CSCI 5352, Network Analysis and Modeling **Prof. Aaron Clauset**
Problem Set 2 **Fall 2021**

There are 100 regular points and 20 extra points possible on this assignment.

1. (20 pts) Consider the following simple and rather unrealistic model of a network: each of n vertices belongs to one of g groups. The mth group has n_m vertices and each vertex in that group is connected to others in the group with independent probability $p_m = A(n_m - 1)^{-\beta}$, where A and β are constants, but not to any vertices in other groups. Thus, this network takes the form of a set of disjoint groups of communities.

   (a) Calculate the expected degree $\langle k \rangle$ of a vertex in group m.

   (b) Calculate the expected value $\langle C_m \rangle$ of the local clustering coefficient for vertices in group m.

   (c) Hence show that $\langle C_m \rangle \propto \langle k \rangle^{-\beta/(1-\beta)}$. What value would β have to assume for the expected value of the local clustering coefficient to fall off as $\langle k \rangle^{-0.75}$, as has been conjectured by some researchers?

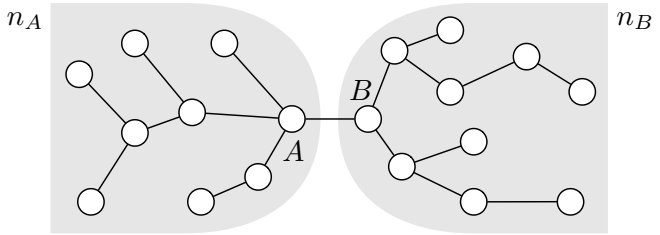
2. (20 pts) Consider the random graph $G(n, p)$ with average degree c.

   (a) Show that in the limit of large n the expected number of triangles in the network is $\frac{1}{6}c^3$. In other words, show that the number of triangles is constant, neither growing nor vanishing in the limit of large n.

   (b) Show that the expected number of connected triples in the network, as defined on page 184 of *Networks*, is $\frac{1}{2}nc^2$.

   (c) Hence, calculate the clustering coefficient C, as defined in Eq. (7.28) in *Networks*, and confirm that it agrees for large n with the value given in Eq. (11.11) in *Networks*.

3. (15 pts) Consider an undirected, unweighted network of n vertices that contains exactly two subnetworks of size n_A and n_B, which are connected by a single edge (A, B), as sketched here:



Show that the closeness centralities C_A and C_B of vertices A and B, as defined by Eq. (7.21) in *Networks*, are related by

$$\frac{1}{C_A} + \frac{n_A}{n} = \frac{1}{C_B} + \frac{n_B}{n} \ .$$

4. (25 pts) Using a uniformly random 25% subset of the FB100 networks, test the degree to which the "scaling" behavior of the clustering coefficient C, i.e., how the clustering coefficient varies as network size increases $C(n)$, can be explained by a random graph null model based on matching either (i) the density of edges alone, or (ii) the degree structure alone. Present your results in a single log-log figure by plotting $C(n)$ for the empirical data, and both null models. Then discuss what conclusions you can draw from this plot about the relative roles of edge density and degree structure in explaining the clustering observed in these social networks as they scale up in size.

   Hints: the clustering coefficient function in `igraph` is about 100x faster than the equivalent function in `networkx`. First measure the empirical function $C(n)$, i.e., compute the 25 pairs of coordinates (n, C) and make a preliminary plot; then, under each null model, and for each of the 25 FB100 networks G_i, generate 100 corresponding random graphs to calculate that network's $(n_i, \langle C \rangle)$. Use the Fosdick et al. MCMC to correctly sample simple graphs with the specified degree sequence.

   (10 pts *extra credit*) Repeat the same experiment as above, but for the network's mean geodesic path length $\langle \ell \rangle$. Discuss what you learn.

5. (20 pts) The Medici family was a powerful political dynasty and banking family in 15th century Florence. The classic network explanation of their power[1] claims that they established themselves as the most central players within the network of prominent Florentine families, occupying the structurally most important position in the network (shown in Fig. 1 below).

   Visit the *Index of Complex Networks* (ICON) at `icon.colorado.edu` and obtain a copy of the `Medici network` data file, under the "Padgett Florentine families" ICON entry.

