Problem Set 1

CSCI 5352

Ben Braun

1.27.2025

Problem 1

- A
  - Properties
    - Unweighted
    - Undirected
    - Bipartite
    - Sparse (probably, unless the university is very small)
    - Disconnected (probably. If there are different majors with little overlap, there may be no path connecting some students.)
  - Domains
    - Social
- B
  - Properties
    - Weighted
    - Directed
    - Multigraph (if edges can go both ways e.g. (i, j) and (j, i))
    - Nodes have metadata
    - Sparse (Given that there are many distinct sectors and workers tend to flow mostly between similar sectors)
    - Projection
    - Disconnected (since some sectors have highly specialized workforces. Could also be connected if there is enough flow across sectors.)
  - Domains
    - Economic
- C
  - Properties
    - Weighted
    - Undirected

        - Nodes have metadata
        - Potentially a multigraph if there are multiple experiments with different results
        - Sparse
        - Projection
        - Disconnected (unless it's a very small set of proteins in a pathway)
        - Potentially a hypergraph (if protein complexes are included)
    - Domains
        - Biological
- D
    - Properties
        - Unweighted
        - Directed (i infects j)
        - Temporal (could also be thought of as multiplex if each snapshot is represented as a layer)
        - Timestamped edges
        - Nodes have metadata
        - Sparse
        - Projection
        - Disconnected (almost certainly, since most infectious diseases spread over time and need to incubate before they can be spread further, and recovered individuals generally become immune)
        - Acyclic (assuming recovered individuals gain immunity)
    - Domains
        - Social
        - Biological
- E
    - Properties
        - Signed
        - Directed
        - Multigraph (since we can consider both directions of trust in each relationship)
        - Sparse (assuming people only provide opinions about people they know well. Could also be dense if there's a lot of gossip and people have opinions about most other people)
        - Connected (assuming it's a small community without too much isolation or assortativity)

    - Domains
        - Social

Problem 2

A

$$A_i = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

B

$$\begin{aligned} A_i = [1] &\to (2, 5) \\ [2] &\to (3) \\ [3] &\to (1) \\ [4] &\to (1, 5) \\ [5] &\to (3, 4) \end{aligned}$$

C

$$A_{\text{top}} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix} \qquad A_{\text{bottom}} = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Problem 3

A

- $k_{\max} = n - 1$
    - Each node has exactly one edge to every other node
- $k_{\min} = n - 1$
    - Same as maximum
- $C = \frac{\binom{n}{3} \times 3}{\binom{n}{3} \times 3} = 1$
    - Every set of 3 nodes forms a triangle, so the number of triangles is the same as the number of connected triples. With n nodes there are $\binom{n}{3}$ triples, which we multiply by 3 to account for the symmetry of each triangle.
- $l_{\max} = 1$

- The network is not weighted, so each edge has length 1. Since each vertex has an edge to every other vertex, the longest path between any two vertices is trivially 1.

B

- $k_{\max} = 3$
  - An internal node has exactly 3 edges (one to its parent and two to its children)
- $k_{\min} = 1$
  - A leaf node has exactly 1 edge (to its parent)
- $C = \frac{0 \times 3}{\frac{3n}{2} - 3.5} = 0$
  - There are no triangles in a tree. The denominator is the number of connected triples in a tree with n nodes. I arrived at $\frac{3n}{2}$ because at depth > 1, each pair of children adds 3 triplets. 3.5 is subtracted because this pattern does not hold for the first two levels of the tree. I arrived at 3.5 by checking a few levels manually - I'm not sure why it works out to exactly 3.5.
- $l_{\max} = 2(\log_2(n+1) - 1)$
  - The longest path in a perfect binary tree is any path from a leaf to another leaf on the opposite side of the tree. So we need to travel up to the root, which is $\log_2(n+1) - 1$ edges away, and then travel back down to the opposite leaf, which is another $\log_2(n+1) - 1$ edges.
- $\langle k \rangle = \frac{1}{n}(2 + 3(\frac{n-1}{2} - 1) + (\frac{n+1}{2})) = 2 - \frac{2}{n}$
  - To find the mean degree, we separate the nodes into three categories. The root has degree 2, the internal nodes have degree 3, and the leaf nodes have degree 1. The tree has $\frac{n-1}{2} - 1$ internal nodes and $\frac{n+1}{2}$ leaf nodes, so the mean degree is simply the sum of these nodes divided by n.

C

- $k_{\max} = k_{\min} = 2$
  - Every node has exactly 2 edges
- $C = \frac{0 \times 3}{n} = 0$
  - There are no triangles (when $n \geq 3$) and there are exactly n connected triples since each node has exactly two neighbors.
- $l_{\max} = \lfloor \frac{n}{2} \rfloor$
  - The furthest distance between any two nodes is halfway around the cycle, so half the number of nodes. We round down if there are an odd number of nodes to select the most direct path.

