CS 144 – Homework 2

Marco Yang

1. **Coauthorship Visualization**

   (a) Plot the histogram and ccdf of node degrees in coauthorship network. Calculate the average clustering coefficient, overall clustering coefficient, maximal diameter, and average diameter.

Solution: Code is here.

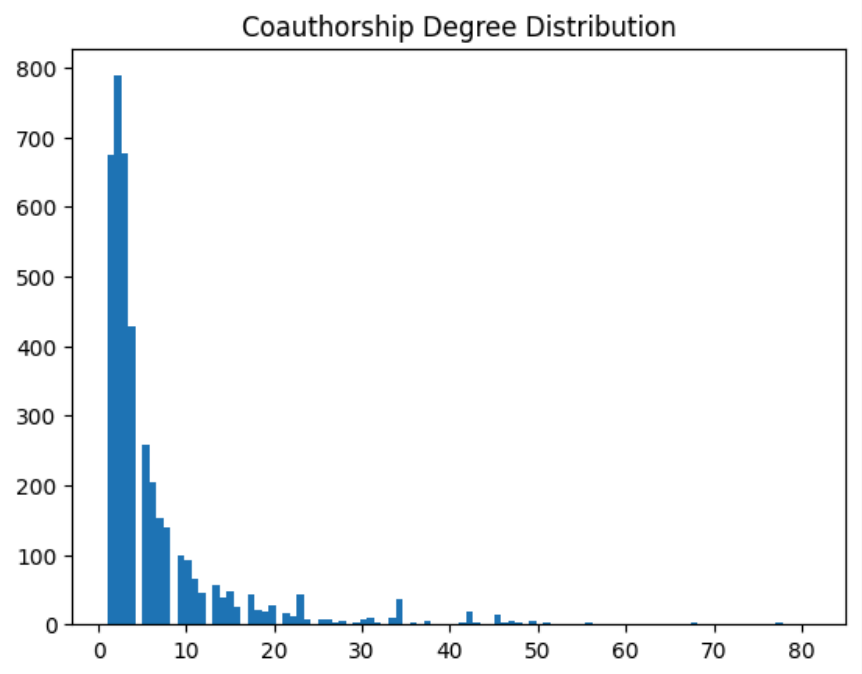

Figure 1: Histogram

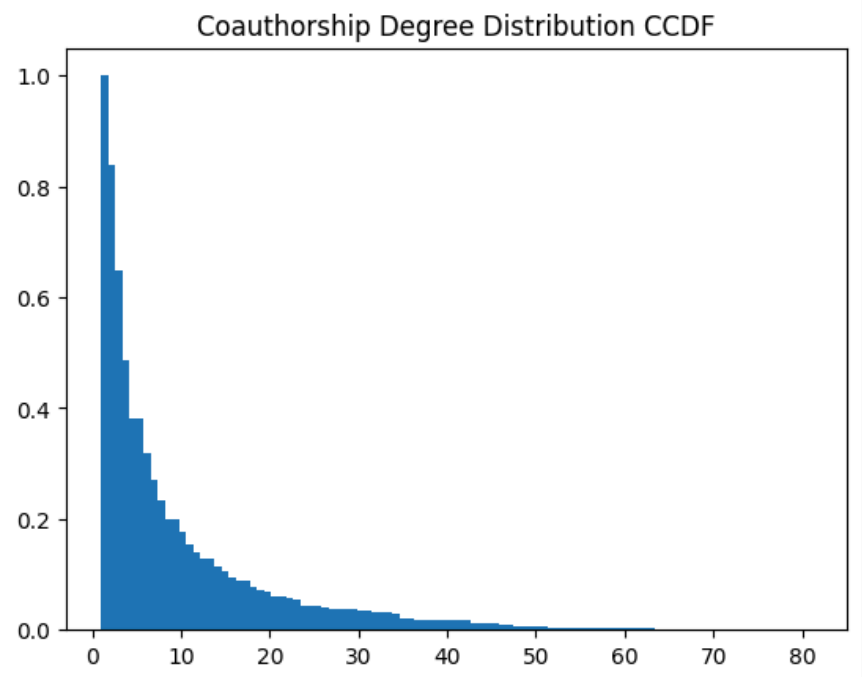

Figure 2: CCDF

- Average clustering coefficent: 0.5568782161697919
- Overall clustering coefficient: 0.6288944756689877
- Maximal diameter: 17
- Average diameter: 6.049380016182999

(b) Calculate the number of triangles. Assuming $T = \mathbb{E}[T]$, the expected number of triangles in an Erdos-Renyi graph with the same number of nodes, calculate the p in $G(n, p)$ for that E-R graph.

