Small-World Network Models and Their Average Path Length

by

Samah M.Osman Taha

Thesis presented in fulfilment of the requirements for the degree of Master of Science in Mathematics at Stellenbosch University

Department of Mathematical Sciences,
Faculty of Science,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Supervisors: Prof. J.W. Sanders
Dr. D. Ralaivaosaona

December 2014

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the owner of the copyright thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature: .....
Samah Taha

Date: 2014/4/8 .....

Abstract

Small-World Network Models and Their Average Path Length

Samah Taha

Department of Mathematical Sciences,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis:

April 2014

Socially-based networks are of particular interest amongst the variety of communication networks arising in reality. They are distinguished by having small average path length and high clustering coefficient, and so are examples of small-world networks. This thesis studies both real examples and theoretical models of small-world networks, with particular attention to average path length.

Existing models of small-world networks, due to Watts and Strogatz (1998) and Newman and Watts (1999*a*), impose boundary conditions on a one dimensional lattice, and rewire links locally and probabilistically in the former or probabilistically adding extra links in the latter. These models are investigated and compared with real-world networks. We consider a model in which randomness is provided by the Erdős-Rényi random network models superposed on a deterministic one dimensional structured network. We reason about this model using tools and results from random graph theory.

Given a disordered network $C(n, p)$ formed by adding links randomly with probability p to a one dimensional network $C(n)$. We improve the analytical result regarding the average path length by showing that the onset of small-world behaviour occurs if pn is bounded away from zero. Furthermore, we show that when pn tends to zero, $C(n, p)$ is no longer small-world. We display that the average path length in this case approaches infinity with the network order. We deduce that at least εn (where ε is a constant bigger than zero) random links should be added to a one dimensional lattice to ensure average path length of order $\log n$.

Uittreksel

Small-World Network Models and Their Average Path Length

Samah Taha

Departement Wiskundige Wetenskappe,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.

Tesis:

April 2014

Sosiaal-baseerde netwerke is van besondere belang onder die verskeidenheid kommunikasie netwerke. Hulle word onderskei deur 'n klein gemiddelde skeidingsafstand en hoë samedrommingskoëffisiënt, en is voorbeelde van klein-wêreld netwerke. Hierdie verhandeling bestudeer beide werklike voorbeelde en teoretiese modelle van klein-wêreld netwerke, met besondere aandag op die gemiddelde padlengte.

Bestaande modelle van klein-wêreld netwerke, te danke aan Watts en Strogatz (1998) en Newman en Watts (1999a), voeg randvoorwaardes by tot een-dimensionele roosters, en herbedraad nedwerkskakels gebaseer op lokale kennis in die eerste geval en voeg willekeurig ekstra netwerkskakels in die tweede. Hierdie modelle word ondersoek en vergelyk met werklike-wêreld netwerke. Ons oorweeg 'n prosedure waarin willekeurigheid verskaf word deur die Erdös-Renyi toevalsnetwerk modelle wat op 'n een-dimensionele deterministiese gestruktureerde netwerk geimposeer word. Ons redeneer oor hierdie modelle deur gebruik te maak van gereedskap en resultate toevalsgrafieke teorie.

Gegewe 'n wanordelike netwerk wat gevorm word deur skakels willekeurig met waarskynlikheid p tot 'n een-dimensionele netwerk $C(n)$ toe te voeg, verbeter ons die analitiese resultaat ten opsigte van die gemiddelde padlengte deur te wys dat die aanvang van klein-wêreld gedrag voorkom wanneer pn weg van nul begrens is. Verder toon ons dat, wanneer pn neig na nul, $C(n, p)$ nie meer klein-wêreld is nie. Ons toon dat die gemiddelde padlengte in hierdie geval na oneindigheid streef saam met die netwerk groote. Ons lei af dat ten minste εn (waar εn n konstante groter as nul is) ewekansige skakels bygevoeg moet word

by 'n een-dimensionele rooster om 'n gemiddelde padlengte van orde $\log n$ te verseker.

Acknowledgements

I would sincerely like to express my sincere gratitude to my supervisors Prof. J.W. Sanders who has been one of the most patient and helpful guides one can ask for. Working with him in this thesis was a great learning experience for me and I am grateful to him for giving me this opportunity. Your advice on both research as well as on my career have been priceless. Also, I would like to express my special appreciation and thanks to Dr. D. Ralaivaosaona. Having him as a co-supervisor has been a privilege. I would like to thank him for being a tremendous mentor for me. Without their endless supports and thoughts I could not have successfully completed my thesis.

Thanks to Dr. Paul Taylor, who took time out of his busy schedules for careful reading my work. Thanks to my office-mate Mr. Ahmed Hassan, who was always ready to help when I needed help.

Finally, my debt in gratitude to all the people -family, friends, teachers, colleagues, and others - who have helped me reach this point. The support I had is greater than I could ever possibly repay.

Dedications

To

My parents:

M.Osman and Ibtesam

Contents

List of Figures

List of Tables

Nomenclature

Variables

$\bar{l}$	Average path length	[ ]
diam	The diameter	[ ]
cc	Clustering coefficient	[ ]

Chapter 1

Introduction

This chapter briefly introduces small-world networks, specifies our research question, presents the related work, summarizes the contribution, and finally gives a summary of the content of the thesis.

1.1 Small-World Network

Complex networks are everywhere; they are present in every aspect of our life, and emerge in a wide range of disciplines in nature and social science. Complex networks, in fact, are all systems of either physical (real) and/or logical (virtual) connected components. They are thus considered complex because of their non-trivial structure which is often called their topology (Jamakovic, 2008). Understanding the topological properties of real-world networks and being able to model them gives us the proper knowledge that is required to understand how the components interact and enables us to predict the overall network performance. Small average path length and high transitivity are two properties that are present in most real-world networks. The standard random network model possesses only the small average path length property. On the contrary, regular network models have high transitivity. Watts and Strogatz (1998) combined these two properties in one network model called ***Small-world networks*** by analogy with the social psychologist Milgram's experiment, popularly known as six degrees of separation.

1.2 Research question

So, small-world properties, short average path length and high transitivity, have been observed in real-world networks. Models of small-world networks combine structured topology represented by one dimensional lattice networks with periodic boundary conditions, i.e., the lattices form a ring. Each vertex is connected to its $2k$ nearest neighbours, k in each side. This structured topology has high transitivity which is order independent and depends only

on k. It is quantified by the clustering coefficient cc. The average path length in this structure is not short, i.e., the average path length scales linearly with the network order. Randomness can be added to structured base networks to reduce the average path length so that it scales logarithmically with the network order. In the literature of small-world models, there are many methods to add such randomness. Our method is to superpose regular base networks and the Erdős-Rényi random network.

Our research question is then: Given a disordered network $C(n, p)$ formed by adding links randomly with probability p to a one dimensional network $C(n)$, what is the average minimum value of p such that the disordered network, $C(n, p)$, is a small-world network? That is, what is the average minimum number of extra random links that can be added to a one dimensional lattice such that the average path length is small. This is equivalent to determining the threshold of the onset of small-world behaviour of a one dimensional lattice.

1.3 Related Work

Although the origin of the study of network can be traced back to Leonard Euler's mathematical solution of the Köningsberg bridge problem in 1735, research in complex networks has expanded broadly in recent years, supported largely by massive real-world data and the availability of powerful computational analysis techniques which makes examining large-scale real-world network structure possible. The discovery of the structure of the internet (Faloutsos *et al.*, 1999), the structure of scientific collaboration (Newman, 2001*c*,*a*,*b*; Barabási *et al.*, 2002), and more recently the Facebook network (Backstrom *et al.*, 2012; Ugander *et al.*, 2011) are good examples of large-scale real-world networks which have been examined recently. Furthermore, in the modelling of networks, the growing observation that real-world networks do not follow the standard Erdős-Rényi random network model has motivated research in proposing other models. The Watts and Strogatz (1998) observation of small average path length and high clustering in social networks leads to small-world networks and the Barabási and Albert (1999) observation of hubs in the world wide web networks leads to scale-free networks.

This thesis contributes to modelling complex networks, in particular, small-world networks. For a general view of the structure of complex networks we recommend the extensive review (Albert and Barabási, 2002). It reviews the recent advances in the field of complex networks focusing on the statistical mechanics of network topology and dynamics as well as covering random network, small-world and scale-free network models. In addition, (Newman, 2003*b*) reviews the same three models as well as models of network growth and preferential attachment. Lastly, Newman (2010)'s book is a good reference for network science in general.

Answering our research question (mentioned in Section 1.2) requires understanding a random network's structure since small-world networks are examples of random networks which mimic empirically observed features in social networks. Then, we investigate the small-world network model as defined by its specific properties. Finally, we examine the cross-over phenomenon where the scale of the average path length in structured networks changes from being linear to logarithmic with the network order. We first survey and discuss the closely related work in those areas.

One of the first quantitative studies of the structure of social networks was given in (Milgram, 1967) by attempting to measure the social-size of the world. The conjecture that was investigated is that in the world of more than 6 billion people, while everyone personally knows only a relatively small sample of the total population, and this small sample is mostly highly connected among each other, everyone in the world is, nonetheless, connected by only a few steps to every one else. Milgram (1967) concluded that, despite the sparsity and the homophily in the structure of the human social network, most of the population is within six degrees of connections from any given stranger. This phenomenon has been labelled "six degrees of separation", a phrase which has since then passed to popular folklore. That number six is not very accurate. However, the general result is that two randomly chosen human beings are connected by a surprisingly short chain of intermediate acquaintances.

One could attempt to explain this by modelling social networks as sparse random networks. Chung and Lu (2001) studied the diameter of a random network $R(n, p)$ for various ranges of p close to the phase transition point for connectivity. They prove that the diameter of $R(n, p)$ is almost surely equal to the diameter of its giant component, i.e., $\mathcal{O}(\log n)$, if $np > 3.6$.

Although a random network does possess a small average path length, it lacks network clustering, which is an essential property of social networks and many other real-world networks. A network is said to have clustering if the probability of two vertices being connected by a link is higher when the vertices in question have a common neighbour. On the contrary, in random networks the probability of two vertices being connected by an link is, by definition, independent of any other connections. That is, the probability of two vertices being directly connected when they have a common neighbour is the same as the probability of being directly connected when they do not have a common neighbour. Newman (2003*a*) discussed random networks as model of real-world networks and reviewed some recent work on the generalization of the random network aimed at correcting random network shortcomings as a model of real-world networks. He concluded that if clustering is introduced in random networks, then any analytical approach becomes significantly harder. Moreover, only a few approximate analytic results for this type of network are known.

The small-world problem is somehow related to a class of problems regarding the study of extremal diameter alteration, i.e., the possible change of

diameter in a network where links are added or deleted. The study of the extremal diameter alteration was initiated in (Chung and Garey, 1984) where the diameter for altered networks after adding t links was studied. Among other results, they showed that if t links are added to an n-cycle, then the least possible diameter of the resulting network for large n was essentially $\frac{n}{t+2}$ when t is even and $\frac{n}{t+1}$ when it is odd. This bound was obtained after adding links in specific way, i.e., the links are not randomly added. A similar work in (Bollobás and Chung, 1988) studied how small the diameter can be made by adding a random matching to an n-cycle. They showed that this construction has diameter about $\log_2 n$. These kinds of study partially discuss the small-world problem but do not contribute to the literature of small-world network since the constructions are different and the intention was not a social network model.

Watts and Strogatz (1998) realized that the small-world phenomenon is not a coincidence and the distinctive combination of high clustering with short path length cannot be captured by the traditional random network models or even by a regular network models. Watts and Strogatz (1998) reconciles those two properties that at first may appear to be contradictory on one network and then call it a small-world network by analogy with the small-world phenomenon (Milgram, 1967). They proposed a specific construction of small-world model by randomly rewiring the links of a "periodic" one dimensional lattice, i.e., a ring. Rewiring means shifting one end of the links and attaching it to a randomly chosen vertex with probability p. With p varying, the model's structural properties can be considered through the two key characteristics: the average path length, $\bar{l}(p)$, and the clustering coefficient, $cc(p)$. Small-world networks lie in the intermediate region between $p = 0$, when the network is still regular, and $p = 1$, when the network is similar to a random network. Watts and Strogatz (1998) concluded that there is a substantial range of values of p in which the value of $\bar{l}$ is small while the values of cc are still high; this range is where the small-world networks lie.

Because of the difficulties of the mathematical treatment of the rewiring process, Newman and Watts (1999a) proposed a variation of the Watts-Strogatz model where extra links, often called short-cuts, are added instead of rewiring the original links in the base network. The Newman and Watts (1999a) model is equivalent to the original model where the average number of the rewiring links in the Watts and Strogatz (1998) model is equal to the average number of the extra links in the Newman and Watts (1999a) model which is nkp, where each vertex in the ring has degree $2k$. The pioneering work in (Watts and Strogatz, 1998) triggered an avalanche of research on the properties of small-world networks, though most results are based on Newman and Watts (1999a) variation. As our interest in this thesis is the average path length $\bar{l}$, in the following we will discuss the closely related work regarding to $\bar{l}$.

Newman (2003b) stated that there is no exact solution for the value of $\bar{l}$, but a number of partial exact results are known, including scaling forms as well

as some approximation of behaviour as a function of the models's parameter. When p is close to zero, the network is large-world, i.e., $\bar{l}$ scales linearly with the network order; by contrast, a small-world network is characterized by the logarithmic scaling of $\bar{l}$, which happens for larger p. Barthélémy and Amaral (1999) study the cross-over from large- to small-world and conjectured that $\bar{l}$ satisfies the scaling form $l(n,p) = nf(\frac{n}{n*})$, where n^* diverges as $n^* \sim p^{-\tau}$ and defines the cross-over. Based on numerical simulations Barthélémy and Amaral (1999) calculated that $\tau = 2/3$. However, simple physical arguments in (Barrat, 1999) obtained that τ cannot be less than one. Newman and Watts (1999*a*) used an asymptotically exact real-space re-normalization group, together with the analytical argument in (Barrat and Weigt, 2000), to show that the exact value of τ is one. This implies that the cross-over occurs at fixed n when $p \sim \frac{1}{n^{1/\tau}}$. The notation $\sim$ used for the onset of small-world behaviour is ambiguous. It actually means $\lim_{n\to\infty} \frac{\log p}{\log n} = \frac{-1}{\tau}$, then $\tau = 1$. This result leaves the threshold area large where we can decide if the network is small-world or not, the result indicates that when $\tau > 1$ the network is small-world and when $\tau < 1$ the network is not small-world, otherwise we can not decide.

Regarding the scaling function $f(x)$, a mean field-treatment is given in (Newman *et al.*, 2000) to show that $f(x) = \frac{1}{2\sqrt{x^2+2x}} \tan^{-1} \sqrt{\frac{x}{x+2}}$. Newman and Watts (1999*b*) calculated the form of the scaling function to fifth order in the short-cut density using a series expansion and to third order using Padé approximants. Making any further progress has proved difficult. Dorogovtsev and Mendes (2000) presented an exact description of a crossover between the two different regimes of simple small-world networks where there are no short-cuts, but instead random vertices are connected to a single central extra vertex or hub. Newman (2000) provides a short review of small-world models.

In addition to the small-word studies, a new accurate result of the diameter was obtained by Flaxman and Frieze (2004), providing a network with $\mathcal{O}(\log n)$ diameter. The result holds for a strongly connected directed network with bounded degree that is perturbed by adding a random network $R(n, \frac{\varepsilon}{n})$ where ε is a constant and $\varepsilon > 0$. The resulting network denoted by D is a network formed by taking the union of the links of base directed network with bound degree network D' and random network $R(n, \varepsilon/n)$. Although the motivation of this result is not the restriction of a social network and does not guarantee the high transitivity in the network, the construction is close to small-world network in the sense that its diameter is of logarithmic order. The result of the diameter is general such that it can be applied to small-world network models.

1.4 Contribution

As we have mentioned in Section 1.3, the threshold value of onset of small-world behaviour has been determined as, $\lim_{n\to\infty} \frac{\log p}{\log n} = \frac{-1}{\tau}$. This result leaves a wide range of p values where we cannot decide whether the network is small

or not, i.e., when $\tau = 1$. In this work, we use Erdős-Rényi random networks which have the advantage of being analytically solvable in many of their average properties and being one of the oldest and best studied network models, to construct a solvable small-world network model. Using random network structure $R(n, p)$ when np is bounded away from zero, we narrow the threshold to be $p = \frac{\varepsilon}{n}$ where ε is a constant $\varepsilon > 0$, and we construct a small-world network with diameter $\mathcal{O}(\log n)$ which bounds the average path length below by $\mathcal{O}(\log n)$. In addition, we study the random network structure when pn tends to zero , and we use the result to study average path length of small-world models. We show that in this case $\bar{l} > \mathcal{O}(\log n)$, i.e., the network is not small-world. We deduce that at least εn where ε is a constant and $\varepsilon > 0$ random links should be added to a one dimensional lattice to ensure average path length of order $\log n$.

In addition to the analytical work, we conduct an empirical study of the scientific collaboration, citation and Facebook social network data by analysing their average path lengths, clustering and degree distributions. The results show the unsuitability of random networks as models of real-world networks, as they do a poor job of capturing real-world network properties like the high clustering and the scale-free degree distribution. In addition, we simulate Watts and Strogatz (1998) and Newman and Watts (1999*a*) models of small-world networks to confirm the results and to compare their properties with real-world network properties. Small-world network models do a good job in capturing real-world network properties but they do not have the scale-free degree distribution. However, it should be noted that the models were never intended to mimic real-world degree distribution.

1.5 Thesis Organization

This thesis organized as follows:

The second chapter introduces the foundations on which the other chapters are based. Particularly, it introduces random network models and investigates their properties compared with three examples of real-world networks.

The third chapter defines small-world properties and studies Watts and Strogatz (1998) and Newman and Watts (1999*a*) models of small-world network, with particular attention to average path length. The properties of small-world network are also comparable with real-world networks.

The fourth chapter considers a small-world model provided by the Erdős-Rényi random network, superposed on a deterministic structured network to provide new results regarding average path length of small-world network models.

Finally, the fifth chapter summarizes our thesis and presents the conclusion and further direction of research.

Chapter 2

Foundation

Traditionally, complex networks have been modelled as random networks. As complex network science has become a focus of attention and continues to grow in importance and popularity (because of the availability of the empirical studies of real-world networks), other models of networks have been developed in order to bridge the gap between random networks and real-world networks. This chapter introduces random networks and presents an empirical study of some real-world networks to make visible the unsuitability of random network in modelling real-world networks. Before giving the formal description of random network models, we concisely present the basics of graph theory and complex network terminologies.

2.1 Introducing Graphs

In this section we will recall the basics of graph theory that are necessary for a basic understanding of this thesis. There are a variety of books studying graph theory from different viewpoints; this section relies on (Bollobás, 1998).

