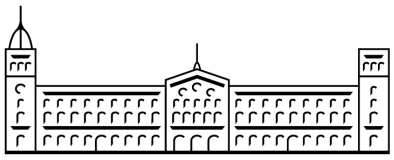


Treball Final de Grau

GRAU DE MATEMÀTIQUES

Facultat de Matemàtiques
Universitat de Barcelona

Random Graphs and Applications

Autor: Adrià Parés
Director: Dr. Javier Soria
Realitzat a: Departament de Matemàtica Aplicada i Anàlisi

Barcelona, June 30, 2015

Contents

Chapter 1

Abstract and outline

Graph theory has been a wide area of study of discrete mathematics since the publication of the Königsberg bridge solution by Leonard Euler in 1736. Despite its historical background and very exciting developments since its birth, graph theory was unable to prove useful when studying complex networks. These networks contain large amounts of vertices and edges, and the sheer quantity of data that has to be handled makes the classical approach not optimal at best and impossible in most cases.

The last decade has seen an uprising of the random network theory, which attempts to study the topology of these complex networks via a statistical approach. This theory has proven very successful at modeling these networks, particularly when applied to the degree distribution of the vertices of the graphs.

This manuscript will attempt to summarize the two most important models from a historical point of view. First, we will describe the model created by Paul Erdös and Alfréd Rényi in 1960, which is considered one of the first to attempt to describe these networks. The second model was introduced by Lázlo Barabási and Réka Albert in the late 1990s. This model can be considered the spiritual successor of the Erdös-Rényi model, while expanding it with some additional properties that were not considered at first. It is particularly important because it motivates a new wave of scientific study about complex network due to essentially two factors. First, there was more data available on complex networks compared to the times when the Erdös- Rényi model was proposed. Lastly, computers were now capable of handling the calculations required to model these networks and were available to the majority of the scientific community, which made the topic much more accessible for new research to be conducted.

This manuscript is organized as follows.

Chapter 2 will present the basic definitions of graph theory and other concepts that are essential for the understanding of the manuscript.

Chapter 3 will present the Erdös-Rényi model (ER model) [11, 12], the first proposed model to analyze random networks. We begin with a set number of isolated vertices and we add edges following a predetermined probability. We will prove that a graph constructed this way will have a degree sequence that follows a Poisson distribution (Theorem 3.1). Further, this graph will also contain a giant connected

component (Theorem 3.4), whose size (Theorem 3.5) and evolution (Theorem 3.6) will also be studied. We will also create a computer simulation and test the theoretical results with it. Finally, at the end of the chapter we discuss the small-world property and the small-world experiment.

Chapter 4 will present the Barabási-Albert model [5, 6], an evolution of the previous one that adds the properties of growth and preferential linking of the network. We start with one vertex and a loop, and we add more vertices and edges, with highly connected vertices being more likely to get and edge than the less connected ones. We will prove that a graph constructed this way has a degree sequence that follows a power law distribution (Theorem 4.2). Finally, we will create a computer simulation to test the theoretical results.

Chapter 5 will present some example cases to exhibit the utility and potential of using random network theory to study some real life networks, such as the citation network [15], the metabolic network [14] and the WWW [1, 7]. We also comment some other case studies and future improvements on the established models.

Chapter 2

Introduction to graph theory and additional preliminaries

2.1 Graph Theory. Definitions and notations

This chapter is intended to serve as a brief introduction to some elementary graph theory concepts and properties, which will appear throughout the manuscript and are needed to understand most of what will be exposed. For a more in depth study of these concepts, as well as some extra results, see [9, 13].

Definition 2.1. A *graph* $G_m^n = (M, N)$ consists of a pair of sets M and N, such that $N \neq \emptyset$ and M is a set of pairs of elements of N. The set M has m elements, and the set N has n elements.

Definition 2.2. We call the elements of N the *vertices* and the elements of M the *edges*. We also say that the edges *connect* the vertices.

Definition 2.3. A *subgraph* of a graph G is a graph that has vertices and edges from the graph G.

Definition 2.4. A *loop* is an edge that connects a vertex with itself.

Definition 2.5. If two or more edges connect the same vertices, we call them *multiple edges*.

Definition 2.6. A *simple* graph is a graph that has no loops and no multiple edges.

Definition 2.7. An *ordinary* or *undirected* graph is a graph where the edges are defined in terms of unordered vertices, i.e., the edge $(1, 2)$ is equivalent to the edge $(2, 1)$.

Definition 2.8. A *directed* graph is a graph where the edges are defined in terms of ordered vertices, i.e., the edge $(1, 2)$ and the edge $(2, 1)$ are considered as different edges. We call them *directed* edges.

Remark 2.9. We will make no distinction between directed edges and "normal" edges, and we will simply call them edges.

Definition 2.10. Given a directed edge (n_1, n_2), we say that the edge *starts* at n_1 and *ends* at n_2.

Definition 2.11. Given a directed graph, the *indegree* of a vertex is the number of edges that end in the vertex. Similarly, the *outdegree* of a vertex is the number of edges that start in the vertex.

Definition 2.12. The *total degree* of a vertex is the number of edges that connect that vertex with another, i.e., the sum of the indegree and outdegree of the vertex. In an undirected graph, we call it simply the *degree* of the vertex.

Remark 2.13. Loops add 1 to the indegree and to the outdegree, and 2 to the total degree of the vertex.

Definition 2.14. The *degree sequence* of a graph is the sequence obtained by ordering all the degrees of all the vertices of the graph, in a growing order.

Definition 2.15. We say that two vertices v_1, v_k are connected by a *path* if we can find a sequence of edges $(v_1, v_2), (v_2, v_3), (v_3, v_4), \ldots, (v_{k-1}, v_k)$. It is noted as $c = v_1 v_2 v_3 v_4 \ldots v_{k-1} v_k$.

Definition 2.16. We say that a graph is *weighted* if there exists a function such that $w : M(G_m^n) \to \mathbb{R}^+$, that is, w assigns a *weight* to the edges of the graph.

Definition 2.17. The *length* of a path $c = v_1 v_2 \ldots v_k$ in a weighted graph is defined as:

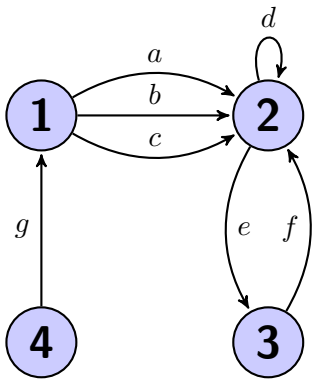
$$L_w(c) = \sum_{i=1}^{k-1} w((v_i, v_{i+1}))$$

Remark 2.18. Throughout the manuscript we will assume that all edges have weight 1. This means that the length of a path is simply the number of edges that form the path.

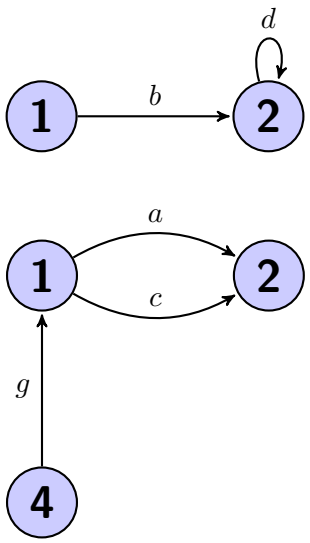
Definition 2.19. A *geodesic* between 2 vertices is the path with the least length that connects those 2 vertices.

Definition 2.20. The *diameter* of a graph is the maximum length of all the geodesics in the graph, i.e., the maximum minimal length between all the vertices in the graph.

Example 2.21.



- There are 4 vertices, $N = \{1, 2, 3, 4\}$, labeled from 1 to 4.
- There are 7 edges, $M = \{(1, 2), (1, 2), (1, 2), (2, 2), (2, 3), (3, 2), (4, 1)\}$, labeled from a to g.
- The edges a, b and c, described as (1,2), are multiple edges.
- The edge d, described as (2,2), is a loop.
- Two subgraphs of this graph could be:



- This graph is directed, since edges e and f are not the same: edge e is (2,3) and edge f is (3,2).
- Edge a, (1,2), starts at vertex 1 and ends at vertex 2.
- Vertex 1 has indegree 1, outdegree 3 and total degree 4.
- The degree sequence of this graph is $(1, 2, 4, 7)$.
- Vertices 4 and 3 are connected, since we can create the path $\{(4, 1), (1, 2), (2, 3)\}$ through edges g, a and e.

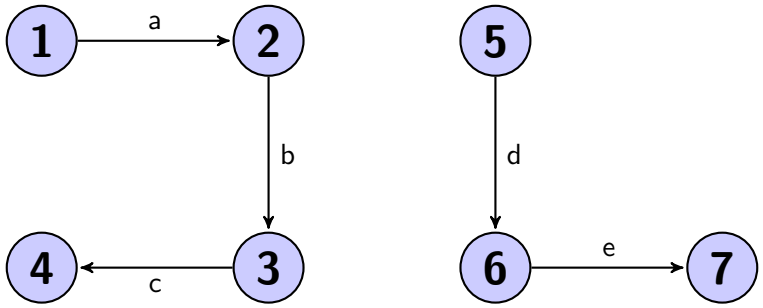
Definition 2.22. A *tree* is an undirected graph in which any two vertices are connected by exactly one path.

Definition 2.23. A *connected graph* is a graph where there exists a path connecting any given pair of vertices. If a graph is not connected, we call it a *disconnected graph*.

Example 2.21 is an example of connected graph.

Definition 2.24. Given a disconnected graph, a *connected component* is a subgraph in which any two vertices are connected, and which is connected to no additional vertices in the graph.

Example 2.25.



- This graph contains 2 trees, $\{1, 2, 3, 4\}$ and $\{5, 6, 7\}$.
- This graph is disconnected, since there are no paths connecting vertices from the first tree to the second one.
- Both trees are connected components of this graph, and hence it has 2 of them.

2.2 Additional preliminaries

In this section we will list all definitions (mainly from probability theory) that cannot be listed under graph theory but are important throughout the manuscript nonetheless. For a more in depth study of these concepts, as well as some extra results, see [2, 3, 10].

Definition 2.26. We say that two non zero functions f and g are *proportional* if

$$\frac{f}{g} = constant.$$

We denote it as $f \propto g$.

Definition 2.27. We say that a function $f(x)$ is *asymptotically equivalent* to $g(x)$, and we denote it by $f(x) \sim g(x)$ if

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1.$$

Definition 2.28. A *random variable* $X : \Omega \to E$ is a measurable function from the set of possible outcomes Ω to some set E, typically $E = \mathbb{R}$. Intuitively, it is a variable whose value is subject to variations due to chance. A random variable is *discrete* if it takes values from a countable list of values, and is *continuous* if it takes values from an interval or a collection of intervals.

It should be assumed that, unless stated otherwise, random variables are discrete throughout the manuscript.

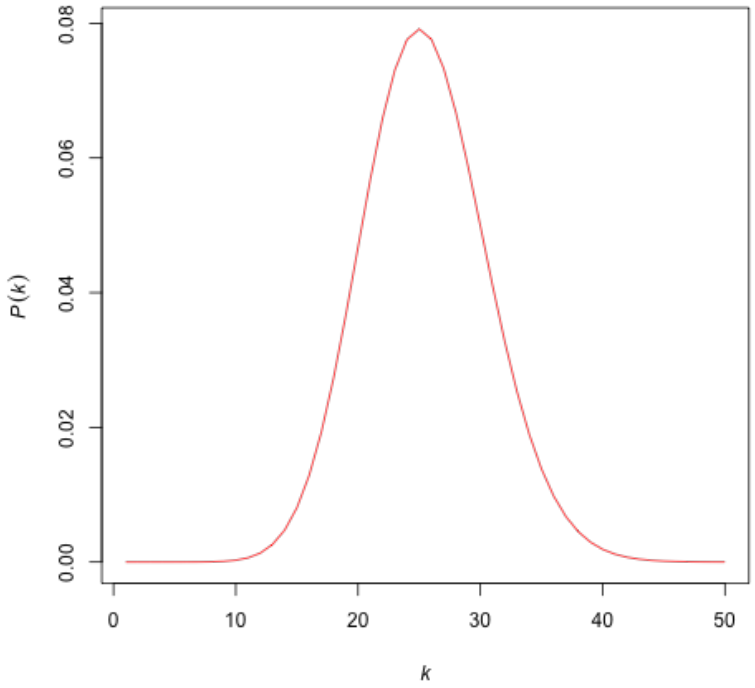


Figure 2.1: Representation of $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$, using $\lambda = 25.5$.

Definition 2.29. A discrete random variable X is said to have a *Poisson distribution* with parameter $\lambda > 0$ if, for $k = 0, 1, 2, \ldots$ the probability of X is given by

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

Figure 2.1 shows a graphical representation of a Poisson distribution.

Definition 2.30. The *expectation* or *expected value* of a random variable X is the probability-weighted average of all possible values. We denote it by $\mathbf{E}(X)$.

If we assume that X is a discrete random variable that can take value x_i with probability P_i, $i \in \mathbb{N}$, then the expectation is defined as

$$\mathbf{E}(X) = \sum_{i \in \mathbb{N}} x_i P_i$$

Example 2.31. Let X represent the outcome of a roll of a six-sided dice. Assuming that it's a fair dice, $P_i = \frac{1}{6}$ and $x_i = i$, $i = 1, 2, 3, 4, 5, 6$. Therefore, we have:

$$\mathbf{E}(X) = \frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6} = \frac{7}{2}.$$

Definition 2.32. The *conditional expectation* of a random variable X given an event H is another random variable equal to the average of the former over all possible outcomes in H. We denote it by $\mathbf{E}(X|H)$.

It should be noted that the same idea can be applied to the conditional expectation of a random variable X over a discrete random variable Y, denoted by $\mathbf{E}(X|Y)$.

Definition 2.33. A *scale-free network* is a graph whose degree sequence follows a power law distribution. That is, the fraction $P(k)$ of nodes in the network having k connections to other nodes goes as

$$P(k) \sim k^{-\gamma},$$

when the network has a big enough amount of vertices.

Figure 2.2 shows a graphical representation of a power law distribution.

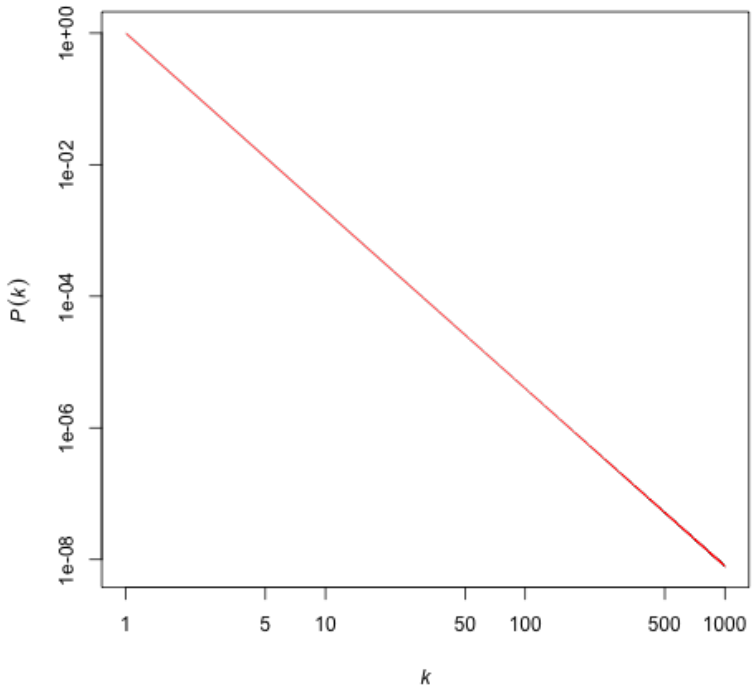


Figure 2.2: Representation of $P(X = k) = k^{-\gamma}$, using $\gamma = 2.7$. Note that the axis are in logarithmic scale.