Solution: Code is here.

- Number of triangles: 47779
- $p = 0.0159$

(c) What distribution should the node degrees of a E-R graph take on? Is E-R a good model for coauthorship?

Solution: An Erdos-Renyi graph should have a binomial distribution, but the coauthorship network clearly does not follow that. Erdos-Renyi is not a good model nodes since authors in similar fields are likely coauthors, so treating every edge as equally likely and random is wrong.

2. **Visualizing Networks**

(a) **Solution**: Code is here.

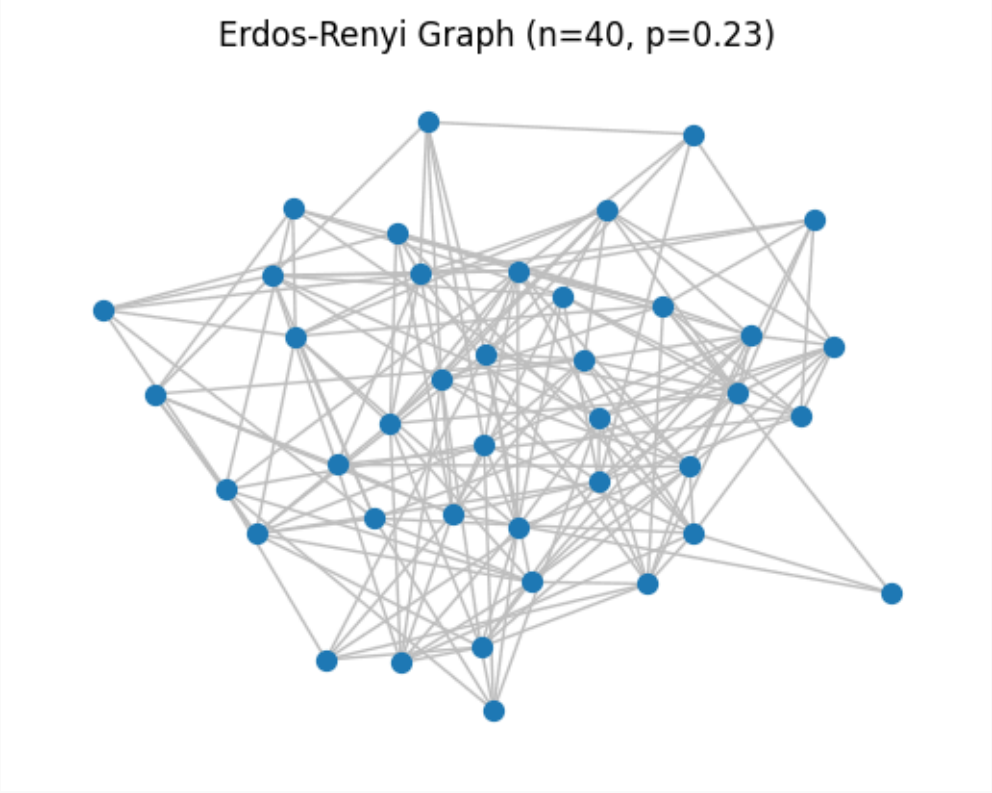

Figure 3: Erdos-Renyi network with $n = 40, p = 0.23$.

Analysis: yeah this graph looks ugly. Very random. I used the spring (nodes repel each other, edges are springs that pull together) layout since it shows just how boring this is.

(b)

Solution: Code is here.