A graph G is a pair (V, E) where $V(G)$ is a set whose elements are called vertices, and $E(G)$ is a set of pairs (u, v) called edges or links, where $u, v \in V$. The cardinality of V is called the graph's order and is denoted by $|V|$. Unless stated otherwise, in this work we use the notation $n \equiv |V|$. A pair $(u, u) \in E$ is called a loop. Multiple edges occur when more than one edge is connected with the same vertices. In this case the graph is said to be multi-graph. If we distinguish (u, v) and (v, u), then the graph is said to be directed (and all its edges are directed); otherwise the graph is undirected. Most graphs in this thesis are undirected graphs, and thus the term graph refers to simple undirected graphs unless we state otherwise. Note that there are $\binom{n}{2}$ possible links in a simple undirected graph. A graph where all possible edges are present is called a complete graph and denoted by K_n. The neighbourhood of vertex v is the set N_v of all vertices u such that $(v, u) \in E(G)$. The degree of vertex v, denoted by $\deg(v)$, is the cardinality of N_v, i.e., $|N_v|$.

2.2 Complex Networks Properties

Understanding complex networks becomes a fundamental problem in graph theory. What does it mean for two large graphs to be similar to each other, when they may differ in obvious ways such as their numbers of vertices? There are many types of networks (social, biological, economic, physical, technological, etc.), whose details vary widely, but they have similar structural and growth phenomena. For this reason, it is natural to consider the network as a sequence of graphs with order tending to infinity and ask whether these graphs converge to any meaningful limit.

Studying real-word complex networks using graph theory techniques is simple, yet powerful, because it treats complex networks from different fields using the same mathematical tools and methods. Moreover, complex networks are difficult to visualize and describe. To be able to compare networks and classify them according to the properties that they display, this section introduces important and robust metrics that are used to characterize and provide meaningful insight into the structure of complex networks. There are a variety of books introducing complex networks and studying them from different viewpoints. We have chosen to rely on (Van Steen, 2010) and (Jackson, 2010).

2.2.1 The Average Path Length and The Diameter

A path P is a finite sequence of edges which connects a sequence of vertices, and it forms its own network of the form $V(P) = \{v_0, v_2, \ldots, v_l\}$ and $E(P) = \{(v_0, v_1), (v_1, v_2), ...(v_{l-1}, v_l)\}$. The path length $l = (v_0, v_l)$ is $|E(P)|$. The network is called connected if there is a path between any two vertices. If the network is not connected, the isolated parts individually comprise connected components, and the "unique" largest component with $\mathcal{O}(n)$ vertices usually called the giant component, In other words for each n there is a connected subset $C(n)$ of $G(n)$ in the network $G = G(n)$ satisfying $|C(n)| = O(n)$ in the usual sense that eventually $sup_n|G(n)|/n, \infty$. So, a connected network has only one connected component.

Since the network might have many paths of different length between v_i and v_j, we reserve the notation l to refer to the shortest path length between v_i and v_j. For brevity we say path length instead of shortest path length. If there is no path between the vertices, we say the path length is infinite.

Definition 2.1. *Let $G(n)$ be a connected network and let $\bar{l}(u)$ be the average path length from vertex u to other vertices in $G(n)$:*

$$\bar{l}(u) = \frac{1}{n-1} \sum_{v \in V, v \neq u} l(u, v).$$

The average path length of the network, $\bar{l}(G)$, is defined as

$$\begin{aligned}\bar{l}(G) &= \frac{1}{n}\sum_{u\in V}\bar{l}(u)\\ &= \frac{1}{n}\sum_{u\in V}\left(\frac{1}{n-1}\sum_{v\in V}l(u,v)\right)\\ &= \frac{1}{n(n-1)}\sum_{u,v\in V}l(u,v)\end{aligned} \tag{2.2.1}$$

The diameter of a network G is the length of the longest path,

$$\text{diam}(G) = \max\{l(u,v)|u,v\in V(G)\}.$$

Notice that these definitions apply to undirected as well as directed networks. Although the diameter gives useful information by indicating the length of the longest path in the network and also provides an upper bound for the average path length, it is not very powerful in distinguishing networks meaningfully. However, the average path length does, and it will come up as a quantity of interest in this work frequently. In the case when the network is not connected, the diameter and the average path length are computed for the largest component in the network.

2.2.2 Transitivity or Clustering Coefficient

A common property of social and many other complex networks is transitivity. In the context of social networks this means that the probability between people to be friends is not equal; two people with a mutual friend are more likely to be friends than any two people chosen at random, i.e., there is a high density of triangles in the networks. This tendency to cluster is quantified by the clustering coefficient, which can be interpreted as the probability that two randomly chosen vertices are connected given that they have a common neighbour. Formally, the clustering coefficient, cc, of a network $G(n)$ is defined as follows (Watts and Strogatz, 1998):

$$cc \equiv cc(G) = \frac{1}{n}\sum_{v\in V}cc(v), \tag{2.2.2}$$

where $cc(v)$ is the local clustering coefficient of vertex v and is defined as follows: suppose that vertex v has k_v neighbours. For $k_v = 0$ or 1 the local clustering coefficient, $cc(v)$, is defined to be zero. When $k_v \geq 2$, then there are $\frac{k_v(k_v-1)}{2}$ possible undirected links among the neighbours of v. We define $cc(v)$ as the ratio between the actual existent links to all possible links for vertex v.. Let N_v be the set of neighbours of v and $k_v \geq 2$. Then

$$cc(v) = \frac{2|E\cap(N_v\times N_v)|}{k_v(k_v-1)}. \tag{2.2.3}$$

The clustering coefficient is equal to 1 for a network where any pair of vertices are linked, i.e., complete network, and it is equal to 0 in trees.

Figure 2.1 shows an example of the global clustering coefficient, using the formula in (2.2.2)

$$cc(v_1) = 0,\ cc(v_2) = 1, cc(v_3) = 1,\ \text{and } cc(v_4) = 1,$$

then,

$$cc(G) = \frac{1}{4}\{0 + 1 + 1 + 1\} = \frac{3}{4}.$$

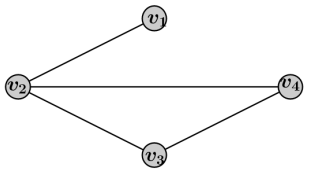


Figure 2.1: The clustering coefficient of the above network is 3/4.

2.2.3 The Degree Distribution

Since not all vertices in the network have the same degree, the average degree of a network $G(n)$ measures how many edges are in set E compared with the number of vertices in set V. This quantity is given by:

$$\deg(G) = \frac{1}{n}\sum_{v \in V} \deg(v). \tag{2.2.4}$$

The spread of vertex degree k is given by its distribution over the whole network which is denoted by $\mathbb{P}_k$.

Directed networks have two different degrees for each vertex: the in-degree and the out-degree, which are respectively the number of edges in-going and out-going at the vertex in question. Therefore, there are two different degree distributions in a directed network, the in-degree and out-degree distributions.

2.3 Random Network

Traditionally, networks with complex topology and unknown organizing principle are considered to be random; i.e., networks in which the edges are randomly distributed. Historically, these were founded and greatly detailed by

Erdős and Rényi (1960), and an in-depth review of the field is available in the classic book of Bollobás (2001), as well as in Janson *et al.* (2011). In the 20th century, random network theory has become more statistical and algorithmic and a particularly rich effort has been dedicated to random networks as one of the earliest theoretical models of real-world networks. In the following we briefly describe the aspects of random network theory that are of direct relevance to our work.

2.3.1 Random Network Models

Among several models of random networks, there are two basically equivalent models, and both are formed in the simple model by Erdős and Rényi (1960). Mainly, these models only differ in the way the sample space is chosen (or the way the edges are chosen). This section is confined to those two models.

Definition 2.2. *Let V be the vertex set, i.e., the vertices are labelled, and a real number p, $0 \leq p \leq 1$. The binomial random network denoted by $R(n,p)$ is constructed by connecting each pair of vertices with probability p independently*

$$\mathbb{P}((v_i, v_j) \in E) = p,$$

for $v_i, v_j \in V$. Therefore, all such networks with n vertices and $|E|$ links have equal probability of

$$p^{|E|}\ (1-p)^{\binom{n}{2}-|E|}.$$

Placing the links between each pair of vertices in a binomial random networks can be seen as a result of a coin flipping, i.e., $\binom{n}{2}$ flips in total, with probability of success (drawing an edge) equal to p. This model is often used in the field of complex networks where it is assumed that the links in the real-world occur randomly. The main advantage of this model is the independence of the presence of links, but the drawback is that the number of edges is not fixed which makes this model hard to analyse and enumerate.

The number of the edges in a binomial random network varies according to the binomial distribution with expectation $p\binom{n}{2}$. If we impose the condition that $|E(R(n,p))| = M$, then a uniform space exists and this leads us to the other random networks model; the uniform random network model.

Definition 2.3. *Let V be the vertex set and an integer M, $0 \leq M \leq \binom{n}{2}$, the uniform random network, denoted by $R(n,M)$, is defined by taking the sample space as the family of all networks on the vertex set $V(n)$ with exactly M edges. Equivalently, all networks with n vertices and M links have equal probability of*

$$1 \Big/ \binom{\binom{n}{2}}{M}.$$

This model is closely related to enumerative combinatorics, which deals with the number of ways that certain patterns can be formed.

These two basic models of random networks are in many cases asymptotically equivalent. This equivalence indicates that in a large majority of cases the properties of random networks $R(n,p)$ and $R(n,M)$ are very similar provided that $M \sim \binom{n}{2}p$ and the investigated property is a monotone property. The details appear in (Bollobás, 2001, p. 14-17).

Remark 2.4. *The abbreviation, a.a.s, meaning asymptotically almost surely, as well as, w.h.p, meaning with high probability both are used to denote an event that holds with probability tending to 1 as* $n \to \infty$.

2.3.2 Random Networks Evolution

Imagine a random network as an organism which grows with time. It is born as a set of n isolated vertices and develops by successively adding edges at random. Our objective in this section is to study the global structure of random networks at different stages. A brief intuitive and imprecise description with no details of the fundamental results of random networks evolution is given by answering the following questions: given a specific probability p, how do the connected components look and what is their order? Random network evolution has been studied generously in (Bollobás, 2001, p. 96-159). Theorem 2.5 summarizes random network evolution.

Theorem 2.5. *Consider a random network* $R \in R(n,p)$ *where* $p = \frac{\varepsilon}{n}$ *for some constant* ε. *Then:*

1. *If* $\varepsilon < 1$, *then a.a.s, most of the components in the network are isolated trees and the largest component of* $R(n,p)$ *has order* $O(\log n)$.
2. *If* $\varepsilon > 1$, *then a.a.s,* $R(n,p)$ *has a unique giant component of order* $O(n)$. *All other components of order* $O(\log n)$.
3. *If* $\varepsilon = 1$, *then a.a.s,* $R(n,p)$ *has a component of order* $O(n^{2/3})$.

2.3.3 Random Networks Properties

Random graphs are considered one of the earliest theoretical network models. This section relies on (Albert and Barabási, 2002) to summarize the properties of random networks as they are defined in Section 2.2.

Degree Distribution

As we have mentioned earlier, the average number of edges in a $R(n,p)$ random network with n vertices is $\frac{n(n-1)}{2}p$. Since the edges occur randomly, most of

the vertices have the same degree which is close to the average degree. The average degree in the binomial random networks is defined as follows:

$$z = \frac{n(n-1)p}{n} = (n-1)p \sim np, \quad \text{as} \quad n \to \infty.$$

The degree distribution in a $R(n,p)$ random network is binomial by definition. $\mathbb{P}_k$ denotes the probability that the degree of a randomly chosen vertex is k, i.e., $p(\deg = k)$, and thus,

$$\mathbb{P}_k = \binom{n-1}{k} p^k (1-p)^{n-1-k}.$$

In the limit when $np \sim z$ where z is a constant, this becomes

$$\mathbb{P}_k = \frac{z^k e^{-z}}{k!},$$

which is the Poisson distribution.

Clustering Coefficient

Recall that a network is said to show clustering if the probability of two vertices being connected by an edge is higher when the vertices have a mutual neighbour. On the other hand, the probability of edge occurrence in Erdős-Rényi random networks are by definition independent. Therefore, Erdős- Rényi random networks have low clustering coefficient.

In a $R(n,p)$ random network, the probability that two neighbours of a given vertex is connected is p. Therefore, we expect

$$cc = p.$$

The Diameter and The Average Path Length

A random network is likely to be spread out as the probability of two vertices to be connected is independent of any other nodal connections. If the average degree of a random network is z, with large probability the number of vertices at a distance l from a given vertex is not much smaller than z^l. Then, equating this with n, one obtains:

$$z^{\text{diam}} = n$$
$$\text{diam} = \frac{\ln(n)}{\ln(z)}.$$

This means that a random network tends to have a small diameter as its diameter scales logarithmically with n. Projecting this to social networks, if a random person A knows z people on average, and each of A's friends also has

z friends, then A has about z^2 second friends. Extending this argument, in l steps A can reach z^l acquaintances. Note that if every person has 100 direct friends, person A can reach most of the population in the world in 5 steps (the world population in 2014= 7.155 billion). This logic does not hold for social networks where the acquaintances tend to overlap.

A variety of studies have investigated the diameter of random networks see (Chung and Lu, 2001; Bollobás, 1981). A general conclusion is that for most values of p, almost all random networks with the same n and p have precisely the same diameter usually concentrated around

$$\text{diam} = \frac{\ln(n)}{\ln(z)}.$$

Fronczak *et al.* (2004) confirmed the observations that the average path length for Erdős- Rényi random networks scales logarithmically with n,

$$\bar{l} \sim \ln(n).$$

2.4 Touching upon Real-World Networks

One motivation for the study of complex networks has been the need to understand real-world networks. Understanding and visualizing the structure of real-world networks stimulates their empirical study. In this section we study briefly three classes of real-world networks: scientific collaboration network, the citation network and social network, e.g., Facebook. Beyond a description of the data set, we shall focus on three important measures of a network's topology as we have discussed in Section 2.2: average path length, clustering coefficient and degree distribution.

2.4.1 Going from Scientific Collaboration network

In the scientific collaboration networks the vertices are scientists and two scientists are connected if they have at least one joint publication. One of the first studies that investigated the properties of the scientific collaboration network was done by Newman (2001*c*,*a*,*b*) and investigated four different networks. Table 2.1 gives a summary of some of the results of (Newman, 2001*c*).

Looking at Table 2.1, we can see that despite the fact that the networks of interest of Newman (2001*c*) are large, they have a small average path length and high clustering, meaning that two scientists are much more likely to collaborate if they have a third common collaborator. Newman (2001*c*) found that the degree distribution of the scientific collaboration network is well fitted by power-law forms with an exponential cut-off, i.e, $\mathbb{P}_z \sim z^{-\tau} e^{-z/z_c}$, where τ and k_c are constants, and z is the number of collaborators. Fits to this form are shown as the solid lines in Figure 2.2.

	MEDLINE	Los Alamos	SPIRES	NCSTRL
Total number of authors	1,520,251	52,909	56,627	11,994
Collaborators/ author	18.1	9.7	173	3.59
clustering coefficient cc	0.066	0.43	0.726	0.496
average path length $\bar{l}$	4.6	5.9	4.0	9.7

Table 2.1: The analysis results for different collaboration network (Newman, 2001*c*). MEDLINE: papers on biomedical research; Los Alamos: preprints in theoretical physics; SPIRES: papers and preprints in high energy physics and NCSTRL: preprints in computer science. The analysis is done during a five years period 1995-1999.

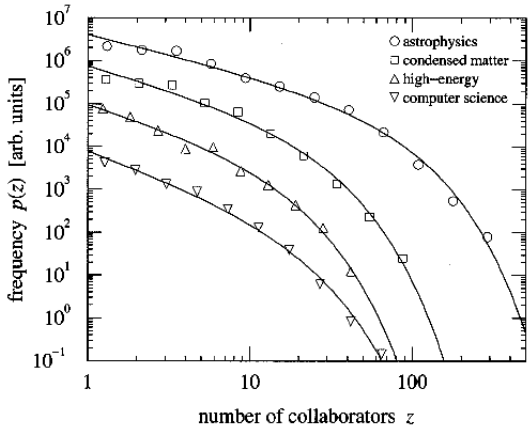


Figure 2.2: The degree distribution of a scientific collaboration done by (Newman, 2001*c*).

A similar study was done in (Barabási *et al.*, 2002) where the interest was more in the evolution of scientific collaborations. The electronic database that is used contained all relevant journals in mathematics and neuroscience for an eight year period (1991-98).

In this section we investigate the Arxiv ASTRO-PH (Astrophysics category) collaboration network. The data was provided from the Stanford large network dataset Collection's site[1] and it covers papers in the period from January 1993 to April 2003. Table 2.2 gives a summary of the results of investigating the properties of the Astrophysics collaboration network. Figure 2.3 shows the degree distribution. A link between two scientists represents collaboration regardless of the number of papers. Table 2.2 shows that the

[1]http://snap.stanford.edu/data/

Astrophysics collaboration network possesses high clustering as well as a small average path length compared to the order of the network. Figure 2.3 shows that the resulting distribution of the network closely follows a scale free degree distribution which mathematically means that the degree distribution follows a power-law, i.e., $\mathbb{P}_z = z^{-\tau}$, where τ is a constant exponent. In a scale-free network we find that some vertices, which are called "hub", have many more connections than others; this can be explained in the case of the scientific collaboration network as the majority of the scientists have a few collaborators while very few have many collaborators.

Number of vertices	17903
Number of edges	197031
Average degree	22.01
Maximum degree	504
Clustering coefficient	0.6328
Average path length	4.19
The diameter	14

Table 2.2: Summary of the analysis of Astro physics collaboration network.

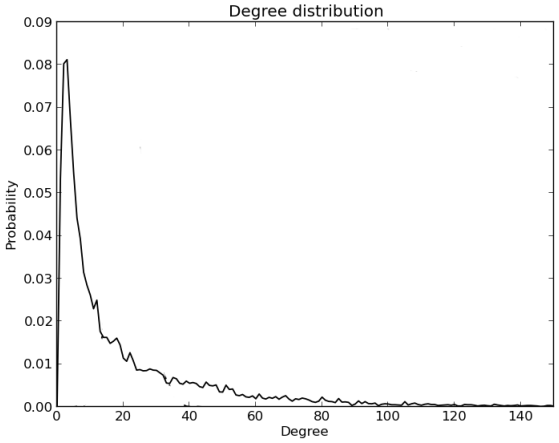


Figure 2.3: The degree distribution of Astro physics collaboration network in the period from January 1993 to April 2003.