Definition 2.34. Given a probability space (Ω, F, P) and measurable space (S, Σ), an S-valued *stochastic process* is a collection of S-valued random variables on Ω,

indexed by a totally ordered set T, called *time*. That is, a stochastic process X is a collection

$$\{X_t : t \in T\},$$

where each X_t is an S-valued random variable on Ω. The space S is then called the *state space* of the process.

Definition 2.35. A *discrete-time* stochastic process is one for which the index variable takes a discrete set of values.

Definition 2.36. A *discrete-time martingale* is a discrete-time stochastic process that satisfies, for any given $n \in \mathbb{N}$,

$$\mathbf{E}(|X_n|) < \infty, \quad \mathbf{E}(X_{n+1} \mid X_1, \ldots, X_n) = X_n.$$

That is, the conditional expected value of the next observation, given all the past observations, is equal to the last observation. Due to the linearity of expectation, this second requirement is equivalent to:

$$\mathbf{E}(X_{n+1} - X_n \mid X_1, \ldots, X_n) = 0$$

or

$$\mathbf{E}(X_{n+1} \mid X_1, \ldots, X_n) - X_n = 0,$$

which states that the average "winnings" from observation n to observation $n+1$ are 0.

Example 2.37. We shall show an example called de Moivre's martingale. Let us suppose and "unfair" coin with probability p of heads and probability $q = 1 - p$ of tails. Let

$$X_{n+1} = X_n \pm 1,$$

with + in case of heads and - in case of tails. Let

$$Y_n = \left(\frac{q}{p}\right)^{X_n}.$$

Then, $\{Y_n : n \in \mathbb{N}\}$ is a martingale with respect to $\{x_n : n \in \mathbb{N}\}$. We can see that:

$$\begin{aligned}\mathbf{E}(Y_{n+1}|X_1, \ldots, X_n) &= p(q/p)^{X_n+1} + q(q/p)^{X_n-1} \\ &= p(q/p)(q/p)^{X_n} + q(p/q)(q/p)^{X_n} = q(q/p)^{X_n} + p(q/p)^{X_n} \\ &= (q/p)^{X_n} = Y_n.\end{aligned}$$

Definition 2.38. Let f and g be two functions defined on some subset of the real numbers.

- We write

  $$f(x) = O(g(x)) \text{ as } x \to \infty$$

  if and only if $\exists x_0 \in \mathbb{R}$ such that $\forall x \geq x_0$, $\exists M > 0$ such that

  $$|f(x)| \leq M|g(x)|.$$

  This is referred to as *big O notation*.

- We write
$$f(x) = o(g(x)) \text{ as } x \to \infty$$
if and only if $\forall M > 0, \exists x_0$ such that $\forall x \geq x_0$,
$$|f(x)| \leq M|g(x)|.$$
This is referred to as *little o notation*.

Note the difference between the definitions: while big O notation has to be true for at least one M, little o notation must hold for all $M > 0$, however small. In this sense, little o notation makes a stronger statement than big O notation.

Definition 2.39. Let x be a real number.

- The *floor function* is a function $\lfloor x \rfloor : \mathbb{R} \to \mathbb{Z}$ that maps a real number x to the largest previous integer.
- The *ceiling function* is a function $\lceil x \rceil : \mathbb{R} \to \mathbb{Z}$ that maps a real number x to the smallest following integer.

Theorem 2.40. *(Weak law of large numbers). The weak law of large numbers states that the sample average converges in probability towards the expected value*
$$\overline{X}_n \xrightarrow{P} \mu \quad \text{when } n \to \infty.$$
That is to say, for any $\varepsilon > 0$,
$$\lim_{n\to\infty} \mathbf{P}\left(|\overline{X}_n - \mu > \varepsilon|\right) = 0.$$

Theorem 2.41. *(Stirling's formula)*
$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \left(1 + O\left(\frac{1}{n}\right)\right).$$

Chapter 3

The Erdös-Rényi model

Contemporary science has seemingly stumbled upon an insurmountable wall: complexity. Its inability to describe complex systems, particularly those composed of non-identical elements, with diverse interactions between them, currently cripples development in many disciplines, ranging from social studies to computer science and molecular biology. The fact that complexity appears in such unrelated topics is interesting on itself, and any breakthrough in its study may provide tools to solve a wide range of problems in multiple disciplines.

The main difficulty lies in the system's topology: due to the size and complexity of these networks, it is mainly unknown. During several years, many models have been presented to try and establish some main properties about these systems' topology.

The Erdös-Rényi model (from now on, ER model) is probably the oldest and most studied model to date; its approach is mainly probabilistic. Introduced in 1959 [11] and deeply studied in 1960 [12], this model constructs a random network by adding edges to a graph consisting of N isolated vertices. Each edge has the same probability of being added to the graph, which is one of the main assumptions of the model.

While this approach would be proved too simplistic, it is nevertheless worth considering, if only because its finding would motivate and influence future research in the topic.

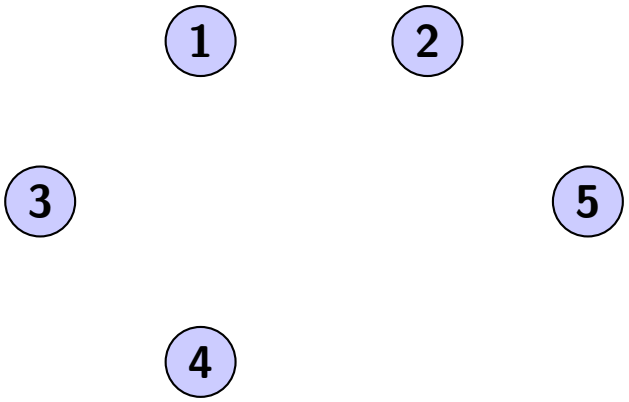
In this chapter we shall describe the simplified version of the ER model, as well as the most important results about the random network connectivity: the degree sequence distribution and the biggest connected component.

3.1 The model

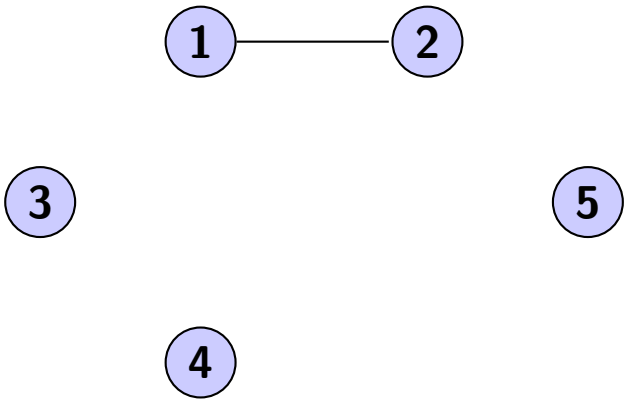
There are multiple models that can be called ER models, but since they coincide as $N \to \infty$, we shall describe the one that is more suited for analytical calculations.

We start with a graph with N vertices and no edges and a fixed probability $0 < p < 1$. For each pair of vertices, we will add an edge that connects them with probability p. For simplicity, we will not allow loops or multiple edges. We do this once for every pair of vertices until all of them have been tried.

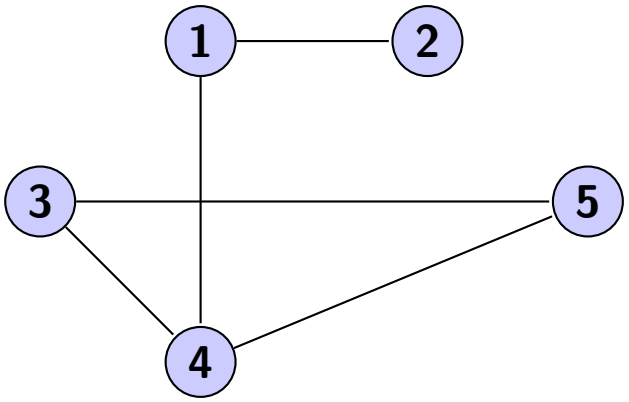
- We start with 5 vertices for this example, and we set $p = 0.5$.



- We start by checking the edge $(1, 2)$. For this example, we will toss a coin. If we get heads, we will add the edge. Otherwise, we don't add any edge to the graph and proceed to the next pair. Note that this graph is not directed, so checking for the edge $(1, 2)$ is the same as checking for the edge $(2, 1)$. We flip it and get heads, so we add the edge and move to the next pair of vertices.



- We repeat this process for every pair of vertices.



Note that there are a lot of different random graphs that can be generated with these parameters (5 vertices, $p = 0.5$), since every iteration will probably generate a different graph.

3.2 Degree sequence and Poisson distribution

The most studied property of random networks is the degree sequence, since it is one of the easiest to calculate. From the analysis of this sequence we can determine essentially all of the significant statistical properties of the graph. In [12], the following theorem sums up the most important result about the degree sequence in random graphs, and will motivate future works in [5, 6, 8]. As we will see in the next chapter, this result is not accurate, since random graphs don't generally follow it, but it is important within the model nevertheless.

Theorem 3.1. *Let $d_{n,N(n)}(P_k)$ denote the degree of the vertex P_k in $G^n_{N(n)}$ (i.e., the number of vertices of $G^n_{N(n)}$ which are connected with P_k by any edge). Put*

$$\underline{d}_n = \min_{1\le k\le n} d_{n,N(n)}(P_k) \text{ and } \overline{d}_n = \max_{1\le k\le n} d_{n,N(n)}(P_k).$$

Suppose that

$$\lim_{n\to+\infty} \frac{N(n)}{n\log n} = +\infty.$$

Then, we have for any $\varepsilon > 0$

$$\lim_{n\to+\infty} \mathbf{P}\left(\left|\frac{\overline{d}_n}{\underline{d}_n} - 1\right| < \varepsilon\right) = 1.$$

Further, for $N(n) \sim cn$ and for any k

$$\lim_{n\to+\infty} \mathbf{P}(d_{n,N(n)}(P_k) = j) = \frac{(2c)^j e^{-2c}}{j!} \quad (j = 0, 1, \dots).$$

Proof. The probability that a given vertex P_k shall be connected by exactly r others in G^n_m is

$$\frac{\binom{n-1}{r}\binom{\binom{n-1}{2}}{N-r}}{\binom{\binom{n}{2}}{N}} \sim \frac{(\frac{2N}{n})^r e^{-\frac{2N}{n}}}{r!}.$$

Thus if $N(n) \sim cn$ the degree of a given vertex has approximately a Poisson distribution with mean value $2c$. The number of vertices having the degree r is thus in this case approximately

$$n\frac{(2c)^r e^{-2c}}{r!} \quad (r = 0, 1, \dots).$$

If $N(n) = (n\log n)w_n$ with $w_n \to +\infty$ then the probability that the degree of a vertex will be outside the interval $\frac{2N(n)}{n}(1-\varepsilon)$ and $\frac{2N(n)}{n}(1+\varepsilon)$ is approximately

$$\sum_{|k-2w_n\log n|>2\varepsilon w_n\log n} \frac{(2w_n\log n)^k e^{-2w_n\log n}}{k!} = O\left(\frac{1}{n^{\varepsilon^2 w_n}}\right)$$

and thus this probability is $o\left(\frac{1}{n}\right)$, for any $\varepsilon > 0$. Thus the probability that the degrees of not all n vertices will be between the limit $(1\pm\varepsilon)2w_n\log n$ will be tending to 0. Thus the theorem is proven. □

This essentially means that the probability that a vertex has k edges follows a Poisson distribution (Definition 2.29)

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

where λ is the mean value of the degrees in the graph, that is:

$$\lambda = \frac{1}{n}\sum_{k=1}^{n} d_{n,N(n)}(P_k)$$

3.3 Biggest connected component

Another important result from [12] is the study of the biggest connected component in the random graph, called simply the *giant component*. The existence, size and evolution of this component is described in the next 3 theorems. First, we prove the following Lemma.

Lemma 3.2. *Let $a_1, a_2, \ldots, a_r$ be positive numbers such that $\sum_{j=1}^{r} a_j = 1$.*

If $\max_{1\leq j\leq r} a_j \leq \alpha$, then there is a value k, $(1 \leq k \leq r-1)$ such that

$$\frac{1-\alpha}{2} \leq \sum_{j=1}^{k} a_j \leq \frac{1+\alpha}{2}$$

and

$$\frac{1-\alpha}{2} \leq \sum_{j=k+1}^{n} a_j \leq \frac{1+\alpha}{2}.$$

Proof. Put $S_j = \sum_{i=1}^{j} a_i$, (j=1, 2,. . . ,r). Let j_0 denote the smallest integer for which $S_j > 1/2$. If $S_{j_0} - 1/2 > 1/2 - S_{j_0-1}$, we define $k = j_0 - 1$. If $S_{j_0} - 1/2 \leq 1/2 - S_{j_0-1}$,we define $k = j_0$. In both cases we have

$$\left| S_k - \frac{1}{2} \right| \leq \frac{a_{j_0}}{2} \leq \frac{\alpha}{2},$$

which proves our lemma. □

We need one last theorem before we proceed.

