# 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*

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

**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$(\lambda)$ 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$(\lambda)$, 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*

$$deg(1) = \sum_{j=1}^{100} A_{1j} \ ,$$

*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 $i$th 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?*

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 2.2.** *Prove this result.*

From this result, we know a lot about the expected degree of a node and its variance. In particular, we know that $E(d(i)) = (n-1)p$ (which you proved last week) and $\text{Var}(d_i) = p(n-1)(1-p)$. We can also compute the probability that a node is *isolated*; that is, that is connects to no other node. By setting $k = 0$, it is given by $\mathbb{P}(i \text{ is isolated}) = (1-p)^{n-1}$. Similarly, by setting $k = n-1$, we see that that the probability that node $i$ is connected to all other nodes in the graph is given by $p^{n-1}$.

**Exercise 2.3.** *Simulate 1000 ER graphs with $n = 10,\ p \in \{1/10, 1/2, 9/10\}$. For each graph and each value of $p$, compute the probability that the first node is isolated and that is is connected to all other nodes in the graph, and compare to the values given in the paragraph above.*

As is the case for most of the results in this project, we are interested in the behavior of the ER model as the graph gets larger (i.e., as $n \to \infty$). To explore this, we assume that $p = p(n)$ is a function of $n$ and we simulate $G \sim G(n, p(n))$ as $n \to \infty$. We have the following result about the degree distribution of such a graph:

**Proposition 2.2.** *Assume that $np(n) \to c$ as $n \to \infty$. Then,*

$$\mathbb{P}\left(d(i) = k\right) \to \frac{c^k e^{-c}}{k!} .$$

*That is, the degree distribution of any node approaches a Poisson distribution with parameter $c$.*

**Exercise 2.4.** *Prove this result. (Hint: First show that if $x_n \to x$ as $n \to \infty$, then $\left(1 - \frac{x_n}{n}\right)^n \to e^{-x}$ as $n \to \infty$. To show this, use the fact that $\left(1 - \frac{1}{n}\right)^n \to e^{-1}$ and then use the continuity of limits of continuous functions.*

**Exercise 2.5.** *Assume that we take $p(n) = \frac{c}{n}$ for some real number $c$. Prove that $\mathbb{P}(\text{node 1 is isolated}) \to e^{-c}$ as $n \to \infty$. Now, set $c = 1$ and simulate 1000 graphs with $n \in \{100, 1000, 10000\}$. For each value of $n$, count the number of times node 1 is isolated and compare this to $e^{-1}$. Does it seem to be getting closer to $e^{-1}$?.*

**Exercise 2.6.** *Pick a sequence $p_1(n)$ such that $np_1(n) \to 1$ and a sequence $p_2(n)$ such that $np_2(n) \to 2$. For each sequence, simulate 1000 ER graphs with $n \in \{100, 1000, 10000\}$. For each value of $n$, plot a histogram of the degree of node 1 and compare it to distribution from the proposition above.*

Finally, we will prove two basic probability theory tools that will be used often in the following sections.

**Proposition 2.3** (Markov's Inequality)**.** *If $X$ is a non-negative random variable and $a > 0$, then $P(X \geq a) \leq \frac{E(X)}{a}$ .*

**Exercise 2.7.** *In this exercise we will prove Markov's inequality for discrete random variables (i.e., random variables taking values in $\{0, 1, \ldots, \}$.) To do this, first express $E(X)$ as*

$$E(X) = \sum_{x=0}^{\infty} xp(x) = \sum_{x=0}^{a-1} xp(x) + \sum_{x=a}^{\infty} xp(x) \; .$$

*Then, drop one of the two terms above and bound the remaining term to obtain the term $aP(X \geq a)$. Then, divide by $a$ to finish the proof.*

**Exercise 2.8** (Boole's Inequality)**.** *The following inequality holds for any events $A_1, \ldots, A_n$ :*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i) \; .$$

*To prove that this inequality holds, note that if the $\{A_i\}$ are mutually disjoint (meaning that for each pair $A_i,\ A_j$ it holds that $A_i \cap A_j = \emptyset$), then the above inequality is in fact an equality. Now, by representing events as circles in a Venn diagram, argue why the above inequality always holds. See the Wikipedia page on "Inclusion-Exclusion principle" to see what it meant by Venn diagram and how it is helpful for this problem.*

## 3. Increasing Properties and Threshold Functions

We begin with a definition of *monotone increasing*. Informally, a property is monotone increasing whenever adding an edge to a graph does not destroy the property.

