PERCOLATION AND ITS RELATION TO OTHER PROCESSES IN NETWORKS

by

Ali FAQEEH

under the supervision of

Prof. James P. GLEESON

A thesis submitted for the degree of Doctor of Philosophy in the
Mathematics Application Consortium for Science and Industry (MACSI)
Department of Mathematics and Statistics
University of Limerick
September 2016
To my wife and to my parents.
Percolation and its Relation to Other Processes in Networks

Ali Faqeeh

Abstract

Many of the systems we observe in nature, in societies, or in infrastructures are in the form of a network of interacting units. This underlying network structure shapes the behavior of such systems and is an indispensable factor in maintaining their correct function. Likewise, the processes that operate on these systems are largely influenced by their network structure. In this thesis, we investigate the theoretical approaches for investigation of the properties of percolation processes on networks.

Percolation processes investigate the alteration of network connectivity. Two such processes that constitute the main focus of this thesis are bond and site percolation, which are simple models for the robustness of a network to random failures of (or intentional attacks to) its constituting units. They also have been used to provide better insight on some other more complicated processes such as spread of epidemic diseases or stability of genetic networks, because some important features of these processes can be mapped to percolation properties.

In this thesis, we first consider the so-called $A_{ij}$ theories developed for percolation and several other processes that operate on networks. We investigate the effect of the presence of high density of short loops (a property observed in many real-world networks) on the accuracy of $A_{ij}$ theories and show its impact on the performance of these theories. We then show that another phenomenon, the emergence of coexisting percolating clusters, can also cause significant inaccuracy in the $A_{ij}$ theory for bond percolation on certain synthetic and real-world networks. Moreover, we introduce a new theoretical approach that takes into account this phenomenon and improves upon the state-of-the-art $A_{ij}$ theory. Then, we develop a theoretical framework for calculation of percolation cluster sizes and discuss its potential applications in studying the properties of neuronal avalanches.
Declaration

I hereby declare that this thesis has been written by me, and that it has not been submitted in any previous application for an academic award. Wherever the contributions of others were involved, every effort has been made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions.

Ali Faqeeh
Limerick, 2016
Publications

Some of the results of this thesis have appeared in the following publications:


Acknowledgements

I gratefully thank my supervisor Prof. James Gleeson for all the support he provided me, for his patient, encouraging and thorough guidance, and for providing me excellent opportunities to excel in my research and academic career.
I also thank all my colleagues and friends in MACSI, especially Sergey Melnik, Davide Cellai, Peter Fennell, Peg Hanrahan, Mel Devine, Adam Hacket, David O’Sullivan, Kevin O’Sullivan, Kevin Moroney, and Catherine Timoney for their help and support. Also thanks to Pol Colomer and Marián Boguñá for great discussions and their hospitality during my visit to Barcelona.

This work was funded by Science Foundation Ireland (under programmes 11/PI/1026 and 12/IA/1683) and from the European Commission FET-Proactive project PLEX-MATH (FP7-ICT-2011-8; Grant No. 317614).
## Contents

1 Introduction

1.1 Complex Networks ................................. 1
    1.1.1 Social networks ............................. 2
    1.1.2 Technological Networks ..................... 5
    1.1.3 Biological networks .......................... 8

1.2 Percolation and networks .......................... 14
    1.2.1 Bond and site percolation .................... 14

1.3 The thesis structure ............................. 16

2 Mathematical background ........................... 17

2.1 Structural measures and representations .......... 17
    2.1.1 Adjacency matrix ............................ 17
    2.1.2 Nodes’ degrees and related quantities .... 18
    2.1.3 Generating functions ........................ 19
    2.1.4 Short loops and clustering .................. 20
    2.1.5 Communities ................................. 21
    2.1.6 Geodesic path ............................... 22

2.2 Some models of random networks .................. 22
    2.2.1 Erdős-Rényi and the configuration model networks ................................. 23
    2.2.2 An ensemble of clustered networks ............ 24
    2.2.3 LFR networks ................................ 25
2.3 Binary-state processes ........................................... 26  
 2.3.1 Percolation .................................................... 27  
 2.3.2 SI epidemic model .......................................... 29  
 2.3.3 The Ising model ............................................. 30  

3 Network Cloning and accuracy of $A_{ij}$ theories 33  
 3.1 Introduction ..................................................... 33  
 3.2 Cloned networks ............................................... 35  
    3.2.1 Design and description .................................. 36  
    3.2.2 The effect of $L$-cloning on clustering ................. 37  
    3.2.3 Extension of $L$-cloning that preserves loops of certain lengths 39  
    3.2.4 Partial $L$-cloning ....................................... 40  
 3.3 Dynamical processes on networks and $A_{ij}$ theories .......... 40  
    3.3.1 Percolation ................................................. 41  
    3.3.2 SI epidemic model ....................................... 45  
    3.3.3 The Ising model ......................................... 47  
 3.4 Conclusion ..................................................... 48  

4 Emergence of coexisting percolating clusters 51  
 4.1 Introduction ..................................................... 51  
 4.2 Synthetic modular networks ................................... 53  
 4.3 The message-passing (MP) approach .......................... 54  
 4.4 Evidence for the emergence of CPCs .......................... 55  
 4.5 The modular message-passing (MMP) approach ............... 59  
 4.6 Results on LFR benchmark networks .......................... 61  
 4.7 Results on real-world networks ............................... 68  
 4.8 Conclusion ..................................................... 70  

5 Size distribution of percolation clusters and neuronal avalanches 71  
 5.1 Introduction ..................................................... 71  
 5.2 The governing equations for the generating functions of the cluster size distributions ........................................... 73  
 5.3 Derivation of cluster size distribution ....................... 76  
    5.3.1 Using the derivatives of the generating function $H_0(z)$ .... 76
5.4 Cluster size distribution in bond percolation .......................... 79
  5.4.1 Using direct differentiation of $G_1(z)$ ............................... 79
5.5 Using asymptotic expansions ............................................. 81
  5.5.1 Method (I) (based on the method of Cohen et al.) .............. 83
  5.5.2 Method (II) ......................................................... 88
  5.5.3 Method (III) ....................................................... 89
  5.5.4 Method (IV) ....................................................... 91
5.6 Numerical results .......................................................... 93
5.7 Conclusion ................................................................. 96

6 Conclusion ........................................................................... 101

A Appendix A: Relation between $L$-cloned and multilayer networks 105
B Appendix B: Calculation of the probability that two CPCs are connected 107
C Appendix C: Accuracy of the MMP approach for LFR networks with small modules 109
D Appendix D: Application of the MMP theory to real-world networks 113
E Appendix E: Derivation of Eq. (5.40) ...................................... 117
F Appendix F: $(a - z)^\rho$ non-analyticity .................................. 121

References .............................................................................. 123
1.1 Complex Networks

Many systems we see around us, we use or from which we benefit, being in nature, society, or in the technological domain, are composed of a large number of interacting entities. Some examples are [116] the societies which are built on the interaction between individuals, the nervous systems where its finest units, neurons, are frequently reacting to the signals sent from other parts of the system, or the Internet in which the communications between routers make possible the flow of information at each instant from one part of the globe to the other. Each of these systems can be thought of as a network of connected units; the connections, depending on the system type, can represent an interaction, a type of communication, a relationship between the units, etc. Each unit is called a node or vertex in the graph theory terminology, and it is shown by a point when the network is drawn as a graph. Each connection is called a link or edge and can be drawn by a line between two connected (interacting) nodes in the sketch of the network.

The pattern of connections is an important characteristic of a network that influences its behavior. For many real-world systems, the connection pattern is non-trivial in the sense that it lies somewhere between order and disorder. An example of a fully ordered network is a regular lattice (e.g., a square lattice), in which each node is connected to a fixed number of neighbors located in specific locations in its (geometrical)
proximity. At the other end of the spectrum is a fully random network, in which con-
nections between nodes are created by chance and not any other constraint. A complex
network is one that has a structure that is neither fully ordered nor fully random.

One of the main aims of network science is to understand the properties of the
structure of complex networks and determine the effect of this structure on the behavior
of the system, or on the behavior of its entities. To accomplish this aim, simple models
are developed in order to uncover the possibly that simple rules govern the behavior of
different systems. Some of these models can be potentially employed to a wide variety
of systems with networked structure. For example, it has been shown [26] that many
networks, with fundamentally different natures and applications (for example being
biological or social systems), have principal similarities in their structure. Moreover,
some different dynamics that operate on very distinct systems can be described by a
more general single dynamical model [64]. These examples show the unifying picture
that network science can provide to understand the properties and behavior of a variety
of systems and phenomena.

In Secs. 1.1.1-1.1.3, we introduce examples of social, technological, and bio-
logical systems that can be viewed and studied as complex networks. Also, we discuss
how a network representation suits such systems and summarize some of the proper-
ties of these networks and their entities. This brief overview is mainly based on the
reviews provided in Refs. [4, 18, 26, 39, 107, 116]. We will use some of the real-
world networks, that we introduce in the following sections, for our investigations in
the subsequent chapters of this thesis.

1.1.1 Social networks

A significant portion of the studies in complex networks science has been performed
on social systems; in fact, network analysis of social behavior has a long history in so-
cial sciences [18, 57, 137]. A social network is constituted by people in a society (or a
class or group of them) connected by one or more types of relations or interactions that
exist between individuals. These interactions include, for example, friendship, col-
laboration, affiliation, business or political relations, sexual relations, and interactions
through communication services [18, 107, 116]. For example in one of the earliest
studies of social networks (1926), Wellman investigated the companionships between
1.1 Complex Networks

school children [150]. Another pioneering study was that of Mayo [137] where he researched social interactions between workers of a factory. Today, the study of social networks has grown significantly and interests researchers from a broad range of specialties, including sociology, physics, computer science, and mathematics. In the following, an overview of some examples of social systems studied using network science and the insight these studies can provide are presented.

**Networks built on friendships**

A popular type of social networks is the network of friends. Investigations on friendship networks have found applications in preventing the spread of disease, e.g., in schools [60, 61], the effect of friendships on adolescence delinquency [73], and describing the patterns of interaction between criminals in jail [85, 90].

**Collaborations**

Another class of social networks includes networks formed by common collaboration between individuals. An example of collaboration networks is the network of co-authorships in which scientists are connected to each other if they coauthored one or more manuscripts [26]. The amount of interaction between co-authors can be related to the number of joint manuscripts they wrote together. To include the number of manuscripts that two authors collaborated to prepare, we can assign a weight to the link that connects the two author in the co-authorship network and hence construct a *weighted* network of co-authorships. Databases of different disciplines are used to construct co-authorship networks[9, 68, 69, 102, 103, 104]. Co-authorship networks can be used to quantify interactions among scientists in a particular area and also to investigate the overlap between different fields.

Members of the boards of company directors can be assumed to form a board membership network in which two board members are related (connected) if they sit at the same board [13, 26, 40]. It has been shown that [13, 18, 20, 40, 127] the structure of the network of board directors can affect the alliances, influences, and investment strategies by determining the spread of ideas and attitudes of the corporations. Similar to the co-authorship networks, the networks of board directors can be assumed to be a
1. INTRODUCTION

weighted network in which the weights of the link between two directors is the number of boards on which they both sit.

Communications

One way to study the interactions between people is to look at their communications, through, for example, phone calls, emails, online social networks (such as Twitter, Facebook, and Instagram), and online communication services (such as Telegram, Viber, and Whatsapp). The datasets for such communication systems are recorded in servers of the service provider. There are huge amounts of data available about communications of millions or billions of people through the aforementioned media. While the traditional methods of obtaining information about social interactions, like using questionnaires [93], have problems such as subjectivity, inaccuracy and small sample size, the electronic and online data does not suffer from such problems [107].

One of the first studies of communication networks was performed on long-distance phone calls in the AT&T network on a single day [2, 3]; in the network constructed using this data, the nodes are people (or more accurately phone numbers) and each directed edge represents a call from one number to the other. An application of such studies is, for example, detection of areas affected by extreme events using the pattern of communication activities during emergencies caused by such events [8].

Email networks are another type of communication systems. Email networks can be constructed by considering each email address as a node, and a directed edge is considered between two addresses if an email is sent from one address to the other [45, 116, 144]. Another method to construct email networks is to use address books which are lists of regular correspondences of users saved by each of them for convenience. Similar to the previous case, in such email networks, there is a directed edge from an email address to the other, if the first has sent a message to the second address.

The youngest type of communication systems are online social communities. In the last few years, many datasets have been gathered for such systems and various studies have been performed on their properties and how the network structure of these systems affect interactions, decisions and reactions of people. An example of such a social network is Facebook, in which if two persons are Facebook friends, an undirected link is considered between them [44]. The Twitter network is another example in which
1.1 Complex Networks

a directed edge exists from the followee to the follower [67]. There are many other online communication platforms such as those provided by Viber, Whatsapp, Line, and Instagram applications with a large population of users. Although there might not exist much published research on these networks, the fact that users are asked to grant consent for collecting the data on their activities shows that the service providers realize the importance and benefits of studying the human behavior in such networks. In Chapter 4, we use two examples of social networks (a Facebook friendship network and a collaboration network) to investigate the accuracy of certain theoretical approaches on these real-world networks.

1.1.2 Technological Networks

In this section, examples of networks brought to our lives by technological achievements are outlined. In particular, some technical details are provided regarding the appropriate choices of the nodes and links in these systems according to the problems that can be addressed by network science. The technological networks considered in this section are the Internet, the World Wide Web, and power grids.

Internet

The Internet is composed of computers and other devices connected through physical connections such as optical fibers. In the Internet network some nodes are the end users, e.g., personal computers at homes, offices, and companies. These devices comprise the outer layer of the Internet, almost all of them having a single link to the device through which they connect to the Internet. At an internal layer, routers are located; the routers are connected to the end users and other routers and transmit data packages from the end users to the other parts of the Internet and vice versa. Routers are powerful computers manufactured for data transmission in the Internet. The innermost layer of the Internet, or its backbone, is composed of powerful routers that are distributed across the globe and are connected to each other through fast pipelines of the most advanced optical fibers. These routers, usually developed and managed by governments or major communication companies, supply the intermediate layer routers, for example local ISPs, that are service providers to the end users.
1. INTRODUCTION

There is no central authority to monitor or control the Internet. New routers or connections are added constantly to the network or are removed autonomously from the network. New end users are also added to (removed from) the Internet all the time as they turn on (off) their computer. Hence, there is no comprehensive registry for the constituents of the Internet, and data about this enormous network must be collected experimentally. The available data vary in their scale and in the method of collection. As the traffic flow on the Internet is mainly imposed by the structure of the router layer, and as most of the features of the Internet network such as the efficiency, performance and robustness are mainly determined by the structure of this layer, network science studies commonly focus only on the routers and omit the users layer [116]. The layer of routers itself can be coarse-grained in different ways, depending on the level of detail required in a particular study. Three main coarse-graining schemes include subnets [70], domains [50], and autonomous systems [23, 29], that we briefly describe in the following.

The coarse-graining in the subnet scheme is done according to the unique IP address of each router. Each IP address is in the format of four numbers separated with a dot, each number being between 0 and 255; e.g., one of the routers at the University of Limerick has the IP address 193.1.103.10. At the first level of coarse-graining, all the addresses that their first three numbers are the same are considered as an individual node of the network. This is reasonable as many organizations have more than one router to supply the traffic they need; hence, the IP addresses of these routers have some of their former numbers in common and their last number(s) varies. For example, University of Limerick owns all the IP addresses of the format 193.1.103.xxx, as well as some other groups of addresses. In this level of coarse-graining, one avoids many of the links that are located inside individual organizations. Further coarse-grained levels considers only the second or the first number of the IP address as the identifier of a node. In the coarse-grained network, a link is assumed between two nodes if there is at least one router of the first node that is connected to one router of the second node. The network data are collected by a method that involves sending packets from different parts of the Internet to different destinations and then analyzing the paths that the packets took (composed of routers that transferred the package) on their way to the destinations.
Another way to construct a coarse-grained picture of the Internet is to consider domains. A group of routers and computers controlled by a single organization that have a unique domain name are recognized as a domain. The domain name is written in human readable format. For example, tcd.ie is the domain name for the devices that operate at Trinity College Dublin. The domain name to which an IP address belongs can be obtained easily according to publicly available data in the web.

The other coarse-graining scheme is considering the autonomous systems (AS). An AS is a group of devices controlled by the same authority that carries out itself its internal Internet communications. Autonomous systems often coincide with domains. The method of collecting information about autonomous systems is different from that of domains. Autonomous systems and the connections between them are obtained from routing tables, which are lists of paths that each router records. These paths are those that the data packets, that passed through the router, traversed to go from an AS to the AS in which their destination locates. At the destination AS, the transfer of data packet to its final destination is handled internally by the organization. We will discuss, in Chapter 4, the properties of a network process operated on an Internet network at the level of autonomous systems.

**WWW**

The World Wide Web (WWW) is probably the most popular example of a technological network; millions of people across the globe use it on a daily basis. In this network, the nodes are webpages and the directed links are the hyperlinks in one webpage that point to another page in WWW. The size of WWW is estimated to be more than 45 billion web pages and it is changing all the time as new pages are added or some old pages are removed from the network. Similar to the Internet, the data for WWW is extracted experimentally. Such data are obtained using computer algorithms called crawlers.

Crawlers start from a webpage in WWW and follow its links to find other pages that it is connected to and then repeat this process for the pages found. At each step, they neglect hyperlinks that were detected in previous crawls. By repeating this process for many starting webpages an appreciable portion of the WWW can be observed. However, not all the webpages can be detected using crawlers, for several reasons: (i)
1. INTRODUCTION

some webpages do not allow the use of crawlers; (ii) some webpages are dynamic and change over time; and most importantly (ii) not all the web is accessible from a starting point because of the specific structure of this directed network [43].

Power Grids

Our final example of technological networks is the power grid. In a power grid network, the links are high voltage transmission lines transporting electric power over long distances within a country or outside its boundaries. The nodes are the power generators and electrical substations. The data for power grids are extracted from the tables and maps maintained by the supervising authorities. The low voltage lines and end users are not usually included in the power grid networks. The power grid is a special network, similar to the Internet, as geographical locations can be assigned to the nodes. The effect of properties corresponding to the location of nodes, such as the economic, social and geographical conditions, can be useful in studying the shape and growth of power grid networks [116]. Also, robustness of power grids has been the focus of several network studies [120]. In Chapters 3 and 4, we discuss the results of our investigations on examples of technological networks, including power grids and the Internet.

1.1.3 Biological networks

In this section, we briefly review some of the networks of interactions that take place in living organisms. First of all, we consider the processes that are essential to the correct functioning and survival of living cells. Then, we consider networks of interactions between neurons.

Intricate processes that involve interactions between different constituents or contents of the living cell enable the correct function of the cell and provide the conditions necessary for its survival [18, 35]. Three main classes of these interactions studied by network science include: metabolic pathways, protein-protein interactions, and gene regulation. Below, each of these classes is briefly described.
1.1 Complex Networks

Metabolic pathways

“Metabolism” refers to the different chemical reactions taking place in a living cell to produce the materials and energy that it needs [18]. The substances from the external environment are converted to other materials useful to the cell through such reactions. There are two main types of reactions in a cell. In the first type, complex molecules are broken by the cell into smaller ones. Such reactions provide the substances needed for other reactions and/or supply their energy. An example is the conversion of glucose into simpler molecules in the cell. In the second type of reactions, small molecules are merged to construct complex molecules such as proteins [26].

Some of these reactions are facilitated or impeded by an enzyme. Enzymes are proteins that catalyze the reactions in the cell. A series of chemical reactions and the enzymes that take part in these reactions are referred to as metabolic pathways. The materials consumed or produced in different stages of a metabolic pathway are referred to as metabolites. In reality, the end product may not merely result from a single metabolic pathway but its production depends on other chemical reactions (more precisely the products of such reactions) taking place at the same time. Hence, the various materials consumed and produced in the cell are related through a network of chemical reactions, i.e., a metabolic network. In such a network, a directed edge connects the substance(s) of a chemical reaction to its product(s). Network studies have helped, for example, to identify the key metabolites in the cell [76, 147] and to understand the robustness of metabolic networks to the changes of external conditions [147].

Protein-protein interactions

Proteins are among the important constituents of the cell [35, 77]. Each protein is comprised of a sequence of a particular class of molecules called amino acids. There are only 20 specific amino acids from which proteins are created, while the number of amino acid units in some proteins is larger than 25000 [26]. Proteins can interact with other molecules inside the cell or in the external environment, including other proteins, DNA, neurotransmitters, and other smaller molecules. The most frequent interaction between proteins is the physical type [116]; two proteins are considered interacting physically [38] if they are part of the same protein complex, i.e., a macromolecule...
1. INTRODUCTION

Figure 1.1: The structure of an example of a protein complex. Each protein is specified by a certain color. The figure is taken from Wikimedia Commons (please see Ref. [54]).

constructed from several proteins [26] (see, for example, Figure 1.1). Protein-protein interactions are essential in living cell processes, functions and the cell organization [25, 155].

A complex networks approach is helpful to gain a better understanding of the complex system of interactions between proteins in the cell [35]. An example application of networks is in the identification of the functions of each protein. It is known that each protein is essential to a specific function in the cell; examples of such functions include catalyzing metabolic reactions, transporting materials, growing of the cell, signal transduction [39], and neutralizing the harmful effect of intruders, such as viruses and bacteria, to the cell by attaching to their antigens [116]. Although the functions of many proteins are known, the function of a considerable fraction of proteins in even the simplest organisms is yet unknown [39]. Network studies have provided methods to discover the functions of such proteins [18, 39]. Moreover, network techniques have shown potential application in studying the robustness of the system of protein-protein interactions against random faults of the proteins or targeted attacks [18, 26, 75, 152], the adaptation of this system to new conditions [5, 18] and also in studying the evolution of such systems [18, 122, 134]. In Chapter 4, we show the results for a theoretical method we developed on an example of protein-protein interaction network.
1.1 Complex Networks

Gene regulatory networks

To maintain its function, a cell needs specific species of proteins; different proteins may be required at different times. Hence, the rate at which each protein is produced in the cell is regulated by a precise mechanism that involves the cell DNA [116]. The information needed to construct each protein is included in a double strand biopolymer called the DNA. Each strand includes a sequence composed of four types of nucleotide molecules (denoted by A, C, G, and T). Each amino acid in a protein is coded by a specific triplet of these molecules on a DNA strand. Such triplets are called codons. Each specific protein is coded by a particular sequence of codons on the DNA strand. The start point and the end point of such a sequence is marked by two specific codons; the sequence between the start and end codons is called a gene.

The production of a protein in the cell is performed in two stages. At the first stage, called transcription, an enzyme referred to as RNA polymerase attaches to the DNA and generates a copy of the gene that codes for the protein. The copy, called messenger RNA, has a sequence similar to the gene, but its chemical structure is slightly different. At the second stage, called translation, the protein is synthesized in ribosome, an adroit molecular machinery that is composed of the messenger RNA, protein complexes and another RNA called transfer RNA. Hence, each protein species is produced when these two stages are fulfilled following the activation of a specific gene; in molecular biology terminology, when this procedure is performed it is stated that the gene is “expressed”.

The expression of a gene is regulated by other proteins: there are proteins that facilitate or inhibit the attachment of specific RNA polymerases to the DNA. Such proteins are themselves produced when their corresponding genes are expressed. Hence, some genes control the activation of other genes. These interactions between genes are represented by a gene regulatory network in which each node is a gene (or equivalently the protein produced by that gene) and an edge symbolizes an interaction. In some cases, the different types of promotion or inhibition interactions are represented by two different types of edges. The study of complex networks of gene regulation helps to understand better the function of these systems, their robustness and adaptation to external conditions, and how their malfunction can cause diseases [18, 142].
1. INTRODUCTION

Figure 1.2: A simplistic sketch of a typical neuron. The figure is taken from Wikimedia Commons (please see Ref. [55]).

Brain

The brain is another good example of a complex biological system where the interactions of its constituents are pivotal to perform its unique functions [141]. The number of neurons in the nervous system differs for different organisms; while the human brain has around a hundred billion neurons, the nervous system of simple organism can have as few as a handful of neurons [74]. Neurons receive input signals from other neurons (some specific classes of neurons also receive signals from other types of cells). After processing its inputs, a neuron itself may send a signal (specific in its characteristics to each neuron) to other neurons.

There are many different types of neurons in the human brain; typically they have four main parts in common (see Fig. 1.2) [116]. The cell body of a neuron is called the soma, and it possesses a relatively large number of projections called dendrites. The cell receives input mainly through dendrites. An axon is a long protrusion from the neuron that carries the output signal. Most neurons have only one axon. Each axon itself branches, usually close to its end; the tip of each branch is in close proximity of the dendrite of another neuron. The small gap between the end of each branch and the dendrite (which itself has a special sophisticated structure) is called a synapse.

The signal sent through an axon is an electric pulse produced by movements of positive ions in and out of the axon channels. This electrochemical signal is called an
action potential. The arrival of the action potential at the branches of the axon stimulates production of chemicals that are defused to the dendrites. When dendrites detect these chemicals, their potential is also altered by movement of ions into and out of them. In order for the receiver neuron to send a signal through its axon, the sum of the input voltages of its dendrites should exceed a certain threshold. Some of the inputs create positive potential at the dendrite; such inputs are referred to as excitatory. On the other hand, some neurons send negative potentials and hence have a inhibitory signal that decreases the chance that the receiving neuron reaches its threshold potential and fires a signal. Hence, when considering a network of neurons connected by synapses, one can assume two types (excitatory and inhibitory) of edges. Nevertheless, it was shown [153] that certain layers of the brain are composed of nearly independent sub-networks of excitatory neurons embedded in the larger-scale network that includes both excitatory and inhibitory groups of neurons. Moreover, some studies has focused on networks of solely excitatory populations of neurons (See, for example, Refs. [1, 72]).

