RANDOM GRAPHS PROJECT

SHANE LUBOLD
DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON

Description: In this project, we will analyze a common random graph model, the Erdős–Rényi model. We will use basic probability theory and simulations to understand the properties and behavior of this random graph model, with an emphasis on these properties as the size of the graph grows. To begin, we will first review basic graph theory and probability theory. Then, we will examine the degree distribution of nodes in this model, the expected number of cliques, and study an important property in random graph theory: the "threshold property". In addition, we will examine the probabilistic method, an interesting proof technique that is frequently used to prove statements about random graph models.

Overview of Project

(1) Week 1: Review graph theory fundamentals and basic probability.
(2) Week 2: Introduce Erdős–Rényi graph model and basic simulations.

1. Graph Theory and Probability Review (Week 1)

We first review basic graph theory concepts.

Definition 1.1 (Graph)**.** *A graph $G = (V, E)$ is a set of vertices (or nodes) $V = \{1, \dots, n\}$ and a set of edges $E \subseteq V \times V$.*

A simple way to represent the edges in a graph is with an adjacency matrix.

Definition 1.2 (Adjacency matrix)**.** *For a graph $G = (V, E)$ on n nodes, the adjacency matrix A of G is an $n \times n$ matrix such that $A_{ij} = 1$ if there is an edge between nodes i and j and zero otherwise.*

Exercise 1.1. *Let $V = \{1, 2, 3\}$ be the vertex set of a graph G. How many graphs are there on this vertex set? Draw all possible graphs on this vertex set. Determine the total number of graphs on $V = \{1, 2, 3, 4\}$ and try to extend this formula to the general case when $V = \{1, 2, \dots, n-1, n\}$ (This is more challenging. To do this, think about how many $n \times n$ symmetric matrices there are with zero on the diagonal and only 0 or 1 on the off-diagonal. Then note that for each adjacency matrix, there is a corresponding graph).*

Definition 1.3 (Path)**.** *A path between nodes i and j is a sequence of distinct edges $(e_1, \dots, e_m)$ connecting i and j such that all nodes along the sequence of edges are distinct.*

Informally, a path is a sequence of edges that allows one to "walk" from node i to node j along edges in G.

Definition 1.4 (Connected Nodes and connected Graph)**.** *Two nodes i and j are connected if there exists a path between i and j. A graph is called connected if all $\binom{n}{2}$ pairs of nodes are connected.*

Exercise 1.2. *Draw an example of a graph on 5 nodes $V = \{1, 2, 3, 4, 5\}$ such that three of the nodes are connected and the other two are not connected.*

Definition 1.5 (Isolated Node)**.** *A node is called isolated if it does not have an edge with any other node in the graph.*

Definition 1.6 (Clique)**.** *A clique $C \subseteq V$ of a graph $G = (V, E)$ is a set of nodes such that for every pair (i, j) in C, i and j are connected. That is, the subgraph induced by C is complete.*

Exercise 1.3. *Draw an example of a graph on 5 nodes $V = \{1, 2, 3, 4, 5\}$ with a 3 clique and a graph on 5 nodes with a 5 clique. (Note that a graph on n nodes containing an n clique is called a complete graph on n nodes, denoted by K_n).*

We now introduce the degree of a node.

Definition 1.7. *Let $i \in V$ be a node of a graph $G = (V, E)$ with adjacency matrix A. Then, the degree of node i, denoted by $d(i)$, is the number of edges connecting i to other nodes in the graph; that is,*

$$d(i) = \sum_{j \in V} A_{i,j} \ .$$

We now review some basic probability theory that will be used throughout this project.

Definition 1.8 (Expected value)**.** *The expected value of random variable $X \in \{0, 1, 2, 3, \dots\}$ with probability mass function p is defined as*

$$E(X) = \sum_{x=0}^{\infty} x \cdot p(x)$$

For a sequence $\{X_n\}$ of random variables, it always holds that $E(\sum_{n=1}^{N} X_n) = \sum_{n=1}^{N} E(X_n)$, even if the $\{X_n\}$ are dependent.

