University of Aberdeen
Institute of Mathematics

MX4023 Project Report

2020-2021

Random Graphs

Author : **Marcell Veiner**
Supervisor : **Dr Mark Grant**

Declaration

I hereby declare that this report has been composed by me. I also declare that all sources of information have been specifically acknowledged and all quotations distinguished by quotation marks.

(signed) ....

Abstract

In this Honours Project we offer an introduction to the vast world of random graphs, which combines graph theory and probability theory into one subject. Throughout this report we will explore what we mean by a random graph, and present the definitions of the classical Erdős-Rényi and Gilbert random graphs. We will then spend most of our time deriving properties of Gilbert random graphs, including its degree distribution, connectivity and average path length. We will also investigate some of these properties by measurement. These experiments were written in the programming language Python, using the NetworkX package[1]. Having explored these properties we will be in a position to compare them with those of real world networks. More specifically, we will explore the small-world property of networks, and study an alternative proposal to random graphs, called the Watts-Strogatz model, which more closely resembles networks observed in real life.

[1] Code available on my GitHub: `https://github.com/Vejni/RandomGraphs`

Contents

CHAPTER 1

Introduction

The study of random graphs lies at the intersection of graph theory and probability theory. Although random graphs have been used to study networks in Sociology and Biophysics [12; 10], the groundwork of the field has been laid down by Paul Erdős and Alfréd Rényi in their paper "On random Graphs" published in 1959 [4]. The contribution of Edgar Gilbert is often neglected, who proposed another approach in his paper "Random Graphs" in the same year [6]. In the literature these two processes are often referred to as Erdős-Rényi or Erdős-Rényi-Gilbert model, due to their similarity. We will not adopt this terminology, and will instead address them accordingly.

The study of random graphs deals with asymptotic properties of graphs drawn from a certain probability distribution. For example, it studies how the connectivity of a random graph evolves as the number of edges increases. Similarly to Graph Theory, the appeal of which was to model fundamental aspects of complex phenomena, the theory of Random Graphs has also found its way into other sciences, and has applications from nature to society to the brain [8].

For example, random graphs may be used to describe one of the most complex structures found in nature: the brain. Modelling a network of interconnected neurons as a random graph, different properties of the network may be observed, e.g. brain graphs have been observed to possess a short average path length, but clusters are more likely to appear than in most random graphs [11]. Networks possessing these properties are also called "small-world" networks, by analogy of the small-world phenomenon or more popularly known as "six degrees of separation" [17].

In the following chapters we will introduce the basic models of random graphs, and investigate some of their properties, relevant to the study of (real world) networks. We will then examine an alternate definition proposed in [17] and its applications. We will start, however, with some elementary definitions and results from Probability and Graph theory, but note that some basic knowledge of set theory (sets, unions, functions, etc.), analysis (limits, series, integrals, etc.) and combinatorics is assumed.

CHAPTER 2

The Preliminaries

1. Graph Theory

Definition 2.1. A **graph** is an ordered pair $G = (V,\ E)$, where V is the set of vertices or nodes, and E is the set of edges, consisting of unordered pairs of vertices, satisfying:

$$E \subseteq \{\ \{x, y\}\ \mid\ x,\ y \in V \text{ and } x \neq y\}$$

Note: In particular, this defines an **undirected simple graph**, as no self-loops, multiple and directed edges are allowed. However, as these are the only types of graphs we examine, from now on we will refer to them simply as graphs.

Definition 2.2. Let $G = (V,\ E)$ and $G' = (V',\ E')$ be two graphs. We call G' a **subgraph** of G if $V' \subseteq V$, and $E' \subseteq E$.

Definition 2.3. Let $G = (V,\ E)$ be a a graph and $u, v \in V$. We say that u is **incident to** v, and equivalently v is incident to u if there is an edge $\{u, v\} \in E$. Similarly we can say that u, v are **adjacent**.

Definition 2.4. We will now define how to traverse a graph:

- A **walk** is a finite or infinite sequence of vertices that are all pairwise adjacent, i.e. given a graph $G = (V,\ E)$ a walk is $(w_0, w_1, ...)$, where $w_i \in V$ and $\{w_i, w_{i+1}\} \in E$ for every i.
- A **trail** is a walk in which all edges are distinct.
- A **path** is a trail in which all vertices (and therefore also all edges) are distinct.

Definition 2.5. Let $G = (V,\ E)$ be a a graph and $u, v \in V$. Then u and v are **connected** if there is a path from u to v. Otherwise, they are said to be **disconnected**.

Definition 2.6. A graph $G = (V,\ E)$ is called **connected**, if every pair of vertices in G are connected. A graph that is not connected is also called **disconnected**.

Definition 2.7. Let G be a graph and G' its subgraph. We say that G' is a **connected component** of G, if it is connected and if it is connected to no additional vertices in G.

Definition 2.8. A graph $G = (V,\ E)$ is called **complete**, if for every vertex $u \in V$ and for every vertex $u \neq v \in V$, there is an edge between u and v, that is $\{u, v\} \in E$.

Definition 2.9. Let $G = (V,\ E)$ be a graph. The **degree** of a vertex $v \in V$ is the number of edges that are incident to the vertex denoted $deg(v)$, i.e.:

$$deg(v) = |\ \{\{v, u\} \in E \mid u \in V \setminus \{v\}\}\ |$$

Example 2.10. For example, consider the following graph $G = (V, E)$ (see also Figure 1), where $V = \{A, B, C, D, E, F\}$ and with edges

$$E = \{\{A, B\}, \{A, D\}, \{A, E\}, \{E, D\}, \{E, F\}, \{D, F\}, \{D, C\}, \{C, F\}, \{C, B\}\}$$

.

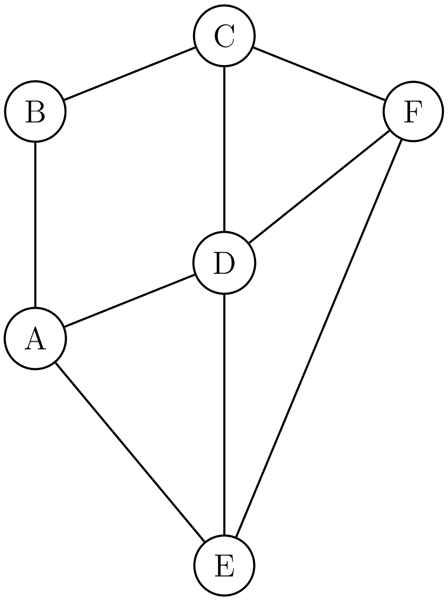

Figure 1. Example Graph

Then a subgraph of G is $G' = (\{A, E, D\}, \{\{A, E\}, \{E, D\}, \{D, A\}\})$. This subgraph is complete, it is not a connected component however, as there are vertices for example B, that it is connected to but are not in the subgraph. The only connected component of G is in fact itself.

The degree of the vertex E is $deg(E) = 3$, as there are 3 vertices adjacent to E. The sequence (E, D, F, D, C) is a walk, but (E, D, F, D, B) is not. (D, F, C, D) is a trail (but not a path), and (D, F, C) is a path (and therefore a trail and a walk).

2. Probability Theory

This section is a distilled version of the lecture notes from MA2010 Probability (2019-2020). In order to keep this part short, I have omitted the proofs of many results that could be found in the lecture notes.

Definition 2.11. A pair (S, P) is called a **discrete probability space**, where S is a countable set called "the sample space", and $P : \mathcal{P}(S) \to \mathbb{R}$ is a function, called the "probability function" satisfying:

- P(S) = 1.
- $0 \leq P(E) \leq 1$ for every **event** $E \subseteq S$.
- If $E_1, E_2, ... \subseteq S$ are pairwise disjoint, then $P(\bigcup\limits_{n=1}^{\infty} E_n) = \sum\limits_{n=1}^{\infty} P(E_n)$.

The following simple results can all be deduced from the three properties of the probability function and subsequently from each other.

Theorem 2.12. *For any probability space (S, P), the following hold:*

- $P(\emptyset) = 0$.
- *Let $E_1, ..., E_n$ be disjoint events, then $P(\bigcup\limits_{i=1}^{n} E_i) = \sum\limits_{i=1}^{n} P(E_i)$.*
- *$P(S \setminus E) = 1 - P(E)$, for any event E.*
- *For events E_1, E_2, $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$.*

Definition 2.13. A probability space (S, P) is called a **uniform probability space** if S is finite and for every $s \in S$ $P(s) = \frac{1}{|S|}$.

It follows directly from the above definitions, in particular the third property of the probability function, that:

Proposition 2.14. *For any event E in a uniform probability space (S, P), $P(E) = \frac{|E|}{|S|}$.*

Theorem 2.15. ***The Inclusion-Exclusion Principle**: Let $A_1, A_2, ..., A_n$ be finite sets, then:*

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{k=1}^{n}(-1)^{k+1} \sum_{i_1<...<i_k} | A_{i_1} \cap ... \cap A_{i_k}| = \sum_{\emptyset \neq I \subseteq \{1,...,n\}} (-1)^{|I|+1} \left|\bigcap_{i \in I} A_i\right|$$

Definition 2.16. Two events E_1 and E_2 are called **independent** if $P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$.

Definition 2.17. Let (S, P) be a probability space, then a **random variable** is a function $X : S \to \mathbb{R}$. For $a \in \mathbb{R}$, we will write $\{X = a\}$ for the event $\{s \in S \mid X(s) = a\}$, and $P(X = a)$ for the probability of such an event. Similarly for $X \leq a$, etc. We will also frequently denote the image of X as $Val(X) = \{X(s)|s \in S\}$.

Definition 2.18. The **expectation** of a random variable X with image $Val(X)$ is:

$$\mathbb{E}(X) = \sum_{x \in Val(X)} x \cdot P(X = x)$$

provided that the series converges absolutely.

Definition 2.19. A random variable X has the **binomial distribution** with parameters n, p, where $0 \leq p \leq 1$, and $1 \leq n$ if $Val(X) = \{0, 1, ..., n\}$ and for all $k \in Val(X)$ it holds that:

$$P(X = k) = \binom{n}{k} p^k q^{n-k}$$

where $q = 1 - p$. Moreover, as the individual events are independent, it can easily be shown that $\mathbb{E}(X) = np$.