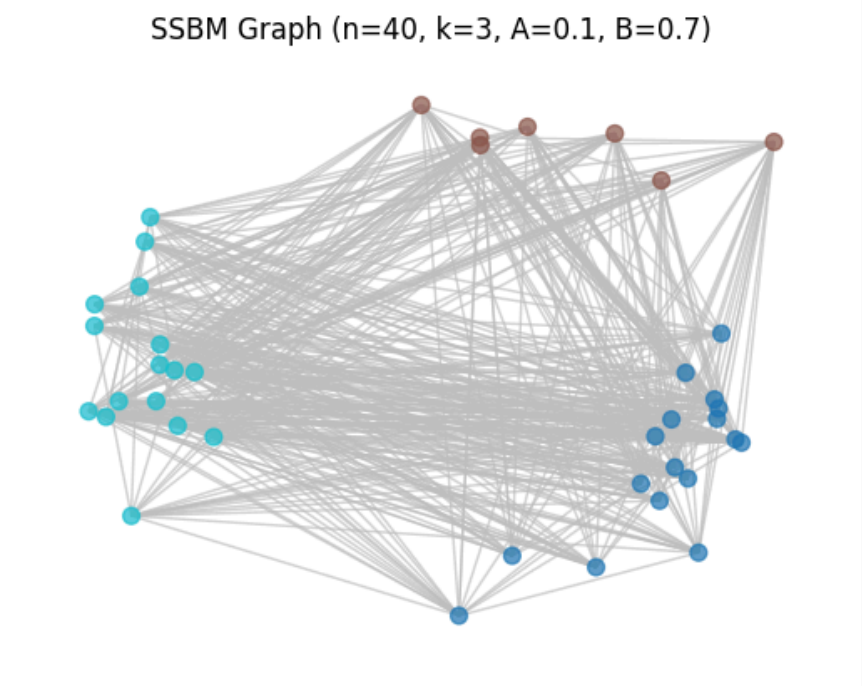

Figure 4: SSBM with $n = 30, k = 3, A = 0.7, B = 0.1$.

Analysis: much prettier this time. I had Claude implement a reverse spring layout (edges push apart, nodes attract) to show the clustering based on the community assignments. Each cluster/community is not very connected to itself but is very connected to the other clusters. I played around with parameters for the spring simulation until I got the clustering I wanted. I also highlighted nodes of the same community with a single color and made edges translucent so that overlapping edges could be distinguished from very sparse edges (e.g. heavy edge coloring in middle of the clusters, light edge coloring within each community) .

(c)

Solution: Code is here.

