CSCI 5352, Network Analysis and Modeling **Prof. Aaron Clauset**
Problem Set 3 **Fall 2021**

There are 100 regular points and 47 extra points possible on this assignment.

1. (50 pts total) *Predicting missing node labels.* In this problem, you will implement and systematically evaluate methods for predicting missing node labels (metadata), in two settings. In the real world, node labels (either categorical or scalar) may be missing for a variety of reasons, e.g., because the labels were sampled (even if the edges were not), or, for social networks, the nodes may not have disclosed their label.

   Suppose we have a graph $G = (V, E)$ where each node i has a categorical label x_i. We say that the labels exhibit *homophily*, or an *assortative* mixing pattern, if two labels x_i, x_j are more likely to be similar (or the same) if $(i, j) \in E$ than if not. In this case, we can use the "local smoothing" heuristic from the lectures to make a simple guess about any particular missing label.

   Obtain the following networks from the *Index of Complex Networks* at `icon.colorado.edu`:

   – ICON entry: "Norwegian Boards of Directors (2002-2011, projection)"
   network: `net1m_2011-08-01`
   metadata: `data_people` (gender variable)

   – ICON entry: "Malaria var DBLa HVR networks"
   network: `HVR_5`
   metadata: `metadata_CysPoLV`

   (a) (25 pts) In the first experiment, you will systematically evaluate the local smoothing heuristic as a function of how many node labels are observed, keeping the network fixed.

   - Implement the local smoothing heuristic as described in the lecture notes.
   - For each network, design and run a numerical experiment that allows you to plot the *accuracy* of the local smoothing heuristic as a function of the fraction $f \in (0, 1)$ of the empirical labels observed. For instance, if $f = 0.8$, then for each i, pretend $x_i = \emptyset$ with probability $p = 0.2$, use the "observed" labels $\{x_i \neq \emptyset\}$ and G to predict the "unobserved" labels, and then compute the accuracy.
   - Define *accuracy* as the average fraction of correct guesses.
   - Make one nice figure showing these two relationships.
   - Discuss what you learn about the local smoothing heuristic, how its performance differs between these networks, how its performance varies with f, and any insights you can gain about the structure of these networks from the shape of these curves.
   - (7 pts *extra credit*) Derive mathematically the expected accuracy for a "baseline" predictor of guessing a missing label uniformly at random, without using the graph G; then calculate that baseline value for the two networks, and comment on your numerical results in the context of this baseline.

- (10 pts *extra credit*) Add another node label predictor to the experiment; explain how it works and discuss its performance relative to the local smoothing heuristic.

Hint: to get a nice figure, for each choice of f, you will want to measure the average fraction of correct guesses over many repetitions for choosing which node labels are observed (training) and which are not (testing).

(b) (25 pts) In the second experiment, you will again evaluate the local smoothing heuristic, but now in the context of how randomized the network is. That is, we will keep the labels fixed, but vary the network structure by combining in a synthetic network some edges taken from the real network with some taken from a comparable Erdős-Rényi null model. The larger the portion we take from the latter, the more random the synthetic graph becomes. This way of "mixing" together an empirical network with a random graph is a common technique in network science for evaluating methods, since the fully random graph provides an interpretable baseline.

For this experiment, set $f = 0.8$, so that a random 20% of the node labels will be your "test" set for calculating accuracy. Then, to "randomize" the network by a variable amount, choose a value $\alpha \in (0, 1)$ and create an empty synthetic graph G_o. Then, add the non-random part by selecting $(1-\alpha)m$ edges uniformly at random from the empirical network G; then add the random part by connecting $\alpha\, m$ of the remaining unconnected pairs. (Note: the goal here is to ensure that the mean degree of G_o matches the mean degree of G.)

- For each network, design and run a numerical experiment that allows you to plot the *accuracy* of the local smoothing heuristic as a function of $\alpha \in (0, 1)$, the degree of randomization. For instance, if $\alpha = 0.2$ and G is a simple graph with $n = 100$ and $\langle k \rangle = 5$; then each G_o will contain $m_o^{\text{non-random}} = 200$ edges chosen from G and a uniformly random $m_o^{\text{random}} = 50$ of the remaining $\binom{n}{2} - m_o^{\text{non-random}}$ pairs as edges.
- Make one nice figure showing accuracy as a function of randomization α.
- Discuss what you learn about the local smoothing heuristic, how its performance differs between these networks, how its performance varies with α, and any insights you can gain about the structure of these networks from the shape of these curves.
- (5 pts *extra credit*) Identify and explain the conditions under which these two experiments should produce the same accuracy scores.

Hint: to get a nice figure, you will want to average over several draws for each α, and over a few different draws for the given f.

2. (50 pts) *Predicting missing edges.* In this problem, you will implement and systematically evaluate methods for predicting missing links, using the same two networks as in question 1.

Recall the definitions of the *degree product* and *Jaccard coefficient* link predictors from the lecture notes. To these, we add the *shortest path* predictor, defined as follows. Let $\sigma(i, j)$ be the length of a geodesic path between i and j. Then, *shortest path*: $\text{score}(i, j) = 1/\sigma(i, j) + \epsilon$, where ϵ is a small amount of random noise. (Recall that we define $\sigma(i, j) = \infty$ if there is not path from i to j.)

- Using the same two networks as in question 1, set up and run a numerical experiment in which you measure the accuracy of these three heuristics as a function of the fraction $f \in (0, 1)$ of the edges observed.
- Define *accuracy* as the AUC.
- Make one nice figure for each network, show these relationships for the 3 predictors.
- Discuss what you learn about them, how their performance differs between these networks, and any insights you can gain about the structure of these networks from the shape of these curves.
- (5 pts *extra credit*) Include the full ROC curves for these three predictors (on a single plot), and comment on their relative shapes.
- (10 pts *extra credit*) Add another link predictor to the experiment; explain how it works and discuss its performance relative to the others.

3. (10 pts *extra credit*) Reading the literature.

Choose a paper from the Supplemental Reading list on the external course webpage . Read the whole paper. Think about what it says and what it finds. Read it again, if it's not clear. Then, write a few sentences for each of the following questions in a way that clearly summarizes the work, and its context.

- What was the research question?
- What was the approach the authors took to answer that question?
- What did they do well?
- What could they have done better?
- What extensions can you envision?

Do not copy text from the paper itself; write your own summary, in your own words. (Using terminology from the paper is okay, of course.) Be sure to answer each of the five questions. The amount of extra credit will depend on the accuracy and thoughtfulness of your answers.