Definition 2.20. Let $1 \leq n$, and $x_1, ..., x_n \in \mathbb{R}$, where $x_i \neq x_j$, whenever $i \neq j$. Let X be a random variable with image $Val(X) = \{x_1, ..., x_n\}$, and $P(X = x_i) = \frac{1}{n}$, for all $1 \leq i \leq n$. Then X said to have the **uniform distribution**, and its expectation is:

$$\mathbb{E}(X) = \sum_{k=1}^{n} x_k \cdot P(X = x_k) = \frac{\sum\limits_{k=1}^{n} x_k}{n}$$

Example 2.21. Consider rolling a fair dice once. The possible outcomes make the sample space $S = \{1, 2, 3, 4, 5, 6\}$. Then this is a uniform probability space, as S is finite and each outcome has the same probability. Moreover, consider an event of rolling an even number, then $E = \{2, 4, 6\}$, and $P(E) = \frac{|E|}{|S|} = \frac{3}{6}$. The outcome of the dice may also be considered as a random variable $X = id$, in which case $Val(X) = S$, and $P(X = x) = \frac{1}{6}$, for any $x \in S$. Therefore X has the uniform distribution, and its expectation value is $\mathbb{E}(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$.

Example 2.22. Now consider a biased coin, with probability 0.2 that it comes up heads. Now, let X be the number of heads in 10 tosses. Then X is a random variable with $p = 0.2$ and $n = 10$, and has the binomial distribution. In particular seeing

$k = 3$ heads in 10 tosses has probability $P(X = 3) = \binom{10}{3} \cdot 0.2^3 \cdot 0.8^7 \approx 0.201$ and its expectation can be calculated as $\mathbb{E}(X) = \sum\limits_{k=1}^{10} k \cdot \binom{10}{k} 0.2^k 0.8^{10-k}$.

Definition 2.23. A random variable X has the **Poisson distribution** with parameters $\lambda > 0$, if $Val(X) = \{0, 1, ..., n\}$ and for all $k \in Val(X)$ it holds that:

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Theorem 2.24. *Let X be a random variable, having the binomial distribution with $0 \leq p \ll 1$, $0 \ll n$ and values $Val(X) = \{0, ..., n\}$, Then for small values of $k \in Val(X)$, the probability of the event $X = k$ can be approximated by the Poisson distribution, i.e.:*

$$P(X = k) \approx e^{-\lambda} \frac{\lambda^k}{k!}$$

where $\lambda = np$.

CHAPTER 3

Classical Random Graph Models

1. Uniform Random Graphs

By random graph we should not think of a graph in its own right, but rather a probability space with graphs as its elements. We can describe a random graph by a probability distribution or the random process that generates it. For example, consider all possible (simple undirected) graphs with vertices u, v, w with exactly 2 edges between them, this generates following three graphs visible on Figure 1:

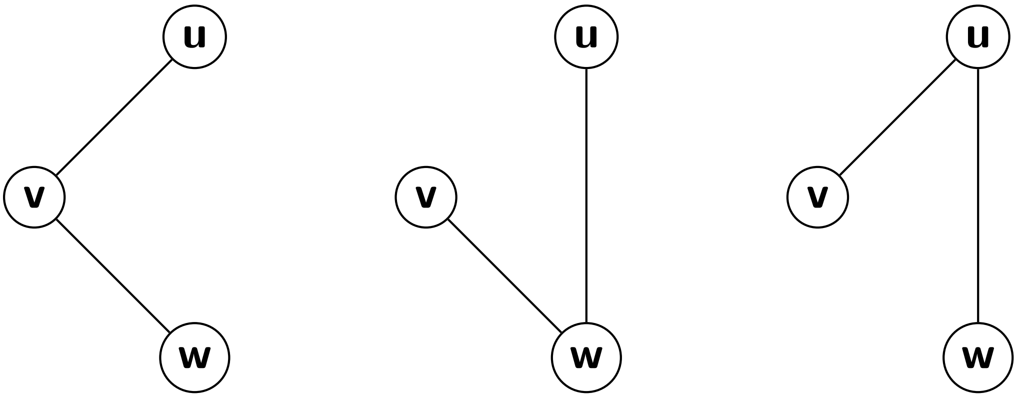

FIGURE 1. Random graph on u, v, w with 2 edges

If we let all three graphs above to be sampled with equal probabilities, i.e. $\frac{1}{3}$ we obtain the uniform random graph with 3 vertices and 2 edges. This process is analogous to the one proposed by Erdős-Rényi, but for arbitrary number of edges and vertices. Let us now define this process rigorously.

DEFINITION 3.1. Take V a set vertices with $|V| = n$, and let $m \in \mathbb{N}$ denote the number of edges. Then a **Erdős-Rényi random graph**, also called **uniform random graph**, $G(n,\ m)$ is the uniform probability space on the set of all graphs with vertices V and m edges.

Let $\mathcal{G}$ be the set of all possible graphs with n vertices and m edges. There are $\binom{n}{2}$ possible edges, of which we can choose m. So the probability that $G(n,\ m)$ assigns $G \in \mathcal{G}$, given that $G = (V,\ E)$ and $|E| = m$ is:

$$P(G(n,\ m) = G) = \binom{\binom{n}{2}}{m}^{-1}$$

Although this model offers an intuitive way of thinking about random graphs, many characteristics of random networks are easier to calculate for graphs obtained from a similar definition, called binomial random graphs.

2. Binomial Random Graphs

In essence, to obtain a binomial or Gilbert random graph we take n vertices and construct a graph by connecting them randomly. Each edge is included in the graph with probability p independent from the other edges. Figure 2 shows 3 binomial graphs obtained with $n = 10$ and $p = 0.3$.

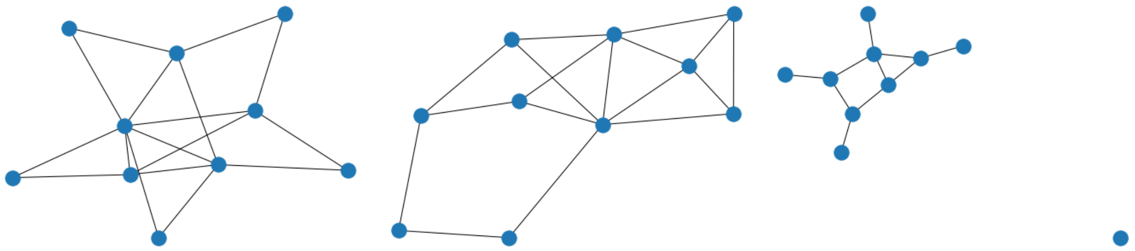

Figure 2. 3 Gilbert random graphs with $n = 10$ and $p = 0.3$

Definition 3.2. Take V a set vertices with $|V| = n$, and let p be the common probability of joining any two of these vertices. Then a **Gilbert random graph**, also called **binomial random graph**, $G(n,\ p)$, is the binomial probability space on the set of all graphs with vertices V.

Let $\mathcal{G}$ be the set of all possible graphs with n vertices and p be the common probability that there is an edge between any two of these vertices. Similarly to the relevant literature, we will write $G(n,\ p) = G$ for the event of sampling G from $\mathcal{G}$. Then there are $\binom{n}{2}$ possible edges, thus $|\mathcal{G}| = 2^{\binom{n}{2}}$ and each $G \in \mathcal{G}$ is sampled by the following probability:

$$P(G(n,\ p) = G) = p^m\ (1-p)^{\binom{n}{2}-m} \tag{3.1}$$

Where m is the number of edges and $0 \leq p \leq 1$. We will also frequently denote $q = 1 - p$. Note that in this model only the number of vertices is fixed in advance, and the number of edges can vary. Furthermore we have a tuning parameter p, which we can make use of in many applications.

Example 3.3. Consider for example the rightmost graph in Figure 1 with vertices u, v, w and edges between v, u and u, w. The probability that the edge $\{v, u\}$ is included in the graph is p, and similarly for $\{u, w\}$. Notice however that the edge $\{v, w\}$ is not included in the graph, which has probability $q = 1-p$. So the probability of this particular graph in the Gilbert model is p^2q. In fact so is the probability of the other 2 graphs, as they have the same number of edges.

Proposition 3.4. *Let X be a random variable denoting the number of edges in a binomial random graph, with n nodes, and parameter p. Then:*

$$P(X = m) = \binom{\binom{n}{2}}{m} p^m \, (1-p)^{\binom{n}{2}-m} \quad \textit{and} \quad \mathbb{E}(X) = p\binom{n}{2}$$

Proof. Let X be the random variable, denoting the number of edges in G. Then there are $\binom{\binom{n}{2}}{m}$ graphs with m edges and n vertices. Obtaining either of these graphs is an independent event and has probability as in Equation 3.1. Thus the probability that $X = m$ is given by $P(X = m) = \binom{\binom{n}{2}}{m} p^m \, (1-p)^{\binom{n}{2}-m}$. Therefore X has the binomial distribution, and following from Definition 2.19 the expected number of edges in a Gilbert random graph is $\mathbb{E}(X) = p\binom{n}{2}$. □

3. Comparing the Uniform and Binomial Model

Lastly let us ask how the two models compare? Note that in the Erdős-Rényi model we fix the number of vertices and the number of edges beforehand, whereas in the Gilbert model before, we only fixed the number of vertices. The following proposition taken from [15] shows how the two models are related.

Proposition 3.5. *Let $n \in \mathbb{N}$. Then the Erdős-Rényi random graph $G(n,\ m)$ is a Gilbert random graph $G(n,\ p)$ given that the number of edges of $G(n,\ p)$ is m.*

Proof. Let G be a graph with m edges. Then by Proposition 3.4:

$$P(G(n,\ p) = G \mid |E(G(n,\ p))| = m) = \frac{P(G(n,\ p) = G)}{P(|E(G(n,\ p))| = m)} =$$

$$\frac{p^m \, (1-p)^{\binom{n}{2}-m}}{\binom{\binom{n}{2}}{m} p^m \, (1-p)^{\binom{n}{2}-m}} = \binom{\binom{n}{2}}{m}^{-1}$$

□

CHAPTER 4

Simple Properties of Binomial Random Graphs

1. Degree Distribution

Proposition 4.1. *The expected degree of an arbitrary vertex in a binomial random graph, with n edges, and parameter p, is $p(n-1)$.*

Proof. Let $G(n,\ p)$ be a Gilbert random graph and v one of its vertices. In order to have $deg(v) = k$ we must have edges between v and k other vertices out of the remaining $n-1$, as loops are excluded. There are $\binom{n-1}{k}$, ways to pick these k vertices, and in each case the probability that v is incident to at least k vertices is given by p^k.

We also require v to not be incident to any of the remaining vertices, as otherwise we would have $deg(v) > k$. The joint probability of these two events is given by $p^k q^{(n-1)-k}$. The final probability is then given by the union of these events. As they are all mutually exclusive, we get the equation:

$$P(deg(v) = k) \ = \ \binom{n-1}{k} p^k \, q^{(n-1)-k} \tag{4.1}$$

In particular $deg(v)$ can also be viewed as a random variable having the binomial distribution. Following again from Definition 2.19 the expected degree of an arbitrary vertex in G with n vertices is $\mathbb{E}(deg(v)) = p(n-1)$, which in the literature is often denoted as $\langle k \rangle$.

