Dakota Bryan
Network Analysis and Modeling
Aaron Clauset
January 2025

1. Properties that are certain I simply state. Properties that could be possible but don't have to be, I start with (maybe). These include an explanation of the ambiguity.

   a) i. Graph Properties
      - Edge:
        - Unweighted
        - Undirected
        - No self loops
        - (Maybe) Multi graph. The network could just be for one semester, or be a multi graph over multiple semesters with tags for each semester that a student and class are connected.
      - Node:
        - (Maybe) Each node is either a class or student, so this could be node attributes/ metadata. We could also label the nodes with the metadata of what class or student it is. We could also include additional details like demographic data.
      - Network:
        - Sparse (bi-modal distribution)
        - Bipartite
        - (Maybe) Connected. If this network is connected, that means you have a class with someone, who has a class with someone... that is in a class, for every class. This is an interesting question to ask of this network. I would guess it is connected, unless it is a very highly silo-ed, maybe technical school.
        - (Maybe) Temporal. If it is taken over multiple semesters with edge tags.
        - (Maybe) Multiplex. Same as above.

      ii. Econmic (could also be seen as social, but I think more economic)

   b) i. Graph Properties
      - Edge:
        - Weighted
        - Directed
        - (Maybe) Multi graph. Could be up until now, or for every year (or other time unit). Would be multi graph if it is for every year.
      - Node:
        - Metadata. Total number of workers and industrial sector.
      - Network:
        - (Maybe) Sparse. This could go either way. If the companies included cover a large geographic area and many sectors, it would be a relatively sparse graph. The smaller the geographic area and less sectors, the denser, and it could get quite dense. Possibly complete if it is in one city, and say only the tech sector with large-ish companies.
        - Projection
        - (Maybe) Connected. Similar to the school network above. An interesting question to ask, and like above, the smaller the geographic range and the less sectors, the more likely it is to be connected. The bounds for being connected are much smaller than high density.
        - (Maybe) Temporal and Multiplex. If it is taken over time instead of up to a point or in a specific time bounds.

      ii. Economic

   c) i. Graph Properties
      - Edge:
        - Weighted

        - Undirected
    - Node:
        - Metadata (annotated with molecular weight)
    - Network:
        - Sparse. I do not know enough about this to know how sparse it would be, but I would assume since it is likely disconnected, it is also sparse.
        - Projection
        - Disconnected. I would assume there are certain types of proteins that can interact. I would also assume that there are different parts of the body where this takes place, so disconnected that way too.
        - (Maybe) Temporal and Multiplex. Like all of the above, you could take snapshots of this process over time. For ex. all the interactions per day.

  ii. Biological

d) i. Graph Properties
    - Edge:
        - Unweighted
        - (Maybe) Directed. Most likely, we could represent who infected who.
        - Multi graph. Because it is timestamps, there could be multi edges over different times. (You can infect the same person twice)
    - Node:
        - Metadata: persons age and sex.
    - Network:
        - Sparse
        - Projection
        - (Maybe) Connected. If the nodes are not directed, this graph would be connected, assuming 1 person started it all. (wild)
        - Temporal and Multiplex.
        - (Maybe) Hyper graph. You could represent this as a hyper graph if you wanted to. A hyper edge would be all the people one person infected.

   ii. Biological (could also be consider social)

e) Like usual, this could be a snap shot, and a multigraph over time.

   i. Graph Properties
    - Edge:
        - Signed
        - Directed
    - Node:
        - (Maybe) Could add metadata.
    - Network:
        - (Maybe) Sparse. Depends on the size. If it is only one friend group, it would be dense. Assuming you only have an edge if you are friends.
        - Projection.
        - (Maybe) Connected. Depends on the size, and assuming the "base" graph is a friendship network. If it is for the whole world, not connected. For a small school, maybe.

   ii. Social

2. a) Since network (i) is directed, the rows represent an out-edge. That is, a 1 in row 3 column 2 means that node 3 has an edge *to* node 2.