Obtaining data on connections between individual neurons is a cumbersome task; currently, the experimental studies with neuron level resolution are limited to populations of a few hundreds of neurons, for example, in tissues of brain of different animals [15, 59, 139], or in the nervous system of simple organisms [18, 116]. For larger scales, noninvasive methods have provided plenty of data on interactions between different regions of the brain of humans and other primates [22, 121]. In such networks the nodes are specific regions of the brain or voxels of a few millimeters in size. The edges are either anatomical connections by axonal pathways between populations of neurons at different brain sites or functional connections representing existence of an interaction between the brain activities of two brain sites.

An area of interest in network science is the properties of neuronal avalanches [141]. A neuronal avalanche is defined as a consecutive sequence of firing activities of neurons in the network, such that the activity starts from a neuron or a group of neurons, which in turn causes, with some probability, activity in some of the neighboring neurons [14, 59, 88, 89]. The avalanche stops when the activity of neurons in the avalanche ends and the propagation of activity is interrupted. The size of an avalanche is then the number of neurons that became activated before the sequence (cascade) of activities stops, i.e., the total number of neurons fired during the avalanche. The duration of an avalanche is defined as the number of time steps between the start of the
activity and the time the activity is interrupted. In Chapter 5, we develop a simplis-
tic theoretical method to calculate the distribution of the sizes of neural avalanches in
networks.

1.2 Percolation and networks

In sections 1.1.1-1.1.3, we discussed examples of complex networks and some of their
important features. In particular, we described for each network, the constituting units
that can be regarded as the interacting nodes of the network, the nature of the inter-
action/relation that the connections represent, and other information that is crucial or
useful for correctly representing each network (e.g., if the connections are directed,
weighted, or temporal). After thinking of a system as a network, one might use this
perception to characterize and inspect the network properties that determine or influ-
ence the behavior of the system. One particular area of interest in this direction is
understanding the behavior of the processes that operate on or may occur in networks.
These processes include, for example, percolation (which is the alteration of the con-
nectivity of the network), the spread of disease, the magnetic interactions (such as those
of the Ising model [82]), synchronization [7], and the flow of information/opinions.

This thesis investigates percolation and its relation with examples of other pro-
cesses. The interest in studying percolation mainly arises due to the following facts:
(i) it is a simple model for network robustness to random failures or attacks [80]; (ii)
despite its simplicity, it presents very interesting physical phenomena [41]; (iii) it has
found applications in describing several other processes due to the fact that some prop-
erties of such processes can be mapped to percolation [78, 106]. Hence, it provides a
simpler and more evident description of the behavior of more involved processes.

1.2.1 Bond and site percolation

The simplest models of percolation are bond and site percolation. In bond percolation,
each link of the network becomes occupied (functional) with a probability $p$, otherwise
it is nonfunctional (or supposedly removed from the network). In site percolation, each
node is occupied with probability $p$. For very small $p$, most of the links are unoccupied
and the network disintegrates into small components whose sizes are vanishingly small compared to the network size. Hence, in this case, the network does not have its global connectivity anymore and its function will be extremely limited if it can function at all.

For $p$ close to 1, almost all the nodes can be reached from the other nodes and the system is practically fully connected. Such nodes are located in a giant connected component (GCC) whose size is the same order of magnitude as the size of the system, or more precisely its size scales linearly with the network size as the number of nodes approaches infinity. There also exist other smaller components (groups of connected nodes), isolated from each other and the GCC, whose sizes are vanishingly small compared to the network size\(^1\). Most of the theoretical approaches to analyze the percolation behavior assume that the network is infinitely large; for an infinitely large network, the GCC is also called the percolating cluster. The network percolating cluster is also infinitely large. The main quantity of interest in a percolation process is the relative size of the GCC which determines what fraction of the network nodes are in the GCC.

It is well-known [32, 118] that there exists a threshold value of $p$ (namely $p_c$) at which a transition occurs between the two regimes mentioned above (i.e., the fully connected regime and the disintegrated network). The critical point is another commonly used name for $p_c$. A lower $p_c$ for a network means that it maintains its global connectivity for more incidents of failures of its constituents and hence it is more robust to random failures or attacks. The location of the transition depends on the structural properties of the network. Although for some types of networks $p_c$ is independent of the type of the percolation [116] (i.e., if it is bond or site percolation), in general, site and bond percolation can have different threshold values [128].

The distribution of the sizes of the small (finite) components may also be a quantity of interest in a percolation process. The shape of this distribution depends on the structural properties of the network. At the critical point, the distribution of component sizes has a power-law form [118]. A system with power-law distribution of its components is referred to as a scale-free system [109]. In a scale free system, a representative size cannot be chosen for the components of the system because the components can have sizes of extremely different magnitudes. For example, at the critical point of

\(^1\)It has been the common belief that complex networks have only one GCC. We will show in Chapter 5 that this is incorrect for some real-world networks.
1. INTRODUCTION

a percolation process, while there are many components with very small sizes, there also exist a very small number of extremely large components. This means that there is an appreciable number of components whose sizes are very different from the average component size.

1.3 The thesis structure

In this thesis, the accuracy of a class of theoretical approaches for percolation behavior is investigated, an improved percolation theory is developed, and the application of percolation theories in studying neural avalanches is assessed. The structure of the thesis is as follows. In Chapter 2, the definitions, concepts and the mathematical background of the material in this thesis are presented. In Chapter 3, a class of theories used for percolation (in particular, for calculation of the size of the GCC) and also for some other processes is considered; then a method is developed to inspect the effect of specific structural properties of the network on the accuracy of these theories. In Chapter 4, a previously unaddressed phenomena causing the inaccuracy of percolation theories is introduced and a novel theory to take into account this phenomenon is developed. In Chapter 5, the application of percolation for understanding better the properties of neural dynamics is considered. In this regard, the calculation of the sizes of neural avalanches is mapped onto the derivation of the distribution of component sizes in a percolation process. Then, the theoretical approaches for calculation of this distribution and their accuracy are investigated. Chapter 6 includes a summary of the results and the conclusion.
In this chapter, we provide an overview of the information necessary or helpful to the rest of the thesis. In particular, we present the definitions for most commonly used concepts and the mathematical background that will be employed frequently in other chapters. The research we present in this thesis is focused on undirected unweighted networks, i.e., networks in which the links do not have a direction since, presumably, the interaction between nodes is mutual and of the same type. Therefore, the information provided in this chapter concerns such networks.

In the following, firstly, we introduce some structural measures of networks. Then, we describe some useful models to construct synthetic networks. Lastly, we overview a class of theories, for bond and site percolation and two other binary-state dynamics.

### 2.1 Structural measures and representations

#### 2.1.1 Adjacency matrix

The adjacency matrix (denoted usually by $A$) of a network includes the information on connections between nodes. For an undirected network, if nodes $i$ and $j$ are connected, the element $A_{ij}$ (and also the element $A_{ji}$) of $A$ takes the value 1, otherwise $A_{ij} = 0$. For a directed network (where each link has a direction that represents an effect one node has on its neighbour but not necessarily vice versa) $A_{ij} = 1$ means that there is
2. MATHEMATICAL BACKGROUND

a directed link from node $i$ to node $j$. Note that for a directed network $A_{ij}$ and $A_{ji}$ are independent.

2.1.2 Nodes’ degrees and related quantities

The simplest, but yet very important, property of a node in a network is the number of links it has to the other nodes or, in other words, the number of its nearest neighbours. This quantity is called the degree of a node. The mean degree $\langle k \rangle$ of a network is then the average of the degree of all nodes.

The degree distribution of a network is denoted by $P(k)$; is the probability that a randomly chosen node has degree $k$. The degree distribution has been a very important quantity in understanding the properties and behaviour of large networks (see Sec. 2.3). For a network with $N$ nodes, $P(k) = N_k / N$, where $N_k$ is the number of nodes with degree $k$. The probability distribution function $P(k)$ is normalized to 1, hence $\sum_k P(k) = 1$; also, the sum $\sum_k kP(k)$ equals $\langle k \rangle$, the average degree of the network.

Another important concept is the degree-degree correlation, which can be represented by the conditional probability $P(k'|k)$, the probability that following a link of a randomly chosen node with degree $k$ one reaches a node with degree $k'$. To calculate $P(k'|k)$ for all existing pairs of degree values, let’s define firstly by $E_{kk'}$ the symmetric matrix whose non-diagonal element with $k \neq k'$ is the number of edges connecting a degree-$k$ node to a node with degree $k'$ and the diagonal element $E_{kk}$ is twice the number of links connecting nodes with identical degree $k$. Then [18]

$$P(k'|k) = \frac{E_{kk'}}{kN_k}.$$  \hspace{1cm} (2.1)

We can use $P(k'|k)$, for example, in calculating other structural measures of the network such as the mean nearest neighbours degree $d$, which is the mean of the average degree of the first neighbours of the nodes [66]:

$$d = \sum_k P(k) \sum_{k'} k'P(k|k').$$  \hspace{1cm} (2.2)

A quantity related to degree-degree correlation is the joint degree distribution characterized by $P(k,k')$ and defined such that $(2 - \delta_{kk'})P(k,k')$ is the probability that the node at one end of a randomly chosen link has degree $k$ and the node at the other end...
2.1 Structural measures and representations

has degree \( k' \). Here, \( \delta_{kk'} \) is the Kronecker delta function. For \( P(k, k') \), we can write
\[ P(k, k') = \frac{E_{kk'}}{\langle k \rangle N_k}. \]  
(2.3)

For certain classes of networks, the percolation behaviour (as well as the properties of some models of disease spread) can be predicted using just the information on the network degree distribution or its degree distribution and degree-degree correlations (see Sec. 2.3).

2.1.3 Generating functions

A class of functions useful to calculate different properties of networks are called generating functions [151]. A generating function, also referred to as a z-transform, for a given probability distribution \( p_s \) is defined as
\[ G(z) = \sum_{s=1}^{\infty} p_s z^s. \]  
(2.4)

For example, the generating functions for the degree distribution \( P(k) \) and excess degree distribution \( q(k) \) of a network are, respectively, \( G_0(z) = \sum_{k=0}^{\infty} P(k) z^k \), and \( G_1(z) = \sum_{k=0}^{\infty} q(k) z^k \). We will use such generating functions at different points of this thesis.

For a given generating function, the probability mass function \( p_s \) can be obtained by differentiating, i.e.,
\[ p_s = \frac{1}{s!} \left[ \frac{d^s G(z)}{dz^s} \right]_{z=0}. \]  
(2.5)

One of the useful properties of generating functions, that we will use in Chapter 5, is the rule for powers of generating functions. This rule is as follows: suppose that \( G(z) \) is the generating function for the probability distribution function \( p_s \). Then, the sum of \( m \) numbers drawn randomly from the distribution \( p_s \) has a generating function \( [G(z)]^m \). For example, if we denote by \( G_0(z) \) the generating function for the degree distribution of a network, the generating function for the sum of the degrees of \( m \) randomly chosen nodes from that network will be \( [G_0(z)]^m \).
2. MATHEMATICAL BACKGROUND

2.1.4 Short loops and clustering

A loop, formally referred to as a simple cycle in the graph theory terminology, is a sequence of adjacent (connected) nodes such that the start and the end nodes are identical and all the other nodes appear only once in the sequence \([18]\). The length of a loop is the number of edges of which it is comprised. In Chapter 2, we will investigate the effect of short loops on the accuracy of theoretical predictions for dynamical processes. Many of the studies (see, for example, Refs. \([63, 97, 113]\)), that investigate the effect of short loops on dynamical processes, focus on the shortest loop (loop with the smallest length), which, in an undirected network, is a triangle (i.e., three nodes are connected by three links).

The clustering coefficient measures the density of triangles in a network and has two slightly different definitions. In one definition \([148]\), firstly a local clustering coefficient \(c_i\) is calculated for each node \(i\); then, the network clustering coefficient is the mean of \(c_i\) values of all nodes. The local measure \(c_i\) is the ratio between \(e_i\), the number of neighbours of \(i\) that are connected to each other\(^1\), and the number of possible connections between the neighbours of \(i\). That is,

\[
c_i = \frac{e_i}{k_i(k_i-1)/2}, \tag{2.6}
\]

where \(k_i\) is the degree of \(i\). Thus, the clustering coefficient of a network with \(N\) nodes is given by

\[
C = \frac{1}{N} \sum_{i=1}^{N} c_i, \tag{2.7}
\]

The other definition of the clustering coefficient, also referred to as the transitivity, considers the total number of triangles in the network with respect to the total number of connected triplets (a connected triplet is a connected subgraph with three nodes and two edges) \([10, 102]\); accordingly, for a given network, the transitivity is defined as:

\[
C = \frac{3 \times \text{the number of triangles}}{\text{the number of connected triplets}}, \tag{2.8}
\]

\(^1\)since the links from \(i\) to two of its connected neighbours together with their connection compose a triangle.

\(^2\)which counts the total number of triangles that can be constructed on \(i\) and a pair of its neighbours.
2.1 Structural measures and representations

In Chapters 3 and 4, when discussions and results on the effect of short loops on networked behaviour are provided, we consider the clustering coefficient as a measure for the density of triangles in networks.

2.1.5 Communities

In a network, a community (also called a module) is a group of nodes such that each node in this group is more likely to be connected to other nodes inside the group than to the nodes outside the community [53, 56, 108]. This means the connections inside communities are denser than those between the communities. Hence, communities represent groups of entities that have more interactions with each other in comparison to their interaction with those outside the community.

Identification of communities in complex networks is an ongoing topic of research in the field and a variety of methods of partitioning a network into communities [56, 84, 126] have been proposed. The performance of these methods have been tested on synthetic benchmark networks [86, 117] to assess their performance and advantages over the other methods.

The quality of partitioning proposed by a community detection method can be tested using a measure called modularity [108, 111] defined as follows: Let’s consider a partition of a network that divides the network into $C$ communities and denote by $c_i$ the label of the community to which node $i$ is assigned. Then the modularity $Q$ is defined as

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_ik_j}{2m}\right) \delta(c_i, c_j),$$

(2.9)

where, $m$ is the total number of links in the network, $A$ is the adjacency matrix, $k_i$ is the degree of node $i$, and $\delta(c_i, c_j)$ is the Kronecker delta function. The value of $Q$ lies between 0 and 1. If the identified communities have denser connections inside, $Q$ will be closer to 1. On the other hand, if the network does not have a community structure or if the detected modules do not match the actual communities then $Q$ will be close to 0.

\[1\] Usually when a community detection method is proposed, the results of applying it to some real-world networks are presented (see, for example, Refs. [53, 56, 132, 154]; however, the quantitative calculation of their performance can be carried out to a better extent on synthetic benchmark networks. This is because the communities are very well defined on benchmarks and also the performance of the method can be tested against different structural variables of networks.
2. MATHEMATICAL BACKGROUND

0. This is because, if two nodes \(i\) and \(j\) belong to a community then we expect that they are more likely to be connected compared to the case that the connections are made randomly (hence, avoiding any connection preference due to a community structure). The probability that the latter case holds, i.e., the two nodes are connected completely at random is \( \frac{k_i k_j}{2m} \). Thus, for each community, it is expected that the sum in Eq. (2.9) is positive; for a denser community, the deviation of the sum from zero will be more appreciable. Hence, for a partitioning with a strong community structure \(Q\) will be close to 1 [108].

2.1.6 Geodesic path

A path between nodes \(i\) and \(j\) in a network is a sequence of consecutively connected nodes through which \(j\) can be reached from \(i\) [116]. The length of a path is the number of links it contains. A geodesic path or shortest path between two nodes is the path whose length is not larger than any of the other paths between them. Accordingly, the geodesic distance or shortest distance between two nodes is the length of the shortest path between them. It is possible that there are more than one geodesic path between two nodes; it is also possible that there is no path between two nodes. The latter happens when the two nodes are located in components that are isolated from each other. In such cases where the geodesic distance is not meaningful, it can be set to infinity, merely for representation purposes.

2.2 Some models of random networks

Some properties of real-world networks are reproduced in networks synthesized according to certain stochastic rules. Investigating the properties and the behaviour of such synthetic random networks helps better understand the behaviour of real-world networks. In this section, several models of random networks are introduced and some of their properties are discussed.
2.2 Some models of random networks

2.2.1 Erdös-Rényi and the configuration model networks

A pioneering model of random networks [18, 27] is that of Erdös and Rényi [48, 49]; in the model of Erdös and Rényi (ER), a network of \(N\) nodes and \(E\) links is constructed by assigning the links to randomly selected pairs of nodes. Alternatively, the network can be formed stochastically by considering \(N\) nodes where each pair is connected with a probability \(p\). In order for the stochastic ER network to produce on average the same number of edges as that of the deterministic ER network (the network with a fixed number of links \(E\)) the relation \(p = 2E / [N(N - 1)]\) should hold. The average degree of an ER network is \(\langle k \rangle = p(N - 1)\) and this network has a binomial degree distribution:

\[
P(k) = \binom{N - 1}{k} p^k (1 - p)^{N - 1 - k}.
\]  

(2.10)

In the limit \(N \to \infty\) and \(\langle k \rangle\) remains constant, i.e., for large ER networks, \(P(k)\) is approximated by a Poisson distribution:

\[
P(k) = \frac{\langle k \rangle^k}{k!} e^{-\langle k \rangle}.
\]  

(2.11)

Another popular method for generating random networks is the configuration model (CM) proposed by Bender and Canfield [16] that constructs networks with a pre-defined degree distribution in the following way [18, 100]: Firstly, a degree sequence for the nodes is derived from the given degree distribution; the sequence includes \(N\) entities, each is a degree value drawn from the degree distribution. Secondly, for each node a number of stubs (each stub can be assumed to be half of a link) equal to its degree is considered. Then, two stubs are chosen randomly and connected together to form a link. This is repeated until no unconnected stub remains. The accidental self links or repeated links are rare for a large number of nodes. Nevertheless, if they occur, either the self links are deleted and each repeated link is considered as one link, or the repeated and self links are rewired with other links until none is left [19].

Both the ER and CM networks are tree-like. The clustering coefficient (Eq. (2.7)) of an ER network is \(C = p = \langle k \rangle / (N - 1)\) since the fraction of connected neighbours of each node is \(p\) [148]. Therefore, as \(N \to \infty\), the density of short loops tends to zero. For CM networks the clustering coefficient is given by [37]:

\[
C = \frac{\langle k \rangle}{N} \left[ \frac{\langle k^2 \rangle - \langle k \rangle^2}{\langle k \rangle} \right]^2
\]  

(2.12)
Thus, for CM networks the clustering coefficient vanishes as $N \to \infty$; nevertheless for finite size networks that have a large \( \frac{(k^2)-\langle k \rangle^2}{\langle k \rangle^2} \) (which is the case for networks that have a degree distribution with a very long tail) $C$ becomes appreciable. We will investigate in Chapter 3, how the accuracy of tree-based theories for network processes is affected by network clustering (i.e., the presence of short loops such as triangles in the network).

Another important characteristic of the ER and CM networks is that they have an uncorrelated structure: In an uncorrelated network the degree of the neighbours of a node is independent of its own degree; in other words, the quantity $P(k'|k)$ is independent of $k$. For an uncorrelated network, given the degree distribution $P(k)$, we have $P(k'|k) = k'P(k')/\langle k \rangle$ [138]. However, many real-world networks have non-trivial degree-degree correlations [105, 138]. Moreover, real-world networks can be composed of communities, while this is not the case for the ER and CM networks. To include these properties of real-world networks, extensions of the configuration model have been proposed that generate networks with a given degree distribution and degree-degree correlation [105], embedded community structure [62, 87], or both [96]. In Sec. 2.3, we overview the theoretical methods that can accurately describe the behaviour of percolation and some other binary-state processes in each of these networks.

### 2.2.2 An ensemble of clustered networks

The models introduced in the previous section create networks that are tree-like (meaning they have a vanishingly small density of short loops in the limit of infinitely large networks). Here, we describe a class of synthetic networks that are clustered, i.e., they possess a considerable density of short loops. This class is defined by a joint probability distribution, $\gamma(k, c)$, which gives the probability that a randomly chosen node has degree $k$ and is a member of a $c$-clique (an all-to-all connected subgraph of $c$ nodes) [63]. In Chapter 3, we will investigate the results for a specific case with $\gamma(3, 3) = 1$ (and $\gamma(k, c) = 0$ for other values of $k$ and $c$), and we dub such networks $\gamma(3, 3)$ networks. In $\gamma(3, 3)$ networks all nodes have degree 3. Each node is part of exactly one triangle and has exactly one single edge that is not part of any triangle. The single edges randomly connect pairs of nodes from different triangles. Figure 2.1 illustrates...
2.2 Some models of random networks

Figure 2.1: Schematic of a $\gamma(3, 3)$ network described in the text. All nodes have degree 3. Each node is part of exactly one triangle and has exactly one single edge that is not part of any triangle. The triangles are randomly connected via single edges. The dashed lines represent connections to nodes not included in the schematic.

A part of a $\gamma(3, 3)$ network. We note that $\gamma(3, 3)$ networks are equivalent to the $p_{1,1} = 1$ case in the clustered random graph model of Refs. [97, 113], where $p_{sx}$ is the probability that a randomly chosen node is part of $t$ different triangles and in addition has $s$ single edges (which do not belong to the triangles).

2.2.3 LFR networks

A difference between some real-world networks and the networks described in Sec. 2.2.1 is that, while it is highly unlikely for an ER or a configuration model network to have a community structure, many real-world networks possess communities [108]. An example of a class of synthetic networks that takes into account the community structure is the LFR networks proposed by Lancichinetti, Fortunato, and Radicchi [87], developed for benchmarking community detection algorithms.

The LFR networks are parameterized by the number of nodes $N$, degree distribution $P(k) \propto k^{-\alpha}$, module (community) size distribution $P(N_m) \propto N_m^{-\beta}$, average degree $k_{av}$, the mixing parameter $\mu$ (the expected fraction of connections each node has to nodes outside its module), the minimum size of modules $N_m^{(\text{min})}$, and the maximum size of modules $N_m^{(\text{max})}$. The method of construction of LFR networks is similar to that of the configuration model in the sense that, according to the network degree sequence, nodes are assigned stubs that will be connected randomly. The difference is that in LFR networks the stubs of each node are divided into two groups: one group can only
be connected to nodes inside the same community and the others can be connected to stubs of the nodes outside the community. More precisely, on average, a fraction $1 - \mu$ of the stubs of each node are connected to nodes inside the same community and the rest are connected to stubs outside their community. We will use LFR networks in Chapter 4 to investigate how the number of inter-community connections can have a strong impact on the percolation properties of modular networks.

### 2.3 Binary-state processes

Percolation models such as bond and site percolation (as introduced in Sec. 1.2.1) can be regarded as an example of a larger class of models referred to as binary-state processes. As the name indicates, a binary-state process is one in which each node can have one of the two possible states. In percolation, for example, a node can have two states: either it is part of the giant connected component (GCC) or it is not in the GCC. In this section, we give an overview of a class of tree-based theories that has found applications in many areas, including the investigation of the properties of binary-state dynamics such as bond percolation [78, 80], susceptible-infected (SI) model for epidemics (Sec. 17.10 of Ref. [116]) and the Ising model (using belief propagation technique or the Bethe-Peierls method) [41].

These theories, which we refer to as adjacency tree-based theories (or $A_{ij}$ theories for short), can often be recognized by the appearance of the adjacency matrix (or its elements $A_{ij}$) in the governing equations. The $A_{ij}$ theories use information on the adjacency of individual nodes and, for each node, they calculate the probability that it is found in a given state. Then the mean of these probabilities over all nodes gives the expected fraction of nodes in the network that are in a given state.

The $A_{ij}$ theories are generally more accurate than their reductions to, for example, degree-based approximations [116, 125] which use excess degree or conditional probability distributions as approximations to the exact connection information provided by the network adjacency matrix. However, $A_{ij}$ theories are not exact on networks with short loops as, in these theories, the state of each node is considered to be independently related to the states of its immediate neighbours (neglecting possible effects of interaction between the neighbours).
In Secs. 2.3.1–2.3.3, we give an overview of the $A_{ij}$ theories for bond and site percolation, SI epidemics, and the Ising model. Then, in Chapter 3, we propose the $L$-cloning method to investigate the accuracy of these theories on clustered networks. In chapter 4, we discuss the accuracy of the message-passing approach for bond percolation [78, 80] on modular networks and improve upon this state-of-the-art $A_{ij}$ theory.

### 2.3.1 Percolation

As described in Sec. 1.3, in bond (site) percolation a fraction $1 - p$ of the links (nodes) in the network are removed and the remaining fraction $p$ of links (nodes) constitute the structure of a new damaged (and possibly disconnected) network; $p$ is called the occupation probability. If $p$ is sufficiently large, a giant connected component (GCC) with a size that scales linearly with the network size exists; otherwise the network is collapsed into isolated small components with vanishing fractional sizes in the limit of infinite network size. The main quantity of interest is the probability $S$ that a randomly selected node is part of the GCC (i.e., the fraction of nodes in the GCC), and how it depends on the occupation probability $p$.