   Conduct the following tests of their structural importance hypothesis. Define the *harmonic centrality* of a node as

$$C_i = \frac{1}{n-1} \sum_{j=1; j \neq i}^{n} \frac{1}{\ell_{ij}} \ , \tag{1}$$

[1] Padgett and Ansell, "Robust Action and the Rise of the Medici, 1400–1434." *American J. Sociology* **98**(6), 1259–1319 (1993).

where ℓ_{ij} is the length of the shortest path from node i to node j; if there is no such path, i.e., because i and j are in different components, then we define $\ell_{ij} = \infty$.

(a) Calculate and report the harmonic centrality of each node in the Medici network, and comment on where in the corresponding ranking the Medici family appears. Then discuss the degree to which your findings agree with the network explanation of the Medici's power, and what, if anything, the scores say about the second most important family.

(b) Use a null model to determine whether the Medici's structural importance can be explained by the degree sequence $\vec{k}$ of the network G alone, or includes effects beyond just degrees. Produce an ensemble of 1000 random graphs $\mathcal{G}$, each with the same degree sequence $\vec{k}$ as G, and for each node i in G, extract a distribution from $\mathcal{G}$ of i's harmonic centrality scores.

Hint: use the Fosdick et al. MCMC to correctly sample simple graphs with the specified degree sequence, and note that the networkx `configuration_model` function only produces stub-labeled loopy multigraphs.

Make two figures, each using a slightly different null model: (i) the vertex-labeled simple graph space, and (ii) the stub-labeled loopy multigraph space, in which one then removes self-loops and collapsed multiedges to produce a simple graph. For each, show the difference between a node's harmonic centrality on G and its average harmonic centrality in the corresponding ensemble $\mathcal{G}$. On each figure, include lines showing the 25% and 75% quantiles of the distribution around the mean for all vertices (as in the figure below, for the Zachary's Karate Club network).

Discuss what your results here mean for the network explanation of the Medici's power and your results from part (5a), and how important the selection of the null model is.

6. (10 pts *extra credit*) Reading the literature.

Choose a paper from the Supplemental Reading list on the external course webpage . Read the whole paper. Think about what it says and what it finds. Read it again, if it's not clear. Then, write a few sentences for each of the following questions in a way that clearly summarizes the work, and its context.

- What was the research question?
- What was the approach the authors took to answer that question?
- What did they do well?
- What could they have done better?
- What extensions can you envision?

Do not copy any text from the paper itself; write your own summary, in your own words. Be sure to answer each of the five questions. The amount of extra credit will depend on the accuracy and thoughtfulness of your answers.

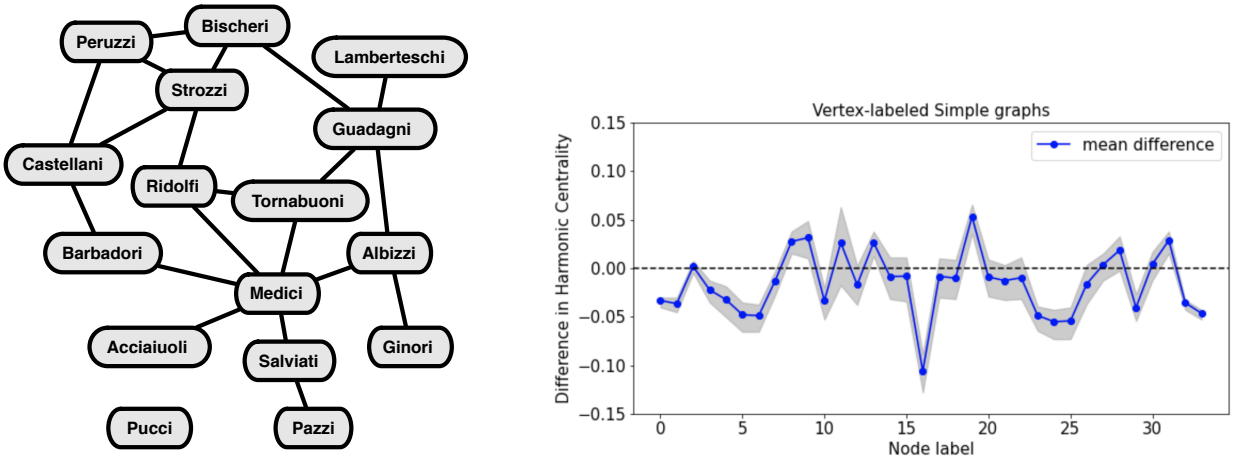


Figure 1: (left) The Medici family alliance network, from Padgett and Ansell (1993). (right) For the Zachary's Karate Club network, the difference between the observed centrality and its mean expected under a simple graph configuration model, along with 25% to 75% quantiles shown as a grey envelope.