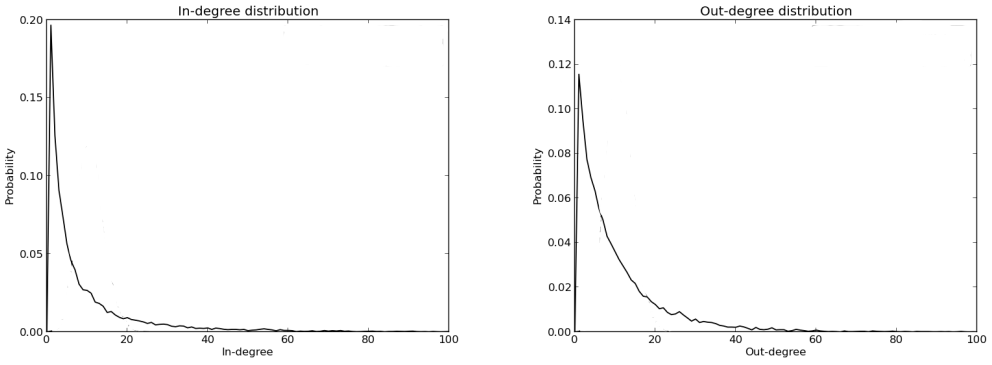
Furthermore, Erdős numbers have been a part of the folklore of mathematicians throughout the world for many years. The Erdős number describes

the collaborative distance between a person and the mathematician Erdős as measured by authorship of mathematical papers. The idea was created by the mathematician's friends as an honour to his enormous contributions in mathematics. The collaboration network for Erdős numbers is mostly based on information of the American Mathematical Society's mathematical reviews of July, 2004. It contains 401,000 different authors as its vertices and 676,00 edges. The links between two vertices represent the existence of collaboration. The average path length of the Erdős collaboration network is 4.65 and the diameter is 13. More information is provided in the Erdős number's project site. [2]

2.4.2 ...Over The Citation Network

Another complex real-world network is formed by the citation of scientific publications. The networks here are directed, the vertices are published papers and a directed edge represents a reference to a previously published paper. Redner (2004); Vázquez (2001) studied the degree distribution of the citation network of papers in *Physical Review* between 1975-1994.

In this section we investigate the properties of the Arxiv HEP-PH (high energy physics phenomenology) citation network. The data is provided from the Stanford large network dataset Collection's site which covers all the citations in the period from January 1993 to April 2003. Table 2.2 gives a summary of the properties of the high energy physics citation network. Figure 2.4 shows the in-degree and out-degree distributions.



(a) In-degree distribution (b) Out-degree distribution

Figure 2.4: The degree distribution of the high energy physics citation network in the period of 1993-2003.

Table 2.3 and Figure 2.4 show that the results are similar to the one in the collaboration network, although it shows a larger value for average path

[2]http://www.oakland.edu/enp/

Number of vertices	12711
Number of edges	139981
Average in-degree	11.01
Average out-degree	11.01
Maximum in-degree	377
Maximum out-degree	322
Clustering coefficient	0.28
Average path length	14.4
The diameter	49

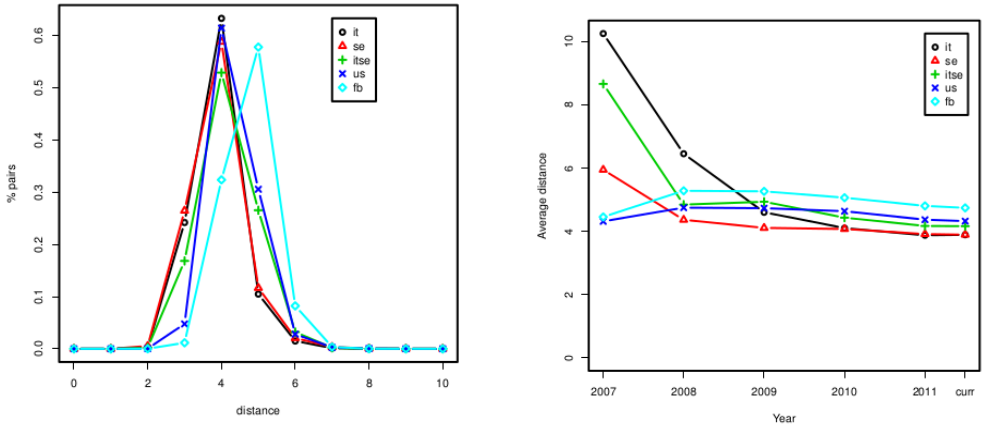
Table 2.3: Summary of the analysis of the high energy physics citation network in the period of 1993-2003.

length but it still has a logarithmic scale. The result shows that the citation network has small $\bar{l}$ and a high cc with scale-free distribution.

2.4.3 To Facebook

Facebook is one of the online social network services. It is considered to be the largest social network ever analysed containing $\approx$ 721 million active users and $\approx$ 69 billion friendship links. Backstrom *et al.* (2012); Ugander *et al.* (2011) investigated the Facebook networks and computed numerous features of the network. In particular, they examined the path length distribution and also some interesting geographic subgraphs by observing the evolution of the Facebook over time. They analysed the entire Facebook network (fb), the USA subgraph (us), the Italian subgraph (it), the Swedish (se) subgraph, and the combination of the Italian and Swedish network (itse). In the latter case, they were specifically checking whether combining two regional but distant networks could significantly change the average distance, in the same spirit as in Milgram's original experiment which was performed in different regions. Backstrom *et al.* (2012); Ugander *et al.* (2011) reported that the whole Facebook network is within four degrees of separation. The result of (Backstrom *et al.*, 2012) regarding the path length in the Facebook network is depicted in Figure 2.5. Figure 2.5(a) shows that the average path length in the regional networks concentrates around four while the average path length of whole Facebook network is concentrated around five. On the other hand, Figure 2.5(b) shows the fast growth of the Facebook network through a quick decrease in the average path length, which appears to be stabilizing now. Moreover, although

the order of the (se) network is much less than the (it) network (see Table 2.4), however, they have similar values for the average path length, 3.89 and 3.90 respectively (see table 2.5) which indicates that the average path length is indeed dependent on the geographical closeness of users, more than on the actual order of the network. This is confirmed by the higher average path length of the (itse) network which has a larger value of $\bar{l}$.



(a) The distribution of the path length. (b) The average path length.

Figure 2.5: Backstrom *et al.* (2012)' result regarding the path length of Facebook network 2012.

	it	se	itse	us	fb
2012	8.3	1.2	9.7	68.5	344.9

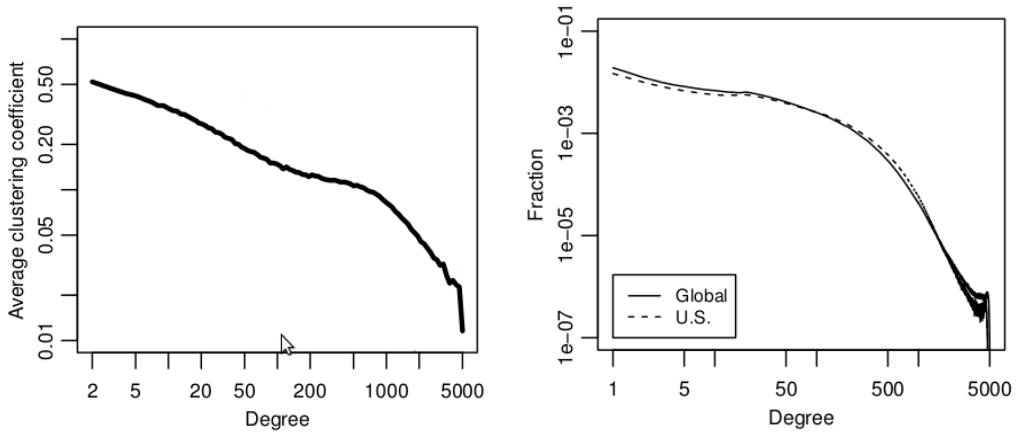
Table 2.4: The size in G-bytes of different Facebook networks in 2012 (Ugander *et al.*, 2011).

	it	se	itse	us	fb
2012	3.89	3.90	4.16	4.32	4.74

Table 2.5: The average path length of different Facebook network in 2012 (Ugander *et al.*, 2011).

(Ugander *et al.*, 2011) investigated the clustering coefficient of the Facebook network. As shown in Figure 2.6(a), the clustering coefficient of the Facebook

network is very large essentially independent of the degree of the vertices. They demonstrated that, as in Figure 2.6(b), the degree distribution is nearly monotonically decreasing, except for a small deviation near 20 friends due to the Facebook strategy to encourage people with few friends to gain more until they reach 20 friends. The distribution shows a clear cutoff at 5000 friends which is also a limit imposed by Facebook on the number of friends at the time of measurement.



(a) Clustering (b) Degree distribution

Figure 2.6: The clustering coefficient and the degree distribution of Facebook network (Ugander *et al.*, 2011).

In this section we investigate Facebook data collected from the Mynetwork application in Facebook[3]. The data collected represent the friendship of one active user and his/her friendships. Table 2.6 summarises the result of our analysis. Facebook sub-network has a large clustering coefficient and a small average path length. Figure 2.7 shows the degree distribution of our Facebook data. We can see that the degree distribution in our case does not agree with the degree distribution of the whole Facebook data shown in Figure 2.6(b); this difference may be expected as our sample of data is for a single example case.

2.4.4 Random Networks Suitability

In this section we compare the properties of the random network of order n with our results for the collaboration network, the citation network and the Facebook social network , particularly regarding the degree distribution, average path length and clustering coefficient. We choose the value of p in the random network such that the random network has the same average degree

[3]https://www.facebook.com/MyFnetwork.

Number of vertices	7532
Number of edges	62079
Average degree	16.4
Maximum degree	207
Clustering coefficient	0.4733
Average path length	5.908
The diameter	16

Table 2.6: Summary of the analysis of Facebook data collected from Mynetwork application in 2013. The data represents one active user's friendship and his/her friendships.

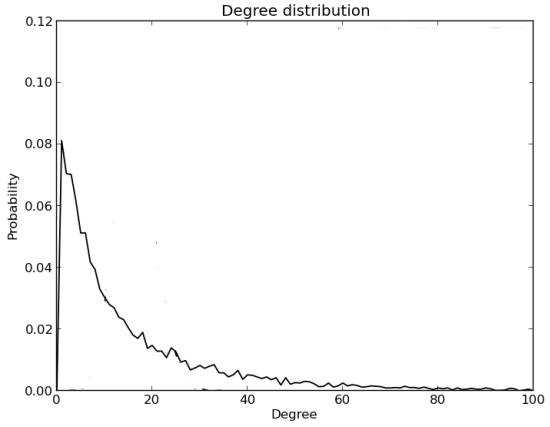
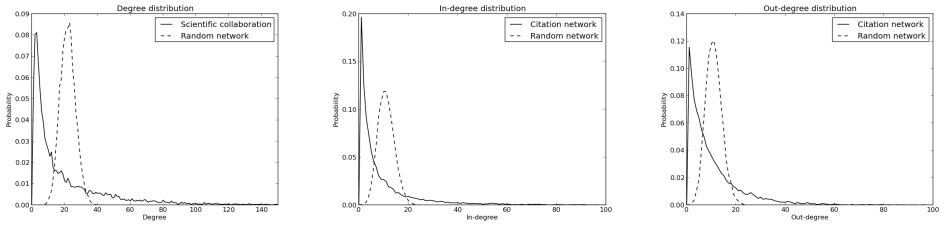


Figure 2.7: The degree distribution of the Facebook data collected from MyNetwork application in 2013. The data represents one active user friendship.

as the network in question. See comparison in Table 2.7 as well as in Figure 3.10 for degree distribution.

Network	$\|V\|$	E	E_{rand}	z	$\bar{l}$	$\bar{l}_{rand}$	cc	cc_{rand}
Collaboration network	17903	197031	196353	22.0195	4.194	3.5165	0.6328	0.00118
Citation network	12711	139981	140117	22.0	14.494	3.387	0.2848	0.00183
Facebook network	7532	62079	61841	16.484	5.90	3.51	0.473	0.00212

Table 2.7: Comparison between Random networks and real-world networks.



(a) Collaboration vs. Random network (b) Citation in-degree vs. Random network (c) Citation out-degree vs. Random network

Figure 2.8: Comparison between Random networks and real-world networks regarding the degree distribution.

As Table 2.7 and Figure 3.10 show, random networks display good agreement with real-world networks regarding average path length. On the other hand, random networks possess very low clustering coefficients compared with real-world networks. Moreover, real-world networks show scale-free degree distribution, which is in opposition to random networks, which possess binomial degree distribution.

2.5 Conclusion

In this chapter three topics have been discussed. Firstly, we introduced complex network terminologies by defining three important properties that are used to characterize complex networks. We then described random networks and their properties. Finally, we investigated three examples of real-world networks and compared their properties with random networks. The comparison showed that Erdős and Rényi random networks do a poor job as a model of real-world networks.

The three metrics in question are: the average path length, the clustering coefficient and the degree distribution. The importance of these quantities has been emphasised by empirical studies of real-world networks which have recently revived network modelling and resulted in an enormous number of studies in network science. First, random networks: despite the fact that their properties deviate from real-world networks, random networks are still widely used in many fields and serve as a standard for many modelling and empirical studies. In addition there are many studies devoted to overcoming the shortcoming of random networks as a model of real-world networks; see (Newman, 2003*a*; Newman *et al.*, 2002). Second, stimulated by the high clustering property observed in real-world networks, a class of models called small-world models has been proposed which interpolate between the high clustering regular lattices and random networks (Watts and Strogatz, 1998). Finally, the discovery of the power-law degree distribution has led to the construction of various scale-free models (Barabási and Albert, 1999) (not part of this study).

Our study in this thesis contributes to small-world network models by using the theory of random networks, as we shall see in the following chapters.

Chapter 3

Small-World Networks

Empirical studies of many large-scale real-world networks such as social networks, the internet, biological networks, etc. have shown that most real-world networks have common properties: their order is large, links are globally sparse, they have a high degree of local transitivity, and despite the last three properties (bigness, sparsity, and transitivity) they have a small characteristic path length, i.e., their path length grows slowly with their order. In 1989 Duncan Watts and Steven Strogatz realized that the distinctive combination of high transitivity and small path length cannot be captured by the traditional random network models, which possess small path lengths on average but lack transitivity. On the other hand, a completely ordered lattice "regular network" lacks small characteristic path length, i.e., its path length grows linearly with the system order. Watts and Strogatz (1998) combined those two properties that appear to be contradictory into one model, which is referred to as small-world network model by analogy with the small-world experiment (Milgram, 1967). Their paper "Collective dynamics of 'small-world' networks" was published in Nature June 1998 and ranks at 6 among highly cited papers in Physics. This paper has garnered 2,700 citations between January 1, 1998 and August 31, 2008. This chapter presents small-world networks: their structure and definition, models and properties. The objective is to go beyond the empirical studies presented in Chapter 2 and to show that the small-world network phenomenon is not coincidence. The main results studied have been confirmed by numerical simulation as well.

3.1 Structure and Definition

The term small-world network is used frequently to refer to the Watts (2004)' model, which was originally based on random perturbations of a regular lattice. A small-world network is characterized by "small" average path length, $\bar{l}$, and " high"clustering coefficient, *cc*. More elegantly, it is a network which is highly clustered and in which the average path length increases sufficiently slowly as

the number of vertices n increases. Usually the average path length of a small-world network must be comparable to the average path length in a random network ($\mathcal{O}(\log n)$).

The words "small" and "high" have not been precisely quantified in the Watts and Strogatz (1998) definition, which emphasizes the two most important properties of small-world network: the average path length $\bar{l}$ and the clustering coefficient cc. Although small-world networks are well studied, the precise definition is controversial. In the study of small-world networks, most of the attention has been dedicated to computing $\bar{l}$. This study is no exception. Defining small-world networks in terms of limiting $\bar{l}$ is a shortcoming since not every network with small $\bar{l}$ can be seen as a small-world network. Bear in mind that we use the expression "small-world effect" to describe any network with a small average path length $\bar{l}$. Even a regular lattice can have the two properties of small-world; high clustering and small $\bar{l}$, if the degree is not bounded yet. In addition, restricting the definition of small-world network to the specific construction of Watts-Strogatz model cannot answer the question whether a given network is small-world or not unless we have a precise definition of small-world network.

We conclude that a precise definition of small-world networks is somewhat baroque. It is not confined by the average path length $\bar{l}$ or even by the Watts and Strogatz (1998) construction. For further discussion on this point see Examples 3.2, 3.3 and 3.4.

Cont and Tanimura (2007) proposed an intrinsic definition of small-world networks based on the scaling properties of the network. Their definition is more reasonable than the intuitive one by Watts and Strogatz (1998). It does not rely on an underlying lattice nor on any particular construction and includes most properties of small-world networks.

Definition 3.1. *Let $G(n)$ be a network, $G(n)$ is said to be small-world if the following three conditions are satisfied:*

- ***Degree property:*** *$\mathcal{O}(\log n)$ is an upper scaling bound of the average degree;*
$$\frac{1}{n}\sum_{0\leq i<n} deg(i) = O(\log n).$$
- ***Clustering property:*** *The clustering coefficient $cc(G(n))$ is bounded away from zero;*
$$\exists\ \varepsilon > 0\ .\forall n : \mathbb{N}^{+}.\ cc(G(n)) > \varepsilon$$
- ***Path length property:*** *$\mathcal{O}(\log n)$ is an upper bound for the characteristic path length;*
$$\bar{l}(G(n)) = \mathcal{O}(\log n)\ \textit{and}\ \operatorname{diam}(G(n)) = \mathcal{O}(\log n)$$

The definition of Cont and Tanimura (2007) is the closest one existing in the literature to our view of a small-world network.

Example 3.2. *It is known that the average path length of a random network scales logarithmically with the system order. Is a random network considered a small-world network?*
Solution. *According to Definition 3.1, the answer is no. We know that the clustering coefficient of the random network is $cc = p$, and if for example $\frac{1}{n} \leq p \leq \frac{\log n}{n}$, the clustering property of small-world networks is violated since the value of cc is not bounded away from zero. Rather it tends to zero as n approaches infinity.*

Example 3.3. *Consider a regular network, with high clustering and diameter bounded by a constant. Is such a network a small-world network?*
Solution. *To satisfy the condition of the diameter mentioned in the example, such a regular network must have an average degree of linear order, i.e., $\bar{deg}(G) = O(n)$ which violates the degree property of small-world network according to Definition 3.1.*

Example 3.4. *A complete network has the minimum value of $\bar{l}$ and the maximum value of cc. Is it a small-world network?*
Solution: *No, complete networks violate the degree property of small-world networks according to Definition 3.1.*

3.2 Small-World Network Models

The small-world network model refers to the construction of a network that meets the properties discussed in Section 3.1. Although small-world networks have been a subject of considerable research, models discussed in literature most often refer to the original one by Watts and Strogatz (1998) which depends on a specific scheme of rewiring links in one dimensional lattices. Another model by Newman and Watts (1999*a*) is where links are added to the lattice instead of rewiring the original links. The Newman and Watts (1999*a*) model is equivalent to the original model in which the same average number of the rewiring links are added. Therefore, the model has the same properties and behaviour. This section studies both models, investigates their properties analytically and by simulation and compares it to the empirical results of Chapter 2.

3.2.1 Watts-Strogatz: The Rewired model of Small-World

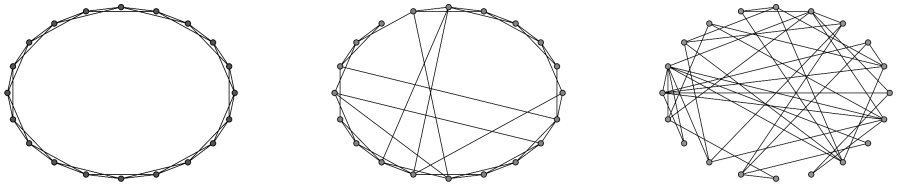
The construction of Watts and Strogatz (1998) has been discussed in (Watts and Strogatz, 1998). Van Steen (2010) explains the construction of Watts and

Strogatz (1998), which we present briefly here.
Watts and Strogatz construction (Van Steen, 2010): Consider a set of n vertices and positive number number $k > 0$. In order to ensure the degree property in Definition 3.1 in the network, choose n and k such that $n \gg k \gg \log(n) \gg 1$.

