

Homework 2: Visualizing Networks

Assigned: 01/15/26

Due: 01/22/26 5:00pm PT

Collaboration is allowed for all problems and at the top of your homework sheet, please list all the people with whom you discussed. Crediting help from other classmates will not take away any credit from you. The details of the collaboration policy for this course are available in the Resources tab on Piazza.

*You must turn in your homework electronically via Gradescope. **Be sure to submit your homework as a single file.***

Start early and come to office hours with your questions! We also encourage you to post your questions on Piazza, as well as answer the questions asked by others on Piazza.

Coding + Data Analysis [70 points]

1. Working with real data [30 points]

We'll now look for first hand evidence of the 'universal properties' by studying a real world dataset. We'll work with a collaboration network between researchers working in the area of General Relativity and Quantum Cosmology. If an author i co-authored a paper with author j , the graph contains a undirected edge from i to j . If the paper is co-authored by k authors this generates a completely connected (sub)graph on k nodes. The original dataset contains 5242 nodes, but we'll work with the largest connected component of this network, consisting of 4158 nodes. You can download the text file from the file called **gr_qc_coauthorships.txt** on piazza under the resources tab, each line of which contains an edge in the network.

- (a) **[10 points]** Plot the histogram, as well as a complementary cumulative distribution function (ccdf) of the node degrees. Compute the average clustering coefficient, the overall clustering coefficient, the maximal diameter, and the average diameter.
- (b) **[10 points]** Calculate the number of triangles T . Assuming $T = \mathbb{E}[T]$, the expected number of triangles of a Erdős-Rényi graph with the same number of nodes (4158), calculate the parameter p of the Erdős-Rényi graph $G(n, p)$.
- (c) **[10 points]** What distribution should the node degrees of a Erdős-Rényi graph take on? Is the Erdős-Rényi model a good model for this graph (use the histogram from part a). Do we need the histogram to conclude this? Think of where the data came from.

Note: It is okay (recommended!) to use existing network libraries (e.g., networkx for python) to do the calculations. If you do this, please specify the library and the functions you are using.

2. Visualize your network [15 points]

In this problem, we will familiarize ourselves with the importance and challenges of visualizing networks. More specifically, we will be visualizing the *Erdős–Rényi* model (which we’ve seen in class), the symmetric stochastic block model (a generalization of the *Erdős–Rényi* model we will study later in the class), and the web graph that we generated in Problem Set 1.

Before starting your task, we’ll first introduce the stochastic block model, which is an important tool when studying clustering algorithms. The *stochastic block model (SBM)* is a random graph model with planted clusters. The model $SBM(n, p, W)$ defines an n -vertex random graph with labeled vertices for positive integers n , k , a probability vector p of dimension k , and a symmetric matrix W of dimension $k \times k$ with entries $W_{i,j} \in [0, 1]$. Each vertex is assigned a community label in $\{1, \dots, k\}$ independently under the community prior p , and pairs of vertices with community labels i and j connect independently with probability $W_{i,j}$. The SBM is called *symmetric* if p is uniform and if W takes the same value A on the diagonal and the same value B outside the diagonal. We can represent the symmetric SBM as $SSBM(n, k, A, B)$.

Note: For this problem, you only need to consider simple graphs with no self-loops or parallel edges.

Your task: Visualize the following three networks:

- (a) **Erdős–Rényi [4 points]:** Sample a network from the *Erdős–Rényi* model with $n = 40$ vertices where every possible edge between vertices occurs independently with probability $p = 0.23$.
- (b) **Symmetric Stochastic Block [4 points]:** Sample a network from the symmetric stochastic block model with $n = 30$ vertices, $k = 3$ communities, $A = 0.7$, and $B = 0.1$.
- (c) **The web [7 points]:** Create two networks using the first n nodes crawled by the web crawler that you created in Problem Set 1. One with $n = 100$ nodes and one with $n = 300$ nodes.

Specifically, for each network, you should:

- Write a function that can generate a network for both $G(n, p)$ and $SSBM(n, k, A, B)$. Please avoid using direct functions (e.g., `networkx.generators.random_graphs.erdos_renyi_graph`). Include your code as part of your pdf submission.
- Visualize the network. You may use existing network libraries (e.g., `networkx` for Python). Play around with the visual presentations of the networks, such as layout (e.g., `random`, `radial`, `spring`), colors, node sizes, edge thickness, etc. Consider the most effective way to convey the information presented by this type of network. Present two visualizations that provide contrasting perspectives on the network.

Note: In addition to `networkx.drawing.layout`, `Graphviz` (`networkx.drawing.nx_pydot.graphviz_layout`) also has some useful layouts to explore.

- Provide analysis of the visualization. Discuss the effectiveness of the different forms of network visualizations you explored, including whether different visualizations were more effective for the different networks. Explain any advantages or drawbacks of the approaches that you experimented with.

