CSCI 5352 Problem set 1

Jake Krol

January 2025

1 Q1

1.1 Q1a

i)

- Nodes: students & courses
  - **Attributed**
    - Labeled as student OR course
- Edges: connect students to their enrolled courses
  - **Unweighted**
  - **Undirected**
- Network: students connected to their enrolled courses
  - **Bipartite**. Two distinct node sets with edges only across sets, not within.
  - **Sparse**. All student-to-student and course-to-course edges are zero.
  - The network is likely **disconnected** if both the student population and number of courses is large. For instance, a fourth year university would have students in specialized courses; for instance, there may not exist a path from a student studying electrical engineering to a student studying history. If there are any cases like this, then the graph is disconnected, otherwise connected.

ii) The network falls into both the **social** and **information** domains since the vertices include both people and their enrolled courses.

1.2 Q1b

i)

- Nodes: companies

  - **Attributed**
    - Industrial sector
    - Total number of workers

- Edges: worker migration counts
  - **Directed**
  - **Weighted**
  - **Multigraph**. Company i and company j can have weighted, directed edges reciprocated showing how many employees left from i to j, and vice versa.

- Network: worker migration counts across companies
  - **Sparse**. Since there are presumably multiple company industrial sectors (given by the node attributes), it is unlikely that most companies would exchange employees. Meaning, most edges are 0.
  - Likely **Disconnected**. It's unlikely that there's a path from, say, a clothing factory and an oil drilling company. Impossible paths imply disconnected graph.

ii) The network domain is **economic** since it depicts employment counts at companies.

1.3 Q1c

i)

- Nodes: proteins
  - **Attributed**
    - Molecular weight

- Edges: bindable protein pairs
  - **Weighted**
    - Binding affinity

- Network: protein binding affinities
  - **Sparse**. Only a few proteins can bind with one another, not most.
  - **Disconnected**. The sparsity of the graph implies there are paths $i \to j$ that are impossible. Plus, some proteins may have no binding partners at all, singletons.

ii) The network domain is **biological** since it depicts the binding affinity of protein bio-molecules and protein molecular weights.

1.4 Q1d

i)

- Nodes: people
  - **Attributed**
    - Age
    - Sex
- Edge layers: layers are indexed by time intervals
  - Edges: infection from source to target
    - **Directed**
- Network: transmitted infections during different time intervals
  - **Multiplex**
  - **Temporal**
  - **Acyclic**. Assuming no-reinfection is allowed in a time interval.
  - **Sparse**. I assume each infection has only one source and target in a given time interval, so most nodes/people would have at most only 1 edge in a time interval.
  - **Connected**. The infection spreads by people, and, assuming no singletons (uninfected individuals), there is a path leading from the first infected individual to the last.

ii) The network falls into both the **biological** and **social** domains since it depicts how a communicable disease (biological) transmits from person to person (social).

1.5 Q1e

i)

- Nodes: people
  - No special node properties
- Edges: trustworthiness score from source to target person
  - **Directed**
  - **Signed**
  - **Multigraph**
- Network: social trust

- **Sparse**. Assuming a large network, most people only know a small fraction of others in the network. So, most trustworthiness edges do not exist (are 0).
- **Disconnected or connected**. I would expect a social network to be connected. However, if the nodes/people is a subset of people from distant communities, say 100 people sampled from different countries, then it's possible to have a disconnected graph.

ii) The network is **social** since is depicts how people (nodes) trust each other (edges).

2 Q2

2.1 Q2a

0	1	0	0	1
0	0	1	0	0
1	0	0	0	0
1	0	0	0	1
0	0	1	1	0

Table 1: Adjacency matrix for network i

2.2 Q2b

Node	**Edges**
1	{2, 5}
2	{3}
3	{1}
4	{1, 5}
5	{3, 4}

Table 2: Adjacency list for network i

2.3 Q2c

0	1	1	1	1
1	0	0	1	1
1	0	0	0	0
1	1	0	0	0
1	1	0	0	0

Table 3: Adjacency matrix of one-mode projection of dark nodes.

0	1	1	0	1	0
1	0	1	0	0	1
1	1	0	1	1	0
0	0	1	0	0	0
1	0	1	0	0	0
0	1	0	0	0	0

Table 4: Adjacency matrix of one-mode projection of light nodes.