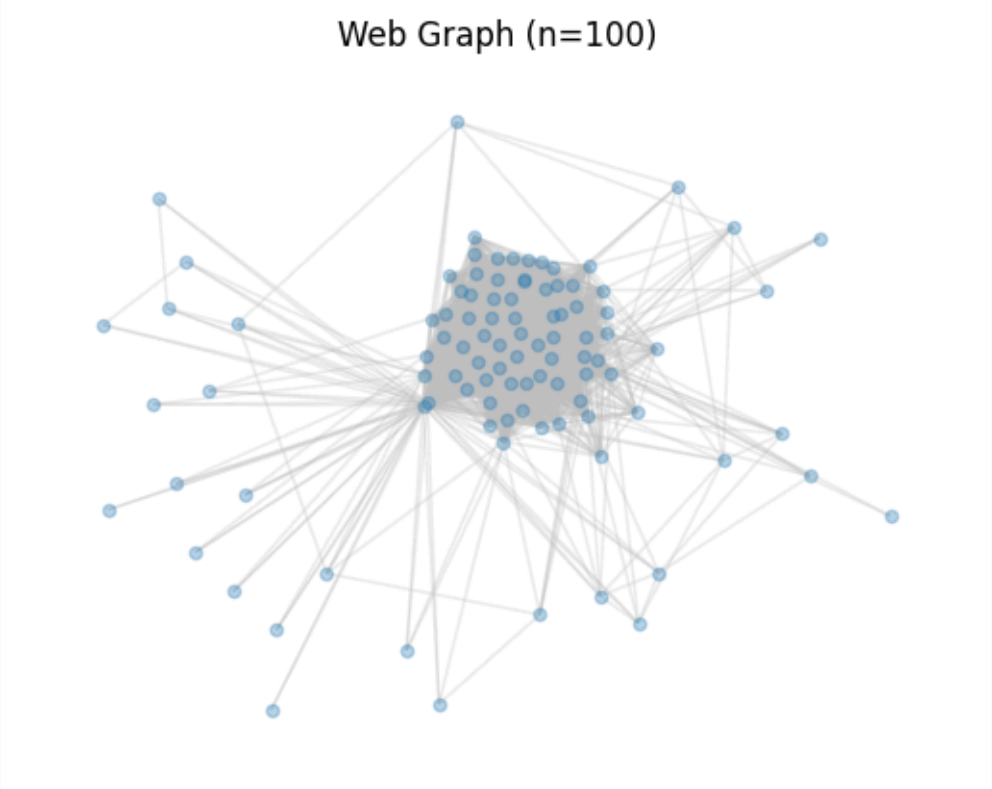

Figure 5: Web crawler network with $n = 100$

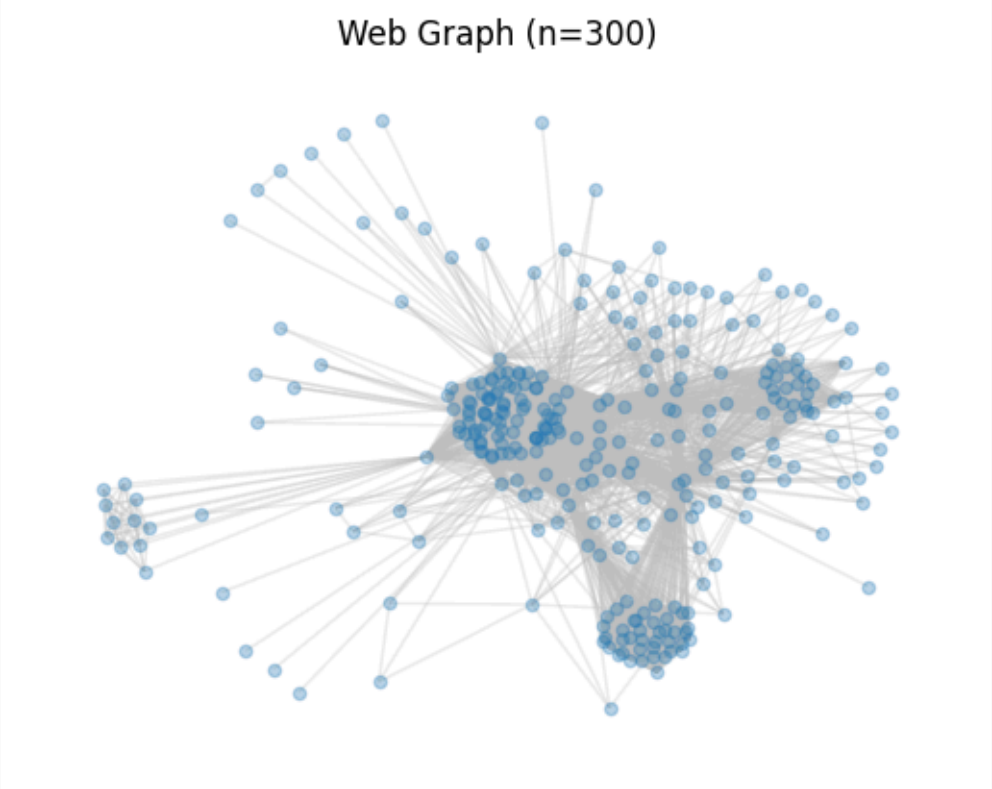

Figure 6: Web crawler network with $n = 300$

Analysis: $n = 100$ has one super well connected component and lots of little satellites. Seems kinda similar to the Internet visualized (one SCC). $n = 300$ gave more insight to

how there were actually individual clusters within the big central cluster. Once again, a lot of satellites. Spring layout with translucent nodes + edges was perfect for this since highly connected clusters would get pulled together, while the less connected websites get pushed far away from the center, and areas of overlapping nodes and edges were well-visualized by the color intensity.

3. **The Navigation Paradox**

Analyze the average shortest path using a greedy algorithm for a Watts-Strogatz graph with $n = 1000, k = 10, p = 0.1$.