Problem 4

In a bipartite network, edges can only connect nodes in n_1 to nodes in n_2. Therefore, for every edge in the one-mode projection of graph 1 there is a corresponding edge in the one-mode projection of graph 2, meaning that the mean degrees are equal. More formally, we calculate the mean degrees as

$$c_i = \frac{m_i}{n_i}$$

where m_i is the number of edges in each graph. Solving for m_i, we get

$$m_i = c_i n_i$$

Because every edge connects a node of type 1 to a node of type 2, we know that

$$m_1 = m_2$$

Then we can plug in the two sides of the equation to get

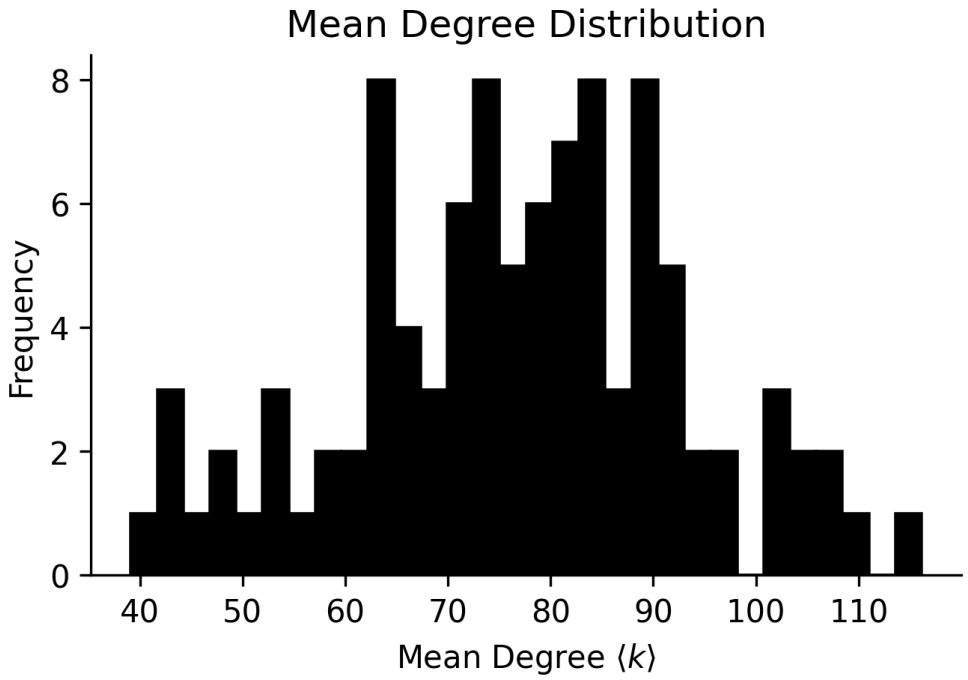
$$m_1 = m_2 = c_1 n_1 = c_2 n_2$$

Which trivially transforms to

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

Problem 5

Part A



We find a wide range of $\langle k \rangle$ from 39 to 116, which makes sense since we input a wide variety of networks. Smaller schools have smaller networks, which will tend to result in generally lower degrees i.e. fewer Facebook friends. Additionally, some schools may have been slower to adopt Facebook, particularly those that were less technology-oriented or had more limited internet connections. It also takes time to establish a large user base and build up a friend network, so the schools for which Facebook launched earlier would have had more time to increase the mean degrees of their networks.

Part B

We can define the degree of a fixed node v as

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

which forms a part of our equation for $\langle k_v \rangle$. We can use this to simplify to

$$\begin{aligned} \langle k_v \rangle &= \frac{1}{2m} \sum_{u=1}^{n} \sum_{v=1}^{n} k_v A_{uv} \\ &= \frac{1}{2m} \sum_{v=1}^{n} k_v^2 \end{aligned}$$