3 Q3

3.1 Q3a

- The network is fully connected, and each node is connected to all other $n-1$ nodes. The degree of every node is $n-1$, and $\mathbf{kmin} = \mathbf{kmax} = \mathbf{n} - \mathbf{1}$.
- The clustering coefficient of a fully connected network is 1, $\mathbf{C} = \mathbf{1}$, **if** $\mathbf{n} \geq \mathbf{3}$, since all possible triangles must exist. **If** $\mathbf{n} < \mathbf{3}$, then no triads nor triangles exist, $\mathbf{C} = \mathbf{0}$. Formally for a network with n nodes, the maximum number of possible triangles is $\binom{n}{3}$ and the maximum number of non-redundant triads is $\frac{n!}{(n-3)!}/2$ (permutations without replacement and divide by 2 to get rid of redundant cases where the tail nodes are simply swapped and the triad is not unique). By definition of the clustering coefficient,

$$C = \frac{3 \cdot \text{\# of triangles}}{\text{\# of triads}} = \frac{3(\frac{n!}{(n-3)!3!})}{\frac{n!}{(n-3)!}/2} = 1$$

- The diameter is 1 since every node is directly connected to all others: $\mathbf{l_{max}} = \mathbf{1}$.

3.2 Q3b

- $\mathbf{k_{max} = 3}$ since, for perfect binary trees, the largest degree are intermediate nodes (non-root and non-leaf) which have edges with 2 children nodes and 1 parent node.
- The minimum degree is $1, \mathbf{k_{min}} = 1$. The minimum degree nodes are all leaf nodes which only have an edge with their parent node.
- The clustering coefficient of a perfect binary tree is 0, $\mathbf{C = 0}$, since the structure of a perfect binary tree does not allow triangles. Binary trees do not allow edges between nodes on the same level; therefore, triangle formation is not possible. Formally,

$$C = \frac{3 \cdot 0}{\text{\# of triads}} = 0$$

- The diameter of a perfect binary tree is $\mathbf{l_{max} = 2 \cdot depth}$. The longest path, l_{max}, is to traverse from a leaf node on the left-hand side of the root to a leaf node on the right-hand side of the root node. Ascending the tree takes depth number of steps and descending also takes depth number of steps, for a total of $2 \cdot \text{depth}$.

Extra credit: The average degree of a perfect binary tree is

$$\langle k \rangle = \frac{(2^d \cdot 1) + (1 \cdot 2) + ((2^d - 2) \cdot 3)}{2^d + 1 + (2^d - 2)}$$

Each parenthesized term in the numerator corresponds to the count of all three possible node types (leaf, root, internal) scaled by their node degrees, given a perfect binary tree. Where d denotes the depth of the tree, $2^d \cdot 1$ corresponds to 2^d leaf nodes each with degree 1. $1 \cdot 2$ corresponds to the single root node with degree 2. And, $(2^d - 2) \cdot 3$ corresponds to the number of intermediate nodes each with degree 3. The denominator is the total number of nodes written as the sum of counts of each node type (leaf, root, internal).

3.3 Q3c

- In the described ring graph, every node is connect to exactly two neighbors $\mathbf{k_{min} = k_{max} = 2}$.
- The **clustering coefficient** of a ring depends on the number of nodes

$$C(n) = \begin{cases} 1 & \text{if } n = 3, \\ 0 & \text{else} \end{cases}$$

For $\mathbf{n = 3}$, there is one triangle and three triads, $\mathbf{C = 1}$. However for $\mathbf{n > 3}$, no triangles will ever exist ($\mathbf{C = 0}$) since each node is only connected to its clockwise and counter-clockwise neighbor. Therefore, each node is part of three triads, yet no triangles.

- The diameter of a ring network is $\mathbf{l_{max}} = \lfloor \frac{\mathbf{n}}{\mathbf{2}} \rfloor$. For networks with an even number of nodes, geometrically speaking, if the nodes are equidistantly arranged to circumscribe a circle, then the furthest target node of any source is directly on the other side (rotating π) of the circle. And, $\frac{n}{2}$ nodes need to be traversed to get to the target, target included in path length. For odd networks, there is no node directly opposite of the source, so either a clockwise or counter-clockwise path has a better path, given the equidistant spacing of nodes along the circumscribed circle. Therefore, odd networks with n nodes have the same l_{max} as the previous even network with $n-1$ nodes. Using the floor equivalence, this is generalized as $\lfloor \frac{n}{2} \rfloor$.

4 Q4

For a bipartite network with two vertex sets V_1 and V_2, proof by contradiction can show that

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

where c_i is the average degree of V_i and n_i is the number of nodes in V_i.