1. **The structured base network:** Order the n vertices into a ring and connect each vertex to its first k left-hand (clockwise) neighbours and to its k right-hand (counter-clockwise) neighbours. This procedure will construct a network in which each vertex has degree $2k$.

2. **The randomness:** With probability p replace each link (v_1, v_2) with a link (v_1, v_i), where v_i is a randomly chosen vertex other than v_1 and any neighbour of v_1.

The resulting network is called the Watts and Strogatz (1998) network, $WS(n, k, p)$. This construction allows us to interpolate between regularity ($p = 0$) and a random network ($p = 1$) and to study the intermediate region $0 < p < 1$ of the disordered networks. The Watts and Strogatz (1998) construction is depicted in Figure 3.1.



(a) $p = 0$ (b) $0 < p < 1$ (c) $p = 1$

Figure 3.1: The Watts and Strogatz (1998) small-world network model in which the crossover from a regular lattice to a random network is realized. When p is small, the network is close to the original regular lattice; for large enough p the network is similar to a random network.

The Watts and Strogatz (1998) result, which was based on numerical simulation, revealed that, with small probability p of rewiring, the clustering coefficient of the network is still nearly the same as that of the original regular lattice, and the average path length drops quickly to the order of the one in random networks. This result is vitally important since it implies that there is a broad interval of p over which the average path length is almost as small as the one in random networks, yet the clustering coefficient is much bigger than that of a random network. We confirm this result by numerical simulation when we examine the properties of the small-world network models in Section 3.3. We summarize the Watts and Strogatz (1998) result in Conjecture 3.5.

Conjecture 3.5 (Watts and Strogatz (1998))**.** *There exists a class of networks that are highly clustered yet whose average path length scales as in Erdős-Rényi random networks. These networks are called small-world networks by analogy with the small-world experiment (Milgram, 1967).*

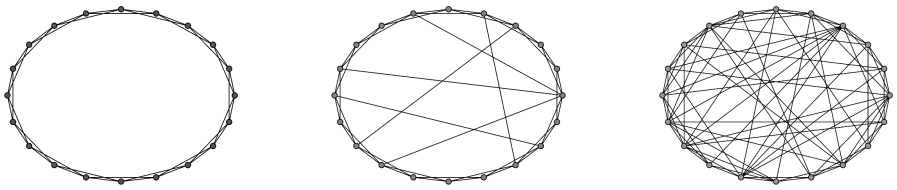
3.2.2 Newman-Watts: Extra links model of small-world

The rewiring procedure in the Watts and Strogatz (1998) model can disconnect the network. In addition, the fact that only one end of each chosen link is rewired, not both, and self-loop and multiple links are not possible, makes it hard to enumerate or average over the networks. Hence, the original model of Watts and Strogatz (1998) is difficult to treat analytically. For that reason, the original model of Watts and Strogatz (1998) has been modified by Newman and Watts (1999*a*). In this variation of the model no links are rewired; rather extra links, often called short-cuts, are added between pairs of vertices chosen randomly; the parameter p governs the density of these short-cuts. To keep the equivalence between this variation and the original Watts and Strogatz (1998) model, each vertex chooses $2k$ vertices at random, and short-cuts are established with each one of the $2k$ vertices with probability p. Therefore the total number of links added to the network by using this procedure is nkp. This way of constructing a network with small-world properties leaves the original lattice intact. The Newman and Watts (1999*a*) construction is depicted in Figure 3.2.

The structured base network: The same as Watts and Strogatz (1998) model explained in Section 3.2.1.

The randomness: add nkp extra random short-cuts.

We examine both models in this chapter, yet the Newman and Watts (1999*a*) construction is our preference due to its ability to be treated analytically.



(a) $p = 0$ (b) $0 < p < 1$ (c) $p = 1$

Figure 3.2: Newman and Watts (1999*a*) Small-World network model, where the small-world network is realized in the intermediate region where $0 < p < 1$.

As we can see from Figure 3.2, this version of the model no longer becomes similar to a random network, even with $p = 1$; rather it becomes a random network on top of the original lattice. However, this turns out not to be a significant problem since most of the interest in the model lies in the limit of large n and small p, where these two models are equivalent. This will become clear when we discuss the properties of the models in Section 3.3. The term Small-world model refers to both cases, although the reader should bear in mind that there are two slightly different constructions of small-world models.

3.3 Small-World Network Properties

The network models that we discussed in the last section have trigged a lot of attention. Many results have been derived for them, and many of their properties have been explored numerically and analytically. In the following, we will summarize the main results regarding the properties of the small-world network models, with particular attention to the average path property, as it is our main concern in this thesis. The results have been also confirmed by simulation in this section.

3.3.1 Degree Distribution

In the Watts and Strogatz (1998) model, for $p = 0$ each vertex has degree $2k$. On the other hand a non-zero value of p introduces disorder into the network, in the form of a non-uniform degree distribution, while maintaining a fixed average degree $2k$. The degree distribution of the rewiring model has been computed in (Barrat and Weigt, 2000) as

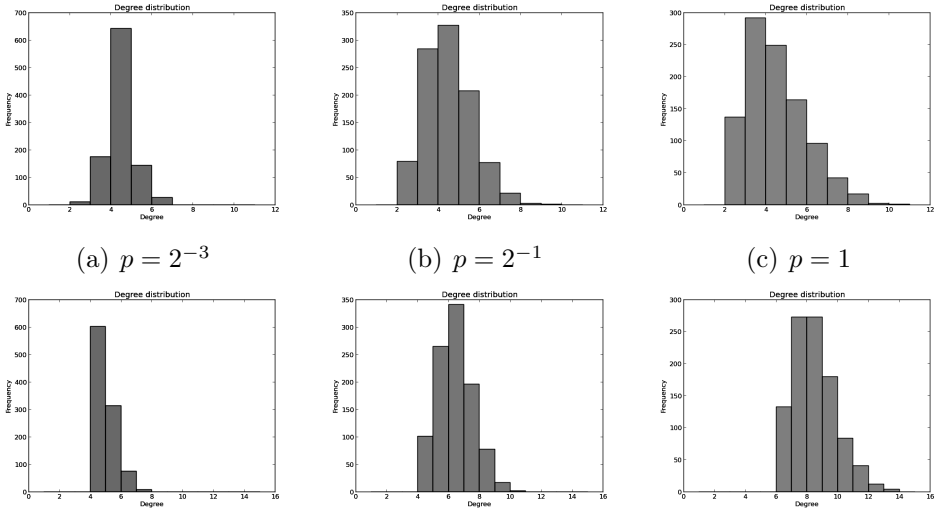
$$\mathbb{P}_j = \sum_{n=0}^{min(j-k,k)} \binom{k}{n} (1-p)^n p^{k-n} \frac{(pk)^{j-k-n}}{(j-k-n)!} e^{-pk},$$

For the Newman and Watts (1999a) model, the degree distribution has been computed in (Newman, 2003b). Every vertex has degree $2k$ on the underlying regular lattice plus a binomially distributed number of links. Hence the degree distribution $\mathbb{P}_j$ is

$$\mathbb{P}_j = \binom{n}{j-2k} \left(\frac{2kp}{n}\right)^{j-2k} \left(1 - \frac{2kp}{n}\right)^{n-j+2k}.$$

Figures 3.3 shows the degree distributions of the Watts and Strogatz (1998) and Newman and Watts (1999a) models. We can see that the distributions of degree in both models are quite similar. With increasing p the degree distribution goes far from the uniform distribution of the lattice. The distribution has an unusual peaked shape, which is significantly different from the empirical degree distribution seen in Chapter 2. Thus, the small-world models

do not meet the property of the degree distribution that has been observed in real-world networks. However, the model was never intended to mimic a real-world degree distribution, but rather it mimics the clustering coefficient and the average path length properties of real-world networks. This has been another direction of research in the science of networks started simultaneously with the small-world networks: degree distribution in real-world networks; the pioneering work was by Barabási and Albert (1999).



(a) $p = 2^{-3}$ (b) $p = 2^{-1}$ (c) $p = 1$

(d) $p = 2^{-3}$ (e) $p = 2^{-1}$ (f) $p = 1$

Figure 3.3: The degree distribution of Watts and Strogatz (1998) model on the top and Newman and Watts (1999*a*) on the bottom, when $n = 1000$ and $k = 4$ with different values of p as shown in the caption. The result is averaged over 10 realizations.

3.3.2 Transitivity or Clustering Coefficient

One of the most important properties of social networks which is also useful in other real-world networks, is transitivity: the likelihood for two neighbours of a given vertex to be neighbours of one another. In a social network, two of your friends are far more likely to be friends of one another than any two people chosen at random from the population. This feature distinguishes social networks from random networks, where all people have the same probability to be friends of one another. Transitivity is a local quantity, and it is quantified by the clustering coefficient as defined in Equation (2.2.3) and (**??**).

For the rewiring model of Watts and Strogatz (1998), and starting with a regular lattice when $p = 0$ we have $cc(0) = \frac{(3k-3)}{(4k-2)}$. When $p > 0$ two neighbours of a given vertex i that were connected at $p = 0$ are still neighbours of vertex i and connected by a link with probability $(1-p)^3$, since there are three links

which need to be intact to maintain the clustering coefficient. Hence, the clustering coefficient given by Barrat and Weigt (2000) is

$$cc(p) = \frac{3(k-1)}{2(2k-1)}(1-p)^3.$$

In the Newman and Watts (1999*a*) model where extra links are added, the clustering coefficient has been calculated in (Newman, 2003*b*) as

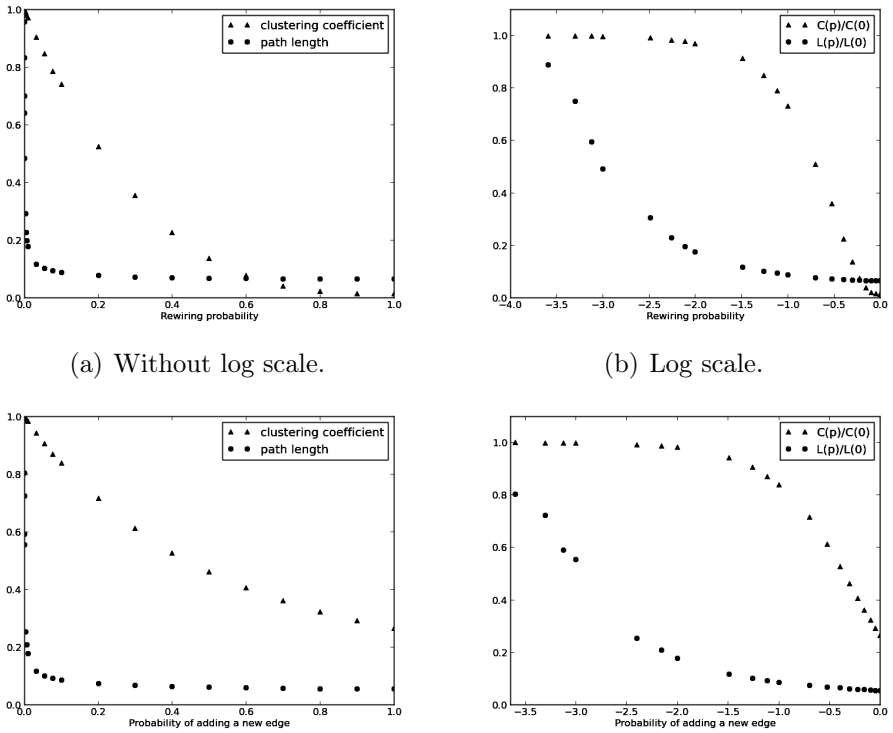
$$cc(p) = \frac{3(k-1)}{2(2k-1)+4kp(p+2)}.$$

As we have explained earlier, the construction allows small-world models to exist for a wide range of values of p, since there is a wide range of values of p that do not change much the clustering coefficient of the network. To view this feature, we have examined a range of small-world networks in both models in Figure 3.4 varying p from very small to relatively large. We compute the clustering coefficient $cc(p)$, and normalize it by $cc(0)$. Likewise, we compute the average path lengths $\bar{l}$, again normalized by dividing by the average path length in case $p = 0$. Figure 3.4 shows that, with increasing p, the average path length drops rapidly, but the clustering coefficient remains relatively high. We notice also that the clustering coefficient of the Newman and Watts (1999*a*) in Figure 3.4(d) model is slightly different from the behaviour of Watts and Strogatz (1998) model in Figure 3.4(b). The difference is more pronounced when we do not take the logarithmic scale as in Figures 3.4(a) and 3.4(c). The drop in the clustering coefficient curve in Newman and Watts (1999*a*) model happens more slowly than the one in Watts and Strogatz (1998) model.

3.3.3 Average Path Length

This section focuses on our main concern, the non-local quantity of a network: the average path length, i.e., the average of the minimum number of links between two vertices. We denoted this quantity by $\bar{l}$ and it is defined in Equation (2.2.1).

Watts and Strogatz (1998) showed by numerical simulation that $\bar{l}$ decreases very rapidly as p increases, and it is comparable to the one in random networks as soon as $p \neq 0$. This result is confirmed by our simulation as Figures 3.4 shows. Also, Figure 3.5 shows, for a random network with $n = 1000$ and $k = 10$, the average path length is about $\bar{l} \approx 3.2$. For the Watts and Strogatz (1998) model, the average path length is only slightly greater, $\bar{l} \approx 3.6$, when the rewiring probability $p = 0.25$, compared with $\bar{l} = 50$ when $p = 0$ for the network with no rewired links at all. Even for $p = 0.0156$, $\bar{l} = 7.4$, which is twice as large as the value for the random networks. The results are averaged over 10 realizations for the randomness. This shows that with fixed n, there is a critical range of p in which a small increase in p results in a dramatic drop



(a) Without log scale.

(b) Log scale.

(c) Without log scale.

(d) With log scale.

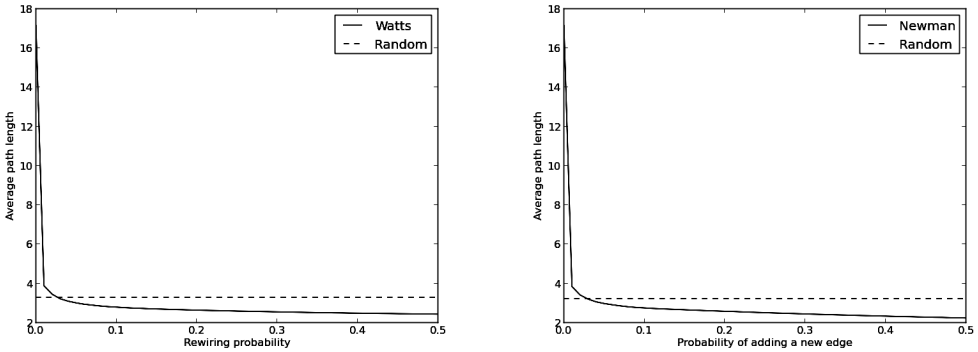
Figure 3.4: The clustering coefficient and average path length in the Watts and Strogatz (1998) model, on the top, and Newman and Watts (1999*a*) model, on the bottom, for increasing values of p. The result is averaged over 10 realizations. Notice that there is a substantial range of values of p in which the value of $\bar{l}$ is low while the value of cc is high. The logarithmic horizontal scale in 3.4(b) and 3.4(d) has been used to resolve the rapid drop in $\bar{l}$.

in $\bar{l}$ and consequently changes the model from being large-world where $\bar{l}$ scales linearly with n to small-world where $\bar{l}$ scales logarithmically with n. The same applies to the Newman and Watts (1999*a*) model.

This sudden change of behaviour of the models from large to small-world has been confirmed by simulation as well as some analytic work which is discussed in the followings.

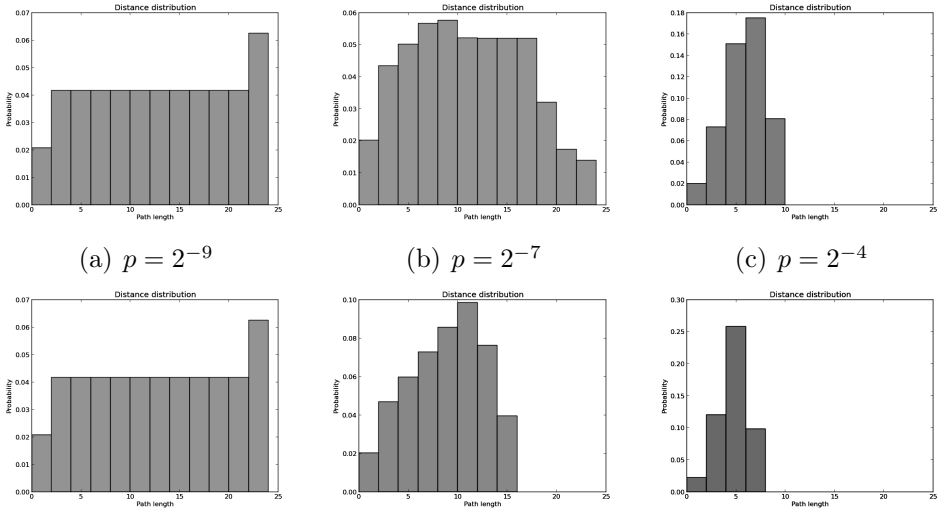
Figure 3.6 shows examples of path length distribution in small-world models. We can see that for very small values of p, the curve is close to a uniform distribution where the base structure does not change from being regular; where we can easily guess that such network is not a small-world. However, as p increases the curve becomes more peaked and skewed left as Figure 3.6(c) shows.

In the following, we shall review some of the important results regarding the average path length. The notation $\sim$ used in the following means: if $f(n) \sim g(n)$, then $\lim_{n\to\infty} \frac{\log(f(n))}{\log(g(n))} = 1$.



(a) Watts-Strogatz model. (b) Newman-Watts model.

Figure 3.5: The average path length in small-world models with fixed $n = 100$ and $k = 10$. A small change of p can make a huge drop in the average path length curve and it becomes comparable to the average path length in random network. The result is averaged over 10 realizations.



(a) $p = 2^{-9}$ (b) $p = 2^{-7}$ (c) $p = 2^{-4}$

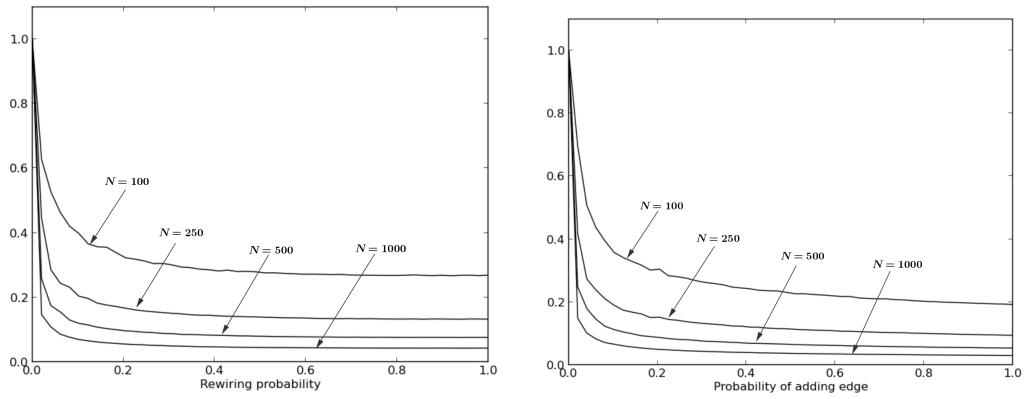
(d) $p = 2^{-9}$ (e) $p = 2^{-7}$ (f) $p = 2^{-4}$

Figure 3.6: The average path length distribution in Watts and Strogatz (1998) model on the top and Newman and Watts (1999*a*) model on the bottom. The curve becomes more and more peaked and skewed left as p grows. The result is averaged over 10 realizations.

