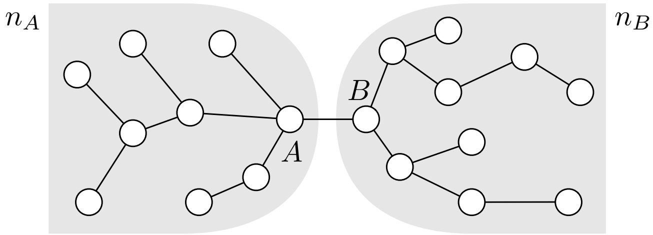
CSCI 5352, Network Analysis and Modeling **Prof. Aaron Clauset**
Problem Set 2 **Spring 2025**

There are 105 regular points and 40 extra points possible on this assignment.

1. (15 pts) *Centrality* measures are node-level statistics that are meant to quantify how centrally positioned (vs. peripherally located) a node i is.[1] They are often interpreted as a measure of node "importance" within the network; hence, nodes with higher centrality values are structurally more important in the system.

   Closeness centrality is defined only on connected graphs, and is given by Eq. (7.21) in *Networks*. 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 mathematically that the closeness centralities C_A and C_B of vertices A and B are related by

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

2. (20 pts) Consider the *double-edge swap* graph operation $s : \; G \to G'$ as defined in class, which preserves the degree sequence and graph space of an input graph G. The double-edge swap is a powerful tool and can be adapted to preserve additional node and graph properties. In this problem, you will develop double-edge swaps for certain types of non-simple networks. Assume in each case that the input graph space forbids multigraphs and self-loops.

   For each, state (i) the criteria for selecting two input edges $(u, v), (x, y)$ at random, (ii) enumerate all the allowed output configurations that meet the requirements, and (iii) state any checks that need to performed on G' to ensure it is in the same graph space as G.

   (a) (10 pts) A double-edge swap operation for *directed networks* that preserves both (i) node in-degree k^{in} and (ii) node out-degree k^{out}.

   (b) (10 pts) A double-edge swap operation for (undirected) *bipartite networks* that preserves (i) the degree sequence $\vec{k}$ and (ii) the node type of every node x_u.

[1] There are *many* centrality measures. The most common are closeness, betweenness, harmonic, and eigenvector. Generally, all centrality measures correlated with each other and with node degree (a.k.a. degree centrality), but each is based on different theoretical assumptions of what it means to be "structurally important."

3. (20 pts) The double-edge swap can also be used to insert a specific *amount* of randomness into a network. In this way, we can use it to define a parametric network model that contains a specified amount of "real" structure, which we can get from an empirical network. Let r be the number of double-edge swaps we have applied to some input graph G.

   Using the UC Berkeley social network from the FB100 data set, design and carry out a numerical experiment to answer the following question: as a function of r, how does the clustering coefficient C and mean path length $\langle \ell \rangle$ relax onto those of the corresponding configuration model? As references, overlay in your plots horizontal lines for the C and $\langle \ell \rangle$ when you have applied $r = 20m$ swaps (which will be the values expected under the configuration model). Comment on how "random" Berkeley's social network was to begin with, in what ways, and on the rate at which randomization destroys the empirical patterns.

4. (30 pts) In network data science, the configuration model serves as the key null model for deciding whether some network measure $x = f(G)$ is big or small or typical or unusual. Visit the *Index of Complex Networks* (ICON) at `icon.colorado.edu` and obtain your choice of

   - 1 online / social network
   - 1 food web / biological network, and
   - 1 connectome / biological network.

   Treating them as simple graphs, design and carry out a numerical experiment to answer the following question for each network: to what degree can the network's (i) clustering coefficient C and (ii) mean path length $\langle \ell \rangle$ be explained by its degree structure? Display your results using plots showing the null distributions of these statistics under the configuration model, and overlay on that distribution a vertical line showing the empirical value for that network. Discuss (i) whether the empirical values are big, small, typical, or unusual, (ii) what that conclusion implies about the structure of these networks, and (iii) what hypotheses it suggests about the underlying data generating process that produced these networks.

