Network Analysis and Modeling
CSCI 5352, Fall 2014
Prof. Aaron Clauset
Problem Set 4, due 10/22

1. (100 pts total) The Kernighan-Lin (KL) heuristic is a classic method in graph partitioning, and it can be adapted to optimize any partition score function, e.g., modularity Q or the stochastic block model's likelihood function. The variation we will use here works as follows.

   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 pair i, j such that (i) $\eta_i \neq \eta_j$, i.e., the endpoints are in different groups, and (ii) both vertices are currently marked as unswapped. If we can choose such a pair, we swap the group labels between these vertices, mark both as swapped, and record the partition. When no such pair exists, we then compute the score for each of the stored partitions, and choose the partition with the largest score to be the initial state for the next round. Rounds continue until no improvement is possible.

   (a) (40 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.

   Implement and apply the following KL heuristic to fit the simple stochastic block model with $k = 2$ using 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
bestL = maximum over all j for L(j,t)
bestP = corresponding P
```

   Using this algorithm, do the following:

   - Make a figure showing the best log-likelihood as a function of t the number of rounds considered, averaged over several runs of the above algorithm. Report the log-likelihood of the best-scoring partition.
   - Visualize the best-scoring partition found at the end of the algorithm. Interpret your results with respect to the social division, and describe what pattern the SBM is capturing.

(b) (20 pts) Now suppose that we are given the "social labels" of some but not all vertices, and we want to estimate the remaining unknown labels conditioned on what we know. This situation could arise because we have spent some resources to measure the unknown labels for some vertices, and we want to make educated, model-based guesses about the unknown values.

Using the KL heuristic from question (1a), fit the SBM with $k = 2$ to the karate club while fixing the labels of the five vertices with highest degree to the value given in the social division. (This can be achieved by permanently marking these vertices as already swapped.) Make the same figures as in (1a); then identify the differences in both the partition and the log-likelihood score relative to the results in (1a), and explain why we see these differences.

(c) (20 pts) Now consider a simple two-group SBM with $n_1 = 30$ and $n_2 = 20$, where p_{in} gives the density of edges within each block and p_{out} gives the density between blocks.

Using $p_{\text{in}} = 0.2$ and $p_{\text{out}} = 0.05$ (strongly assortative groups), first generate a large number of networks drawn from this SBM. Then, using this ensemble of networks and your KL heuristic, estimate the distributions of $\hat{p}_{\text{in}}$ and $\hat{p}_{\text{out}}$.[1]

Show a visualization of one such network, with the groups labeled. Then, show a visualization of the estimated parameter distributions and discuss the degree to which these agree with the population values.

(d) (20 pts) Using the same model as question (1c), now let $n = 500$ and $p_{\text{in}} = p_{\text{out}} = 2/(n-1)$, which is a model with no real community structure and mean degree $\langle k \rangle = 2$. What happens when we apply our KL heuristic to these networks? Illustrate and discuss.

2. (10 pts extra credit) Consider a "ring graph" made of k cliques, each containing c vertices, arranged in circle, where each clique connects to each its two nearest neighbors via single edge. 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 the number of edges with both endpoints in group i and d_i the number of edges with at least one endpoint in group i.

3. (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

[1] "Hatted" variables like $\hat{\theta}$ denote estimates from data; unhatted variables like θ denote true or "population" values of those parameters.

constant and λ is a parameter. These assumptions reduce the SBM to a 3-parameter model, given a choice of k.

Derive as simple mathematical expression as possible 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.