Theorem 3.3. *Let $V(n, m)$ denote the number of vertices of a given graph G_m^n that belong to an isolated tree contained within G_m^n. Let us suppose that*

$$\lim_{n\to+\infty} \frac{m(n)}{n} = c > 0.$$

Then, we have

$$\lim_{n\to+\infty} \frac{\mathbf{M}[V(n,m(n))]}{n} = \begin{cases} 1 & \text{for } c \leq \frac{1}{2} \\ \frac{x(c)}{2c} & \text{for } c > \frac{1}{2} \end{cases}$$

where $m(n)$ is a function of n, $\mathbf{M}[V(n,m(n))]$ is the mean value of $V(n,m(n))$ and $x(c)$ is the only root of the equation

$$xe^{-x} = 2ce^{-2c}$$

that satisfies $0 < x < 1$. We can also obtain $x(c)$ as

$$x(c) = \sum_{k=1}^{\infty} \frac{k^{k-1}}{k!}(2ce^{-2c})^k. \tag{3.1}$$

Proof. We shall need the well known fact that the inverse function of the function $y = xe^{-x}$, with $0 \leq x \leq 1$, has the power series expansion

$$x = \sum_{k=1}^{\infty} \frac{k^{k-1}y^k}{k!}, \tag{3.2}$$

convergent for $0 \leq y \leq \frac{1}{e}$. Let τ_k denote the number of isolated trees of order k contained in G_m^n. Then, we clearly have

$$V(n,m) = \sum_{k=1}^{n} k\tau_k,$$

and thus

$$\mathbf{M}[V(n,m(n))] = \sum_{k=1}^{n} k\mathbf{M}(\tau_k).$$

Then [12], we have that

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{M}(\tau_k) = \frac{1}{2c}\,\frac{k^{k-2}}{k!}\,(2ce^{-2c})^k,$$

and thus, for $c \leq \frac{1}{2}$,

$$\liminf \frac{\mathbf{M}[V(n,m(n))]}{n} \geq \frac{1}{2c}\sum_{k=1}^{s} \frac{k^{k-1}(2ce^{-2c})^k}{k!}$$

for any $s \geq 1$. Therefore, we can also claim that

$$\liminf_{n\to\infty} \frac{\mathbf{M}[V(n,m(n))]}{n} \geq \frac{1}{2c}\sum_{k=1}^{\infty} \frac{k^{k-1}(2ce^{-2c})^k}{k!}. \tag{3.3}$$

But according to (3.2), for $c \leq \frac{1}{2}$ we have

$$\sum_{k=1}^{\infty} \frac{k^{k-1}(2ce^{-2c})^k}{k!} = 2c.$$

It follows from (3.3) that for $c \leq \frac{1}{2}$

$$\liminf_{n\to\infty} \frac{\mathbf{M}[V(n, m(n)]}{n} \geq 1.$$

On the other hand, $V(n, m(n)) \leq n$ and thus

$$\limsup_{n\to\infty} \frac{\mathbf{M}[V(n, m(n))]}{n} \leq 1.$$

All of these combined yield that, for $c \leq \frac{1}{2}$,

$$\lim_{n\to\infty} \frac{\mathbf{M}[V(n, m(n))]}{n} = 1.$$

We now consider the case $c > \frac{1}{2}$. Similarly as with the previous case, we can obtain

$$\mathbf{M}[V(n, m(n))] = \frac{n^2}{2m} \sum_{k=1}^{n} \frac{k^{k-1}}{k!} \left(\frac{2m(n)}{n} e^{-\frac{-2m(n)}{n}} \right)^k + O(1),$$

where the bound of the term $O(1)$ depends only on c. Since

$$\sum_{k=n+1}^{\infty} \frac{k^{k-1}}{k!} \left(\frac{2m(n)}{n} e^{-\frac{-2m(n)}{n}} \right)^k = O(n^{-\frac{3}{2}})$$

for $m(n) \sim cn$ with $c > \frac{1}{2}$, it follows that

$$\mathbf{M}[V(n, m(n))] = \frac{n^2}{2m(n)} x\left(\frac{m(n)}{n}\right) + O(1),$$

where $x = x\left(\frac{m(n)}{n}\right)$ is the only solution of the equation

$$xe^{-x} = \frac{2m(n)}{n} e^{-\frac{2m(n)}{n}}$$

with $0 < x < 1$. Thus, if (3.2) holds with $c > \frac{1}{2}$, we have

$$\lim_{n\to\infty} \frac{\mathbf{M}[V(n, m(n))]}{n} = \frac{x(c)}{2c},$$

with $x(c)$ defined in (3.1). □

We can now proceed to prove the existence of the giant component.

Theorem 3.4. *(Existence) Let $H_{n,m}(A)$ denote the set of those vertices of G_m^n which belong to components of size $> A$, and let $h_{n,m}(A)$ denote the number of elements of the set $H_{n,m}(A)$. If $m_1(n) \sim n(c-\varepsilon)$, where $\varepsilon > 0$, $c-\varepsilon \geq 1/2$ and $m_2(n) \sim cn$ then, with probability tending to 1 for $n \to +\infty$, from the $h_{n,m_1(n)}(A)$ vertices belonging to $H_{n,m_1(n)}(A)$, more than $(1-\delta)h_{n,m_1(n)}(A)$ vertices will be contained in the same component of $G^n_{m_2(n)}$ for any δ with $0 < \delta < 1$ provided that*

$$A \geq \frac{50}{\varepsilon^2\delta^2}.$$

Proof. From Theorem 3.3 we can see that the mean value of the number of vertices belonging to trees of order $\leq A$ is, with probability tending to 1 for $n \to +\infty$, equal to

$$n\left(\sum_{k=1}^{A}\frac{k^{k-1}}{k!}[2(c-\varepsilon)]^{k-1}e^{-2k(c-\varepsilon)}\right)+o(n),$$

where we have changed $c \to (c-\varepsilon)$ and the sum only extends up to A, since that is the maximum order we are looking for.

On the other hand, the number of vertices of $G^n_{m_1(n)}$ belonging to components of size $\leq A$ and containing exactly one cycle is $o(n)$ for $c-\varepsilon \geq 1/2$ (with probability tending to 1 [12]), while it is easy to see, that the number of vertices of $G^n_{m_1(n)}$ belonging to components of size $\leq A$ and containing more than one cycle is also bounded with probability tending to 1. Let $E_n^{(1)}$ denote the event that

$$|h_{n,m_1(n)}(A) - nf(A, c-\varepsilon)| < \tau n f(A, c-\varepsilon), \tag{3.4}$$

where $\tau > 0$ is an arbitrary small positive number which will be chosen later and

$$f(A,c) = 1 - \frac{1}{2c}\sum_{k=1}^{A}\frac{k^{k-1}}{k!}(2ce^{-2c}) > 0,$$

and let $\overline{E}_m^{(1)}$ denote the contrary event. It follows that

$$\lim_{n\to+\infty}\mathbf{P}(\overline{E}_n^{(1)}) = 0.$$

We consider only such $G^n_{m_1(n)}$ for which (3.4) holds. Now, it is clear that $G^n_{m_2(n)}$ is obtained from $G^n_{m_1(n)}$ by adding $m_2(n) - m_1(n) \sim n\varepsilon$ new edges at random to $G^n_{m_1(n)}$. The probability that such new edge should connect two vertices belonging to $H_{n,m_1(n)}(A)$ is, at least,

$$\frac{\binom{h_{n,m_1(n)}(A)}{2} - m_2(n)}{\binom{n}{2}}$$

and thus by (3.4) is not less than $(1-2\tau)f^2(A, c-\varepsilon)$, if n is sufficiently large and τ sufficiently small. As these edges are chosen independently from each other, it

follows by the law of large numbers (Theorem 2.40) that denoting by v_n the number of those of the $m_2 - m_1$ new edges which connect two vertices of H_{n,m_1} and by $E_n^{(2)}$ the event that

$$v_n \geq \varepsilon(1-3\tau)f^2(A, c-\varepsilon)n \tag{3.5}$$

and by $\overline{E}_n^{(2)}$ the contrary event, we have

$$\lim_{n\to+\infty} \mathbf{P}(\overline{E}_n^{(2)}) = 0.$$

We consider now only such $G^n_{m_2(n)}$ for which $E_n^{(2)}$ takes place. Now let us consider the subgraph $g^n_{m_2(n)}$ of $G^n_{m_2(n)}$ formed by the vertices of the set $H_{n,m_1(n)(A)}$ and only those edges of $G^n_{m_2(n)}$ which connect two such vertices.

Let the sizes of the components of $g^n_{m_2(n)}$ be denoted by $b_1, b_2, \ldots, b_r$. Let E_n^3 denote the event

$$\max b_j > h_{n,m_1(n)(A)(1-\delta)}$$

and $\overline{E}_n^{(3)}$ the contrary event. Applying Lemma 3.2 with $\alpha = 1-\delta$ to the numbers $a_j = \frac{b_j}{h_{n,m_1(n)}(A)}$ it follows that if the event $\overline{E}_n^{(3)}$ takes place, the set $H_{n,m_1(n)}(A)$ can be split in two subsets H_n^1 and H_n^n containing h_n^1 and h_n^n respectively, such that $h_n^1 + h_n^n = h_{n,m_1(n)}(A)$ and

$$h_{n,m_1(n)}(A)\frac{\delta}{2} \leq \min(h_n^1, h_n^n) \leq \max(h_n^1, h_n^n) \leq h_{n,m_1(n)}(a)\left(1-\frac{\delta}{2}\right). \tag{3.6}$$

Further, no vertex of H_n^1 is connected with a vertex of H_n^n in $g^n_{m_2(n)}$. It follows that if a vertex P of the set $H_{n,m_1(n)}(A)$ belongs to H_n^1 then all other vertices of the component of $G^n_{m_1(n)}$ to which P belongs are also contained in H_n^1. As the number of components of size $> A$ of $G^n_{m_1(n)}$ is clearly $< \dfrac{h_{n,m_1(n)}(A)}{A}$ the number of such divisions of the set $H_{n,m_1(n)}(A)$ does not exceed $2^{\frac{1}{A}h_{n,m_1(n)}(A)}$. This is also true for H_n^n. If further $\overline{E}_n^{(3)}$ takes place then every one of the v_n new edges connecting vertices of $H_{n,m_1(n)}(A)$ connects either two vertices of H_n^1 or two vertices of H_n^n. The possible number of such choices of these edges is clearly

$$\binom{\binom{h_n^1}{2} + \binom{h_n^n}{2}}{v_n}.$$

As by (3.6),

$$\frac{\binom{h_n^1}{2} + \binom{h_n^n}{2}}{\binom{h_n}{2}} \leq \frac{\delta^2}{4} + \left(1+\frac{\delta}{2}\right)^2 = 1 - \delta + \frac{\delta^2}{2} \leq 1 - \frac{\delta}{2}$$

it follows that

$$\mathbf{P}(\overline{E}_n^{(3)}) \leq 2^{\frac{1}{A}h_{n,m_1(n)}(A)}\left(1-\frac{\delta}{2}\right)^{\varepsilon(1-3\tau)f^2(A,c-\varepsilon)n}$$

and thus by (3.4) and (3.5),

$$\mathbf{P}(\overline{E}_n^{(3)}) \leq \exp\left[nf(A, c-\varepsilon)\left(\frac{(1+\tau)\log 2}{A} - \frac{\varepsilon(1-3\tau)f(A,c-\varepsilon)\delta}{2}\right)\right]. \quad (3.7)$$

Thus, if

$$A\varepsilon\delta(1-3\tau)f(A,c-\varepsilon) > (1+\tau)\log 4$$

then

$$\lim_{n\to+\infty} \mathbf{P}(\overline{E}_n^{(3)}) = 0.$$

In case $c-\varepsilon > 1/2$ we have $f(A, c-\varepsilon) \geq G(c-\varepsilon) > 0$ for any A, while in case $c-\varepsilon = 1/2$ we have

$$f\left(A, \frac{1}{2}\right) = 1 - \sum_{k=1}^{A} \frac{k^{k-1}}{k!e^k} = \sum_{k=A+1}^{\infty} \frac{k^{k-1}}{k!e^k} \geq \frac{1}{2\sqrt{A}} \text{ if } A \geq A_0$$

the inequality (3.7) will be satisfied provided that $\tau < 1/10$ and $A > \frac{50}{\varepsilon^2\delta^2}$. □

Clearly, the giant component of $G^n_{m_2(n)}$ the existence of which has now been proven, contains more than $(1-\tau)(1-\delta)nf(A, c-\varepsilon)$ vertices. By choosing ε, τ and δ sufficiently small and A sufficiently large, $(1-\tau)(1-\delta)nf(A, c-\varepsilon)$ can be brought as near to $G(c)$ as we want. Thus, we have also proven the following.

Theorem 3.5. *(Size) Let $l_{n,m}$ denote the size of the greatest component of G^n_m. If $N(n) \sim cn$ where $c > 1/2$ we have, for any $\delta > 0$,*

$$\lim_{n\to+\infty} \mathbf{P}\left(\left|\frac{l_{n.N(n)}}{n} - G(c)\right| < \delta\right) = 1,$$

where $G(c) = 1 - \dfrac{x(c)}{2c}$ and $x(c) = \displaystyle\sum_{k=1}^{\infty} \frac{k^{k-1}}{k!}(2ce^{-2c})^k$ is the solution satisfying $0 < x(c) < 1$ of the equation $x(c)e^{-x(c)} = 2ce^{-2c}$.

The last theorem indirectly refers to the evolution of the component, which grows by absorbing smaller connected components, which ultimately can be simplified as isolated trees.

Theorem 3.6. *(Evolution) The probability that an isolated tree of order k which is present in $G^n_{m_1(n)}$ where $m_1(n) \sim cn$ and $c > 1/2$ should still remain an isolated tree in $G^n_{m_2(n)}$ where $m_2(n) \sim (c+t)n$, with $t > 0$, is approximately an exponential distribution with mean value $\frac{n}{2k}$ and is independent of the "age" of the tree.*

Proof. The probability that no vertex of the tree in question will be connected with any vertex is

$$\prod_{j=m_1(n)+1}^{m_2(n)} \left(\frac{\binom{n-k}{2} - j + k}{\binom{n}{2} - j}\right) \sim e^{-2kt}.$$

□

Hence it is proven that these trees do not tend to survive and are therefore absorbed by small components. Ultimately, we have established that there is a giant connected component in the graph that tends to grow by absorbing smaller components.

3.4 Computational model

We will now simulate the theoretical model. The code for the ER model is the following, using R as the programming language:

```
require(igraph)
er_vertices<-150 #We can put any positive integer here
er_edges<-er_vertices*(er_vertices-1)/3
er_matrix<-integer(er_vertices*er_vertices)
dim(er_matrix)<-c(er_vertices,er_vertices)
er_degree<-integer(er_vertices)
for(i in 1:er_edges){
          er_rowcol<-sample(1:er_vertices,2)
          if(er_matrix[er_rowcol[1],er_rowcol[2]]==0){
                    er_matrix[er_rowcol[1],er_rowcol[2]]<-1
                    er_matrix[er_rowcol[2],er_rowcol[1]]<-1
                    er_degree[er_rowcol[1]]<-er_degree[er_rowcol[1]]+1
                    er_degree[er_rowcol[2]]<-er_degree[er_rowcol[2]]+1
          }else{
                    i<-i-1
          }
}
er_k<-0:(er_vertices-1)
d1<-dpois(er_k,mean(er_degree))
plot(density(er_degree),xlab=expression(italic(k)),
ylab=expression(italic(P(k))),col="red",main = "ER Model")
lines(er_k,d1,col="green")
dev.copy(png,paste('ER_dist_',er_vertices,'.png'))
dev.off()
er_graph<-graph.adjacency(er_matrix,mode="undirected")
plot.igraph(er_graph,layout=layout.circle)
dev.copy(png,paste('ER_graph_',er_vertices,'.png'))
dev.off()
```

Note that we have to define how many vertices and edges we want to add in the graph, and those are the two only variables that we can modify. The amount of edges cannot be bigger than $\frac{n(n-1)}{2}$, with n being the amount of vertices of the graph. These can be modified at will to produce new simulations.

