Network Analysis and Modeling
CSCI 5352, Fall 2013
Prof. Aaron Clauset
Problem Set 3, due 10/7

1. (10 pts) *Claim*: the clustering coefficient of an undirected, unweighted and connected graph G may be computed exactly by running a single-source shortest-path (SSSP) tree from any vertex. Specifically, let E denote the set of edges in G, and let T denote a single-source shortest-path (SSSP) tree grown from some vertex s. Every edge in the set $E - T$, i.e., edges not in the SSSP tree, is part of a triangle, and every edge in T is part of a connected triple.

   Prove this claim true or false.

2. (30 pts) Consider the random graph $G(n, p)$ with average degree c.

   (a) (10 pts) Show that in the limit of large n the expected number of triangles in the network is $\frac{1}{6}c^3$. In other words, show that the number of triangles is constant, neither growing nor vanishing in the limit of large n.

   (b) (10 pts) Show that the expected number of connected triples in the network, as in Eq. (7.41) in *Networks*, is $\frac{1}{2}nc^2$.

   (c) (10 pts) Hence, calculate the clustering coefficient C, as defined in Eq. (7.41) in *Networks*, and confirm that it agrees for large n with the value given in Eq. (12.11) in *Networks*.

3. (30 pts) We can make a simple random graph model of a network with clustering or transitivity as follows. We take n vertices and go through each distinct trio of three vertices, of which there are $\binom{n}{3}$, and with independent probability $p = c \,/ \binom{n-1}{2}$, with c a constant, connect the members of the trio using three edges to form a single triangle.

   (a) (10 pts) Show that the mean degree of a vertex in this model is $2c$.

   (b) (10 pts) Show that the degree distribution is

   $$p_k = \begin{cases} \mathrm{e}^{-c}c^{k/2} \,/ (k/2)! & \text{if } k \text{ is even,} \\ 0 & \text{if } k \text{ is odd.} \end{cases}$$

   (c) (10 pts) Show that the clustering coefficient, Eq. (7.41) in *Networks*, is $C = 1/(2c + 1)$.

   (d) (10 pts extra credit) Show that when there is a giant component in the network, its expected size S as a fraction of network size satisfies $S = 1 - \mathrm{e}^{-cS(2-S)}$.

   (e) (10 pts extra credit) What is the value of the clustering coefficient when the giant component fills half of the network?

4. (40 pts) When faced with a weighted graph, 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 $p(w) = \lambda \mathrm{e}^{-\lambda w}$, with $\lambda > 0$ and $w \geq 0$. For $\lambda = 1$, conduct the following two numerical experiments and visualize the results. (No credit if you do not label your axes and data series.) For each experiment, give brief explanations of (i) how you set up and ran the experiment, and (ii) what properties of the network model produce the pattern you see.

- Fix $n = 500$ and characterize the way the component size distribution changes as a function of threshold t, over its dynamical range.

  Hint 1: 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. It can be shown that $x = -(1/\lambda) \log(1 - r)$ is such a variable.

  Hint 2: There are three variables you should try to display together on a single figure: the threshold t, the size of a component s, and the probability of observing a component of that size $\Pr(s)$.

  Hint 3: Because the input graph is a random variable, in order to get a clean characterization, you will need to combine the results of many independent instances of the input graph before you visualize the results.
- Determine whether the pattern found in (a) varies with increasing n.

5. (10 pts extra credit) Equation (13.74) in *Networks* tells us on average how many vertices are a distance d away from a given vertex under the configuration model.

   (a) (5 pts extra credit) Assuming that this expression works for all values of d (which is only a rough approximation to the truth), at what value of d is the average number of vertices equal to the number n in the whole network?

   (b) (5 pts extra credit) Hence, derive a rough expression for the diameter of the network in terms of c_1 and c_2, and so argue that configuration model networks display the small-world effect, in the sense that typical geodesic distances between vertices are $O(\log n)$.