   Hint: A good null distribution requires around 1000 configuration model random graphs. Applying a double-edge swap operation $r = 20m$ times, starting with the empirical graph at $r = 0$, is sufficient to generate a single corresponding configuration model random graph. The bigger the graphs you choose, the longer your compute times will be.

5. (20 pts) Null models can also be used to test hypotheses about individual nodes and their positions within the network. For example, is a high centrality value for node i actually just a function of the network's degree structure? Here, we use this approach to revisit a classic paper in social network analysis.

   The Medici family was a powerful political dynasty and banking family in 15th century Florence. The classic network explanation of their power[2] argues that their influence came from

[2] Padgett and Ansell, *American J. Sociology* **98**(6), 1259–1319 (1993). `https://www.jstor.org/stable/2781822`

positioning themselves as the most central node within the network of prominent Florentine families (see the network figure below).

Visit the *Index of Complex Networks* and obtain a copy of the `Medici network` data file, under the "Padgett Florentine families" ICON entry. Conduct the following tests of the Medici structural importance hypothesis. Define the *harmonic centrality* of a node as

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

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 the scores say about how important is second most important family.

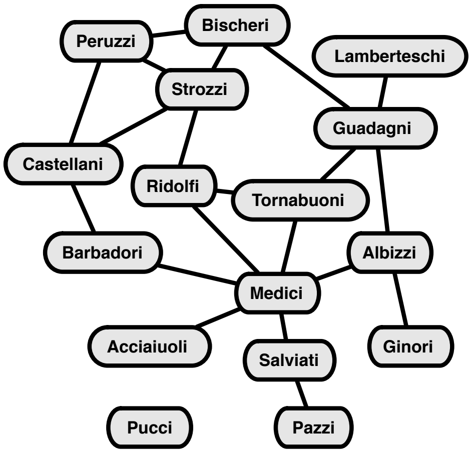
(b) Design and carry out a numerical experiment to answer the question: to what degree can each family's structural importance be explained as purely a function of the network's degree sequence $\vec{k}$? Use these results to assess the network explanation of the Medici's power, and explain how your results answer that question. Comment on which other families are more (or less) important than we expect under the null model.

Hint: If a family's centrality score is 'typical' in the null model, then we conclude that its centrality can be explained entirely by the network's degree sequence; if it not typical, then we conclude that other factors (beyond degrees) are needed to explain its centrality, e.g., it could be more central than we expect, or less!

Hint: As above, 1000 random graphs should give you a good null distribution, and it should be sufficient to apply $r = 20m$ double-edge swaps to produce one such graph. Because you need to look at all the families, a good visualization would arrange the node indices on the x-axis and plot for each the corresponding distribution of $(C_i^{\text{null}})_j - C_i^{\text{data}}$, where j indexes the random graphs. The degree to which each distribution spans $y = 0$ is then informative.

(c) (10 pts *extra credit*) Repeat the above experiment but change the null model to one that uses "the wrong" graph space. Specifically, use the stub-matching algorithm to construct stub-labeled loopy multigraphs, which you then "simplify" by removing self-loops and collapsing multiedges.

Make the same plot as before, and discuss how your results here differ (if at all) from when you use the correct null model (one that matches the data's graph space), and what conclusions would change (if any).



6. (20 pts *extra credit*) In the Watts-Strogatz model, the first few rewired edges play an outsized role in collapsing the path-length distribution, as they act like highways for shortest paths to cross the center of the ring. But as we rewire more edges, each highway carries progressively less traffic because we're creating many alternate avenues and the shortest-path traffic gets distributed more evenly until, when $p = 1$, every edge should carry roughly an equal fraction of all shortest paths.

   Design and carry out a numerical experiment that shows how the distribution of *betweenness centrality* values relaxes across the network as we rewire progressively more edges. Let r count the number of edges rewired.

7. (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 paper did you choose?
   - 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.