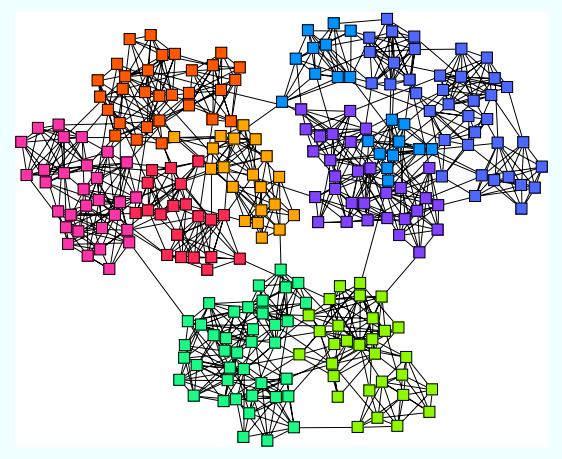
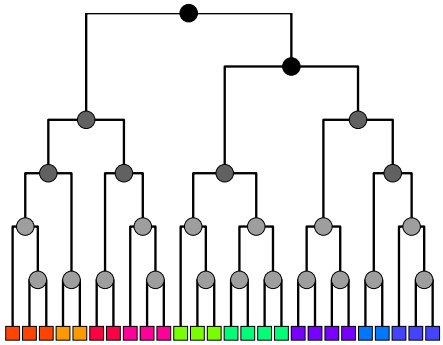
**Inference, Models and Simulation for Complex Systems**
**CSCI 7000-001, Fall 2011**
**Prof. Aaron Clauset**
**Problem Set 4, due 10/25**





For some of these problems, you will likely need to refer to our networks text *Networks: An Introduction* by M.E.J. Newman for additional contextual details, definitions and mathematical explanations.

1. **The Sciences of the Artificial** (15 points)

   Read Chapters 7 and 8. Write no more than three paragraphs that clearly and concisely discuss the following. (i) What roles do hierarchy and independence play in structuring complexity? And (ii) among complex systems, which of their aspects are likely to be predictable and which aspects are likely to be unpredictable? What does this say about the likelihood of prediction biological or social evolution?

2. **Mathematical exercises** (30 points)

   (a) (15 pts) 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) (15 pts) One definition of the *closeness centrality* is $c_i' = n \Big/ \sum_j d_{ij}$, where $d_{ij}$ is the length of the shortest path from $i$ to $j$. 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

$$\frac{1}{c_A'} + \frac{n_A}{n} = \frac{1}{c_B'} + \frac{n_B}{n} \ .$$

(c) (5 pts extra credit) 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

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

(d) (5 pts extra credit) Among all pairs of vertices in a directed network that are connected by an edge or edges, suppose that half are connected in only one direction and the rest are connected in both directions. What is the reciprocity $r$ of the network?

(e) (10 pts extra credit) For a directed network in which in- and out-degrees are uncorrelated, show that it takes time $O(m^2/n)$ to calculate the reciprocity of the network. Why is the restriction of uncorrelated degrees necessary? What could happen if they were correlated?

3. **Modular graphs** (30 points)

Given a graph, a *partition* of the graph is a division of its vertices into groups such that every vertex belongs to one and only one group. The number of groups $k$ is naturally bounded on the interval $[1, n]$. If we are looking for modules in the graph, we can score each possible partition of the network according to how well it places edges within the groups. A "good" partition should define groups with "large" internal densities, while a "bad" partition should define groups with "few" internal connections. One score function that satisfies these requirements is the Newman-Girvan *modularity* $Q$. Given a partition, its modularity score can be calculated as

$$Q = \sum_i \left[ \frac{e_i}{m} - \left( \frac{d_i}{2m} \right)^2 \right] \ , \tag{1}$$

where $i$ indexes the groups (modules), $e_i$ counts the number of (undirected) edges within the $i$th module and $d_i$ is the total degree of vertices within the $i$th module. Thus, the first term $e_i/m$ measures the fraction of edges within module $i$ and $(d_i/2m)^2$ is the expected fraction under a random graph with the same degree distribution.

(See additional discussion in *Networks: An Introduction*.)

(a) (10 pts) Consider a "line graph" consisting of $n$ vertices where each vertex connects to exactly two others, in a line, except for the two end points, which have degree one. That is, a network of $n$ vertices and $m = n - 1$ edges whose diameter is $m$. Show that if we divide this network into any two contiguous groups, such that one group has $r$ connected vertices and the other has $n - r$, the modularity takes the value

$$Q = \frac{3 - 4n + 4rn - 4r^2}{2(n-1)^2} \ .$$

(b) (10 pts) Again considering the line graph, show that when $n$ is even, the optimal division, in terms of modularity $Q$, is the division that splits the network exactly down the middle, into two parts of equal size.