Solution: Code is here.

- Average shortest path length: 4.5
- Average greedy path length: 11.69

The greedy algorithm isn't able to look ahead and take small steps towards a node that provides a much more optimal shortcut to the destination.

4. **Getting to Know Erdos Renyi**

(a) **Clustering**

(i) Calculate the expected number of triangles $\mathbb{E}[T]$, that $G(n, p)$ contains.

Solution: The number of triplets is $\binom{n}{3}$. The probability that any triple is connected in a triangle is p^3 since the probability of any edge existing is p. Thus,

$$\mathbb{E}[T] = \binom{n}{3} p^3 = \frac{n(n-1)(n-2)}{6} p^3.$$

(ii) Prove that $E[T]$ has a threshold. Specifically, find a function $\pi(n)$ s.t. $\lim_{n\to\infty} \mathbb{E}[T] = \infty$ if $p \in \omega(\pi(n))$ and $\lim_{n\to\infty} \mathbb{E}[T] = 0$ if $p \in o(\pi(n))$.

Solution: For large n, the above expected value is roughly

$$\mathbb{E}[T] \approx \frac{n^3 p^3}{6}.$$

Let $\pi(n) = \frac{1}{n}$. Since n and π multiply out to a constant, if $p \in \omega(\pi(n))$, $\mathbb{E}[T]$ explodes, while for $p \in o(\pi(n))$, $\mathbb{E}[T]$ disappears.

Let X be the event that a triangle is contained in G. Note that the previous part is insufficient to show that $\pi(n)$ is a threshold for X. But we can show this fact with second moment method – the same way we proved the result for isolated vertices in class. Since the general case is a bit nasty, we will focus on a specific setting here: $p(n) = \pi(n)\log(n)$.

(iii) Show that $\mathrm{Var}(T) \in \Theta(\log^3(n))$. Hint: Think about how to express T as a sum of binary r.v.s. Then, use casework to understand the covariance terms that result.

Solution: For each triplet i, let X_i be the random variable that is 1 if the triplet forms a triangle and 0 otherwise. We can express T as a sum of all X_i. Then, we know that the variance is

$$\operatorname{Var}(T) = \sum_i \operatorname{Var}(X_i) - 2\sum_{i<j} \operatorname{Cov}(X_i, X_j).$$

The variance of T is the variance of a Bernoulli variable with probability p^3, which is $p^3(1-p^3)$. The number of triplets scales with n^3.

The covariance is defined as

$$\operatorname{Cov}(X_i, X_j = E[X_i X_j] - E[X_i]E[X_j].$$

Notice that if the triplets i and j do not share a side, they are independent, and thus the covariance for them is 0. So for our calculations, we only consider triplets that share a side. The number of pairs of triplets that share a side is the number of combinations two pairs of points, which is

$$n_{\text{adjacent triangles}} = \binom{n}{2} \cdot \binom{n-2}{2}.$$

The probability of both triangles i and j being formed is the probability that all 5 lines that form the two triangles are present, which is p^5. Thus,

$$\begin{aligned}\operatorname{Cov}(X_i, X_j) &= E[X_i X_j] - E[X_i]E[X_j] \\ &= p^5 - p^3 p^3 \\ &= p^5 - p^6\end{aligned}$$

Now, evaluating the variance altogether (and making some approximations since we are using Θ notation),

$$\begin{aligned}\operatorname{Var}(T) &= n^3p^3(1-p^3) - 2\binom{n}{2}\binom{n-2}{2}(p^5-p^6) \\ &\approx n^3p^3 - n^3p^6 - n^4\left(p^{5-p^6}\right) \\ &\approx n^3p^3 - n^4p^5\underbrace{-n^3p^6 + n^4p^6}_{\text{trivial in magnitude}} \\ &\approx n^3p^3 - n^4p^5.\end{aligned}$$

Plugging in $p(n) = \frac{1}{n}\log(n)$, we have

$$\begin{aligned}\operatorname{Var}(T) &= n^3\frac{\log^3(n)}{n^3} - n^4\frac{\log^5(n)}{n^5} \\ &= \log^3(n)\underbrace{-\frac{\log^5(n)}{n}}_{\text{trivial}} \\ &\approx \log^3(n) \\ &\in \Theta(\log^3(n)).\end{aligned}$$

(iv) Use Chebyshev's Inequality to show that $\Pr(T = 0) \in o(1)$. Conclude that $\lim_{n\to\infty} \Pr(X) = 1$ for this particular p.