Karrer et al. [80] showed that bond percolation can be formulated as a message-passing process on locally treelike networks. The name “message-passing” refers to the way the calculations are done, in which messages are assumed to be passed among network nodes. In the case of bond percolation, the message is the probability $u_{jk}$ that node $j$ is not connected to the GCC via node $k$ (see Eq. 2.13). We employ the results of Karrer et al. to calculate the size of the GCC for bond and site percolation: Let’s consider a network with adjacency matrix $A$. If we then assume that the network is locally treelike, for bond (site) percolation we have

$$u_{ij} = 1 - p + p \prod_{k \neq i} A_{jk}^{u_{jk}},$$

where the term $1 - p$ denotes the probability that the link $i - j$ (node $j$) is not occupied. The last term in Eq. (2.13) is the product of $p$, the probability that the link $i - j$ (site $j$) is occupied, and the probability that none of the other links leaving $j$ lead to the GCC. Then, for the case of bond percolation, node $i$ is not connected to GCC if none of its links lead to the GCC which happens with probability $U_i = \prod_j u_{ij}^{A_{ij}}$; hence, it is in the GCC with probability $s_i = 1 - U_i$. For site percolation, the probability that $i$
2. MATHEMATICAL BACKGROUND

is part of the GCC is \( s_i = p(1 - U_i) \) where the pre-factor \( p \) is the probability that \( i \) is itself occupied. Now, let us denote by \( \bar{S} \) the expected relative size of the GCC. Then, averaging \( s_i \) over all nodes, we obtain \( \bar{S} \) for bond and site percolation, respectively,

\[
\bar{S}_{\text{bond}} = 1 - \frac{1}{N} \sum_i \prod_j u_{ij}^{A_{ij}},
\]

\[
\bar{S}_{\text{site}} = p \left( 1 - \frac{1}{N} \sum_i \prod_j u_{ij}^{A_{ij}} \right).
\]

Equations (2.14) and (2.15) give the average (expected) relative size of the GCC, \( \bar{S} \), over an infinite number of realizations of the percolation process [128]. Equation (2.14) resembles Eq. (24) of Ref. [78], which is derived for the late time behaviour of the susceptible-infected recovered model using the message-passing approach and is the same as Eq. (7) of Ref. [80].

The above approach is an example of an \( A_{ij} \) theory which uses the information on the exact connectivity of pairs of nodes represented in the network adjacency matrix. If we assume that the edges leaving nodes with degree \( k \) have approximately the same probability \( u_k \) of not leading to the GCC, then Eqs. (2.13) and (2.14) can be reduced to a degree-based approximation as described by Eqs. (1) and (2) of Ref. [145]:

\[
u_k = 1 - p + p \sum_{k'} P(k'|k)(u_{k'})^{k'-1},
\]

\[
\bar{S} = 1 - \sum_k P(k)(u_k)^k,
\]

where \( P(k) \) is the degree distribution, and \( P(k'|k) \) is the conditional probability that a neighbour of a degree-\( k \) node is a degree-\( k' \) node. This assumption is correct when the degree of neighbours of a node with any degree can be well approximated using the conditional probability; this is the case, for example, for configuration model and \( P \) - rewired networks [95]. In the absence of degree-degree correlation, all edges have the same probability of leading to the GCC and Eq. (2.13) is further reduced to Eq. (16.4) or (17.27) of Ref. [116] and Eqs. (2.14) and (2.15) are reduced to Eqs. (16.2) and (17.28) of [116] respectively.

Although Eqs. (2.13)–(2.15) are more general and precise than their aforementioned reductions, they still employ the assumption that the network is locally treelike. Hence, they can be inaccurate for clustered networks [66, 95]. We will discuss about the effect of the clustering on the accuracy of these \( A_{ij} \) theories in Chapter 3.
2.3 Binary-state processes

2.3.2 SI epidemic model

In the Susceptible-Infected (SI) model of disease spread, two possible states are considered for a node: It is either infected or susceptible [116]. In this model, an infected node transmits infection to each of its susceptible neighbours with rate $\beta$, i.e., during the time interval $dt$ a susceptible node catches infection from an infected neighbour with probability $\beta dt$. The macroscopic quantity of interest, which we denote by $I(t)$, is the fraction of network nodes that are in the infected population at time $t$ after the disease begins spreading from a very small fraction of infected nodes. The probability $S_i(t)$ that node $i$ is in the susceptible population at time $t$ can be obtained by using a pair approximation method and assuming that the network is treelike [see Eqs. (17.54) and (17.55) of Ref. [116]]:

\begin{align}
\frac{dS_i}{dt} &= -\beta S_i \sum_j A_{ij} p_{ij}, \\
\frac{dp_{ij}}{dt} &= \beta (1 - p_{ij}) \left[ -p_{ij} + \sum_{k \neq i} A_{jk} p_{jk} \right],
\end{align}

where $p_{ij}$ is the conditional probability that node $j$ is infected given that node $i$ is susceptible. Then, at time $t$, the fraction of the population that is infected is just $I(t) = 1 - \sum_i S_i(t)$. This approach is an $A_{ij}$ theory which employs the full information on the connectivity of nodes, included in the adjacency matrix.

On configuration model networks, on the other hand, a degree-based approximation [11, 12, 123, 124], whose equations are less involved than that of the above $A_{ij}$ theory, can still provide accurate results. In the degree-based approximation proposed by Pastor-Satoras and his colleagues [11, 12, 123, 124], the probability of being infected is considered to be approximately the same for all nodes with the same degree. Then, for a configuration model network with degree distribution $P(k)$, excess degree distribution $q(k)$, and a fraction of initially infected nodes $I_0 = 1 - s_0$, $I(t)$ is given by:

\begin{equation}
I(t) = \sum_{k=1} P(k) \left[ 1 - s_0 (u(t))^k \right],
\end{equation}

where $u$ (which is a function that appears in the solution for the probability $s_k(t)$ that a node with degree $k$ is infected at time $t$) is obtained from

\begin{equation}
\frac{du}{dt} = -\beta u [1 - s_0 G_1(u)],
\end{equation}

where $G_1(u)$ is a generating function.
where $G_1(u)$ is the generating function for the excess degree distribution of the network. The solution of $u(t)$ in Eq. (2.21) is substituted in Eq. (2.20) to give the fraction of infected nodes in this degree-based approximation.

The $A_{ij}$ theory represented by Eqs. (2.18) and (2.19) is more accurate than the degree-based approximation and performs better, for example, on correlated networks. Nevertheless, as Eq. (2.19) assumes that the network is locally treelike, we expect the predictions made by Eqs. (2.18) and (2.19) to be inaccurate for networks that have an appreciable density of short loops. We investigate, in Chapter 3, the accuracy of this theory in clustered networks.

### 2.3.3 The Ising model

The Ising model is a simplified theoretical framework describing the local interactions between magnetic moments of atomic or multi-atomic particles in a solid [82]. In the Ising model each node can be in one of the two spin states called up or down (here denoted by +1 and −1), and the spin of each node is affected by the spins of its neighbours in the network through the Ising Hamiltonian [41, 82]. This Hamiltonian determines the equilibrium configuration of spin states in the network. The total magnetic moment $M$ of the system at equilibrium is defined as the sum of local spins. For a locally treelike network the expected spin of nodes at equilibrium according to the belief propagation algorithm [41] is given by the set of equations described in Sec. VI.A.2 of Ref. [42]. For constant coupling $J$, and in the absence of external magnetic field\(^1\), these equations are:

\[
\begin{align*}
\mu_{ji}(S_i) &= R \sum_{S_j=\pm 1} e^{\beta J S_i S_j} \prod_{n \in N(j) \setminus i} \mu_{nj}(S_j), \\
b_i(S_i) &= R \prod_{j \in N(i)} \mu_{ji}(S_i), \\
M_i &= \sum_{S_i=\pm 1} S_i b_i(S_i),
\end{align*}
\]

where $\beta = 1/(kT)$ with $k$ being the Boltzmann constant, $T$ is the temperature, $R$ is a normalization constant, $S_i$ is the spin value at node $i$, and $\mu_{ji}$ are called *messages* in the

---

\(^1\)These assumptions are not necessary. However we consider the simplest case which makes the equations shorter and simpler, but still supports the argument.
2.3 Binary-state processes

belief propagation method. The product in Eq. (2.22) is over all neighbours of node $j$ except $i$. The fixed point of Eq. (2.22) is used to calculate $b_i(S_i)$, the probability that node $i$ is in state $S_i$ at equilibrium. Accordingly, the expected local magnetic momentum $M_i$ is calculated from Eq. (2.24). The result of the belief propagation method for the Ising model is equivalent to that of the Bethe-Peierls approach [42]. This can be shown by writing $\mu_{ji}$ in a general form as:

$$\mu_{ji}(S_i) = \frac{e^{\beta h_{ji} S_i}}{2 \cosh \beta h_{ji}}, \quad (2.25)$$

which transforms Eqs. (2.22)–(2.24) to the Bethe-Peierls equations of Ref. [42].

These two methods are $A_{ij}$ theories; hence, as in the previous examples, their predictions for clustered networks are prone to errors. We will include as well, in Chapter 3, the investigation of the results obtained from Eqs. (2.22)–(2.24) on clustered networks.
2. MATHEMATICAL BACKGROUND
3.1 Introduction

The behaviour of processes such as percolation (as a model for network resilience) or susceptible-infected (SI) disease spreading depends on the structure of the underlying network on which they operate. Degree distribution and degree-degree correlation (introduced in Sec. 2.1.2) are two important structural properties of a network that influence its dynamics [116]. Moreover, the presence of an appreciable number of short loops (see Sec. 2.1.4) in the network (referred to as clustering) is known to significantly affect dynamics [63, 97, 113, 116]. Real-world networks often have a relatively large clustering coefficient [18, 107, 116], hence they possess a relatively large number of loops of length 3 (i.e., simple cycles of length 3). Some recent network models are able to produce random networks with a desired number of short loops [63, 65, 71, 79] and provide theoretical frameworks for the analysis of bond and site percolation properties, as well as calculating the sizes of $k$-cores and global cascades. However, the state-of-the-art theoretical methods cannot capture the effect of clustering on dynamics running on real-world networks.

Commonly used theories for dynamical processes running on real-world networks are tree-based, i.e., they assume (in one way or another) that the network has a locally treelike structure. Examples of tree-based theories include mean-field theories (which represent the network by its degree distribution or degree-degree correlation)
and a class of pair-approximation methods \[64, 66, 95\]. More sophisticated and accurate tree-based theories (which we call adjacency tree-based or \(A_{ij}\) theories for short) use information on the adjacency of individual nodes. As discussed in Sec. 2.3, these theories have been used to study different dynamical processes including site percolation on directed networks \[130\], bond percolation \[80\] and SI disease spread (see Sec. 17.10 of Ref. \[116\]). The belief propagation, the Bethe-Peierls, and the cavity methods (used, for example, to study the Ising model \[41, 140\]) are also \(A_{ij}\) theories. Although tree-based methods are expected to fail on clustered networks, they perform reasonably well on some clustered networks which casts doubts on the origins of inaccuracy \[95\]. In a study by Gleeson and coworkers \[66\] it was demonstrated that tree-based theories perform worse on clustered networks that have low values of average degree and nearest neighbours degree. This study showed that there are important characteristics other than the network clustering coefficient that affect the accuracy of the tree-based theories. However, the net effect of the short loops on the accuracy of theories was not addressed.

In this chapter\(^1\), we investigate how short loops influence the network dynamics and inspect the accuracy of existing theories for dynamical processes operating on real-world networks. In this regard, we propose the \(L\)-cloned networks that can be constructed for any given real-world or artificial network via a simple process that we call \(L\)-cloning. We show that the ensemble of \(L\)-cloned networks mimics some of the important properties of the original network, including degree distribution and degree-degree correlation between and beyond nearest neighbours, while the network clustering coefficient is reduced by a factor of \(L\) compared to that of the original network. We also introduce an extension of the \(L\)-cloning method in which densities of loops of specified lengths are also preserved. We use this extension to highlight the net effect of short loops on network dynamics. These features make network cloning an ideal framework for isolating the effect of short loops on dynamical processes and for evaluating the performance of theoretical models in the presence of network clustering.

This chapter is organized as follows. In Sec. 3.2, we introduce and discuss the design and some of the structural properties of networks constructed by \(L\)-cloning and

\(^1\)This chapter is based on the paper “Network cloning unfolds the effect of clustering on dynamical processes”, A. Faqeh, S. Melnik, J.P. Gleeson, Phys. Rev. E 91, 052807 (2015) \[52\].
its extensions. In Sec. 3.3, we consider several dynamical processes operating on networks and their corresponding $A_{ij}$ theories. We then investigate the effect of short loops on dynamical processes and the accuracy of corresponding theoretical predictions by applying network cloning methods to some synthetic and real-world clustered networks. In Sec. 3.4, we present a summary and point out potential future work and possible applications of $L$-cloning.

3.2 Cloned networks

In this section, we first describe the ensemble of $L$-cloned networks and explain how they can be constructed from a network of interest (e.g., a real-world network). We then describe their structural properties, focusing on how short loops in the network are affected by the $L$-cloning. Afterward, we introduce two extensions of $L$-cloning, i.e., $L_{X}$-cloning, and $L_{f}$-cloning (defined in Secs. 3.2.3 and 3.2.4).

The main goal of proposing the $L$-cloned networks ensemble is to have networks that are very similar, in most respects, to the network of interest, while having the effect of short loops reduced or practically removed. As we show, this isolates the effect of clustering from other structural properties of the network, and assists in understanding how networked behaviour can be influenced solely by clustering.

In the remainder of this section we give a detailed description of the construction and properties of $L$-cloned networks, but to summarize briefly, $L$-cloned networks have the following features:

1. An $L$-cloned network can be constructed for any given network and any positive integer $L$.

2. An $L$-cloned network has $L$ times as many nodes and edges as the original network.

3. The degree distribution and degree-degree correlation between and beyond the nearest neighbours in an $L$-cloned network are identical to those of the original network.
4. The expected clustering coefficient of an $L$-cloned network is $L$ times smaller than that of the original network. More generally, the density of loops of any fixed length approaches zero for sufficiently large values of $L$.

5. The first three statements above also hold for $L_X$ and $L_f$ cloned networks. On the other hand $L_X$-cloning preserves short loops of specified lengths.

3.2.1 Design and description

In order to build an $L$-cloned network, we start with $L$ clones (i.e., identical copies) of the original network. In particular, for each node $i$ of the original network, there exist $L$ copies $i_1, i_2, ..., i_L$ each placed in one of the $L$ layers as we show in Fig. 3.1(a). An $L$-cloned network is then created by interweaving the $L$ layers by reassigning the existing links uniformly at random, subject to the following restriction: If nodes $i$ and $j$ are connected in the original network, then each copy of $i$ is connected to exactly one copy of $j$ and each copy of $j$ is connected to exactly one copy of $i$ (see Fig. 3.1(b)). The ensemble of $L$-cloned networks is comprised of realizations of this $L$-cloning procedure for a fixed number of layers $L$.

An $L$-cloned network is $L$ times larger than the original network; nevertheless it has the same degree distribution and nearest neighbours degree-degree correlation. Moreover degree-degree correlation beyond nearest neighbours is also preserved by $L$-cloning. That is, for any node in the original network, the set of degrees of its neighbours, its second neighbours, or any group of nodes at a distance $d$ from the node, is identical to that of any of its clones in the $L$-cloned network. In other words, the set of degrees of nodes that are $d$ links away from a node $i$, denoted by $\{K\}_{(i,d)}$, is identical for all clones of the original node $i$ for all $d$ values. Moreover, for each path in the original network, there exist exactly $L$ related paths in the $L$-cloned network; the sequence of nodes in each of these $L$ paths consists of identically ordered copies of nodes of the original path. Hence, such paths have identical sequences of node degrees.

By contrast, an $L$-cloned network has a different density of loops (clustering) than the original network. The density of triangles is decreased: The shuffling of links between layers in the $L$-cloned network breaks the triangles in each layer, constructing a new structure in which triangles are less likely to happen. For example, instead of 3 triangles in Fig. 3.1(a), there is 1 triangle $(i_1 - j_3 - k_2 - i_1)$ in Fig. 3.1(b). In fact, a
3.2 Cloned networks

Figure 3.1: (a) To create an $L$-cloned network, we start with $L$ clones of the original network. In particular, for each node $i$ of the original network, there exist $L$ copies $i_1, i_2, ..., i_L$ each placed in one of the $L$ layers. (b) We then reassign existing links uniformly at random, subject to the following restriction: If nodes $i$ and $j$ are connected in the original network, then each copy of $i$ is connected to exactly one copy of $j$ and each copy of $j$ is connected to exactly one copy of $i$.

Fraction of the broken triangles create new triangles while the rest of them merge into longer loops. In Fig. 3.1(b), for example, a loop of length 6 ($j_1 - k_1 - i_3 - j_2 - k_3 - i_2 - j_1$) is created in addition to the aforementioned new triangle. In Sec. 3.2.2, we show that the density of triangles is decreased by a factor of $L$ and discuss the effect of $L$-cloning on longer loops.

3.2.2 The effect of $L$-cloning on clustering

In this section, we discuss how the density of short loops is affected by $L$-cloning. In particular, we provide an analytical expression for the change in the clustering coefficient [18] and verify it by numerical simulations.

Consider a triangle in the original network. Figure 3.1(a) shows $L$ copies of the triangle each located in one of the $L$ separate layers. Disregarding the node sub-indices for the moment, each of these $L$ triangles consists of a dyad (a connected subgraph consisting of 2 links and 3 nodes) of edges $i - k - j$ completed by a third (thick red) edge $i - j$. Importantly, in the $L$-cloned network depicted in Fig. 3.1(b), the number
3. NETWORK CLONING AND ACCURACY OF $A_{ij}$ THEORIES

Figure 3.2: The dependence of clustering coefficient $C$ of $L$-cloned networks on $L$ (blue circles). Here $C_1$ is the value of the clustering coefficient of the original ($L = 1$) network. The red dashed line is $C_1/L$.

of dyads $i – k – j$ remains the same. However, the probability that each dyad is now completed by a third edge $i – j$ (and hence forms a triangle) is $1/L$ because links are reassigned at random and each copy of node $i$ can be connected to any of $L$ copies of node $j$. This means that the density of triangles in an $L$-cloned network is reduced by a factor of $L$. (Note that, in an $L$-cloned network, the expected number of triangles is the same as in the original one, but an $L$-cloned network has $L$ times as many nodes and edges.)

Therefore, if the clustering coefficient of the original network is $C$, the expected clustering coefficient of its $L$-cloned version is $C/L$. This relation holds for both common definitions of the clustering coefficient, given respectively by Eqs. (2.8) and (2.9) of Ref. [18]. Figure 3.2 illustrates the changes of the clustering coefficient (defined by Eq. (2.9) of Ref. [18]) with respect to $L$ for $L$-cloned networks constructed from a $\gamma(3,3)$ network.

Similarly, the density of loops of length 4 or 5 (or any prime number above 5) in an $L$-cloned network is $L$ times smaller than that in the original network. However, the change in the density of loops of other lengths (e.g., 6 or 8) is more complicated and depends on both $L$ and the density of some shorter loops. The reason for this is that $L$-cloning can transform a loop of length $m$ that exists in the original networks into loops of length $m$, $2m$, ..., $mL$ in the $L$-cloned network.$^1$ Nevertheless, for sufficiently large $L$, $^1$To verify this, suppose that we have a loop of length 3 (a triangle) $i – j – k – i$ in the original network as shown in Fig. 3.1(a). Every loop that can be created by $L$-cloning from this one will have the same
the density of loops of any fixed length will approach zero. Accordingly, by increasing $L$, one can investigate how diminishing the density of short loops in $L$-cloned networks affects networked behaviour such as dynamical processes.

It is also worth mentioning that $L$-cloned networks can be regarded as multilayer networks, which were reviewed and categorized in Ref. [83]. We discuss this in more detail in Appendix A. Using the multilayer representation of $L$-cloned networks is not the main focus of this thesis; however, it can benefit further studies.

### 3.2.3 Extension of $L$-cloning that preserves loops of certain lengths

It is possible to modify the edge reassignment rules of the $L$-cloning process to preserve additional network characteristics. A useful and informative extension of $L$-cloning is one that preserves loops of specified lengths, e.g., triangles and/or loops of length 4 (squares). We refer to this extension of cloning as $L_\mathcal{X}$-cloning, where the set $\mathcal{X}$ specifies the lengths of the loops whose edges are excluded from the reassignment stage of $L$-cloning. For example, to preserve the density of triangles, we exclude edges that are part of one or more triangles from the edge reassignment. Hence, only edges which are not part of any triangle are reassigned (rewired) and interweave the $L$ layers of the network. This type of cloning is accordingly referred to as $L_{-3}$-cloning. By only rewiring the edges that are not part of any triangle, $L_{-3}$-cloning keeps the triangles intact and can only change the densities of longer loops.

Similarly, to keep the density of squares it is sufficient to avoid rewiring the edges that are part of such loops. Moreover, in order to preserve the density of both triangles and squares, the edges that are part of either of such loops are not rewired; we call this $L_{-3,-4}$-cloning. In a similar way, the cloning process can be modified to preserve the densities of loops of any chosen lengths.

It is worth mentioning that, similar to $L$-cloning, these extensions preserve the degree distribution and degree-degree correlation between and beyond nearest neighbor sequence of letters $i - j - k - i$, possibly repeated up to $L$ times with different subscripts (denoting the layer of the copy): $i - j - k - i - j - k - i - ... - i$. Observe, for example, the loops $i_1 - j_3 - k_2 - i_1$ and $i_2 - j_1 - k_1 - i_3 - j_2 - k_3 - i_2$ in Fig. 3.1(b). Conversion of triangles to loops of lengths that are not multiples of 3 would require a different sequence of connected copies of nodes which is impossible by the definition of the $L$-cloning procedure. In general, from loops of length $m$ in the original network, only loops of lengths that are multiples of $m$ can be created in the $L$-cloned ensemble.
bours of the original network from which they are constructed. Additionally, some other characteristics are also preserved. For example, denote the number of triangles and squares that contain node $i$ by $t_i$ and $s_i$ respectively and consider $L_{-3,-4}$-cloning. All clones of node $i$ will have the same $t_i$ and $s_i$ values. Furthermore, the sequence of these values on any path in $L_{-3,-4}$-cloned networks is identical to that of the corresponding path in the original network.

### 3.2.4 Partial $L$-cloning

In this section, we introduce another extension of $L$-cloning that will be used to inspect the effect of short loops. Let $f$ denote the fraction of links that are reassigned (rewired) during network cloning. Accordingly, $f = 1$ in the case of regular $L$-cloning where all the links are rewired, but $f < 1$ when we $L_{-3}$-clone a network with non-zero clustering coefficient because we do not rewire links that form triangles. In order to check whether the fraction of rewired links itself has an effect on the dynamics, we propose another extension of $L$-cloning that we call $L_f$-cloning. For $L_f$-cloning, a fixed fraction $f$ of the links of the original network are randomly selected. Then $L_f$-cloning proceeds as the regular $L$-cloning with the difference that only the randomly selected links are reassigned: For each randomly selected link $i - j$ of the original network, the links between clones of $i$ and $j$ (located in the $L$ identical layers) are reassigned such that each clone of $i$ is connected to exactly one randomly chosen clone of $j$ and vice versa.

An $L_f$-cloning in which the fraction of randomly selected links to be reassigned is identical to the fraction of links reassigned in the $L_{-3}$-cloning (or similarly $L_{-3,-4}$-cloning) is dubbed $L_f(3)$-cloning (or $L_f(3,4)$-cloning). In $L_f$-cloning, the reassigned links are randomly selected from the set of all links. Hence, they may be part of triangles or squares. Therefore, in contrast to $L_{-3}$ and $L_{-3,-4}$-cloning, $L_f$-cloning reduces the density of triangles and squares. Comparing the simulations on the different types of cloned networks (e.g., $L_{-1}$, $L_{-3}$, $L_{-3,-4}$ and $L_f$-cloned networks) constructed from the same original network can provide insight into the effects of short loops on the network dynamics and structure.
3.3 Dynamical processes on networks and $A_{ij}$ theories

In this section, we consider the dynamical processes introduced in Sec. 2.3 (i.e., bond percolation [78, 80], SI epidemics (see Sec. 17.10 of Ref. [116]), and the Ising model [41]), and inspect their behaviour on cloned versions of several synthetic and real-world clustered networks. By comparing the results against theoretical predictions we investigate the effect of clustering on the accuracy of adjacency tree-based ($A_{ij}$) theories developed for these dynamics.

As we described in Sec. 2.3, the $A_{ij}$ theories are generally more accurate than their reductions to, for example, degree-based approximations [116, 125] (see, for example, Secs. 2.3.1 and 2.3.2), which use excess degree or conditional probability distributions as approximations to the exact connection information provided by the network adjacency matrix. Still, $A_{ij}$ theories are not exact on networks with short loops. In this section, we demonstrate that $A_{ij}$ theories, for the aforementioned binary-state processes, are accurate on $L$-cloned versions of a clustered network for sufficiently large $L$. To investigate the effect of short loops on the accuracy of $A_{ij}$ theories, we also use the other variants of cloning described in Secs. 3.2.3 and 3.2.4.

### 3.3.1 Percolation

In this section, we inspect the role of short loops in the accuracy of the $A_{ij}$ theory for percolation, i.e., the message-passing (MP) approach introduced in Sec. 2.3.1. We consider $\gamma(3,3)$ networks (introduced in Sec. 2.2.2) and the western United States power grid networks as examples of, respectively, synthetic and real-world clustered networks. To do so, we apply $L$-cloning to reduce clustering while preserving degree distribution and degree-degree correlation between and beyond nearest neighbours in our networks.

In Fig. 3.3(a), we show how the size of the GCC, in bond percolation, depends on the occupation probability $p$ for the $\gamma(3,3)$ network and its $L$-cloned versions with different $L$, and compare the numerical results with the theoretical predictions of the $A_{ij}$ theory from Eqs. (2.13) and (2.14). The theory does not make an accurate prediction
3. NETWORK CLONING AND ACCURACY OF $A_{ij}$ THEORIES

for the original network ($L = 1$). However, when $L$ is increased the numerical curves approach the result of the $A_{ij}$ theory for the original network. It is worth mentioning that the result of the $A_{ij}$ theory is the same for the original network and any of its $L$-cloned versions.

