Network Analysis and Modeling
CSCI 5352, Fall 2013
Prof. Aaron Clauset
Problem Set 6, due 11/18

1. (100 pts total) Recall that the Kernighan-Lin heuristic begins with a random partition of the vertices into k groups and proceeds in rounds, iteratively choosing a pair i, j with different labels, neither of which has yet been swapped in this round, and swapping their labels. At the end of a round, when no more swappable pairs remain, the heuristic chooses the best-scoring partition from within the round and begins a new round at that state.

   (a) (50 pts) Let a $(j, n-j)$-partition denote a division of the network into $k = 2$ groups, one containing j vertices and the other containing the remaining $n - j$ vertices. Because the KL heuristic swaps the labels of a pair of vertices, it can only explore $(j, n-j)$-partitions for a given choice of j. However, we can explore the full range of group sizes by running the KL heuristic for each value of j and then taking the best partition across these different values for a given round.

   Use the following KL-based algorithm to fit the stochastic block model with $k = 2$ to the karate club network.

```
for each of j=1 to n/2,
  P(j,0) = choose a random (j,n-j)-partition
  L(j,0) = log-likelihood of P
  for t=1 to O(n) rounds
     initialize KL heuristic with P(j,t-1) and L(j,t-1)
     run the KL heuristic until no swappable pairs remain
     [L(j,t),P(j,t)] = [best log-likelihood, corresponding partition]
  end
end
for t=1 to O(n) rounds
   bestL(t) = maximum over all j for L(j,t)
   bestP(t) = corresponding P
end
```

   Make (i) a figure showing the best log-likelihood, averaged over several runs of the above algorithm, as a function of t, the number of rounds considered, and (ii) a figure showing the best-scoring partition at the end of the algorithm. Discuss your results with respect to the social division (which is given in the PS6 data file).

   (b) (50 pts) Now suppose that we know the "true" labels of several vertices, and want to estimate the remaining unknown group labels conditioned on what we know. This situation can arise because we have spent some resources to measure the latent variables for some vertices, and now we want to make educated, model-based guesses as to the remaining values.

   To investigate this idea, use the same algorithm from part (a) to fit the SBM with $k = 2$ to the karate club. Now, however, fix the labels of the five vertices with highest degree to

be their "true" values according to the social division. (You can do this by permanently marking these vertices as already swapped.) Make the same figures as in (a) and discuss any differences in the resulting partition and log-likelihood scores.

2. (30 pts extra credit) Suppose that we restrict the stochastic block matrix M so that all diagonal elements have the same value p_{in}, and all off-diagonal elements have a different value p_{out}. Furthermore, let the number of vertices n_i with a particular label i be a random variable with a geometric distribution of the form $p_{n_i} = C\mathrm{e}^{-\lambda n_i}$, where C is a normalization constant and λ is a parameter. These assumptions reduce the SBM to a 3-parameter model, given a choice of k.

   Derive a mathematical expression in terms of k, p_{in}, p_{out} and λ for the expected degree distribution of the entire graph. Show your work and then show a figure comparing your theoretical expression to the empirical distribution for networks generated from this model.

   Hint: Start with the expected degree for a particular vertex.

3. (60 pts extra credit) In a spatial network, each of the n vertices is assigned some position within a metric space, e.g., $\mathbf{z}_i \in \mathbb{R}^N$ where typically $N = 2$. In many such networks, the vertices' locations are fixed and our task is to build a network that both connects them and minimizes some kind of cost function over the properties of the network. For instance, in an airline network, the locations of airports are fixed but we can choose which airports to connect by flights. Cost functions are typically some tradeoff of the total length of all edges in the network (which we seek to minimize) against the efficiency of the network (which we seek to maximize).

   Consider the following spatial network growth mechanism. Place $n-1$ points (vertices) uniformly at random on the unit square (i.e., $0 \leq \mathbf{z}_x, \mathbf{z}_y \leq 1$) and one point (vertex) in the exact center. This point is vertex 0. Now, add $n-1$ edges, one at a time, so that each edge connects one of the still disconnected vertices to the growing network. At each time step, add the edge (i, j) that has minimum weight under the function

$$w_{ij} = d_{ij} + \alpha \left(\frac{d_{ij} + \ell_{j0}}{d_{i0}} \right) \ ,$$

   where d_{ij} is the Euclidean distance between vertices i and j, ℓ_{ij} is the distance along the shortest path in the network between i and j, and α is a free parameter. This cost function represents the sum of the length of the prospective edge (the first term) and the routing time to the center of the network (second term).

   Study the functional relationship between the *route factor*, defined as

$$q = \frac{1}{n} \sum_{i=1}^{n} \frac{\ell_{i0}}{d_{i0}} \ ,$$

   and the free parameter α in the weight function. Present your results on a single figure. Include example visualizations of the networks grown for a few values of α.