□

In summary the number of edges in a random network varies between realizations. Its expected value is determined by n and p. If we increase p a (Gilbert) random network becomes denser: The average number of edges increase linearly from $m = 0$ to $\binom{n}{2}$ and the average degree of a node increases from $deg(v) = 0$ to $deg(v) = n-1$.

Remark. Recall also that by Theorem 2.24 for a binomial random variable X, with $0 \leq p \ll 1$ and $k \ll n$, the Poisson distribution may be used as an approximation.

In particular, evidence shows that many real world networks are sparse [2], implying both $k \ll n$ and $p \ll 1$. Therefore for large sparse networks we may approximate the degree distribution by the Poisson distribution, and the probability of $deg(v) = k$ is given by:

$$P(deg(v) = k) \approx e^{-p(n-1)} \frac{p^k(n-1)^k}{k!} = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!} \tag{4.2}$$

where $\langle k \rangle = p(n-1)$. This form is often preferable as some properties of random graphs have much simpler forms in this limiting case.

2. Hubs

Intuitively, hubs are vertices with a large degree, often having a significantly larger number of links in comparison with other nodes in the network. Hubs often arise in real world networks, but are absent from classical random graphs. A more formal definition has been formalised by Walsh [16] as follows.

DEFINITION 4.2. Let $G = (V, E)$ be a graph. A hub H of G is a set of vertices $(H \subseteq V)$ with the property that for any pair of vertices outside of H, there is a path between them with all intermediate vertices in H.

PROPOSITION 4.3. *Let $G(n, p)$ be a Gilbert random graph with $0 \leq p \ll 1$, $0 < k \ll n$, and v an arbitrary vertex. Then*

$$\lim_{k \to \infty} P(deg(v) = k) = 0$$

PROOF. Using the Poisson approximation and Stirling's approximation for the factorial [21] we obtain:

$$P(deg(v) = k) \approx e^{-\langle k \rangle} \frac{\langle k \rangle^k}{\sqrt{2\pi k}} \cdot \frac{e^k}{k^k} = \frac{e^{-\langle k \rangle}}{\sqrt{2\pi k}} \cdot \left(\frac{\langle k \rangle e}{k} \right)^k$$

Taking the limit as $k \to \infty$ we get that

$$\lim_{k \to \infty} P(deg(v) = k) = \lim_{k \to \infty} \frac{e^{-\langle k \rangle}}{\sqrt{2\pi k}} \cdot \lim_{k \to \infty} \left(\frac{\langle k \rangle e}{k} \right)^k = 0 \cdot \left(\lim_{k \to \infty} \frac{\langle k \rangle e}{k} \right)^k = 0$$

As $\langle k \rangle$ denotes the average degree, and is fixed as k varies, and therefore $\frac{\langle k \rangle e}{k} < 1$. Overall this predicts that in a random graph the probability of observing a hub decreases faster than exponentially.

□

3. Clustering

A more local property of random graphs than hubs are so called, clusters. The idea comes from social networks. For example, consider a group of friends. Then we expect them to all know each other, whereas in a group of acquaintances some introductions are yet to be made. In order to capture this phenomenon first we need the following definitions.

Definition 4.4. Let $G = (V, E)$ be a graph. The **neighbourhood** of a vertex $v \in V$ can then be defined as the vertices adjacent to v, i.e.:

$$N_v = \{u \mid \{v, u\} \in E\}$$

Definition 4.5. Let $G = (V, E)$ a graph. The **local clustering coefficient** for a vertex $v \in V$ is then given by the proportion of links between the vertices within its neighbourhood (which excludes v) divided by the number of links that could possibly exist between them, i.e. let $k = deg(v)$, then:

$$C_v = \frac{|\{\{w, u\} \mid w, u \in N_v \text{ and } \{w, u\} \in E\}|}{\frac{k(k-1)}{2}}$$

Note that the above definitions can only be interpreted whenever $N_v > 1$, therefore we define $C_v = 0$, for v, s.t. N_v.

Definition 4.6. The **average clustering coefficient** of a graph $G = (V, E)$, with $V = \{v_1, ..., v_n\}$, is defined as:

$$\overline{C} = \sum_{i=1}^{n} \frac{C_{v_i}}{n}$$

Example 4.7. Consider a graph

$$G = (\{A, B, C, D, E\}, \{\{A, B\}, \{A, C\}, \{B, C\}, \{C, D\}, \{A, D\}, \{D, E\}\})$$

see also (Figure 1). The Neighbourhood of vertex A is $N_A = \{B, C, D\}$. There are 2 edges $\{B, C\}$ and $\{C, D\}$ between nodes of N_A, while there could be a total of $\frac{3 \cdot 2}{2} = 3$. Thus the local clustering coefficient of node A is $C_A = \frac{2}{3}$.

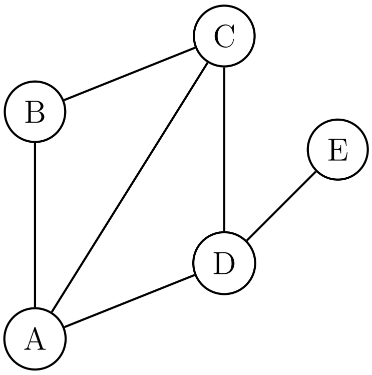

Figure 1. Clustering Coefficient Example

Similarly, $C_B = 1, C_C = \frac{2}{3}, C_D = \frac{1}{3}$ and $C_E = 0$ by definition. Therefore the average clustering coefficient is $C = \left(\frac{2}{3} + \frac{2}{3} + \frac{1}{3} + 1 + 0\right) \cdot \frac{1}{5} = \frac{8}{15}$

Proposition 4.8. *Let $G(n,p)$ be a binomial random graph and v one of its vertices, and let C_v be the random variable denoting the local clustering coefficient of v. Then:*

$$\mathbb{E}(C_v) = p$$

Proof. By Proposition 3.4, we know that the expected number of edges between n vertices is $p\binom{n}{2}$. Therefore among k vertices we expect to have $p\binom{k}{2} = p\frac{k(k-1)}{2}$ edges. Thus:

$$\mathbb{E}(C_v) = \mathbb{E}\left(\frac{|\{\{w,u\} \mid w,u \in N_v \text{ and } \{w,u\} \in E\}|}{\frac{k(k-1)}{2}}\right)$$

$$= \frac{\mathbb{E}(|\{\{w,u\} \mid w,u \in N_v \text{ and } \{w,u\} \in E\}|)}{\frac{k(k-1)}{2}} = \frac{p\frac{k(k-1)}{2}}{\frac{k(k-1)}{2}} = p$$

□

CHAPTER 5

The Connectivity of Random Graphs

We have seen how the number of edges or the degree of vertices are likely to behave in random graphs, but let us now consider a more complicated property. Given a random binomial graph with parameters n, p can we determine the probability that the graph is connected? Or that two specific vertices of the graph u and v are connected? We will denote these probabilities by P_n and R_n respectively, given that G has n vertices. As it turns out that the answer to both of these questions is yes, and they were given by Gilbert in his paper from 1959 [6].

1. Generating Series

Let $n, m \in \mathbb{N}$. Then $C_{n,m}$ will denote the number of connected graphs on n vertices and m edges. Each of these graphs have a probability of $p^m\ q^{\binom{n}{2}-m}$ of being selected. As the selection of any of these is mutually exclusive $P_{n,m} = C_{n,m}\ p^m\ q^{\binom{n}{2}-m}$, where $P_{n,m}$ is the probability that a graph with n vertices and m edges is connected.

To get to P_n, we need to note that graphs having m edges and $m-1$ edges, etc. are mutually exclusive events as well. Thus we can sum over these cases and arrive at:

$$P_n = \sum_{m=n-1}^{\binom{n}{2}} C_{n,m}\ p^m\ q^{\binom{n}{2}-m} \tag{5.1}$$

The range of the summation in equation 5.1 comes from the fact that no graph with n vertices and less than $n-1$ edges can be connected, and that a complete graph of n edges has a total of $\binom{n}{2}$ edges. Having more than this would mean that the corresponding graph is not simple. Using the generating series given in [7], we can write:

$$\sum_{n,m}^{\infty} C_{n,m}\ \frac{x^n y^m}{n!} = ln \left(1 + \sum_{i=1}^{\infty} \frac{x^i (1+y)^{\binom{i}{2}}}{i!} \right) \tag{5.2}$$

Followed by the substitution $y = \frac{p}{q}$, we may derive:

$$\sum_{n,m}^{\infty} C_{n,m}\,\frac{x^n y^m}{n!} = \sum_{n}^{\infty}\sum_{m}^{\infty} C_{n,m}\,\frac{x^n y^m}{n!} =$$

$$\sum_{n}^{\infty}\sum_{m=n-1}^{\binom{n}{2}} C_{n,m}\,\frac{x^n y^m}{n!} = \sum_{n}^{\infty}\sum_{m=n-1}^{\binom{n}{2}} C_{n,m}\,\frac{x^n p^m q^{-m}}{n!} =$$

$$\sum_{n}^{\infty}\left(\sum_{m=n-1}^{\binom{n}{2}} C_{n,m}\, p^m q^{\binom{n}{2}-m}\right)\frac{x^n q^{-\binom{n}{2}}}{n!} = \sum_{n}^{\infty} P_n \frac{x^n q^{-\binom{n}{2}}}{n!}$$

The second equality holds, since all other terms in the summond are zeros. In particular we arrive at the following:

$$\sum_{n}^{\infty} P_n \frac{x^n q^{-\binom{n}{2}}}{n!} = ln\left(1 + \sum_{i=1}^{\infty}\frac{x^i q^{-\binom{i}{2}}}{i!}\right) \tag{5.3}$$

Using the Taylor series expansion of $ln(1+x)$, we get the following:

$$\sum_{n}^{\infty} P_n \frac{x^n q^{-\binom{n}{2}}}{n!} = ln\left(1 + \sum_{i=1}^{\infty}\frac{x^i q^{-\binom{i}{2}}}{i!}\right) = \sum_{j=1}^{\infty}(-1)^{j+1}\,\frac{\left(\sum_{i=1}^{\infty}\frac{x^i q^{-\binom{i}{2}}}{i!}\right)^j}{j} =$$

$$= \sum_{i=1}^{\infty}\frac{x^i q^{-\binom{i}{2}}}{i!} - \frac{\left(\sum_{i=1}^{\infty}\frac{x^i q^{-\binom{i}{2}}}{i!}\right)^2}{2} + \frac{\left(\sum_{i=1}^{\infty}\frac{x^i q^{-\binom{i}{2}}}{i!}\right)^3}{3} - ...$$

$$\Longrightarrow P_1 x + P_2\frac{x^2}{2q} + P_3\frac{x^3}{6q^3} + ... = x + \frac{x^2}{2q} - \frac{x^2}{2} + \frac{x^3}{6q^3} - \frac{x^3}{2q} + \frac{x^3}{3} + ...$$

Equating the coefficients for each x^k we get that:

$$P_1 = 1$$
$$P_2 = 1 - q$$
$$P_3 = 1 - 3q^2 + 2q^3$$
$$...$$

In particular this is a computationally heavy task. As n increases there are an increasing number of coefficients to group in order to express P_n.