In the inset of Fig. 3.3(a), we illustrate the effect of $L$-cloning on two $\gamma(3, 3)$ networks with different sizes. Since the curves in the inset that correspond to the same value of $L$ match each other, we conclude that the shift towards the $A_{ij}$ theory in the main Fig. 3.3(a) with increasing $L$ is not due to the change in the network size, but because the density of triangles decreases for larger $L$. Moreover, the results of bond percolation on the $L$-3-cloned versions of $\gamma(3, 3)$ network (not shown) coincide with the numerical curve for the original network regardless of the number of layers. This is due to the fact that triangles are the only short loops present in $\gamma(3, 3)$ networks and $L$-3-cloning keeps triangles intact. Hence, $L$-3-cloned $\gamma(3, 3)$ networks have effectively the same structure as the original $\gamma(3, 3)$ networks, i.e., triangles randomly connected via single edges (see Fig. 2.1). These observations confirm that eliminating the short

![Figure 3.3](image-url)

**Figure 3.3:** Bond percolation on (a) a $\gamma(3, 3)$ network with approximately $10^4$ nodes, and (b) the western US power grid network and their respective $L$-cloned networks. Here $L = 1$ indicates the result of numerical simulations on the original network. The result of the $A_{ij}$ theory is shown by the dashed line. The result of numerical simulation on $L$-cloned networks approaches the $A_{ij}$ theory as $L$ is increased. On the power grid network the prediction made by the $A_{ij}$ theory is different from and more accurate than that of the $P_{kk'}$ theory studied in Refs. [95, 145]. Nevertheless, the two theories have the same result on $\gamma(3, 3)$ networks. The numerical results were derived by averaging over at least 100 realizations of the bond percolation process.
3.3 Dynamical processes on networks and $A_{ij}$ theories

loops in the $\gamma(3, 3)$ network, by $L$-cloning, transforms its structure to a treelike network for which the theory is designed, and that the difference between the theory and the numerics on the original $\gamma(3, 3)$ network is mainly due to the triangles present in the network.

Figure 3.3(b) shows the bond percolation results for the power grid network of the western United States \cite{148} and its $L$-cloned versions. It has been previously observed that mean-field theories are very inaccurate for this network \cite{66, 95}, an example of which is $P_{k,k'}$ theory for bond percolation which employs Eqs. (2.16) and (2.17) \cite{95}. Figure 3.3(b) shows that $A_{ij}$ theory is also inaccurate for this network; however, as $L$ is increased, the numerical result for the $L$-cloned version of the network approaches the prediction of the $A_{ij}$ theory. We conclude that the $A_{ij}$ theory for bond percolation will be accurate for $L$-cloned networks with sufficiently large $L$.

We also run numerical simulations of bond percolation on $L_{-3}$- and $L_{-3,-4}$-cloned versions of the power grid network. In Fig. 3.4 we show that the difference between the $A_{ij}$ theory and numerics on $L_{-3}$-cloned networks decreases as we start to increase the number of layers $L_{-3}$. However, the numerical results on $L_{-3}$-cloned networks do not approach the $A_{ij}$ theory curve even for very large $L_{-3}$. Moreover, the numerical results on $L_{-3,-4}$-cloned networks differ from the $A_{ij}$ theory even more than the results on $L_{-3}$-cloned networks.

Due to the fact that, in $L_{-3}$- and $L_{-3,-4}$-cloned networks, short loops of specified lengths are preserved, to create these networks only the edges that are not part of such loops (hence, only a fraction of the total number of edges) are reassigned (rewired); in contrast all edges are rewired in $L$-cloned networks. For example, in the power grid network, 79% of links are not part of loops of length 3 and 67% of links are not part of loops of length 3 or 4. By running simulations on $L_{f}$-cloned networks, we show that if the same fraction of links, as in $L_{-3,-4}$-cloned networks, for example, are randomly selected and rewired between the layers, the numerical results are appreciably closer to the theoretical prediction, compared to the numerical results for the $L_{-3,-4}$-cloned networks. This implies that the preserved short loops, and not the smaller fraction of rewired links, is the reason that the numerics for $L_{-3,-4}$-cloned (and similarly for $L_{-3}$-cloned) networks do not match the $A_{ij}$ theory even for large number of layers.

In Fig. 3.4 the numerical results for $L_{f(3,4)}$-cloned power grid are compared with the results for other variants of $L$-cloning. It is observed that the result for $L_{f}$-cloned
networks, with a large number of layers $L_f$, matches very well with the $A_{ij}$ theory. This indicates that even with fewer rewired links, since the clustered structure is effectively destroyed for large $L_f$, the numerical results approach the adjacency tree-based theory. Similar results (not shown) were observed for site percolation.

These results demonstrate that triangles and also squares play a significant role in the inaccurate prediction of the $A_{ij}$ theory for the power grid network. On the other hand, as the numerical results for $L_{3,-4}$-cloned networks are closer to the theory than the numerics for the original network, there must exist other sources of error too. One source is the finite size effect,\(^1\) as the original network has a size smaller than any of its

---

\(^1\)The finite size of the networks is known to cause deviations from the theoretical predictions; for
3.3 Dynamical processes on networks and $A_{ij}$ theories

![Figure 3.5: Numerical simulations and $A_{ij}$ theory for SI model on (a) a $\gamma(3, 3)$ network with approximately $10^4$ nodes, and (b) the US power grid network and their corresponding $L$-cloned versions. Here $L = 1$ indicates the results of numerical simulations on original networks. The $A_{ij}$ theory is not accurate on the original clustered networks, however its performance improves with $L$ on $L$-cloned networks.

cloned versions. Other short loops with lengths larger than 4 may also have an effect on the accuracy of the $A_{ij}$ theory.

3.3.2 SI epidemic model

As we discussed in Sec. 2.3.2, since Eq. (2.19) assumes that the network is locally treelike, we expect the predictions made by the $A_{ij}$ theory for SI dynamics (Eqs. (2.18) and (2.19)) to be inaccurate for clustered networks. Although there are cases where the theoretical results and actual behaviour match quite well, such as in Fig. (17.5) of Ref. [116], the accuracy of prediction of Eq. (2.18) is not generally guaranteed for clustered networks. We demonstrate this in Fig. 3.5 by simulations of SI dynamics on the $\gamma(3, 3)$ and the US power grid networks and their corresponding $L$-cloned versions. The theoretical predictions for these two clustered networks are only accurate for sufficiently large $L$.

In order to compare the contribution of short loops of certain length against any other possible source of inaccuracy we employ the extensions of $L$-cloning introduced in Secs. 3.2.3 and 3.2.4. The numerical results of SI dynamics on $L_{-3}$-cloned versions of the $\gamma(3, 3)$ networks (not shown) coincide with the results on the original network. In particular, the finite-size effect is shown to cause error in the $P_{ll,k}$ theory for binary-state dynamics [96].
This indicates that the presence of triangles is the source of inaccuracy of $A_{ij}$ theory for $\gamma(3, 3)$ networks.

The numerics for the SI dynamics on $L_{-3}$ or $L_{-3,-4}$-cloned versions of the power grid network are closer to the $A_{ij}$ theory than those of the original network (Fig. 3.6). However, the theory is still inaccurate on the $L_{-3}$ and $L_{-3,-4}$-cloned networks (even when the number of layers is large) because triangles (or both triangles and squares) are preserved in these networks. The difference between the numerics on $L$-cloned networks and the numerics on $L_{-3}$ or $L_{-3,-4}$-cloned networks is not due to the difference in the number of reassigned (rewired) links because the numerics on $L_{f(3,4)}$-cloned networks (where the fraction of rewired links is the same as in $L_{-3,-4}$-cloned networks) matches the $A_{ij}$ theory very well (Fig. 3.6). Similar to the results shown for bond percolation on the power grid network in Sec. 3.3.1, the SI numerics on $L_{-3}$ and $L_{-3,-4}$

![Figure 3.6](image.png)

**Figure 3.6:** For a sufficiently large number of layers, the results of numerical simulations for SI epidemics on $L_{-3}$ or $L_{-3,-4}$ cloned versions of the power grid network approach a specific curve which is different from the $A_{ij}$ theory curve; this is in contrast to the convergence of the results on $L$-cloned networks to the $A_{ij}$ theory. Although the fraction of reassigned (rewired) links in $L_{-3}$ or $L_{-3,-4}$ cloned networks is smaller than that of $L$-cloned networks, this is not the cause of difference in the corresponding numerical results, as the numerics on $L_{f(3,4)}$-cloned networks (in which the fraction of reassigned links is the same as that of the $L_{-3,-4}$-cloned network) with a large number of layers match very well with the $A_{ij}$ theory.
cloned networks of the power grid are closer to the theoretical prediction; this may be due to finite size effects and the presence of loops of length greater than 4 in the original network. Nevertheless, the loops of lengths 3 and 4 are shown to have a significant effect on the accuracy of the $A_{ij}$ theory for SI dynamics.

### 3.3.3 The Ising model

In this section we investigate the accuracy of the $A_{ij}$ theory for the Ising model introduced in Sec. 2.3.3 on a $\gamma(3, 3)$ network. Figure 3.7 illustrates the behaviour of the magnetization $M$ versus inverse temperature $1/T$ (for $J = 1$ and $k = 1$) in a $\gamma(3, 3)$ network and its $L$-cloned versions; the results are compared to the prediction of the $A_{ij}$ theory. As expected, the tree-based approach does not show the correct behaviour for the original $\gamma(3, 3)$ network. However, for sufficiently large $L$, the prediction of $A_{ij}$ theories matches the numerical simulations on the corresponding $L$-cloned versions of the network. It is worth mentioning that, as expected, the numerical results on $L_{-3}$-cloned $\gamma(3, 3)$ network (not shown) match the numerics on the original $\gamma(3, 3)$ network. This

![Figure 3.7: Magnetization $M$ versus inverse temperature for the Ising model on a $\gamma(3, 3)$ network and its $L$-cloned versions. Here $L = 1$ indicates the result of numerical simulations on the original $\gamma(3, 3)$ network with approximately $10^6$ nodes averaged over 20 realizations of the Ising model. The $A_{ij}$ theory is inaccurate on this network; however, its performance improves appreciably with increasing $L$ on the $L$-cloned versions of this network.](image-url)
implies that the improved agreement between the numerical results on $L$-cloned networks and the $A_{ij}$ theory is not due to the larger size of $L$-cloned networks. Hence, the presence of triangles is the source of inaccuracy in the $A_{ij}$ theory for the Ising model on this network.

### 3.4 Conclusion

We introduced the so-called $L$-cloning of networks, a new technique to create random ensembles of networks with certain properties based on real-world or synthetic complex networks. We discussed in Sec. 3.2 that $L$-cloned networks have degree distribution and degree-degree correlation between and beyond nearest neighbours that are identical to those in the original network from which they are constructed. However, the density of triangles in $L$-cloned networks, and hence the clustering coefficient $C$, is reduced by a factor of $L$. Moreover, the expected density of any short loops approaches zero for sufficiently large $L$. Some useful extensions of $L$-cloned networks, i.e., $L_{-3}$-, $L_{-3,-4}$-, and $L_{ij}$-cloned networks, were also introduced; similar to $L$-cloned networks, these networks have degree distribution and degree-degree correlations between and beyond nearest neighbours identical to those of the original network from which they are constructed. On the other hand, in $L_{-3}$-cloned networks the triangles are preserved, and in $L_{-3,-4}$-cloned networks both triangles and squares are preserved.

We used these properties of network cloning to investigate the effect of short loops on dynamical processes running on networks and to inspect the applicability of tree-based theories to clustered networks (i.e., networks with appreciable density of short loops). In this regard the accuracy of theories for percolation, SI epidemics, and the Ising model were investigated by comparing the theoretical predictions against numerical simulations on examples of clustered networks and their cloned versions.

We demonstrated that $L$-cloned networks with sufficiently large $L$ are the ensemble of networks for which the $A_{ij}$ theories are designed. Hence, by running numerical simulations on $L$-cloned networks, one can predict the outcome and potential benefits of $A_{ij}$ theories for dynamical processes where such theories do not yet exist. In addition, by comparing the numerics for the dynamics on $L_{-3}$ and $L_{-3,-4}$-cloned networks with the numerics on $L$-cloned networks we highlighted the effect of triangles and squares
on the deviation of numerical results from the theoretical predictions. Using \( L_f \)-cloned networks, we ruled out any possible effect due to the different fraction of reassigned (rewired) links in \( L_\cdot \), \( L_{-3} \) and \( L_{-3,4} \)-cloned networks and demonstrated the significant effect of short loops on the accuracy of the \( A_{ij} \) theories.

Nevertheless, it was previously shown [66] that mean-field theories can perform well on clustered networks with high values of mean nearest-neighbour degree. Moreover, the existence of a double percolation phase transition on some clustered networks has recently been reported [33], an effect that is not considered in the tree-based theories. These findings imply that, although network clustering can cause inaccuracy in tree-based theories, its net effect and strength depend also on other factors. In this regard, inspecting the organization of clustered structures in networks (e.g., as in Ref. [34]) can contribute to the understanding of the effect of network clustering on dynamical processes and its mitigation by, for example, high values of mean nearest-neighbour degree.

It is worth noting that the belief propagation and Bethe-Peierls methods used for deriving analytical results for the Ising model have also been employed in many other problems in various areas [41]. According to the tree-based nature of these approximations, the models for which they are used are subject to inaccuracies due to network clustering (short loops). This suggests that network \( L \)-cloning and its extensions can be applied to gain a better understanding of the accuracy of these models as well.

Another potential application of cloned networks is the analysis of the finite size effect on real systems, where a small sample of a large network is available. By inspecting the result of numerical simulations on the sample and on its \( L \)-cloned versions, one can examine the sensitivity of results to changes in network size. If the network of interest is clustered, the changes made in network clustering should be considered in addition to the finite size effect. The extensions of \( L \)-cloning that preserve the density of loops of specified lengths can help achieve this goal. For example, the results of bond percolation on the \( L \)-cloned versions of the western US power grid network (Fig. 3.4) showed that although short loops cause a significant deviation of the theoretical prediction from the actual numerical results, there is an additional factor/phenomenon that produces further inaccuracy of the theory. This interesting finding motivated us to investigate what other property of the networks such as the power grid is not captured.
3. NETWORK CLONING AND ACCURACY OF $A_{ij}$ THEORIES

by the $A_{ij}$ theory for bond percolation, hence resulting in poorer performance of this state-of-the-art theory. We present our findings in this regard in the next chapter.
4.1 Introduction

Percolation theories are among the most studied in network science [6, 116, 135], as well as in several other areas [30, 143], providing insights for a broad range of applications such as robustness of a network to random failures or attacks [28], epidemics in contact processes [101], vaccination strategies [116], neuronal avalanches [58], and stability of gene regulatory networks [142].

Theoretical approaches and extensive numerical simulations play pivotal roles in understanding and describing the behaviour of percolation processes on networks. As described in Sec. 2.3.1, the $p_k$ theory for bond percolation [95, 116], for example, can accurately reproduce the results of numerical simulations on configuration model [100] networks using only the network degree distribution. On networks with degree-degree correlations, more accurate results are obtained using the $P(k, k')$ theory [145] which employs the joint degree distribution. The $P^{ij}_{kk'}$ theory [96] can provide a more accurate description of dynamics on modular networks as it considers the joint degree distributions within and between modules. The message-passing (MP) approach [80] provides more accurate results than the aforementioned theories as it uses the full information on the adjacency of individual nodes, and reduces to the above degree-based approximations in special cases [52] (see Sec. 2.3.1).

In the previous chapter, we investigated the accuracy of the MP approach, as well
as the performance of two other $A_{ij}$ theories for binary-state processes, on clustered networks. We proposed and examined the $L$-cloning methods to investigate the effect of short loops on the behaviour of these dynamics. We showed also that although short loops can have a significant effect on the accuracy of tree-based theories, there can exist other equally important sources of inaccuracy in such theoretical predictions. These results indicate the presence of an unexplained source of error and possibly a phenomenon not captured by the theories.

In this chapter\(^1\), we focus on bond percolation and investigate the accuracy of the MP approach in certain networks. We show the appearance of coexisting percolating clusters (CPCs) in different modular networks, and demonstrate that this phenomenon causes significant errors in the aforementioned theories. We show that CPCs appear in modular networks with limited mixing, i.e., networks with a finite and sufficiently small (limited) number of interlinks between modules. We verify these observations by developing the modular message-passing (MMP) theory which takes into account the presence of CPCs. We show that the MMP theory provides accurate predictions on treelike modular networks with limited mixing and also improves over the predictions of MP on several real-world clustered networks.

This chapter is organized as follows. In Sec. 4.2 we consider sausage-like and LFR [87] networks, two ensembles of synthetic modular networks, on which we will later demonstrate the emergence of CPCs. In Sec. 4.3 we briefly discuss the MP theory [80]. In Sec.4.4, we show the emergence of CPCs on sausage-like networks (SLNs) with limited mixing and discuss why the MP theory performs poorly in the presence of CPCs. In Sec. 4.5, we develop the MMP theory to describe and verify the emergence of CPCs, and show that it provides accurate predictions on SLNs. In Sec. 4.6 we show that the MMP theory performs very well on LFR networks, and in Sec. 4.7 we demonstrate that MMP improves significantly upon the MP theory on several real-world networks. We conclude in Sec. 4.8.

\(^1\)The results of this chapter have also appeared in the paper “Emergence of coexisting percolating clusters in networks”, A. Faqeeh, S. Melnik, P. Colomer-de-Simn, J. P. Gleeson, Phys. Rev E. 93, 062308 (2016) [51].
In this section, we introduce sausage-like networks (SLNs), a simple ensemble of random modular networks on which we can show later in this chapter a rather unsophisticated example of the emergence of CPCs.

To create an SLN we first pick a connected undirected unweighted graph with \(N_m\) nodes. Then we make \(M\) identical copies of that graph (Fig. 4.1(a)) which will become modules in the SLN. We assign to each of these modules a unique label \(m \in \{1, 2, \ldots, M\}\) and connect each pair of modules with consecutive labels \(m\) and \(m + 1\) by \(I\) links, where \(I\) is an even number. To do so, exactly \(I/2\) links are selected randomly from module \(m\). For each selected link \(i_m - j_m\) we consider its copy \(i_{m+1} - j_{m+1}\) in module \(m + 1\), and rewire these two links to create two new links \(i_m - j_{m+1}\) and \(i_{m+1} - j_m\) instead. The resulting SLN is comprised of a chain of modules (Fig. 4.1(b)), each pair of consecutive modules connected with exactly \(I\) interlinks. Moreover, an SLN has a degree distribution and degree-degree correlations between and beyond the nearest neighbours identical to those of the original graph [52]. Similarly, one can construct SLNs from non-identical modules by rewiring links that are not copies but randomly selected from each module.

Because of their structural simplicity, SLNs are suitable to show and describe how CPCs may emerge on networks. Nonetheless, it will also be very useful to investigate LFR networks [87] (introduced in Sec. 4.6) that are a more realistic network ensemble: In LFR networks, the node degrees and community (module) sizes have a power-law

\[\text{(a)}\]
\[\text{(b)}\]

\textbf{Figure 4.1:} (a) A schematic of \(M = 3\) identical copies of a graph, (b) that are connected in a chain to construct an SLN with \(M\) modules.
4. EMERGENCE OF COEXISTING PERCOLATING CLUSTERS

distribution. Hence, LFR networks mimic the heterogeneity of node degrees and also community sizes observed in real-world networks [87]. Also, as opposed to SLNs, in LFR networks, all pairs of modules can be connected. Moreover, in SLNs with very small number of interlinks and sufficiently large modules, each boundary node (a node at one end of an interlink) has, with high probability, only one interlink. However, in LFR networks (or real-world networks) it is possible that a boundary node is connected to more than one node outside its own module.

4.3 The message-passing (MP) approach

In this section, we briefly recap the equations for the MP approach [78, 80] for bond percolation, introduced in Sec. 2.3.1. Then we present an alternative way to write one of the equations for modular networks; hence, we show an assumption embedded in this approach that can cause inaccuracy in the predictions of the MP theory for certain modular networks.

As described in Sec. 2.3.1, in the MP approach, firstly the probability \( u_{ij} \) that a node \( i \) is not connected to the network PC via its link to \( j \) is calculated:

\[
    u_{ij} = 1 - p + p \prod_{k \neq i} A_{jk}, \quad (4.1)
\]

where \( p \) is the occupation probability, and \( A \) is the adjacency matrix of network. Then, the relative size of the network percolating cluster \( S \) is given by

\[
    S = \frac{1}{N} \sum_{i=1}^{N} s_i = 1 - \frac{1}{N} \sum_{i=1}^{N} \prod_{j} u_{ij}^{A_{ij}}, \quad (4.2)
\]

where \( N = M \times N_m \) is the size of the network, and \( s_i = 1 - \prod_{j} u_{ij}^{A_{ij}} \) is the probability that node \( i \) is in the PC. The set of coupled equations for \( u_{ij} \) of all pairs of connected nodes (represented by Eq. (4.1)) is solved by iteration from a random initial guess [52, 80, 128]. Then the solution for \( u_{ij} \) values is substituted in Eq. (4.2) to obtain \( S \).

In a modular network, Eq. (4.2) can also be written in a different format as we will explain in a moment. For a modular network, let us denote by \( S_m \) the fraction of nodes
4.4 Evidence for the emergence of CPCs

In module $m$ that are in the network PC predicted by a given theoretical approach. Hence, for the MP theory

$$S_m = \frac{1}{N} \sum_{i \in \hat{m}} s_i,$$

(4.3)

where $\hat{m}$ denotes the set of nodes located in module $m$. Accordingly, one can rewrite Eq. (4.2) as:

$$\bar{S} = \sum_{m=1}^{M} S_m,$$

(4.4)

where $M$ is the number of modules in the network.

An assumption embedded in Eq. (4.2) or (4.4) is that there is only one percolating cluster in the network and $s_i$ is the probability that node $i$ is part of that cluster. Hence, summing over the $s_i$ of all nodes (explicitly in Eq. (4.2) or implicitly in Eq. (4.4)) gives the probability that a randomly chosen node is part of the network PC, which is the expected relative size of the PC ($\bar{S}$) over an infinite number of realizations of the percolation process [128].

In Secs. 4.4-4.5, we will show that in modular networks with limited mixing the MP theory becomes inaccurate because the above assumption does not hold anymore, and hence, Eq. (4.4) (or equivalently Eq. (4.2)) should be modified appropriately. We introduce the MMP theory that corrects this assumption by taking into account the emergence of CPCs in networks with limited mixing and provides a replacement for Eq. (4.2).

4.4 Evidence for the emergence of CPCs

In this section we present the results of direct numerical simulations on SLNs and investigate the accuracy of the MP theory for these networks. The numerical $\bar{S}$ values are obtained from at least 500 realizations of the Newman-Ziff algorithm [119]. In Figs. 4.2(a)-4.2(c), we illustrate the bond percolation results on SLNs constructed from a 3-regular graph. Numerical simulations show that the behaviour of $\bar{S}$ in the SLNs depends substantially on the number of interlinks $I$ and the number of modules $M$. Surprisingly, the result of the MP theory for the SLNs is independent of $I$ and $M$.
4. EMERGENCE OF COEXISTING PERCOLATING CLUSTERS

Figure 4.2: (a) Bond percolation results on SLNs constructed from a 3-regular graph of size $N_m = 1000$ for different $I$. (b) The mean absolute error $E$ versus $I$ and (c) versus $N_m$ for the MP theory on such 3-regular SLNs. In (a) $M = 50$, in (b) $N_m = 5000$, and in (c) $M = 10$.

and coincides with its prediction for a single 3-regular graph (Fig. 4.2(a)); it is worth
mentioning that in such SLNs the result of MP theory is the same as the results of
any of its degree-based reductions mentioned above (e.g., $p_k$ theory, $P(k,k')$ theory,
etc.). The numerical results, on the other hand, deviate from the theoretical prediction
as the number of interlinks $I$ is decreased (Fig. 4.2(a)). We quantify the difference
between the theoretical and numerical results by calculating the mean absolute error
between the two: $E = 1/R \sum_{j=1}^{R} | \overline{S}_{MP}(p_j) - \overline{S}_{num}(p_j) |$, where the sum is over $R = 100$
equally spaced occupation probabilities $p_j = j/R$.

Figure 4.2(b) shows that the error increases dramatically for lower number of interlinks $I$; on the other hand, the error increases only slightly for smaller module sizes
$N_m$ (Fig. 4.2(c)). Figure 4.2(c) highlights also that the error increases for lower $I$ rather
4.4 Evidence for the emergence of CPCs

than for lower ratio of $I$ to all edges.

To understand these observations, it is necessary to inspect also the numerical results for $S$, the fractional size of the network PC for single realizations of the bond percolation process. Figure 4.3(a) shows the results for a 3-regular SLN with $I = 4$; the MP theory overestimates $\bar{S}$, which is the expected (average) value of $S$. For a sufficiently large network, we normally expect that $S$ fluctuates slightly around $\bar{S}$. Surprisingly, we observe in Fig. 4.3(a) that, for a fixed value of $p$ (in a certain interval), $S$ can take one of several possible values which can be significantly different from $\bar{S}$. The different possible values for $S$ can be explained by considering the contribution $S_m$ (calculated from Eq. (4.3)) of each of the modules of the network to the value of $S$. According to the MP theory (see Eq. (4.4)), the expected value of $S$, i.e., $\bar{S}$, is the sum of $S_m$ of all the modules. However, only the largest possible value of $S$ in Fig. 4.3(a) matches this prediction. The lowest possible value of $S$, on the other hand, coincides with $S_m$ of only one module$^1$. The next 3 larger possible values of $S$ coincide with the sum of $S_m$ for 2, 3, and 4 modules respectively (Fig. 4.3(a)). These results suggest that Eq. (4.4) of the MP theory is not a correct way to calculate the size of the PC here, and we need an alternative description of how the PC is formed in this network.