**Definition 3.1.** *Let $P$ be a property of a graph $G = (V, E)$. We say that $P$ is* monotone increasing *if $P$ also holds for the graph $\tilde{G} = (V, E \cup \{e\})$ where $e$ is any vertex not in $G$.*

Some examples of monotone increasing functions include the existence of a $k$-clique, connectivity. A monotone increasing property is called non-trivial whenever the empty graph never has this property and the complete graph always has this property. We use the notation $\mathbb{P}(G_{n,p} \in P)$ to denote the probability that $G_{n,p}$ has property $P$.

**Definition 3.2.** *A function $\tilde{p}(n)$ is a* threshold *for a monotone increasing property $P$ if*

$$\lim_{n\to\infty} \mathbb{P}\left(G_{n,p} \in P\right) = \begin{cases} 1, \frac{p(n)}{\tilde{p}(n)} \to \infty \\ 0, \frac{p(n)}{\tilde{p}(n)} \to 0 \; . \end{cases}$$

Note that the threshold function $\tilde{p}$ above are not unique, since $\tilde{p} + o(\tilde{p})$ also satisfies the definition. For example, $\tilde{p} + 1$ is a threshold function for $P$ whenever $\tilde{P}$ is a threshold function for $P$.

**Proposition 3.1.** *Every non-trivial monotone increasing property has a threshold function.*

We will now start investigating how to find threshold functions for various non-trivial graph properties. To begin, we will consider the property $P$ "the graph has at least one edge ." To do this, we will use the following method:

**Proposition 3.2.** *The function $\tilde{p}(n) = 1/n^2$ is a threshold function for the property "the graph has at least one edge."*

*Proof.* To show this, for $1 \leq i \leq \binom{n}{2}$, we define a random variable

$$X_{n,i} = \mathbf{1}(\text{edge } i \text{ exists})$$

And we define $X_n = \sum_{i=1}^{\binom{n}{2}} X_{n,i}$ to be the sum of these $X_{n,i}$. Note that $X_n$ counts the number of edges in $G_{n,p}$. By linearity of expectation, we know that

$$E(X_n) = \sum_{i=1}^{\binom{n}{2}} E(X_{n,i}) = \binom{n}{2} E(X_{n,1}) = \binom{n}{2} p \ .$$

Similarly, since the $X_{n,i}$ are mutually independent, we have that

$$\mathrm{Var}(X_n) = \sum_{i=1}^{\binom{n}{2}} \mathrm{Var}(X_{n,i}) = \binom{n}{2} \mathrm{Var}(X_{n,1}) = \binom{n}{2} p(1-p) \ .$$

We can also derive these expressions from the fact that $X_n$ is a sum of $\binom{n}{2}$ independent and identically distributed Bernoulli random variables, each with probability $p$. If $p = o(1/n^2)$, then

$$E(X_n) = \binom{n}{2} p = \frac{n(n-1)}{2} p \sim n^2 p \to 0 \ ,$$

as $n \to \infty$.

□

## 4. Existence of triangles and other cliques (Week 3)

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 4.1.** *Simulate 1000 ER graphs with $n = 100$ and $p(n) = \frac{1}{n}$. For each graph, compute the number of triangles and compute the average number of triangles in the 1000 graphs. Compare it to $\frac{1}{6}$.*

**Exercise 4.2.** *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, that $X$ would have a binomial distribution. However, these bernoulli random variables are not independent. Why?*

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

**Exercise 4.3.** *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?).*

**Exercise 4.4.** *Simulate 1000 ER graphs with $n = 100, p = 1/3$. Compute the number of cliques of size $k \in \{3, 5, 7\}$ and compare that to the expected value in the proposition above.*

Above, we used a very common probabilistic technique to count the number of cliques in an ER graph by writing $X$ as a sum of indicator functions. To appreciate how useful this technique is, consider the following exercise, in which we will use this same approach to count the number of fixed points in a random permutation.