3.4.1 Degree distribution

We can compare the degree distribution with a Poisson distribution. Here, we take λ as the mean value of the degree of all vertices. Sadly, we are unable to get past 10^5 vertices using this algorithm and our computer, although it yields a nice correlation with the theory. Figure 3.1 shows some of the resulting graphics (red is our model, green is the associated Poisson distribution).

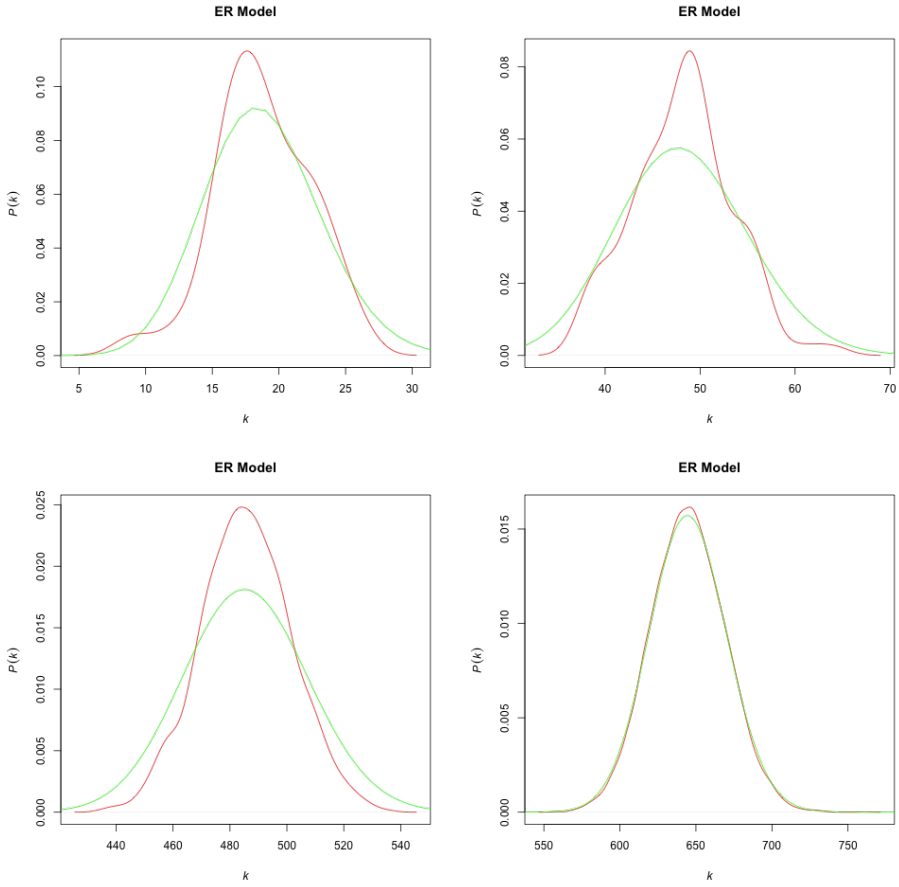


Figure 3.1: Probability distribution for graphs with 40, 100, 1000 and 10000 vertices, constructed via the ER model.

We can see that for low number of vertices the distributions are similar, although there are significant differences between them, particularly around the mean value, where the computational model has a much higher probability than expected. Nevertheless, the theoretical results are proven for a high amount of vertices, and we can see that the correlation between the two distributions gets "better" as we use more and more vertices, until we get to 10000 vertices, where the lines are almost the same.

It should also be noted that we have chosen a very high number of edges when modeling, although similar results are found when reducing that number by a factor. For a low vertex count, it is not that interesting to test the simulation with fewer

edges, since there are few edges to begin with, hence the reduced value remains very similar to the original.

A much more interesting approach is to reduce the amount of edges for 10000 vertices, which is shown in Figure 3.2.

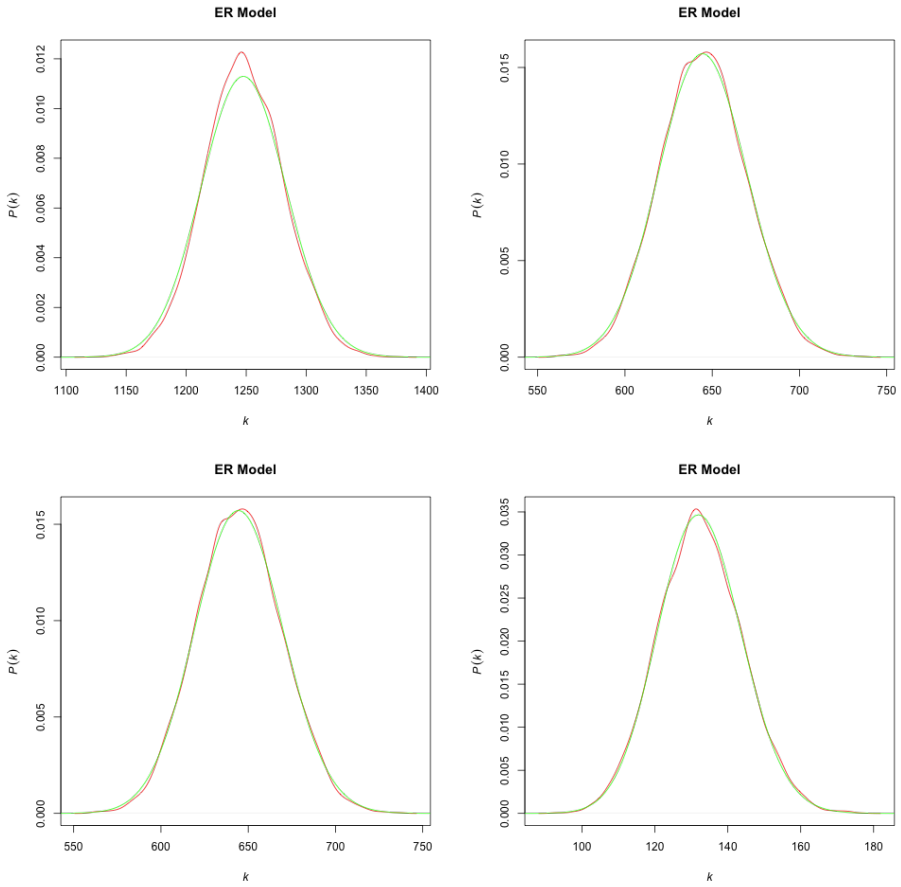


Figure 3.2: Probability distribution for graphs with 10000 vertices and a decreasing factor of 5, 10, 25 and 50 edges, constructed via the ER model.

We can observe that the prediction from the theory still holds even when significantly reducing the amount of edges to add. Sadly, even when this reduction significantly drops the simulation time, adding more vertices again yields the computation impossible for our computer, hence we stop the study of the distribution at this point.

3.4.2 Graph Plot

The igraph package allows us to draw graphs, however even with a low number of vertices (≤ 30) the graphs become unreadable. Nevertheless, it can be instructive to visualize the structure of the graphs with few vertices. Figure 3.3 shows some examples for graphs with few vertices and a high edge count.

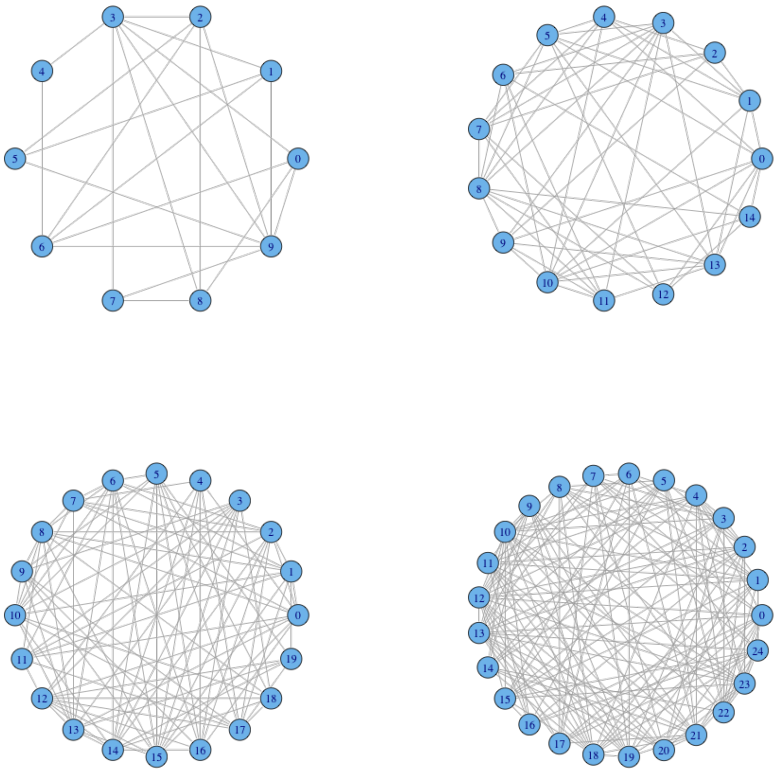


Figure 3.3: Graphs with 10, 15, 20 and 25 vertices generated through the ER model.

These graphs appear very clustered, and it can be interesting to see the structure of the graphs with less edges. Figure 3.4 shows graphs with the same number of vertices as the previous figure, but with an edge count 5 times smaller.

For the graphs with 20 and 25 vertices, we can go one step further and reduce the number of edges by a factor 10. Figure 3.5 shows this.

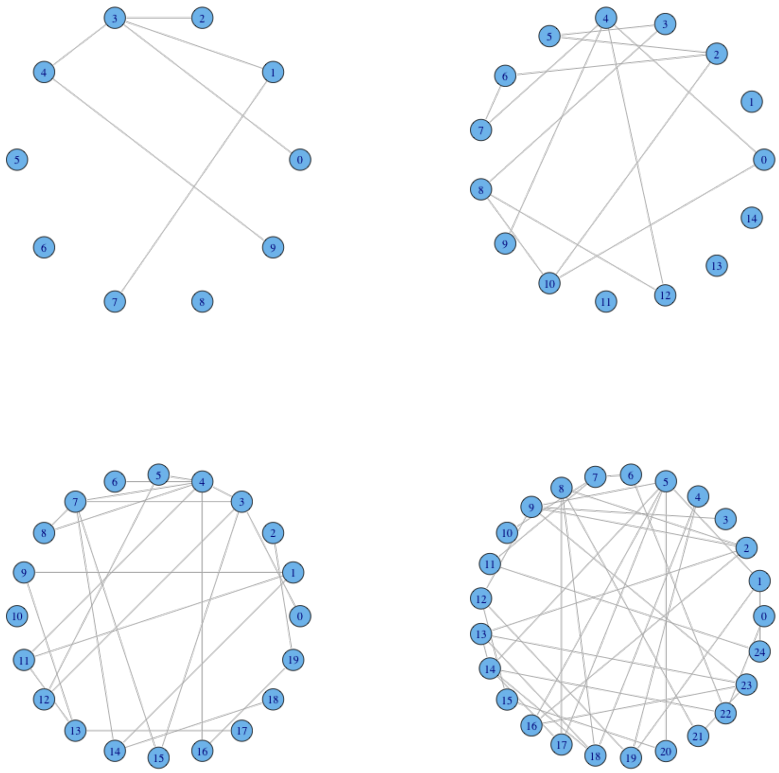


Figure 3.4: Graphs with 10, 15, 20 and 25 vertices generated through the ER model, but with 5 times less edges.

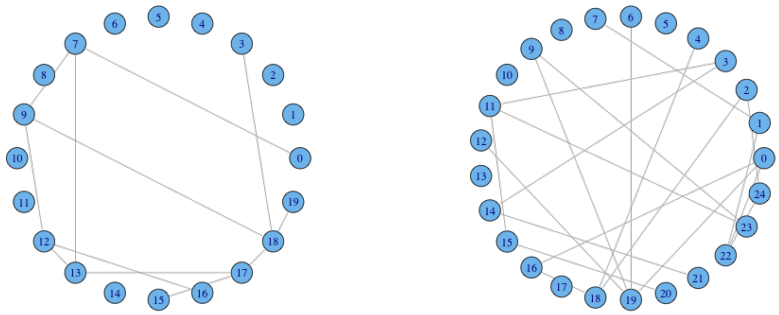


Figure 3.5: Graphs with 20 and 25 vertices generated through the ER model, but with 10 times less edges.

3.5 Diameter and the small world property

A small world network is a type of graph in which the majority of nodes have low degree, but most nodes can be reached from every other node in the graph by a small length path. More specifically, if D is the average length of the path between any two vertices of the graph (i.e., the diameter), then the graph is a small world network if

$$D \propto \log n,$$

where n is the number of vertices of the graph. Another property typical of these networks is a high clustering coefficient, that is, a tendency of the nodes of the network to cluster together and create groups. The nodes that are contained in these groups or "cliques" have a high density of links between them, so that most of the nodes of the group are directly connected.

The small world networks were introduced by Watts and Strogatz [16] to attempt to model the social networks and the Internet connectivity. Although it does account for many of the networks properties, it produced an unrealistic degree distribution that did not match the observed power law distribution that was later modeled with the BA model. Nevertheless, the small-world properties are still hugely relevant, since they are also prevalent in most social networks. Especially important is the small-world experiment carried by Milgram [21], as is one of the first to address the small-world properties of the social network.

Milgram's experiment attempted to measure the probability that two random people knew each other. In practice, however, this was implemented by measuring the average path length between any two people. The experiment's procedure is as follows:

1. We create a correspondence chain, with Nebraska and Kansas as the starting points, and Boston as the end point. (These were thought to represent a big distance in the U.S., both geographically and socially).

2. We send letters to "randomly selected people" from the starting points. The letter contains an explanation of the experiment and the basic information of a target person in Boston, as well as a paper to write their names in and business reply cards pre-dressed to Harvard.

3. If the person knew the target, they were to send the letter directly at them.

4. On the likely case that they did not know the target, they were to think about another person they knew that might know the target, and send them the letter, as well as a postcard to Harvard to keep track of the chain.

5. When and if the letter arrived at their destination, the length of the path could then be analyzed. Also, for the letters that never reached their destination, the tracking allowed to identify the break points.

As expected, reality (and humanity) were not keen on collaborating, and 232 of the 296 letters never reached their destination solely because the last receiver refused

to keep the chain going. From the remaining letters that did arrive, the estimated diameter was around 5.5.

Several problems and critiques arose despite the experiment's ingenuity. For instance, the reported path length was an average, not a maximum or a minimum. There is also no reason to believe that all the successful chains were in fact the shortest. Given that the participants had no access to the complete social network, they could be extending the path length by sending the letter away from the receiver. Some critics also pointed out the fact that some communities are isolated, and even when "discovered" their relationship with the outside world is almost nonexistent. These communities disrupt the global chains and are to be taken into account.