First, assume the equality is not true.

$$c_2 \neq \frac{n_1}{n_2} c_1$$

Let k_i be the degree of node i and expand the average degree terms (c_1, c_2) into scaled summation form.

$$\frac{1}{n_2} \sum_{i=1}^{n_2} k_i \neq (\frac{n_1}{n_2}) \frac{1}{n_1} \sum_{i=1}^{n_1} k_i$$

Simplifying the equation reveals that each side is the degree sum of the disjoint vertex sets.

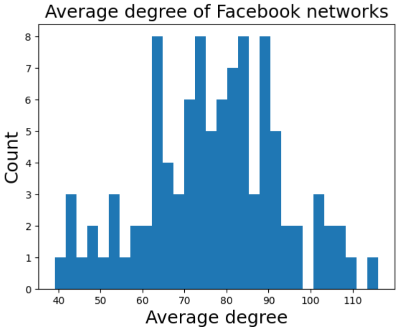
$$\frac{1}{n_2} \sum_{i=1}^{n_2} k_i \neq (\frac{n_1}{n_2}) \frac{1}{n_1} \sum_{i=1}^{n_1} k_i$$
$$\frac{1}{n_2} \sum_{i=1}^{n_2} k_i \neq \frac{1}{n_2} \sum_{i=1}^{n_1} k_i$$
$$\sum_{i=1}^{n_2} k_i \neq \sum_{i=1}^{n_1} k_i$$

However, the degree sum of the two vertex sets in a bipartite network *must* be equal since each $(u, v) \in E$ has one node from each set. By contradiction, it must be the case that the average degree c_2 is equal to $\frac{n_1}{n_2} c_1$.

$$\sum_{i=1}^{n_2} k_i = \sum_{i=1}^{n_1} k_i \rightarrow c_2 = \frac{n_1}{n_2} c_1$$

5 Q5

5.1 Q5a



The **range** of average degree for all college Facebook networks is roughly $[40, 120]$. **Intuitively, I would reasonably believe that an "average" person has any where between 40-120 friends**, depending on their college. Maybe, mean degree increases with student population, but there's no evidence to support this.

5.2 Q5b

The mean neighbor neighbor degree (MND), $\langle k_v \rangle$, can be written in terms of the mean of squared degrees, $\langle k^2 \rangle$, and the mean degree, $\langle k \rangle$, as

$$\langle \mathbf{k_v} \rangle = \frac{\langle \mathbf{k^2} \rangle}{\langle \mathbf{k} \rangle}$$

Given MND $= \langle k_v \rangle = \frac{1}{2m} \sum_{u=1}^{n} \sum_{v=1}^{n} k_v A_{uv}$, expanding the double summation shows that $\sum_{u=1}^{n} \sum_{v=1}^{n} k_v A_{uv} = \sum_{v=1}^{n} k_v^2$.

Expand the inner sum.

$$\sum_{u=1}^{n} \sum_{v=1}^{n} k_v A_{uv} =$$

$$\sum_{u=1}^{n} k_{v=1} A_{u,v=1} + \sum_{u=1}^{n} k_{v=2} A_{u,v=2} + \ldots + \sum_{u=1}^{n} k_{v=n} A_{u,v=n}$$

Factor out the constant degrees, k_v.

$$k_{v=1} \sum_{u=1}^{n} A_{u,v=1} + k_{v=2} \sum_{u=1}^{n} A_{u,v=2} + \ldots + k_{v=n} \sum_{u=1}^{n} A_{u,v=n}$$

Notice that $\sum_{u=1}^{n} A_{u,v=j}$, is just a column sum over the adjacency at column $v = j$. For a simple graph, this is equivalent to the degree of node $v = j$.

$$k_{v=j} \sum_{u=1}^{n} A_{u,v=j} = k_v \cdot k_v = k_v^2$$

Therefore, the double summation in the MND formula can be substituted with a summation over the squared degree of nodes.

$$\sum_{u=1}^{n} \sum_{v=1}^{n} k_v A_{uv} = \sum_{v=1}^{n} k_v \cdot k_v = \sum_{v=1}^{n} k_v^2$$