2. Recurrence Relation

We can actually do better than that, let us now derive recurrence relations for P_n and R_n.

Theorem 5.1. *Let $G(n,\ p)$ be a Gilbert random graph. Then P_n can be expressed by the following recurrence relation:*

$$1 - P_n = \sum_{m=1}^{n-1} \binom{n-1}{m-1} P_m \, q^{m(n-m)} \tag{5.4}$$

Proof. Let $\mathcal{G}$ be the set of all possible graphs with n vertices. Assume $G(n,\ p)$ assigns $G \in \mathcal{G}$. We would like to know the probability that G is <u>not</u> connected, i.e $\bar{P}_n = 1 - P_n$. These events are of course mutually exclusive.

Let $G_m = (V_m,\ E_m)$ be the connected component such that $v \in V_m$. We can then express the probability that G_m has exactly m vertices, i.e. $|V_m| = m$. There are exactly $\binom{n-1}{m-1}$, ways to pick the remaining $m-1$ vertices. The probability that each one of these graphs is connected is again denoted by P_m.

Since we assumed that G_m is a connected component, we require that none of the vertices in G_m are connected to the remaining $n-m$ vertices, by definition. This happens with probability $q^{m(n-m)}$. Therefore, the probability that G_m has exactly m vertices is given by:

$$Q_m := \binom{n-1}{m-1} P_m \, q^{m(n-m)} \tag{5.5}$$

That means that given a graph G with n vertices, the probability that G has a connected component of size m is as in equation 5.5. Notice that when $m \neq m'$, then Q_m and $Q_{m'}$ are mutually exclusive. In other words, the probability that G is not connected is can be expressed as:

$$1 - P_n = \bar{P}_n = \sum_{m=1}^{n-1} Q_m = \sum_{m=1}^{n-1} \binom{n-1}{m-1} P_m \, q^{m(n-m)}$$

□

We can express R_n, the probability that two specific vertices are connected in the graph that is assigned similarly:

Theorem 5.2. *Let $G(n,\ p)$ be a Gilbert random graph. Then R_n can be expressed by the following recurrence relation:*

$$1 - R_n = \sum_{m=1}^{n-1} \binom{n-2}{m-1} P_m\, q^{m(n-m)} \tag{5.6}$$

Proof. The proof is similar to the one above. Let $\mathcal{G}$ be the set of all possible graphs with n vertices. Assume $G(n,\ p)$ assign $G \in \mathcal{G}$. Let $G = (V,\ E)$, and pick $v, u \in V$. Assume they do not belong to the same connected component. Let $G_m = (V_m,\ E_m)$ be the connected component such that $v \in V_m$, but $u \notin V$.

We can then express the probability that G_m has exactly m vertices, i.e. $|V_m| = m$. There are exactly $\binom{n-2}{m-1}$, ways to pick the remaining $m-1$ vertices, as we cannot pick u. The probability that each one of these graphs is connected is again denoted by P_m.

Since we assumed that G_m is a connected component, we require that none of the vertices in G_m are connected to the remaining $n-m$ vertices (u included), by definition. This happens with probability $q^{m(n-m)}$. Therefore, the probability that G_m has exactly m vertices is given by:

$$Q_m := \binom{n-2}{m-1} P_m\, q^{m(n-m)} \tag{5.7}$$

Using the same reasoning as in the proof of theorem 5.1 the probability that two specific vertices u, v of G are not connected may be expressed as:

$$1 - R_n = \bar{R_n} = \sum_{m=1}^{n-1} Q_m = \sum_{m=1}^{n-1} \binom{n-2}{m-1} P_m\, q^{m(n-m)}$$

□

Using these results Gilbert was able to set up converging bounds for P_n and R_n and showed that$P_n = 1 - nq^{n-1} + O(n^2 q^{\frac{3n}{2}})$ and $R_n = 1 - 2q^{n-1} + O(n^2 q^{\frac{3n}{2}})$. Which for large enough n (depending on q) they can be approximated by:

$$P_n \approx 1 - nq^{n-1} \quad \text{and} \quad R_n \approx 1 - 2 \cdot q^{n-1} \tag{5.8}$$

At a previous point Gilbert calculates several of P_n and R_n, for some small ns and varying q. We may do the same, and obtain the values as in Figures 1, 2, compared with the obtained approximations. It can be observed that the smaller q is, the bigger n we require for the approximations to converge to the observations. As these recurrence relations require heavy computation, only data points for small n have been calculated.

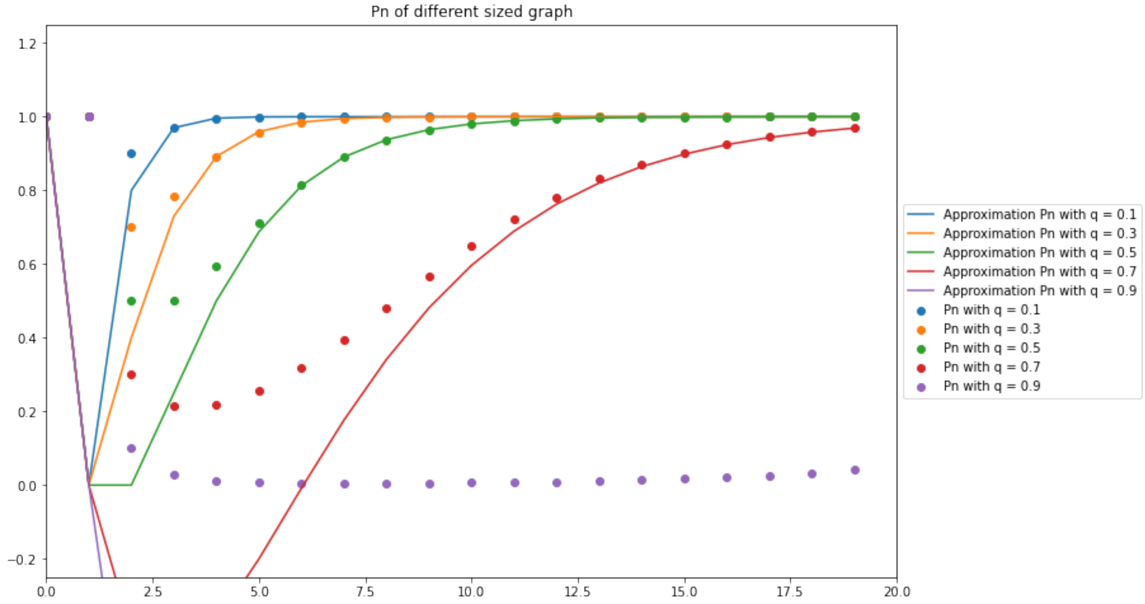

Figure 1. Probability of a Connected Graph

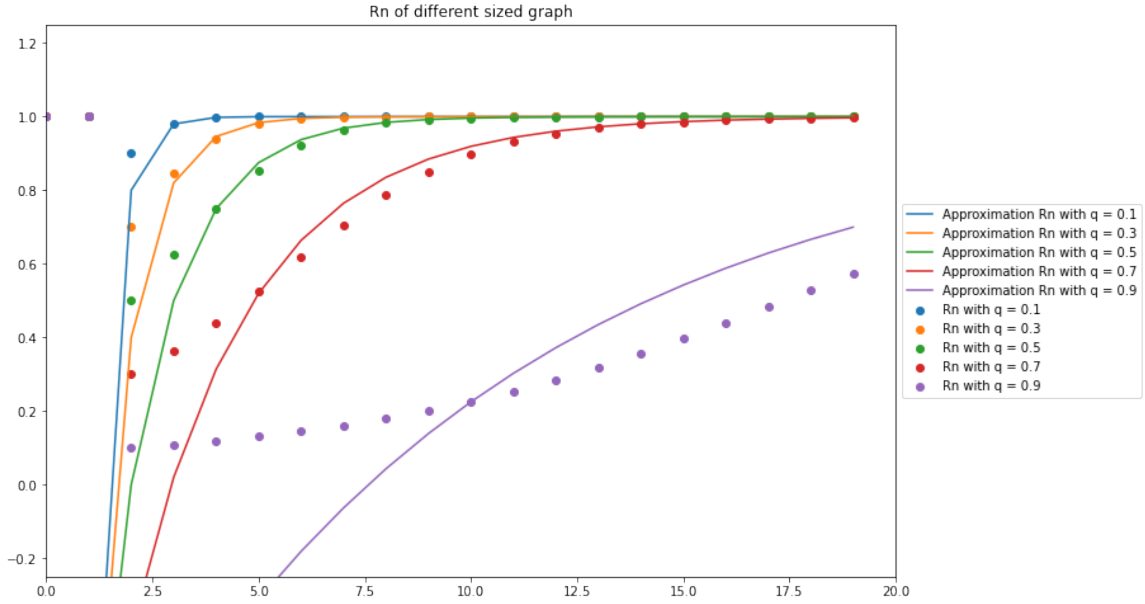

Figure 2. Probability of two Vertices being Connected

Remark. In particular $q \in [0,1]$ and therefore $\lim_{n\to\infty} n \cdot q^{n-1} = 0$, similarly $\lim_{n\to\infty} 2 \cdot q^{n-1} = 0$. Therefore:

$$\lim_{n\to\infty} P_n = \lim_{n\to\infty} R_n = 1 \tag{5.9}$$

Which means that as the graph grows, it is expected to be connected for even low values of p. This is a rather surprising result!

CHAPTER 6

Path Lengths

Having shown that a binomial random graph is almost certain to be connected as n grows large. It makes sense to ask, about the distance between two nodes as $n \to \infty$. We will start with the following definitions:

Definition 6.1. Let $G = (V,\ E)$ be a graph, and $v, u \in V$ two of its vertices. The **distance** between v and u is the length, i.e. the number of vertices on the shortest path. By convention, if there is no path between u and v, the distance, denoted $d(u, v)$ is said to be 0.