Recall that the mean degree of a network is defined as

$$\langle k \rangle = \frac{2m}{n}$$

Which we can transform to

$$2m = n\langle k \rangle$$

Subsituting this into our equation for $\langle k_v \rangle$ gives us

$$\langle k_v \rangle = \frac{1}{n\langle k \rangle} \sum_{v=1}^{n} k_v^2$$

Note that

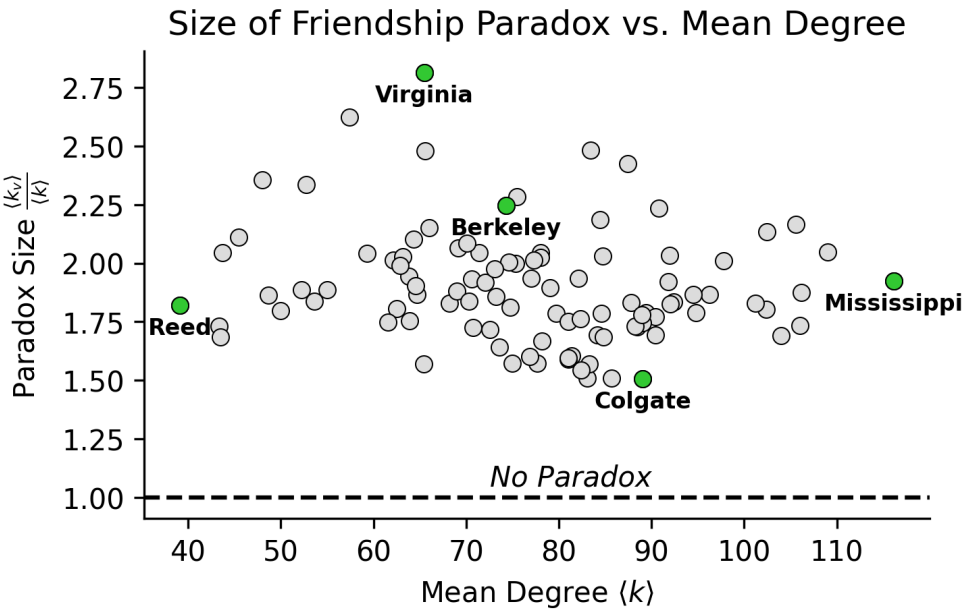
$$\frac{1}{n} \sum_{v=1}^{n} k_v^2 = \langle k^2 \rangle$$

So we can simplify to

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

Very cool!

Part C



We observe the friendship paradox in all of the Facebook networks, with no networks having a paradox size lower than 1.5.

The highlighted schools are largely notable because they represent the extremes of each axis. Reed has the lowest $\langle k \rangle$ and Mississippi has the highest $\langle k \rangle$, but their paradox sizes are very similar. Virginia has the largest friendship paradox while Colgate has the smallest, but their $\langle k \rangle$'s are not particularly unusual. Berkeley is the exception, with a $\langle k \rangle$ close to the average and a paradox size that's only slightly higher than most.

There seems to be no correlation between the mean degree and the paradox size, indicating that $\langle k_v \rangle$ tends to scale with $\langle k \rangle$ as might be expected in a typical social network.

We should expect friendship paradoxes here because the alternative is that $\frac{\langle k_v \rangle}{\langle k \rangle} = 1$, which is only possible if every node has the same degree i.e. everyone has exactly the same number of friends. In any real social network, this is unlikely to the point of being virtually impossible.

