CSCI 5352, Network Analysis and Modeling **Prof. Aaron Clauset**
Problem Set 3 **Spring 2025**

There are 100 regular points and 75 extra points possible on this assignment.

1. (50 pts total) *Predicting missing node labels.*

   Let $G = (V, E)$ be a graph and let $\vec{x}$ be a vector of categorical node attributes. Recall that labels exhibit *assortative* mixing if labels x_i, x_j are more likely to be similar (or the same) if $(i, j) \in E$ than if not. When this is true, we can use the "local smoothing" heuristic from the lectures to make a simple unsupervised guess about any particular missing label.

   Go to the *Index of Complex Networks* at `icon.colorado.edu` and obtain the following:

   - 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) (50 pts) Design and carry out an experiment to systematically evaluate the local smoothing heuristic as a function of α, the fraction of node labels we observed for some G.

   - Implement the local smoothing heuristic as described in the lecture notes; don't forget to default to the baseline when necessary, and break ties randomly.
   - Let α vary between 0 and 1 in fixed increments, e.g., $\Delta\alpha = 0.02$, and measure the average accuracy (ACC) over some number of repetitions at each α. (Hint: more repetitions makes a smoother curve.)
   - Plot the average ACC functions for the two networks on a single nice figure.
   - Discuss: (i) How are the accuracy curves similar/different between the two networks (be sure to at least discuss the performance in the very low, mid-range, and very high α ranges)? (ii) What if anything do you learn about the local smoothing heuristic from this experiment? And, (iii) What if any insights do you gain about the structure of these networks from the shape of these curves?
   - Derive mathematically the expected accuracy for the baseline predictor (guessing a missing label uniformly at random from $\vec{x}^o$); calculate the baseline expectation for the two networks, and comment on how the results of your experiment compare to this baseline.

   (b) (25 pts *extra credit*) Design and carry out an experiment to evaluate the local smoothing heuristic as a function of how randomized the network is. That is, keep the fraction of observed labels α fixed, and instead vary the fraction β of the network's edges that are randomized. Test the hypothesis that as the network structure becomes increasingly randomized, the local smoothing heuristic's accuracy degrades toward the baseline predictor.

For this experiment, set $\alpha = 0.8$, and use the removed 20% of the node labels as the "test" set for calculating accuracy (ACC). Use the double-edge swap algorithm to parameterize the extent to which the structure is randomized by defining $\beta = r/2m$, where r is the number of double edge swaps that have been applied to the original network G.

- Decide on a set of values spanning $r \in [0, 2m]$ that show off the overall pattern of how the local smoothing heuristic's accuracy varies from $\beta = 0$ (fully empirical G) to $\beta = 1$ (fully randomized G).
- Run your experiment to measure the average accuracy for the 80/20 split of node labels as a function of β, over some number of repetitions for a particular value β. (Hint: more repetitions makes a smoother curve.)
- Make one nice figure showing accuracy vs. β curves of the two networks. Indicate on the figure the expected baseline accuracy on the original network G (e.g., with a horizontal line).
- Discuss: (i) To what degree do the accuracy curves differ between these networks? (ii) How does the accuracy curve compare to the baseline on the original graph G? And, (iii) What if any insights does this experiment give you about the structure of these networks from the shape of these curves?

2. (50 pts) *Predicting missing edges.*

Recall the definitions of the *degree product* (DP) and *Jaccard coefficient* (JC) link predictors from the lecture notes. To these, add the *shortest path* (SP) predictor, defined as follows. Let $\sigma(i, j)$ be the length of a geodesic path between i and j. Then, the SP predictor is defined as $\text{score}(i, j) = 1/\sigma(i, j) + \epsilon$, 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.)

(a) (50 pts) *Unsupervised link prediction.*

- Implement the DP, JC, and SP score functions, each of which takes as input a node pair i, j and an observed graph G', and returns its respective score for that pair.
- Define *accuracy* as the AUC, and implement a function that takes as input the completed table over the set of potential missing links X (see lecture notes), and calculates the corresponding AUC. (Hint: you can use one table with multiple score columns, one per predictor, which can then be row-sorted by a particular column to get a version you can use to calculate that predictor's AUC.)
- Design and carry out a numerical experiment, using the same two networks as in question 1, in which you measure the accuracy of these three heuristics as a function of the fraction $f \in (0, 1)$ of the edges that are observed. (You will need to write a function to generate an observed graph G', given G and choice of f.)

- Let f vary in fixed increments, e.g., $\Delta f = 0.05$, and measure the average AUC for each of the three heuristics at each f. (Hint: averaging over more repetitions at a choice of f makes for smoother curves.)
- First, make a pair of figures (one for each network) that plot the corresponding accuracy curves for the 3 predictors; include a line at AUC $= 0.5$ as a reference.
- Second, make a nice figure showing all three full ROC curves for one of your networks, for $f = 0.8$. Indicate which network you chose.
- Discuss: (i) Why does one predictor perform much better than others when f is very small, and worse than others when f is larger? (ii) What if anything does the difference in performance across the two networks tell you about how those networks' structures differ? And, (iii) what do their relative shapes of the ROC curves imply about the accuracy of these algorithms?

(b) (40 pts *extra credit*) *Supervised link prediction.*

- Implement a *stacking model* that uses the DP, JC, and SP predictors as level-0 algorithms, and uses either a random forest (RF) or logistic regression (LR) algorithm (your choice) as the supervised meta-learner for the level-1 algorithm.
- To evaluate the stacked model's accuracy, we will need to 'observe' the network in a 2-step process because we need a training/test split to train the supervised algorithm, but we also need held-out data to evaluate the trained algorithm.
  - Recall that $G' = (V, E')$ where E' is the observed set of edges (each observed with probability f). Using the observed network G', create the validation data set by calculating the level-0 algorithms for each pair $i, j \in X$, the set of possible missing links in G'. (Recall that $Y = E - E'$ tells you which pairs $i, j \in X$ are missing links.) Store the validation data as matrix S.
  - Now define the training graph $G'' = (V, E'')$, where E'' is a fraction f sample of the observed edges. Create the training data set by selecting uniformly at random c pairs from $Y' = E' - E''$ (true positives for G'') and c pairs from $X' - Y'$ (true negatives for G''); then run the level-0 algorithms for each of these $2c$ pairs, using G''. Store the training data as matrix T.
  - Train the level-1 model to predict , on the basis of the training data T, whether $i, j \in Y'$.
  - Apply the trained model to the validation data S, and evaluate its predictions relative to Y.
- Choose just one of the two networks you've used so far and make one nice figure for each network plotting AUC values for $f = \{0.5, 0.6, 0.7, 0.8, 0.9\}$. Add to this figure the corresponding AUC values for the DP, JC, and SP algorithms from question 2a, and legend indicating which is which.

- For $f = 0.8$, make a nice figure showing the three full ROC curves for DP, JC, and SP, and add a fourth line showing the ROC curve for the stacked model.
- Discuss: (i) To what extent does the stacked model's AUC improve upon, or not, the level-0 model AUCs? And, (ii) to what extent does the stacked model's ROC curve improve upon, or not, the level-0 model ROC curves?

Hint: for the training data, $c = 1000$ is a good number. However, for modest-sized networks, there may not be so many true positives; in this case, we *upsample* the true positives by sampling with replacement until we have $c = 1000$ examples.

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 paper did you choose?
   - 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.