Definition 6.2. Let $G = (V, E)$ be a graph with $|V| = n$, then the **average (shortest) path length** is defined :

$$\ell_G = \sum_{i \neq j} \frac{d(v_i, v_j)}{n(n-1)}$$

1. Initial Calculations and Estimates

Let X be the random variable given by the distance in $G(n\ ,p)$ between two of its vertices u and v. Then what is the probability that $P(X = k)$? The answer to this is more complicated than it seems at first. In particular, we compute the boundary cases $k = 1, 2$ and $k = n - 1$.

k = 1. The case $X = 1$ is straightforward. By our definition of the Gilbert model, the edge between any two vertices u and v is included in the graph by probability p. As we do not have other restrictions on the graph this gives us

$$P(X = 1) = p$$

k = 2. This is already more complicated, and we will motivate the calculation by the following picture:

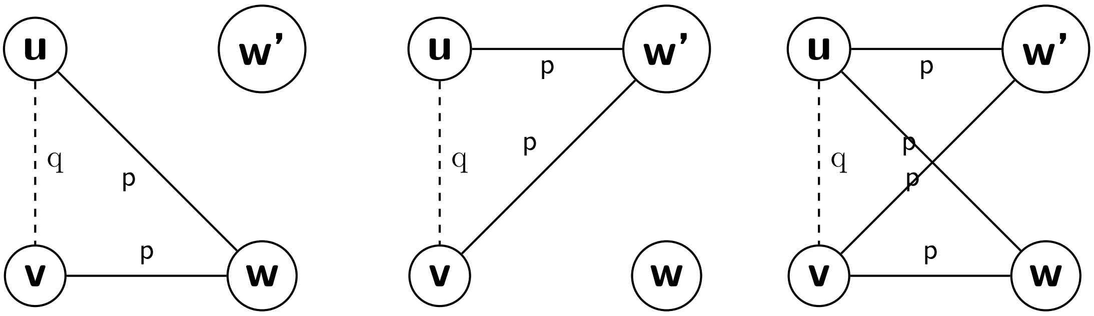

Figure 1. Path of length 2 through a vertex w, w' or both

As we can see for the distance of u and v to be 2, we can a path through any of the remaining $n-2$ vertices, however we require u and v not to be incident. So we can rewrite $P(X=2)$ as follows:

$$P(X=2) = P(X>1 \cap X \leq 2) = P(X>1)P(X\leq 2) = qP(X\leq 2) = q\left(\bigcup_{w\neq v,u} E_w\right)$$

Where E_w is the event such that there is a path of length 2 through vertex w. To get the union, we need to use inclusion-exclusion. There is $\binom{n-2}{i}$ ways to pick i vertices through which there can be a path of length 2. The probability of this event is p^{2i}. So by Theorem 2.15 we get that:

$$q\left(\bigcup_{w\neq v,u} E_w\right) = q\left(\binom{n-2}{1}p^2 - \binom{n-2}{2}p^4 + ... + (-1)^{n-1}\binom{n-2}{n-2}p^{2(n-2)}\right) =$$

$$q\sum_{i=1}^{n-2}(-1)^{i+1}\binom{n-2}{i}p^{2i}$$

Which by the binomial theorem gives:

$$P(X=2) = q(1-(1-p^2)^{n-2}) \tag{6.1}$$

k = n - 1. Let us first present this case rather than when $k=n-2$. We will see that $k=n-2$ follows from $k=n-1$. In this case, the graph is a path graph with vertices u and v as its endpoints. We can reorder the vertices in the path and require the path to be in the graph, but all other edges that would create shortcuts not part of the graph. This yields the probability:

$$P(X = n - 1) = (n - 2)!p^{n-1}q^{\binom{n}{2}-(n-1)} \tag{6.2}$$

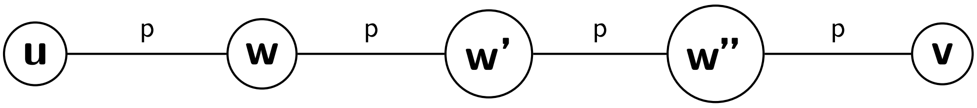

Figure 2. Path Graph of Length $n - 1$

1.1. The Diameter of a Random Graph. To get an estimate for the average path length ℓ, or APL, recall that by Proposition 4.1 the average degree of a vertex is $p(n-1) = \langle k \rangle$. This average degree is often held constant as n grows large, which is used to get some control over p. In particular we will define $p(n)$ as:

$$p(n) = \frac{\langle k \rangle}{n} \tag{6.3}$$

for fixed $\langle k \rangle \in \mathbb{N}_{>0}$. We follow with a definition and an estimate.

Definition 6.3. The **diameter** of a graph $G = (V,\ E)$ is the maximal distance between any pair of its nodes:

$$diam(G) = max(\{d(u,\ v) \ \mid\ u, v \in V\})$$

The diameter of random graphs has been studied by many authors [3] and the general conclusion is that the diameter d of a Gilbert random graph is concentrated around:

$$d = \frac{ln(n)}{ln(\langle k \rangle)} = \frac{ln(n)}{ln(p(n-1))}$$

We can use this as a rough estimate for the APL as well, as one expects that it scales with the number of vertices in the same way as the diameter:

$$\ell \sim d = \frac{ln(n)}{ln(\langle k \rangle)} \tag{6.4}$$

2. Analytic Solution

We will now turn our attention to an analytic solution for the average path length in Gilbert random graphs found by Fronczak, Fronczak and Holyst [5] in the limiting case $n \to \infty$. Fronczak et al. let the probability p vary between any pair of edges according to hidden variables. Fortunately in the Gilbert model we do not allow this probability to vary depending on the vertices it connects, so the computation simplifies greatly. We will however make the assumption that p is a function of the number of vertices n.

LEMMA 6.4. *If $A_1, A_2, ..., A_n$ are mutually independent events and their probabilities fulfill $\forall i P(A_i) \leq \varepsilon$ then:*

$$P(\bigcup_{i=1}^{n} A_i) = 1 - exp\left(-\sum_{i=1}^{n} P(A_i)\right) - Q$$

where $0 \leq Q \leq \sum_{j=0}^{n+1} (n\varepsilon)^j / j! - (1+\varepsilon)^n$.

PROOF. Using the Inclusion-exclusion principle 2.15, we may write $P(\bigcup\limits_{i=1}^{n} A_i) = \sum\limits_{j=1}^{n} (-1)^{j+1} S(j)$, with

$$S(j) = \sum_{1 \leq i_1 < i_2 < ... < i_j \leq n}^{n} P(A_{i_1}) P(A_{i_2}) ... \, P(A_{i_j}) \tag{6.5}$$

where $S(j), for\ j = 1, ..., n$, can be thought of as terms from the multinomial expansion of $(P(A_1) + P(A_2) + ... + P(A_n))^j = (\sum\limits_{i=1}^{n} P(A_i))^j$, which is given by:

$$\left(\sum_{i=1}^{n} P(A_i)\right)^j = \sum_{k_1+k_2+...+k_n=j} \frac{j!}{k_1!k_2!...k_n!} P(A_1)^{k_1} ... \, P(A_n)^{k_n} = j!(S_j + Q_j) \tag{6.6}$$

Where Q_j denotes the sum of all the terms with $P(A_i)^k$, where $k > 1$ for some i, divided by $k_1!k_2!...k_n!$. Remember that in Equation (6.5) the index j runs from 1 to n, so in Equation (6.6) we will always have more k_i than j. So S_j contains the terms where $k_{i_1} = k_{i_2} = ... = k_{i_j} = 1$ and and the rest are 0. Therefore:

$$S(j) = \frac{1}{j!} \left(\sum_{i=1}^{n} P(A_i)\right)^j - Q_j \tag{6.7}$$

Now what is Q_j? We claim that since $\forall i P(A_i) \leq \varepsilon$ we can bound it by some quantity. Disregarding the coefficients in the multinomial expansion, there are a total of $\frac{n^j}{j!}$ summands, which all consists of the product of j number of $P(A_i)$, where $i = 1, ..., n$.

As the order does not matter, we divide by $j!$ to get the total number of terms. But we have $\binom{n}{j}$ of these already in S_j, so there are $\frac{n^j}{j!} - \binom{n}{j}$ terms in the sum Q_j. As $\forall i P(A_i) \leq \varepsilon$, we know that $0 \leq Q_j \leq (\frac{n^j}{j!} - \binom{n}{j})\varepsilon^j$.

Then it follows from equation 6.7 that:

$$1 - P(\bigcup_{i=1}^{n} A_i) = 1 - \sum_{j=1}^{n}(-1)^{j+1}S(j) = 1 - \sum_{j=1}^{n}(-1)^{j+1}\left(\frac{1}{j!}\left(\sum_{i=1}^{n} P(A_i)\right)^j - Q_j\right)$$

$$= \sum_{j=0}^{n} \frac{(-1)^j}{j!}\left(\sum_{i=1}^{n} P(A_i)\right)^j - \sum_{j=1}^{n}(-1)^{j+1}Q_j \tag{6.8}$$

The first sum in Equation (6.8) is the first $(n+1)$ terms (from the 0th to the nth)in the MacLaurin expansion of e^{-x}, where $x = \sum_{i=1}^{n} P(A_i)$. Therefore we may write:

$$\sum_{j=0}^{n} \frac{(-1)^j}{j!}\left(\sum_{i=1}^{n} P(A_i)\right)^j = exp\left(-\sum_{i=1}^{n} P(A_i)\right) - R_n \tag{6.9}$$

Where R_n in Equation (6.9) is the remainder, and can be written in Lagrange's form by Theorem 7.43 in [14], so that $R_n = \frac{f^{(n+1)}(c)}{(n+1)!} \cdot x^{n+1}$ with $c \in (0, x)$. We first note that e^{-x} is a monotone decreasing function, and so restricting it over the non-negative numbers it attains its maximum at $x = 0$. Then we obtain the following bounds on the remainder:

$$R_n \leq |R_n| = \left|\frac{f^{(n+1)}(c)}{(n+1)!}x^{n+1}\right| = \left|\frac{exp(c)}{(n+1)!}x^{n+1}\right| \leq \left|\frac{exp(0)}{(n+1)!}x^{n+1}\right| \leq \left|\frac{(n\varepsilon)^{n+1}}{(n+1)!}\right| \tag{6.10}$$

The last inequality in 6.10 holds, since $\forall i P(A_i) \leq \varepsilon$ we have that $\sum_{i=1}^{n} P(A_i) \leq n\varepsilon$. Using the MacLaurin expansion in Equation (6.8) and rearranging we get that:

$$P(\bigcup_{i=1}^{n} A_i) = 1 - exp\left(-\sum_{i=1}^{n} P(A_i)\right) - Q \tag{6.11}$$