Consider another example of an SLN consisting of one 3-regular module and one 4-regular module each having the same number of nodes $N_m$, and let us denote by $S_1$ and $S_2$ the $S_m$ of the 4-regular and 3-regular graph respectively. For this SLN, $\bar{S}$ from numerics and from the MP theory both match $S_1$ up to $p_{3\text{reg}} = 0.5$, the percolation threshold of a 3-regular graph. Above this value, the MP prediction deviates from the numerical result (Fig. 4.3(b)). This deviation can be better understood by looking at single realizations of the Newman-Ziff algorithm [119] where, starting with no occupied links, we occupy links one by one in random order. As the network is large, $p$ is approximately equal to the fraction of occupied links. In Fig. 4.3(b), we can observe that up to $p_{3\text{reg}}$ single realization values of $S$ match $\bar{S}$. However, above $p_{3\text{reg}}$, while the predicted $\bar{S}$ is $S_1 + S_2$, $S$ will remain equal to $S_1$ until some larger value of $p$ and then suddenly jump to $S_1 + S_2$.

This implies that although $S_2$ is macroscopic for $p \geq p_{3\text{reg}}$, it does not represent the probability that a node in module 2 belongs to the network PC, and accordingly $S_1 + S_2$

$^1$Since all of the modules of this network are similar, i.e. they are 3-regular graphs with identical sizes, and as the modules are sufficiently large, $S_m$ is approximately the same for all the modules.
4. EMERGENCE OF COEXISTING PERCOLATING CLUSTERS

Figure 4.3: (a) Percolation on an SLN with $M = 5$ and $N_m = 20000$. For each $p$, $S$ is shown separately for 20 single realizations of percolation (triangles). $S$ does not match $\bar{S}$ (circles) averaged over 500 realizations; nevertheless, it coincides with one of the 5 possible values denoted by the dashed line (MP theory) or the solid lines. (b) Percolation on an SLN with a 3-regular and a 4-regular module, $I = 4$, and $N_m = 20000$. Blue lines represent the results from 20 single realizations of the Newman-Ziff algorithm. The vertical line on the left (right) denotes the percolation threshold of a 4-regular (3-regular) graph. (c) The MMP prediction for $\bar{S}$ on the SLN of panel (b) matches the numerical results on this SLN (circles) as well as on similar SLNs with identical $I$ and any sufficiently large $N_m$. (d) Predictions for the SLN of panel (a). The 5th and 95th percentiles of $S$ are indicated by the green shade (numerical results) and by the dashed lines (MMP theory).

is not the expected size $\bar{S}$ of the network PC. In fact, the nodes in module 2 are part of a PC with size $S_2$ and those in module 1 are part of another PC with size $S_1$. In a single realization these two PCs may be unconnected. Then, when more links are occupied one at a time, at a value of $p$ with $p \geq p_{3\text{reg}}$, the two PCs become suddenly connected, and the size of the network PC changes abruptly from $S_1$ to $S_1 + S_2$. We refer to such PCs, which exist independently of each other in the network and are connected with
4.5 The modular message-passing (MMP) approach

Next, we develop the modular message-passing (MMP) theory to describe and verify the phenomenon of CPCs appearing in networks. The two main assumptions are (i) modules (which are internally well connected) can percolate independently (hence, the appearance of independent CPCs), and (ii) CPCs of neighbouring modules are connected with probability $\eta \leq 1$. This is a new concept different from the common assumption (see for example [33, 96, 116]) that there exists only one monolithic PC in a network; here we show that in networks with limited mixing, the network PC is polylithic, i.e., constituted by connected CPCs.

For SLNs we assume that each boundary node (a node with links to other modules) has exactly one link to a neighbouring module. Then for an SLN with two modules, our MMP theory is comprised of two simple equations. First we calculate $\eta_{mn}$, the probability that the CPCs of the two modules $m$ and $n$ are connected:

$$\eta_{mn} = 1 - (1 - p v_m v_n)^I. \quad (4.5)$$

Here $p$ is the occupation probability and $v_m$ is the probability that a boundary node in $m$ is part of the CPC of $m$. Then $(1 - p v_m v_n)^I$ gives the probability that the two CPCs are not connected via any of the $I$ interlinks. Equation (4.5) is independent of $N_m$...
4. EMERGENCE OF COEXISTING PERCOLATING CLUSTERS

and, hence, in the thermodynamic limit \((N \to \infty)\), if \(I\) is a sufficiently small constant, (independent) CPCs emerge, connected with a probability \(\eta_{mn}(p) < 1\). On the other hand, for large \(I\), since \(\eta_{mn} \to 1\), the two CPCs are connected with high probability, leading to a monolithic PC. For the SLNs described above, \(v_m\) and \(v_n\) can be simply calculated using the \(p_k\) theory (see Appendix B). For an SLN with two modules, the size of network PC is then:

\[
\bar{S} = \eta_{12} (S_1 + S_2) + (1 - \eta_{12}) S_1, \tag{4.6}
\]

where \(S_1\) and \(S_2\) are the fractional sizes of, respectively, the larger and the smaller CPC of the SLN (calculated using the MP theory or one of its appropriate degree-based reductions such as the \(p_k\) theory). Figure 4.3(c) shows that the prediction of the MMP theory (Eqs. (4.5)-(4.6)) matches perfectly the numerical result for the SLN of Fig. 4.3(b).

For SLNs with more than two modules Eq. (4.5) can still be used to calculate \(\eta_{mn}\) for each pair of modules \(m\) and \(n\). In the case where boundary nodes have more than one interlink, or when the connection pattern of the modules can not be well approximated using the \(p_k\) theory, Eq. (4.5) should be extended to include more information on the network structure. We can use the full information on the adjacency of individual nodes to write a general formula for the connection probabilities \(\eta_{mn}\) between CPCs in a treelike network:

\[
\eta_{mn} = 1 - \prod_{i=1}^{I} \left[ 1 - \left(1 - \prod_{j \in A_m(i)} u_{ij} \right) \left(1 - \prod_{k \in A_n(i)} u_{ik} \right) \right], \tag{4.7}
\]

where \(i\) is a boundary node of module \(m\) and \(A_m(i)\) denotes the set of neighbours of \(i\) in module \(m\). Here, \(u_{ij}\) and \(u_{ik}\) are the probabilities that \(i\) is not connected to the CPC of, respectively, modules \(m\) and \(n\) via its links to nodes inside each of those modules; the results of Eq. (4.1) of the MP theory are substituted for \(u_{ij}\) values in Eq. (4.7)\(^1\). Hence, within the first (second) set of parentheses in Eq. (4.7) is the probability that \(i\) is in the CPC of \(m\) (the CPC of \(n\)), and in the square brackets we have the probability that the two CPCs are not connected via the interlinks of \(i\).

\(^1\)Here we assume that \(u_{ij}\), calculated using Eq. (4.1), is the probability that node \(i\) is connected, through its link to \(j\), to the CPC of the module in which \(j\) is located.
4.6 Results on LFR benchmark networks

To calculate $\bar{S}$ for networks with more than two modules Eq. (4.6) should be extended as well. In such networks, different CPCs (with sizes $S_m$) are connected together with some probability, creating larger polylithic PCs. The polylithic PC $l$ has a size $S_{l}^{(\text{pol})} = \sum_{m \in \hat{l}} S_m$, where $\hat{l}$ denotes the set of CPCs that constitute $l$. Then, in a single realization, at a fixed value of $p$, $S$ is $\max_l \left( S_{l}^{(\text{pol})} \right)$, i.e., the size of the largest polylithic PC. Then the expected size of the network PC is

$$\bar{S} = \sum_{S} P_S S,$$

(4.8)

where $P_S$ is the probability that in a single realization the size of the largest polylithic PC is $S$. To calculate $P_S$, for each $p$ we first calculate $S_m(p)$ values using the MP theory or an appropriate degree-based reduction of MP. Then we assume a meta-network which is comprised of meta-nodes; each meta-node $m$ represents a module of the original network and has a weight $S_m(p)$. For a sufficiently large number of meta-network realizations, we calculate $P_S$ using the Newman-Ziff algorithm [119] with the following modifications: (i) each link between the meta-nodes $m$ and $n$ is added with probability $\eta_{mn}(p)$, and (ii) the size of a cluster $l$, comprised of connected meta-nodes, is the sum of the weights of meta-nodes it includes, i.e., $S_{l}^{(\text{pol})}$.

Figure 4.3(d) illustrates that, using Eqs. (4.7)-(4.8), the MMP theory performs very well for the SLN of Fig. 4.3(a). As shown in Figs. 4.3(a) and 4.3(b), in the presence of CPCs, $S$ can deviate considerably from $\bar{S}$. In the MMP theory, the variability of $S$ is determined by $P_S$ defined above. The percentiles calculated from $P_S$ match very well the numerical values in SLNs (see Fig. 4.3(d) for example) which confirms that the high variability of $S$ originates from the presence of CPCs that suddenly merge together. Similar results (not shown) are obtained for SLNs with different values of $I$, $M$, $N_m$, and also for SLNs constructed from modules with a heterogeneous structure, i.e., with a power-law degree distribution.

Recall that we introduced LFR networks in Sec. 2.2.3. As LFR networks were developed to better reflect degree distribution and modular structure found in real-world
networks [87], in this section we inspect the results on these networks to better understand the behaviour of percolation on real-world modular networks. In Table 4.1, we summarize some structural information of the networks exemplified here and in the following sections and show the measurement of the performance of the MP and MMP theories for these networks. The predictions of the MMP method in this section are from Eqs. (4.7) and (4.8).

Figure 4.4 illustrates the results on LFR networks 1 and 2 of Table 4.1; all the structural measures of these two networks are approximately the same, except for the significantly different average number of interlinks $\langle I \rangle$, dictated by their different mixing parameters $\mu$. Both the MP and MMP theories perform very well in predicting $\overline{S}$

![Diagram](image)

**Figure 4.4:** Bond percolation results on two LFR networks which have similar structural properties except for the mixing parameter $\mu$ which determines the number of interlinks. (a) The MP theory performs very well for the LFR network with $\mu = 0.05$ as the number of interlinks is large in this network (see Table 4.1); in this case the prediction of the MMP theory matches that of MP. (b) For the network with $\mu = 0.0005$ the prediction of MP for $\overline{S}$ is inaccurate; however MMP predicts accurately $\overline{S}$ and also percentiles for the distribution of $S$. The 5th and 95th percentiles of $S$ are indicated by the green shade (numerical results) and by the dashed lines (MMP theory). The two networks have a degree distribution $P(k) \propto k^{-3}$, and community size distribution $P(N_m) \propto N_m^{-2.5}$. 

62
### 4.6 Results on LFR benchmark networks

<table>
<thead>
<tr>
<th>No.</th>
<th>Network</th>
<th>$N$</th>
<th>$M$</th>
<th>$\langle N_m \rangle$</th>
<th>$\sigma_{N_m}$</th>
<th>$N_m^{\text{min}}$</th>
<th>$N_m^{\text{max}}$</th>
<th>$k_{av}$</th>
<th>$C$</th>
<th>$\langle I \rangle$</th>
<th>$\sigma_I$</th>
<th>$E_{MP}$</th>
<th>$E_{MMP}$</th>
<th>$\sigma_{E_{MMP}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fig. 4.4(a), $\mu = 0.05$</td>
<td>10000</td>
<td>10</td>
<td>10000</td>
<td>3775.1</td>
<td>6732</td>
<td>16104</td>
<td>5.17</td>
<td>0.0043</td>
<td>2605.8</td>
<td>1003.4</td>
<td>0.0039</td>
<td>0.0019</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Fig. 4.4(b), $\mu = 0.0005$</td>
<td>10000</td>
<td>10</td>
<td>10000</td>
<td>3784.2</td>
<td>6118</td>
<td>16017</td>
<td>5.10</td>
<td>0.0038</td>
<td>27.4</td>
<td>11.4</td>
<td>0.026</td>
<td>0.0006</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Fig. 4.5(a), $\mu = 0.01$, $N_m^{\text{min}} = 100$</td>
<td>5000</td>
<td>10</td>
<td>135.1</td>
<td>28.3</td>
<td>100</td>
<td>194</td>
<td>8.01</td>
<td>0.22</td>
<td>10.8</td>
<td>3.3</td>
<td>0.067</td>
<td>0.0011</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Fig. 4.5(b), $\mu = 0.01$, $N_m^{\text{min}} = 10$</td>
<td>5000</td>
<td>10</td>
<td>244</td>
<td>20.5</td>
<td>22.7</td>
<td>10</td>
<td>189</td>
<td>8.06</td>
<td>0.61</td>
<td>1.9</td>
<td>3.6</td>
<td>0.3378</td>
<td>0.0022</td>
</tr>
<tr>
<td>5</td>
<td>Fig. 4.5(c), $\mu = 0.1$, $N_m^{\text{min}} = 10$</td>
<td>5000</td>
<td>10</td>
<td>241</td>
<td>20.8</td>
<td>20.8</td>
<td>10</td>
<td>199</td>
<td>8.03</td>
<td>0.51</td>
<td>16.7</td>
<td>26.4</td>
<td>0.027</td>
<td>0.012</td>
</tr>
<tr>
<td>6</td>
<td>Fig. 4.5(d), $\mu = 0.1$, $N_m^{\text{min}} = 50$</td>
<td>5000</td>
<td>10</td>
<td>61</td>
<td>82</td>
<td>32.5</td>
<td>50</td>
<td>194</td>
<td>8.06</td>
<td>0.29</td>
<td>65.4</td>
<td>26.1</td>
<td>0.0069</td>
<td>0.0016</td>
</tr>
<tr>
<td>7</td>
<td>Fig. C.1(a), $\mu = 0.05$, $N_m^{\text{min}} = 10$</td>
<td>5000</td>
<td>10</td>
<td>127</td>
<td>39.4</td>
<td>48.3</td>
<td>10</td>
<td>361</td>
<td>3.11</td>
<td>0.15</td>
<td>6.1</td>
<td>8.2</td>
<td>0.090</td>
<td>0.0061</td>
</tr>
<tr>
<td>8</td>
<td>Fig. C.1(b), $\mu = 0.01$, $N_m^{\text{min}} = 50$</td>
<td>50000</td>
<td>10</td>
<td>142</td>
<td>352.1</td>
<td>280.3</td>
<td>50</td>
<td>994</td>
<td>5.10</td>
<td>0.061</td>
<td>18.2</td>
<td>3.3</td>
<td>0.039</td>
<td>0.00034</td>
</tr>
<tr>
<td>9</td>
<td>Real world</td>
<td>4039</td>
<td>7</td>
<td>577</td>
<td>432.3</td>
<td>60</td>
<td>1341</td>
<td>43.69</td>
<td>0.61</td>
<td>89.1</td>
<td>98.2</td>
<td>0.014</td>
<td>0.0019</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Fig. D.1, Facebook [94]</td>
<td>1287</td>
<td>3</td>
<td>429</td>
<td>524.5</td>
<td>79</td>
<td>1032</td>
<td>37.22</td>
<td>0.50</td>
<td>72.7</td>
<td>47.3</td>
<td>0.0067</td>
<td>0.0035</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Fig. 4.6, AS Internet [92]</td>
<td>6474</td>
<td>7</td>
<td>924.9</td>
<td>1320.1</td>
<td>5</td>
<td>3695</td>
<td>3.88</td>
<td>0.25</td>
<td>254.9</td>
<td>289.4</td>
<td>0.00078</td>
<td>0.00097</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Real world</td>
<td>4941</td>
<td>33</td>
<td>149.7</td>
<td>56.2</td>
<td>66</td>
<td>260</td>
<td>2.67</td>
<td>0.080</td>
<td>72.7</td>
<td>47.3</td>
<td>0.0067</td>
<td>0.0035</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Fig. 4.6(b), Western US power grid [148]</td>
<td>4941</td>
<td>7</td>
<td>149.7</td>
<td>56.2</td>
<td>66</td>
<td>260</td>
<td>2.67</td>
<td>0.080</td>
<td>72.7</td>
<td>47.3</td>
<td>0.0067</td>
<td>0.0035</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Fig. 4.6(c), Polish power grid [47, 156]</td>
<td>3374</td>
<td>26</td>
<td>129.8</td>
<td>42.1</td>
<td>82</td>
<td>258</td>
<td>2.41</td>
<td>0.019</td>
<td>15.4</td>
<td>6.8</td>
<td>0.033</td>
<td>0.012</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Fig. D.1(d), NetSci coauthorships [110]</td>
<td>1174</td>
<td>44</td>
<td>26.7</td>
<td>28.9</td>
<td>2</td>
<td>83</td>
<td>2.41</td>
<td>0.017</td>
<td>6.1</td>
<td>7.9</td>
<td>0.047</td>
<td>0.018</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Fig. D.1(e), E-roads [146]</td>
<td>512</td>
<td>7</td>
<td>73.1</td>
<td>23</td>
<td>51</td>
<td>123</td>
<td>4.82</td>
<td>0.74</td>
<td>14.6</td>
<td>2.9</td>
<td>0.041</td>
<td>0.023</td>
<td></td>
</tr>
</tbody>
</table>

**Table 4.1:** Basic summary statistics for the networks that we used in this chapter. We have treated all real-world data sets as undirected, unweighted networks and have computed the following properties: total number of nodes $N$; the number of modules in the network $M$; the average size of modules $\langle N_m \rangle$ and its standard deviation $\sigma_{N_m}$; the size of the smallest module $N_m^{\text{min}}$ and the size of the largest module $N_m^{\text{max}}$ observed in the network; mean degree $k_{av}$; clustering coefficients $C$ defined by Eq. (2.10) of [107]); the average number of interlinks $\langle I \rangle$ and its standard deviation $\sigma_I$; the difference between the theoretical and the numerical results for the MP theory $E_{MP}$, and for the MMP theory $E_{MMP}$ measured by the mean absolute error $E_{\text{theo}} = 1/R \sum_{j=1}^{R} |S_{\text{theo}}(p_j) - S_{\text{num}}(p_j)|$, where the sum is over $R = 60$ equally spaced occupation probabilities $p_j = j/R$. 

Real world
4. EMERGENCE OF COEXISTING PERCOLATING CLUSTERS

for the LFR network of Fig. 4.4(a) that have a large $\langle I \rangle$. However, the MP method becomes inaccurate for the LFR network of Fig. 4.4(b) which has a very small $\langle I \rangle$. On the other hand, the MMP method provides an accurate prediction for the LFR network of Fig. 4.4(b). Moreover, the MMP method accurately predicts the percentiles for the distribution of $S$ in this network (Fig. 4.4(b)). The networks of Fig. 4.4 and their modules are sufficiently large, so they have very low clustering coefficients. Hence, it is evident that the deviation of the prediction of the MP theory in Fig. 4.4(b) is not due to the presence of short loops, but—as verified by the high accuracy of the MMP theory—it is due to the emergence of CPCs in the network of Fig. 4.4(b).

An assumption made in the MMP theory was that in networks with finite number of interlinks, modules can have (independent) CPCs. To have a percolating cluster on its own, a module should be sufficiently large. However, real-world and LFR networks can have a wide distribution of module sizes where modules of very small sizes are likely to appear. In Figs. 4.5(a)-4.5(d) we investigate the percolation results on LFR networks that contain relatively small modules. The LFR networks of Fig. 4.5 all have the same size, degree distribution $P(k) \propto k^{-2}$, community size distribution $P(N_m) \propto N_m^{-3}$, and approximately the same average degree, but the minimum possible module size and/or the mixing parameter are varied.

The MP method performs poorly on the LFR network of Fig. 4.5(a) which has a small number of interlinks ($\langle I \rangle=10.8$). Although the modules in this network are not very large ($100 < N_m < 200$), MMP accurately predicts $\bar{S}$ and the percentiles for the distribution of $S$. This shows that the modules do not need to be extremely large in order for the MMP theory to perform accurately. In fact, the MP theory does a surprisingly good job in obtaining the $S_m$ values for such small modules; however, it misses the CPCs formed independently in the modules of this networks. Interestingly, this network has a relatively large clustering coefficient of 0.22; however, the accurate prediction of the MMP theory shows that CPCs are the main source of inaccuracy of the MP method for this network and not the presence of the short loops.

The LFR networks of Figs. 4.5(b) and 4.5(c) have extremely small modules with as few as 10 nodes. Despite the very small module sizes, the MMP method still performs very well for the network of Fig. 4.5(b) which has a mixing parameter of $\mu = 0.01$ but its accuracy decreases for the network of Fig. 4.5(c) which has a larger $\mu = 0.1$. 
4.6 Results on LFR benchmark networks

Figure 4.5: Percolation results on four LFR networks with degree distribution $P(k) \propto k^{-2}$, community size distribution $P(N_m) \propto N_m^{-3}$, similar average degrees $k_{av} \approx 8$ and $N_{m}^{(\text{max})} \approx 200$, but different values of $N_{m}^{(\text{min})}$ and/or $\mu$ (see Table 4.1). (a) The MMP theory can perform very well in a network whose modules are not so large ($100 < N_m < 200$). (b) Even on a network with many modules with sizes as small as 10 nodes the MMP theory can be accurate. This implies that, despite the extremely small module sizes, the assumption that modules can have independent CPCs is a very good approximation in this network that has (because of its relatively low mixing parameter $\mu = 0.01$) a very small number of interlinks. (c) On a network with extremely small modules and a mixing parameter $\mu = 0.1$ (larger by an order of magnitude), the MMP theory becomes inaccurate. (d) For $\mu = 0.1$, the MMP theory performs very well again if the minimum size of modules is set to 50 nodes. For the network of panel (b), we also ran the percolation process separately on the largest component of the network (which had a relative size $\frac{3865}{5000}$). The $S$ of this component, when multiplied by $\frac{3865}{5000}$, matches the $S$ of the network (for both the numerics (circles) and the MMP theory (blue solid line)). The $S$ of this component, multiplied by $\frac{3865}{5000}$, obtained by the MP theory is shown by the dotted green line.

We will first discuss the results of Fig. 4.5(b), and then explain the effect of larger $\mu$ in LFR network of Fig. 4.5(c).

There is another interesting point about the results on the network of Fig. 4.5(b): this network has isolated components\(^1\). Hence at $p = 1$ the value of $S$ is less than 1.

\(^1\)This network is comprised of 84 isolated components. The largest component has 3865 nodes. The other components are very small and their average size is 13.7 nodes with a standard deviation of 4.5. Out of the 244 modules of the network, 158 modules belong to the largest component and their average size is 15.8 with a standard deviation of 24.9.
4. EMERGENCE OF COEXISTING PERCOLATING CLUSTERS

and equals the relative size of the largest component of this network. Nevertheless, the MP method predicts that $\bar{S} = 1$ at $p = 1$; in fact, MP substantially overestimates $\bar{S}$ everywhere above the percolation threshold of this network. The PC of this network is that of the largest component as, for all $p$, the $\bar{S}$ of the network matches the $\bar{S}$ of the largest component scaled by the relative size of this component (i.e., by $3865/5000$) (Fig. 4.5(b)). The MMP method accurately predicts the $\bar{S}$ for this network and if we run the MMP method only on the graph of the largest component it also accurately predicts the $\bar{S}$ of this graph (Fig. 4.5(b)). On the other hand, the MP method is inaccurate for both the whole network and the largest component of the network.

The MP theory predicts an $\bar{S}$ for this network that is appreciably larger than the $\bar{S}$ it predicts for the largest component (Fig. 4.5(b)). This difference between the two $\bar{S}$ values is due to the contribution from the nodes isolated from the largest component and located in the other smaller components. It is worth noting that, in this network, the components other than the largest component have very small sizes (on average 13.7 nodes with a standard deviation of 4.5). This indicates that although the nodes in components other than the largest component can not be in the network PC, the MP theory still obtains a nonzero $s_i$ (the probability that node $i$ is in the network PC) for such nodes. Then, summing over all the $s_i$ (Eq. (4.2) of the MP method) leads to the incorrect prediction $\bar{S}(p = 1) = 1$. This shows that the $s_i$ calculated by the MP method can be a local measure: in the MP theory a node, well connected to the other nodes in an isolated small component, may obtain an appreciable probability of being connected to the “network PC”. This helps to understand why the MP method becomes inaccurate when the CPCs emerge: the sum in Eq. (4.2) (or Eq. (4.4)) misses the fact that the $s_i$ values calculated using Eq. (4.1) can be local measures representing the probability that $i$ is in the CPC of its module and not necessarily in the network PC$^1$.

$^1$By stating that the $s_i$ calculated by the MP theory is a local measure, we mean that it represents the probability that node $i$ is part of the PC of its immediate neighbourhood. In a modular network this neighbourhood is the module in which the node is located: because of the high density of connections and relatively small shortest path lengths inside the module, $s_i$ is strongly coupled with the $s_j$ of other nodes in its neighbourhood. On the other hand, $s_i$ is very loosely correlated with the $s_j$ of nodes in other modules because of the small number of connections between the two modules. This property allows us to take the $s_i$ calculated in the MP theory for the whole network and then sum these $s_i$ values for each module (Eq. (4.3)) to obtain $S_m$, the size of the CPC located only in module $m$. Hence, in a
We observed that MMP performed very well for the LFR network of Fig. 4.5(b), although the network had modules of very small sizes (and, interestingly, despite the very high clustering coefficient $C = 0.61$ of this network). It is because of the very small number of interlinks that the assumption of formation of CPCs is still a good approximation even for the extremely small modules of this network. On the other hand, the MMP method underestimates $\tilde{S}$ for the LFR network of Fig. 4.5(c) whose mixing parameter (and consequently its number of interlinks) is much larger. Because of the larger mixing parameter, the modules are less dense than those of the LFR network of Fig. 4.5(b), and they are more connected to other modules. Hence, it is less likely that such small modules can have a CPC of their own, while it is more likely that they form a joint cluster with other modules.