A better (and theoretical) approximation can be reached through the following theorem:

Theorem 3.7. *Let G be a connected graph, Δ the biggest degree of any vertex of G, and D the diameter of G. Then, if n is the number of vertices of G, we have:*

$$n \leq N(\Delta, D) := \frac{\Delta(\Delta-1)^D - 2}{\Delta - 2}$$

Proof. We pick a vertex v from the graph. By hypothesis, there are at most Δ vertices connected to v. For each of those vertices, there are at most $\Delta - 1$ new connected vertices, and therefore there are $\Delta(\Delta-1)$ new vertices. On the next step, at distance 2 from v, we find at most $\Delta(\Delta-1)^2$ new vertices. Since the diameter of G is D, at most we can iterate this process D times, at which point we will have all of the vertices of G. To summarize:

- Distance 0: We have 1 vertex, v.
- Distance 1: We have $1 + \Delta$ new vertices.
- Distance 2: We have $1 + \Delta + \Delta(\Delta-1)$ new vertices.
- Distance D: We have $1 + \Delta + \Delta(\Delta-1) + \cdots + \Delta(\Delta-1)^{D-1}$ new vertices.

After summing this amount, we get that, for distance D, the amount of vertices is:

$$1 + \Delta\frac{1-(\Delta-1)^D}{1-(\Delta-1)} = \frac{\Delta(\Delta-1)^D - 2}{\Delta - 2}$$

which proves the result. □

From this equation, we can deduce:

$$D \geq \frac{\log\left[\dfrac{n(\Delta-2)+2}{\Delta}\right]}{\log(\Delta-1)}.$$

If we take $n = 6 \cdot 10^9$ as the number of people on Earth, we can see that a value of only $\Delta = 50$ yields $D \geq 6$, somewhat consistent with the results from Milgram.

The small-world experiment inspired other analysis, focused on subsets of the world population graph. One case, which will be discussed in chapter 5, studies the citation network, i.e., the graph created by all the publications of scientific articles. In this network, two articles are connected if one of them cites the other. Within this network we can define the Erdös number [22] as the minimum distance between any article written by Erdös and the article of another author. Hence, Erdös is the only one with Erdös number 0, all of his collaborators have Erdös number 1, and so on.

A similar network is the one created by the actors and actresses of the film industry, where two actors are connected if they appear in the same movie. Here, the number we can define is the Bacon number, which measures the distance between the actor and Kevin Bacon.

A more recent experiment based on the email network, where two people are connected if one of them is in the contacts list of the other one [23]. Although it started as a scientific initiative, the idea behind the website inspired the new social media websites, such as Friendster, MySpace and, more recently, Facebook and Twitter.

Chapter 4

The Barabási-Albert model

One of the major flaws about the ER model was that, at the time, there was not enough data to check if the model could hold for larger N. As more and more data on the complex networks became available and better computers were developed, it became clear that the ER model was too simple and could not replicate the experimental results. The assumption that all edges were equally likely was not in agreement with most of the results from real networks. Further, growth was not contemplated in the ER model, which severely crippled its predictive power. A new model was then proposed in 1999 by Barabási-Albert [5, 6], (from now on BA model), that covered both these flaws:

- Growth: The ER model assumes that the number of vertices is constant, however most networks evolve with time and this must be taken into account to be more accurate. Therefore, in the BA model the number of vertices can increase and decrease during the network's lifetime.

- Preferential Attachment: The ER model assumes that the probability of linking is always the same, however real networks show signs of preferential attachment, that is, new vertices are more likely to link to a vertex with a large number of edges.

Both of these additions yield a scale-free network model (Definition 2.33), that is, $P(k)$ follows a power law distribution that is scale-free.

We shall explain in more detail an extended version of this model, described by Bollobás *et.al.* [8], which allows the existence of loops and multiple edges, and prove the result. The proof can be difficult to follow, and we add a second "heuristic" proof that yields the same result, but with a different approach.

It should be noted that, although the BA model succeeds in predicting the power law nature of the scale-free networks, it fails when predicting the value of γ for the data that was available. Models that consider additional properties, such as erasing edges or vertices, have yielded more accurate results.

In this chapter we shall describe the model, as well as prove that the degree sequence of the network follows a power law distribution.

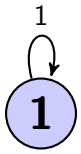
4.1 The model

Let $d_G(v)$ be the degree of vertex v in the graph G. We consider a fixed sequence of vertices $v_1, v_2, \ldots$ We shall inductively define a random graph process $(G_1^t)_{t\geq 0}$ so that G_1^t is a directed graph on $\{v_i : 1 \leq i \leq t\}$, as follows. We shall start with G_1^1, the graph with one vertex and one loop. Given G_1^{t-1}, form G_1^t by adding v_t along with an edge directed from v_t to v_i, such that v_i is chosen randomly following the probability

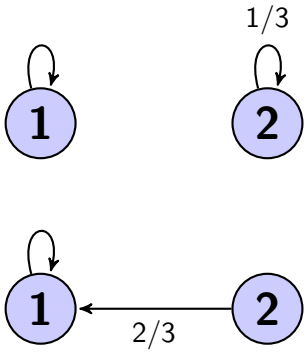
$$P(i=s) = \begin{cases} \dfrac{d_{G_1^{t-1}}(v_s)}{2t-1}, & 1 \leq s \leq t-1, \\ \\ \dfrac{1}{2t-1}, & s = t. \end{cases}$$

To simplify, when we add the vertex, we send an edge from this vertex to another existing vertex in the graph, where the probability of linking is proportional to its total degree at the time. This model can be easily expanded to adding m edges per step instead of one, but we take $m = 1$ for simplicity and because it will not affect our results.

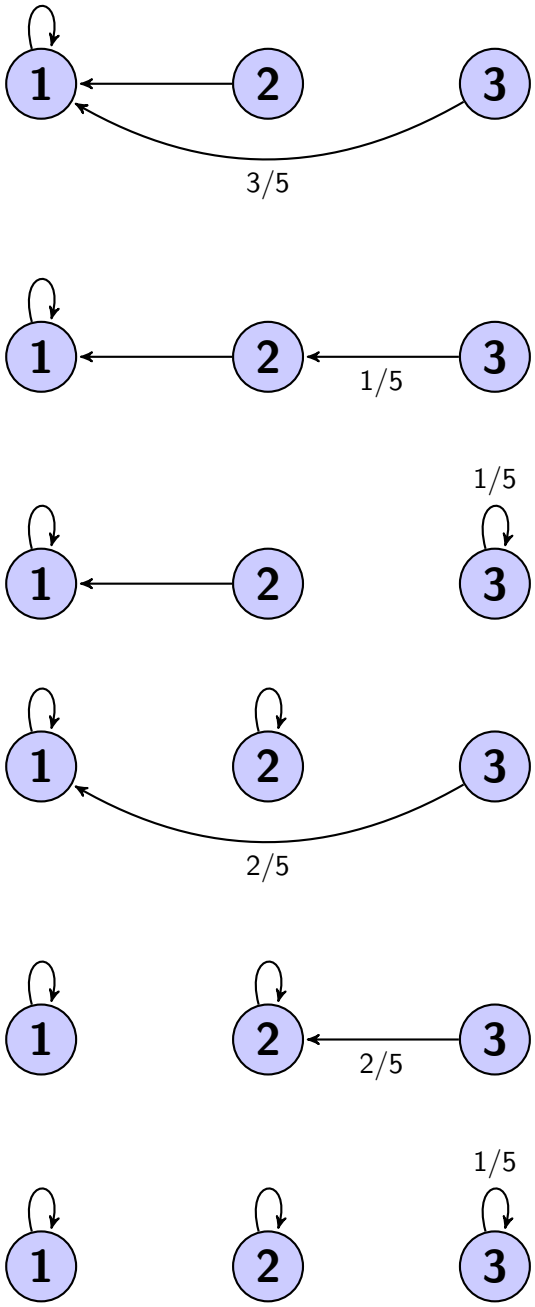
- We start with $t = 1$, a vertex with a loop. The number above the loop is the probability that this is the edge we add to the graph.



- For $t = 2$, we add vertex number 2 and we can add an edge from 1 to 2 or we can add a loop on vertex 2.



- For $t = 3$, we add vertex number 3 and we can add an edge from 3 to 1, from 3 to 2 or we can add a loop on vertex 3.



We should note that in the BA model the biggest connected component has little interest, since by construction most of the nodes are connected to each other and only a small portion of nodes can become isolated (those that add loops and are then never connected to any other node) or create smaller connected components. This last case is very unlikely, since the biggest connected component will most likely contain most of the higher connected vertices, and as such it acts as a black hole for new edges.

4.2 Degree sequence and power law distribution

We want to analytically calculate the probability $P(k)$, that is, the probability that a vertex has degree k. This will allow us to determine the exact value of the scaling exponent γ. For that purpose, we need the following lemma [4]:

Lemma 4.1. *(Azuma-Hoeffding inequality). Let $(X)_{t=0}^n$ be any martingale with $|X_{t+1} - X_t| \leq c$ for $t = 0, \ldots, n-1$. Then,*

$$P\left(|X_n - X_0| \geq x\right) \leq \exp\left(-\frac{x^2}{2c^2 n}\right).$$

This result will help us in proving the following:

Theorem 4.2. *Let $m \geq 1$ be fixed, and let $(G_m^n)_{n\geq 0}$ be the random graph process defined in the BA model. Let*

$$\alpha_{m,d} = \frac{2m(m+1)}{(d+m)(d+m+1)(d+m+2)},$$

and let $\varepsilon > 0$ be fixed. Let $\#_m^n(d)$ be the number of vertices with indegree d (i.e. total degree $m+d$). Then, with probability tending to 1 as $n \to \infty$, we have

$$(1-\varepsilon)\alpha_{m,d} \leq \frac{\#_m^n(d)}{n} \leq (1+\varepsilon)\alpha_{m,d},$$

for every d in the range $0 \leq d \leq n^{1/15}$.

Proof. As mentioned, the results for general m follow from those of $m = 1$. Let D_k be the sum of the first k degrees; we want to find its distribution and the distribution of the next degree, $d_{G_1^n}(v_{k+1})$, given D_k. We show that D_k is concentrated about a certain value and hence find approximately the probability that $d_{G_1^n}(v_{k+1}) = d$. Summing over k gives us the expectation of $\#_1^n(d)$, and concentration follows from Lemma 4.1. We must first note that the expectations of the distributions of the total degrees are easy to calculate:

$$\mathbf{E}\left(d_{G_1^t}(v_t)\right) = 1 + \frac{1}{2t-1}.$$

Also, for $s < t$,

$$\mathbf{E}\left(d_{G_1^t}(v_s) | d_{G_1^{t-1}}(v_s)\right) = d_{G_1^{t-1}}(v_s) + \frac{d_{G_1^{t-1}}(v_s)}{2t-1},$$

which yields the iterative formula

$$\mathbf{E}\left(d_{G_1^t}(v_s)\right) = \frac{2t}{2t-1}\,\mathbf{E}(d_{G_1^{t-1}}(v_s)).$$

Thus, for $1 \leq s \leq n$,

$$\mathbf{E}\left(d_{G_1^n}(v_s)\right) = \sum_{i=s}^{n} \frac{2i}{2i-1} = \frac{4^{n-s+1} n!^2 (2s-2)!}{(2n)!(s-1)!^2} = \sqrt{n/s}(1+O(1/s)),$$

using Stirling's formula (Theorem 2.41):

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \left(1 + O\left(\frac{1}{n}\right)\right).$$

If every degree was equal to its expectation we would end here. Unfortunately, degrees can be far from their expectations. Let us write d_i for $d_{G_1^n}(v_i)$. We aim to describe the distributions of the individual d_i. Let us consider their sums, $D_k = \sum_{i=1}^{k} d_i$.

Consider the event $\{D_k - 2k = s\}$, $0 \leq s \leq n-k$. This is the event that the last $n-k$ vertices of G_1^n send exactly s edges to the first k vertices. This corresponds to pairings $\mathcal{P}$ in which the kth right endpoint is $2k+s$. Consider any pairing $\mathcal{P}$ with this property. We shall split $\mathcal{P}$ into two partial pairings, the *left partial pairing* $\mathcal{L}$ and the *right partial pairing* $\mathcal{R}$, each consisting of some number of pairs together with some unpaired elements. For $\mathcal{L}$ we take the partial pairing on $\{1, \ldots, 2k+s\}$, induced by $\mathcal{P}$, for $\mathcal{R}$ that on $\{2k+s+1, \ldots, 2n\}$. From the restriction on $\mathcal{P}$, in $\mathcal{L}$ the element $2k+s$ must be paired with one of $\{1, \ldots, 2k+s-1\}$, precisely s of the remaining $2k+s-2$ elements must be unpaired, and the other $2(k-1)$ elements must be paired off somehow. Any of the

$$(2k+s-1)\binom{2k+s-2}{s}\frac{(2k-2)!}{2^{k-1}(k-1)!}$$

partial pairings obtained in this way may arise as $\mathcal{L}$. Similarly, for $\mathcal{R}$ there are

$$\binom{2n-2k-s}{s}\frac{(2n-2k-2s)!}{2^{n-k-s}(n-k-s)!}$$

possibilities. Any possible $\mathcal{L}$ may be combined with any possible $\mathcal{R}$ to form $\mathcal{P}$ by pairing off the unpaired elements of $\mathcal{L}$ with those of $\mathcal{R}$ in any of $s!$ ways. Multiplying together and dividing by the total number $(2n)!/(2^n n!)$ of n-pairings we can see that for $1 \leq k \leq n$ and $0 \leq s \leq n-k$,

$$\mathbf{P}\left(D_k - 2k = s\right) = \frac{(2k+s-1)!(2n-2k-s)!n!2^{s+1}}{s!(k-1)!(n-k-s)!(2n)!}.$$

We can deduce a concentration result for D_k. For k with $1 \leq k \leq n$ let us write $p_s = p_{s,k}$ for the probability defined above, and let

$$r_s = \frac{p_{s+1}}{p_s} = 2\frac{(2k+s)(n-k-s)}{(s+1)(2n-2k-s)}.$$

Note that r_s is a decreasing function of s. Allowing s to be a real number, the unique positive solution to $r_s = 1$ is given by

$$s = -2k + \sqrt{4kn - 2n + \frac{1}{4}} + \frac{1}{2}.$$

Thus, $s_0 = \lceil -2k + \sqrt{4kn - 2n + \frac{1}{4}} + \frac{1}{2} \rceil$ is one of the at most two most likely values of $D_k - 2k$. Also, for n larger than some constant we have

$$\begin{aligned}\frac{r_{s+1}}{r_s} &= \left(1 - \frac{2k-1}{(s+2)(2k+s)}\right)\left(1 - \frac{n-k}{(2n-2k-s-1)(n-k-s)}\right) \\ &\leq \left(1 - \frac{2k-1}{2n^2}\right)\left(1 - \frac{n-k}{2n^2}\right) \leq \exp\left(-\frac{2k-1}{2n^2}\right)\exp\left(-\frac{n-k}{2n^2}\right) \\ &\leq \exp\left(-\frac{1}{2n}\right).\end{aligned}$$

As $r_{s_0} \leq 1$ it follows that $r_{s_0+x} \leq \exp(-x/(2n))$ for $x > 0$ and therefore $p_{s_0+x} \leq \exp(-x(x-1)/(4n))$. A similar bound on p_{s_0-x} shows that

$$\mathbf{P}\left(|D_k - (2k + s_0)| \geq 3\sqrt{n \log n}\right) = o(n^{-1}).$$

In fact, since $|s_0 - (2\sqrt{kn} - 2k)| \leq 2\sqrt{n}$ for each k, we obtain

$$\mathbf{P}\left(|D_k - 2\sqrt{kn}| \geq 4\sqrt{n \log n}\right) = o(n^{-1}). \tag{4.1}$$