Substituting $\sum_{v=1}^{n} k_v^2$ for the summation portion of MND,

$$\text{MND} = \langle k_v \rangle = \frac{1}{2m} \sum_{v=1} k_v^2$$

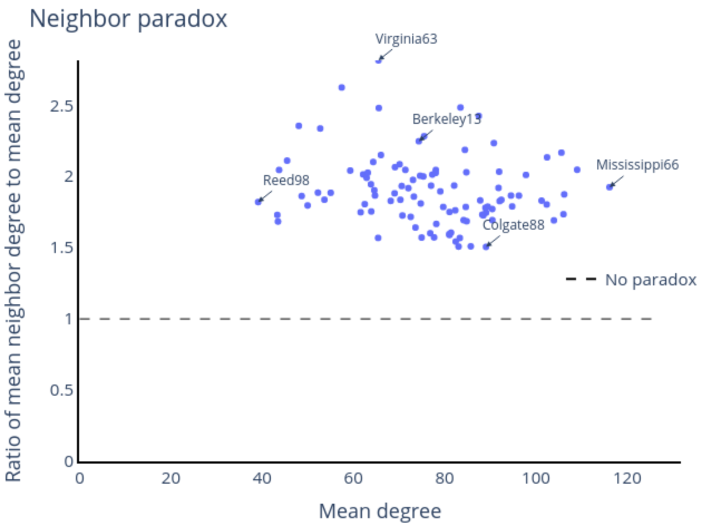
Next, we use the fact that the number of edges, m, is half of the degree sum: $m = \frac{1}{2} \sum_{v=1}^{n} k_v$ to rewrite m in terms of $\langle k \rangle$.

$$\begin{aligned} m &= \frac{1}{2} \sum_{u=1}^{n} k_u \\ \frac{m}{n} &= \frac{1}{2n} \sum_{u=1}^{n} k_u \\ \frac{m}{n} &= \frac{\langle k \rangle}{2} \\ m &= \frac{n\langle k \rangle}{2} \end{aligned}$$

Substituting this result for m in MND, shows how $\langle k_v \rangle$ relates to $\langle k^2 \rangle$ and $\langle k \rangle$.

$$\begin{aligned} \langle \mathbf{k_v} \rangle = \frac{1}{2\frac{n\langle k \rangle}{2}} \sum_{v=1}^{n} k_v^2 &= \\ \frac{1}{n\langle k \rangle} \sum_{v=1}^{n} k_v^2 &= \\ \frac{1}{\langle k \rangle} \frac{1}{n} \sum_{v=1}^{n} k_v^2 &= \\ \frac{\langle \mathbf{k^2} \rangle}{\langle \mathbf{k} \rangle} \end{aligned}$$

5.3 Q5c



i) The **friendship/neighbor paradox is prevalent in the Facebook networks** where the ratio of mean neighbor degree to mean degree ranges (vertical axis) from about $[1.5, 3.0]$. **No friendship paradox would correspond to a ratio of** 1, the dashed line of the figure.

The **five annotated colleges show how the paradox is invariant to whether the mean degree of the college's network is low or high**, shown by Reed98 and Mississippi66. The largest ratio of mean neighbor degree to mean degree was Virginia63.

ii) **No, there is not a noticeable dependency between the value of the mean neighbor degree to mean degree ratio as the mean degree** (horizontal axis) changes. The ratio is roughly uniformly distributed over the range college mean degrees.

Extra credit: **The friendship paradox would not be present if every node in the network had the same degree**. Since the nodes in each Facebook network do not have constant degrees, the friendship paradox does exist.

5.4 Q5d

I modeled each x_u as a Bernoulli random variable, and related the fraction of neighbor attributes to the mean neighbor degree in question 5b, but I did not find that $\langle x_v \rangle$ could be $> p$ to support the majority illusion phenomenon.

First, I **relate the fraction of neighbor vertices, which is equivalent to the average $\langle x_v \rangle$, with the attribute to the computation of mean neighbor degree (MND), defined in Q5b**. Except, k_v is substituted for x_v.

$$\text{MND} = \frac{1}{2m}\sum_{u=1}^{n}\sum_{v=1}^{n} k_v A_{uv} \rightarrow \langle x_v \rangle = \frac{1}{2m}\sum_{u=1}^{n}\sum_{v=1}^{n} x_v A_{uv}$$

Second, I use the fact that the **binary valued attribute, $x \in 0, 1$, is distributed over the nodes as a Binomial random variable**. The distribution is parameterized by the number of nodes, n, and the probability, p, of the attribute equaling 1 for a given node.

$$X \sim \text{Binomial}(n, p)$$

Third, I tried **showing that $\langle x_v \rangle > 0.5 > p$ by using the linearity and scalar properties of expectation**; however, I instead find that $\langle x_v \rangle = p$, not the desire result of $\langle x \rangle > 0.5 > p$.

I rewrite the summation using the fact that a column sum of the adjacency matrix is equal to the degree of the neighbor node $\sum_{u=1}^{n} A_{uv} = k_v$.

