a structure needs to exist), the k-club (where all nodes in a structure need to be within a distance of k from one another), etc. In a clique, all nodes need to be adjacent to every other node; what if we relaxed this property? What if we allowed all nodes to be connected to every other node but $k-1$ of them? That is, what if for any structure S, we want each node to be adjacent to at least $|S| - k$ nodes in the same structure?
Formulate the problem of finding the maximum cardinality such structure as an integer program and solve it for the les misérables network. What is the structure of maximum cardinality you find?

Question 4: Non-adjacent groups of nodes

Consider a setup where we want to separate the nodes into groups such that no two nodes that are adjacent are in the same group. For example, see Figure 3, where both graphs on the left and right can be separated into three groups each (let them be the red, blue, green groups). Answer the following questions.

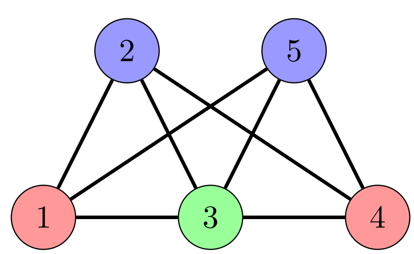
Figure 3: Two examples of graphs where the nodes can be partitioned into 3 groups each.

(a) Write a mathematical formulation for finding the smallest number of groups (such that no two nodes in the same group are adjacent) you can find. Use Gurobi and networkx to identify the number of groups in the Les Misérables network.

(b) Removing a node from a graph may or may not decrease the number of groups you can build this way. For example, removing any node in the graph on the left in Figure 3 will lead to a decrease in the number of groups (from 3 to 2) whereas removing nodes 4 or 5 in the graph on the right will not decrease the number of groups (it is still 3 even after that removal). Write a mathematical formulation to identify the k nodes you should remove to decrease the number of groups you can build in the Les Misérables network. Use $k = 1, 2, 3, 4, 5$ and report the number of groups you can originally build versus the number of groups you can build after removal.

(c) Similarly, adding an edge in a graph may or may not increase the number of groups you can build. For example, adding an edge between nodes 3 and 5 will not change the number of groups in the graph on the right in Figure 3. On the other hand, consider the graph in Figure 4, where any edge addition will lead to increasing the number of groups from 3 to 4. Write a mathematical formulation to identify the k edges you should add to increase the number of groups you can build in the Les Misérables network. Use $k = 2, 4, 6, 8, 10$ and report the number of groups you can originally build versus the number of groups you can build after adding the necessary edges.

Figure 4: One example of a graph where adding any edge will increase the number of groups.



Good luck!