$$A_{\mathrm{i}} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

b)

$$\begin{array}{rclcl} A_{\text{i}} & = & [1] & \rightarrow & (2,5) \\ & & [2] & \rightarrow & (3) \\ & & [3] & \rightarrow & (1) \\ & & [4] & \rightarrow & (1,5) \\ & & [5] & \rightarrow & (3,4) \end{array}$$

c)

$$A_{\text{gray}} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

$$A_{\text{white}} = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

3. Assuming we will not consider the trivial case of $n = 1$ for all graphs. For completeness, everything equals 0 for any graph with $n = 1$ except the clustering coefficient which is undefined.

   a) Complete graph with n nodes.

      i. $\boxed{k_{max} = n - 1}$
      In a complete graph with n nodes, every node has degree $n - 1$.

      ii. $\boxed{k_{min} = n - 1}$
      In a complete graph with n nodes, every node has degree $n - 1$.

      iii. $\boxed{C = 1}$
      Every node is connected to every other node. Therefore, everything that could be a triangle is a triangle (it is all triangles). Each triangle has three connected triples, so the number of triangles times 3 equals the number of connected triples, and C is 1.

      iv. $\boxed{\ell_{max} = 1}$ when $n \geq 3$, undefined otherwise.
      Every node is connected to every other node, so the length of the path from any node to any other node is 1, and the diameter is 1.
      If the graph has two nodes, there is no 2-path, so divided by zero is undefined.

   b) Perfect binary tree containing n nodes.

      i. k_{max}.
      For $n > 3$, $\boxed{3}$
      Leaf nodes have degree 1, source node has degree 2, and all other nodes have a edge to two children and one parent, so degree 3. Max is 3.
      For $n = 3$, $\boxed{1}$
      There is one parent and two children, so max and min degree 1.
      There is no perfect binary tree containing 2 nodes.

      ii. $\boxed{k_{min} = 1}$
      All leaf nodes have degree 1.

      iii. $\boxed{C = 0}$
      There are no triangles in a perfect binary tree. 0 times 3 divided by anything is 0.

      iv. ℓ_{max}.
      $\boxed{\ell_{max} = 2\log_2(n + 1) - 2}$
      A tree with depth d has $2^{d+1} - 1$ nodes. Consider the tree with depth 2. There are 4 leaf nodes. $2^2 = 4$. However, the other levels above will always have $4 - 1$ nodes. $2^3 - 1 = 7$. This holds for all d. The largest shortest path for depth d would be a node on the left side of the tree, then the path back to the source, then down to a leaf on the right side. This is $2 \times d$.

We need an equation for d in terms of n. We have $n = 2^{d+1} - 1$.
$n + 1 = 2^{d+1}$
$\log_2(n+1) = d + 1$
$d = \log_2(n+1) - 1$
$\ell_{max} = 2 \times d = 2 \times (\log_2(n+1) - 1) = 2\log_2(n+1) - 2$.

v. (extra)
$\boxed{\langle k \rangle = 2 + \frac{1}{n}}$
Every leaf has degree 1. There are 2^d leaves. Every interior node has degree 3, there are $2^d - 1$ interior nodes. The source has degree 2.
Therefore, sum of all degrees $2^d + 3(2^d - 1) + 2 = 2^d + 3(2^d) - 3 + 2 = 4 \times 2^d - 1$. Substitute for our formula for d in terms on n from above — $4 \times 2^{\log_2(n+1)-1} - 1 = 4 \times (n+1/2) - 1 = 2n+2-1 = 2n+1$ is the edges times 2 (or sum of all degrees). Divided by n for average, and we get $(2n + 1/n) = 2 + (1/n)$

c) Ring graph with $n \geq 3$

i. $\boxed{k_{max} = 2}$
Every node has degree 2. It has two neighbors, the node to the left around the ring, and the node to the right around the ring. Therefore, the max degree is 2.

ii. $\boxed{k_{min} = 2}$
Every node has degree 2. It has two neighbors, the node to the left around the ring, and the node to the right around the ring. Therefore, the min degree is 2.

iii. C
If $n = 3$, $\boxed{1}$. In this case, the ring is a triangle, and n is 3. Therefore, everything that could be a triangle is one, hence, $C = 3$.
If $n > 3$, $\boxed{0}$. In this case, the ring has no triangles. Therefore, C is 0.

iv. $\boxed{\ell_{max} = \lfloor n/2 \rfloor}$
Take any node v. The node, say u that is "furthest" from v is the one half way around the ring. Any other node will have a shorter path to v. This can be seen because it will be visited on the way to u. If n is even, then going either counter-clockwise or clockwise around the ring will have the same path length, $n/2$. If n is odd, then there will be two paths, and one will be 1 shorter than the other. Therefore, we must take the floor of $n/2$ in the case that n is odd.