Cross-Over Phenomena in small-world networks models

We have seen that with n fixed, $\bar{l}$ decreases rapidly as soon as p is not zero, and there is a critical range of p in which the change from large-world to small-world takes place. Now, we investigate whether the change from large-world to small-world occurs by a transition at a certain finite critical value of p, or

if there is a cross-over phenomenon at any finite value of n, with transition occurring only at $p = 0$. This question was first posed in (Barthélémy and Amaral, 1999). We investigate this question by numerical simulation to study the behaviour of $\bar{l}(n, p)$ varying n and p as in Figure 3.7. We use values of n= 100, 250, 500 and 1000, with different values of p, averaging over 10 realizations for each value of p. It is clear that as n becomes larger, the drop in the curve of $\bar{l}(n, p)$ occurs for smaller values of p, showing that there is no finite critical values of p that can be determined for all n. This is an indication of the cross-over phenomenon of the small-world network model which first has been investigated by numerical simulation in (Barthélémy and Amaral, 1999) and which we are going to discuss in more detail.



(a) Watts-Strogatz model (b) Newman-Watts model

Figure 3.7: The average path length $l(n, p)$ normalized by $\bar{l}(n, 0)$ versus p, for $k = 4$ and $n = 100, 250, 500$ and 1000: the drop in the curve occurs at lower values of p as n grows.

Barthélémy and Amaral (1999) test the hypothesis that the appearance of small world behaviour is a cross-over phenomenon which depends both on network order n and the degree of disorder p. That is, for any value of p, there is a cross-over order $n^*(p)$ below which the network is a large-world and above which it is small-world, i.e., the transition between two regimes takes place at some intermediate system order $n = n^*$. They conjectured that the average path length $\bar{l}$ of the network obeys the scaling form

$$\bar{l}(n, p) = nf\left(\frac{n}{n^*}\right), \tag{3.3.1}$$

where $f(x)$ is a universal scaling function

$$f(x) \sim \begin{cases} \text{constant} & : \text{for} \quad x \ll 1 \\ \log x & : \text{for} \quad x \gg 1 \end{cases} \tag{3.3.2}$$

and

$$f(u \ll 1) \sim u, \text{and } f(u \gg 1) \sim \ln u. \tag{3.3.3}$$

Naively, one would expect to see small-world behaviour when the expected number of the rewired or the extra links is much greater than one, that is $pnk \gg 1$, and the network will be in the large-world regime when $pnk \ll 1$. The reason for that is one would expect the average path length to decrease when some links are successfully rewired. Hence the cross-over should occur for $n * p = \mathcal{O}(1)$, which implies that

$$n^* \sim p^{-\tau} \tag{3.3.4}$$

with $\tau = 1$. This result relies on the fact that the crossover from large to small worlds is obtained with only a small but finite fraction of rewired links.

On the basis of numerical results, Barthélémy and Amaral (1999) further conjectured that $\tau = \frac{2}{3}$. Using a simple physical argument, Barrat (1999) disproved the estimated $\tau = \frac{2}{3}$. He assumed that $\tau < 1$, then taking α such that $\tau < \alpha < 1$, according to Equation (3.3.1) and using the scaling function as defined in (3.3.3) he obtained

$$l(n, n^{-1/\alpha}) \sim n^{\tau/\alpha} f(n^{1-\tau/\alpha}) \sim n^{\tau/\alpha} \ln(n^{1-\tau/\alpha}),$$

since $\tau/\alpha < 1$ and $n^{1-\tau/\alpha} \gg 1$ for large n. However, the average number of rewired or extra links in this case is equal $kn^{1-\frac{1}{\alpha}}$, which goes to zero for large n. The immediate conclusion is that a change in the behaviour of $\bar{l}$ (from $\bar{l} \sim n$ to $\bar{l} \sim n^{\tau/\alpha} \ln(n)$) could occur by the rewiring or adding a vanishing number of links! This is physically not valid, showing that, with the initial assumption, $n^* \sim p^{-\tau}$, the τ cannot be lower than 1.

After Barthélémy and Amaral (1999) had been alerted to the possibility of an error in their estimate of τ, they performed a new calculation using a different algorithm that allowed them to study the system with a bigger order up to 5500, since the reason for the incorrect numerical result was the small order, 1000, which was too small to show the true scaling behaviour.

Re-normalization Group Analysis

We have seen earlier, the characteristic path length n^* diverges according to $n^* \sim p^{-\tau}$. Barrat (1999) has shown using a simple physical argument that τ can not be less than 1. Newman and Watts (1999*a*) have improved the estimation of n^* using an asymptotically exact real-space re-normalization group analysis as well as numerical simulations, and have shown that τ is exactly one. In theoretical physics, the re-normalization group refers to a mathematical apparatus that allows systematic investigation of the changes of a physical system as viewed at different distance scales (Wiki01, 2013). Newman and Watts (1999*a*) have studied the small-world network model using re-normalization

group analysis; they calculated the scaling forms and the single critical exponent describing the behaviour of the model in the critical region using Equations (3.3.1), (3.3.2) and (3.3.4) from (Barthélémy and Amaral, 1999), which proved to be correct.

Assuming that in the critical region n^* has the form of Equation (3.3.4), with the value of τ in unknown, then Equation (3.3.1) can be written as

$$\bar{l} = nf(p^{\tau}n). \tag{3.3.5}$$

We consider the real-space re-normalization group transformation as follows: when $k = 1$, adjacent vertices in a small-world network are blocked to create a one-dimensional lattice of half as many vertices, i.e., the lattice order is assumed to be even, yet the transformation works for any order. Two vertices are connected on the renormalized network if any of the vertices in the original network was connected to any of them including the short-cuts. The transformation is shown in Figure 3.8 for a lattice of order $n = 24$.

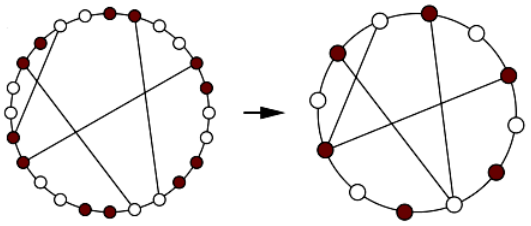


Figure 3.8: Re-normalization transformation for $k = 1$. From (Newman and Watts, 1999a).

The number of short-cuts on the lattice is conserved under the transformation, so the fundamental parameters n and p are re-normalized according to

$$n' = \frac{n}{2}, \qquad p' = 2p. \tag{3.3.6}$$

Since the probability of finding a short-cut between any two vertices i and j is independent of i and j both before and after re-normalization, the transformation generates all possible configurations of short-cuts on the renormalized lattice with the correct probability. In addition, the geometry of the shortest path between any two points is unchanged under the transformation. However, the length of the path is, on average, halved along those portions of the path which run around the perimeter of the ring, and remains the same along the short-cuts. For large n and small p, the portion of the length along the short-cuts tends to zero and so can be neglected. Thus

$$l' = \frac{l}{2}. \tag{3.3.7}$$

Using the relation in Equation (3.3.5) for n', p' and l', then substituting definitions from Equations (3.3.6) and (3.3.7) leads to

$$\begin{aligned} \bar{l}' &= n' f(p'^{\tau} n') \\ \frac{\bar{l}}{2} &= \frac{n}{2} f\left((2p)^{\tau} \frac{n}{2}\right) \\ \bar{l} &= n f\left(2^{\tau-1} p^{\tau} n\right) \end{aligned} \tag{3.3.8}$$

Finally, comparing the relations in Equations (3.3.5) and (3.3.8), equality demands that $2^{\tau} - 1 = 1$, so that $\tau = 1$.

For $k > 1$, Newman and Watts (1999*a*) defined a slightly different re-normalization group as follows: different vertices in a group of order k are grouped, as shown in Figure 3.3.3 for $n = 24$ with $k = 3$. Again the number of short-cuts is preserved under the transformation which gives the following re-normalization equations for the parameters:

$$n' = n/k, \quad p' = k^2 p, \quad k' = 1, \quad l' = l. \tag{3.3.9}$$

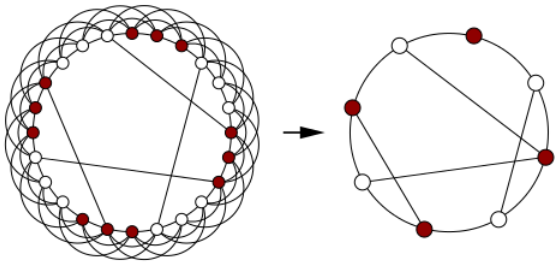


Figure 3.9: Re-normalization transformation for $k = 3$. From (Newman and Watts, 1999*a*).

Note that in the limit of large n and small p, $\bar{l}$ is not affected at all since the number of vertices along the path joining two distant vertices is reduced by a factor k, but the number of vertices that can be traversed in one step is reduced by the same factor and the two cancel out. For the same reasons as before, this transformation is exact in the limit of large n and small p. This transformation can be used to turn any network with $k > 1$ into a corresponding network with $k = 1$, which can then be treated using the arguments given before. We conclude that the correct value of the exponent τ is one. For all values of k and by substituting Equation (3.3.9) into Equation (3.3.1), the small-world network must satisfy the scaling form

$$l = \frac{n}{k} f(pkn). \tag{3.3.10}$$

This form holds under re-normalization for $n' \gg 1$ and $p' \ll 1$ which implies that $n/k \gg 1$ and $k^2 p \ll 1$. The first of these conditions is trivial. The

second condition is necessary to ensure that the average path travelled along short-cuts in the network is small compared to the path travelled around the perimeter of the ring.

Probabilistic Method

In this section we shall discuss the contribution (Flaxman and Frieze, 2004) which is very close to small-world networks and obtains a new result different from what we discussed in Section 3.3.3 but with a motivation other than social networks. The central observation of this paper is that if εn random links are added to any strongly connected directed networks with bounded degree then the resulting network has diameter $\mathcal{O}(\log n)$. The construction of $\mathcal{O}(\log n)$ directed networks in (Flaxman and Frieze, 2004) satisfies the properties of small-world networks discussed in Section 3.1 and is quite similar to our construction of small-world networks which we shall discuss in Chapter 4. Their result holds for a strongly connected directed network D' with degree $\mathcal{O}(n^{\varepsilon/100})$ that is perturbed by adding a random directed network $R \sim \mathbb{D}_{n,\varepsilon/n}$. Here $R \sim \mathbb{D}$ means R is distributed according to distribution $\mathbb{D}$ (there is no connection between this notation and the asymptotic equality defined later), and $\mathbb{D}_{n,p}$ is the distribution of directed network of order n where each possible link appears independently with probability p. The notation $D = D' + R$ means that D is the network formed by taking the union of the links of D' and R, and the vertex sets are the same.

Theorem 3.6. *Let ε be a positive constant with $\varepsilon \leq 1$, let $\triangle = n^{\varepsilon/100}$ and let D' be a strongly connected n-vertex directed network with in-degree and out-degree at most $\triangle$. Let $D = D' + R$ where $R \sim \mathbb{D}_{n,\varepsilon/n}$ Then with high probability the diameter of D is at most $100\varepsilon^{-1} \log n$.*

Proof. We sketch the proof which can be found in (Flaxman and Frieze, 2004). The idea of the proof is to show that with high probability D contains short paths of special form, alternating between some links from D' and random links from R. They use a similar approach to (Bollobás and Chung, 1988) where sets of vertices are grouped if they are within distance $d = 5\varepsilon^{-1}$ in the base network D'. Since the network is directed in this case, the direction of the links is considered. Then random links from R are found to link between these groups. Vertices are called useful and the created set denoted by U if it is not within distance d of any vertex which is previously placed in any set. Algorithm 3.3.3 explains the procedure of creating the sets of vertices denoted by S_i and T_i. The notation ${N_d}^+(S)$ denotes the set of vertices reachable in D' in at most d steps starting from some vertex of S, ${N_d}^-(S)$ denotes the set of vertices from which some vertex of S is reachable in at most d steps in D', and $N_d(S) = {N_d}^+(S) + {N_d}^-(S)$. Finally let $l = \lceil \log_2 n \rceil$. Using probabilistic analysis we can see that when the algorithm *GenerateSets* halts we have

$$\mathbb{P}(|S_i| \leq n^{2/3} \text{ or } |T_j| \leq n^{2/3}) = o(n^{-2}),$$

Algorithm 1 Procedure *GenerateSets*$[(s,t)]$-path P

S_0 :=first $32\varepsilon^{-1}\log n$ vertices of P
T_0 :=last $32\varepsilon^{-1}\log n$ vertices of P
$U := V \backslash N_d(S_0 \cup T_0)$
$i = 0$
$j = 0$
while $(|S_i| \leq n^{2/3}$ and $i \leq l)$ or $(|T_j| \leq n^{2/3}$ and $j \leq l)$ **do**
if $|S_i| \leq n^{2/3}$ and $i \leq l$ **then**
$S_{i+1} := \emptyset$
for $\forall s' \in S_i$ **do**
if $|S_{i+1}| \leq n^{2/3}$ and there exists $s'' \in U$ such that $(s', s'') \in R$ **then**
$S_{i+1} := S_{i+1} \cup N_d^+(\{s''\})$
$U := U \setminus N_d(S_{i+1})$
end if
end for
$i : i + 1$
end if
if $|T_j| \leq n^{2/3}$ and $j \leq l$ **then**
$T_{j+1} = \emptyset$
for $\forall t' \in T_j$ **do**
if $|T_{j+1}| \leq n^{2/3}$ and there exists $t'' \in U$ such that $(t'', t') \in R$ **then**
$T_{j+1} := T_{j+1} \cup N_d^-(\{t''\})$
$U := U \setminus N_d(T_{j+1})$
end if
end for
$j := j + 1$
end if
end while

this implies that with high probability *GenerateSets* halts when $l = \lceil \log_2 n \rceil$. Now, to finish the short path from s to t, random links of R between S_i and T_j are generated,

$$\mathbb{P}((R \cap (S_i \times T_j) = \phi) | (|S_i| \leq n^{2/3} \wedge |T_j| \leq n^{2/3})) \leq o(n^{-2}).$$

Putting all the pieces together, we have an (s,t)-path consisting of a path of length at most $32\varepsilon^{-1}\log n$, followed by at most $2l$ paths in D of length at most $d+1$ joined by links from R and finishing with a path of length at most $32\varepsilon^{-1}\log n$, for a total length which is less than $100\varepsilon^{-1}\log n$ as numerical calculation shows. Since there are only $n(n-1)$ choices for (s,t) the theorem follows by the union bound. □

3.4 Conclusion

In this chapter, three constructions of small-world models have been studied. The first two are the most popular and they are equivalent: Watts-Strogatz and Newman-Watts small-world network models. Regarding the average path length $\bar{l}$, our main interest in this thesis, the cross-over from large-world to small-world occurs with fixed n at $p \sim n^{-1/\tau}$. The onset of small-world behaviour occurs if

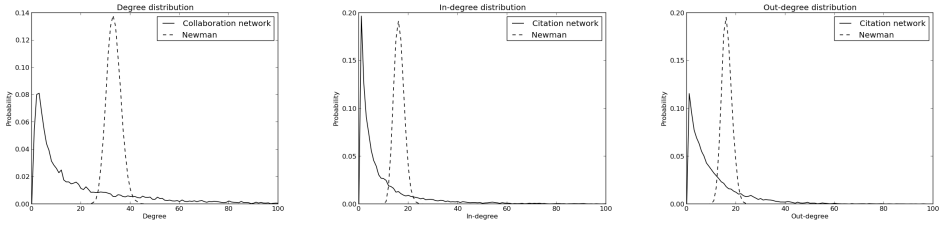
$$\lim_{n\to\infty} \frac{\log p}{\log n} = \frac{-1}{\tau}, \text{then } \tau = 1.$$

This result leaves the threshold area large, where we can decide whether the network in this area is small-world or not. Note that the network is a small-world when $\tau > 1$, and large-world when $\tau < 1$, otherwise we can not decide.

Although these models are probably generic for many large, sparse networks found in nature, they are not realistic representations of real-world networks since they depend on the lattice structure as base network to construct small-world network models. In addition, they do not have the scale-free degree distribution observed in real-world networks. Table 3.1 compares the (Newman and Watts, 1999a) model with the empirical results of real-world networks.

Network	$\|V\|$	E	E_{NW}	z	p	z_{NW}	$\bar{l}$	$\bar{l}_{NW}$	cc	cc_{NW}
Collaboration network	17903	197031	197052	22.01	0.00055	22.01	4.192	26.58	0.63	0.71
Citation network	12711	139981	140793	22.02	0.007	22.15	14.4	7.92	0.28	0.7
Facebook network	7532	62079	60340	16.48	0.001	16.0	5.91	20.9	0.47	0.698

Table 3.1: Comparison between Random networks and real-world networks.



(a) Collaboration vs. Newman-Watts network (b) Citation in-degree vs. Newman-Watts network (c) Citation out-degree vs. Newman-Watts network

Figure 3.10: Comparison between Newman and Watts (1999a) and real-world networks regarding the degree distribution.

(Flaxman and Frieze, 2004) is the third construction we have discussed in this chapter. Although, this construction is more reasonable since it does not depend on lattice structure, this construction is not made as a social-network

model, hence, does not guarantee the high transitivity observed in small-world networks. They obtain a directed network with diameter $\mathcal{O}(\log n)$ when $p = \frac{\varepsilon}{n}$.

The question remains: is it possible to obtain better results regarding the average path length $\bar{l}$ for small-world network models?

Chapter 4

Average Path Length in Small-World Networks

As we have seen, models of small-world networks combine structured base topology and randomness represented by probabilistically rewiring of links as in Watts and Strogatz (1998) model, or probabilistically adding extra links between each pair of vertices of the base structure as in (Newman and Watts, 1999*a*). In this chapter we consider a model in which randomness is provided by the Erdős-Rényi random network model. In contrast with most other complex network models where the analytical solution is hardly possible, the Erdős-Rényi random network model is one of the oldest and best studied network models, and possesses a considerable advantage of being analytically solvable in many of its average properties. We propose a solvable small-world model in Section 4.1. In addition, this chapter shows the vitally important role of random network evolution and structure in obtaining theoretical results for small-world networks regarding its average path length.