Solution: Since 0 is $n^3 \frac{p^3}{6}$ away from from the mean, the number of standard deviations it is away from the mean is

$$k \in \Theta\left(\frac{n^3}{6} \cdot \frac{\log^3(n)}{n^3} \cdot \frac{1}{\sqrt{\log^3(n)}}\right) = \Theta\Big(\sqrt{\log^3(n)}\Big).$$

Thus,

$$\Pr(T = 0) = \frac{1}{\Theta\Big(\sqrt{\log^3(n)}\Big)} \in o(1).$$

Since the probability of no triangles grows infinitely small as n increases, $\lim_{n\to\infty} \Pr(X) = 1$.

(b) **Diameter**

Suppose $p \in (0, 1)$ is held constant. Prove that the maximal diameter of $G(n, p)$ equals 2 with a probability that approaches 1 as n becomes large:

$$\lim_{n\to\infty} P(\text{diameter}(G(n, p)) = 2) = 1.$$

Solution: Notice that the maximal diameter being more than 2 is the same as there being two nodes that aren't in a triangle. For a pair of nodes to node be in a triangle, we need that there is no edge between them and no edge between them and a common node. This probability that two nodes i and j aren't in a triangle $p_{\text{no triangle}(i,j)}$ is

$$p_{\text{no triangle}}(i, j) = \underbrace{1 - p}_{\text{no direct edge}} * \underbrace{\left(1 - p^2\right)^{n-2}}_{\text{no common neighbor}}.$$

The $\left(1 - p^2\right)^{n-2}$ term comes from the fact that for a third node k, the probability that it doesn't have an edge with both i and j is $1 - p^2$.

The probability that at least one of the pairs of nodes isn't part of a triangle is less than or equal to the sum of the probabilities of no triangle for each of the pairs of nodes (union bound). That is,

$$P(\text{diameter}(G(n, p)) > 2) \leq \binom{n}{2}(1 - p)\left(1 - p^2\right)^{n-2} \approx \frac{n^2}{2}(1 - p)\left(1 - p^2\right)^{n-2}$$

Since exponential decay grows faster than quadratic increase,

$$\lim_{n\to\infty} \binom{n}{2}(1 - p)\left(1 - p^2\right)^{n-2} \approx \frac{n^2}{2}(1 - p)\left(1 - p^2\right)^{n-2} \to 0.$$

Thus,

$$\lim_{n\to\infty} P(\text{diameter}(G(n, p)) > 2) \to 0.$$

Now, we just have to show that the probabilty of all nodes being one away from each other (max diameter of 1) is 0 as $n \to \infty$. The probability that every node is connected to every other is just $p^{\binom{n}{2}}$ which obviously goes to 0 as $n \to \infty$. Thus,

$$\lim_{n\to\infty} P(\text{diameter}(G(n,p)) = 2) \to 1.$$

5. **It's a small world after all**

Consider the model in Figure 7. There are n nodes in a directed ring, labelled $A_1, ..., A_n$. Every node is connected to the next with a directed link of length 1. Central node B is connected to each of the nodes via a undirected link of length $\frac{1}{2}$ with probability p.