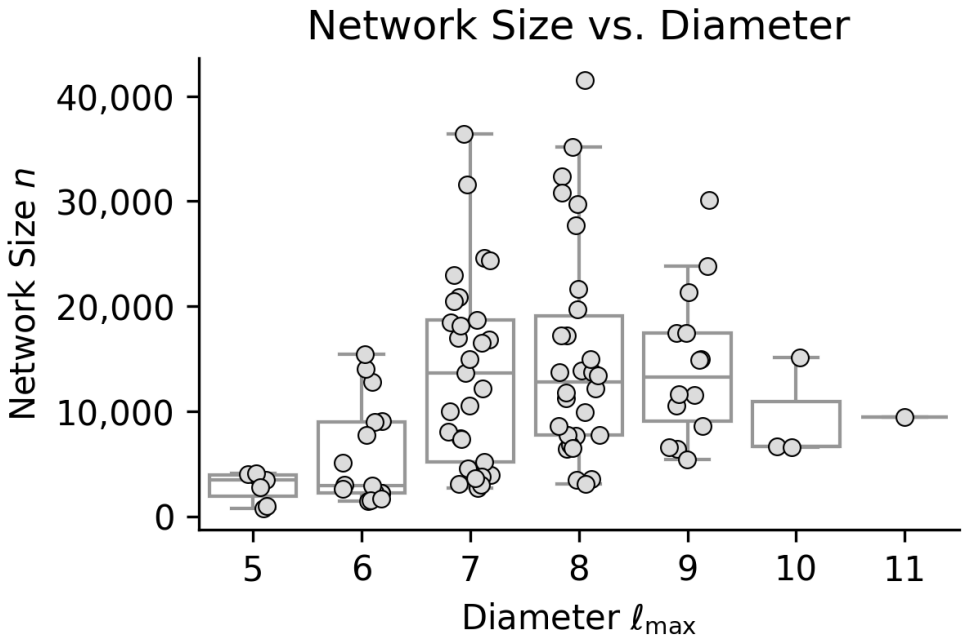
Part D

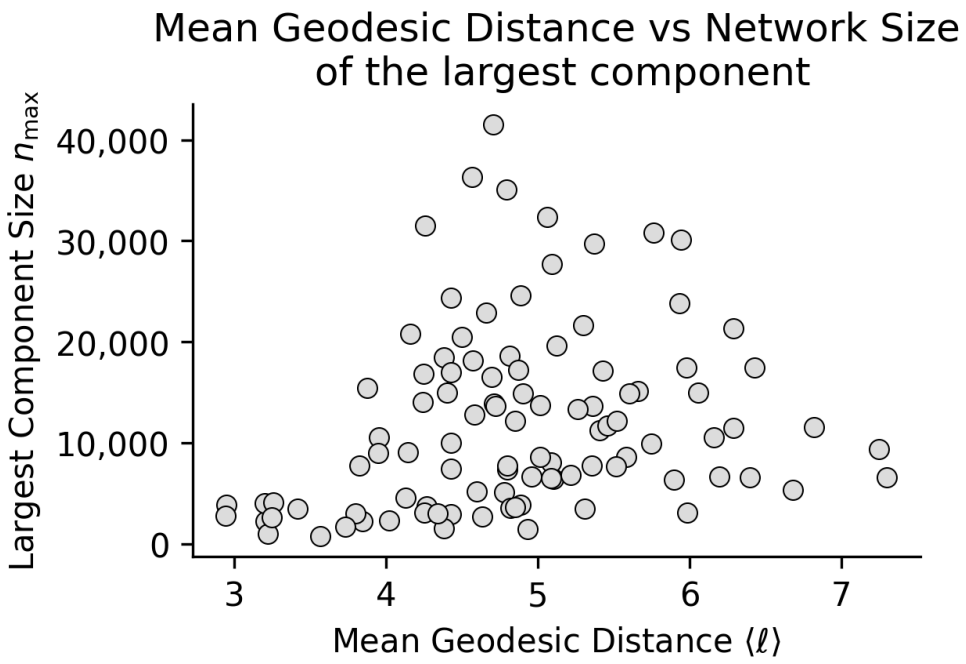
The majority illusion can occur when nodes of disproportionately high degree frequently exhibit the property of interest. The friendship paradox tells us that most nodes will tend to have neighbors with more friends, so if the property is exhibited by high-degree nodes, we can end up with the majority of nodes (which tend to have smaller degree) experiencing the majority illusion. In other words, most nodes will have few friends, so requires relatively few friends to form a majority. In a network with high-degree nodes

exhibiting the property, it's possible for the majority of nodes to experience this illusion despite most nodes not having the property. This becomes less likely as the variance in degrees decreases and as q decreases further below 0.5.

For example, we could construct a network with odd n in which all nodes with $k > \langle k \rangle$ have the property (so $q < 0.5$). The friendship paradox tells us that $\langle k_v \rangle > \langle k \rangle$, and in this network, majority of the average node's friends will have $k > \langle k \rangle$ and will thereby also exhibit the property. This creates a perceived prevalence of the property greater than 0.5 despite $q < 0.5$.

Part E





According to the first plot comparing network size and diameter, the six degrees of separation idea does not hold universally for most of the networks. The maximum degree of separation was 11, and the mean diameter was 7.4, indicating that these social networks are not well-connected enough to support six degrees of separation. However, the second plot shows that for the average pair of nodes in the largest component of most networks, the degrees of separation are indeed lower than 6, with a mean of 4.9 and a maximum of 7.3. This means that "six degrees" holds for the majority of users, but there are plenty of exceptions.

I think that Facebook's diameter has likely increased dramatically since 2005 since it has grown enormously and is no longer separated by universities/geography. Even if the $\langle \ell \rangle$ is low, the $\ell_{\max}$ is likely to be very high since one can likely find long strings of very poorly-connected users or users who do not follow well-connected accounts. I could go on Facebook right now and create 30 accounts that follow each other to form a chain, which would make Facebook's $\ell_{\max} > 30$. Given how large the modern network is, I don't think it's unlikely that structures like that exist for whatever reason. It's worth noting that according to the first plot, larger networks do tend to have greater diameters, although it isn't clear how this trend scales at higher orders of magnitude. As a result, there may not be enough information in this dataset to make any sound predictions about the diameter of modern Facebook.

Acknowledgments

- I solved all of these problems on my own but checked some of them with a few classmates afterwards
- I used Wikipedia to research a few relevant concepts and algorithms