If a module does not have a CPC then the $s_i$ (calculated by Eq. (4.2) of the MP theory) is not any more the probability that node $i$ is in the CPC of that module; however, it represents the probability that node $i$ is part of a monolithic percolating cluster to which the nodes of this module can connect. Accordingly, the relative size of that monolithic cluster is the direct sum over the $s_i$ of nodes forming that cluster, or equivalently, it equals the sum over the $S_m$ of the modules on which the monolithic cluster is formed. Hence, the MMP theory underestimates the size of this cluster as it assumes that $S_m$ are the sizes of CPCs and instead of summing over all the $S_m$ values, only adds the $S_m$ of CPCs that are connected with a probability $\eta < 1$. As we mentioned earlier, this happens since the modules are very small and the number of interlinks is not small enough to assume (independent) CPCs for such small modules.

Increasing the module sizes improves the accuracy of the MMP theory; interestingly, the module sizes do not need to be extremely large for the MMP theory to perform well. Figure 4.5(d) shows that the MMP theory has reasonable accuracy if the minimum module size is set to 50 instead of 10 as in the network of Fig. 4.5(c), network with limited mixing, we interpret $s_i$ as the probability that node $i$ is in the CPC of that module. This interpretation can still be employed when the network is comprised of well-mixed modules which share many interlinks because, in this case, we can still assume that the well-mixed modules possess (independent) CPCs that are connected with probability 1. If all the modules are well-mixed, we will have only one polylicthic cluster comprised of CPCs of all the modules. Then Eq. (4.8) is reduced to Eq. (4.4) or, equivalently, to Eq. (4.2), as the only polylicthic cluster (that has a size $S$) occurs with probability $P(S) = 1$ and the MMP theory is reduced to the MP theory.
although both networks have the same mixing parameter $\mu = 0.1$. Further investigations show that on LFR networks with modules as small as 50 nodes the MMP theory performs very well (see Appendix C).

### 4.7 Results on real-world networks

We use the above observations to design a simple method for applying the MMP theory to real-world networks where the modular structure is not known. In real-world networks we first identify the best representations of the network modular structure using a multiresolution module detection method (e.g., see Refs. [17, 53, 131, 132]). We are interested in a representation that (i) captures the modular structure of the network, hence has the maximum modularity $Q$ [117], and (ii) has the minimum number of very small modules, and accordingly the identified modules are large enough so that they can have an (independent) CPC. We observe that, in LFR networks the performance of MMP is not affected appreciably by the finite size of the modules as long as the minimum modules size is $N^\text{(min)}_m = 50$. As LFR networks were designed to resemble real-world modular networks structurally [87], we suppose that in real-world networks the performance of MMP is not affected by the module sizes if modules are larger than $N^\text{(min)}_m = 50$. Therefore, for each representation of modular structure of a real-world network, we calculate the modularity $Q$, the number of modules $M$ and also $M_s$, the number of modules with sizes less than a threshold value of $N^\text{(min)}_m = 50$. We then choose the representation that maximises $\frac{M - M_s}{M} Q$. This representation is optimal for both modularity and module sizes. Using this simple measure, we avoid having modules of very small sizes in the representation we use for the MMP theory.

In Appendix D, we present the results for the module detection using the Ronhovde-Nussinov method [131] for the real-world networks exemplified in Table 4.1, as well as the percolation results for these networks.

We can show that on several real-world networks the bond percolation results are affected by the emergence of CPCs if the number of interlinks is sufficiently low. Figure 4.6(a) shows the results for percolation process on the AS Internet network [92]. Both the MP and MMP methods perform well on this network. The MP method is accurate on this network as the number of interlinks is relatively large. The modular
Figure 4.6: (a) Bond percolation results on the AS Internet network. The result of the MP and MMP methods match each other and both match the result of direct numerical simulations on this network. As the number of interlinks is sufficiently large in this network (see Table 4.1) the CPCs on different modules connect with probability 1 and MMP reproduces the result of the MP theory. (b) Results on the western United States power grid network. Because of the small number of interlinks in this network, the emergent CPCs can only connect with a probability \( \mu \) smaller than 1. This is a source of inaccuracy of the MP theory on this network. Taking the CPCs into account, the MMP method improves significantly over the MP theory. The MMP predictions match the numerical results for the \( P_{k,k'}^{i,i'} \) rewired version of this network.

structure considered for this network has, on average, relatively large modules; hence the inaccuracy caused by the small module(s) is negligible and the MMP theory result is also accurate.

Figure 4.6(b) shows that MMP improves significantly over the MP prediction of \( \bar{S} \) on the western United States power grid network [95, 148], and also provides a prediction for the variability of \( S \) according to the percentiles of \( P_{(S)} \). The results of MMP on the power grid network match the numerics for the \( P_{k,k'}^{i,i'} \) rewired [96] version of this network in which the links are rewired inside each module; \( P_{k,k'}^{i,i'} \) rewiring preserves the modular structure and degree-degree correlations but effectively destroys the short
4. EMERGENCE OF COEXISTING PERCOLATING CLUSTERS

loops. This shows that MMP provides a highly accurate prediction in the absence of short loops and when the modular structure is identified accurately. Similar results are shown in Appendix D for several other examples of real-world networks in Table 4.1.

4.8 Conclusion

We demonstrated that CPCs can emerge in ensembles of random networks and in real-world networks, when the network modules are connected via a sufficiently small number of interlinks. Moreover, we showed that CPCs are an important source of error in the theories for bond percolation and proposed the MMP theory that accurately captures the impact of CPCs on percolation results.

An important implication of the appearance of CPCs is the uncertainty they cause in determining the network robustness: when the CPCs emerge, the size $S$ of the network PC can be highly variable. This implies the prominent role of interlinks in network robustness, even in the absence of module-based targeted attacks [36], and subject to only random failures. Another implication is that the eventual size of an epidemic spread in the SIR model [78, 106] may not be best represented by $\bar{S}$, which is the expected size of the largest polylithic PC, since even CPCs not in the largest polylithic PC represent (independent) outbreaks of comparable sizes located in different modules. Hence, the total size of an epidemic outbreak may better be represented by the sum of the sizes of all CPCs.

It is worth mentioning that the MMP theory is exact on tree-like networks in the limit of infinitely large network modules. Nevertheless, we showed that if the network modules are not extremely small the MMP theory still provides accurate predictions.

It is also worth noting that some of the networks we exemplified had very high clustering coefficient, and we showed that the inaccuracy of the MP theory on these networks is not due to the high density of short loops but because of the emergence of CPCs. Hence, the clustering coefficient of a network may not be the best measure for the effect that short loops can have on the accuracy of tree-based theories. Studying the organization of short loops on networks [34], and its effect on the accuracy of tree-based theories [52], may shed more light on how short loops affect network processes such as percolation.
5.1 Introduction

In Chapters 3 and 4, we focused on the theoretical analysis of the size of the percolating cluster of networks. We noted that this quantity is important to investigate the robustness of networks to failures or attacks [28]. Moreover, we saw in Chapter 3 that the size of the percolating cluster can be regarded as the size of an epidemic outbreak in the SIR model [78, 106].

In addition to the size of the percolating cluster, other quantities measured in a percolation process can be interesting to study and useful in understanding other processes that operate on networks. In this chapter we investigate the distribution of sizes of finite clusters in a percolation process, i.e., clusters other than the one that percolates. An interesting application of finding the distribution of cluster sizes is understanding the behaviour of neuronal avalanches.

An early application of bond percolation to describe neural dynamics was proposed by Breskin et al. [21, 46]. Breskin and his colleagues [21, 46] considered gradual weakening of synaptic connections between neurons as a bond percolation process. Using this analogy, they obtained the distribution of cluster sizes in the corresponding percolation process and studied the network connectivity in cultures of hippocampal neurons.
5. SIZE DISTRIBUTION OF PERCOLATION CLUSTERS AND NEURONAL AVALANCHES

Later, Eric Friedman and Adam Landsberg \[58\] showed that a simplified version of the integrate-and-fire spiking neurons model \[24\] can be mapped to bond percolation. Accordingly, they could associate some measures of neuronal avalanches with quantities in the bond percolation process and provide analytical and numerical calculations for them. In particular, they showed that in a class of hierarchical modular networks a power-law distribution of neuronal avalanche sizes arises due to the structure of such networks. Using this evidence and calculations for other measures of the avalanche dynamics, they showed that the special structure of these networks can cause critical behaviour in neural dynamics. Their results emphasized that the network structure may play a significant role in producing or reinforcing the critical dynamics observed for more intricate models of neural dynamics such as that of Ref. \[133\].

In another study, Nir Friedman et al. \[59\] used numerical simulations for a similar simplified integrate-and-fire model to explain the collective behaviour of neurons observed in experimental measurements for cultured cortical slices. Using this model, Friedman and his colleagues \[59\] reproduced the avalanche sizes, avalanche durations, and the mean temporal profile of the avalanches\[1\] obtained from the experiments and observed distinct types of behaviour for different samples of cultured cortical networks. The authors argued that \[59\] the cortical slices can operate in distinct regimes, i.e., subcritical, critical, or supercritical phases, where the phase is determined by the number and the density of connections in an all-to-all weighted graph assumed for the cortical networks.

In this chapter, we aim at providing a framework based on the bond percolation process to shed more light on the behaviour of neuronal avalanches. In order to do so, we consider the following model for neural dynamics which is similar the model described by Friedman and Landsberg \[58\]. In this model, we consider the nodes that have fired as active, and assume that each avalanche starts with a single random active neuron. When a neighbour of an active node receives the fired signal, it is considered to be in one of two states: it is primed with probability $p$, or quiescent with probability $1 - p$. If a neuron is primed, it becomes active after receiving the signal and fires a signal to its neighbours, but if it is quiescent, it does not fire. It is also assumed that, in the timescale of interest, each node can fire only one time and repeated activation is neglected. Hence, in this model, the activity propagates through the neurons that

\[1\]i.e., the average number of neurons fired at time $t$ for avalanches with duration $T$.\[72\]
were primed when they received the signal and the neurons in an avalanche comprise a connected component of primed nodes.

The only difference between this model and that of Friedman and Landsberg [58] is that, in our model, if a node that was quiescent when it received a signal from a neighbouring neuron receives a signal from another neighbour at a later time, it can now be either in the quiescent or the primed state. However, in Ref. [58] a quiescent neuron remains absolutely at the same state. On the other hand, similar to the model of Friedman and Landsberg [58], the above model can be thought as a simplified version of the integrate-and-fire model of neural dynamics [24] as, despite its simplicity, it captures the fast time-scale behaviour of the integrate-and-fire model. Moreover, the above model can be mapped to a bond percolation process with occupation probability $p$. This is because each primed node that is activated can be assumed to activate only the nodes to which it has occupied links; in other words, each connection of an active node is occupied and connected to a primed neighbour with probability $p$. Then each avalanche, which is a connected component of primed nodes, can be considered as a cluster of connected nodes in a bond percolation process. Thus, we can map the problem of finding the distribution of the sizes of neuronal avalanches to the calculation of the distribution of cluster sizes in a bond percolation process.

Accordingly, in this chapter, we investigate the problem of finding the distribution of bond percolation cluster sizes and analyse the results to find out the potential applications and implications this approach can have for understanding the properties of neuronal avalanches. In this regard, we introduce an analytical approach to obtain the distribution of cluster sizes for undirected unweighted networks with arbitrary degree distribution. In particular, we study networks with scale-free degree distribution as neural networks were shown to have a heavy-tailed distribution of neural connections [139]. We examine the accuracy of our approach through direct numerical simulations of the percolation process. Then we argue how this approach can help to shed more light on the effect of network structure on dynamics of neural systems.
5. SIZE DISTRIBUTION OF PERCOLATION CLUSTERS AND NEURONAL AVALANCHES

5.2 The governing equations for the generating functions of the cluster size distributions

In this section, we introduce the method developed by Newman et al. [114, 118] to obtain the distribution of cluster sizes in configuration model networks. This distribution is related to \( \pi_s \), the probability that a randomly chosen node belongs to a cluster with size \( s \). In the method of Newman et al., one first obtains two main relations that determine the generating function for the distribution of cluster sizes,

\[
H_0(z) = \sum_{s=1}^{\infty} \pi_s z^s; \tag{5.1}
\]

then \( \pi_s \) can be calculated using the relations for \( H_0(z) \).

To obtain such relations, let us also denote by \( \rho_s \), the probability that following an edge of a randomly chosen node, we reach a cluster with size \( s \), and by \( H_1(z) \) the generating function for \( \rho_s \), where

\[
H_1(z) = \sum_{s=0}^{\infty} \rho_s z^s. \tag{5.2}
\]

Now consider a randomly chosen node \( i \) that has a degree \( k \), and denote by \( P(s|k) \) the probability that when \( i \) is removed, the sum of the sizes of the \( k \) clusters reached from each of its \( k \) neighbours is \( s \). Accordingly, the probability that node \( i \) (that has a degree \( k \)) is part of a cluster of size \( s \) is represented by \( P(s-1|k) \). Therefore, for a tree-like network, the probability \( \pi_s \) that a node with any degree is part of a cluster with size \( s \) is given by

\[
\pi_s = \sum_{k=0}^{\infty} p_k P(s-1|k), \tag{5.3}
\]

where \( p_k \) is the probability that a randomly chosen node has a degree \( k \). Then, substituting the above relation for \( \pi_s \) into Eq. (5.1) gives

\[
H_0(z) = \sum_{s=1}^{\infty} \sum_{k=0}^{\infty} p_k P(s-1|k) z^s = z \sum_{k=0}^{\infty} p_k \sum_{s=1}^{\infty} P(s-1|k) z^{s-1}
\]

\[
= z \sum_{k=0}^{\infty} p_k \sum_{s=0}^{\infty} P(s|k) z^s \tag{5.4}
\]
The second sum in Eq. (5.4) is the generating function for the probability that the sizes of the clusters of the $k$ neighbours add up to $s$. As these sizes are independent from each other (under the tree-like assumption), the second sum can be written as the generating function for the size of the cluster for any of the neighbours, i.e., $H_1(z)$ to the power of $k$, according to the power rule (see Chapter 1). Hence,

$$H_0(z) = z \sum_{k=0}^{\infty} p_k [H_1(z)]^k$$

(5.5)

The next step is to obtain a relation that determines $H_1(z)$. To do so, firstly a relation for $\rho_s$ is obtained: the quantity $\rho_s$ defines the probability that an edge of a randomly chosen node leads to a cluster of size $s$. Consider the neighbour at the end of such an edge that has an excess degree $k$, i.e., it has $k$ edges other than the edge through which we arrived at that node. Then, the size of the cluster the neighbour can reach through its $k$ other edges is $P(s - 1|k)$. The probability of reaching such a neighbour is determined by the excess degree distribution $q_k$. Hence, the probability that a randomly chosen edge reaches a cluster of size $s$ through one of its neighbours is

$$\rho_s = \sum_{k=0}^{\infty} q_k P(s - 1|k).$$

(5.6)

Therefore,

$$H_1(z) = \sum_{s=1}^{\infty} \sum_{k=0}^{\infty} q_k P(s - 1|k)z^s = z \sum_{k=0}^{\infty} q_k \sum_{s=1}^{\infty} P(s - 1|k)z^{s-1}$$

(5.7)

$$= z \sum_{k=0}^{\infty} q_k \sum_{s=0}^{\infty} P(s|k)z^s$$

$$= z \sum_{k=0}^{\infty} q_k [H_1(z)]^k$$

$$= z G_1(H_1(z)),$$

where the last sum in the second line is the generating function for $P(s|k)$, which according to the power rule is equal to $[H_1(z)]^k$. In conclusion, the generating function for $\pi_s$ can be obtained using the two equations below:

$$H_1(z) = z G_1(H_1(z)),$$

(5.8)

$$H_0(z) = z G_0(H_1(z)).$$

(5.9)
5. SIZE DISTRIBUTION OF PERCOLATION CLUSTERS AND NEURONAL AVALANCHES

5.3 Derivation of cluster size distribution

5.3.1 Using the derivatives of the generating function $H_0(z)$

One method to drive the cluster size distribution is to calculate $\pi_s$ from the definition of its generating function (Eq. (5.1)): for $s > 1$, $\pi_s$ is given by

$$\pi_s = \frac{1}{(s-1)!} \left[ \frac{d^{s-1}}{dz^{s-1}} \left( \frac{H_0(z)}{z} \right) \right]_{z=0}. \tag{5.10}$$

We follow the method of Newman [112] for driving a rather simple equation for $\pi_s$ using Eq. (5.10): Substituting $H_0(z)/z$ in Eq. (5.10) with $G_0(H_0(z))$ (according to Eq. (5.9)) gives

$$\pi_s = \frac{1}{(s-1)!} \left[ \frac{d^{s-1}}{dz^{s-1}} G_0(H_1(z)) \right]_{z=0}$$

$$= \frac{1}{(s-1)!} \left[ \frac{d^{s-1}}{dz^{s-1}} \left[ G_0'(H_1(z))H_1'(z) \right] \right]_{z=0}. \tag{5.11}$$

Then, the Cauchy formula for the $n$th derivative of a function can be used:

$$\frac{d^n f}{dz^n} \bigg|_{z=z_0} = \frac{n!}{2\pi i} \oint \frac{f(z)}{(z-z_0)^{n+1}} dz,$$ \tag{5.12}

where $z$ is considered to be a complex variable and the integral is over a contour in the complex $z$-plane, that encloses the point $z_0$, but does not enclose any singularities of $f(z)$. Now, using Eq. (5.12), Eq. (5.11) becomes

$$\pi_s = \frac{1}{2\pi i (s-1)} \oint \frac{G_0'(H_1(z))}{z^{s-1}} \frac{dH_1}{dz} \frac{dH_1}{dz} \tag{5.13a}$$

$$= \frac{1}{2\pi i (s-1)} \oint \frac{G_1(H_1)}{z^{s-1}} dH_1, \tag{5.13b}$$

where Eq. (5.13b) is obtained using the relations below:

$$G_1(z) = \frac{G_0'(z)}{G_0'(1)}, \quad \langle k \rangle = G_0'(1). \tag{5.14}$$

An infinitesimal loop around $z = 0$ is considered for the contour of Eq. (5.13a). Then, because $H_1(z) \to 0$ as $z \to 0$, the contour of Eq. (5.13b) is also an infinitesimal
5.3 Derivation of cluster size distribution

loop around the origin. Now, using Eq. (5.8), we substitute $z$ in Eq. (5.13b) with $H_1/G_1(H_1)$ and, hence, we get

$$\pi_s = \langle k \rangle \frac{1}{2\pi i (s-1)} \oint \frac{[G_1(H_1)]'}{H_1^{s-1}} dH_1. \quad (5.15)$$

Using the Cauchy formula (Eq. (5.12)) once more, yields

$$\pi_s = \langle k \rangle \frac{1}{(s-1)!} \left[ \frac{d^{s-2}}{dz^{s-2}} [G_1(z)]' \right]_{z=0} \quad \forall \ s > 1. \quad (5.16)$$

For $s = 1$, where Eq. (5.16) becomes invalid, one should note that a cluster has a size equal to 1 if and only if it has no connection to other nodes. Hence, the probability that a node belongs to a cluster of size 1 equals the probability that a randomly chosen node has a degree zero, that is,

$$\pi_1 = p_0. \quad (5.17)$$

Equations (5.16) and (5.17) provide an analytic solution for the entire distribution of cluster sizes, once the derivatives in Eq. (5.16) are carried out. It is possible to carry out any finite number of these derivatives [112]. However, one should note that numerical differentiation is inaccurate for a large number of derivatives [67]. Gleeson et al. used inversion of the probability generating function by fast Fourier transform techniques to obtain the derivatives of generating functions (see the supplemental material of Ref. [67] and references therein). For some degree distributions it is possible to find a closed-form for the derivatives of $G_1(z)$. In the following, we present the results for three such degree distributions [112].

Examples

As shown in Ref. [112], it is possible to find a closed-form expression for $\pi_s$ for the Poisson and exponential degree distributions; however, this is not the case for a pure power-law degree distribution [116]. Instead, it was shown [112] that for a Yule distribution, whose tail can be well approximated by a power-law, it is possible to find a closed-form expression for $\pi_s$. For example, the Yule distribution $q_k \propto \frac{\Gamma(k+1)}{\Gamma(k+2)}$ behaves as $k^{-3/2}$ for large $s$; this can be shown using Stirling’s approximation: for large $n$, $\Gamma(n) \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$. In Table 5.1, we summarize the results of direct calculation of $\pi_s$ for the aforementioned degree distributions [112].
5. SIZE DISTRIBUTION OF PERCOLATION CLUSTERS AND NEURONAL AVALANCHES

Table 5.1: Summary of the analytical results for the distribution of cluster sizes in networks with different degree distributions. The Yule degree distribution defined in the last column has a power-law tail with exponent $-5/2$. The parameter $c$ for Poisson degree distribution is equal to the average degree of the network, and $C$ in the Yule distribution is a normalization constant.

<table>
<thead>
<tr>
<th></th>
<th>Poisson</th>
<th>Exponential</th>
<th>Yule (power-law tail)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_k$</td>
<td>$e^{-c/k} k!$</td>
<td>$(1 - e^{-A}) e^{-Ak}$</td>
<td>$C \frac{\Gamma(k + \frac{1}{2})}{k \Gamma(k + 2)}$</td>
</tr>
<tr>
<td>$G_1(z)$</td>
<td>$e^{e(z-1)}$</td>
<td>$\left(\frac{e^z - 1}{e^z - z}\right)^2$</td>
<td>$\frac{1}{1 + \sqrt{1 - z}}$</td>
</tr>
<tr>
<td>$\pi_s$</td>
<td>$\frac{e^{-cs} (cs)^{s-1}}{s!}$</td>
<td>$(3s - 3)! e^{-A(s-1)} (1 - e^{-A})^{2s-1} \frac{1}{(s-1)!(2s-1)!}$</td>
<td>$(3s - 5)! s^{2s-3s} \frac{1}{<a href="s-1">1 - \ln 2</a>!(2s-2)!}$</td>
</tr>
</tbody>
</table>

Table 5.2: The parameters in the formula $\pi_s \approx As^{-3/2} e^{-\mu s}$ obtained for the large $s$ behaviour of cluster size distribution in networks with different degree distributions. As for the Yule distribution considered here we have $\mu > 1$, the network with this degree distribution is in the super-critical phase component and many vanishingly small clusters.

<table>
<thead>
<tr>
<th></th>
<th>Poisson</th>
<th>Exponential</th>
<th>Yule</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>$1/\sqrt{2\pi c^2}$</td>
<td>$\sqrt{\frac{3}{4\pi}} \left(\frac{1 - e^{-A}}{2}\right)$</td>
<td>$\frac{1}{6 \sqrt{\pi (1 - \ln 2)}}$</td>
</tr>
<tr>
<td>$\mu$</td>
<td>$c - 1 - \ln(c)$</td>
<td>$-\ln\left[3^3 e^{-A} \left(\frac{1 - e^{-A}}{2}\right)^2\right]$</td>
<td>$5 \ln(2) - 3 \ln(3)$</td>
</tr>
<tr>
<td>Critical point</td>
<td>$c = 1$</td>
<td>$\lambda_c = \ln(3)$</td>
<td>N/A</td>
</tr>
</tbody>
</table>

The behaviour of cluster size distribution for large values of $s$ can be obtained by approximating $\pi_s$ using Stirling’s approximation. This calculation was performed before in Ref. [112] for the exponential and the Yule distribution of Table. 5.1; however, the final results reported in [112] for these distributions were erroneous.\(^1\) We per-

\(^1\)We discussed these calculations with Mark Newman via email and he agrees with our results.
formed the calculations again for these two distributions and also for the Poisson distribution and found out that for all the three distributions $\pi_s$ has the form $\pi_s \sim A s^{-3/2} e^{-\mu s}$ at large values of $s$. In Table 5.2, we illustrate the values of the parameters we obtained for the behaviour of $\pi_s$ at large values of $s$. At a critical value of $\mu = \mu_c = 0$, the tail of the distribution of cluster sizes takes a pure power-law form. At such a critical point, the network percolating cluster appears; this can be checked using the Molloy-Reed criterion [100] that states the percolating cluster appears at a critical point at which $\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} = 1$ [116]. This criterion gives the same values for the parameters of the degree distribution as those obtained for $\mu = 0$ (see Table 5.2).

### 5.4 Cluster size distribution in bond percolation

In this section we derive the cluster size distribution for bond percolation on networks, where each link of the network is occupied with probability $p$. If we denote by $G_0(z)$ and $G_1(z)$, respectively, the generating functions for the degree and excess degree distribution of the network, then for a percolation process on the same network, these generating functions are, respectively, $G_0(1 - p + pz)$ and $G_1(1 - p + pz)$ [106]. We use this fact and the calculations performed in Secs. 5.2 and 5.3 to obtain the cluster size distribution for bond percolation in the following sections.

#### 5.4.1 Using direct differentiation of $G_1(z)$

For a network on which a bond percolation process operates, Eq. (5.16) is changed to [112]:

$$\pi_s(p) = \frac{\langle k \rangle p^{s-1}}{(s-1)!} \left[ \frac{d^{s-2}}{dz^{s-2}} [G_1(\hat{z})]^i \right]_{\hat{z}=1-p} \quad \forall \ s > 1, \quad (5.18)$$

where $\hat{z} = 1 - p + pz$. Equation (5.18), which was proposed in Ref. [112], was used to demonstrate how the cluster size distribution in a network with Poisson degree distribution changes if a bond percolation process is performed on such a network.

Here we use Eq. (5.18) to obtain a closed form relation for the cluster size distribution $\pi_s(p)$ of bond percolation on the networks with each of the degree distributions
5. SIZE DISTRIBUTION OF PERCOLATION CLUSTERS AND NEURONAL AVALANCHES

of Table 5.1. We also calculate the asymptotic (large \( s \)) behaviour of the obtained \( \pi_s \) functions.