This chapter is divided into two parts: In the first one we construct a small-world network with diameter $\mathcal{O}(\log n)$, and we show that this happens when pn is bounded away from zero, i.e., $\exists\, \varepsilon > 0\ .\ \forall n : \mathbb{N}^+ . pn > \varepsilon$. The threshold of the giant component in a random network is the main tool used to obtain this result. In the second part we show that εn is the minimum average number of links that should be added to a one dimensional lattice so that the average path length has a logarithmic scale. We explain the behaviour of $\bar{l}$ when pn tends to zero and show that in this case $\bar{l} > \omega(n) \log n$ as $n \to \infty$, where $\omega(n)$ is a function $\omega(n) \to \infty$. We rely on the structure of random networks when $pn \to 0$ to obtain this result.

4.1 Small-world Networks

In this section, we construct a solvable model of small-world networks based on Erdős-Rényi random networks, which allow us to develop mathematical tools

to treat small-world networks analytically using results of random network theory.

Assume a one dimensional lattice $C(n)$ with vertices $\{v_i|0 \leq i < n\}$ and with consecutive vertices adjacent, and an Erdős-Rényi random network $R(n,p)$ with vertices $\{r_i|0 \leq i < n\}$. We superpose the two networks by identifying their vertices using a matching operation

$$M : R(n,p) \to C(n)$$

which is one of the $n!$ bijection from $V(C(n))$ to $V(R(n,p))$. The bijection function M is used to update the structure of the lattice $C(n)$ to give the disordered network $C(n,p)$ with the same vertices, i.e. $(V(C(n,p)) = V(C(n)))$, but enhanced edges

$$E(C(n,p)) = E(C(n)) \cup \{(M(u), M(v)) \in V(C(n)) \times V(C(n)) | (u,v) \in E(R(n,p))\}.$$

Note that, we use the term "disordered network" instead of small-world network, because small-world network is quantified by the logarithmic scale of the average path length, which is dependent on the value of p of the random network. In the subsequent sections we will show that the disordered network $C(n,p)$ is a small-world network when pn is bounded away form zero. On the contrary, when pn tends to zero the network $C(n,p)$ does not have the logarithmic scale of the average path length, hence, it is not a small-world network yet.

4.2 The diameter of small-world networks

In this section we show that when pn is bounded away from zero, the diameter of $C(n,p)$ is bounded above by $\mathcal{O}(\log n)$. Our method is to use the threshold of the giant component of the random network $R(n,p)$.

We use "Vinogradov" notation: $f(n) \gg g(n)$, which means there exist a constant $c > 0$ such that $f(n) > cg(n)$ as $n \to \infty$.

In the following proof we use a simple case when the base network is a cycle. The result will be generalized for the ring later.

Lemma 4.1. *For a constant $\delta > 0$, if δn vertices are deleted uniformly at random from the cycle $C(n)$, then the maximum arc length left is at most k with probability at least $1 - ne^{-\delta k}$.*

Proof. If S is an arc of length $k-1$ in the cycle, then we have

$$\mathbb{P}(S \text{ is uncut}) = (1-\delta)^k.$$

The event "the arc S is uncut" means that all the k vertices of S are left. Since there are exactly n such arcs in $C(n)$, then by the union bound and using the inequality $1 - x \leq e^{-x}$ we have

$$\mathbb{P}(\text{ at least one arc of length } k-1 \text{ uncut}) \leq n(1-\delta)^k \leq ne^{-\delta k}$$

Hence, the probability that there is no arc of length $k-1$ left is

$$\mathbb{P}(\text{no arc of length } k-1) \geq 1 - ne^{-\delta k}.$$

which proves the lemma. □

The following theorem shows that $\mathcal{O}(\log n)$ is an upper bound for small-world networks $C(n,p)$ when np is bounded away from zero.

Theorem 4.2. *For a constant $\varepsilon > 0$,*

$$\operatorname{diam}\Big(C(n, \tfrac{\varepsilon}{n})\Big) = \mathcal{O}(\log n).$$

Proof. Consider $C(n,p)$ as defined in Section 4.1 using the random network $R(n, \frac{\varepsilon}{n})$ where $\varepsilon > 0$.

Consider the case when $\varepsilon > 1$. Recalling from Section 2.3.2 a.a.s, $R(n,p)$ has a unique giant component of order n and diameter $\mathcal{O}(\log n)$. Conditioning the fact that the random network $R(n,p)$ has this property w.h.p., denote the giant component of $R(n,p)$ by C_1 and assume that $|C_1| = \delta n$. According to the definition of the model in 4.1 the vertices of the giant component of the random network are uniformly distributed on the cycle. Using Lemma 4.1 for $k = \frac{2}{\delta}\log n$, we deduce that with probability at least $1 - \frac{1}{n}$ every arc of length more than $\frac{2}{\delta}\log n$ contains at least a vertex of C_1.

Now, for any pair of vertices u and v in the cycle, we can find vertices u' and v' in C_1 such that $d(u,u')$ and $d(v,v')$ are both at most $\frac{2}{\delta}\log n$. Since the diameter of C_1 is at most of order $\log n$, we deduce that the diameter of $C(n, \frac{\varepsilon}{n})$ is also of order at most $\log n$.

If $\varepsilon \leq 1$, then we partition the cycle into arcs of k vertices where k is a fixed positive integer with

$$k > \frac{1}{\varepsilon}. \tag{4.2.1}$$

Note that in this grouping we assume that n is divisible by k. If n is not divisible by k then we would have one arc of order at most $k-1$, and our argument remains valid even if one vertex in the $C(n)$ behaves differently from the other vertices provided that n is large enough.

We can think of these arcs as vertices of a cycle of order $\frac{n}{k}$. We say that two arcs are adjacent if there is at least one edge connecting them. So this model is reduced to $C(N,q)$ with

$$N = \frac{n}{k} \qquad \text{and} \qquad q = 1 - \Big(1 - \frac{\varepsilon}{n}\Big)^{k^2} \sim \frac{\varepsilon k}{N}.$$

Thus, since $\varepsilon k > 1$, we can apply the same argument as before to deduce that in this case the diameter is also of order at most $\log n$, which completes the proof. □

4.3 The Average Path Length

We have seen in Section 4.2 when pn is bounded away from zero, we have $\mathcal{O}(\log n)$ as an upper bound for the diameter, which also bounds the average path length $\bar{l}$ below by $\mathcal{O}(\log n)$. In this section, we answer our research question mentioned in Section 1.2 and show that εn is the average minimum number of extra random links that can be added to a one dimensional lattice such that the average path length scales logarithmically with the network order. We study the average path length $\bar{l}(C(n,p))$ when pn tends to zero and we show that the network is not a small-world, i.e., the average path length is not of logarithmic order. We make use of the random network structure when pn tends to zero to derive our result.

4.4 Random Network Structure

In this section we study the structure of the random network $R(n, \frac{\varepsilon}{n})$ on the assumption that

$$\varepsilon \gg n^{-\delta}, \quad \text{for all} \quad \delta > 0. \tag{4.4.1}$$

The other cases have been already covered in the literature as discussed in Chapter 3.

As we have seen in Section 2.3.2, asymptotically almost surely when pn tends to zero, most of the components of random networks are trees and the largest component of $R(n,p)$ has order $\mathcal{O}(\log n)$. Assuming that $R(n,p)$ has this property, let N_k be the number of the tree components of $R(n,p)$ of order $k \geq 1$ and I_S is a random variable such that

$$I_S = \begin{cases} 1 & \text{if } S \text{ is the vertex set of a tree component in } R(n,p) \text{ and } |S| = k \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$N_k = \sum_{S \subseteq V(R(n,p)), |S|=k} I_S, \tag{4.4.2}$$

and

$$\mathbb{E}(N_k) = \sum_{S \subseteq V(R(n,p)), |S|=k} \mathbb{E}(I_S). \tag{4.4.3}$$

Also,

$$\begin{aligned} {N_k}^2 &= \left(\sum_{S \subseteq V(R(n,p)), |S|=k} I_S \right)^2 \\ &= \sum_{S \subseteq V(R(n,p)), |S|=k} {I_S}^2 + \sum_{\substack{|S_1|=k, |S_2|=k \\ \text{and } S_1 \neq S_2}} I_{S_1} I_{S_2} \\ &= N_k + \sum_{\substack{|S_1|=k, |S_2|=k \\ \text{and } S_1 \neq S_2}} I_{S_1} I_{S_2}. \end{aligned} \tag{4.4.4}$$

Therefore,

$$\mathbb{E}({N_k}^2) = \mathbb{E}(N_k) + \sum_{\substack{|S_1|=k, |S_2|=k \\ \text{and} S_1 \neq S_2}} \mathbb{E}(I_{S_1} I_{S_2}). \tag{4.4.5}$$

Lemma 4.3. *W.h.p.*

$$N_1 = n + \mathcal{O}(\varepsilon n)$$

Proof. The expected number N_1 of trees of order one (i.e., isolated vertices in $R(n,p)$) is

$$\mathbb{E}(N_1) = n(1-p)^{n-1}.$$

So, using the binomial theorem,

$$\mathbb{E}(N_1) = n + \mathcal{O}(\varepsilon n). \tag{4.4.6}$$

Given that $S_1, S_2 \in V(R(n,p))$, such that $S_1 \neq S_2$,

$$\mathbb{P}(S_1 \text{ and } S_2 \text{ are isolted vertices in } R(n,p)) = (1-p)^{2(n-1)-1}.$$

Hence, using Equation (4.4.5) for $k = 1$ and

$$\mathbb{E}({N_1}^2) = \mathbb{E}(N_1) + n(n-1)(1-p)^{2(n-1)-1}.$$

The variance, $Var(N_1)$, is

$$\begin{aligned} Var(N_1) &= \mathbb{E}({N_1}^2) - \mathbb{E}(N_1)^2 \\ &= \mathbb{E}(N_1) + n(n-1)(1-p)^{2(n-1)-1} - n^2(1-p)^{2(n-1)} \\ &= \mathbb{E}(N_1) + n(1-p)^{2(n-1)-1}[(n-1) - n(1-p)] \\ &= \mathbb{E}(N_1) + n(1-p)^{2(n-1)-1}[np - 1] \\ &= \mathbb{E}(N_1) - n(1-p)^{2(n-1)-1} + n^2 p(1-p)^{2n-1}, \end{aligned}$$

since $p = \frac{\varepsilon}{n}$,

$$\begin{aligned} Var(N_1) &= n + \mathcal{O}(\varepsilon n) - (n + \mathcal{O}(\varepsilon n)) + \mathcal{O}(\varepsilon n) \\ &= \mathcal{O}(\varepsilon n). \end{aligned} \tag{4.4.7}$$

By Chebyshev's inequality for any real number $t > 0$, we obtain

$$\mathbb{P}\left(|N_1 - \mathbb{E}(N_1)| > t\sqrt{Var(N_1)}\right) \leq \frac{1}{t^2}. \tag{4.4.8}$$

Using the variance value in Equation (4.4.7), and letting $t = \frac{n^{1/4}}{\sqrt{\log n}}$, there exists a constant c_1 such that

$$\mathbb{P}\left(|N_1 - \mathbb{E}(N_1)| > n^{1/4}\sqrt{\frac{c_1 \varepsilon n}{\log n}}\right) \leq \frac{\log n}{n^{1/2}}.$$

Thus,

$$\mathbb{P}\left(|N_1 - \mathbb{E}(N_1)| \leq n^{1/4}\sqrt{\frac{c_1 \varepsilon n}{\log n}}\right) > 1 - \frac{\log n}{n^{1/2}}.$$

Hence, N_1 is concentrated around its mean, i.e.,

$$N_1 = \mathbb{E}(N_1) + \mathcal{O}(n^{1/4}\sqrt{\frac{\varepsilon n}{\log n}})$$

w.h.p. By the condition on ε in (4.4.1) and the estimated value of $\mathbb{E}(N_1)$ in (4.4.6) we have, w.h.p.,

$$N_1 = n + \mathcal{O}(\varepsilon n). \tag{4.4.9}$$

□

Lemma 4.4. *a) For $k = 2$, w.h.p.*

$$N_k = \mathcal{O}(\varepsilon n)$$

b) For $k \geq 3$

i) if $\mathbb{E}(N_k) \geq \sqrt{n}$, then w.h.p.

$$N_k < 2\mathbb{E}(N_k)$$

ii) if $\mathbb{E}(N_k) < \sqrt{n}$, then w.h.p.

$$N_k < n^{\frac{1}{4}}\mathbb{E}(N_k).$$

Proof. In a manner similar to the proof of Lemma 4.3. For any $S \subseteq V(R(n,p))$ such that $|S| = k$ and $k \geq 2$

$$\mathbb{P}\left(S \text{ is the vertex set of a tree component in } R(n,p) \text{ and } |S| = k\right) = k^{k-2}\, p^{k-1}\, (1-p)^{\binom{k}{2}-(k-1)+k(n-k)},$$

where k^{k-2} represents Cayley's formula of the number of trees on k labelled vertices (Harris *et al.*, 2009, p. 43-45), p^{k-1} means that $k-1$ links should be

present in the tree and $(1-p)^{\binom{k}{2}-(k-1)+k(n-k)}$ for the tree to be an isolated connected component. Hence,

$$\mathbb{E}(N_k) = \binom{n}{k} k^{k-2}\ p^{k-1}\ (1-p)^{\binom{k}{2}-(k-1)+k(n-k)}. \tag{4.4.10}$$

Using Equation (4.4.5) for given S_1 and S_2 such that $|S_1| = k$, $|S_2| = k$ and $S_1 \neq S_2$. If $S_1 \cap S_2 \neq \emptyset$, then

$\mathbb{P}(S_1$ and S_2 are the vertex set of a tree components on $R(n,p)$ of order $k) = 0$.

If $S_1 \cap S_2 = \emptyset$ then

$\mathbb{P}(S_1$ and S_2 are the vertex set of a tree component on $R(n,p)$ of order $k) = k^{2(k-2)}p^{2(k-1)}(1-p)^{2\binom{k}{2}-2(k-1)+2k(n-k)-k^2}$.

Hence,

$$\begin{aligned}
\mathbb{E}({N_k}^2) &= \binom{n}{k,k} k^{2(k-2)}\ p^{2(k-1)}\ (1-p)^{2\binom{k}{2}-2(k-1)+2k(n-k)-k^2} + \mathbb{E}(N_k) \\
&= \mathbb{E}(N_k)^2 \left(\frac{\binom{n}{k,k}(1-p)^{-k^2}}{\binom{n}{k}^2} + \mathbb{E}(N_k)^{-1} \right) \\
&= \mathbb{E}(N_k)^2 \left(\frac{((n-k)!)^2}{(n-2k)!n!}(1-p)^{-k^2} + \mathbb{E}(N_k)^{-1} \right) \\
&= \mathbb{E}(N_k)^2 \left((1-p)^{-k^2} \prod_{j=0}^{k-1} (1 - \frac{k}{n-j}) + \mathbb{E}(N_k)^{-1} \right) \\
&= \mathbb{E}(N_k)^2 \left(1 + \mathcal{O}\left(\frac{k^2}{n} + \mathbb{E}(N_k)^{-1} \right) \right).
\end{aligned} \tag{4.4.11}$$

For $k = 2$, by substituting $k = 2$ in Equation (4.4.10), the number of the trees of order two, i.e., isolated edge in the network, is

$$\mathbb{E}(N_2) = \binom{n}{2} p(1-p)^{2(n-2)} = \Theta(\varepsilon n). \tag{4.4.12}$$

Using the formula of $\mathbb{E}({N_k}^2)$ in Equation (4.4.11), we get

$$\mathbb{E}({N_2}^2) = (\mathbb{E}(N_2))^2 \left(1 + \mathcal{O}\left(\frac{1}{n} + \frac{1}{\varepsilon n} \right) \right),$$

and

$$Var(N_2) = \mathbb{E}(N_2)^2 \left(1 + \mathcal{O}\left(\frac{1}{\varepsilon n} \right) \right) - \mathbb{E}(N_2)^2 \tag{4.4.13}$$

$$= \mathbb{E}(N_2)^2 \mathcal{O}\left(\frac{1}{\varepsilon n} \right). \tag{4.4.14}$$

Using Chebyshev's inequality, for any real number $t > 0$

$$\mathbb{P}(N_k \geq t\mathbb{E}(N_k)) \leq \frac{Var(N_k)}{(t-1)^2\mathbb{E}(N_k)^2}.$$

Setting $t = 2$ and using the variance of N_2 in Equation (4.4.13), there exists a constant c_2 such that

$$\mathbb{P}\left(N_2 \geq 2\mathbb{E}(N_2)\right) \leq \frac{c_2}{\varepsilon n},$$

which means that

$$\mathbb{P}\left(N_2 < 2\mathbb{E}(N_2)\right) > 1 - \frac{c_2}{\varepsilon n},$$

using the expectation of N_2 in Equation (4.4.12) which implies that w.h.p

$$N_2 < 2\mathbb{E}(N_2) = \Theta(\varepsilon n). \tag{4.4.15}$$

For all $k \geq 3$, using Stirling's approximation (Mitzenmacher and Upfal, 2005), which states that

$$\log(n!) = n\log n - n + \frac{1}{2}\log n + \mathcal{O}(1),$$

then

$$\begin{aligned}
&\log\binom{n}{k} = k\log n - k\log k + k - \frac{1}{2}\log k + \mathcal{O}(1)\\
&\log(k^{k-2}) = k\log k - 2\log k\\
&\log(p^{k-1}) = (k-1)\log\varepsilon - k\log n + \log n \quad \text{as} \quad p = \frac{\varepsilon}{n}\\
&\log(1-p)^{\binom{k}{2}-(k-1)+k(n-k)} = -\varepsilon k + \mathcal{O}(1)
\end{aligned}$$

Note that all those asymptotic estimates only hold when $k = \mathcal{O}\sqrt{n}$, which is the case here. Putting all pieces together on Equation (4.4.10) leads to

$$\begin{aligned}
\log\mathbb{E}(N_k) &= \log n + k(1-\varepsilon) + (k-1)\log\varepsilon - \frac{5}{2}\log k + O(1)\\
&= \log n + (k-1)(1-\varepsilon+\log\varepsilon) - \frac{5}{2}\log k + \mathcal{O}(1).
\end{aligned}$$

Since $(1-\varepsilon)$ is absorbed in $\mathcal{O}(1)$,

$$\mathbb{E}(N_k) = \Theta\left(\frac{n}{k^{5/2}}e^{(k-1)(1-\varepsilon+\log\varepsilon)}\right). \tag{4.4.16}$$

Knowing that $1 - \varepsilon + \log\varepsilon \sim \log\varepsilon$ as $\varepsilon \to 0$,

$$1 - \varepsilon + \log\varepsilon \leq \frac{1}{2}\log\varepsilon - 1,$$

$$(k-1)(1-\varepsilon+\log\varepsilon) \leq \frac{k-1}{2}\log\varepsilon + 1 - k \leq \log\varepsilon - k + 1. \tag{4.4.17}$$

Therefore, we can use Equation (4.4.17) in Equation (4.4.16) to obtain

$$\mathbb{E}(N_k) = \mathcal{O}\left(\frac{\varepsilon n}{k^{5/2}}e^{-k}\right). \tag{4.4.18}$$