**Exercise 4.5.** *Let $S_n = \{\pi : \pi$ is a permutation on $\{1, \dots, n\}\}$, where a permutation on $\{1, \dots, n\}$ is a bijection from this set to itself. We say that $i \in \{1, \dots, n\}$ is a fixed point of $\pi$ if $\pi(i) = i$; that is, $\pi$ does not change the value of $i$. First, show that $|S_n| = n!$. Then, by drawing $\pi$ uniformly from $S_n$, show that the expected number of fixed points of $\pi$ is 1.*

**Proposition 4.1** (First Moment Method)**.** *Let $X$ be an integer-valued random variable. Then, $P(X > 0) \leq E(X)$.*

*Proof.* Note that since $X$ only takes integer values, we have that $P(X > 0) = P(X \geq 1)$. By Markov's inequality, we have that $P(X \geq 1) \leq E(X)$, so that $P(X > 0) \leq E(X)$, as claimed. □

We will use this proposition to prove the following result.

**Proposition 4.2.** *If $X_n$ is a sequence of integer-valued random variables and $E(X_n) \to 0$, then $P(X_n = 0) \to 1$ as $n \to \infty$.*

*Proof.* By the result above, we have that $0 \leq \lim P(X_n > 0) \leq \lim E(X_n) = 0$, so that $\lim P(X_n = 0) = 1$, as claimed. □

Now, a natural question to ask is then, if $E(X_n) \to \infty$ then is is true that $P(X_n > 0) \to 1$? The answer is no, as the following example shows.

**Example 4.1.** *Let $X_n$ be an integer-valued random variables such that*

$$X_n = \begin{cases} n^2, & \text{with probability } 1/n \\ 0, & \text{with probability } 1 - 1/n . \end{cases}$$

*Then it is easy to see that $E(X_n) = n \to \infty$ but $P(X_n = 0) \to 1$.*

We now study the threshold properties of ER models. For example, what is the probability that $G(n, p(n))$ contains a clique of size 4 as $n \to 0$? Clearly, as $p \to 0$, this will happen with smaller and smaller probability, but as $p \to 1$, then this will happen with probability closer and closer to 1. Therefore, one can ask, what is the "cutoff" for this behavior? That is, what is the behavior of $p(n)$ that will guarantee that $G(n, p(n))$ has a four clique? Before doing this, we first introduce some notation.

**Definition 4.1.** *We say that a sequence* $\{x_n\}_{n=1}^{\infty}$ *is* $o(n^\alpha)$ *if* $\frac{x_n}{n^\alpha} \to 0$ *as* $n \to \infty$.

**Exercise 4.6.** *Give an example of a sequence that is* $o(n)$ *and one that is* $o(n^2)$. *Now prove that if* $\{x_n\}$ *is* $o(n^\alpha)$ *then* $\{x_n\}$ *is* $o(n^\beta)$ *when* $\alpha < \beta$. *Finally, prove that if* $\{x_n\} = o(n^\alpha)$, *then* $\{x_n^\beta\} = o(n^{\alpha\times\beta})$.

