Network Analysis and Modeling
CSCI 5352, Fall 2013
Prof. Aaron Clauset
Problem Set 5, due 11/4

1. (50 pts) Consider Price's model of a citation network (Chapter 14.1 of *Networks*), applied to publications in a single field.

   (a) (10 pts) Implement the simulation algorithm described in Chapter 14.1.1. Choosing $c = 3$, $a = 1.5$ and $n = 10^6$, produce a figure showing the complementary cumulative distribution function for in-degree.

   (b) (30 pts) Now use your numerical simulation to recreate Figure 14.4 from *Networks*, which shows the average in-degree of vertices as a function of their rescaled time of creation.

   (c) (10 pts) Reasonable values of the model parameters for real citation networks are $c = 20$ and $a = 5$. For these choices, use your numerical simulation to calculate (i) the average number of citations to a paper (in-degree) in the first 10% published and (ii) the average number for a paper in the last 10%. Briefly discuss the implications of your results with respect to the "first-mover advantage," and the corresponding bias in citation counts for the first papers published in a field.

   Hint: Choose a reasonably large value of n.

2. (50 pts) The modularity function (Eqs. (7.69) and (7.76) in *Networks*) provides a simple but principled measure of the degree to which edges fall within specified groups, controlling for vertex degrees. We can choose these groups using the network structure itself by searching over partitions of vertices to find one (or several) that maximize the corresponding modularity score Q.

   The Kernighan-Lin heuristic is a simple method for searching over partitions. To begin, let η_i give the group label of vertex i, and choose a random partition of the vertices into K groups. The algorithm proceeds in rounds. At the beginning of a round, we mark each vertex as being "unswapped." During a round, we repeatedly choose a uniformly random edge (i, j) such that (i) $\eta_i \neq \eta_j$, i.e., the endpoints are in different groups, and (ii) both vertices are currently unswapped. If we can choose such an edge, we swap the group labels between these vertices, mark both as swapped, and record the partition. When no such edge exists, we then compute Q for each of the stored partitions, and choose the partition with the largest score to be the initial state for the next round.

   - Using the PS5 network file on the class website (an undirected, simple network) and choosing $K = 2$ groups, implement and run the Kernighan-Lin heuristic to find a high-scoring partition. Let the algorithm run for $O(n)$ rounds.
   - Make a figure showing the modularity Q as a function of the number of partitions stored. (If each round stores on average ℓ partitions and you run the algorithm for $3n$ rounds, there should be $3n\ell$ values in this data series.) On top of this function, draw a series of vertical lines to demarcate the rounds. Briefly discuss the efficiency of this algorithm at finding high-scoring partitions.

- Make a figure showing the network where vertices are labeled according the to final partition. Briefly comment on how well this partition manages to find groups in this network.

3. (50 pts total extra credit) Consider a model of a growing directed network similar to Price's model, but without preferential attachment. That is, vertices are added one by one to the growing network and each has c outgoing edges, but those edges now attach to existing vertices uniformly at random, without regard for degrees or any other vertex properties. It can be shown that the in-degree distribution for this model follows an exponential form $p_q = C\mathrm{e}^{-\lambda q}$, where C is a normalization constant and $\lambda = \ln(1 + 1/c)$.

   (a) (10 pts extra credit) Derive a maximum likelihood estimator for λ. Show your work. Hint: Remember that q is a discrete variable.

   (b) (30 pts extra credit) Simulate the uniform attachment model for $n = 10^6$ and apply your maximum likelihood estimator for λ to the final in-degree sequence. Report your estimated value $\hat{\lambda}$ and compare it to the theoretical value λ.

   (c) (10 pts extra credit) Make a figure showing the empirical in-degree distribution from your simulation and its theoretical form. Briefly discuss how this distribution compares to the expected distribution for an equivalent Erdős-Rényi random graph.

4. (15 pts extra credit) Consider a "line graph" consisting of n vertices in a single component with diameter $n - 1$ composed of $n - 2$ vertices with degree 2 and 2 vertices with degree 1. Show mathematically 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 Q takes the value

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

5. (15 pts extra credit) 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.