3. The Navigation Paradox [25 points]

In Lecture 4, we saw that adding random edges to a ring lattice drastically reduces the diameter, creating a “small world.” However, we also discussed a critical flaw: while short paths exist, decentralized agents (like people passing a letter) cannot necessarily find them if the shortcuts are purely random. In this problem, you will demonstrate this paradox by simulating a greedy search on a 1-dimensional Watts-Strogatz ring.

- (a) **[5 points]** Use a library (like `networkx`) to generate a Watts-Strogatz small world graph. Set $n = 1000$ nodes, $k = 10$ (each node connects to 5 neighbors on each side), and rewiring probability $p = 0.1$. (If the generated graph is disconnected, regenerate it until it is connected.)
- (b) **[5 points]** Since the nodes are arranged in a ring labeled 0 to $n - 1$, the distance between two nodes u and v is the minimum steps along the ring circumference:

$$d_{ring}(u, v) = \min(|u - v|, n - |u - v|)$$

Implement a helper function to calculate this distance.

- (c) **[10 points]** Select 100 random pairs of source and target nodes (s, t) and calculate the path length using two different methods:
- **Method 1:** Use the built-in shortest path function (i.e., `nx.shortest_path_length`) to find the actual distance between s and t utilizing all shortcuts optimally.
 - **Method 2:** Implement a simple greedy routing algorithm:
 - Start at node s .
 - At each step, examine the current node’s neighbors (both local ring neighbors and random shortcut neighbors).
 - Move to the neighbor that minimizes the ring distance d_{ring} to the target t .
 - Repeat until you reach t . Count the number of hops.

Implementation Note: Since the Watts-Strogatz model rewires edges, it is possible (though rare) for the greedy agent to get stuck. If this happens, simply discard that specific trial and generate a new (s, t) pair to ensure you collect 100 successful paths.

Calculate the average path length for method 1 and method 2 over your 100 samples.

- (d) **[5 points]** You should find that the true shortest path is very small (characteristic of a small world), but the greedy search is significantly worse. Why does the greedy agent fail to utilize the random shortcuts effectively?

Theory [30 Points]

4. Getting to know *Erdős–Rényi* [20 points]

In class, we’ve talked a bit about the *Erdős–Rényi* random graph model and begun to explore how well it models the four “universal” properties that we’ve been focusing on. In this problem, you will revisit two of these four properties and prove some additional results.

Recall, the *Erdős–Rényi* random graph model creates an undirected random graph, denoted by $G(n, p)$, by considering n vertices and letting every possible edge between vertices occur independently with probability p .

Since we already showed in class that the degree of any vertex in a $G(n, p)$ is binomially distributed, in this problem we will focus on clustering and diameter.

(a) Clustering [10 points]

- (i) [1 point] Calculate the expected number of triangles, denoted $E[T]$, that $G(n, p)$ contains.
- (ii) [2 points] Prove that $E[T]$ has a threshold. Specifically, find a function $\pi(n)$ such that $\lim_{n \rightarrow \infty} \mathbb{E}[T] = \infty$ if $p \in \omega(\pi(n))$ and $\lim_{n \rightarrow \infty} \mathbb{E}[T] = 0$ if $p \in o(\pi(n))$?¹

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 (not for credit, but think about why!). 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) [5 points] Show that $\text{Var}(T) \in \Theta(\log^3(n))$. **Hint:** First think of how to express T as sum of 0-1 random variables. Then, use case-work to understand the covariance terms that result.
- (iv) [2 points] Use Chebyshev’s Inequality to show that $\Pr(T = 0) \in o(1)$. Conclude that $\lim_{n \rightarrow \infty} \Pr(X) = 1$ for this particular p .

(b) Diameter [10 points]

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

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

¹In this problem, all functions are with respect to n unless otherwise stated. For a reminder on the notation, please consult https://en.wikipedia.org/wiki/Big_O-notation#Family_of_Bachmann%E2%80%9993Landau_notations

5. It's a small world after all [10 points]

We will see a few examples of small world models in class. Here, you'll develop another simple model that has small world properties.

Consider the model in Figure 1. There are n nodes arranged on a directed ring, labeled A_1, \dots, A_n . Every node is connected to the next node with a directed link of length 1. We add a new central node B to the network, and connect it with each of the nodes on the ring with probability p , using an undirected link of length $1/2$.

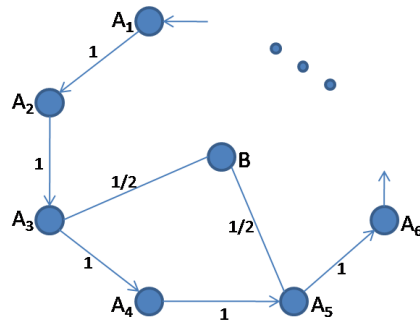


Figure 1: Graph model

For the following, **feel free to use Mathematica and the FullSimplify function if necessary to obtain closed-form expressions for the final answers, as long as you attach your code.**

- (a) **[5 points]** 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 node A_i to node A_j has length l . What is the expected value of the shortest path length from A_i to A_j ?

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

- (b) **[5 points]** Compute the expected average shortest path length between the nodes on the ring of the graph. How does this quantity scale with n ? Contrast this with the case where the graph has no central node.