Network Analysis and Modeling
CSCI 5352, Fall 2017
Prof. Aaron Clauset
Problem Set 5, due 11/7

1. (100 pts total) *Predicting missing information using networks.* Here, you will explore techniques for predicting missing node labels and missing edges in networks.

   (a) (40 pts) *Missing node labels.* In real-world networks where nodes have labels (either categorical or scalar), such labels may be missing for a variety of reasons. For instance, the labels may have been sampled (even if the network was not), or, for social networks, the nodes may not have disclosed their label.

   If the mixing pattern of labels is assortative, i.e., noded exhibit *homophily* with respect to this label, then we can use a simple "guilt by association" (GbA) heuristic to make a reasonable guess about any particular missing label. The GbA heuristic works as follows: for a node i with no label, guess ("impute") that its missing label is the mode of the distribution of non-missing labels observed among i's nearest neighbors (breaking ties randomly).

   Visit the *Index of Complex Networks* at `icon.colorado.edu` and obtain these files:

   - 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`

   Then:

   - For each network, set up and run a numerical experiment in which you measure the accuracy of the GbA heuristic as a function of the fraction $f \in (0, 1)$ of the empirical labels observed, chosen uniformly at random.
   - Define accuracy as the average fraction of correct guesses.
   - Make one nice figure showing these two relationships.
   - Discuss what you learn about the GbA heuristic, how its performance differs between these networks, and any insights you can gain about the structure of these networks from the shape of these curves.

   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) (60 pts) *Missing edges.* Heuristics for predicting missing edges score each possible pair $\{i, j\} \notin E$ in some way and then we say that high-scoring pairs are more likely to be missing connections than low-scoring pairs. There are any number of possible functions score(i, j); in this question, we will explore the following three: degree product (which is

related to both the preferential attachment model, and to a random graph with specified degree structure), a normalized common neighbor measure, and the shortest path.

Let $\Gamma(i)$ denote the set of neighbors of i in the network, and let $\sigma(i,j)$ be the length of a geodesic path between i and j.

- degree product: $\text{score}(i,j) = k_i k_j$
- normalized common neighbors: $\text{score}(i,j) = |\Gamma(i) \cap \Gamma(j)| \,/\, |\Gamma(i) \cup \Gamma(j)|$
- shortest path: $\text{score}(i,j) = 1/\sigma(i,j)$

Then:

- Using the same two networks as in question (1a), 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.[1]
- Make one nice figure for each network, show these relationships for the 3 heuristics.
- 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.

Hint: To ensure that ties are broken randomly, define and use instead $\text{score}'(i,j) = \text{score}(i,j) + U(0,1)/n$, where $U(0,1)$ is a uniformly random variable and n is the number of nodes.

2. (20 pts extra credit) Using your *SI* simulation from Problem Set 4, construct a new centrality measure, which we will call *spreading centrality*. We define the spreading centrality s_i of a node i to be the average size of a cascade (number of infected nodes when the epidemic is complete) that is seeded at i and when the transmission probability is $1/c$, for the network's mean degree c. Thus, the bigger the average cascade that a node i tends to produce, the more important it is under this measure.

   Choose two moderately-sized networks ($n > 500$) from the *Index of Complex Networks* and numerically calculate s for all vertices in each network. Produce a table listing the names of the top 10 nodes, ordered by their spreading centrality, and report their centrality scores and their degree, for each network. Briefly comment on what you discover, with respect to the two networks you chose.

[1] See the ROC wikipedia page: `http://bit.ly/2ehXHrb` . The AUC is mathematically equivalent to the probability that your binary classifier assigns a better score to a randomly chosen true positive (TP) than to a randomly chosen true negative (TN). When AUC= 0.5, the classifier cannot distinguish between a TP and a TN. The standard way to calculate the AUC for a given binary classifier is to first apply your score(.) function to each member x_i of the set on which you are making classifications $\mathcal{S}$. Let $y_i = 1$ if $x_i \in \mathcal{S}$ is a TP, and $y_i = 0$ otherwise. Then, construct a vector of tuples $(\text{score}(x_i), y_i)$, and sort the vector in increasing order of the first value (the score). From this data structure, it is straight forward to either tabulate the ROC curve or calculate the AUC summary statistic.