4. Consider a bipartite graph with n_1 vertices of type 1, and n_2 vertices of type 2. The types are bipartite. Show the mean degrees c_1 and c_2 of the two types are given by

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

Proof.

Let m denote the total number of edges.

Since the graph is bipartite, the sum of degrees in type 2 is m. (Because by definition, every edge must have one node in type 2)

Thus, the $c_2 = \frac{m}{n_2}$.

The same can be said for c_1, $c_1 = \frac{m}{n_1}$.

We can derive a formula for m by rewriting c_1... $n_1 \times c_1 = \frac{m}{n_1} \times n_1 \rightarrow m = n_1 \times c_1$.

Substitute into the formula for c_2... $c_2 = \frac{c_1 \times n_1}{n_2} \rightarrow c_2 = \frac{n_1}{n_2} c_1$.

Qed.

5. Julia code available upon request (private repo for now)

a) Figure 0.1 shows the figure I generated. We see a fairly normal distribution, we the min around 37, and the max around 118. This range does agree with my intuition given what I know about these schools. They given list of schools is fairly diverse, but a strong preference for more prestigious, and smaller schools. My intuition which is largely based on my bias of going to a small liberal arts school, and hearing about bigger schools, is that smaller, more prestigious school would have higher mean degree. The smaller the school, the more people know "everyone on campus". The data seems to mostly agree with this assumption, with a few exceptions.

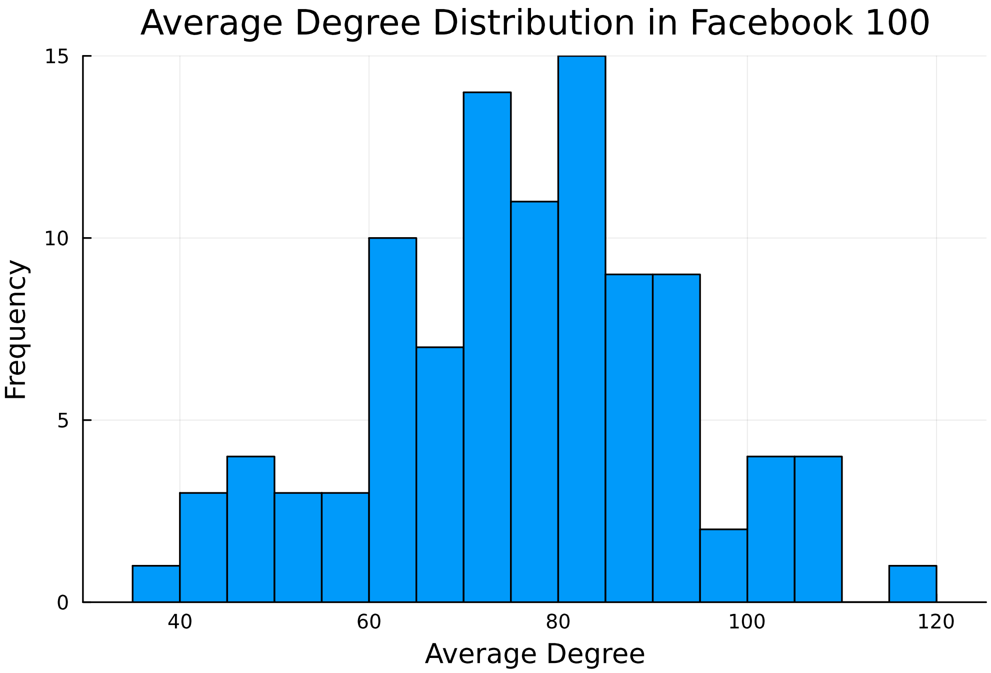

Figure 0.1: Distribution of Mean Degree (5.a. plot)

b) Let $\langle k_u \rangle$ denote the mean degree in the network, $\langle k_u^2 \rangle$ denote the mean squared degree, and $\langle k_v \rangle$ denote the mean degree of a neighbor. Also let n denote the total number of vertices, and m denote the total number of edges. Furthermore A_{uv} is 1 if $(u, v) \in E$, and 0 otherwise. Consider the formula for calculating $\langle k_v \rangle$

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

Informally, we are adding up the degrees of all neighbor's for each vertex, and adding those together, then multiplying by 1/2m.

