Network Analysis and Modeling
CSCI 5352, Fall 2014
Prof. Aaron Clauset
Problem Set 5, due 11/5

1. (50 pts total) When faced with an edge-weighted network, a common procedure is to "binarize" the data by applying a threshold t to each edge. If w_{ij} is the weight of edge (i, j) in the input graph, we define the binarized or unweighted adjacency matrix as

$$A_{ij} = \begin{cases} 1 & \text{if } w_{ij} \geq t \\ 0 & \text{otherwise} \end{cases} .$$

Choosing $t = \min_{ij} w_{ij}$ produces a binarized graph in which edge weights are simply ignored. As we increase t, progressively more edges are thrown out until all edges are excluded when $t > \max_{ij} w_{ij}$.

Consider an undirected complete graph with n vertices in which each edge (i, j) has a weight drawn from an exponential distribution $p(w) = \lambda e^{-\lambda w}$, with $\lambda > 0$ and $w \geq 0$. For parts (1b) and (1c), fix $\lambda = 1$ and investigate the question via numerical simulation and visualize your results one more nice figures. (No credit if you do not label your axes and data series.)

(a) (10 pts extra credit) To generate such a graph, you will need to be able to convert a continuous uniform random deviate $r \sim U(0, 1)$—which most pseudorandom number generators produce—into an exponentially distributed one. Show (via a mathematical derivation) that $x = -(1/\lambda) \log(1 - r)$ is such a variable.

Hint: If $F(x)$ is a continuous cumulative distribution function for x, and r is a uniform random deviate from $[0, 1]$, then $x = F^{-1}(r)$ is a random deviate with distribution F. This very useful trick is sometimes call the "transformation method" for deriving pseudo-random number generators.

(b) (30 pts) Fix $n = 500$ and characterize the way the component size distribution changes as a function of threshold t, over its dynamical range. Give a brief explanation of (i) how you set up and ran the experiment, and (ii) what pattern you observe.

Hint: Consider the following three variables: the threshold t, the fractional size of a component s/n, and the logarithm of the probability of observing a component of that fractional size $\log \Pr(s/n)$. A 3d-surface figure can work well, if you average your results over a sufficiently large number repetitions for each choice of t, and choose a sufficiently fine grid on t.

(c) (20 pts) Determine whether the pattern found in (1b) varies with increasing n. Give a brief explanation of (i) how you set up and ran the experiment, and (ii) why we are seeing this particular pattern.

Hint: Is this model related to another model we have studied in class?

(d) (10 pts extra credit) Derive a mathematical expression in terms of n and λ that fully explains the patterns observed in questions (1b) and (1c).

2. (50 pts) In a spatial network, each of the n vertices is assigned some position within a metric space, e.g., $\mathbf{z}_i \in \mathbb{R}^d$ where typically $d = 2$. For many real-world spatial networks, the vertices' locations are fixed and our task is to build a network that both connects them and minimizes some kind of cost function over the properties of the network, as in light-rail, subway, or air travel systems, where the locations of stations or airports are fixed but we can choose which stations to connect by transport. The cost we optimize in such situations is typically some tradeoff of the total length of all edges in the network (which we seek to minimize) against the efficiency of the network (which we seek to maximize).

   Consider the following spatial network growth mechanism. Place $n - 1$ points (vertices) uniformly at random on the unit square (i.e., $0 \leq \mathbf{z}_x, \mathbf{z}_y \leq 1$) and one point (vertex) in the exact center. Label this point as vertex 0. Now, add $n - 1$ edges, one at a time, so that each edge connects one of the still disconnected vertices to the growing network. At each time step, add the edge (i, j) that has minimum weight under the function

$$w_{ij} = d_{ij} + \alpha \left(\frac{d_{ij} + \ell_{j0}}{d_{i0}} \right) \ ,$$

   where d_{ij} is the Euclidean distance between vertices i and j, ℓ_{ij} is the distance along the shortest path in the network between i and j, and α is a free parameter. This cost function represents the sum of the length of the prospective edge (the first term) and the routing time to the center of the network (second term).

   Study the functional relationship between the *route factor*, defined as

$$q = \frac{1}{n} \sum_{i=1}^{n} \frac{\ell_{i0}}{d_{i0}} \ ,$$

   and the free parameter α in the weight function. Present your results on a single figure. Include example visualizations of the networks grown for a few values of α. Briefly discuss the implications of your findings.