We now take a look to the probability that $d_{k+1} = d + 1$, i.e., that the indegree of v_{k+1} is d, given D_k. Suppose that $1 \leq k \leq n-1$ and $0 \leq s \leq n-k$, and consider a left partial pairing $\mathcal{L}$ as above. We have already seen that each such $\mathcal{L}$ has

$$s!\binom{2n-2k-s}{s}(2n-2k-2s-1)!!$$

extensions to an n-pairing. Such an extension corresponds to a graph with $d_{k+1} = d+1$ if and only if $2k+s+d+1$ is a right endpoint, and each of $2k+s+1, \ldots, 2k+s+d$ is a left endpoint. We note here that the element paired with $2k+s+d+1$ must be either of the unpaired elements in $\mathcal{L}$ or one of the $2k+s+1, \ldots, 2k+s+d$, and that $s-1+d$ pairs start before $2k+s+d+1$ and end after this point, each $\mathcal{L}$ has exactly

$$(s+d)(s+d-1)\binom{2n-2k-s-d-1}{s+d-1}(2n-2k-2s-2d-1)!!$$

such extensions, and for $0 \leq d \leq n-k-s$, we have

$$\mathbf{P}\left(d_{k+1} = d+1 \mid D_k - 2k = s\right) = (s+d)2^d \frac{(n-k-s)_d}{(2n-2k-s)_{d+1}}, \tag{4.2}$$

where we write $(a)_b$ for $\frac{a!}{(a-b)!}$. This also applies when $k = s = 0$, yielding

$$\mathbf{P}(d_1 = d+1) = \frac{d2^d(n)_d}{(2n)_{d+1}}.$$

For $k \geq 1$, we shall use (4.1) and (4.2) to estimate the expectation of $\#_1^n(d)$, the number of vertices of G_1^n with indegree d. Let $M = \lfloor n^{4/5}/\log n \rfloor$, let $k = k(n)$ be any function satisfying $M \leq k \leq n - M$, and let $d = d(n)$ be any function satisfying $0 \leq d \leq n^{1/15}$. For any D with $|D - 2\sqrt{kn}| \leq 4\sqrt{n \log n}$ we can use (4.2) to write $\mathbf{P}(d_{k+1} = d+1 | D_k = D)$ as

$$(2\sqrt{kn} - 2k + O(\sqrt{n\log n}))2^d \frac{(n + k - 2\sqrt{kn} + O(\sqrt{n\log n}))^d}{(2n - 2\sqrt{kn} + O(\sqrt{n\log n}))^{d+1}}$$

Using the bounds on d and k, we find that the ratio of $n + k - 2\sqrt{kn} = (\sqrt{n} - \sqrt{k})^2$ to $d\sqrt{n\log n}$ tends to infinity as $n \to \infty$, as does $(2n - 2\sqrt{kn})/(d\sqrt{n\log n})$. Also, $\sqrt{n\log n} = o(2\sqrt{kn} - 2k)$, so the probability above is equal to

$$(1 + o(1))\frac{2\sqrt{kn} - 2k}{2n - 2\sqrt{kn}} \left(\frac{2(\sqrt{n} - \sqrt{k})^2}{2(n - \sqrt{kn})} \right)^d \sim \sqrt{\kappa}(1 - \sqrt{\kappa})^d,$$

where $\kappa = k/n$. As this estimate applies uniformly to $\mathbf{P}\left(d_{k+1} = d+1 \mid D_k = D\right)$ for all D with $|D - 2\sqrt{kn}| \leq 4\sqrt{n\log n}$, we see from (4.1) that

$$\mathbf{P}(d_{k+1} = d+1) = o(n^{-1}) + (1 + o(1))\sqrt{\kappa}(1 - \sqrt{\kappa})^d.$$

Keeping n and d fixed and varying k in the range $M \leq k \leq n - M$, as the estimate above is uniform in k we find that the expected number of vertices v_{k+1}, $M \leq k \leq n - M$, with degree equal to $d + 1$ can be written as

$$o(1) + \sum_{k=M}^{n-M} (1 + o(1))\sqrt{\frac{k}{n}} \left(1 - \sqrt{\frac{k}{n}}\right)^d.$$

As all terms in the sum are positive, it follows that

$$\mathbf{E}(\#_1^n(d)) = O(M) + o(1) + (1 + o(1)) \sum_{k=M}^{n-M} \sqrt{\frac{k}{n}} \left(1 - \sqrt{\frac{k}{n}}\right)^d. \tag{4.3}$$

Writing $f = \sqrt{\kappa}(1 - \sqrt{\kappa})^d$, we have

$$\frac{1}{f}\frac{\mathrm{d}f}{\mathrm{d}\kappa} = \frac{\kappa^{-1}}{2} - \frac{d}{2}\frac{\kappa^{-1/2}}{(1 - \kappa^{1/2})}.$$

Provided $n\kappa$ and $n(1-\kappa)$ tend to infinity, the proportional change in f as κ changes by $1/n$ is thus $o(1)$ uniformly in κ. It follows that the sum in (4.3) can be written as

$$(1 + o(1))n \int_{(M+1)/n}^{1-M/n} \sqrt{\kappa}(1 - \sqrt{\kappa})^d \mathrm{d}\kappa \sim n \int_0^1 \sqrt{\kappa}(1 - \sqrt{\kappa})^d \mathrm{d}\kappa.$$

Substituting $\kappa = (1-u)^2$ we obtain

$$\mathbf{E}(\#_1^n(d)) = O(M) + (1+o(1))\frac{4n}{(d+1)(d+2)(d+3)} \sim \frac{4n}{(d+1)(d+2)(d+3)},$$

which is the required form of the distribution.

Let us return to the general case $m \geq 1$. Suppose that m is a fixed constant and let d'_k be the degree of v_k in the graph G_m^n. We shall estimate $\mathbf{P}(d'_{k+1} = d+m)$, keeping the notation d_K for degrees in the graph G_1^{nm} from which G_m^n is obtained. We look at the distributions of $d_{K+1}, \ldots, d_{K+m}$ in G_1^N, where $K = mk$ and $N = mn$. The argument giving the conditional probability estimate (4.2) also applies to the conditional probability given the entire sequence of earlier degrees. For $M \leq k \leq n - M$ and $d \leq n^{1/15}$ our earlier estimates show that, provided no $|D_{K'} - 2\sqrt{K'N}|$ is too large,

$$\mathbf{P}(d_{K+j+1} = d+1 \mid d_1, d_2, \ldots, d_{K+j}) \sim \sqrt{\frac{K+j}{N}}\left(1 - \sqrt{\frac{K+j}{N}}\right)^d$$

which can be written as

$$\mathbf{P}(d_{K+j+1} = d+1 \mid d_1, d_2, \ldots, d_{K+j}) \sim \sqrt{\kappa}(1-\sqrt{\kappa})^d,$$

when $N \gg j$, with $\kappa = k/n = K/N$. Thus, using (4.1),

$$\begin{aligned}\mathbf{P}(d'_{k+1} = d+m) &= o(n^{-1}) + (1+o(1)) \sum_{a_1+\cdots+a_m=d} \prod_{j=1}^{m} \sqrt{\kappa}(1-\sqrt{\kappa}^{a_j}) \\ &= o(n^{-1}) + (1+o(1))\binom{d+m-1}{m-1}\kappa^{m/2}(1-\sqrt{\kappa})^d.\end{aligned}$$

Proceeding as before we can express the expectation of the number $\#_m^n(d)$ of vertices of G_m^n with indegree d in terms of

$$\int_0^1 \kappa^{m/2}(1-\sqrt{\kappa})^d \mathrm{d}\kappa = 2\int_0^1 (1-u)^{m+1}u^d du = 2\frac{(m+1)!d!}{(d+m+2)!},$$

where we have again substituted $\kappa = (1-u)^2$. We find that, for $0 \leq d \leq n^{1/15}$,

$$\mathbf{E}(\#_m^n(d)) \sim \frac{2m(m+1)n}{(d+m)(d+m+1)(d+m+2)} \tag{4.4}$$

uniformly in d. Let us consider again the graph G_m^n as one from the process $(G_m^t)_{t\geq 0}$. We fix $m \geq 1$, $n \geq 1$ and $0 \leq d \leq n^{1/15}$, and consider the martingale

$$X_t = \mathbf{E}(\#_m^n(d) \mid G_m^t),$$

for $0 \leq t \leq n$. We have $X_n = \#_m^n(d)$, while $X_0 = \mathbf{E}(\#_m^n(d))$. We claim that the differences $|X_{t+1} - X_t|$ are bounded by 2. To see this, note that whether at stage t we

join v_t to v_i or v_j does not affect the degrees at later times of vertices v_k, $k \notin \{i, j\}$. More precisely, the joint distribution of all other degrees is the same in either case. Since we are just counting vertices with a particular degree, no matter how much the degrees of v_i and v_j are changed in G_m^n, this changes $\#_m^n(d)$ by at most 2.

Applying Lemma 4.1, we find that for each d with $0 \leq d \leq n^{1/15}$ we have

$$\mathbf{P}\left(|\#_m^n(d) - \mathbf{E}(\#_m^n(d))| \geq \sqrt{n \log n}\right) \leq e^{-\log n/8} = o(n^{-1/15}).$$

Noting from (4.4) that, in this range,

$$\mathbf{E}\left(\#_m^n(d)\right) \sim \frac{2m(m+1)n}{(d+m)(d+m+1)(d+m+2)},$$

and that this is much larger than $\sqrt{n \log n}$, the result follows. □

A more simplistic proof [6] can be obtained by analyzing the time dependence of the connectivity of a given vertex. This can be calculated analytically using a mean-field approach.

We define m_0 as the initial ammount of vertices of the graph. At each step, we add a vertex and $m(\leq m_0)$ edges that don't have to originate from the added vertices. After t steps, we have $N = m_0 + t$ vertices and mt edges.

We want to calculate the probabilty that a vertex has k edges, $P(k)$. Let us assume that k is continuous. Then, the probability that a new vertex will be connected to vertex i at any given step,

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}$$

with k_i being the degree of vertex i, can be interpreted as a continuous rate of change of k_i. Hence, for a vertex i, we can write

$$\frac{\partial k_i}{\partial t} = A\Pi(k_i) = A\frac{k_i}{\sum_{j=1}^{m_0+t-1} k_j}. \tag{4.5}$$

Taking into account that $\sum_j k_j = 2mt$ and the change in connectivities at a timestep is $\Delta k = m$, we obtain that $A = m$, which leads to

$$\frac{\partial k_i}{\partial t} = \frac{k_i}{2t}.$$

The solution of this equation, with the initial condition that vertex i was added to the system at time t_i with connectivity $k_i(t_i) = m$ is

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{0.5}.$$

We can observe that the last equation essentially means that there is a "rich-gets-richer" kind of phenomenon, where older vertices (i.e., those with small t_i) increase

their degree at the expense of the younger vertices (i.e., those with large t_i). This property can be used to calculate γ analytically. This in turn means that the probability that a vertex has connectivity $k_i(t)$ smaller than k, $P(k_i(t) < k)$ can be written as

$$P(k_i(t) < k) = P\left(t_i > \frac{m^2 t}{k^2}\right).$$

Assuming that we add the vertices at equal time intervals to the system, the probability density of t_i is

$$P_i(t_i) = \frac{1}{m_0 + t}.$$

Substituting this into equation (4.5) we obtain that

$$P\left(t_i > \frac{m^2 t}{k^2}\right) = 1 - P\left(t_i \leq \frac{m^2 t}{k^2}\right) = 1 + \frac{m^2 t}{k^2(t + m_0)}.$$

Finally, we can obtain the probability density for $P(k)$ by using

$$P(k) = \frac{\partial P(k_i(t) < k))}{\partial k} = \frac{1}{k^3}\frac{2m^2 t}{m_0 + t}, \tag{4.6}$$

predicting $\gamma = 3$, independent of m and m_0. This means that no matter how many vertices we start with or how many edges we add at every step, we eventually get $P(k) \sim Ak^{-3}$. Furthermore, equation (4.6) also predicts that the coefficient A is proportional to the square average connectivity of the network (the average degree of the vertices in the network)

$$A \propto m^2.$$

We note that this last model was the original one proposed by Barabási-Albert in 1999. We can observe, however, that starting with no edges essentially yields the probability to add any vertex equal to 0, which is why Bollobás *et.al.* expanded it by starting with one vertice and a loop. As we have already seen, the value of γ is not dependant on the initial number of vertices m_0 or the amount of edges added per step m. Hence, it is easier to use the case where $m_0 = m = 1$, which is the way we originally described the model at the beggining of the chapter, and for which we have proven Theorem 4.2.

4.3 Computational model

We will now simulate the theoretical model, adding only one edge per step. The code for the BA model is the following, using R as the programming language:

```
require(igraph)
ba_vertices<-10 #We can put any positive integer here
ba_matrix<-integer(ba_vertices*ba_vertices)
dim(ba_matrix)<-c(ba_vertices,ba_vertices)
```

```
ba_matrix[1,1]<-1
ba_degree<-integer(ba_vertices)
ba_degree[1]<-2
for(j in 2:ba_vertices){
          cont<-cont+1
          aux<-j
          ba_rand<-sample(1:(2*j-1),1)
          for(i in 1:aux){
                    if(ba_rand-ba_degree[i]<=0){
                              ba_matrix[i,aux]<-1
                              ba_degree[i]<-ba_degree[i]+1
                              ba_degree[aux]<-ba_degree[aux]+1
                              ba_rand<-0
                              break
                    }else{
                              ba_rand<-ba_rand-ba_degree[i]
                    }
          }
          if(ba_rand==1){
                    ba_matrix[j,j]<-1
                    ba_degree[j]<-ba_degree[j]+2
          }
}
ba_dseq<-integer(ba_vertices+1)
for(k in 1:(ba_vertices+1)){
          for(i in 1:ba_vertices){
                    if(ba_degree[i]==k){
                              ba_dseq[k]<-ba_dseq[k]+1
                    }
          }
}
ba_dseq<-ba_dseq/ba_vertices
ba_k<-1:(ba_vertices+1)
ba_p<-ba_degree[1]*ba_k^(-3)
matplot(ba_k,cbind(ba_dseq,ba_p),type="l",col=c("red","green"),
lty=c(1,1),log="xy",
xlab=expression(italic(k)),ylab=expression(italic(P(k))),main="BA Model"))
dev.copy(png,paste('BA_dist',ba_vertices,'.png')
dev.off()
ba_graph<-graph.adjacency(ba_matrix)
plot.igraph(ba_graph,layout=layout.circle)
dev.copy(png,paste('BA_graph',ba_vertices,'.png'))
dev.off()
```

4.3.1 Degree distribution

We compare the degree distribution with the expected power law

$$P(k) = Ak^{-3}$$

where A is a correction value that added to make the lines appear closer in the graphic. Figure 4.1 shows some of the results (red is our model, green is the associated power law distribution).