Proposition 1.1. *Let X be a random variable and let I be an indicator random variable that is 1 when $X \in A$ and 0 when $X \notin A$. Then,*

$$E(I) = P(X \in A) \, .$$

Exercise 1.4. *Prove this result.*

As an example, let $X \sim \mathrm{N}(0, 1)$ and let $A = [0, \infty)$. Then, if I is an indicator random variable that is 1 when $X \in A$ and 0 otherwise, then $P(X \in A) = 1/2$ (check this!), so that $E(I) = 1/2$ too.

Definition 1.9. *Let n be an integer and $p \in [0, 1]$. A discrete random variable X has a binomial distribution Binomial(n, p) when the probability mass function of X is*

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, 1, \dots, n-1, n\} \, .$$

Exercise 1.5. *Show that if X has a Binomial(n, p) distribution then $E(X) = np$.*

Definition 1.10. *Let $\lambda > 0$. A discrete random variable X has a Poisson(λ) distribution when the probability mass function of X is*

$$\mathbb{P}(X = k) = \frac{\lambda^k \exp(-k)}{k!} \, .$$

Exercise 1.6. *Prove that if $X \sim$ Poisson(λ), then $E(X) = \lambda$.*

Next week, we will show an interesting relationship between the binomial and Poisson random variables that will be helpful when studying random graphs.

Definition 1.11. *Let A and B be events. A and B are said to be independent when*

$$P(A \text{ and } B) = P(A)P(B)$$

2. Erdős–Rényi model (Week 2)

We now introduce the Erdős–Rényi (ER) model. To do this, we fix an integer n and a value $p \in [0, 1]$, which controls the probability nodes connect to each other. To simulate an ER model, we do the following: For each pair of nodes $(i, j) \in V^2$ with $i < j$, add an edge between nodes i and j independently with probability p. There

are a total of $\binom{n}{2} = \frac{n(n-1)}{2}$ possible edges in the graph.

In the previous section, we described the degree of a node i for a deterministic graph. Now, when discussing a random graph, the degree of a node for each simulation is random, and so we therefore want to talk about the *degree distribution* of a node.

Exercise 2.1. *Simulate 1000 ER graphs with $n = 10, p = 1/5$.*

(1) *For each graph, compute the degree of node 1, using the formula*

$$deg(1) = \sum_{j=1}^{10} A_{1j} \ ,$$

where A is the adjacency matrix of the graph. Make a histogram of these 1000 values using the command hist(deg_1), where deg_1 is the vector that contains the degrees of the 1000 graphs you generated. Compute the expected value of deg_1 and the variance of deg_1 using the R commands mean(deg_1) and var(deg_1). Next we will show that these values are equal to $p(n-1)$ and $p(1-p)(n-1)$.

(2) *We will also examine numerically the probability that node 1 is isolated. Recall that this means that node 1 does not connect to any other node. For each simulation i, compute the variable*

$$isolated_i := I(\text{node 1 is isolated})$$

To check if a node is isolated, we can just check if the sum of row 1 of A is zero, because if this row sum is zero, then node 1 does not connect to any other nodes. So,

$$isolated_i = I\left(\sum_{i=1}^{n} A_{1,j} = 0\right) \ .$$

Compute

$$\frac{1}{1000} \sum_{i=1}^{1000} isolated_i \ , \tag{1}$$

where $isolated_i$ is the value of isolated for the ith simulation. This should approximate the probability that node 1 is isolated, which we will show next week is equal to $(1-p)^{n-1}$. Compare the true value of $(1-p)^{n-1}$ to the approximation you found in (1) when $n = 10$ and $p = 1/5$. Are these numbers close?

To describe the degree distribution of a node in an ER model, we need the following definition.

Definition 2.1 (Binomial Distribution)**.** *Let $X_1, \ldots, X_n$ be a sequence of independent and identically distributed Bernoulli random variables. That is,*

$$X_i = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1-p \ . \end{cases}$$

Then, $\sum_{i=1}^{n} X_i$ is distributed according to a Binomial distribution with parameters n and p.

Exercise 2.2. *Recall that for independent variables $X_1, \ldots, X_n$:*

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) \ .$$

That is, the variance of a sum of independent variables is just the sum of their individual variances. Using this result, and the representation of a Binomial distribution given in Definition 2.1, prove that the variance of a Binomial(n, p) random variable is $np(1-p)$.