Where Q is the total error. Recall that we have shown that $0 \leq Q_j \leq (\frac{n^j}{j!} - \binom{n}{j})\varepsilon^j$, therefore Q can be bounded as in Equation (6.12):

$$Q \leq \sum_{j=1}^{n}(-1)^{j+1}Q_j + \frac{(n\varepsilon)^{n+1}}{(n+1)!} < \sum_{j=1}^{n} Q_j + \frac{(n\varepsilon)^{n+1}}{(n+1)!}$$

$$\leq \sum_{j=0}^{n+1}\left(\frac{n^j\varepsilon^j}{j!} - \binom{n}{j}\varepsilon^j\right) + \frac{(n\varepsilon)^{n+1}}{(n+1)!} = \sum_{j=0}^{n+1}\left(\frac{(n\varepsilon)^j}{j!}\right) - \sum_{j=0}^{n+1}\binom{n}{j}\varepsilon^j$$

$$= \sum_{j=0}^{n+1}\frac{(n\varepsilon)^j}{j!} - (1+\varepsilon)^n \tag{6.12}$$

where the second inequality is strict, because $0 \leq Q_j$ for all j, assuming that at least one $Q_j \neq 0$, for some j. This completes the proof. □

Let us now turn our attention back to path lengths, more specifically to the proof for the analytic formula given in [5]. We wish to show that for a Gilbert random graph the APL is given by:

$$\ell = \frac{ln(n) + \gamma}{ln\langle k\rangle} + \frac{1}{2} \tag{6.13}$$

Where $\gamma \approx 0.5772$ is Euler's constant. Let $(v, w_1, w_2, ..., w_{k-1}, u)$ be a walk of length k. In order to follow the arguments presented in Fronczak et al. let us assume that the probability for the existence of such a walk in a Gilbert random graph to be $p_{vw_1}p_{w_1w_2}\cdots p_{w_{k-1}u} = p^k$.

As we are considering only a walk for now, note that $(v, w_1, w_2, ..., w_{k-1}, u)$ do not have to be distinct vertices, we only require subsequent vertices in the walk to be adjacent. Note that our definition of a walk does not allow for $w_i = w_{i+1}$, i.e. for staying in one place. Then some edges may appear in the walk more than once, in this case the probability of the inclusion of this edge is given by $p_{w_iw_{i+1}}$ and not $p^2_{w_iw_{i+1}}$, as the proof assumes. Therefore we are only able to bound the probability of the existence of a k-walk by p^k from below, and by p from above, as $p \in [0, 1]$.

Nevertheless, by accepting this assumptions we may notice the following. First let E_k denote the event that there is at least one walk of length k between u, v. If this is in fact a shortest path then $d(u, v) = k$. Note however that the existence of a walk of length k between u, v implies that $d(u,v) \leq k$. In particular, this holds for arbitrary k, and so for $k-1$. Letting $\dot{E}_k$ denote the event that $d(u,v) = k$, it follows that $E_k = \dot{E}_k \cup E_{k-1}$, and we obtain the following relation between probabilities:

$$P(\dot{E}_k) = P(E_k) - P(E_{k-1}) \tag{6.14}$$

Let $A_{w_1,w_2,\ldots,w_{k-1}}$ denote the event that there is a walk $(v, w_1, w_2, \ldots, w_{k-1}, u)$. Since we allow vertices to be revisited, and the endpoints are fixed, there are $(n-1)^{k-1}$ such events (self loops are excluded). Furthermore if V is the set of vertices, we may write:

$$E_k = \bigcup_{w_1,w_2,\ldots,w_{k-1} \in V} A_{w_1,w_2,\ldots,w_{k-1}}$$

The question is whether the events $A_{w_1,w_2,\ldots,w_{k-1}}$ are mutually independent or not. We may notice that certain edges may appear in more than one walks, and in this case these events are dependent. Fronczak et al. dismisses this by stating that the fraction of such walks is negligible for $n \gg k$. Although this indeed seems to be the case, it relies on posterior knowledge. We will see that most random graphs, especially ones possessing small-world property, do in fact tend to have a small APL, from which we may deduce that most vertices are of short distance apart, as $n \to \infty$ the probability of having correlated walks can also increase.

In order to proceed with the proof however, we will accept this assumption as well. In particular claiming mutual independence for the events $A_{w_1,w_2,\ldots,w_{k-1}}$ lets us use Lemma 6.4.

$$P(E_k) = P\left(\bigcup_{w_1,\ldots,w_{k-1} \in V} A_{w_1,\ldots,w_{k-1}}\right) = 1 - exp\left(-\sum_{w_1,\ldots,w_{k-1} \in V} P\left(A_{w_1,\ldots,w_{k-1}}\right)\right) - Q$$

$$= 1 - exp\left(-\sum_{w_1,,\ldots,w_{k-1} \in V} p^k\right) - Q = 1 - e^{(-(n-1)^{k-1}p^k)} - Q$$

As to whether the term Q may or may not be ignored, ultimately is tied to the assumption that $P\left(A_{w_1,\ldots,w_{k-1}}\right) = p^k$. Fronczak et al. claims that since the number of vertices with large degree, can be bounded, we may disregard Q. Let us follow this assumption in order to simplify our calculations.

$$P(E_k) = 1 - e^{(-(n-1)^{k-1}p^k)} = 1 - e^{(-((n-1)p)^{k-1}p)} = 1 - e^{-p\langle k\rangle^{k-1}}$$

$$= 1 - e^{-\frac{\langle k\rangle^k}{n}} := 1 - F(k) \tag{6.15}$$

The term $\langle k\rangle$ in Equation (6.15) is the average vertex degree, and is different to the path length we denote by k. Taking advantage of Equation (6.14) and (6.15), we can calculate the expectation value for the average path length between vertices u and v as $n \to \infty$:

$$\ell_{u,v} = \sum_{k=1}^{\infty} kP(\dot{E}_k) = \sum_{k=1}^{\infty} k(P(E_k) - P(E_{k-1})) = \sum_{k=1}^{\infty} k(1 - F(k) - 1 + F(k-1))$$

$$= \sum_{k=1}^{\infty} k(F(k-1) - F(k)) = F(0) - F(1) + 2F(1) - 2F(2) + 3F(2) - 3F(3) + ...$$

$$= F(0) + F(1) + F(2) + ... = \sum_{k=0}^{\infty} F(k) = \sum_{k=0}^{\infty} e^{-\frac{\langle k\rangle^k}{n}} \tag{6.16}$$

This sum, however, is harder to calculate than it seems. In order to obtain the actual limit, let us make use of the Poisson Summation Formula [20] following the proof of Fronczak et al. The formula states that:

$$\sum_{k=0}^{\infty} F(k) = \frac{1}{2}F(0) + \int_0^{\infty} F(k)\ dk + 2\sum_{m=1}^{\infty}\left(\int_0^{\infty} F(k)\ cos(2m\pi k)\ dk\right) \tag{6.17}$$

It is straight forward to see that $F(0) \approx 1$ for large enough n (for example, at $n = 100$ $F(0) = 0.99$). Furthermore Fronczak et al. claims that every integral in the last term of the summation formula in Equation (6.17) is equal to zero, owing to the Generalized Mean Value Theorem, which is a term used for a whole family of results in real analysis. They are skipping some essential details here, which makes it difficult to retrace their steps in this complicated proof. There is a possibility that they refer to the (Second) Mean Value Theorem for Integrals, i.e. 8.3.5. in [14], as this looks promising for what we would like to accomplish. However, using this theorem we are only able to obtain diverging results:

$$\int_0^\infty F(k)\cdot cos(2m\pi k)\; dk = \lim_{a\to\infty} F(c)\cdot \int_0^a cos(2m\pi k)\; dk$$

$$= F(c)\lim_{a\to\infty}\int_0^a cos(2m\pi k)\; dk \approx \lim_{a\to\infty}\frac{1}{2\pi m}\Big[sin(0) - sin(2\pi am)\Big] = \lim_{a\to\infty}\frac{-sin(2\pi am)}{2\pi m}$$

for some $c \in (0, a)$. This limit, however, diverges. In order to finish with this proof, and since we have made so many assumptions already, let us suppose convergence as Fronczak et al. claims, and see what we are able to obtain. Then what is left from Equation 6.17 is only:

$$\int_0^\infty F(k)\; dk = \int_0^\infty e^{\frac{-\langle k\rangle^k}{n}}\; dk = \int_{\frac{-1}{n}}^{-\infty}\frac{e^u}{u\; ln\langle k\rangle}\; du = \frac{1}{ln\langle k\rangle}\left(-\int_{-\infty}^{\frac{-1}{n}}\frac{e^u}{u}\; du\right)$$

$$= \frac{1}{ln\langle k\rangle}\left(\int_{\frac{1}{n}}^{\infty}\frac{e^{-u}}{u}\; du\right) = \frac{Ei(\frac{1}{n})}{ln\langle k\rangle} \tag{6.18}$$

using u-substitution with $u = \frac{-\langle k\rangle^k}{n}$ in the integral of Equation 6.18. Furthermore $Ei(x)$ is the exponential integral [19], and for positive arguments we have that $Ei(y) = -\gamma - ln(y)$. Therefore:

$$\ell = \frac{1}{2} + \frac{Ei(\frac{1}{n})}{ln\langle k\rangle} = \frac{1}{2} + \frac{-\gamma - \ln\left(\frac{1}{n}\right)}{ln\langle k\rangle} = \frac{1}{2} + \frac{ln(n) - \gamma}{ln\langle k\rangle} \tag{6.19}$$

Let us summarise. Under the many assumptions we have made we were able to show the expected APL of a binomial random graph. How does this approximation hold up? We have already seen in Equation 6.4 that $\ell \sim d$, where d denoted the diameter of the graph. Since we hold $\langle k\rangle$ constant as n grows, we see that:

$$\mathcal{O}(\ell) = \mathcal{O}\left(\frac{1}{2} + \frac{ln(n)-\gamma}{ln\langle k\rangle}\right) = \mathcal{O}\left(\frac{ln(n)}{ln\langle k\rangle} - \frac{\gamma}{ln(\langle k\rangle)}\right) = \mathcal{O}\left(\frac{ln(n)}{ln\langle k\rangle}\right) = \mathcal{O}(d)$$

3. Simulation

Having made so many assumptions to arrive at an approximation formula for the APL, it may be beneficial to seek experimental confirmation for our calculations. Inspired by Fronczak et al. we set $\langle k \rangle$ to a fixed integer, which we keep constant over the experiment. This is used to determine $p(n) = \frac{\langle k \rangle}{n-1}$ as n grows. In particular we obtain the results as in Figure 3. The x-axis of the plot is logarithmic.