Using Chebyshev's inequality, for any real number $t > 0$

$$\mathbb{P}(N_k \geq t\mathbb{E}(N_k)) \leq \frac{Var(N_k)}{(t-1)^2\mathbb{E}(N_k)^2}, \tag{4.4.19}$$

then, substituting the value of $\mathbb{E}({N_k}^2)$ in Equation (4.4.11)

$$Var(N_k) = \mathbb{E}(N_k)^2\left(1 + \mathcal{O}\left(\frac{k^2}{n} + \mathbb{E}(N_k)^{-1}\right)\right) - \mathbb{E}(N_k)^2. \tag{4.4.20}$$

Using the variance in Equation (4.4.20) into Equation (4.4.19) we obtain

$$\mathbb{P}(N_k \geq t\mathbb{E}(N_k)) \leq \frac{\mathcal{O}(\frac{k^2}{n} + \mathbb{E}(N_k)^{-1})}{(t-1)^2}. \tag{4.4.21}$$

Consider the case where

$$\frac{n}{k^{5/2}}e^{(k-1)(1-\varepsilon+\log\varepsilon)} \geq \sqrt{n}.$$

Setting $t = 2$, using the above assumption and by substituting the value of $\mathbb{E}(N_k)$ in Equation (4.4.16) into the probability in Equation (4.4.21), we obtain

$$\mathbb{P}(N_k \geq 2\mathbb{E}(N_k)) \ll \frac{1}{\sqrt{n}}. \tag{4.4.22}$$

Then, w.h.p.

$$N_k < 2\mathbb{E}(N_k). \tag{4.4.23}$$

Otherwise, we do the same by setting $t = n^{\frac{1}{4}}$ to obtain

$$\mathbb{P}(N_k \geq n^{\frac{1}{4}}\mathbb{E}(N_k)) \ll \frac{1}{\sqrt{n}}, \tag{4.4.24}$$

$$N_k < n^{1/4}\mathbb{E}(N_k) \tag{4.4.25}$$

w.h.p., i.e, $1 - \frac{1}{\sqrt{n}}$. □

After estimating the number of tree components in a random network $R(n,p)$ when np tends to zero in the last two lemmas, now we need to estimate the average order of the component of a random vertex, $e(R(n,p))$. Toward that end, we define

$$e(R(n,p)) = \sum_{v \in V(R(n,p))} \frac{|C(v)|}{n}, \tag{4.4.26}$$

where $C(v)$ is the connected component of $R(n,p)$ containing v. Given that the number of tree component in $R(n,p)$ is N_k and the order of the largest component is $K\log n$ where $K>0$

$$\begin{aligned} e(R(n,p)) &= \sum_{k=1}^{K\log n} k\mathbb{P}(|C(v)|=k) \\ &= \sum_{k=1}^{K\log n} k\left(\frac{kN_k}{n}\right) \\ &= \sum_{k=1}^{K\log n} \frac{k^2N_k}{n} \end{aligned} \tag{4.4.27}$$

Note that, the average order of the component of a random vertex is different from the normal averaging. The average order of the component in (4.4.27), estimates the order of the component of a random vertex. The standard averaging estimates the order of the component in the whole network. See Figure 4.1.

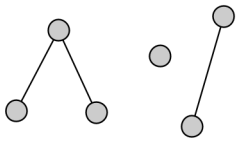


Figure 4.1: The difference between the average order of the component in (4.4.27) and the standard averaging. The formula in (4.4.27) estimates the average order of the component of the random vertex of this random network by $\frac{3+3+3+1+2+2}{6}=2.3$ while the standard averaging estimates the order of the component by $\frac{3+1+2}{3}=2$.

Lemma 4.5. *The average order of the component of a random vertex in* $R(n,\frac{\varepsilon}{n})$ *w.h.p. satisfies*

$$e(R(n,p)) \leq 1+c_1\varepsilon,$$

for some $c_1>0$, *where* ε *satisfies* (4.4.1).

Proof. Using the definition of the average order of the component of a random vertex in $R(n,p)$ given in Equation (4.4.27)

$$e(R(n,p)) = \frac{N_1}{n} + \sum_{k=2}^{K\log n} \frac{k^2N_k}{n},$$

by substituting the value of N_1 in Lemma 4.3, we obtain

$$e(R(n,p)) = 1+\mathcal{O}(\varepsilon) + \sum_{k=2}^{K\log n} \frac{k^2N_k}{n}. \tag{4.4.28}$$

If $k \geq 2$ satisfies the first case in Equation (4.4.23) then

$$\frac{k^2 N_k}{n} < \frac{2k^2 \mathbb{E}(N_k)}{n}.$$

By substituting the value of $\mathbb{E}(N_k)$ in Equation (4.4.18)

$$\frac{k^2 N_k}{n} < \mathcal{O}\left(\frac{\varepsilon e^{-k}}{\sqrt{k}}\right) \tag{4.4.29}$$

with probability at least $1 - \frac{1}{\sqrt{n}}$. By the union bound all such k satisfy the above inequalities with probability at least $1 - O(\frac{\log n}{\sqrt{n}})$. Similarly the second case in Equation (4.4.24)

$$\frac{k^2 N_k}{n} < \frac{2k^2 n^{1/4} \mathbb{E}(N_k)}{n} = \mathcal{O}\left(\frac{(\log n)^2}{n^{3/4}}\right) \tag{4.4.30}$$

with probability at least $1 - \frac{1}{\sqrt{n}}$. By the union bound all $k \geq 2$ satisfy the above inequalities with probability at least $1 - \mathcal{O}(\frac{\log n}{\sqrt{n}})$.

Applying the union bound again,

$$e(R(n,p)) = 1 + \mathcal{O}\left(\varepsilon \sum_{k \geq 2} \frac{e^{-k}}{\sqrt{k}} + \frac{(\log n)^3}{n^{3/4}}\right),$$

with probability $1 - \mathcal{O}(\frac{\log n}{\sqrt{n}})$. □

4.5 Random Process

Random processes are most often used to investigate theoretical questions about random networks. In this section we will investigate the random process of constructing the small-world network model explained in Section 4.1.

We use the term "identified vertex" if the vertex has been superposed to a vertex from the random network $R(n,p)$, otherwise it is not an identified vertex.

Step 0: All the vertices are not identified yet. Choose a random vertex v_0 from the cycle $C(n)$; and we say v_0 is ***open*** vertex.

Step 1: Choose a random vertex w from the random network $R(n,p)$, identify w with v_0 which has been chosen at step 0. If vertex w has other vertices in its component then identify them with randomly chosen vertices from the cycle. The vertices that have been identified are ***closed*** vertices. The neighbours of the closed vertices which have not been identified yet are open vertices now.

Step 2: Restart the process with the open vertices in step 1.

The process continues this way and ends when all vertices have been identified as closed vertices. The resulting network at the end of the process is $C(n,p)$ which generates a disordered network according to the model in 4.1. The vertex v_0 that is picked at random at step 0 will not affect the structure of the generated network $C(n,p)$, however it plays a very important role in our analytical derivation.

Figure 4.2 explains the random process of generating disordered network $C(n,p)$ with $n = 20$. At ***step 0*** all vertices are not identified yet and there is only one open vertex. On the right hand side we have the random network created with specific p. At ***step 1***, the component that contains 3 vertices is picked, a chosen vertex of this component is identified and superposed on the first open vertex that was chosen in step 0. Then, two random vertices in the cycle are identified and superposed on the two vertices from the random network. These newly identified vertices are classified as closed vertices and have the color black. Their neighbours will become open vertices and have color white. The vertices with gray color are not yet identified and also not open yet. The process is repeated starting from the open vertices. The process halts when all vertices are classified as closed. Unless all open vertices generated at step k are finished, we do not consider vertices generated at step $k+1$. It should be noted that the process in Figure 4.2 is not random, it is made to clarify our explanation.

As we have seen, at any step k there are two categories of vertices: the set of open vertices denoted by X_k and the set of closed vertices denoted by T_k. Therefore, the set of the identified vertices, D_k, is the union of the closed vertices until step k. The union of all open vertices until step k, is denoted by S_k.

$$S_k = X_1 \cup X_2 \cdots \cup X_k. \tag{4.5.1}$$

The set of all the vertices at distance at most k from v in $C(n,p)$ is denoted by R_k:

$$R_k = \{w \in C(n,p) : l(v,w) \leq k\}. \tag{4.5.2}$$

Lemma 4.6 compares S_k with R_k in the disordered network $C(n,p)$, and shows that S_k contains all the vertices at distance at most k from v_0 in $C(n,p)$.

Lemma 4.6. *At any step k*

$$R_k \subseteq S_k.$$

Proof. Using induction. At step 0, $R_0 = \{v_0\}$, so $R_0 \subseteq S_0$. Now, assume that for some k, $R_k \subseteq S_k$ is true. Let $w \in R_{k+1}$, i.e., $l(v_0,w) \leq k+1$, which means that there exists $w' \in R_k$ such that w is adjacent to w' in $C(n,p)$. Since $w' \in R_k$, by the induction hypothesis $w' \in S_k$; which means that w' was open at a step $j \leq k$ of the process starting at v. Since w' is adjacent to w in $C(n,p)$, then w is open at step $j+1$, and $j+1 \leq k+1$. Hence, $w \in S_{k+1}$ □

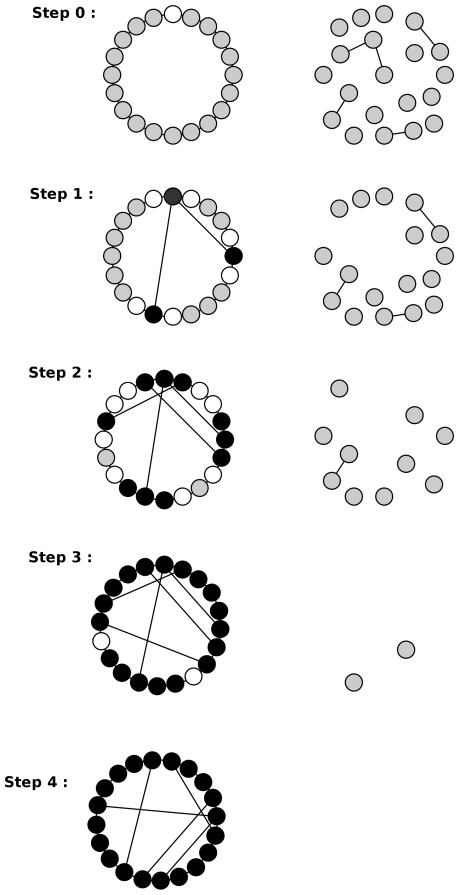


Figure 4.2: The random process of constructing small-world network as it is explained in the text.

Let A be an arbitrary positive number (may depend on n) and let

$$p(A,n) := \mathbb{P}\left(\bar{l}(C(n,p)) \leq A\log n\right).$$

Lemma 4.7. *For a random vertex $v \in V(C(n,p))$*

$$\mathbb{P}\left(|R_{4A\log n}(v)| \geq \frac{n-1}{2}\right) \geq \frac{1}{2}p(A,n).$$

Proof. Given the event

$$\bar{l}(C(n,p)) \leq A\log n.$$

By definition in (2.2.1), we have

$$\frac{1}{n}\sum_{v\in V(C(n,p))} \bar{l}(v) \leq A\log n.$$

This implies that

$$\frac{1}{n}\sum_{\substack{v\in V(C(n,p))\\ \bar{l}(v)\geq 2A\log n}} \bar{l}(v) \leq A\log n,$$

and

$$\frac{1}{n}\sum_{\substack{v\in V(C(n,p))\\ \bar{l}(v)\geq 2A\log n}} \bar{l}(v) \geq \frac{2A\log n}{n}\sum_{\substack{v\in V(C(n,p))\\ \bar{l}(v)\geq 2A\log n}} 1.$$

Hence,

$$\sum_{\substack{v\in V(C(n,p))\\ \bar{l}(v)\geq 2A\log n}} 1 \leq \frac{n}{2}.$$

So, there are at least $\frac{n}{2}$ vertices in $C(n,p)$ with the property $\bar{l}(v) < 2A\log n$.

For such a vertex, using definition in (2.2.1) again we have

$$\frac{1}{n-1}\sum_{\substack{w\neq v\\ w\in V(C(n,p))}} l(v,w) < 2A\log n.$$

And so,

$$\frac{4A\log n}{n-1}\sum_{\substack{w\neq v\\ w\in V(C(n,p))\\ l(v,w)\geq 4A\log n}} 1 \leq \frac{1}{n-1}\sum_{\substack{w\neq v\\ w\in V(C(n,p))}} l(v,w) < 2A\log n.$$

Hence, there are at least $\frac{n-1}{2}$ vertices in $C(n,p)$ with $w \neq v$ and $l(v,w) < 4A\log n$, i.e.,

$$|R_{4A\log n}(v)| \geq \frac{n-1}{2}.$$

This shows that for a random vertex v in $C(n,p)$

$$\mathbb{P}\left(|R_{4A\log n}(v)| \geq \frac{n-1}{2}\,\middle|\,\bar{l}(C(n,p)) \leq A\log n\right) \geq \frac{1}{2}.$$

The proof is completed by noting that

$$\mathbb{P}\left(|R_{4A\log n}(v)| \geq \frac{n-1}{2}\,\middle|\,\bar{l}(C(n,p)) \leq A\log n\right) \leq \frac{\mathbb{P}(|R_{4A\log n}(v)| \geq \frac{n-1}{4})}{\mathbb{P}(\bar{l}(C(n,p)) \leq A\log n)} = \frac{\mathbb{P}(|R_{4A\log n}(v)| \geq \frac{n-1}{4})}{p(A,n)}$$

□

4.6 Early Evolution of The Random Tree Process

In this section we shall study the random tree process at its early stage, i.e., when $k \geq 2$ and S_k is still small. More precisely, we will study the random tree process conditioned by

$$S_k \leq n^{1-\delta}, \tag{4.6.1}$$

for an arbitrary integer k when $\delta > 0$ is a small fixed real number. Lemma 4.8 gives the average order of the component at step k. The remaining network at the early stage for a given k is denoted $R_k(n,p)$. Hence, the average order of the component of a random vertex in $R_k(n,p)$ is

$$e(R_k(n,p)) = \sum_{v \in V(R_k(n,p))} \frac{|c(v)|}{n - \mathcal{O}(n^{1-\delta})}. \tag{4.6.2}$$

Lemma 4.8. *For any $k > 0$*

$$|e(R(n,p)) - e(R_k(n,p))| = \mathcal{O}(n^{-\delta} \log n).$$

Proof.

$$\begin{aligned}
|e(R(n,p)) - e(R_k(n,p))| &= \left| \sum_{v \in V(R(n,p))} \frac{|c(v)|}{n} - \sum_{v \in V(R_k(n,p))} \frac{|c(v)|}{n - \mathcal{O}(n^{1-\delta})} \right| \\
&= \left| \left(\sum_{v \in V(R_k)} \frac{|c(v)|}{n} + \sum_{v \in V(R)/V(R_k)} \frac{|c(v)|}{n} \right) - \sum_{v \in V(R_k)} \frac{|c(v)|}{n - \mathcal{O}(n^{1-\delta})} \right| \\
&\sum_{v \in V(R_k)} |c(v)| \left(\frac{1}{n} - \frac{1}{n - \mathcal{O}(n^{1-\delta})} \right) + \sum_{v \in V(R)/V(R_k)} \frac{|c(v)|}{n}
\end{aligned}$$

assuming that at step k, $n^{1-\delta}$ component has order of $\log n$ then

$$\begin{aligned}
|e(R(n,p)) - e(R_k(n,p))| &= \sum_{v \in V(R_k)} \frac{|c(v)|}{n} \underbrace{\left(1 - \frac{1}{1 - \mathcal{O}(n^{-\delta})}\right)}_{= \mathcal{O}(n^{-\delta}) \text{ by using geometric power series}} + \mathcal{O}\left(\frac{n^{1-\delta} \log n}{n}\right) \\
&= \mathcal{O}(n^{-\delta} \log n) + \mathcal{O}(n^{-\delta} \log n) \\
&= \mathcal{O}(n^{-\delta} \log n).
\end{aligned}$$

□

Lemma 4.9 gives an upper bound for the probability of the growth of S_l defined in formula (4.5.1).

Lemma 4.9. *For any positive integer l and any positive real number t, define $f(\varepsilon) := \log(1+c_2\varepsilon)$, then we have an upper bound on the conditional probability*

$$\mathbb{P}(|S_l| > le^{t+(l-1)f(\varepsilon)+\log|X_1|}||S_{l-1}| \leq n^{1-\delta}) \leq l\exp\left(-c_3\frac{t^2}{l(\log\log)^2}\right),$$

for some constant $c > 0$.

Proof. Assume that we are in the early evolution of the random tree process by conditioning on (4.6.1), i.e., $|S_{l-1}| \leq n^{1-\delta}$. For $k \leq l-1$, using Lemma 4.8, each open vertex at step k will have $e(R(n,p)) + \mathcal{O}(n^{-\delta}\log n)$ vertices in its component on average. We have an upper bound for the conditional expectation

$$\mathbb{E}(|X_{k+1}| \mid |X_k|) \leq 2|X_k|(e(R(n,p)) + \mathcal{O}(n^{-\delta}\log n)) - |X_k| \leq |X_k|(1+c_2\varepsilon).$$

This is because the new vertices from the component will produce at most two open vertices at step $k+1$, and each open vertex at step k will produce at most one open vertex at step $k+1$. By using Lemma 4.5, the above upper bound for the conditional expectation holds.

Let

$$Y_k := \log|X_k|. \tag{4.6.3}$$

Then

$$\begin{aligned}\mathbb{E}(Y_{k+1}|Y_k) &= \mathbb{E}(\log|X_{k+1}| \mid |X_k|)\\ &\leq \log(\mathbb{E}(|X_{k+1}| \mid |X_k|))\\ &\leq \log(|X_k|(1+c_2\varepsilon))\\ &\leq Y_k + \log(1+c_2\varepsilon).\end{aligned}$$

Let

$$Z_k := Y_k - kf(\varepsilon). \tag{4.6.4}$$

Then

$$\begin{aligned}\mathbb{E}(Z_{k+1}|Z_k) &= \mathbb{E}(Y_{k+1} - (k+1)f(\varepsilon)|Y_k)\\ &\leq Y_k + f(\varepsilon) - (k+1)f(\varepsilon)\\ &\leq Y_k - kf(\varepsilon)\\ &\leq Z_k,\end{aligned}$$

which implies that the sequence $(Z_k)_{k\geq 1}$ is a supermartingales (Mitzenmacher and Upfal, 2005, p. 295-313).