For a network with a Poisson degree distribution, \( G_1(1 - p + pz) = e^{c(1 - p + pc - 1)} = e^{cp(1-c)} \). Therefore, \( \pi_s(p) \) can be obtained by substituting \( cp \) instead of \( c \) for \( \pi_s \) of this distribution in Table 5.1, i.e., \( \pi_s(p) = e^{-cps(cps)^{s-1}/s!} \). Hence, for large \( s \), \( \pi_s(p) \) again behaves as \( As^{-3/2} e^{-\mu s} \), where \( A \) and \( \mu \) are now equal to \( \frac{1}{\sqrt{2 \pi}} \frac{1}{c^2} \) and \( cp - 1 - \ln(c) \), respectively. The critical point for the percolation process is located at \( \mu = 0 \); i.e., when \( cp = 1 \). Thus, the critical (threshold) occupation probability is \( p_{th} = 1/c \). To check this result we can use the criterion \( p_{th} \langle k^2 \rangle - \langle k \rangle \langle k \rangle = 1 \) obtained for the threshold at which the giant connected component of a network in a percolation process appears [32, 116]; obviously this is similar to the Molloy-Reed condition with an additional factor \( p_{th} \).

For an exponential degree distribution, we first calculate the derivatives in Eq. (5.18):

\[
\frac{d^{s-2}}{dz^{s-2}} [G_1(\hat{z})]' = \frac{d^{s-2}}{dz^{s-2}} \left[ \left( \frac{e^\lambda - 1}{e^\lambda - \hat{z}} \right)^2 \right] = \frac{(3s - 3)! (e^\lambda - 1)^{2s}}{(s-1)! (e^\lambda - \hat{z})^{3s-2}}.
\] (5.19)

Substituting \( \langle k \rangle = 1/(e^\lambda - 1) \), \( \hat{z} = 1 - p \), and Eq. (5.19) into Eq. (5.18) gives

\[
\pi_s = \frac{(3s - 3)! [p/(e^\lambda - 1)]^{s-1}}{(s-1)! (2s-1)! [1 + p/(e^\lambda - 1)]^{3s-2}}.
\] (5.20)

Now we define \( \psi = p/(e^\lambda - 1) \), and use Stirling’s approximation to obtain the large \( s \) behaviour of \( \pi_s \):

\[
\pi_s = \frac{(3s - 3)! \psi^{s-1}}{(s-1)! (2s-1)! [1 + \psi]^{3s-2}} \\
\approx \sqrt{\frac{3}{4\pi}} s^{-3/2} \frac{3s-3}{2^{s-1}} \frac{\psi^{s-1}}{(1 + \psi)^{3s-2}} \\
= \sqrt{\frac{3}{\pi}} \frac{(1 + \psi)^2}{27 \psi} s^{-3/2} \left[ \frac{4 \left( \frac{1 + \psi}{3} \right)^3}{\psi} \right]^{s-3/2} \\
\equiv As^{-3/2} e^{-\mu s}.
\] (5.21)
At the critical point $\psi = \psi_c = 1/2$, $\mu = 3 \ln \frac{1 + \psi}{3} - \ln \frac{\psi}{4}$ equals zero, and the distribution of cluster sizes becomes a pure power-law. This $\psi_c = p_c/(e^\lambda - 1) = 1/2$ that we obtained for $\mu = 0$ matches the result for the threshold value $p_c$ obtained [115] using the criterion $p_c \frac{(\langle k \rangle - \langle k \rangle)}{\langle k \rangle} = 1$.

5.5 Using asymptotic expansions

As shown in the previous section, for some degree distributions it is possible to find a closed-form solution for the distribution of cluster sizes by direct differentiation in Eq. (5.18); however, is not always the case. For example, the power-law degree distribution, $p_k \propto k^{-\gamma}$, does not have closed-form expression for $G_1(z)$ [116]. Hence, for a power-law distribution with an arbitrary $\gamma$, we cannot find an analytical expression for $\pi_s$ using Eq. (5.18). The alternative solution is to write a Yule distribution that gives a power-law tail and then try to find a closed-form expression for $\pi_s$ using Eq. (5.18), which is a rather involved task.

In this section, we discuss an approach, different from that of Sec. 5.4.1, that can be used to find the cluster size distribution for any degree distribution of interest. In this method, the asymptotic expansion of the corresponding generating function is analyzed to obtain the behaviour of the tail of the probability distribution for which the probability generating function is defined [118]. In other words, for a generating function $f(z) = \sum_n a_n z^n$, the sequence $a_n$ of the coefficients of $z^n$ for large $n$ is calculated [151]; the approaches for these kind of problems are referred to as Tauberian methods (TMs) and the approaches for the inverse problems of finding the behaviour of $f(z)$ from $a_n$, are referred to as Abelian theorems [149]. Newman et al. [112] showed that, using a TM, the cluster size distribution for any degree distribution with finite second moment can be calculated for configuration model networks. Cohen et al. used these approaches to analyze the bond percolation properties at the critical point for undirected [31] and directed [136] scale-free networks that can have a diverging second moment of the degree distribution, and provided results for the shape (i.e., the functional form) of the distribution of cluster sizes. Lee et al. [91] used TMs to find the size of the trees in scale-free branching processes. Gleeson et al. [67] also provided a Tauberian method to obtain the popularity distribution of memes in a model of Twitter.
5. SIZE DISTRIBUTION OF PERCOLATION CLUSTERS AND NEURONAL AVALANCHES

As we mentioned in the above paragraph, the analysis of the percolation properties of networks using methods based on generating functions was first presented by Cohen et al. [31]. Cohen et al. derived results for the critical point of the percolation process; however, for the subcritical and supercritical regimes, they did not present calculations for derivation of $\pi_s$. On the other hand, they assumed that a result Newman et al. proposed [118] (for configuration model networks with certain conditions) should be valid for percolation on configuration model scale-free networks too.

In the next section, we present the method Cohen et al. [31] introduced for calculating the shape of $\pi_s$, for large $s$, at the critical bond percolation occupation probability; in addition to the calculations of Cohen et al., we perform calculations of the multiplicative prefactor that appears in $\pi_s$ to determine the full form of $\pi_s$, for all $s \gg 1$. We show that obtaining the prefactor is essential to determine the validity and accuracy of the theoretical results. At the last part of the next section, using TMs, we extend the method of Cohen et al. [31] to calculate $\pi_s$ away from the critical point and compare it with the shape that Cohen et al. [31] assumed $\pi_s$ should have away from the critical point.

We refer to the above calculations (for both the critical and noncritical regimes), that are based on the method of Cohen et al., as Method (I). We show that the supercritical regime of scale-free networks with exponent $3 < \gamma < 4$ is an exceptional case for which method (I) can not provide a valid solution for $\pi_s$. Moreover, we show that method (I) predicts correctly the shape (functional form) of $\pi_s$, however, its predictions for the multiplicative prefactor are inaccurate.

Next, we attempt to improve upon method (I). In order to do so, we consider different modifications to method (I); thus we propose three other asymptotic methods that resemble method (I) in some parts and differ from it in some other respects. Therefore, we propose alternative solutions for the multiplicative prefactor of $\pi_s$, and also find a valid solution for the supercritical regime of $3 < \gamma < 4$ and discuss our new findings. Afterwards, in Sec. 5.6, we compare the performance of the different methods against direct numerical simulations of the bond percolation process. Therefore, for different ranges of $\gamma$ values and different regimes, we propose the method that produces valid and accurate predictions for both the functional form and the multiplicative prefactor of $\pi_s$. 

82
5.5 Using asymptotic expansions

5.5.1 Method (I) (based on the method of Cohen et al.)

πs at the critical point

Cohen et al. [31] adapted Eqs. (5.8) and (5.9) to the case of a bond percolation process on configuration model networks, and proposed two equations for the relation between the generating functions in a bond percolation process with occupation probability p:

\[ H_1(z) = 1 - p + p z G_1(H_1(z)), \]
\[ H_0(z) = 1 - p + p z G_0(H_1(z)). \]

Eqs. (5.22) and (5.23) are obtained using heuristic methods [19]. We show that, using these relations, the shape of πs is predicted correctly, however the prediction of these equations for the multiplicative prefactor is inaccurate. Hence, in the calculations we present as method (I), we first use Eqs. (5.22) and (5.23) and review the method to obtain πs at the critical point using these equations. Then, based on the method of Cohen et al., we develop an approach to derive πs away from the critical point. Then, in the next sections, we propose the alternative methods and compare their results and their advantages to those of method (I).

In the next step of method (I), \( H_0(z) \) (that appeared Eq. (5.23)) is expanded around \( z = 1 \): For \( z \) near 1, we can write \( z = 1 - w \) and \( H_1(1 - w) = 1 - \phi \), where \( w, \phi \to 0 \).

At the critical occupation probability \( p_c \), we will show that the asymptotic expansion of \( H_0(1 - w) \) has the form

\[ H_0(1 - w) = \text{analytic part} + \sum_{m=1}^{\infty} b_m w^{\beta_m}, \]
\[ \sim \text{analytic part} + b_1 w^{\beta_1} \quad \text{as} \quad w \to 0, \] (5.24)

where \( \beta_m \) are not integers and \( \beta_1 < \beta_2 < \ldots \). Then, according to a Tauberian theorem, for such asymptotic expansion of \( H_0 \), it can be shown [118, 151] (see Appendix F) that for large \( s \), πs has the form

\[ \pi_s \sim \frac{b_1}{\Gamma(-\beta_1)} s^{-(\beta_1+1)}, \] (5.25)

where \( \Gamma \) is the Gamma function. A nice proof for this theorem was also described in Ref. [67] (see Lemma 1 in its supplementary material).
5. SIZE DISTRIBUTION OF PERCOLATION CLUSTERS AND NEURONAL AVALANCHES

Accordingly, the first step is to find the asymptotic expansion of $H_0(1-w)$. Equation (5.23) gives

$$H_0(1-w) = 1 - p + p(1-w) G_0(1)$$

$$= 1 - p + p(1-w) \left[ G_0(1) - \phi G_0'(1) + \frac{1}{2} \phi^2 G_0''(1) + ... \right]$$

$$= 1 - p + p(1-w) \left[ 1 - \phi(k) + \frac{1}{2} \phi^2 G_0''(1) + ... \right]$$

Hence, we get

$$H_0(1-w) = 1 - \langle k \rangle p \phi - qw + O(\phi^2).$$

(5.26)

Next, we need to obtain $\phi$ as a function of $w$, and substitute it in Eq. (5.24) to find the dominant powers of $w$ that determine the leading order behaviour of $H_0(1-w)$. Equation (5.22) gives

$$1 - \phi = 1 - p + p(1-w) G_1(1)$$

(5.27)

Hence,

$$\phi = p - p(1-w) G_1(1)$$

(5.28)

The expansion of $G_1(1-\phi)$ depends on whether the degree distribution $p_k$ is scale-free or not. If $p_k$ in not scale-free, then

$$G_1(1-\phi) = G_1(1) - \phi G_1'(1) + \frac{1}{2} \phi^2 G_1''(1) + ...$$

(5.29a)

$$= 1 - \frac{\langle k(k-1) \rangle}{\langle k \rangle} \phi + \frac{\langle k(k-1)(k-2) \rangle}{2\langle k \rangle} \phi^2 + o(\phi^2) \quad \text{as} \quad \phi \to 0. \quad (5.29b)$$

On the other hand, for $p_k \propto k^{-\gamma}$ and $q_k = Qk^{-(\gamma-1)}$, where $Q$ is a normalization factor, we have [91]:

$$G_1(1-\phi) = 1 - A\phi + B\phi^2 + \begin{cases} C\phi^{\gamma-2} & \text{(for $\gamma \neq \text{integer}$)} \\ D\ln(\phi)\phi^{\gamma-2} & \text{(for $\gamma = \text{integer}$)} \end{cases} + ..., \quad (5.30)$$

where $A = \frac{\langle k(k-1) \rangle}{\langle k \rangle}$, $B = \frac{\langle k(k-1)(k-2) \rangle}{2\langle k \rangle}$, $C = Q\Gamma(2-\gamma)$, and $D = \frac{(-1)^{\gamma-1}}{\Gamma(\gamma-1)}$.

Therefore, for non-integer $\gamma$, Eq. (5.28) gives

$$w = \left( \frac{1 - pA}{p} \right) \phi + \begin{cases} B\phi^2 + A\phi w - B\phi^2 w + ... & \text{(for $\gamma > 4$)} \\ C\phi^{\gamma-2} - C\phi^{\gamma-2} w + ... & \text{(for $3 < \gamma < 4$)} \end{cases} \quad (5.31)$$
It is worth noting that the expansion obtained for $\gamma > 4$ in Eq. (5.31) is the same for other degree distributions which have a finite second moment. Thus, the results that will be obtained for $\gamma > 4$ apply to networks with non-scale-free degree distribution too. At the critical point, $p_c = \frac{\langle k \rangle}{\langle k \rangle - 1}$, $p_c A = 1$; therefore, Eq. (5.31) yields

$$w = \begin{cases} \frac{B\phi^2 + o(\phi^2)}{C\phi^{\gamma - 2} + o(\phi^{\gamma - 2})} & \text{for } \gamma > 4 \\ \frac{\phi^2}{\phi^{\gamma - 2}} & \text{for } 3 < \gamma < 4. \end{cases}$$ \hspace{1cm} (5.32)

Thus,

$$\phi \sim \begin{cases} \frac{w^{1/2}}{B^{1/2}} & \text{for } \gamma > 4 \\ \frac{w^{1/(\gamma - 2)}}{C^{1/(\gamma - 2)}} & \text{for } 3 < \gamma < 4. \end{cases} \quad \text{as } w \to 0 \quad (5.33)$$

Now, substituting Eq. (5.33) in Eq. (5.26) gives

$$H_0(1 - w) - \text{analytic part} \sim \begin{cases} \frac{\langle k \rangle p}{B^{1/2}} w^{1/2} & \text{for } \gamma > 4 \\ \frac{\langle k \rangle p}{C^{1/2/(\gamma - 2)}} w^{1/(\gamma - 2)} & \text{for } 3 < \gamma < 4. \end{cases} \quad \text{as } w \to 0 \quad (5.34)$$

Now, according to Eqs. (5.24) and (5.25):

$$\pi_s \sim \begin{cases} -\frac{\langle k \rangle p}{B^{1/2}} s^{-3/2} & \text{as } s \to \infty \quad \text{for } \gamma > 4, \quad (5.35a) \\ -\frac{\langle k \rangle p}{C^{1/(\gamma - 2)}} s^{-(\gamma - 1)/(\gamma - 2)} & \text{as } s \to \infty \quad \text{for } 3 < \gamma < 4. \end{cases}$$

$$\pi_s \text{ away from the critical point}$$

Away from the critical point, i.e., at $p = p_c + \delta$, for $\gamma > 4$, Eq. (5.31) gives

$$pw = -\delta A\phi + Bp\phi^2 + o(\phi^2);$$ \hspace{1cm} (5.36)

hence, the leading order behaviour of $\phi(w)$ is determined by the quadratic equation

$$Bp\phi^2 - A\delta\phi - pw \sim 0 \quad \text{as } w \to 0. \quad (5.37)$$

Therefore,

$$\phi \sim \frac{A\delta}{2Bp} \pm \frac{1}{2Bp} \sqrt{(A\delta)^2 + 4Bp^2w} \quad \text{as } w \to 0. \quad (5.38)$$
As $H_1$ is bounded above by 1, $\phi$ should be positive; thus, the physical solution for $\phi$ is that with the positive sign for the second term on the rhs of Eq. (5.38). Substituting $\phi$ in Eq. (5.26) we get

$$H_0(1 - w) \sim \text{analytic part} - \sqrt{a + bw} \quad \text{as } w \to 0,$$

(5.39)

where $a = (A\langle k \rangle \delta^2)/(2B)^2$, and $b = \langle k \rangle^2 p^2 / B$. It can be shown (using a method similar to that of Ref. [67] for deriving Eq. (5.25) from Eq. (5.24)) that for $H_0(1 - w)$ in the form of Eq. (5.39), $\pi_s$ behaves as (see Appendix E)

$$\pi_s \sim \sqrt{\frac{b}{4\pi}} s^{-3/2} e^{-a/b s} \quad \text{as } s \to \infty,$$

(5.40)

for large values of $s$, where

$$a/b = A^2 \delta^2 / (4p^2 B) = \frac{\langle k(k - 1) \delta^2 \rangle}{2p^2 \langle k(k - 1)(k - 2) \rangle}.$$

In Eq. (5.40), the strength of exponential decay is determined by $a/b$ which is proportional to $\delta^2$; this is consistent with the result obtained using finite size scaling [31].

For $3 < \gamma < 4$, and $p = p_c + \delta$, Eq. (5.31) gives

$$pw + \delta A\phi - C p\phi^{\gamma - 2} - B p\phi^2 - A p\phi w = o(\phi^2, \phi w) \sim 0 \quad \text{as } w \to 0.$$  

(5.41)

To find the leading order behaviour of $\phi(w)$, we hypothesize that $\phi$ has an asymptotic expansion with the form

$$\phi \doteq \phi_1 + \phi_2 + \ldots = a_1 w^\alpha + a_2 w^\beta + \ldots,$$

(5.42)

where the exponents can be any real nonnegative number such that $\alpha < \beta$. Then, we employ the dominant balance method [98]: inserting $\phi = \phi_1$ in Eq. (5.41) and neglecting $o(\phi^2, \phi w)$ terms yields

$$pw + \delta Aa_1 w^\alpha - C p a_1^{\gamma - 2} w^{\alpha(\gamma - 2)} - B p a_1^2 w^{2\alpha} - A p a_1 w^{\alpha + 1} \sim 0 \quad \text{as } w \to 0.$$  

(5.43)

Here, for $3 < \gamma < 4$, the only dominant balance is obtained by balancing the first and second term that gives $\alpha = 1$ and $pw + \delta Aa_1 w = o(w) \sim 0$ as $w \to 0$. So far, we have
not obtained the leading order nonanalytic term of \( \phi(w) \). Hence, we proceed to the next term in the expansion of \( \phi \). Inserting \( \phi_1 + \phi_2 \) for \( \phi \) in Eq. (5.41) gives

\[
0 = pw + \delta A(a_1w + a_2w^\beta) - C p(a_1w + a_2w^\beta)^{\gamma-2} - B p(a_1w + a_2w^\beta)^2
\]

\[
- A p(a_1w + a_2w^\beta)w
\]

\[
= \delta A a_2w^\beta - C p a_1^{\gamma-2} w^{\gamma-2}(1 + a_2w^{\beta-1}/a_1)^{\gamma-2} - B p a_1^2 w^2 - B p a_2^2 w^{2\beta}
\]

\[
- B p a_1 a_2 w^{\beta+1} - A p a_1 w^2 - A p a_2 w^{\beta+1},
\]

(5.44)

\[
\sim \delta A a_2w^\beta - C p a_1^{\gamma-2} w^{\gamma-2} - C p(\gamma - 2)a_1^{\gamma-3} a_2 w^{\beta+\gamma-3} - B p a_1^2 w^2
\]

\[
- B p a_2^2 w^{2\beta} - B p a_1 a_2 w^{\beta+1} - A p a_1 w^2 - A p a_2 w^{\beta+1}
\]

as \( w \to 0 \). (5.45)

Given that \( 3 < \gamma < 4 \), the only dominant balance for Eq. (5.45) is between the first and second terms which gives \( \beta = \gamma - 2 \) and \( a_2 = C p a_1^{\gamma-2}/(A\delta) = |C p/(A\delta)| (p/A)^{\gamma-2} (-1/\delta)^{\gamma-2} \). As \( p \) and \( A \) are real positive numbers, \( a_2 \) is real only for \( \delta < 0 \). We will show in a moment that only when \( a_2 \) is real will we get a physical solution for \( \pi_s \). Hence, for \( 3 < \gamma < 4 \), Eq. (5.41) can only give acceptable results for the subcritical phase, if \( \delta < 0 \).

Now let us calculate \( \pi_s \) for \( 3 < \gamma < 4 \), and \( \delta < 0 \). According to the above calculations

\[
\phi \sim \phi_1 + \phi_2 = a_1 w + a_2 w^{\gamma-2} = \text{analytic part} + a_2 w^{\gamma-2} \quad \text{as} \quad w \to 0.
\]

Inserting \( \phi \) in Eq. (5.26) yields

\[
H_0(1 - w) \sim 1 - \langle k \rangle p \phi \sim \text{analytic part} - \frac{\langle k \rangle C p^2}{A\delta} \left( \frac{p}{A} \right)^{\gamma-2} \left( -\frac{1}{\delta} \right)^{\gamma-2} w^{\gamma-2} \quad \text{as} \quad w \to 0,
\]

Hence, we substitute \( C = Q \Gamma(2 - \gamma) \) and employ Eq. (5.25) to get

\[
\pi_s \sim -\frac{\langle k \rangle Q p^2}{A\delta} \left( \frac{p}{A} \right)^{\gamma-2} \left( -\frac{1}{\delta} \right)^{\gamma-2} s^{-(\gamma-1)} \quad \text{as} \quad s \to \infty.
\]

(5.46)

The result of Eq. (5.46) is only valid for the subcritical region as \( \pi_s \) is real (and positive) only for \( \delta < 0 \). The result of Eq. (5.46) is very interesting, as it shows that even away from the critical point, in the subcritical region of a scale-free network with \( 3 < \gamma < 4 \), a scale-free distribution of component sizes emerges. This is in sharp contrast with the behaviour previously assumed [31] for cluster size distribution of scale-free networks in the subcritical region.
5. SIZE DISTRIBUTION OF PERCOLATION CLUSTERS AND NEURONAL AVALANCHES

We compare the predictions of method (I), i.e., those of Eqs. (5.35a), (5.35b), (5.40), and (5.46) with the result of direct numerical simulations in Sec. 5.6. As we will see they provide a good prediction for the shape of the distribution $\pi_s$, while the predicted pre-factor is not accurate. To identify the source of this problem and improve our predictions we try alternative methods for calculating $\pi_s$ using an approach similar to that of method (I). In the following we discuss these methods, the difference between them and method (I), and how they will modify the above results for $\pi_s$.

5.5.2 Method (II)

In this method, we first rearrange Eq. (5.28) to get

$$1 - w = \frac{1 - \phi/p}{G_1(1 - \phi)}.$$  \hspace{1cm} (5.47)

Then, for $\gamma > 4$ we will have

$$1 - w \sim \frac{1 - \phi/p}{1 - A\phi + B\phi^2 + o(\phi^2)} \quad \text{as} \quad \phi \to 0 \quad (5.48a)$$

$$\sim (1 - \phi/p)(1 + A\phi - B\phi^2) \quad \text{as} \quad \phi \to 0 \quad (5.48b)$$

$$\sim 1 + \frac{A\delta}{p} \phi - (B + A/p)\phi^2 \quad \text{as} \quad \phi \to 0, \quad (5.48c)$$

where in Eq. (5.48c) we used $Ap - 1 = A(p_c + \delta - 1) = A\delta$. This method differs from method (I) in that a rearranged version of Eq. (5.28) (i.e., Eq. (5.47)) is used and then an extra asymptotic expansion (i.e., the expansion that leads to Eq. (5.48b)) is performed. This is the method used by Lee et al. [91] for critical branching processes; hence, we introduce this method here so that we compare its performance with the other methods.

Now, according to Eq. (5.48c), the leading order behaviour of $\phi$, for $\gamma > 4$, is determined by the relation

$$p(B + A/p)\phi^2 - A\delta\phi - pw \sim 0 \quad \text{as} \quad w \to 0.$$  \hspace{1cm} (5.49)

Equation (5.49) can be changed to Eq. (5.37) if $B$ in Eq. (5.37) is substituted with $B + A/p$. Hence, for $\gamma > 4$, $\pi_s$ in this method is obtained by making the substitution $B \to B + A/p$ in Eqs. (5.35a) and (5.40).
For $3 < \gamma < 4$, the term $B\phi^2$ in Eqs. (5.48a)–(5.48b) is replaced with $C\phi^{\gamma-2}$ and the leading order behaviour of $\phi$ is determined by

$$pw + \delta A\phi - C p\phi^{\gamma-2} \sim 0 \quad \text{as} \quad w \to 0.$$  

Equation (5.50) has the same leading order terms as those of Eq. (5.41). Therefore, for $3 < \gamma < 4$, method (II) gives the same results as those of method (I).

### 5.5.3 Method (III)

In this method, instead of Eqs. (5.22) and (5.23) that are obtained using heuristic methods [19, 31], we use the equations derived by Newman [106]. Newman showed [106] that the generating functions for the degree distribution and excess degree distribution in a bond percolation with occupation probability $p$ are, respectively, $G_0(\hat{z})$ and $G_1(\hat{z})$, where $\hat{z} = 1 - p + pz$ and $G_0(z)$ and $G_1(z)$ are, respectively, the generating functions for the degree distribution and excess degree distribution for the undamaged network, i.e., the network without a percolation process performed on it.