We will use the first moment method to prove the following result:

**Proposition 4.3.** *If* $p(n) = o(n^{-2/3})$, *then as* $n \to \infty$,

$$\mathbb{P}\left(G(n,p) \text{ contains a clique of size at least } 4\right) \to 0 \ .$$

*Proof.* Let $I$ be the set of 4-tuples containing the indices $\{1,\dots,n\}$. Then, we define $X_i$ to be an indicator variable that is 1 when the four nodes in the $i$th 4-tuple are all connected. Then, we define $X \equiv \sum_{i\in I} X_i$, and

$$E(X) = E\left(\sum_{i\in I} X_i\right) = \binom{n}{4} p^6 \sim \frac{n^4 p^6}{24} \ .$$

The second equality follows since $|I| = \binom{n}{4}$. Since $p(n) = o(n^{-2/3})$, we have that $n^4p^6 = o(1)$. To show this, note that if $p(n) = o(n^{-2/3})$, then $p(n)^6 = o(n^{-4})$, so that $n^4p^6 \to 0$. So that $E(X) \to 0$. Therefore, by the first moment method, we have that $0 \le P(X > 0) \le E(X) \to 0$, so that $P(X > 0) \to 0$. Therefore, since $P(X > 0) = P(G(n,p)$ contains a clique of size 4), we have that $\mathbb{P}\left(G(n,p) \text{ contains a clique of size } 4\right) \to 0$, as claimed. □

We will now extend this in the following result:

**Proposition 4.4.** *If* $p(n) = o(n^{-\frac{2}{k-1}})$, *then as* $n \to \infty$,

$$\mathbb{P}\left(G(n,p) \text{ contains a clique of size at least } k\right) \to 0 \ .$$

**Exercise 4.7.** *Prove this result using the same argument as above.*

In this section, we gave sufficient conditions on the $p(n)$ that guarantee $k$ cliques in an ER model. Next week, we will provide necessary conditions and will therefore provide a complete understanding of the relationship between $p(n)$ and the existence of k-cliques.
We can specialize this to the case of a triangle.

**Proposition 4.5.** *If* $p(n) = o(n^{-1})$, *then* $P(G(n,p(n))$ *contains a triangle*$) \to 0$ *as* $n \to \infty$.

We will demonstrate that this is true using simulations. To do this, we will use the following result.

**Proposition 4.6.** *The number of triangles in an undirected graph is* $tr(A^3)/6$.

**Exercise 4.8.** *Prove this result.*

## 5. Connectivity of ER Models

Last week, we talked about the existence of $k$-cliques in ER models as the size of the graph grows with $p = p(n)$. Now, we will study another property of ER models, the notion of "connectedness." Note that if we keep $p$ fixed as $n \to \infty$, then

$$\mathbb{P}\left(G(n,p(n)) \text{ is connected}\right) \to 0 \ .$$

To show this, note that for any node $i$,

$$\mathbb{P}\left(\text{node } i \text{ is isolated}\right) = (1-p)^{n-1} \to 0 \ ,$$

as $n \to \infty$ if $p$ is fixed. In fact, we can prove that if $p$ is fixed, the probability that $G$ is not connected is

$$\mathbb{P}\left(G(n,p) \text{ is not connected}\right) = \mathbb{P}\left(\text{node } i \text{ is isolated}, i = 1,\dots,n\right) \le np^{n-1} \to 0 \ ,$$

where the inequality follows from the Booles' inequality. Therefore, in order for $G$ to be connected as $n \to \infty$, it must be that $p$ changes with $n$. We now investigate what behavior $p$ must have to guarantee that $G$ is connected with probability 1 as $n \to \infty$. That is, we want to characterize the behavior of $p(n)$ that will guarantee that

$$\mathbb{P}\left(G(n, p(n)) \text{ is connected}\right) \to 1 \ ,$$

as $n \to \infty$. To do this, we will use the same method as we did before.