The degree of each vertex gets added to the sum the amount of times that it is an adjacency (which is also the degree). Say node b has degree 5. In the summation, it will contribute 5^2. Let the sum of the degree of every vertex squared be k_{sum}.

Now, we have $\langle k_v \rangle = \frac{1}{2m} k_{sum}$, and
$\langle k_u^2 \rangle = \frac{k_{sum}}{n}$.
Furthermore, the average degree is calculated as follows: $\langle k_u \rangle = \frac{2m}{n}$.
$k_{sum} = \langle k_u^2 \rangle n$.
$\langle k_v \rangle = \langle k_u^2 \rangle \frac{n}{2m} = \langle k_u^2 \rangle \frac{1}{\langle k_u \rangle}$.
Which gives us the mean degree of a neighbor in terms of the mean squared degree and mean degree, namely: $\boxed{\langle k_v \rangle = \langle k_u^2 \rangle \frac{1}{\langle k_u \rangle}}$

c) Figure 0.2 shows the generated scatter plot with appropriate labels.

i. We do observe a friendship paradox across all networks, and to a large degree. The average ratio between the average degree of a neighbor over the average node degree is around 1.9 which is quite high. This means on average across all networks, the average person's friend will have just under twice as many friends as them. The lowest is just above 1.5, meaning that even in the school with the smallest ratio, on average, your friend will have 50 percent more friends that you.
The five labeled points are notable because they represent the boundary. The max and min friendship paradox, and the max and min degree.

ii. There seems to be no dependency between the size of the paradox and the network's mean degree. You could regression fit this to be sure, but it doesn't look like it would be significant.

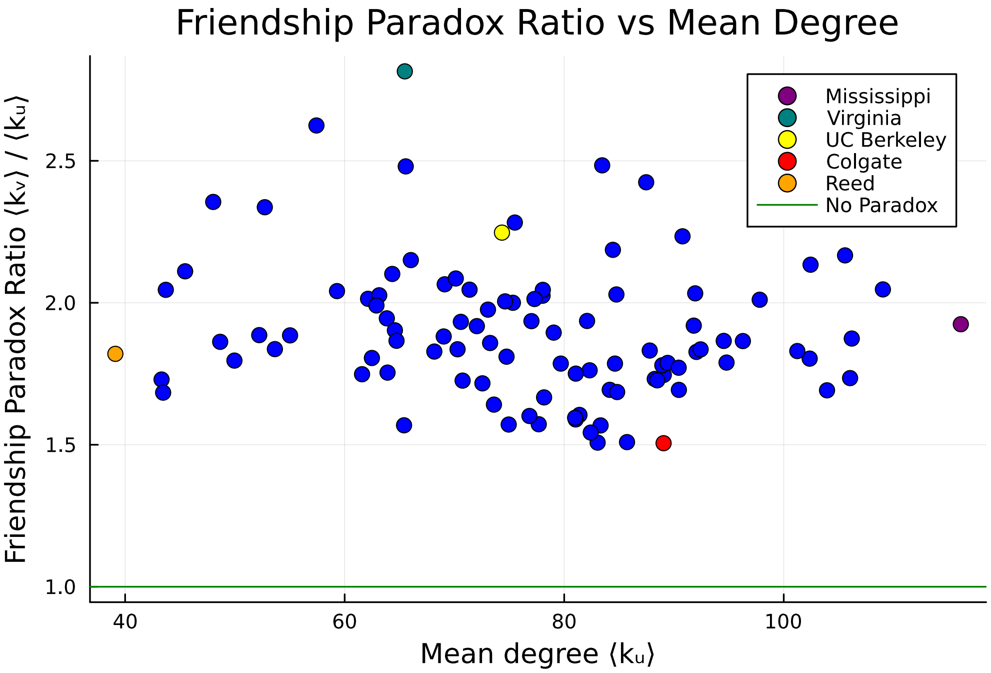

Figure 0.2: Scatter plot of the size of the friendship paradox as a function of the mean degree