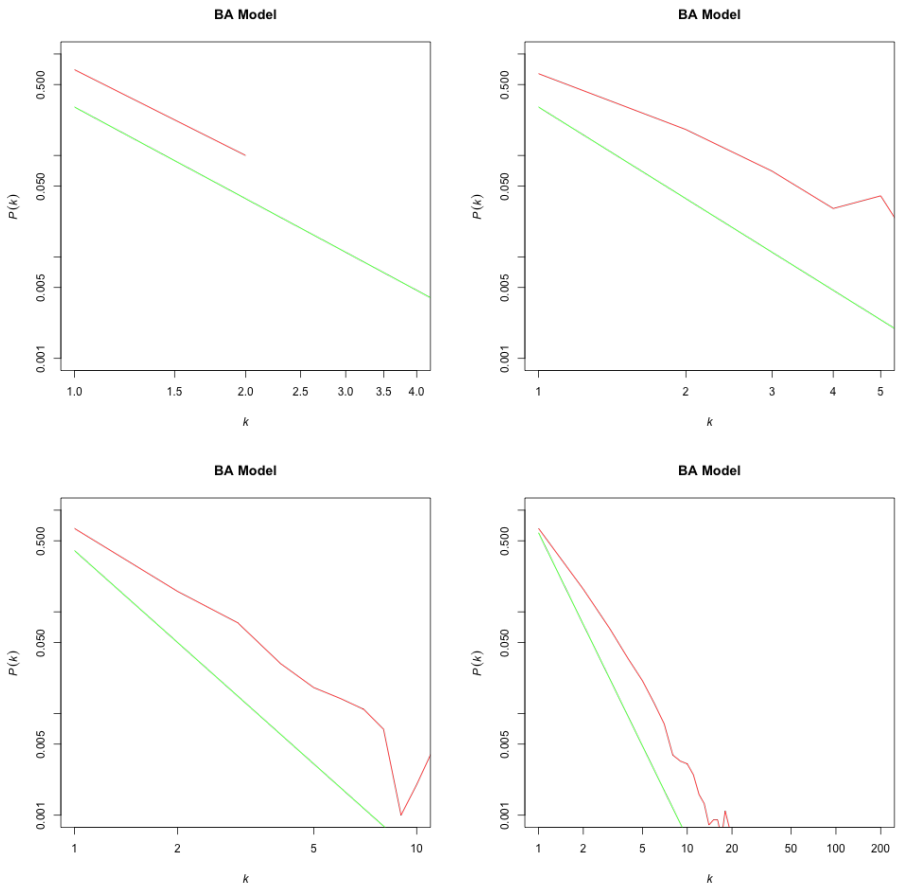


Figure 4.1: Degree distribution for graphs with 10, 100, 1000 and 10000 vertices constructed via the BA model. The graphs are centered on the low values of k and have been amplified to observe the correlation between the two lines.

At low vertices, the two lines are similar, although this can vary significantly when making new simulations. For illustration purposes, we have added the better ones. The model gets more consistent as we increase the number of vertices up to 10000, and all simulations are essentially the same, with few noticable variations. The slopes of the lines with a high number of vertices are still not exactly as predicted. This can be explained because the number of vertices that was considered should go much higher, around $8 \cdot 10^5$, at which point we expect better results, as we

can observe in [6]. Here, we have the same problem as before, where our computer cannot go further, although these simulations show that the trend is pretty accurate.

4.3.2 Graph plot

In contrast with the ER model, the BA model allows for good plotting of graphs up to 35 vertices. Past that point, the clustering of the nodes makes the graph difficult to read. We should especially note that, as expected, the number of loops is very low compared to the amount of steps that we are simulating, emphasizing the fact that their effect is negligible when analyzing the graph. Figure 4.2 shows some examples.

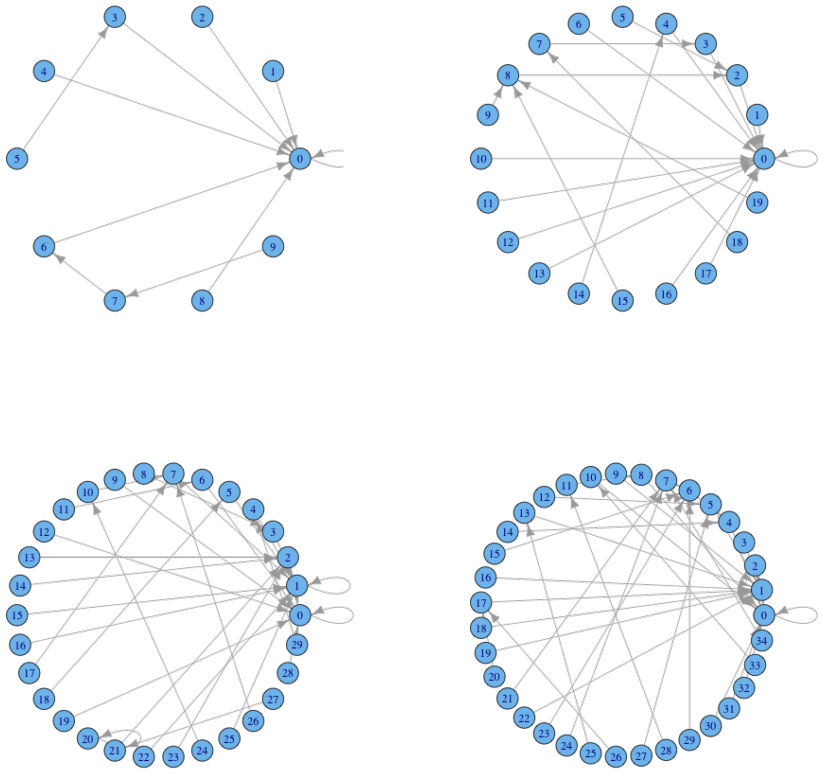


Figure 4.2: Graphs with 10, 20, 30 and 35 vertices generated through the BA model.

Chapter 5

Applications

It is now time to check how our models hold when faced with real networks. There is a massive number of situations that can be studied via simulations, so we shall try to analyze a diverse sample to emphasize the flexibility of random graph theory. In each case we will identify the elements that will act as vertices and as edges, why is the problem important and what kind of advantages we can gain by using random networks to study that particular situation.

We should, however, approach these examples with caution. We have selected them for its historical significance, and some of them are outdated or their data is no longer available, so that we cannot access it to use for our simulations. Moreover, we have focused our attention on proving the theoretical models and getting a good computational simulation to check them, so we will not go too in-depth when analyzing them. These examples have been selected to show the potential of the theory, and that random graph theory can be applied in a lot of diverse cases.

Except for the first example, whose data is still available, the rest of the examples' figures are created just to illustrate and complement the text. This means that we "trust" the results of the original authors of the different cases, and explain them accordingly. Other examples also show that, although the models we have proved in previous chapters give a good approximation of the real networks, they fail when approaching dynamic networks. We will extend on this concept at the end of the chapter.

5.1 The paper citation distribution

The amount of citations that a paper has (i.e., the number of times that an article is cited in a paper) is usually assumed to represent the "influence" of the paper. It is also an important factor when considering merits of the author, both in regard to the academic career of the author and in merit-based considerations, such as grants and partnerships. Despite the importance of the citation network, little to no research has been conducted on quantifying scientific research, its correlation with productivity and the impact of some papers in the network. Further, most of the studies conducted about the subject are either based on heuristic arguments and give no numerical values, or use an insufficient amount of data to extract significant

results.

We will focus on the distribution of citations, i.e. the number of papers with x citations, $N(x)$. In [15], this model is analyzed using two databases: papers published in 1981 in journals catalogued by the Institute of Scientific Information (ISI) (+780,000 papers) [19] and papers published during 20 years of publications in Physical Review D (PRD), volumes 11–50 (+24,000 papers) [20]. The elements of the network are as follows:

- Vertices: The articles or publications.
- Edges: The citations: if an article A is cited in another article B, the edge goes from B to A. This graph is directed.

Analysis of the citation case. The fundamental distribution that we analyze here is the number of articles that have been cited a total of x times, $N(x)$. It should be noted that the data only matches a power law distribution for big values of x (as should be expected, since we have only proved the results asymptotically). Figure 5.1 represents the ISI data (yellow), the PRD data (red), together with a line of slope -3 (green) and slope -2.8 (blue).

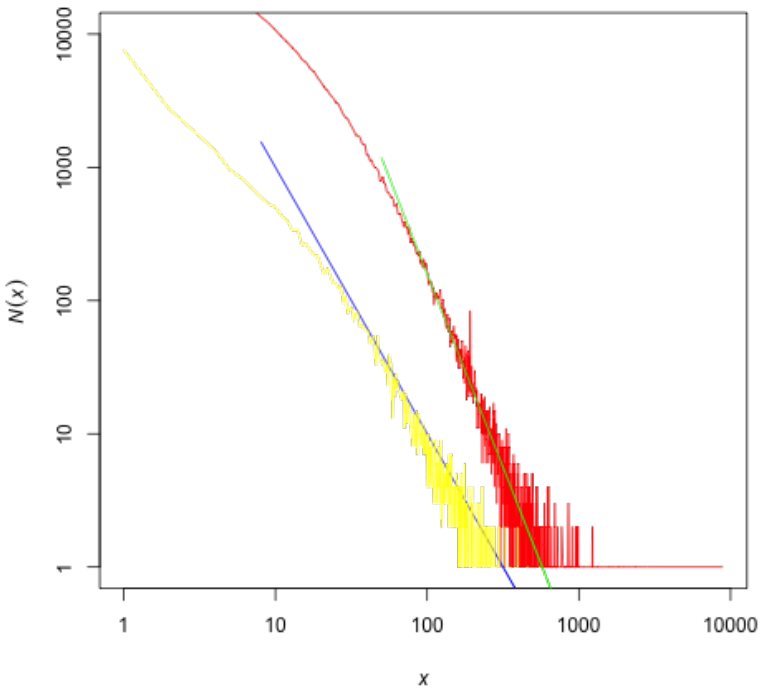


Figure 5.1: The number of articles cited x times, for the ISI and PRD database. Two lines of slope -3 and -2.8 are added as guideline.

We can therefore assume that $N(x)$ follows a power law distribution, $N(x) \propto x^{-\alpha}$. Finally, let M be an ensemble of M publications and the corresponding number of citations for each of these papers in rank order $Y_1 \geq Y_2 \geq ... \geq Y_M$, then we define the number of citations of the *k-th* most cited paper, Y_k, as

$$\int_{Y_k}^{\infty} N(x)dx = k. \tag{5.1}$$

In other words, there are k articles out of M that are cited *at least* Y_k times in the database. Figure 5.2 displays Y_k vs k for the PRD set.

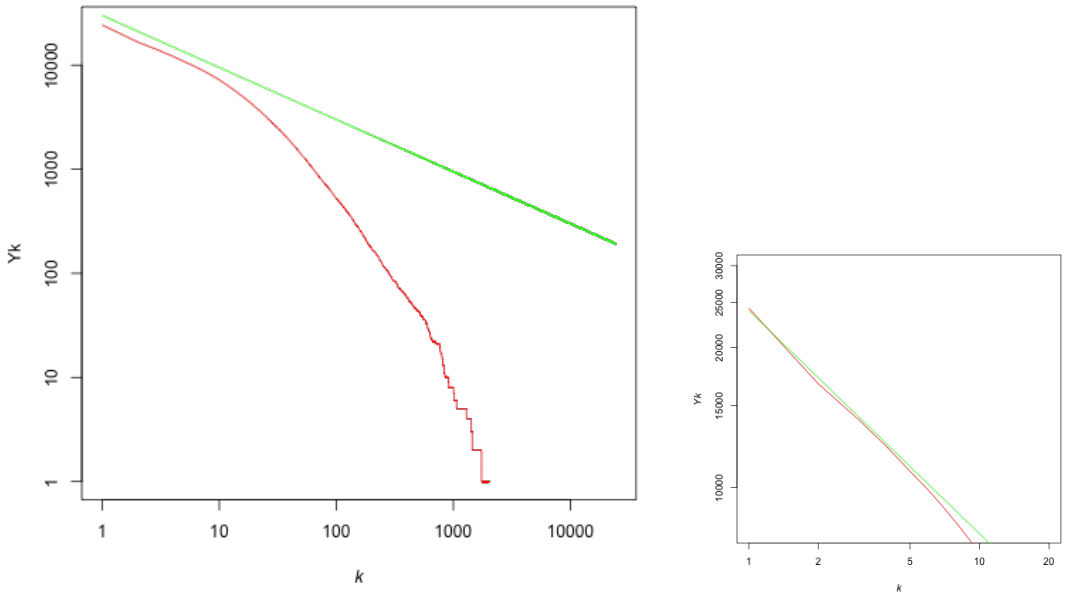


Figure 5.2: The left graphic shows Y_k vs k in red, with a line of slope -0.48 in green. The right graphic shows a closer look for $Y_k > 8000$.

From this representation, we can see that the highly-ranked citations provide a good representation of the asymptotic tail of the citation distribution (those with low k). This data is pretty linear, and a least square data fit yields an exponent of about -0.48. Let us now assume without a loss of generality that $N(x) = x^{-\alpha}$. Then, by (5.1), we have that

$$\int_{Y_k}^{\infty} x^{-\alpha}dx = \left[\frac{x^{1-\alpha}}{1-\alpha}\right]_{Y_k}^{\infty} = k.$$

Now, we assume $\alpha > 1$ and we have that

$$-Y_k^{1-\alpha} = k(1-\alpha)$$

which finally yields

$$Y_k = [(\alpha - 1)k]^{\frac{1}{1-\alpha}}.$$

Since the exponent was approximately -0.48, we have that

$$\frac{1}{1-\alpha} = -0.48 \Longrightarrow \alpha = 1 + \frac{1}{0.48} \approx 3.08,$$

which is in agreement with the theory.

We should note that, as the author points out, this data is somewhat inaccurate (a simple review of the data shows that the total number of citations to individual articles is 524 when there are only 426 articles in the dataset. Even with the disparity, the smoothness of the distribution suggests that these have little influence in the final results).

5.2 The World Wide Web

It is indisputable that the Internet has changed the world. The world wide web plays a fundamental role in a lot of aspects of our daily life and has become the mainstream medium to exchange information worldwide. Despite its massive importance, the huge complexity of the network and the fast and unregulated growth it presents makes the study of the web a rather unappealing and intimidating effort. It is easy to imagine that we could model the web as a graph:

- Vertices: The HTML documents or "the web page".
- Edges: URLs that point to one document. This graph is therefore directed.

The topology of this graph would yield impressive results, such as the connectivity of the web, and consequently how efficiently we can locate information on it. A complete mapping of the web would allow an omniscient agent to interpret all the links and choose the optimal path to navigate the web, thus becoming the ultimate search engine. Although this is unachievable nowadays, the results would nevertheless significantly improve the power of the robots, since they rely on a matching string method that can prove limited in numerous occasions.