**Proposition 5.1.** *If* $p(n) = \lambda \frac{\log(n)}{n}$ *and* $\lambda < 1$*, then*

$$\mathbb{P}\left(G(n, p(n)) \textit{ is connected}\right) \to 0 \ .$$

*If* $p(n) = \lambda \frac{\log(n)}{n}$ *and* $\lambda > 1$*, then*

$$\mathbb{P}\left(G(n, p(n)) \textit{ is connected}\right) \to 1 \ .$$

*Proof.* We will only prove the second claim in the proof. To do this, we will use the first-moment method. To do this, we define $X_i$ to be an indicator variable that is 1 when node $i$ is isolated and zero when node $i$ is not isolated. Define $X = \sum_{i=1}^{n} X_i$ to be the sum of isolated nodes. Then, we will show that $E(X) \to 0$, which will show that $G$ becomes connected with probability 1. Note that

$$E(X) = E\left(\sum_{i=1}^{n} X_i\right) = nE(X_1) \ ,$$

where have used the fact that the $X_i$ are identically distributed. Then,

$$E(X) = n(1-p)^{n-1} \ .$$

We now use the bound $n(1-p)^{n-1} < ne^{-pn}$ to conclude that

$$E(X) \leq ne^{-pn} \ .$$

To show that this goes to 1, we take the log of both sides to conclude that

$$\log(E(X)) \leq \log(n) - pn \ .$$

If $p(n) = \lambda \frac{\log(n)}{n}$, then

$$\log(E(X)) \leq \log(n) - \lambda \log(n) = \log(n)\left(1 - \lambda\right) \ .$$

If $\lambda > 1$, then clearly $E(\log(X)) \to -\infty$, so that $E(X) \to 0$, which means by the first moment method that $P(X = 0) \to 1$, so that the probability that the graph is connected goes to 1. □

We now review the basic facts that we have learned about the ER model:

(1) The degree distribution of any node is Binomial$(n-1, p)$.
(2) If $np(n) \to c$ for a constant $c$, then the degree distribution of any node converges to a Poisson random variable with parameter $c$.
(3) The expected number of $k$-cliques in an ER graph is $\binom{n}{k} p^{\binom{k}{2}}$. This allows us to determine the behavior of $p(n)$ that guarantees that there exists a $k$-clique with probability 1 as $n \to \infty$. We proved this using the first moment method. We showed that the threshold for a $k$-clique is $o(n^{-2/(k-1)})$.
(4) Using the first moment method, we also showed that if $p(n) = \lambda \frac{\log(n)}{n}$ for some $\lambda > 1$ then the ER graphs becomes connected with probability 1, and if $\lambda < 1$ then the ER graphs becomes disconnected with probability 1 as $n \to \infty$.

## 6. Barabasi-Albert model

In this section we now look at a different random graph model called the Barabasi-Albert (BA) model. To generate a random graph according to this model, we start with an arbitrary graph on $m_0$ nodes. To add a new node to this graph, we add a connection between the new node and node $i$ with probability

$$p_i = \frac{d(i)}{\sum_{j=1}^{n} d(j)}\ , \tag{2}$$

where recall that $d(j)$ is the degree of node $j$. That is, the more edges node $i$ currently has, the more likely it is to connect to a new node in the graph. This is an example of a scale-free network, meaning that the degree distribution of a node follows a power law. That is, the degree of a node has a density that is of the form $\mathbb{P}(d(i) = k) = k^{\gamma}$ for some $\gamma > 0$. This is different from the degree distribution of the ER model, in which the degree distribution of a node was Binomial, which is not of this form.

**Exercise 6.1.** *We will generate some BA graphs. To do this, fix $m_0 = 5$. Add edges between nodes 1 and 2, 2 and 4, and 3 and 5. In the steps below, we will see how to add a node to the graph. To do this, consider the following pseudocode:*

1. *Create the adjacency matrix $A_5$ that corresponds to the graph on 5 nodes as described in the paragraph above.*
2. *Create a $6 \times 6$ matrix $A_6$ where the first 5 rows and 5 columns are the same as the entries of $A_5$. Then, we want to add a new node to the graph. To do this, we consider $[A_6]_{1,6}$, the $(1, 6)$ entry of the new adjacency matrix $A_6$. We compute $p_1$ as in (2) and set $[A_6]_{1,6} =$ Bernoulli$(p_1)$. Similarly, for $[A_6]_{i,6}$ for $i = 2, \ldots, 5$, we set $[A_6]_{i,6} =$ Bernoulli$(p_i)$. Now make $A_6$ symmetric by setting $[A_6]_{6,i} = [A_6]_{i,6}$ for $i = 1, \ldots 5$. This is the adjacency matrix of the graph on 6 nodes.*

**Exercise 6.2.** *Read the following article written by Barbasi on scale-free networks and their use.* `https://science.sciencemag.org/content/sci/325/5939/412.full.pdf`.