iii. (Extra Credit) We would except to see no paradox if everyone had about the same amount of friends. However, in real-world networks (particularly social), this is rarely the case. Instead, there is usually a strong right tail in the distribution of degrees, as there are a few nodes which have many connections. The larger the skew, the larger the paradox. Furthermore, on a college campus, there are often some people who act as social hubs which skew the distribution. They could also be strong early adopters of Facebook.

Another way of thinking about it is if people stayed strictly in their circles, or mathematically, cliques, then we would see no friendship paradox. I.e. if all of your friends were friends with each other, and all of you were only friends with each other. This is not how college or life works.

d) The majority illusion is more likely to occur in situations where:

- The property x is positively correlated with node degree. (That is, the higher degree of a node, the more likely it is for the node to have the property)
- The friendship paradox occurs in the network. (That is, the average degree of a neighbor, $\langle k_v \rangle$, is higher than the average degree $\langle k \rangle$). This occurs when the degree distribution has variation, and is strengthened with a right tail distribution (most real-world networks).

In fact, the only situation where the majority illusion is statistically significantly likely to occur is when *both* the above conditions are met. The more of either, the more likely it is to occur (and of course the base probability of the property over any given node increasing).

The basic idea of this is if the property is equally distributed over all nodes with probability p, then the probability that any given neighbor k_u, of any node u, exhibiting the property is p. In general, for any given node v, any other node u is more likely to v has a neighbor the higher the degree of v. When there is no friendship paradox, every node has about the same degree. Therefore, no matter how x is distributed, we are not likely to see a majority illusion (because most degrees are the same).

We see a friendship paradox when nodes do not all have about the same degree, and is strengthened when there is a right tail distribution of degrees. Now, if x is distributed equally among all nodes, then this does not matter in terms of the majority illusion. But if higher degree nodes are more likely to have x, then (in a network with a friendship paradox) your neighbors are more likely to have x. This is because high degree nodes are more likely to have x, and you are more likely to be neighbors with high degree nodes. Of course, the larger the friendship paradox, the larger the majority illusion. Also, the more positively correlated x is with node degree, the larger the majority illusion.

More mathematically:

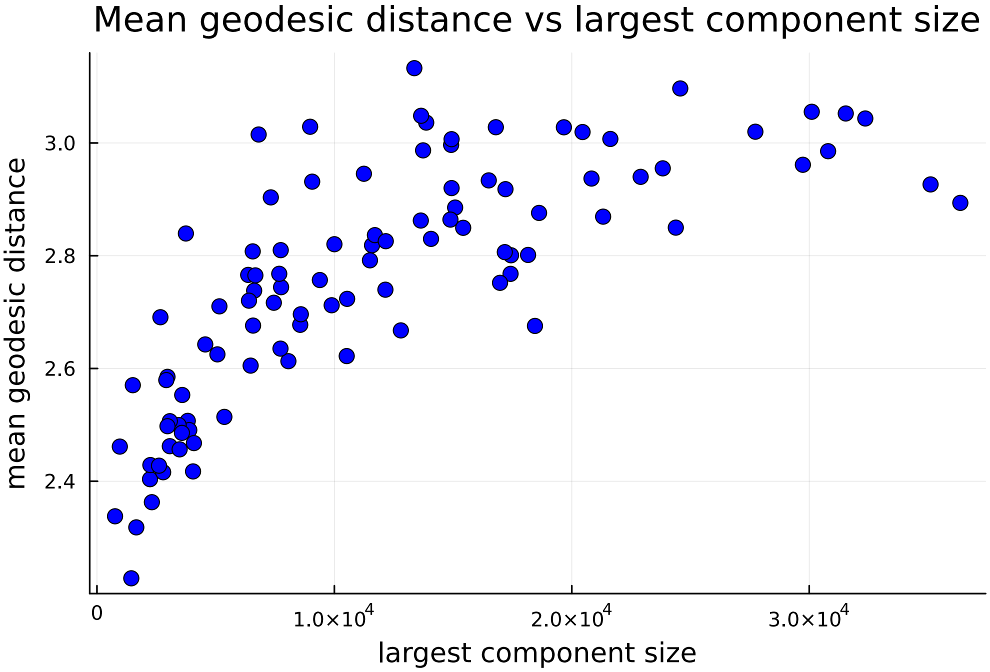

Figure 0.3: Scatter plot of the mean geodesic distance as a function of the largest component size

Let $q = \frac{1}{n}\sum_u x_u$ be the fraction of vertices with the property.

Let's make the q positively correlated with degree, by letting $p > 1$, and $q_u = \frac{k_u p}{r}$, where q_u is the probability any vertex u has the property, k_u is the degree of u, and r is some integer to keep all probabilities under 1.

The expected percentage of a node u's neighbors with the property is $\frac{1}{k_u}\sum_v q_v$ for all neighbors v. Or, $\frac{1}{k_u}\sum_v \frac{k_v p}{r}$. The probability of u having the property is $\frac{k_u p}{r}$.

Since on average with a friendship paradox, $k_v > k_u$, $\frac{k_u p}{r} < \frac{1}{k_u}\sum_v \frac{k_v p}{r}$ on average. This is true when $\frac{1}{k_u}\sum_v k_v > k_u$. The larger the friendship paradox and the larger p is, the bigger the difference between $\frac{k_u p}{r}$ and $\frac{1}{k_u}\sum_v \frac{k_v p}{r}$ is. The difference grows with p and $\langle k_v\rangle/\langle k_u\rangle$.

Note: you could really solve this out for any p with n determined by q, but that is a lot of probability and combinatorics I don't know and beyond the scope of the question (I think).

e) (extra credit) Note: these are 99/100 of the graphs. Penn was very large, and I had proceeded about 35000 of the 40000 nodes when I submitted this, so it is not included.

- Figure 0.3 shows the mean geodesic distance as a function of the largest component size. The mean geodesic distance was computed in the largest component, and does not count distances of length 0. Figure 0.4 shows the diameter as a function of the total network size. The diameter was computed in the largest component. Figure 0.5 shows the counts of diameters as a histogram.

  These figures support the idea that everyone people are relatively few "friend edges" away from each other, but does not support the six degrees of separate principle. It states that *all people* are six degrees or fewer away from *all* other people. This means we are talking about the measure of diameter. In these networks, most of the diameters are eight. However, it is worth noting that these are people who go to the same college. They are probably much further from someone living in Timbuktu. On the other hand, we are only considering Facebook friends, which is not an accurate measure (could miss some true edges, but also add ones that aren't for people who aren't "real" friends). Furthermore, some of these people may be closer to each other if you include friendships outside of the university. Thus, it is tough to say, but I would imagine the diameter for the nodes in these networks are slightly over their actual diameter (in terms of a global social network, with only these nodes, but edges outside of the nodes).

  It does support the idea that people are relatively close, as the average degree is usually between 2.4 and 3. Thus, if we expand this to the world, it may only go up a factor of two or three, supporting

