CSCI 5352, Network Analysis and Modeling **Prof. Aaron Clauset**
Problem Set 1 **Spring 2025**

There are 100 regular points and 40 extra points possible on this assignment.

1. (20 pts total) Read each of the following written descriptions of a network. (i) List all applicable graph properties of the network, from Lecture 2, and (ii) list which of the six domains the network falls into. If there are ambiguities, give a brief explanation of why it could go either way.

   (a) (4 pts) A network of associations among university students and the classes they take. A node is either a student or a class, and each student is connected to all the classes they are enrolled in.

   (b) (4 pts) A network of labor flows, in which nodes are companies, and a directed edge points from company i to company j with weight w_{ij} if w_{ij} workers left jobs at i to take jobs at j. Nodes are annotated by the total number of workers at the company, and its industrial sector.

   (c) (4 pts) A network of proteins and their pairwise interactions. Each node is a protein, and a pair of nodes i, j are connected if we observe *in vivo* that proteins i and j can bind. Edges are annotated with the corresponding (real-valued) binding affinity, and nodes are annotated with their molecular weight.

   (d) (4 pts) A sequence of network snapshots representing the spread over time of a communicable disease (e.g., covid) through a human population. Every snapshot contains the same set of nodes, and the t-th snapshot contains all the edges that occurred in the real-time interval of $[t, t+1)$. Nodes are people, and two people i, j are connected in the t-th snapshot if j was infected by i within the interval $[t, t+1)$. Nodes are annotated by the person's age and sex.

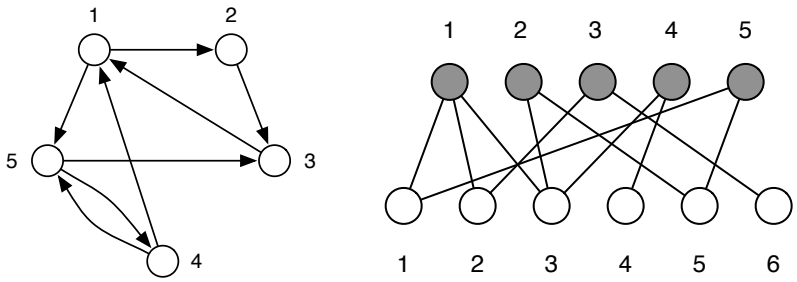
   (e) (4 pts) A network of social trust relationships, where nodes are people, and an edge i, j captures whether i has a positive or negative opinion of the trustworthiness of j.

2. (12 pts total) Consider the following two networks:

   (a) (3 pts) Give the adjacency matrix for network (i).

   (b) (3 pts) Give adjacency list for network (i).

   (c) (6 pts) Give adjacency matrices for both one-mode projections of network (ii).



(i) (ii)

3. (15 pts total) For each network G, calculate these summary statistics by hand:
   - maximum degree $k_{\max}$,
   - minimum degree $k_{\min}$,
   - clustering coefficient C (transitivity),
   - diameter $\ell_{\max}$.

   Explain each calculation and show your work. Express your answers in terms of the variables that parameterize the network (e.g., n).

   (a) (5 pts) Let G be a fully connected simple network, i.e., a complete graph, with n nodes.

   (b) (5 pts) Let G be a perfect binary tree containing n nodes. Hint: how many nodes n does a perfect binary tree contain, for depth $d = 0, 1, 2, \ldots$?
   (*5 pts extra credit*) Calculate the mean degree $\langle k \rangle$.

   (c) (5 pts) Let G be a simple "ring" network with $n \geq 3$ nodes, where nodes are arranged in a cycle, e.g., around the edge of a clock, each one connecting only to its immediate neighbors to the left (counter-clockwise) and right (clockwise).

4. (10 pts) Consider a bipartite network, with its two types of vertices, and suppose there are n_1 vertices of type 1 and n_2 vertices of type 2. Show that the mean degrees c_1 and c_2 of the two types are given by

$$c_2 = \frac{n_1}{n_2} c_1 \ .$$

5. (45 pts total) In this question, we will investigate the statistical properties of online social networks by analyzing the Facebook100 ("FB100") data set (download from the class Canvas). Each of the 100 plaintext ASCII files in the FB100 folder contains an edge list for a 2005

snapshot of a Facebook social network among university students and faculty within some university. Interpret this edge list as a simple graph.[1]

(a) (10 pts) For each of the FB100 networks, calculate the mean degree $\langle k \rangle$ and then make a nice histogram figure showing the empirical distribution of mean degree. Comment on the range of $\langle k \rangle$ you find, and whether it agrees or disagrees with your intuition, given what you know about these schools (feel free to look a few up to learn something about them).