Unfortunately, the size of the Internet has been estimated to be no less than $8 \cdot 10^8$ documents [1], with more modern calculations setting the number around $4 \cdot 10^{12}$. These numbers must be handled with caution, since the algorithms used to calculate them are often not that reliable and tend to overestimate. Also, indexing the web is not an easy task by any means, since the uncontrolled growth of the web makes it a very dynamic network, with vertices and edges appearing and disappearing constantly. Thus, local analysis is not optimal, and instead a large-scale approach is more suitable for a first glance at the infrastructure of the Internet, and therefore we go back to obtaining the topology of the network. A lot of effort is being put in an indexation of the web, with various projects and databases attempting to provide a good approximation, most of them englobed in CAIDA. The indexation of the web is analyzed in a P2P manner, with Skitter [18] and its successor, Ark [17], being the main projects.

Analysis of the WWW case. The analysis will consist on 2 parts.

Distribution. We can use the data to determine the probabilities $P_{out}(k)$ and $P_{in}(k)$ that a document has k outgoing and incoming links [7]. Both follow a power law over several orders of magnitude, as shown in Figure 5.3.

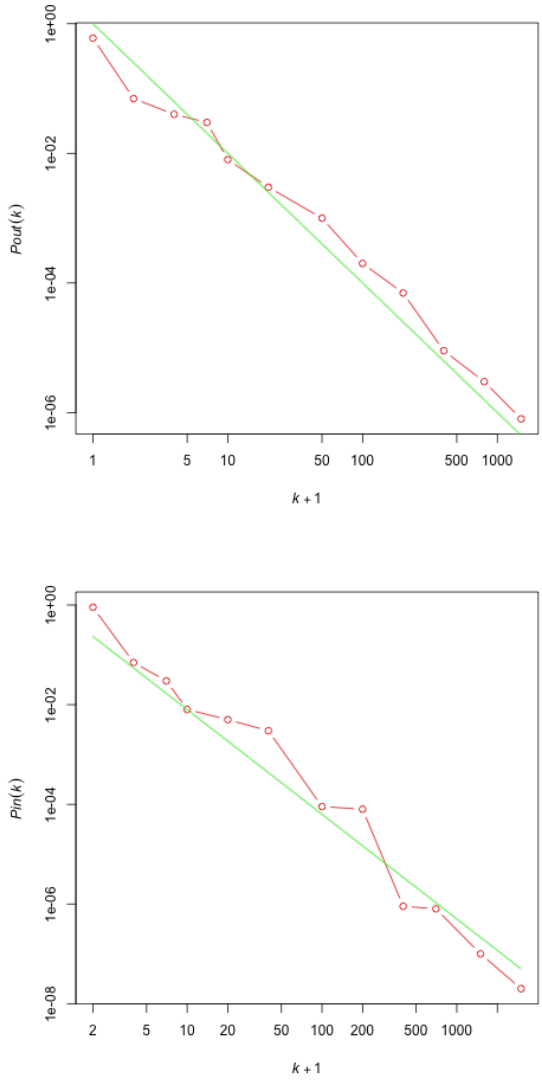


Figure 5.3: Distribution of the links on the World Wide Web, outgoing links and incoming links. The data is in red, the power law distribution is in green. The tail of the distributions follow a power law with $\gamma_{out} = 2.45$ and $\gamma_{in} = 2.1$.

The model predicts $\gamma = 3$, while the obtained results yield $\gamma_{out} = 2.45$ and $\gamma_{in} = 2.1$, significantly different values. We will discuss the reasons behind this discrepancies at the end of the study.

Diameter. A particularly important quantity in a search process is the shortest path between two documents, D, defined as the smallest number of URL links that must be followed to navigate from one document to the other [7]. As Figure 5.4 shows, we find that the average of D over all pairs of vertices follows

$$\langle D \rangle = 0.35 + 2.06 \log(N),$$

indicating that the web forms a small-world network. (See Section 5 of Chapter 3).

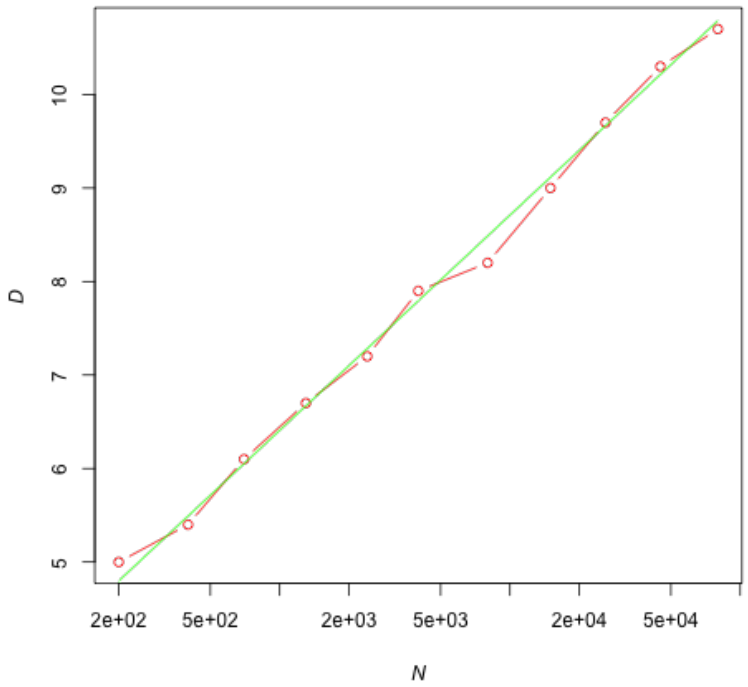


Figure 5.4: Average of the shortest path between two documents vs the size (N) of the system. The data is in red, and we add the line $\langle D \rangle = 0.35 + 2.06 \log(N)$ in green.

For $N = 8 \cdot 10^8$, this yields $\langle D_{www} \rangle = 18.59$. This means that, on average, it takes 19 clicks to go from one random document to another in the web. Even more important than that is the logarithmic dependence of $\langle D \rangle$ with N, since the enormous predicted growth of the Internet will barely affect $\langle D \rangle$. To illustrate, a growth of 1000% over a few years would yield a $\langle D_{www} \rangle = 21$.

This result, however, provides no real advantage for matching strings robots. Assuming that such a robot could locate a document at distance $\langle D \rangle$, it would need to search around 10% of the whole web to do so. This unbearable costly, and as such new search engines are to be developed to take advantage of this high connectivity of the web.

It should be noted however that the web is a much more complex network than this. The BA model, although it gives a good first glance at the infrastructure, makes huge omissions on its premises that, when applied to the web, cannot faithfully describe the topology of the web. As we have already commented, the edges (links) of this graph appear and disappear constantly, which is not considered in the model. Similarly, vertices (documents) are not stable: they are often removed, modified or even change domain. To top it off, web pages are structured in domains, which gives a complex hierarchical structure and promotes clustering, which is also not accounted for in the model.

These discrepancies are notable when analyzing the degree distribution where the obtained values are significantly different than the expected ones. This discrepancy is being studied as of now, and more precise and complex models will surely yield better predictions.

Remark 5.1. It should be noted that this article is older than Google and the newer search engines. Inspired by this article's results (among others), they produced faster and more efficient search algorithms.

5.3 Metabolic networks

In a cell or organism, the processes that sustain life are integrated in a complex network of cellular constituents and reactions. Despite the key roles of these metabolic pathways to sustain cellular functionality, their large scale structure and properties are essentially unknown. Understanding this could not only provide very valuable and perhaps even universal structural information, but also lead to a better understanding of the processes that generated the network itself, i.e., a mathematical approach to evolution. In [14], the topological properties of the core metabolic network of 43 different species are analyzed based on the data gathered at WIT. Unfortunately, this database has been made inaccessible and thus the results are not to be reproduced.

- Vertices: The enzymes or substrates that produce the metabolic processes.
- Edges: The actual metabolic processes. Since under physiological conditions a large number of biochemical reactions are favored towards one direction, this is a directed graph.

Analysis of the metabolic case. As always, we display the probability of a node to have k links, $P(k)$ in Figure 5.5.

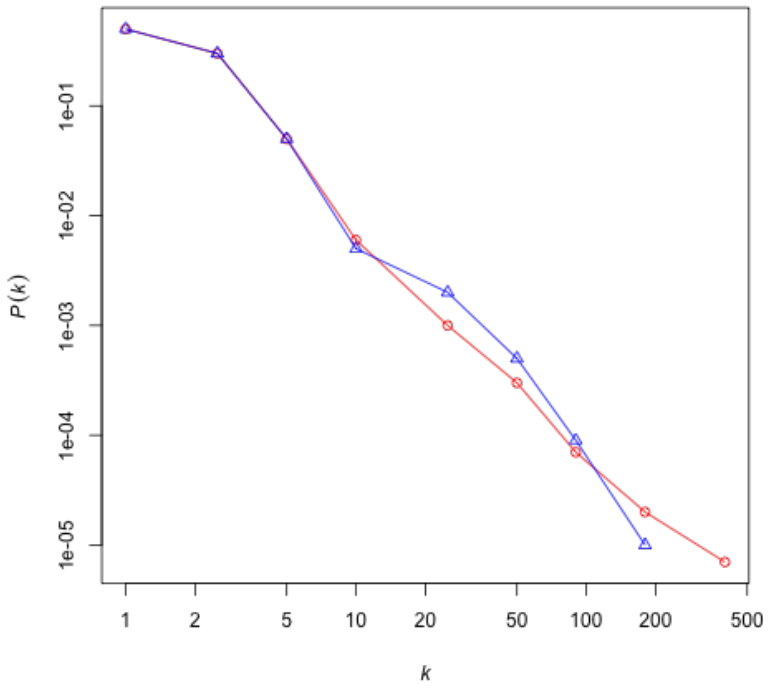


Figure 5.5: $P(k)$ vs k. In red, the outgoing links. In blue, the incoming links.

The results suggest that the probability that a given substrate participates in k reactions follows a power law distribution. For instance, in *Escherichia coli* the probability that a substrate participates in k metabolic reactions follows $P_{in}(k) \sim k^{-2.2}$, and similarly the probability that a substrate is produced by k different reactions is $P_{out}(k) \sim k^{-2.2}$. We can see that the scale-free network is suited to describe all of the organisms analyzed.

Another important aspect is the diameter of the metabolic network. We find that the average diameter is the same for all 43 organisms, independently of the number of nodes of the network, with a few highly linked metabolites acting as the connection between most pathways. These nodes provide important (and already known) biological facts:

- Resistance to random errors Random errors can appear when copying DNA sequences, and thus modify the protein and metabolic structure of the cells. These random mutations can be modeled as eliminating nodes from the network. The diameter remains unaffected after random removal of some non-high linked metabolites, indicating a strong resistance to random errors. This could show an adaptation of these organisms to the fact that random mutations are unpredictable and unavoidable, and ensure that the cell can remain functional even with these deletion.

- Importance of some central metabolites On the other hand, the removal of the highly linked nodes yields a dramatic increase of the diameter, and usually the network becomes disconnected and not functional. This would represent the effect of fatal mutations, where the high connection metabolite cannot be produced by the cell anymore, and thus renders it unable to function normally, since some of the metabolic pathways are now not connected.

Another interesting fact that we obtain is that the ranking of the most connected substrates is practically identical for all 43 organisms. Moreover, only 4% of all substrates that are found in all 43 organisms are present in all species. These represent the most highly connected nodes found in any individual organism, suggesting that the use of these substrates is pretty generic among species. To contrast with this fact, most of the species-specific differences arise from the less connected substrates. To sum it up, most species have the same highly connected nodes to develop "essential" metabolic processes, while the species-specific processes (i.e., specialization) are carried by less connected metabolites, which are typically exclusive to that species.

5.4 Other models

There are many more examples that can be studied applying random graph theory. Networks such as the neuron connectivity in the nervous system, brain structure, social networks and social media, protein structure, economic models, the calling network of phones,...

Although most of these can be analyzed using the BA model, we have seen that the last two examples have yielded $\gamma \neq 3$ by a significant, although small, amount. This can be due to various factors. First of all, most networks are not deliberate constructions, and as such will show some differences (sometimes, these are too big to ignore and a different model has to be used). Also, throughout this manuscript we have left aside those models that make use of rewiring of edges (edges being erased or repositioned), or the possibility of vertices disappearing. These factors offer a degree of non-linear evolution to the networks that the BA model does not take into account, and as such it will fail when applied to networks with these characteristics. To summarize, the BA model is pretty accurate when predicting a power law distribution on this random networks, but is still not good enough to predict the correct γ. More research into these models will help shed light to the topic.

Bibliography

[1] R. Albert, H. Jeong, and A. L. Barabási, *Diameter of the world wide web*, Nature **401** (1999), 130–131.

[2] R. B. Ash, *Probability Theory*, Wiley, 1972.

[3] R. B. Ash, *Real Analysis and Probability*, Academic Press, 1972.

[4] K. Azuma, *Weighted sums of certain dependent variables*, Tohoku Math Journal **3** (1967), 357–367.

[5] A. L. Barabási and R. Albert, *Emergence of scaling in random networks*, Science **286** (1999), 509–512.

[6] A. L. Barabási, R. Albert, and H. Jeong, *Mean-field theory for scale-free random networks*, Physica (Amsterdam) **272** (1999), 173–187.

[7] A. L. Barabási, R. Albert, and H. Jeong, *Scale-free characteristics of random networks: the topology of the world wide web*, Physica A **281** (2000), 69–77.

[8] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády, *The degree sequence of a scale-free random graph process*, Random Struct. Alg. **18** (2001), 279–290.

[9] J. A. Bondy and U.S.R. Murty, *Graph Theory*, Springer, 2008.

[10] Z. Brezniak and T. Zastawniak, *Basic Stochastic Processes*, Springer 2005.

[11] P. Erdös and A. Rényi, *On random graphs*, Publicationes Mathematicae **6** (1959), 290–297.

[12] P. Erdös and A. Rényi, *On the evolution of random graphs*, Publ. Math. Inst. Hung. Acad. Sci. **5** (1960), 17–61.

[13] F. Harary, *Graph Theory*, Addison-Wesley 1969.

[14] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási, *The large-scale organization of metabolic networks*, Nature **407** (2000), 651–655.

[15] S. Redner, *How popular is your paper? An empirical study of the citation distribution*, The European Physical Journal B **4** (1998), 131–134.

[16] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature **393** (1998), 440–442.

[17] Internet Topology Datasets Collected on the Archipelago (Ark) Infrastructure, `http://www.caida.org/projects/ark/topo_datasets.xml`

[18] Skitter AS Links Dataset, `http://www.caida.org/data/active/skitter_aslinks_dataset.xml`

[19] Citation distribution of 1981 Publications Catalogued by the ISI and Cited Between 1981 and June 1997, `http://physics.bu.edu/~redner/projects/citation/isi.html`

[20] Citations to Publications in Physical Review D 1975-1994, `http://physics.bu.edu/~redner/projects/citation/prd.html`

[21] Small-world experiment, `http://en.wikipedia.org/wiki/Small-world_experiment`

[22] The Erdös Number Project, `http://wwwp.oakland.edu/enp/`

[23] The Six Degrees Project, `https://en.wikipedia.org/wiki/SixDegrees.com`