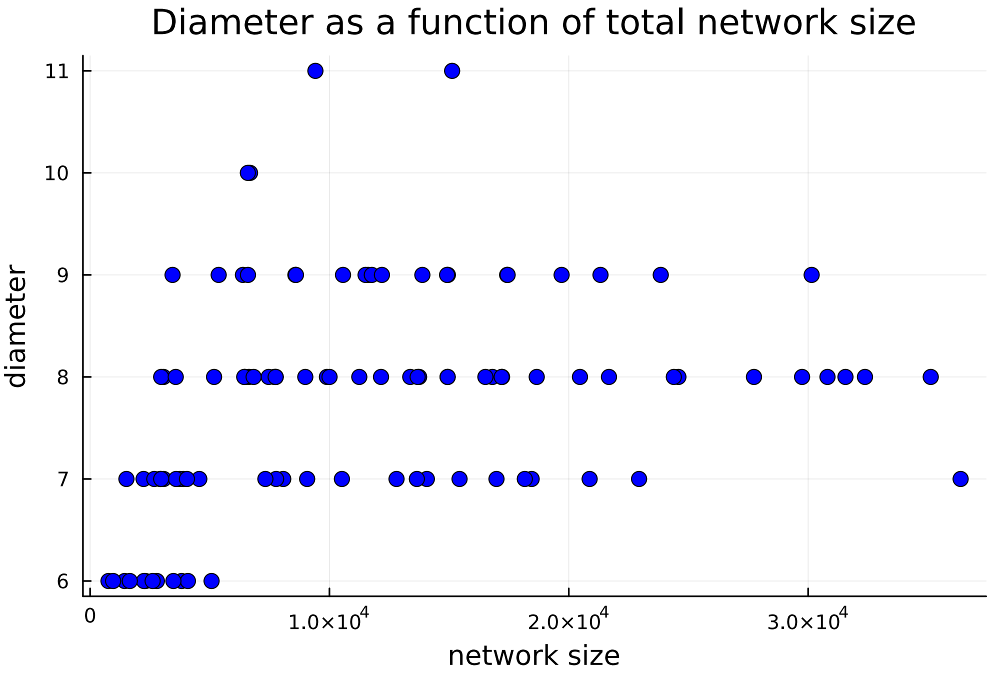

Figure 0.4: Scatter plot of diameter as a function of total network size

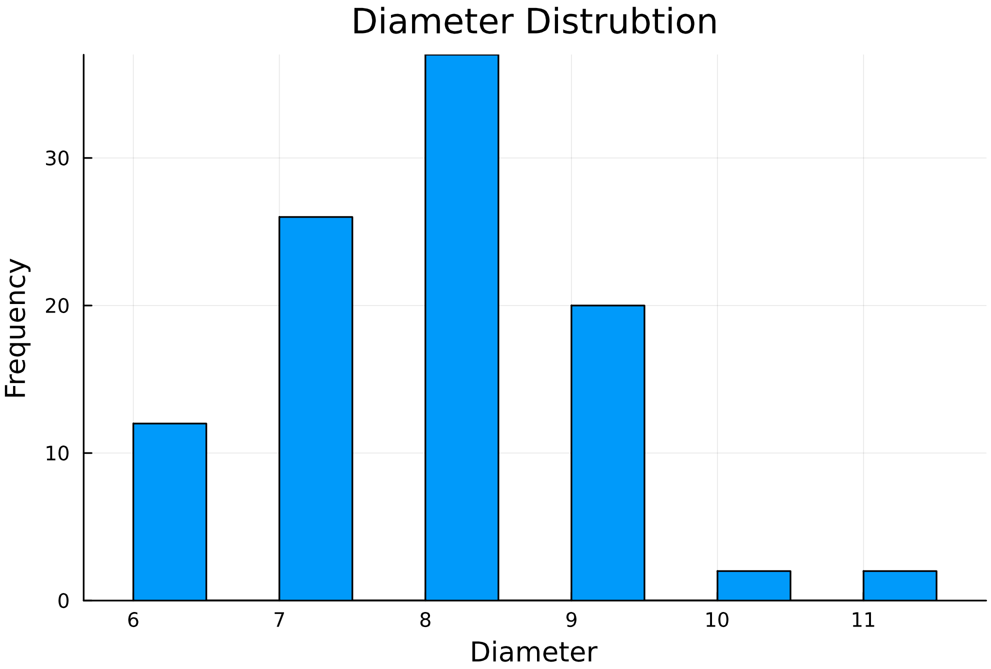

Figure 0.5: Histogram of diameters

the idea that *maybe* people are *on average* 6 or 7 degrees away.

- I think that the diameter of Facebook has increased slightly. I think on average, people have more friends and more friends from different places. However, in these graphs, it is only for one college. So, these are skewed to be smaller than the snapshot of all of Facebook back then. I think now, the average distance and the diameter has decreased from back then (for a full Facebook snapshot), because people have "online-only" friends. There is also many influences that serve as giant hubs shrinking average distances and diameter.