Hint: If you use a standard histogram plotting function, don't use the default number of bins (often just 10). Instead, set the bin width equal to c, for some small-ish integer c.

(b) (5 pts) In most social networks, we observe a surprising phenomenon called the *friendship paradox*. Let k_u denote the degree of a node u, and consider the edge $(u, v) \in E$. The paradox is that the average degree $\langle k_v \rangle$ of a neighbor v is *greater* than the average degree $\langle k_u \rangle$ of a node u (note that in this notation, $\langle k_u \rangle = \langle k \rangle$, the mean degree of the network). Or, colloquially, each friend of yours has more friends than you, on average.

Define the mean neighbor degree (MND) of a network as

$$\langle k_v \rangle = \frac{1}{2m} \sum_{u=1}^{n} \sum_{v=1}^{n} k_v A_{uv} \ . \tag{1}$$

Derive an expression for $\langle k_v \rangle$ in terms of the average squared-degree $\langle k^2 \rangle$ and the average degree $\langle k \rangle$. Show your work.

(c) (15 pts) Using all of the FB100 networks, make a figure showing a scatterplot of the *size of the paradox*, defined as the ratio $\langle k_v \rangle / \langle k_u \rangle$, as a function of the mean degree $\langle k_u \rangle$. Include a horizontal line representing the line of "no paradox," and label the nodes corresponding to Reed, Colgate, Mississippi, Virginia, and UC Berkeley. (Remember: figures without axes labels will receive no credit.) [2]

Now comment on (i) the degree to which we do or do not observe a friendship paradox across these networks as a group, and on what makes the five labeled points notable, and (ii) whether there is any dependency between the size of the paradox and the network's mean degree.

(5 pts *extra credit*) Explain why we should, in fact, expect to see a friendship paradox in these networks, in terms of the conditions under which we should expect to see *no* paradox.

[1] Data kindly provided by A.L. Traud, P.J. Mucha and M.A. Porter, via their paper "Social Structure of Facebook Networks," *Physica A* **391**, 4165–4180 (2012), `http://arxiv.org/abs/1102.2166` or `http://bit.ly/1ztbVoS`.

[2] It took about 40 minutes on my laptop to read in all the data and calculate the x, y points for this plot.

(d) (15 pts) A related phenomenon in social networks is the *majority illusion*. Let $x \in \{0, 1\}$ be a binary-valued vertex-level property, and let $q = \frac{1}{n}\sum_u x_u$ be the fraction of vertices that exhibit this property. If we set $q < 0.5$, then this property appears only in a minority of nodes. The majority illusion occurs when $q < 0.5$, but the majority of a node's neighbors, on average, exhibit that property, that is, $\langle x_v \rangle > 0.5$.

Explain in both words *and* mathematics how this can be possible, in terms of the concepts of node degrees, the friendship paradox, and the distribution of x over nodes in the network.

(e) (20 pts *extra credit*) Most social networks also have very small diameters relative to their total size. This property is sometimes called the "small-world phenomenon" [3] and is the origin of the popular phrase "six degrees of separation".[4] [5]

- For each FB100 network, compute (i) the diameter $\ell_{\max}$ of the largest component of the network and (ii) the mean geodesic distance $\langle \ell \rangle$ between pairs of vertices in the largest component of the network. Make two figures, one showing $\ell_{\max}$ versus network size n and one showing $\langle \ell \rangle$ versus the size of the largest component n. Comment on the degree to which these figures support the six-degrees of separation idea.
  Hint: it's okay to use a library function to compute $\ell_{\max}$ and $\langle \ell \rangle$, but it'll be a slow calculation.
- Briefly discuss whether and why you think the diameter of Facebook has increased, stayed the same, or decreased relative to these values, since 2005. (Recall that Facebook now claims to have roughly 10^9 accounts.)

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 paper did you read?
   - What was the research question?

[3] Unfortunately, "small-world phenomenon" is sometimes also used to refer to a network with both a small diameter *and* a high clustering coefficient. You generally have decide from context which is the intended meaning.

[4] This term originated in a play written by John Guare in 1990, which was turned into a 1993 movie starring Will Smith. The concept, however, was originated by the sociologists Stanley Milgram, working in 1967, who was the first to measure the lengths of paths in large social networks.

[5] Running an All-Pairs-Shortest-Paths algorithm on these networks is computationally expensive. On my laptop, doing all 100 networks took about 10 hours of computing time, using a library implementation of Johnson's algorithm, which runs in $O(n \log(n) + nm) = O(n^2)$ for sparse networks.

- 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.

Warning: Don't ask chatGPT to summarize papers for you; it does a poor job, you don't learn anything, and it's pretty obvious to me as an expert when you do.