Network Analysis and Modeling
CSCI 5352, Fall 2014
Prof. Aaron Clauset
Problem Set 3, due 10/8

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 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?

		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	

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) (10 pts) 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

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

   (b) (10 pts) 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.

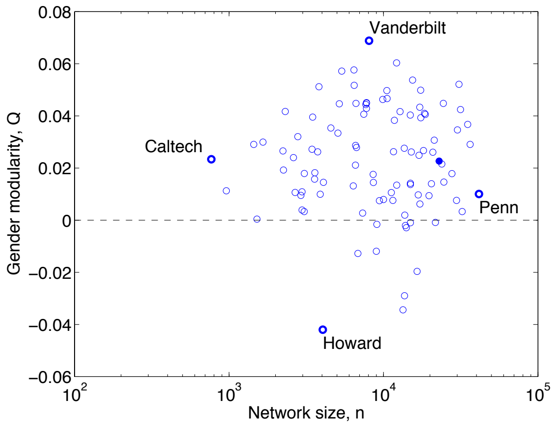
3. (25 pts) Implement the greedy agglomerative algorithm described in the lecture notes for maximizing modularity on an unlabeled simple network. Apply this algorithm to the Karate club network (data file in the class Dropbox).

   Make (i) a plot showing the modularity score Q as a function of the number of merges and (ii) a visualization of the network with vertices labeled according to your maximum modularity partition. Then calculate the normalized mutual information (NMI) between your partition and the "social partition" (second file in the Dropbox).[1] Finally, briefly interpret the degree of agreement or disagreement between the two partitions.

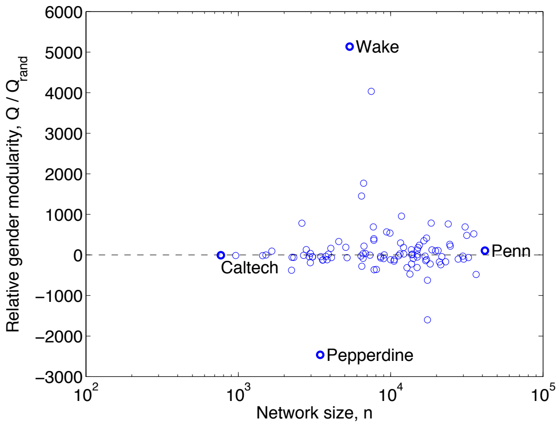
4. (40 pts) Using the FB100 data files from the class Dropbox, investigate the assortativity patterns for the following vertex attributes: student/faculty status, major, and vertex degree. Treat these networks as undirected.

   For each of vertex attribute, produce a pair of figures on log-linear axes. In one, make a scatter plot showing your measure of assortativity as a function of network size n. In the other, make a scatter plot showing the relative size of your empirical value versus its *average* value under a simple null model; include a horizontal line for no relative difference. As a null model for enumerative attributes, take a random permutation of the vertex labels, keeping the graph structure fixed. For degree, use the configuration model, keeping the attributes fixed.

   As an example of what you should produce, below is a pair of figures showing results for the gender attribute on the Facebook 100 networks. The left-hand figure suggest that most schools have gender assortativity very close to what we would expect from random mixing (within 6% in either direction), and that mixing does not covary with network size. The pattern in the right-hand figure shows that the expected value under the permutation model is often *very* close to 0, which produces large ratios in many cases. Thus, even the appearance of marginal assortativity by gender, e.g., Wake, is likely the result of due to noise.



empirical values



versus permutation model

   Based on these results, comment on the degree to which vertices do or do not exhibit assortative mixing on each attribute. Interpret your results by providing a brief discussion of what your findings suggest about the overall pattern of friendships at these schools and the potential mechanisms that structure them. (Figures lacking axis labels will receive no credit.)

[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. (10 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.