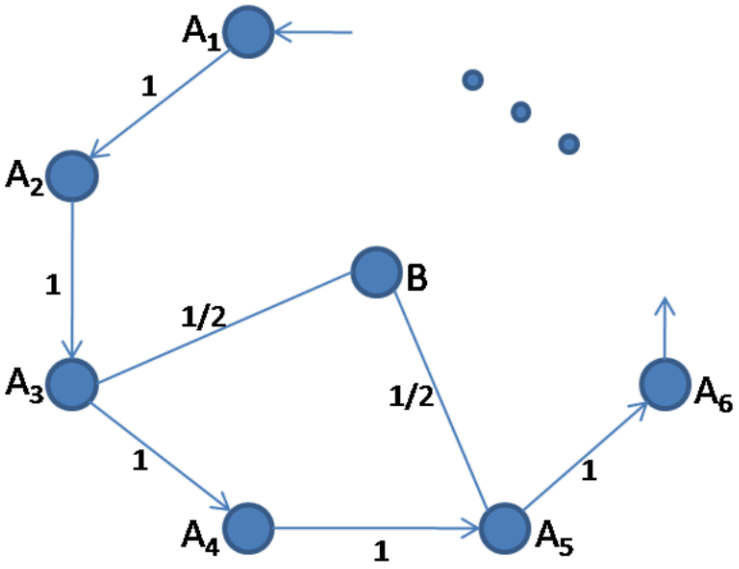

Figure 7: Graph model

(a) Consider nodes A_i and A_j, A_j being k hops away from A_i along the ring. Compute the probability $P(l, k)$ that the shortest path from A_i to A_j has length l. What is the expected value of the shortest path length from A_i to A_j?

Hint: Think about the two cases when $l < k$ and $l = k$ (there are sub-cases in the latter).

Solution: If $l = k$, one of the following has happened:

1. none of the nodes along the path are connected to B.
   - probability of $(1-p)^{k+1}$.
   - $k+1$ is the number of nodes in the path
2. exactly one of the nodes along the path is connected to B.
   - probability of $(k+1)p(1-p)^k$.
   - $k+1$ is the number of choices of the solo node to connect to B.
   - p is the probability that the chosen solo node solo node is connected to B.
   - $(1-p)^k$ is the probability that no other nodes are connected
3. there is a pair of consecutive nodes A_k, A_{k+1} that are both connected to B, and none of the other nodes are connected.
   - probability of $kp^2(1-p)^{k-1}$.
   - k is the number of choices of the pair of adjacent nodes to connect to B.
   - p^2 is the probability that the chosen adjacent pair is connected to B.
   - $(1-p)^{k-1}$ is the probability that no other nodes are connected

Thus,

$$\begin{aligned} p(k,k) &= (1-p)^{k-1}((1-p)^2 + p(1-p)(k+1) + p^2 k) \\ &= (1-p)^{k-1}(1-p+pk). \end{aligned}$$

If $l < k$, then there exists a shortcut path through B between two nodes that are $k-l+1$ apart (and no better shortcut path than this). The probability of this happening is

$$p(l<k,k) = lp^2(1-p)^{l-1}$$

- l is number of choices of pairs that are $k-l+1$ apart
- p^2 is the probability that the chosen pair is connected to B
- $(1-p)^{l-1}$ is the probability there is no better shortcut
  - $l-1$ is the number of nodes before or after the shortcut's endpoints.

Thus, the expected value of the shortest path is

$$\begin{aligned} E[L(k)] &= (1-p)^{k-1}(1-p+pk)\cdot k + \sum_{l=1}^{k-1} lp^2(1-p)^{l-1}\cdot l \\ &= k(1-p)^{k-1}(1-p+pk) + \sum_{l=1}^{k-1} l^2p^2(1-p)^{l-1}. \end{aligned}$$

(b) Compute the expected average shortest path between the nodes on the ring of the graph. How does this quantity scale with n? Contrast this with when the graph has no central node.

Solution: Since there is circular symmetry for each of the nodes, the average shortest path between the nodes is just the average shortest path between a fixed node A_i and any of the other nodes $A_j, i \neq j$ (or alternatively, iterate over all possibilities for the number of hops between A_i, A_j) which is

$$\begin{aligned} \mathbb{E}[l] &= \frac{1}{n-1}\sum_{k=1}^{n-1} E[L(k)] \\ &= \frac{1}{n-1}\left(\sum_{k=1}^{n} k(1-p)^{k-1}(1-p+pk)\right) + \frac{1}{n-1}\sum_{k=1}^{n}\sum_{l=1}^{k-1} l^2p^2(1-p)^{l-1} \end{aligned}$$

My computer is out of storage and I had to delete Mathematica. I am also lonely and have no friends who have Mathematica. I am too lazy to do geometric series by hand. Just take my points (but show some mercy please).