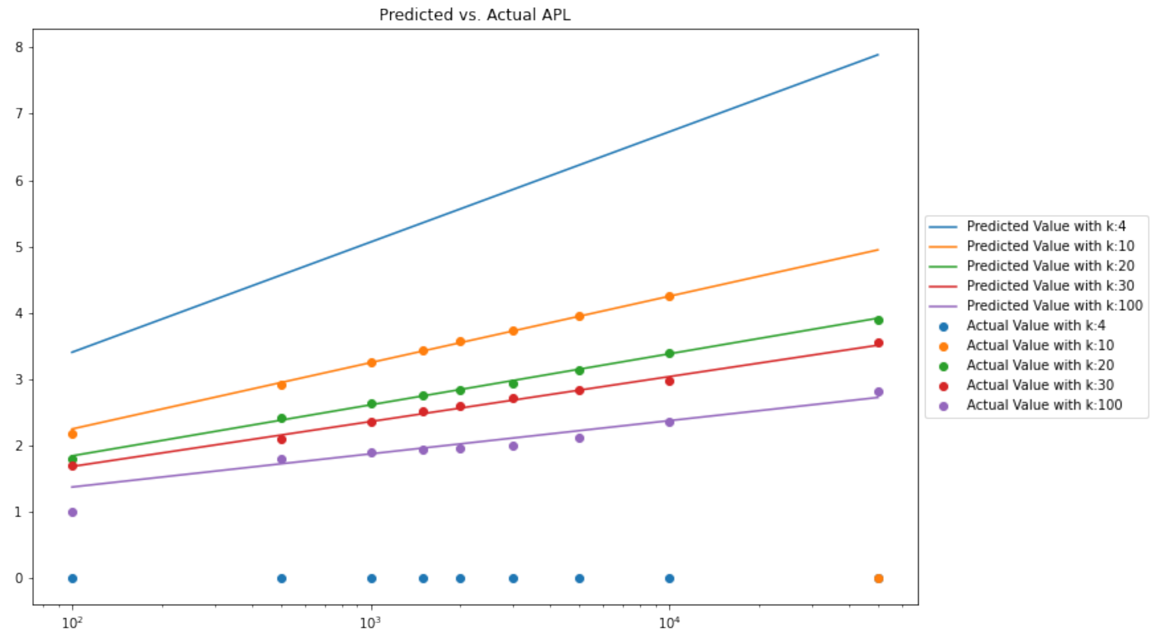

Figure 3. Average Path Length: Observed vs. Predicted

The predictions were made using Equation 6.19. As we see the predicted values were almost identical to the observations in most cases. The datapoints on the horizontal axis, correspond to disconnected graphs, (corresponding to the lowest values of p). This was an assumption of the APL formula used in the simulation, which our definition has sidestepped by letting the distance of disconnected vertices be 0 instead of ∞.

Based on this empirical evidence, one might find it easier to accept the assumptions made by Fronczak et al. In the remaining sections, we will talk about real world networks, and the small-world property.

CHAPTER 7

The Small-World Property

The small-world phenomenon, also known as six degrees of separation, has long fascinated the general public for its surprising statement that if you choose two individuals anywhere on Earth, you will find a path of at most six acquaintances between them [9]. But how does this translate to the study of random graphs?

Although small-world networks do not have a word-by-word definition, more specifically, a small-world network is a graph, whose average path length grows proportional to the logarithm of the number of nodes, while the average clustering coefficient is not low.

As asymptotically $ln(n) \ll n$, this means that in a small-world networks the distances in a random network are orders of magnitude smaller than the size of the network. They also tend to contain highly connected sub-networks, also called cliques, which have connections between almost any two nodes within them. A high number of hubs is also typical, i.e. nodes of higher degrees, which play an important role in creating short path lengths between nodes.

While discovered in the context of social sciences, the small world property applies beyond social networks. In particular many real world networks have small-world properties. Table 1 is adapted from [13], showing key properties of real world networks.

Network	n	$\langle k \rangle$	C	ℓ	C_{rand}	ℓ_{rand}	$0.5 + \frac{ln(n)-\gamma}{ln\langle k \rangle}$
C. elegans	453	8.94	0.65	2.66	0.28	2.50	3.02
Protein interactions	1539	2.67	0.07	6.81	0.04	5.69	7.38
E-mail	1133	9.62	0.22	3.60	0.09	3.27	3.35

Table 1. Real world networks

Where $n, \langle k \rangle, C, \ell$ are properties of the network (number of nodes, avg. degree, avg. clustering and apl respectively), and C_{rand}, ℓ_{rand} are the properties of random networks with the same number of nodes and average degree. The last column shows the approximated APL in such a random network.

Firstly, we may observe that $\ell \sim \ell_{rand}$. This is in fact expected, as we have spent a long time showing that for a binomial random graph we expect the APL to be around $0.5 + \frac{ln(n)-\gamma}{ln\langle k\rangle} \sim ln(n)$. Secondly, there are differences between C and C_{rand}. This is backed up by the fact that in Chapter 2 Section 3 we have shown that the clustering coefficient of an arbitrary vertex of a binomial random graph is p.

That is binomial random graphs do not generate a networks with large average clustering coefficients, and they do not produce a large number of hubs, as a result of the binomial or Poisson degree distribution. In particular the degree distribution of random networks fails to represent that of real world networks. Figure 1 taken from [2] shows the degree distributions of two networks compared to the predicted degree distribution of a random graph of the same size.

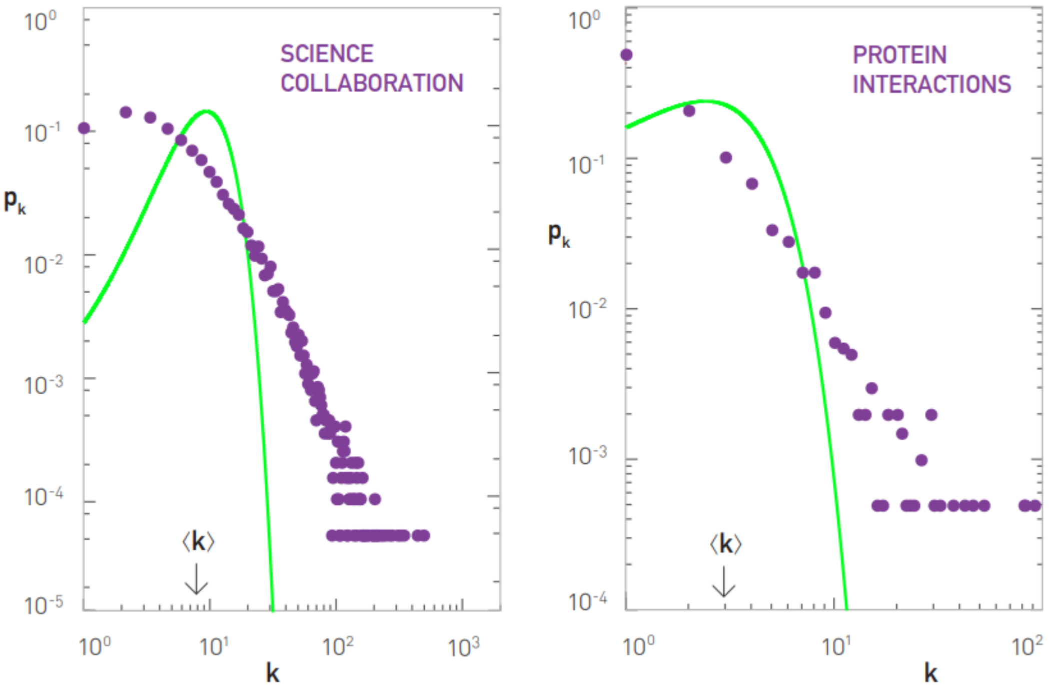

Figure 1. Degree Distribution of Real World Networks [2]

In particular the Poisson distribution for the degree underestimates the size and the frequency of nodes at the two sides of the spectrum. These discrepancies between random and real world networks have motivated the search for other models possessing these desired properties. One such network, designed to address the differences in how clusters are formed, is what we will discuss next.

1. The Watts-Strogatz Model

Although the simplicity of ER and Gilbert random graphs has made them an attractive choice for many applications, their degree distributions and clustering coefficients do not resemble networks observed in the real world. The Watts-Strogatz model was designed to mitigate the differences in the second aspect [17]. It is a generative model, that is it does not define a probability space, but an algorithm for obtaining a random graph.

Definition 7.1. **Watts-Strogatz Random Graph**: Let n denote the number of vertices, and k be an even integer, denoting the average degree of a vertex. Assume that $n \gg k \gg ln(n) \gg 1$, and let $0 \leq p \leq 1$. The algorithm consists of two steps:

(1) Arrange the n vertices in a ring lattice, labeled $v_0, v_1, ...v_{n-1}$, then connect every vertex v_i with $k/2$ of its neighbours on each side, i.e. $\{v_i, v_j\} \in E$ whenever
$$0 < |i-j| \bmod \left(n-1-\frac{k}{2}\right) \leq \frac{k}{2}$$
.

(2) For each vertex v_i, $i = 0, ..., n-1$, consider its closest neighbour, i.e. v_j where $|i-j| \bmod \left(n-1-\frac{k}{2}\right) = 1$. Then replace the edge $\{v_i, v_j\} \in E$, with $\{v_i, v_k\}$, where k is chosen uniformly from $\{0, 1, ..., i-1, i+1, ..., n-1\}$, provided that $\{v_i, v_k\} \notin E$. We then repeat this process with the second closest neighbours of v_i, and so on, until all of the original edges are exhausted.

The conditions $n \gg k \gg ln(n) \gg 1$, is enforced to obtain graphs with a large number of nodes, but sparse connections. However, $k \gg ln(n)$ guarantees that the graph does not become disconnected. Figure 2 visualises the procedure.