We have the following result:

Proposition 2.1. *The degree distribution of any vertex i is Binomial$(n-1, p)$; that is,*

$$\mathbb{P}\left(d(i) = k\right) = \binom{n-1}{k} p^k (1-p)^{n-1-k} \ .$$

Exercise 2.3. *Prove this result using the definition of a binomial random variable given in Definition 2.1 as well as the definition of a degree, which is*

$$deg(i) = \sum_{j=1, j\neq i}^{n} A_{i,j} \ .$$

From this result, we know a lot about the expected degree of a node and its variance, which we prove below.

Proposition 2.2. *Let G be an ER graph on n nodes with parameter $p \in [0,1]$. Then,*

$$E(deg(i)) = p(n-1), \quad var(deg(i)) = p(1-p)(n-1) \ .$$

In addition, we have that $\mathbb{P}(deg(i) = 0) = (1-p)^{n-1}$. From this, we see that $\mathbb{P}(\text{node } i \text{ is isolated}) = (1-p)^{n-1}$.

Exercise 2.4. *Prove the result above, using the expression for the variance of a Binomial random variable from Exercise 2.2.*

3. Cliques

In the ER model, it is possible to write in closed form the expected number of cliques. To do this, we will introduce a common probabilistic method that is often used when determining the expected number of objects (in our case, cliques) that appear.

To make things as concrete as possible to begin, we will first study the expected number of triangles (3-cliques) in an ER model. To do this, note that if we define

$\Delta_{i,j,k}$ to be an indicator that is equal to one if nodes $i,\ j,\ k$ are in a triangle, then we can write

$$X = \sum_{i,j,k}^{n} \Delta_{i,j,k}, \text{ where } \quad \Delta_{i,j,k} = \begin{cases} 1, & \text{with probability } p^3 \\ 0, & \text{with probability } 1 - p^3 \end{cases}$$

To find $E(X)$, the expected number of triangles in G,

$$E(X) = \binom{n}{3} E(\Delta_{i,j,k}) = \binom{n}{3} p^3 \ .$$

For example, set $p(n) = \frac{d}{n}$ for some constant d. Then, $E(X) = \binom{n}{3} \frac{d^3}{n^3} \approx \frac{d^3}{6}$.

Exercise 3.1. *Simulate 1000 ER graphs with $n = 10$ and $p = 1/2$. For each graph, compute the number of triangles and compute the average number of triangles in the 1000 graphs. Compare this average to the true expected value, which by the previous work is $\binom{10}{3}(1/2)^3 = 15$. You can use the following result that relates the trace of the cube of the adjacency matrix to the number of triangles in a graph:*

$$\textit{trace}(A^3)/6 = \textit{number of triangles in } G \ .$$

If one is interested in the distribution of X, then one might try to argue that since X is a sum of identically distributed Bernoulli random variables, then X would have a binomial distribution. However, these Bernoulli random variables are not independent. The following exercise will help show why this is the case.

Exercise 3.2. *To understand why the $\{\Delta_{ijk}\}_{i,j,k=1}^{n}$ are not independent, consider two random variables $\Delta_{1,2,3}$ and $\Delta_{1,2,4}$. Recall that two random variables are dependent if when we condition on one of their values, the distribution of the other variable changes. In this exercise, we will use this definition to show that $\Delta_{1,2,3}$ and $\Delta_{1,2,4}$ are dependent.*

1. *Compute $\mathbb{P}(\Delta_{ijk} = 1)$.*
2. *Now compute $\mathbb{P}(\Delta_{1,2,3} = 1 | \Delta_{1,2,4} = 1)$ by using the formula $\mathbb{P}(A|B) = \mathbb{P}(A \text{ and } B)/\mathbb{P}(B)$. Are these two values the same? Give conditions on the value p such that these two random variables are independent.*

The following exercise will derive the expected number of cliques in a ER model.

Exercise 3.3. *Let X denote the number of k-cliques in an ER model for an integer k. Using the same argument as above, show that*

$$E(X) = \binom{n}{k} p^{\binom{k}{2}} \ .$$

(Hint: try the case when $k = 4$, since $k = 3$ was done above. Think about what must happen for four nodes to all be connected. How many edges must exist between the four nodes?).