Inference, Models and Simulation for Complex Systems
CSCI 7000-003, Fall 2010
Prof. Aaron Clauset
Problem Set 3, due 10/13

1. Erdös-Rényi random graphs

   (**step 1**) Implement an algorithm that generates an instance of an Erdös-Rényi random graph $G(n, p)$, where n is the number of nodes and p is the probability that a given edge (i, j) exists.

   (**step 2**) Implement an algorithm (such as those discussed in Lecture 8) that takes as input an unweighted, undirected simple network and gives as output a vector $\vec{c}$ where c_i is the index of the component to which the ith vertex belongs.

   (a) Using both your $G(n, p)$ generator and components algorithms from above, conduct a numerical experiment to demonstrate the phase transition in the size of the largest component, as a function of c, the average degree. The product of the experiment should be (i) a brief description of what you did, (ii) a figure showing both your numerical results for the size of the largest component S, measured as a fraction of n, and the theoretical prediction we derived in Lecture 9, as c goes from close to zero to well above the critical threshold at $c = 1$, and (iii) some comments about the shape and location of the transition, and its agreement with the analytic prediction.
   (Hint: choose $n > 10^3$, average your results over several graphs at a particular value of c and choose at least 50 values of c on the interval $0 \leq c \leq 5$.)

   (b) Derive analytically or measure numerically the component size distribution at (i) $c < 1$, (ii) $c = 1$ and (iii) $c > 1$. Present and comment on your results. Include a visualization of each type of network for $n = 100$.
   (Hint: For the visualization, write your own or use an existing implementation of the Fruchterman-Reingold spring-embedder algorithm. The course webpage includes links to several existing pieces of software that can do this.)

   (c) **(optional)** For $c = 5$, measure (via numerical simulation) the diameter of a $G(n, p)$ graph as a function of n. Show that the estimated diameter grows like $O(\log n)$ (a straight line on linear-log axes).
   (Hint: use a dozen or more values of n spaced logarithmically over several orders of magnitude, and average over several instances at a given value of n.)

(d) **(optional)** Recall that the configuration model constructs a random graph with a particular degree sequence. Using tools from class, study the case of random graphs with power-law degree distributions (for $2 < \alpha < 3$). Characterize the location of the high-degree vertices with respect to other low- or high-degree nodes. Present your results and discuss.
(Hint: think centrality and assortativity measures.)

(e) **(optional)** As a function of c, derive analytically or measure numerically the average time for an unbiased random walker to return to its starting location in a $G(n,p)$ network for $c > 1$. Here, unbiased means the walker chooses its next location uniformly at random from its current neighbors.

2. Mathematical exercises

   (a) Consider an undirected (connected) tree of n vertices. Suppose that a particular vertex in the tree has degree k, so that its removal would divide the tree into k disjoint regions, and suppose that the sizes of those regions are $n_1, \ldots, n_k$. Show that the betweenness centrality b of the vertex is

$$b = n^2 - \sum_{i=1}^{k} n_i^2 \ .$$

   (b) Consider a bipartite network with n_1 vertices of type 1 and n_2 vertices of type 2. Show that the mean degrees $\langle k \rangle_1$ and $\langle k \rangle_2$ of the two types are related by

$$\langle k \rangle_2 = \langle k \rangle_1 \frac{n_1}{n_2} \ .$$

   (c) An alternative definition of the *closeness centrality* is $c'_i = n \Big/ \sum_j d_{ij}$.
   Consider an undirected, unweighted network of n vertices that contains exactly two subnetworks of size n_A and n_B, which are connected by a single edge (A, B). Show that the closeness centralities c'_A and c'_B of these two vertices are related by

$$\frac{1}{c'_A} + \frac{n_A}{n} = \frac{1}{c'_B} + \frac{n_B}{n} \ .$$

   (d) What is the time complexity, as a function of the number of vertices n and the number of edges m, of the following network operations, if the network in question is stored in an adjacency list format? Give pseudo-code for each solution.

      i. Calculating the mean degree.
      ii. Calculating the median degree.
      iii. Calculating the clustering coefficient.