Network Analysis and Modeling
CSCI 5352, Fall 2016
Prof. Aaron Clauset
Problem Set 4, due 10/18

1. (100 pts total) *The impact of community structure on spreading processes.* In this question, you will explore via a numerical experiment the impact that community structure has on the dynamics of a spreading process.

   (a) (10 pts) The "planted partition" model is a one-parameter version of the stochastic block model that generates *simple* synthetic networks with community structure of varying strength. Let n be a large and even number of vertices, let every vertex have a constant mean degree $c \geq 0$, and let $q = 2$ be the number of equal-sized communities in the model. If we define the probability of an edge existing within a group as $p_{\text{in}} = c_{\text{in}}/n$ and the probability of an edge existing between two groups as $p_{\text{out}} = c_{\text{out}}/n$, then the identity $2c = c_{\text{in}} + c_{\text{out}}$ is implied.

   Derive expressions for p_{in} and p_{out} in terms of constants, c, n, and the parameter $\epsilon = c_{\text{in}} - c_{\text{out}}$ alone, and hence show that this is a one parameter model.

   (b) (40 pts) As discussed in class, a simple SI spreading process is one in which the probability that a vertex in the infected state I transmits its infection to a particular uninfected neighbor is a constant p, and we flip that coin at most once for that edge over the lifetime of the epidemic. Initially, at time $t = 0$, all vertices are in the uninfected state S ("susceptible"), time proceeds in discrete steps and vertex state updates are applied simultaneously when moving from $t \to t + 1$. A simulated epidemic is deemed *complete* when no new infected nodes are produced in a time step. The epidemic's *size* s is the fraction of nodes in the I state when the epidemic is complete, and its *length* ℓ is the time $t = \ell$ at which the last node in the epidemic became infected. To begin the epidemic, at time $t = 1$, choose a node uniformly at random to infect.

   Using the planted partition model, you can generate any number of synthetic networks to use as a substrate for studying the behavior of the above simple SI spreading process. Using $n = 1000$, mean degree $c = 8$, and $\epsilon = 0$, measure the average epidemic size $\langle s \rangle$ and length $\langle \ell \rangle$, as a function of $p \in [0, 1]$.

   Make two figures, showing these measured relationships; on the length figure, include a horizontal line showing $\langle \ell \rangle = \log(n)$. Comment on the qualitative behavior of these measured relationships as a function of the transmission probability. Discuss the results relative to your expectations. Identify any special values of p.

   Hint: To get good figures, you will want smooth functions, which means averaging the measured s and ℓ over multiple draws from your data generating process, which in this case is the network generator *and* the simulation. You will also want to vary p smoothly enough to get good resolution on how the average epidemic size changes, especially in the regions of p where s or ℓ are changing quickly.

(c) (50 pts) Now use the planted partition model to explore how the behavior of the epidemic varies with the "strength" of the community structure, quantified by ϵ. Set $n = 200$ and mean degree $c = 8$. Consider different choices of p, and for each measure the average size s and length ℓ of an epidemic as a function of $\epsilon \in [0, 16]$.

When you have found something interesting, relative to the expectations you built up from part (b), make two good figures showing the measured relationships for that or those values of p. Discuss the qualitative behavior of these functions, how they compare to your results from part (b), and what your results here indicate about how stronger or weaker community structure influences the size and length the epidemics under this simple model.

A small amount of extra credit if your discussion covers how the transmission rules for this simple epidemic model relate to your results.

2. (10 pts extra credit) *The "resolution limit" for modularity maximization.* Consider a "ring graph" made of k cliques, each containing c vertices, arranged in circle, where each clique connects to each of 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.