## 7. Stochastic Block Model and Random Geometric Graph Model:

The stochastic block model is a commonly used extension of the Erdős–Rényi model. The parameters of the stochastic block model are the following:

1. $n$ nodes
2. A partition of the vertices $\{1, \ldots, n\}$ into sets $C_1, \ldots, C_r$ called communities
3. An $r \times r$ matrix $P$ with $0 \leq P_{i,j} \leq 1$ for all $i$ and $j$.

Then to generate the stochastic block model, we pick two nodes $v_i$ and $v_j$ at random from $\{1, \ldots n\}$, where $v_i$ is in $C_i$ and $v_j \in C_j$. We add an edge between these two nodes with probability $P_{i,j}$. It is easy to see that if $P$ is a constant matrix, then the stochastic block model reduces to the Erdős–Rényi model. The stochastic block model is therefore a generalization of Erdős–Rényi to the case when we expect some latent characteristic to affect the probability of edges between nodes. Therefore, the stochastic block model does not assume the edges are identically distributed, but it still assumes that they are independent. In the next exercise, we will generate a stochastic block model matrix.

**Exercise 7.1.** *Take $n = 10, r = 3$ and assume that nodes 1, 2,3 are in community 1, nodes 4, 5, 6, 7 are in community 2, and nodes 8, 9, 10 are in community 3. Assume the following matrix $P$ for the cross-community probabilities:*

$$P = \begin{pmatrix} 1/2 & 1/3 & 1/5 \\ 1/3 & 2/3 & 1/3 \\ 1/5 & 1/3 & 1/2 \end{pmatrix}\ .$$

*Generate a stochastic block model using these parameters.*

An important task people try to solve is community detection. First, they want to determine if there is community structure in the graph they observe? In other words, they try to determine if the matrix $P$ is constant or not. Second, if they believe that there is community structure, they attempt to determine which nodes are in which communities.

The second model we are discussing today is the random geometric graph. To specify this model, we need

(1) $n$ nodes
(2) A distribution over $[0,1)^2$
(3) a parameter $r \in (0,1)$.

To generate a graph from the random geometric graph model, we distribute the $n$ nodes according to the distribution on $[0,1)^2$. For example, one could take this distribution to be uniform on $[0,1)^2$. Then, for each pair of nodes, we add an edge between them if the distance between these locations is less than $r$. Here, for points $(x,y) \in [0,1)$, the distance between $x$ and $y$ is just $d(x,y) = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2}$.

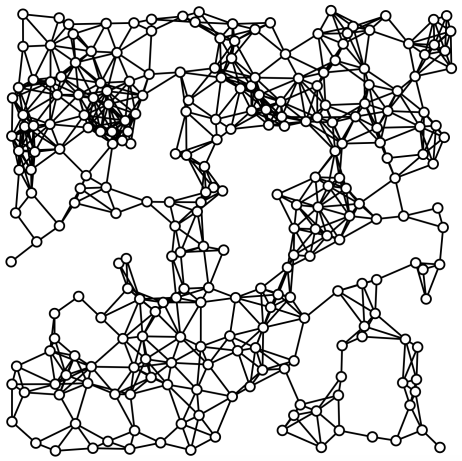

FIGURE 1. Example of a geometric random graph. Image from `https://upload.wikimedia.org/wikipedia/commons/6/66/Random_geometric_graph.svg`

**Exercise 7.2.** *In this exercise we will generate a geometric random graph. To do this, we will fix $n = 25$ nodes with locations distributed uniformly and independently over $[0,1)^2$. Set $r = 1/5$ and generate a geometric random graph using these parameters.*