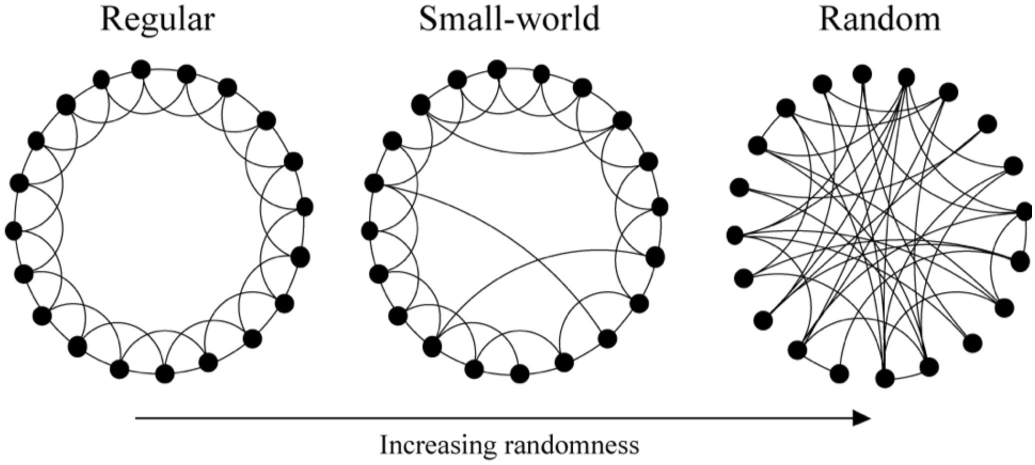

Figure 2. The Watts-Strogatz Model Visualised [17]

Proposition 7.2. *Let $G = (V, E)$ be a Watts-Strogatz "random graph" with n nodes, k neighbours to each vertex, and $p = 0$. In this case, G in fact is deterministic, as no rewiring occurred, and as such it resembles a regular ring lattice (see left side of Figure 2). The average clustering coefficient of G is*

$$C = \frac{3(k-2)}{4(k-1)} \tag{7.1}$$

Proof. As the graph is a regular ring lattice, the local clustering coefficient is the same for each vertex $c \in V$. Therefore $C_v = C$, the average clustering coefficient.

Let $deg(v) = k$, i.e. v has k neighbours. Then the complete graph between these k neighbours has $\binom{k}{2} = \frac{k(k-1)}{2}$ edges. To count the number of edges in the neighbourhood of v, N_v it is easier to count the edges absent between the nodes of N_v and subtract it from the number of possible edges $\frac{k(k-1)}{2}$.

The vertices one distance away from v on the ring lattice, i.e. the vertices drawn right next to v on the lattice. As edges are undirected, it is enough to consider one of these vertices, denoted u. Then u has only 1 edge that is absent, the edge to the node on the other side of v, far at the end. The rightmost neighbour of u has 2 such absent edges, the next 3, etc. The last vertex will have $\frac{k}{2}$ missing edges. Therefore the number of edges present between the nodes of N_v is:

$$\frac{k(k-1)}{2} - \sum_{i=1}^{k/2} i = \frac{k(k-1)}{2} - \frac{\frac{k}{2}\left(\frac{k}{2}-2\right)}{2} = \frac{4k(k-1)}{8} - \frac{k(k-2)}{8} = \frac{3k(k-2)}{8}$$

Therefore the clustering coefficieint is:

$$C = \frac{\frac{3k(k-2)}{8}}{\frac{k(k-1)}{2}} = \frac{3(k-2)}{4(k-1)}$$

□

On the other hand as $p \to 1$, the graph becomes more random. In fact it starts to resemble a binomial random graph with $p = \frac{k}{n-1}$, while not converging to it, because the minimum number of edges in a Watts-Strogatz graph is $\frac{nk}{2}$. Similarly, the average path length interpolates between $\ell = \frac{n}{2k}$ at $p = 0$ and $\frac{ln(n)}{ln(k)}$ at $p = 1$ [17]. The interpolation of these two quantities between the extremes leaves a window of opportunity for small-world networks. Figure 3 shows this phenomenon, as presented by Watts and Strogatz (WS) [17].

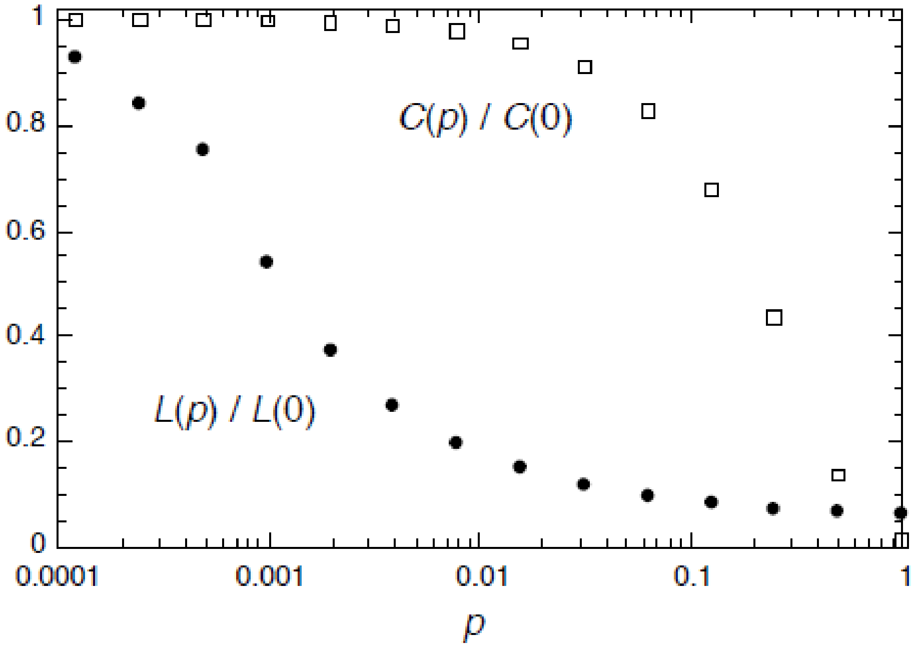

Figure 3. Small-World Networks Obtained by WS [17]

CHAPTER 8

Afterword

We will close by as king the same question as in the closing pages of [2], namely: *Are real world networks truly random?* The asnwer, as Barabási put it, "is clearly no". In reality, it is more likely that complex systems appear as a result of some deep underlying order behind all the chaos. However, by studying random graphs, we were able to quantify some of these differences, and even explore ideas that tried to patch the discrepancies.

In fact the theory of random graphs serves more as a reference when studying real world networks. Upon discovering a new property of a network, a question random graphs can answer is *if it could have emerged by chance?* If the property cannot be found in random graphs, it may represent a signature of order, "requiring a deeper explanation" [2]. This is the relevance of random graphs to the study of networks.

This project is only a tiny glimpse into the study of random networks, and naturally there is a lot more to explore. As a reference point, we have not explored the end of the analysis by Gilbert in [6], deriving the converging bound on P_n and R_n. We have not talked about the emergence of a giant component [2], and there is a whole new world waiting to be explored in [1], including new random graph models, percolation theory and the evolution of random networks.

Bibliography

[1] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[2] Albert-László Barabási. *Network Science*. Chapter 3: Random Networks.

[3] Fan Chung and Linyuan Lu. The diameter of random sparse graphs. *Advances in Applied Math*, 26(4):257–279, 2001.

[4] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[5] Agata Fronczak, Piotr Fronczak, and Janusz A Hołyst. Average path length in random networks. *Physical Review E*, 70(5):056110, 2004.

[6] Edgar N Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[7] Eo N Gilbert. Enumeration of labelled graphs. *Canadian Journal of Mathematics*, 8:405–411, 1956.

[8] Mihyun Kang. Random graphs: from nature to society to the brain. *Número especial del Math. Intelligencer dedicat al SEOUL ICM*, pages 42–44, 2014.

[9] Stanley Milgram. Six degrees of separation. *Psychology Today*, 2:60–64, 1967.

[10] Jacob L Moreno and Helen H Jennings. Statistics of social configurations. *Sociometry*, pages 342–374, 1938.

[11] David Papo, Javier M Buldú, Stefano Boccaletti, and Edward T Bullmore. Complex network theory and the brain, 2014.

[12] Ray Solomonoff and Anatol Rapoport. Connectivity of random nets. *The bulletin of mathematical biophysics*, 13(2):107–117, 1951.

[13] Qawi K Telesford, Karen E Joyce, Satoru Hayasaka, Jonathan H Burdette, and Paul J Laurienti. The ubiquity of small-world networks. *Brain connectivity*, 1(5):367–375, 2011.

[14] Brian S Thomson, Judith B Bruckner, and Andrew M Bruckner. *Elementary real analysis*, volume 1. `ClassicalRealAnalysis.com`, 2008.

[15] Charalampos E. Tsourakakis. C284r: Social data mining. 2005.

[16] Matthew Walsh. The hub number of a graph. *Int. J. Math. Comput. Sci*, 1(1):117–124, 2006.

[17] Duncan J Watts and Steven H Strogatz. Collective dynamics of ‘small-world’networks. *nature*, 393(6684):440–442, 1998.

[18] Wikipedia. Erdős-rényi random graph. `https://en.wikipedia.org/wiki/Erd%C5%91s%E2%80%93R%C3%A9nyi_model`.

[19] Wikipedia. Exponential integral. `https://www.wikiwand.com/en/Exponential_integral`.

[20] Wikipedia. Poisson summation formula. `https://www.wikiwand.com/en/Poisson_summation_formula`.

[21] Wikipedia. Stirling’s approximation. `https://www.wikiwand.com/en/Stirling%27s_approximation`.

Case for Independent Work

Please provide detailed evidence of independent work. For example:

- The examples in the Preliminaries, i.e. Examples 2.10, 2.22 and 2.21 are my own.
- The proof of Proposition 3.4 is my own.
- Proposition 3 has been adapted from the lecture notes [15] but reworded. This in general is a well-known result, many notes mention it.
- The result of Proposition 4.1 was taken from the Wikipedia article for Erdős-Rényi Random Graphs [18], but the proof is my own work.
- The proof of Proposition 4.3 is an expanded version of [2], in particular the computation is my own.
- Example 4.7 is my own.
- The proof of Proposition 4.8 is my own.
- The workings in Chapter 5 and the 2 theorems in section 2 are following the results from Gilbert's original paper [7], but as the proofs and calculations in the paper are omitted, these are my own work.
- Figures 1, 2 are my own measurements. Code available on my GitHub.
- The calculations in Chapter 6 Section 1 are completely my own work.
- Lemma 6.4 was taken from [5], the proof given by Fronczak et al. has given the general direction, but almost all of it had to be reconstructed. This was one of the hardest proofs in the project, even though the result was given.
- The calculations in Chapter 6 Section 2 are adapted from [5]. The paper states a more general result, this was reworded and simplified here, as well as expanded. Doing so revealed a number of assumptions underlying the results, and we ended up consulting with the authors of the paper.
- Figure 3 shows my own measurements, and experiment.
- I had help in the proof of Proposition 7.2, namely to count the number of edges absent, instead of the edges present in the graph.

Checklist

□ Title Page
□ Table of Contents
□ Abstract
□ Introduction
□ Chapters
□ Bibliography
□ Case for Independent Study
□ Checklist