Inference, Models and Simulation for Complex Systems
CSCI 7000-001, Fall 2011
Prof. Aaron Clauset
Problem Set 3, due 10/11

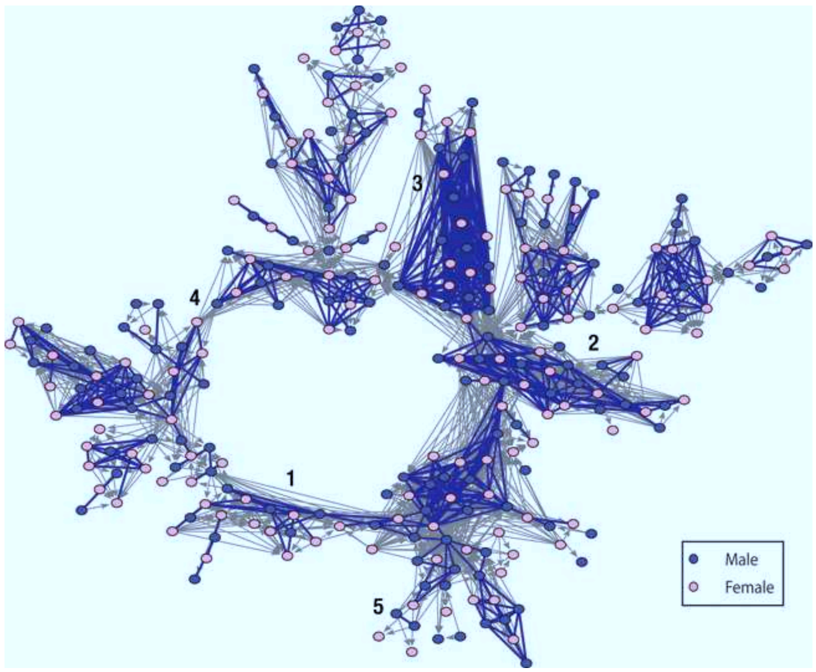

Figure 1: `http://tinyurl.com/68gx2y`

1. **The Sciences of the Artificial** (20 points)

   Read Chapter 4. Write no more than two paragraphs that clearly and concisely explain (i) what *implicit* assumptions Simon makes about learning and inference and (ii) whether, since he wrote the book, we have made any progress on the representation questions. Why is this important?

2. **Erdös-Rényi random graphs** (80 points)

   An Erdös-Rényi random graph is a type of random variable, but instead of being a scalar value like the ones we've encountered previously, it is a discrete object: a graph. A *graph* is a set of nodes or vertices V and a set of pairwise connections $E : V \times V$, i.e., each edge is an ordered pair of vertices. Let n denote the number of vertices and m denote the number of edges. An Erdös-Rényi random graph $G(n, p)$ is a graph with

n vertices where each of the $\binom{n}{2}$ possible pairwise connections (we exclude "self loops" where a vertex connects to itself) exists independently with constant probability p. That is, in $G(n, p)$ edges are iid random variables. Because edges are iid, the average degree of a vertex is simply np.

Note: A graph G can be represented as an $n \times n$ matrix A in which $A_{ij} = 1$ if $(i, j) \in E$ and $A_{ij} = 0$ otherwise. If edges are "undirected," then if $(i, j) \in E$ then $(j, i) \in E$, too; in the matrix, this means $A_{ij} = A_{ji} = 1$. In Erdös-Rényi random graphs, edges are undirected (this is why we consider only the $\binom{n}{2}$ possible unordered pairs rather than the $n^2 - n$ possible ordered pairs on the n vertices).

(a) (25 pts) Write a function that takes as input values of n and p and returns an instance of an Erdös-Rényi random graph $G(n, p)$.

For $n = 100$ and $p = 3/n$, tabulate the *degree distribution*, i.e., the distribution of nodes with k edges attached to it. Make a figure showing this distribution and overlay a Poisson distribution with parameter $\lambda = np$. No credit if you don't label your axes and provide a legend.

(Hint: to get a really nice figure, draw many instances of $G(n, p)$ and compute the average proportion of vertices with each degree value.)

(b) (55 pts) Now we'll investigate a *phase transition* in the structure of $G(n, p)$. To do this, you will need to implement an algorithm that takes as input a graph G (the output of your $G(n, p)$ function from part (a)) and finds the size of (number of vertices within) the largest component in the graph, denoted S. A *component* is a set of vertices $V' \subseteq V$ in which for all pairs $i, j \in V'$, there exists a path from i to j. (Because G is undirected, if there's a path from i to j, there must also be a path from j to i.) This can be done easily by modifying a depth-first or a breadth-first search algorithm.

Mathematically, it can be shown (which we'll do in a later lecture) that the size of the largest component in $G(n, p)$ is given by the solution to this transcendental equation:

$$S = 1 - \mathrm{e}^{-cS} \ , \tag{1}$$

where $c = np$, the average degree. The derivative of the solution function is discontinuous at $c = 1$; the function's value is effectively 0 for $c < 1$ and increases quickly from 0 to nearly 1 for $c > 1$. That is, when the average degree is less than 1, the largest component is a vanishing fraction of the size of the network; most

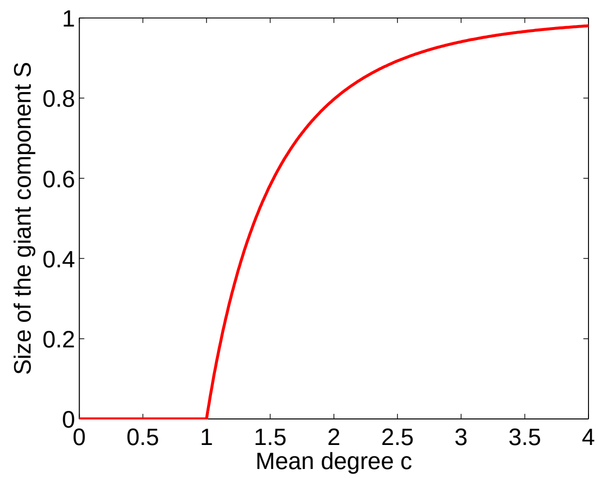

Figure 2: The phase transition in the size of the largest component in $G(n, p)$.

components are small and tree-like. When the average degree is greater than 1, most of these small components connect up forming a single *giant component* that contains nearly all of the nodes in the network. See Figure 2.

Conduct a numerical experiment to show that as the average degree np crosses the critical value of 1, the proportion of vertices in the largest component S/n exhibits the phase transition behavior. Make a figure that plots the S/n as a function of c and overlay on your empirical results the line predicted by Eq. (1).

(Hint 1: first, reproduce Figure 2 by figuring out how to find the value of S that solves Eq. (1), as a function of average degree c.)

(Hint 2: for small values of n, the empirically measured transition will be less discontinuous than Figure 2 shows. To get a good transition, make n large and average your results over many graph instances at each value of c. Also, you'll want to choose many values of c on the interval $0 < c \leq 4$ to get good resolution.)

(c) (15 pts extra credit) Make visualizations of the graphs produced by

i. $G(n = 100, p = 1/5n)$, far below the critical point,

ii. $G(n = 100, p = 1/n)$, at the critical point, and

iii. $G(n = 100, p = 5/n)$, far above the critical point.

Write a few sentences describing the qualitative differences between these graphs and how they relate to the results from part (b). No credit if you don't label which graph is which.

(Hint: use an existing implementation of the Fruchterman-Reingold (or some other) spring-embedder algorithm to get a "natural" organization of the nodes on the 2d plane. The course webpage includes links to several existing pieces of software that can do this.)