Furthermore, because the largest component is of order more than $K\log n$ and each open vertex in step k will have at most $K\log n$ closed neighbours, each generates two open vertices. This leads to

$$\frac{|X_{k+1}|}{|X_k|} \ll 2\log n. \tag{4.6.5}$$

Using Equation (4.6.4) we obtain

$$\begin{aligned}|Z_{k+1} - Z_k| &= |Y_{k+1} - (k+1)f(\varepsilon) - (Y_k - kf(\varepsilon))| \\ &= |Y_{k+1} - Y_k - f(\varepsilon)|.\end{aligned}$$

Using Equation (4.6.3),

$$\begin{aligned}|Z_{k+1} - Z_k| &= |\log|X_{k+1}| - \log|X_k| + f(\varepsilon)| \\ &\leq \left|\log\frac{|X_{k+1}|}{|X_k|}\right| + |f(\varepsilon)|.\end{aligned}$$

Therefore, using Equation (4.6.5) we obtain

$$|Z_{k+1} - Z_k| \ll \log\log n. \tag{4.6.6}$$

Applying the Azuma-Hoeffding inequality for supermartingale (Mitzenmacher and Upfal, 2005, p. 295-313), we obtain for any positive real number t,

$$\mathbb{P}(Z_l - Z_1 > t) \leq \exp\left(-c_3\frac{t^2}{l(\log\log n)^2}\right). \tag{4.6.7}$$

Using the definition of Z_k in Equation (4.6.4), and then substituting from Equation (4.6.3) leads to

$$\begin{aligned}Z_l - Z_1 &= Y_l - lf(\varepsilon) - Y_1 + f(\varepsilon)) \\ &= Y_l - (l-1)f(\varepsilon) - Y_1 \\ &= \log|X_l| - (l-1)f(\varepsilon) - \log|X_1|.\end{aligned} \tag{4.6.8}$$

Thus, $Z_l - Z_1 > t$ is equivalent to

$$\begin{aligned}&\log|X_l| - (l-1)f(\varepsilon) - \log|X_1| > t \\ &|X_l| > e^{t+(l-1)f(\varepsilon)+\log|X_1|}.\end{aligned} \tag{4.6.9}$$

Coming back to the Azuma-Hoeffding inequality in Equation (4.6.7) and using the result of the derivation in Equations (4.6.8) and (4.6.9) implies that

$$\mathbb{P}(|X_l| > e^{t+(l-1)f(\varepsilon)+\log|X_1|}||S_{l-1}| \leq n^{1-\delta}) \leq \exp\left(-c_3\frac{t^2}{l(\log\log n)^2}\right).$$

Recall the definition of S_l from Equation (4.5.1), i.e., $S_l = \sum_{j=0}^{l} X_l$. Then, using the union bound we obtain

$$\mathbb{P}(|S_l| > le^{t+(l-1)f(\varepsilon)+\log|X_1|}||S_{l-1}| \leq n^{1-\delta}) \leq l\exp\left(-c_3\frac{t^2}{l(\log\log n)^2}\right).$$

□

4.7 The result

This section states our result regarding the average path length of small-world networks. It gives an upper and lower bound for the average path length and shows that the disordered network $C(n,p)$ is a small world network when pn is bounded away from zero where the average path length in this case is bounded above by $\mathcal{O}(\log n)$. Otherwise, $C(n,p)$ does not have an average path length of logarithmic order.

Theorem 4.10. *Let $C(n,p)$ be a network formed by adding edges randomly with probability p to the cycle $C(n)$.*

a) If pn is bounded away from zero then

$$\bar{l}(C(n,p)) = \mathcal{O}(\log n).$$

b) If $pn \to 0$ then there is a function $\omega(n) \to \infty$ such that, w.h.p. the average path length of the network $C(n,p)$ satisfies

$$\bar{l}(C(n,p)) > \omega(n) \log n.$$

Proof. a) From Theorem 4.2, the diameter of $C(n,p)$ is $\mathcal{O}(\log n)$, which gives an upper bound for the average path length. So

$$\bar{l}(C(n,p)) = \mathcal{O}(\log n).$$

b) Let A be an arbitrary positive number depending on n, and define $p(A,n)$ to be

$$p(A,n) := \mathbb{P}(\bar{l}(C(n,p)) \leq A \log n).$$

Using Lemma 4.7,

$$\mathbb{P}\left(|R_{4A\log n}(v)| \geq \frac{n-1}{2}\right) \geq \frac{1}{2}p(A,n).$$

Let $k_0 = 4A \log n$. Using Lemma 4.6

$$\mathbb{P}\left(|S_{k_0}(v)| \geq \frac{1}{2}(n-1)\right) \geq \mathbb{P}\left(|R_{4A\log n}(v)| \geq \frac{n-1}{2}\right) \geq \frac{1}{2}p(A,n). \tag{4.7.1}$$

Now we choose A such that the following holds as $n \to \infty$

$$A \to \infty, \quad f(\varepsilon)A \to 0 \quad \text{and} \frac{\log n}{A(\log\log n)^2} \gg \sqrt{\log n}.$$

For every $k \leq k_0$ and $t = t_k$ such that

$$ke^{t+(k-1)f(\varepsilon)+\log|X_1|} = n^{1-\delta}$$

we have

$$t_k \gg \log n.$$

By applying Lemma 4.9, then, uniformly for $k \leq k_0$,

$$\mathbb{P}\left(|S_k| > n^{1-\delta} || S_{k-1}| \leq n^{1-\delta}\right) \leq k \exp\left(-c_1 \frac{(\log n)^2}{k(\log\log n)^2}\right).$$

Since $k \leq k_0 = 4A \log n$, then

$$\mathbb{P}\left(|S_k| > n^{1-\delta} \middle| |S_{k-1}| \leq n^{1-\delta}\right) \leq k \exp\left(-c_2 \frac{\log n}{A(\log\log n)^2}\right).$$

Therefore,

$$\begin{aligned}\mathbb{P}\left(|S_k| > n^{1-\delta}\right) =& \mathbb{P}\left(|S_k| > n^{1-\delta} || S_{k-1}| \leq n^{1-\delta}\right) \mathbb{P}\left(|S_{k-1}| \leq n^{1-\delta}\right) + \mathbb{P}\left(|S_{k-1}| > n^{1-\delta}\right) \\ &\leq \exp\left(-c_2 \frac{\log n}{A(\log\log n)^2}\right) + \mathbb{P}\left(|S_{k-1}| > n^{1-\delta}\right).\end{aligned}$$

Iterating the above recurrence we obtain

$$\begin{aligned}\mathbb{P}\left(|S_{k_0}| > n^{1-\delta}\right) &\leq k_0 \exp\left(-c_2 \frac{\log n}{A(\log\log n)^2}\right) \\ &\ll A \log n \exp\left(-c_3 \sqrt{\log n}\right).\end{aligned}$$

Finally, using Equation (4.7.1), we obtain

$$p(A, n) \ll A \log n \exp\left(-c_3 \sqrt{\log n}\right) \to 0,$$

as $n \to \infty$.

□

Earlier we derived our result regarding the average path length when the base network is a cycle (i.e, each vertex has degree 2). This result can be generalized when $k > 1$ as follows; We group adjacent vertices in groups of order k. Two vertices are connected in the resulting network if any of the original vertices in each group was connected to any of the original in the other. This includes short-cut connections. Notice that the number of the short-cuts is preserved in this grouping process. Therefore, the parameters of the resulting network are:

$$n' = \frac{n}{k}, \quad p' = k^2 p, \quad k' = 1, \quad l' = l.$$

Figure 4.7 shows the grouping technique when $k = 4$. (A similar grouping is done in (Newman and Watts, 1999*a*)).

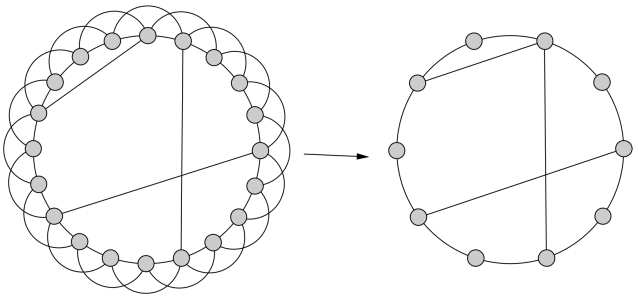


Figure 4.3: The grouping technique when $k > 1$ to obtain a generalized small world result.

4.8 Conclusion

We have started this thesis by asking this question: what is the average minimum number of extra random links that can be added to a one dimensional lattice such that the average path length is small ? In this chapter we have answered this question by showing that at least εn (where ε is a constant bigger than zero) random links should be added to a one dimensional lattice to ensure average path length of order $\log n$.

Chapter 5

Conclusion and Future work

5.1 Conclusion

One of the main concerns of network science is to study and model real-world networks through mathematical abstraction that captures some key realistic features. The three metrics studied in this thesis are: clustering coefficient, cc, degree distribution and average path length, $\bar{l}$. The importance of these quantities has been first defined and emphasised by empirical studies of real-world networks in Chapter 2. Our real interest lies in the third metric because the first two metrics are always easy to compute analytically.

Stimulated by the high clustering and small average path length observed in real-world networks, small-world models have been proposed to interpolate between high clustering regular lattices and random networks (Watts and Strogatz, 1998). This thesis presents a combination of analytical results, empirical work and simulations regarding small-world network models and their average path length.

Models of small-world network combine structured topology very often represented by a periodic one dimensional lattice, and randomness represented in different ways. Firstly, rewiring the links in the base structured network with probability p; this was the original construction of small-world network proposed by Watts and Strogatz (1998). Secondly, adding extra links on top of the base network with probability p; this construction was proposed by Newman and Watts (1999a) to overcome the difficulties regarding the analytical treatment of the original model. The properties of both models are investigated by numerical simulation and compared with empirical results obtained from real-world networks. Both models have small average path length, and high clustering coefficient in a good agreement with the empirical study of real-world networks. However, they lack the scale-free degree distribution observed in real-world networks.

The average path length has been computed analytically for the model Newman and Watts (1999a) model methods from statistical physics and ana-

lytical methods including scaling forms in (Newman and Watts, 1999*a*; Barrat and Weigt, 2000). Following these methods, the onset of small world phenomena where $\bar{l}$ scales logarithmically with n occurs if

$$\lim_{n\to\infty} \frac{\log p}{\log n} = \frac{-1}{\tau}, \text{then } \tau = 1.$$

Making any further progress using those methods has proved difficult.

Despite the fact that the empirical study in Chapter 2 shows that the properties of random networks deviate from real-world networks (specifically the clustering coefficient and the degree distribution), we have used a random network as a mathematical tool to model and study small-world networks. Indeed, our task is not only to study and design a model that meets small-world properties, but also to impose and develop mathematical tools to study small-world networks. For that reason, we have represented randomness in small-world models by Erdős-Rényi random network models which we impose on a base structured network. Our construction allows us to reason about small-world networks using tools and results from random graph theory. We have improved the analytical result regarding the average path length and we have shown that the onset of small-world behaviour occurs if

$$p = \frac{\varepsilon}{n},\ \varepsilon > 0.$$

Furthermore, we have shown that when $p = \frac{\varepsilon}{n}$ and $\varepsilon \to 0$ the disordered network is no longer small-world since $\bar{l} \to \infty$ as $n \to \infty$.

Using random graph theory has proved to be a good mathematical tool to reason analytically about small-world networks and it can be used to overcome some drawbacks of small-world models which we shall briefly mention in the subsequent section.

5.2 Future Work

Although we have constructed small-world models that are analytically solvable, knowing that the best studied case by far is the one dimensional lattice based network, our result of the average path length is obtained for that case also. However, we believe our small-world models can be built on lattices of any dimension and we can derive the same result regarding the average path length.

Small-world models are sadly missing two related properties of real-world networks: the first one is the degree distribution. We have shown that small-world network models do not have the scale-free degree distribution observed in real-world networks. Although small-world models were never intended to mimic real-world networks in terms of degree distribution, the distribution of the extra links on the lattices are important for routing to be efficient and easy

in spite of a lack of global knowledge of the network structure. Searching using local information in small-world networks was realised by Kleinberg (2000). He observed that the small-world experiments of Milgram (1967) not only demonstrate the small average path length between arbitrary pairs of vertices, but also individuals operating with purely local information are capable of finding these paths. Kleinberg (2000) proposed a decentralized mechanism to navigate the underling network without knowledge of the global structure of the network where the underlying network is a d-dimensional lattice and random links are added where the probability p_{ij} of two vertices being connected is

$$p_{ij} \propto d(i,j)^{-\alpha},$$

where $d(i,j)$ is the lattice distance from vertex i and vertex j. A short path of polylogarithmic expected length can be found using a local greedy algorithm only when $\alpha = 2$ and there is no such algorithm in the original small-world models where $\alpha = 0$. Some open research questions in the navigable small-word have been indicated in (Kleinberg, 2006). Following our methods in this work, we can address the navigability of small-world network by using random graphs with specific degree distributions instead of the standard Erdős-Rényi random graph.

Additionally, all small-world models we have investigated including Kleinberg (2000)'s model of navigable small-world networks rely on lattice structure as a base, to guarantee connectivity and high clustering in the former and to provide a distance metric that is independent of the network distance in the latter (Watts, 2004). Certainly, social networks and most real-world networks are not built on a lattice. The question arises, are we able to find a better representation of the base network to model social networks?

Bibliography

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, vol. 74, no. 1, p. 47.

Backstrom, L., Boldi, P., Rosa, M., Ugander, J. and Vigna, S. (2012). Four degrees of separation. In: *Proceedings of the 3rd Annual ACM Web Science Conference*, pp. 33–42. ACM.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, vol. 286, no. 5439, pp. 509–512.

Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A. and Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, vol. 311, no. 3, pp. 590–614.

Barrat, A. (1999). Comment on"small-world networks: Evidence for a crossover picture". *arXiv preprint cond-mat/9903323*.

Barrat, A. and Weigt, M. (2000). On the properties of small-world network models. *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 13, no. 3, pp. 547–560.

Barthélémy, M. and Amaral, L.A.N. (1999). Small-world networks: Evidence for a crossover picture. *Physical Review Letters*, vol. 82, no. 15, p. 3180.

Bollobás, B. (1981). The diameter of random graphs. *Transactions of the American Mathematical Society*, vol. 267, no. 1, pp. 41–52.

Bollobás, B. (1998). *Modern Graph Theory*. Graduate texts in mathematics. U.S. Government Printing Office. ISBN 9780387984889.

Bollobás, B. (2001). *Random Graphs*. Cambridge Studies in Advanced Mathematics. Cambridge University Press. ISBN 9780521797221.
Available at: `http://books.google.co.za/books?id=o9WecWgilzYC`

Bollobás, B. and Chung, F.R.K. (1988). The diameter of a cycle plus a random matching. *SIAM Journal on discrete mathematics*, vol. 1, no. 3, pp. 328–333.

Chung, F. and Garey, M. (1984). Diameter bounds for altered graphs. *Journal of Graph Theory*, vol. 8, no. 4, pp. 511–534.

Chung, F. and Lu, L. (2001). The diameter of sparse random graphs. *Advances in Applied Mathematics*, vol. 26, no. 4, pp. 257–279.

Cont, R. and Tanimura, E. (2007). Small world graphs: characterization and alternative constructions.

Dorogovtsev, S.N. and Mendes, J.F.F. (2000). Exactly solvable small-world network. *EPL (Europhysics Letters)*, vol. 50, no. 1, p. 1.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, vol. 5, pp. 17–61.

Faloutsos, M., Faloutsos, P. and Faloutsos, C. (1999). On power-law relationships of the internet topology. In: *ACM SIGCOMM Computer Communication Review*, vol. 29, pp. 251–262. ACM.

Flaxman, A.D. and Frieze, A.M. (2004). The diameter of randomly perturbed digraphs and some applications. In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pp. 345–356. Springer.

Fronczak, A., Fronczak, P. and Hołyst, J.A. (2004). Average path length in random networks. *Physical Review E*, vol. 70, no. 5, p. 056110.

Harris, J., Hirst, J. and Mossinghoff, M. (2009). *Combinatorics and Graph Theory*. Undergraduate Texts in Mathematics. Springer. ISBN 9780387797113.
Available at: `http://books.google.co.za/books?id=DfcQaZKUVLwC`

Jackson, M. (2010). *Social and Economic Networks*. Princeton University Press. Princeton University Press. ISBN 9780691148205.
Available at: `http://books.google.co.za/books?id=bJbuD2XH1_oC`

Jamakovic, A. (2008). Characterization of complex networks: application to robustness analysis.

Janson, S., Luczak, T. and Rucinski, A. (2011). *Random Graphs*. Wiley Series in Discrete Mathematics and Optimization. Wiley. ISBN 9781118030967.
Available at: `http://books.google.co.za/books?id=RjnqVoB4VmUC`

Kleinberg, J. (2000). The small-world phenomenon: an algorithm perspective. In: *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pp. 163–170. ACM.

Kleinberg, J. (2006). Complex networks and decentralized search algorithms. In: *Proceedings oh the International Congress of Mathematicians: Madrid, August 22-30, 2006: invited lectures*, pp. 1019–1044.

Milgram, S. (1967). The small world problem. *Psychology today*, vol. 2, no. 1, pp. 60–67.

Mitzenmacher, M. and Upfal, E. (2005). *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press. ISBN 9780521835404.
Available at: `http://books.google.co.za/books?id=0bAYl6d7hvkC`

Newman, M. (2010). *Networks: An Introduction*. OUP Oxford. ISBN 9780191500701.
Available at: `http://books.google.co.za/books?id=LrFaU4XCsUoC`

Newman, M.E. (2000). Models of the small world. *Journal of Statistical Physics*, vol. 101, no. 3-4, pp. 819–841.

Newman, M.E. (2001*a*). Scientific collaboration networks. i. network construction and fundamental results. *Physical review E*, vol. 64, no. 1, p. 016131.

Newman, M.E. (2001*b*). Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E*, vol. 64, no. 1, p. 016132.

Newman, M.E. (2001*c*). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 404–409.

Newman, M.E. (2003*a*). 2 random graphs as models of networks. *Handbook of graphs and networks*, p. 35.

Newman, M.E. (2003*b*). The structure and function of complex networks. *SIAM review*, vol. 45, no. 2, pp. 167–256.

Newman, M.E., Moore, C. and Watts, D.J. (2000). Mean-field solution of the small-world network model. *Physical Review Letters*, vol. 84, no. 14, p. 3201.

Newman, M.E. and Watts, D.J. (1999*a*). Renormalization group analysis of the small-world network model. *Physics Letters A*, vol. 263, no. 4, pp. 341–346.

Newman, M.E. and Watts, D.J. (1999*b*). Scaling and percolation in the small-world network model. *Physical Review E*, vol. 60, no. 6, p. 7332.

Newman, M.E., Watts, D.J. and Strogatz, S.H. (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. Suppl 1, pp. 2566–2572.

Redner, S. (2004). Citation statistics from more than a century of physical review. *arXiv preprint physics/0407137*.

Ugander, J., Karrer, B., Backstrom, L. and Marlow, C. (2011). The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*.

Van Steen, M. (2010). *Graph Theory and Complex Networks: An Introduction*. Maarten Van Steen. ISBN 9789081540612.
Available at: `http://books.google.co.za/books?id=V7bMbwAACAAJ`

Vázquez, A. (2001). Statistics of citation networks. *arXiv preprint cond-mat/0105031*.

Watts, D.J. (2004). The" new" science of networks. *Annual review of sociology*, pp. 243–270.

Watts, D.J. and Strogatz, S.H. (1998). Collective dynamics of âĂŸsmall-worldâĂŹnetworks. *nature*, vol. 393, no. 6684, pp. 440–442.

Wiki01, 2013 (Accessed on February 2013). Re-normalization grpup analysis. Wikipedia, the Free Encyclopedia.
Available at: `http://en.wikipedia.org/wiki/Renormalization _group`