(c) (10 pts) Now consider a "ring graph" made of $k$ cliques, each containing $c$ vertices, arranged in circle, where each clique is connected by one edge to each its two neighbors. Let each edge have unit weight; let $k$ be an even number; let $P_1$ be a partition with $k$ groups where each group contains exactly one of the $k$ cliques; and let $P_2$ be a partition with $k/2$ groups where each group contains one pair of adjacent cliques.

Derive an expression for the difference in modularity scores $\Delta Q = Q_2 - Q_1$ and show that this difference is positive whenever $k > 2\left[\binom{c}{2} + 1\right]$. This is the so-called *resolution limit* of the modularity function, which says that at some size of the network, merging smaller module-like structures—here, the cliques—becomes more favorable under the modularity function than keeping them separate. Thus, finding the partition that maximizes $Q$ will miss these small structures.
(Hint: for each partition, begin by writing expressions for $e_i$ and $d_i$ for a group.)

(d) (10 pts extra credit) Again consider the ring graph, but now connect each clique to every other clique with edges of weight $2/(k-1)$. Derive an expression for the difference in modularity scores $\Delta Q = Q_2 - Q_1$. (To generalize $Q$ to a weighted graph, let $e_i$ be the total edge weight within group $i$, let $m$ be the total edge weight in the graph, and let $d_i$ be the total weight of edges with an endpoint in

group $i$.) For what value of $k$ is this expression positive? Briefly discuss what this result means for the resolution limit.
(Hint: for each of the two partitions, again start with expressions for $e_i$ and $d_i$.)

(e) (5 pts extra credit) Suppose that we represent our graph $G$ using an adjacency list data structure and that we represent a partition of the graph as a vector $\rho$, of length $n$, such that $\rho_i$ gives the index of the module to which the $i$th vertex belongs. In big-$O$ notation, how long does it take to compute $Q$, given $G$ and $\rho$?

(f) (15 pts extra credit) The *stochastic block model* is a generative model for modular random graphs. It is defined by a $k \times k$ module matrix $p$, where $p_{ij}$ is the probability that a randomly selected vertex in module $i$ and a randomly selected vertex in module $j$ are connected. Thus, the connective between a pair of modules $i \neq j$ is simply a random bipartite graph with density $p_{ij}$ and the connectivity within a given module $i$ is an Erdős-Rényi random graph with density $p_{ii}$. Let $\eta_i$ denote the number of vertices in the $i$th module (and thus $\sum_i \eta_i = n$).

Derive a mathematical expression for the expected degree distribution of the entire graph, given the matrix $p$ and the vector $\eta$.

(g) (15 pts extra credit) The *hierarchical random graph* model is a generalization of the stochastic block model. Instead of assuming a single level of modular structure, in the HRG model, modules may themselves be composed of modules or be contained within other modules, in a nested fashion. Mathematically, we represent the hierarchical structure using a binary tree $\mathcal{D}$ with $n$ leaves and $n-1$ internal branch points. At each of these branch points (including the root), we place a probability $p_r$ that gives the probability that a randomly selected vertex in its left subtree is connected to a randomly selected vertex in its right subtree.

Assume that $\mathcal{D}$ is a balanced binary tree and that $n$ is some power of 2. Assume that $p_r$ is an exponentially increasing function with the distance from the root such that it has value $p_\epsilon$ at the root and value 1 at the bottom-most level. Derive a mathematical expression for the expected degree distribution of the entire graph.

(h) (15 pts extra credit) For the same model as in part (f), derive a mathematical expression for the expected clustering coefficient $C$ for the entire graph.

4. **Data analysis** (25 points)

   Download the "Political blogs" and "Coauthorships in network science" data sets from Mark Newman's website:

   `http://www-personal.umich.edu/~mejn/netdata/`

   These networks are in the GML format, which is a kind of markup language for graph structures. (If necessary, convert each network to be a simple graph: undirected, unweighted, no self-loops and no multi-edges.) For each data set, provide the following. Include a brief discussion of your results.

   - (10 pts total) A good visualization of each network. ("Good" means that the nodes do not overlap and, to a large extent, neither do edges.)

     (Hint: The Java program yEd (linked from the class webpage) can understand the GML format and has a version of the Fruchterman-Reingold spring embedder under the submenu Layout $\rightarrow$ Organic. Other visualization tools will also suffice.)

   - (15 pts total) For each network, a table giving the names of top ten vertices, in order of their closeness centrality (as defined in *Networks: An Introduction*), each with their closeness score and their degree.

   - (15 pts extra credit) A figure showing the vertex closeness centrality as a function of vertex degree (this may look better on log-log axes). Overlay on this data a line showing the same for a random graph with the same degree sequence as the empirical network (à la the configuration model; see the textbook), but averaged over several instances.