Network Analysis and Modeling
CSCI 5352, Fall 2018
Prof. Dan Larremore
Problem Set 3, due 10/10

1. (15 pts) In a survey of couples in the city of San Francisco in 1992, Catania et al. recorded, among other things, the ethnicity of interviewees and calculated the fraction of couples whose members were from each possible pairing of ethnic groups. The fractions were as follows: Assuming the couples interviewed to be a representative sample of the edges in the undirected

		Women				
		Black	Hispanic	White	Other	Total
Men	Black	0.258	0.016	0.035	0.013	0.322
	Hispanic	0.012	0.157	0.058	0.019	0.246
	White	0.013	0.023	0.306	0.035	0.377
	Other	0.005	0.007	0.024	0.016	0.052
	Total	0.288	0.203	0.423	0.083	

network of relationships for the community studied, and treating the vertices as being of four types—black, hispanic, white, and other—calculate the numbers e_{rr} and a_r that appear in Eq. (7.76) in *Networks* for each type. Hence calculate the modularity Q of the network with respect to ethnicity. What do you conclude about homophily in this community?

2. (20 pts total) Consider an undirected "line graph" consisting of n vertices in a single component, with diameter $n-1$, and composed of $n-2$ vertices with degree 2 and 2 vertices with degree 1.

   (a) 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

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

   (b) Considering the same 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.

3. (30 pts) Implement the greedy agglomerative algorithm described in the lecture notes for maximizing modularity on an unlabeled simple network. (There is no need to make your algorithm particularly efficient, as we will not apply it to large networks; thus, it is okay to compute ΔQ using the adjacency matrix to derive the e matrix at each step.)

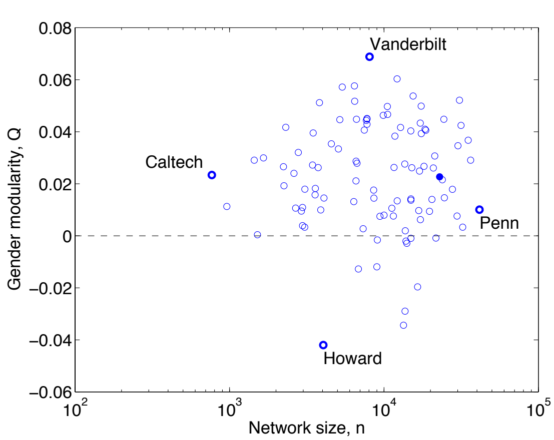
   Visit the Index of Complex Networks (ICON) at `icon.colorado.edu`, and obtain from the "Zachary Karate Club" entry a copy of the 77-edge network data file, along with the associated file that gives the "social partition" of the nodes. Apply your algorithm to this network.

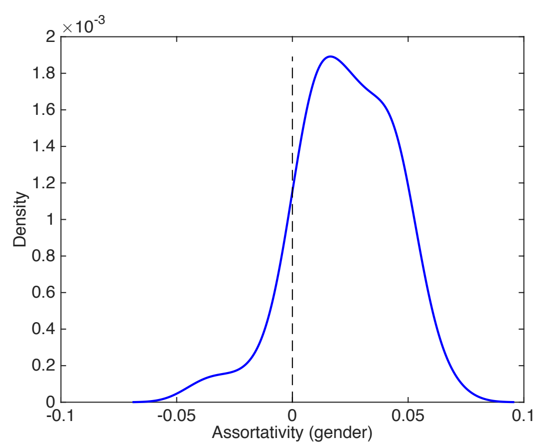
- Make a plot showing the modularity score Q as a function of the number of merges.
- Make a visualization of the network itself with vertices labeled according to your maximum modularity partition.
- Then calculate and report the normalized mutual information (NMI) between your partition and the social partition.[1]
- Finally, briefly discuss the agreement or disagreement between the two partitions, and what that agreement/disagreement implies about the utility of modularity maximization inferring good partitions without knowing such labels.

4. (35 pts) Using the FB100 networks, investigate the assortativity patterns for three vertex attributes: (i) student/faculty status, (ii) major, and (iii) vertex degree. Treat these networks as simple graphs in your analysis.

   For each vertex attribute, make a scatter plot showing the assortativity versus network size n, on log-linear axes, for all 100 networks, *and* a histogram or density plot showing the distribution of assortativity values. In both figures, include a line indicating no assortativity. Briefly discuss the degree to which vertices do or do not exhibit assortative mixing on each attribute, and speculate about what kind of processes or tendencies in the formation of Facebook friendships might produce this kind of pattern.

   For example, below are figures for assortativity by gender on these networks. The distribution of points spans the line of no assortativity, with some values nearly as far below 0 as others are above 0. However, the gender attributes do appear to be slightly assortative in these social networks: although all values are within 6% in either direction of 0, the mean assortativity is 0.02, which is slightly above 0. This suggests a slight amount of homophily by gender ("like links with like") in the way people friend each other on Facebook, although the tendency is very weak. In some schools, we see a slight tendency for heterophily ("like links with dislike"), as one might expect if the networks reflected heteronormative dating relationships.





[1]For details of how to do this calculation, see Equation (11) in Karrer, Levina, and Newman, "Robustness of community structure in networks." *Phys. Rev. E* **77**, 046119 (2008), which is available here `http://arxiv.org/abs/0709.2108`.

5. (10 pts extra credit) As described in Section 13.2 of *Networks*, the configuration model can be thought of as the ensemble of all possible matchings of edge stubs, where vertex i has k_i stubs. Show that for a given degree sequence, the number Ω of matchings is

$$\Omega = \frac{(2m)!}{2^m m!} \ ,$$

which is independent of the degree sequence.

6. (15 pts extra credit) Using the configuration model, investigate the set of random graphs in which all vertices have degree 1 or 3.

   - Calculate via computer simulation the mean fractional size of the largest component for a network with $n = 10^4$ vertices, and with $p_1 = 0.6$, $p_3 = 1 - p_1$, and $p_k = 0$ for all other values of k.
   - Now make a figure showing the mean fractional size of the largest component for values of p_1 from 0 to 1 in steps of 0.01. Show that this allows you to estimate the value of p_1 for the phase transition at which the giant component disappears.

     Hint: The more smooth your line, the better the figure. The more independent instances you average over, the smoother your line.
   - Do your results depend on which graph space you choose for your configuration model?