Network Analysis and Modeling
CSCI 5352, Fall 2019
Prof. Dan Larremore
Problem Set 4, due 10/23

1. (70 pts total) *The impact of community structure on spreading processes.* In this question, you will explore via a numerical simulation 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. In this setting "varying strength" means how distinct the communities are—strong community structure means that most edges fall within the communities and few between, while weak community structure means edges are nearly equally distributed within or between groups.

   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 only constants, c, n, and the parameter $\epsilon = c_{\text{in}} - c_{\text{out}}$, and hence show that this is a one parameter model.
   - Generate and create simple visualizations (using an off-the-shelf spring-embedding algorithm) of three graphs, for $n = 50$, $q = 2$, $c = 5$, and $\epsilon = \{0, 4, 8\}$. Comment on the strength of the community structure each of these graphs exhibits.

   (b) (30 pts) Consider the following rules for a simple discrete time SI spreading process—one which we did *not* explicitly describe in class: the probability that a vertex in the infected state I transmits its infection to a particular uninfected neighbor is a constant p, and we evaluate that transmission possibility 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 from part (1a), 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 $n = 1000$, $c = 8$, and $\epsilon = 0$, measure the average epidemic size $\langle s \rangle$ and epidemic 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 $\langle \ell \rangle = \log(n)$. On both figures, include a vertical line at the "critical value" of p. (A critical value p_{crit} is where a phase transition occurs.)

- Comment on the qualitative behavior of these measured relationships as a function of the transmission probability p, and discuss your results relative to your expectations.
- Give an estimate p_{crit} and an intuitive explanation of why this value is special.

Hint 1: 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 (I used 500 repetitions, which took about 60 CPU minutes), which in this case is the network generator *and* the simulation. You will also want to vary p slowly enough over the $[0, 1]$ interval to get good resolution on how the average epidemic size changes, especially in the regions of p where s or ℓ are changing quickly.

Hint 2: Think carefully about where in $p \in [0, 1]$ the most interesting dynamics will occur. Recall that the critical value or "epidemic threshold" for a spreading process is a mean degree of 1 in the transmission graph.

(c) (30 pts) Now use the planted partition model of part (1a) to investigate whether the strength of community structure enhances, limits, or has no effect on epidemic size s and/or epidemic length ℓ. Let $n = 200$ and $c = 8$, and consider various combinations of the two parameters: $p \in [0, 1]$ and $\epsilon \in [0, 2c]$.

- Present your characterization of how community structure strength ϵ impacts epidemic size s and/or epidemic length ℓ using one or more figures that show the relationship clearly (smooth functions).
- Discuss the qualitative shape of these functions, and how they contrast, if at all, with your results from part (1b).
- Provide a brief intuitive explanation for why community structure strength does or does not impact the shape of the epidemic. (A small amount of extra credit if your discussion covers how the transmission rules for this simple epidemic model relate to your results.)

Hint: Use a non-uniform spacing for choices of ϵ, and note that ϵ does not need to be an integer. I averaged over 2000 repetitions, which took approximately 220 CPU minutes, to produce two 3d plots.

2. (30 pts total) *The impact of long-range edges on spreading processes.* In this question, you will explore via a numerical simulation the impact that long-range links have on the dynamics of a spreading process.

(a) An $n \times n$ grid or lattice is a very simple network, in which each node, except for those on the boundary, connects to its neighbors above, below, left, and right of itself. Using the same SI model from question (1), investigate how the epidemic size s and epidemic length ℓ vary as a function of p, for $n = 50$. Present two figures showing these relationships and provide a brief interpretation of their shape, based on the way the SI model works and the shape of the network.

(b) If a network can be embedded in a metric space, as can a grid or lattice, a "long-range" connection is one that connects two vertices that are separated by many steps on the lattice, i.e., two nodes that are "far" apart. Let q represent the probability that some pair of nodes i, j are connected by a long range link. (Note: these links exist independently

of the lattice edges.) Investigate how the epidemic size s and epidemic length ℓ depend on the variables q and p. Present your results clearly, and include a brief interpretation of how the long-range links change the epidemic dynamics relative to your results in part (2a).

3. (15 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.