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

Expanding the sum and considering each x_v as an independent Bernoulli trial yields

$$\begin{aligned} &\frac{1}{2m}[k_1\langle x_1 \rangle + \ldots + k_n\langle x_n \rangle] = \\ &\quad \frac{1}{2m}[k_1 p + \ldots + k_n p] = \\ &\quad\quad \frac{p}{2m}[k_1 + \ldots + k_n] \end{aligned}$$

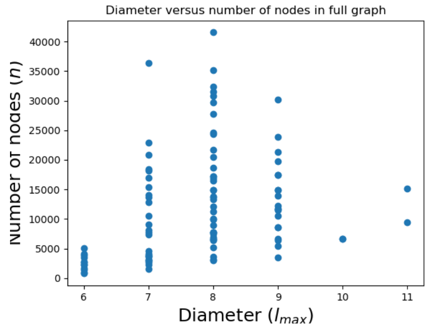
However, the degree sum, $k_1 + \ldots + k_n$ is twice the number of edges which leads to a result of p. Therefore, I instead show that $\langle x_v \rangle = p$, not $\langle x_v \rangle > 0.5 > p$.

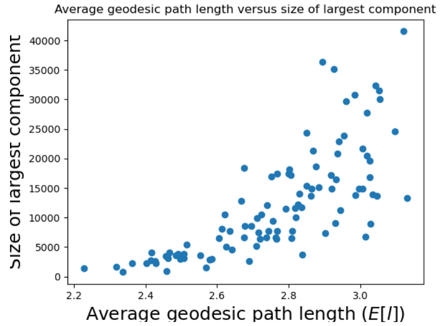
$$\begin{aligned} &\frac{p}{2m}[k_1 + \ldots + k_n] = \\ &\quad \frac{p}{2m} 2m = \\ &\quad\quad p \end{aligned}$$

I would guess my mistake is in how I originally define $\langle x_v \rangle$, and get to the sum over the Bernoulli variables weighted by degrees. But, I'm not too sure.

Disregarding the math, I intuitively think the majority illusion phenomenon is related to the friendship paradox. Where, by changing the way we compute the average attribute of nodes $\langle x_u \rangle$ to the targets $\langle x_v \rangle$ will increase the observed instance of $x = 1$ attribute. I think this occurs because most degree distributions are heavy-tailed, and if a high-degree node has the attribute $x_v = 1$, then it will be counted many times.

5.5 Q5e





No, the Facebook networks do not follow the "six degrees of separation idea" since the diameter ranges from $[6, 11]$. **If the "six degrees of separation" is defined as the idea that *all* nodes can reach each other with a path length** $l \leq 6$, then the diameter versus n plot shows that there exists many paths greater than 6 in the Facebook networks.

However, the average path length versus size of largest component plot suggests that *most* paths are probably well below 6. So, **if the six degrees idea is relaxed to say "most" people in a network are separated by a path** $l \leq 6$, then the average path length versus size of largest component plot supports this.

I think a modern Facebook network's diameter has stayed the same since generally network diameters tend to be asymptotically constant, $O(1)$, not $O(n)$. Meaning, the diameter will not grow as the amount of people (n, nodes) increases overtime.

6 Q6

- I read the paper "To explain or to predict?" (Shmueli, 2010).
- The paper did not have a research question. The paper's objective was to communicate the difference of causal and predictive modeling.

- The approach to communicating the difference of causal and predictive modeling involved stating the definitions of various terms, describe experimental design differences, and discussing examples of causal and predictive research studies.

- I think the paper conveyed well how important it is to define a research objective to decide whether an experimental design should use explanatory or predictive statistics. If the research goal is to explain some phenomena, then the experiment should be designed to test hypotheses related to the phenomena, not necessarily describing data related to the phenomena. While, research goals aiming to build accurate, generalizable predictors should design a study to evaluate a model's predictive power.

- To improve the paper, I would suggest he paper could have excluding some of the many term definitions, and I would move some of the motivation for the subject from the conclusion to the introduction. The words ”explaining”, ”explanatory modeling” and other definitions are provided. While these definitions are useful to make sure the reader and writer are on the same page, I think terminology is subject to context. Instead, the paper should emphasize providing motivation and evidence for the discussion. Overall, the overuse of definitions in the introductory sections and lack of early motivation could have been adjusted to improve the paper.

- An interesting extension of this work would be to investigate how often recent deep/machine learning research is inaccurately used to prove causal hypotheses. The conclusions of this paper focused on the opposite, that researchers were mistakenly using explanatory statistics to quantify predictive power. However, recent research is extremely data driven, so I think there's a good chance to see common cases where predictive models are used to prove explanatory hypotheses.

7 Code

Code at https://github.com/jakekrol/csci-5352-network_analysis_and_modeling/tree/main