The proof is straightforward and interesting: For a network with degree distribution $p_k$, the probability that a randomly chosen node has $m$ occupied edges, after the percolation process, is $\sum_{k=m}^{\infty} p_k \binom{k}{m} p^m (1 - p)^{k-m}$. Hence, the generating function $\widetilde{G}_0$ for the probability distribution of the new degrees (or in other worlds, the new numbers of (occupied) edges) of the nodes is:

$$\widetilde{G}_0(z) = \sum_{m=0}^{\infty} \sum_{k=m}^{\infty} p_k \binom{k}{m} p^m (1 - p)^{k-m}z^m$$

$$= \sum_{k=0}^{\infty} p_k \sum_{m=0}^{k} \binom{k}{m} (pz)^m (1 - p)^{k-m}$$

$$= \sum_{k=0}^{\infty} p_k (pz + 1 - p)^k$$

$$= G_0(1 - p + pz)$$

$$= G_0(\hat{z}).$$

Similarly, it can be shown that $G_1(z)$ will be transformed to $G_1(1 - p + pz)$ after a bond percolation process. Accordingly, substituting $G_0(z)$ and $G_1(z)$ in Eqs. (5.8) and (5.9)
5. SIZE DISTRIBUTION OF PERCOLATION CLUSTERS AND NEURONAL AVALANCHES

with the corresponding functions obtained by Newman, we get

\[ H_1(z) = z G_1(1 - p + pH_1(z)), \]  
\[ H_0(z) = z G_0(1 - p + pH_1(z)). \]  

(5.51)

(5.52)

Now, in this method, we take similar steps to method (I) but use Eqs. (5.51) and (5.52) as our starting point in place of Eqs. (5.22) and (5.23). Again, we define \( z = 1 - w \) and \( H_0 = 1 - \phi \). Then, Eq. (5.52) gives

\[ H_0(z) = (1 - w) G_0(1 - p\phi) \]

\[ \sim 1 - \langle k \rangle p\phi + o(\phi) \quad \text{as } w, \phi \to 0, \]  

(5.53)

which is the same as Eq. (5.26) to the linear order. On the other hand, Eq. (5.51) gives

\[ H_1(z) = 1 - \phi = (1 - w) G_1(1 - p\phi) \]  
\[ \sim (1 - w) \left[ 1 - A p \phi + \begin{cases} B p^2 \phi^2 & \text{(for } \gamma > 4) \\ C p^{\gamma-2} \phi^{\gamma-2} & \text{(for } 3 < \gamma < 4) \end{cases} + ... \right] \quad \text{as } w \to 0. \]  

(5.54a)

(5.54b)

Thus, for \( \gamma > 4 \), the leading order behaviour of \( \phi \) is determined by

\[ B p^2 \phi^2 - A \delta \phi - w \sim 0 \quad \text{as } w \to 0. \]  

(5.55)

The solutions of Eq. (5.55) for \( p = p_c \) have the same form as Eq. (5.35a), we just need to make the substitution \( B \to B p^2 \) in Eq. (5.35a) to get the solution of Eq. (5.55).

Similarly, making the substitutions \( a \to a/p^2 \) and \( b \to b/p^2 \) in Eq. (5.40) gives the solution for Eq. (5.55) at \( p = p_c + \delta \).

For \( 3 < \gamma < 4 \), the leading order behaviour of \( \phi \) according to Eq. (5.54b) is determined by the relation

\[ w + A \delta \phi - C p^{\gamma-2} \phi^{\gamma-2} \sim 0 \quad \text{as } w \to 0. \]  

(5.56)

It is straightforward to show that, at \( p = p_c \) and \( p = p_c - |\delta| \), the solution of Eq. (5.56) is the right hand side of, respectively, Eqs. (5.35b) and (5.46) divided by \( p \).

As shown above, the \( \pi_s \) obtained in method (III) equals one over \( p \) times the \( \pi_s \) that resulted from method (I). This can also be shown by comparing the two different sets of equations for the generating function employed in each method: Let us define the

\[ \underline{\pi_s} = \frac{1}{p} \pi_s. \]
5.5 Using asymptotic expansions

argument of \( G_1 \) in Eq. (5.51) as another generating function \( \tilde{H}_1 = 1 - p + pH_1 \). Now, writing Eq. (5.51) in terms of \( \tilde{H}_1 \), we get \( \tilde{H}_1(z) = 1 - p + pz G_1(\tilde{H}_1(z)) \); thus, \( \tilde{H}_1 \) satisfies Eq. (5.22) of method (I). On the other hand, whether we define \( \phi = 1 - H_1 \) or \( \phi = 1 - \tilde{H}_1 \), both Eqs. (5.23) and (5.52) lead to the same form for \( H_0 \) in terms of \( \phi \), up to the linear order (see Eqs. (5.26) and (5.53)). Hence, the difference between \( \pi_s \) values obtained from the two methods lies in the difference between \( H_1 \) and \( \tilde{H}_1 \). Now, as \( H_1 = (1 - p)/p + \tilde{H}_1/p \), therefore, except for \( s = 0 \), all the coefficients \( \rho_s \) of the generating function \( H_1(z) = \sum \rho_s z^s \) are equal to \( \tilde{\rho}_s \), the coefficients of \( \tilde{H}_1(z) \), divided by \( p \). Consequently, \( \pi_s \) obtained using \( H_1 \) is equal to one over \( p \) times the \( \pi_s \) obtained using \( \tilde{H}_1 \), for \( s \neq 0 \).

5.5.4 Method (IV)

This method is similar to method (II) in the sense that we rearrange the equation for \( H_1(z) \), in this case Eq. (5.54a), to get

\[
1 - w = \frac{1 - \phi}{G_1(1 - p\phi)},
\]

which is similar to Eq. (5.47); the rest of steps are also similar to those following Eq. (5.47). Consequently, for \( 3 < \gamma < 4 \) the leading order behaviour of \( \phi \) is identically determined by Eq. (5.56), and for \( \gamma > 4 \) we will get

\[
(Bp^2 + Ap)\phi^2 - A\delta\phi - w \sim 0 \quad \text{as } \phi \to 0.
\]

Then, for \( \gamma > 4 \) and \( p = p_c \), \( \pi_s \) is obtained by substituting \( Bp^2 + Ap \) for \( B \) in Eq. (5.55b). For \( \gamma > 4 \) and \( p = p_c + \delta \), \( \pi_s \) is given by making the substitutions \( a \to aB^2/(Bp + A)^2 \) and \( b \to bB^2/(Bp + A)^2 \) in Eq. (5.40).

\( \pi_s \) for \( 3 < \gamma < 4 \) and \( \delta > 0 \)

As we showed, when \( 3 < \gamma < 4 \), Method (I) (and also its alternatives described above) could only find a solution for \( \pi_s \) for the critical and subcritical region. In this section we use an approach similar to the method of Lee et al. [91] to obtain \( \pi_s \) for the supercritical region where \( p = p_c + \delta \), with \( \delta > 0 \). We start with Eq. (5.57) which states that

\[
z = \frac{1 - \phi}{G_1(1 - p\phi)}.
\]
Then, for \(3 < \gamma < 4\), we have
\[
z \sim \frac{1 - \phi}{1 - Ap\phi + Cp\phi^\nu + o(\nu)} \quad \text{as} \ \phi \to 0 \quad (5.60a)
\]
\[
\sim (1 - \phi)(1 + Ap\phi - Cp\phi^\nu) \quad \text{as} \ \phi \to 0 \quad (5.60b)
\]
\[
\sim 1 + A\delta\phi - Cp\phi^\nu \quad \text{as} \ \phi \to 0, \quad (5.60c)
\]
where \(\nu \equiv \gamma - 2\). Equation (5.60c) is the same as Eq. (5.56) and has a solution in the form of Eq. (5.42) only for \(\delta \leq 0\). Let’s define \(u = z^{-1}\), hence \(z = H^{-1}(u)\), and consider \(z\) and its derivative:
\[
z = 1 + A\delta(1 - u) - Cp\nu(1 - u)^\nu, \quad (5.61a)
\]
\[
\frac{dz}{du} = -A\delta + C\nu p\nu(1 - u)^{\nu - 1}. \quad (5.61b)
\]
Also consider the point \(u^*\) for which \(\left.\frac{dz}{du}\right|_{u^*} = 0\). According to Eq. (5.61b), \(1 - u^* = \left(\frac{A\delta}{C\nu p\nu}\right)^{\frac{1}{\nu}}\) and \(z^* = z(u^*)\) is obtained from Eq. (5.61a). Then for the expansion of \(z\) around \(u^*\) we get [91]
\[
z = z^* + \sum_{n=2}^{\infty} \frac{M_n}{n!} (u - u^*)^n. \quad (5.62a)
\]
To the leading order,
\[
z - z^* \sim \frac{M_2}{2} (u - u^*)^2 \quad \text{as} \ \ u \to u^*, \quad (5.63)
\]
where \(M_2 = -\nu(\nu - 1)Cp\nu(1 - u^*)^{\nu - 2} = -\nu(\nu - 1)Cp^\nu(\nu - 2)/(\nu - 1)\). Thus, the leading order behaviour of \(H_1(z)\) is determined by
\[
1 - \phi = u \sim u^* \pm \sqrt{\frac{2}{M_2}} (z - z^*)^{1/2} \quad \text{as} \ \ z \to z^*; \quad (5.64)
\]
as a result, the expansion of \(H_0(z)\) around \(z^*\) has a singularity at \(z = z^*\):
\[
H_0(z) \sim \text{analytic part} \pm \langle k \rangle p \sqrt{\frac{2}{-M_2}} (z^* - z)^{1/2} \quad \text{as} \ \ z \to z^*, \quad (5.65)
\]
as opposed to the assumption of Eq. (5.42) that the singularity (closest to the origin) is located at \(z = 1 - w = 1\). It is worth noting that \(M_2\) in Eq. (5.65) is real only for
5.6 Numerical results

nonnegative $\delta$; hence, as we will show in a moment, for $3 < \gamma < 4$, $\pi_s$, obtained from Eq. (5.65) is valid only for $p > p_c$.

Using Eq. (5.65) and the results of Appendix F, for $3 < \gamma < 4$ and $\delta > 0$, we obtain

$$\pi_s \sim \langle k \rangle p e^{\ln(z^*)/2} s^{-3/2} e^{-\ln(z^*)} s^s \quad \text{as } s \to \infty.$$  
(5.66)

Now, it is obvious that as, for $3 < \gamma < 4$, $M_2$ is real when $\delta > 0$, Eq. (5.66) is valid only for the supercritical phase. The rate of exponential decay is determined by $\ln(z^*)$ which is proportional to $\delta^{(\gamma-2)/(\gamma-3)}$, for $3 < \gamma < 4$, consistent with the result obtained according to finite size scaling [31].

It is worth mentioning that the above approach can also be used for $\gamma > 4$; in this case the term $C p^\nu \phi^\nu$ in Eq. (5.60a) is replaced with $B p^\nu \phi^\nu$. Consequently, $\pi_s$ is determined by Eq. (5.66), but with different $z^*$ and $M_2$ values. This result is valid for both the supercritical and subcritical region, and matches the result of method (IV) for $\gamma > 4$ (not shown).

5.6 Numerical results

In this section, we compare the theoretical predictions obtained in Sec. 5.5 with the results of direct numerical simulations for bond percolation on scale-free networks. For each given degree distribution, the network is constructed using the configuration model [100]. If there is any self-link or repeated link, they are rewired with randomly chosen links until there is none left [19]. From the numerical simulations, we first obtain $n_s$, the probability that in a realization of bond percolation process, a randomly chosen cluster has a size $s$. To calculate $n_s$, we obtain the size of clusters for a number ($R = 100$) of realizations of the percolation process on the network, and for each realization we count the fraction of clusters of size $s$. Then, $n_s$ is the mean of this fraction over the different realizations of the percolation process. Afterwards, we calculate the numerical value for $\pi_s$, the probability that a node is in a cluster of size $s$, using the relation $\pi_s = s n_s / \sum(s n_s)$. We also calculate the complementary cumulative distributions $N_s = \sum_{s_{\text{max}}}^s n_s$ and $\Omega_s = \sum_{s_{\text{max}}}^s \pi_s$. We compare the theoretical and numerical results for $\pi_s$ and also for $\Omega_s$. 

93
5. SIZE DISTRIBUTION OF PERCOLATION CLUSTERS AND NEURONAL AVALANCHES

In Fig. 5.1, we illustrate the numerical results for bond percolation on a scale-free network with exponent $\gamma = 4.4$ for three different values of occupation probability: $p < p_c$, $p = p_c$, and $p > p_c$. In Fig. 5.2, we observe that although method (I) underestimates $\pi_s$ and $\Omega_s$, the shape of distributions predicted by this method is similar to the numerical results. If we normalize the $\pi_s$ predicted by method (I) by its sum, we observe that the normalized $\pi_s$ and also $\Omega_s$ obtained accordingly match well the numerics for the large $s$ behaviour of $\pi_s$ and $\Omega_s$ (Fig. 5.2). This indicates that while the shape of the distribution is predicted correctly by method (I), the predicted normalization factor is inaccurate. We demonstrate in Fig. 5.2 that method (II) is even less accurate than method (I), while method (IV) is more accurate than both of them. We also show in Fig. 5.2 that Method (III) is the most accurate and its predictions for the large $s$ behaviour of $\pi_s$ and $\Omega_s$ match the numerical results very well (Fig. 5.2).
5.6 Numerical results

In Fig. 5.3, we illustrate the numerical results for bond percolation on another scale-free network with exponent $\gamma = 3.3$ for three different values of $p$ corresponding to subcritical, critical, and supercritical regimes of bond percolation. Figure 5.4 shows the

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figures/5.2}
\caption{Comparison between the theoretical predictions and numerical results for the scale-free network of Fig. 5.1. The cyan curves are obtained by normalizing $\pi_s$ obtained from method (I) by the sum of $\pi_s$. Method (III) provides accurate predictions for $\pi_s$ and $\Omega_s$. In panel (d), the small bump in the tail of the distribution can be due to finite size effects [67]. The results of using Eq. (5.66) (not shown) for the subcritical and supercritical regime match the results of method (IV).}
\end{figure}
5. SIZE DISTRIBUTION OF PERCOLATION CLUSTERS AND NEURONAL AVALANCHES

![Graphs showing size distribution of percolation clusters and neuronal avalanches.](image)

**Figure 5.3:** The numerical results for a scale-free network with exponent $\gamma = 3.3$. The network has $2 \times 10^7$ nodes and the minimum degree is set to 2. The difference between the subcritical and supercritical behaviour is evident for the complementary cumulative distributions.

A comparison between the theoretical predictions and numerical results. Figures 5.4(a)–5.4(d) show that for the subcritical and critical regimes method (III) performs very well, while method (I) is inaccurate. For the supercritical regime, the prediction of Eq. (5.66) is in good agreement with the numerics (Figs. 5.4(e)–5.4(f)). It is worth mentioning that, the theoretical predictions for the subcritical and supercritical regimes are reasonably accurate as long as $p$ is sufficiently far from the critical occupation probability $p_c$. This is because, in contrast to the case where $\gamma > 4$, for $3 < \gamma < 4$, in each of the subcritical, critical and supercritical regimes, $\pi_s$ has a distinct power-law exponent; hence, close to the critical point a mixed behaviour exists which cannot be simply described by equations for $\pi_s$ away from the critical point.

96
Figure 5.4: The comparison between the theoretical predictions and numerical results for the network of Fig. 5.3. The predictions for the subcritical and critical regimes, obtained by Method (III), are in good agreement with the numerics. The bump at the end of the tail for the cumulative distributions (panels (b) and (d)) may be due to finite size effects. Eliminating the $\pi_s$ values for the highest values of $s$ (yellow line in panel (d)) leads to a better agreement between the theoretical and numerical results for $\Omega_s$ for $s$ values away from the bump. The predictions for panel (e) and (f) are obtained from Eq. (5.66).
5. SIZE DISTRIBUTION OF PERCOLATION CLUSTERS AND NEURONAL AVALANCHES

5.7 Conclusion

In summary, we argued that bond percolation provides a useful paradigm for analysing the size distribution of neuronal avalanches in a simple integrate-and-fire model. We then provided analytical approaches to study cluster size distribution in networks with arbitrary degree distribution. Using the method of Newman [112], we showed that for some degree distributions, such as exponential and Poisson distributions, an exact closed-form relation can be derived for the probability $\pi_s$ that a node is part of a cluster of size $s$ in a bond percolation process (Sec. 5.4).

We also demonstrated that using asymptotic expansions of generating functions and Tauberian theorems, very good approximations can be obtained for $\pi_s$ and its complementary cumulative distribution $\Omega_s$. In this regard, we introduced the asymptotic method of Cohen et al. [31] for derivation of the functional form of $\pi_s$ at the critical point of bond percolation. We extended this method to calculate $\pi_s$ away from the critical point and obtained the multiplicative factor of $\pi_s$ as well. We showed that this set of calculations (referred to as method (I)) cannot provide a valid result for the supercritical phase of scale-free networks with exponent $3 < \gamma < 4$; we showed this by noting that the predicted $\pi_s$ (according to its multiplicative factor) is not a nonnegative real number. For other cases, we showed that method (I) can predict correctly the shape (functional form) of $\pi_s$; however, its prediction for the magnitude of multiplicative factor of $\pi_s$ is not accurate.

To find an accurate prediction of both the shape and the multiplicative factor of $\pi_s$, we considered modifications to method (I) and proposed three other asymptotic methods. We designed method (III) that produces most accurate results for critical and subcritical regimes of scale-free networks with $3 < \gamma < 4$ and for both critical and noncritical regimes of those with $\gamma > 4$. We also proposed method (IV) that produces valid and accurate results for supercritical regime when $3 < \gamma < 4$.

We showed that the relations (5.22) and (5.23), in method (I), obtained using heuristic methods [19] do not provide accurate predictions for $\pi_s$ values. The correct results are obtained by substituting these relations with those derived by Newman [112] (employed in methods (III) and (IV)). Also, we showed adding an extra expansion to method (I) (expanding $1/G_1(z)$ around $G_1(z) = 1$ that was employed in Methods (II) an (IV)), similar to the expansion performed by Lee et al. [91] in the calculation of
critical branches, provides a valid and accurate solution for the supercritical regime of scale-free networks with exponent $3 < \gamma < 4$, but, for $\gamma > 4$ causes some inaccuracy, compared to method (III).

As we mentioned in Sec. 5.1, calculation of the cluster sizes in bond percolation has potential application in studying neuronal avalanches as the cluster sizes can be associated with the size of neuronal avalanches (the number of neurons fired in an avalanche). In this regard, when a neuronal avalanche is modeled by a percolation process [21, 58, 59], the observation of scale-free distributions of cluster (avalanche) sizes may not be due to the criticality of the neural system; power-law distribution of cluster sizes may appear in different situations:

- At the critical point of the percolation process, the cluster sizes have a power-law distribution $n_s \propto \pi_s / s \propto s^{-\alpha}$ with $\alpha = (\gamma - 1)/(\gamma - 2) + 1$ for a scale-free network with degree distribution exponent $3 < \gamma < 4$, or $\alpha = 5/2$ for a non-scale-free network or a scale-free network with exponent $\gamma > 4$.

- In a scale-free network with the degree distribution exponent $\gamma$ between 3 and 4, the cluster size distribution in the subcritical regime is in the form of a power-law with exponent $\gamma$.

- In a non-scale-free network or in a scale-free network with exponent larger than 4 which has a large value of $\langle k \rangle \langle k^3 \rangle / \langle k^2 \rangle^2$: According to Eq. (5.40), $\pi_s \propto s^{-3/2} e^{-s/s^*}$, where $s^* \equiv b/a \propto \langle k \rangle \langle k^3 \rangle / \langle k^2 \rangle^2$. Therefore, the exponential factor in $\pi_s$ tends to 1 for $s \ll s^*$. Thus, if $s^*$ is large, for a large range of $s$ values, a power-law form with exponent $5/2$ is observed for the cluster size distribution.

When studying the criticality of neuronal avalanches, the above scenarios should be taken into account in order to understand the effect of network structure on producing power-laws and to unfold the relation of power-laws with the regime in which the dynamical process is operating. For example, the power-law distribution of avalanche sizes observed in neural data extend only for a few decades [59, 133]. Hence, one can not reject the possibility that such power-laws are pure and are not part of a power-law distribution with exponential decay of the type mentioned in the third case above. Moreover, the results of this chapter indicate that the exponent of the power-law distribution of cluster sizes can take a range of different possible values, depending on the
network type and the regime in which the dynamical process is operating. These can help to understand the different power-law exponents observed in neuronal avalanches data [59].

It is worth mentioning that a similar approach can be employed to calculate the cluster size distribution in directed networks, which are more realistic when neural networks are considered. This calculation has been done by Cohen et al. [136] to obtain cluster size distribution for the critical regime of directed scale-free networks. However, Cohen et al. assumed that away from the critical point the cluster size distribution is always in the form of a power-law with exponential decay, similar to the assumption they made for undirected networks [31]. As we demonstrated, this was not correct for the subcritical regime of undirected scale-free networks. Moreover, it was shown [67], using generating functions methods, that for a model of opinion spread (meme popularity) dynamics on directed networks, the popularity distribution in the subcritical regime of the model also follows a power-law, if the second moment of the degree distribution diverges; this result indicated that, for certain values of the degree distribution exponent, the scale-free structure of directed networks can cause power-law behaviour in the subcritical regime. Therefore, more detailed analytical investigations are required to verify the results of Ref. [136] for percolation behaviour of directed scale-free networks in the subcritical and supercritical regimes.

The methods described in this chapter for analysing the properties of generating functions can find application in studying some other important properties of avalanches such as the duration of avalanches [58]. The avalanche duration can be regarded as the diameter of an out-component in a directed network, and may be related to the diameter of a cluster in an undirected network. The method of Lee et al. [91] for finding the lifetime duration of critical branches can be potentially employed for calculating avalanche durations.
In this thesis, we investigated theoretical approaches for describing the percolation properties of networks. In Chapter 1, we gave a brief overview of different systems that are composed of interacting units and outlined how such a system is represented as a complex network. In particular, we provided several examples from social, biological, and technological systems. We described the fundamental units (nodes and links) perceived for each type of network, and described the meaning of interactions (links) considered between the entities (nodes) in that network. Then, we provided an overview of the network percolation process, which is investigated in further detail in other chapters of this thesis. We described bond and site percolation processes and outlined some quantities of interests in these models.

In Chapter 2, we presented the mathematical background for the material covered in this thesis. In this regard, we introduced some of the measures of networks’ structural properties, and described several models for constructing synthetic networks with specific structural features. Lastly, we presented an overview of adjacency tree-based \((A_{ij})\) theories for three different binary-state processes, i.e., percolation, susceptible-infected model of epidemic spread, and the Ising model, and outlined how these theories employ the information on the connectivity of nodes to predict macroscopic behaviour of such processes.

Chapters 3, 4, and 5 all contain novel results; the results of Chapters 3 and 4 have appeared in two papers (Refs. [52] and [51]). In Chapter 3, we investigated the effect of short loops on the accuracy of the \(A_{ij}\) theory for percolation as well as for the other binary-state processes introduced in Chapter 2. We proposed the \(L\)-cloned networks
ensemble that can be constructed from any given real-world or synthetic network using a certain mechanism that involves rewiring of the links. An $L$-cloned network has $L$ times more nodes and edges than the original network. On the other hand, the degree distribution and degree-degree correlations between and beyond the nearest neighbours of the original network are persevered, while the density of short loops is decreased as $L$ is increased. We also developed two extension of $L$-cloning: (i) $L_X$-cloning that preserves the density of short loops of specified lengths and decreases that of other loops, and (ii) $L_f$-cloning that has the same number of rewired links as in $L_X$-cloning but does not preserve the density of short loops. We demonstrated that $L$-cloned networks with sufficiently large $L$ are the ensemble of networks for which $A_{ij}$ theories are accurate. Using the $L$-cloning method and its extensions, we showed the significant effect of short loops on the accuracy of $A_{ij}$ theories. We also showed that the high density of short loops is not the only major source of inaccuracy of $A_{ij}$ theories and that there possibly exists another phenomenon not captured by such theories.

In Chapter 4, we considered the $A_{ij}$ theory for bond percolation, which is referred to as the message-passing (MP) approach. We revealed an important source of error (other than the presence of short loops) in predictions of the MP theory for the size of the percolating cluster: that is, on modular networks with a finite and sufficiently small number of interlinks, coexisting percolating clusters (CPCs) emerge, and such CPCs are an important source of inaccuracy of the MP approach in several real-world and synthetic networks. We developed the modular message-passing (MMP) approach which takes into account the presence of CPCs and improves on the state-of-the-art MP theory. These results are consistent with and explain our foregoing findings: we had shown in Chapter 3 that $L_X$-cloning of networks (which increases their number of nodes and links but preserves the organization of short loops with specific sizes) can increase the accuracy of the MP theory in networks such as the Western US power grid. This observation can be explained according to the number of interlinks: as $L_X$-cloning increases the number of interlinks too, it reduces the effect of CPCs on the inaccuracy of the MP theory and improves its performance.

The results of Chapter 4 show the substantial effect of the number of interlinks on percolation behaviour and on the accuracy of the state-of-the-art MP theory for this process. The number of interlinks in modular networks may also affect the accuracy of theoretical approaches for other binary-state processes. It was previously shown
that inaccurate results are expected for mean field theories on networks with low mean degree and mean nearest neighbours degree and/or high mean shortest path length [66, 95], and on modular networks with finite number of interlinks [96]; however the underlying phenomenon behind these inaccuracies was not identified. A possible explanation for these observations is according to the number of interlinks: modular networks can have a relatively large mean shortest path length and, moreover, a small number of connections (low mean and mean nearest neighbours degrees) in a modular network indicates the possibility that the network has a small number of interlinks. Therefore, our findings in Chapter 4, together with the results of other studies, suggest the importance of the number of interlinks in determining the behaviour of processes on modular networks.

In Chapter 5, we focused on the theoretical approaches for calculation of another quantity of interest in percolation processes, i.e., the distribution of cluster sizes. In theoretical approaches, this distribution is obtained from $\pi_s$, which is the probability that a randomly chosen node belongs to a cluster with size $s$. We focused on the calculation of $\pi_s$ for undirected unweighted networks. In this regard, we first presented an overview of the Nemwan's method [112] which provides a closed-form solution for $\pi_s$ on some configuration model networks with a given degree distribution. We then corrected the errors in the reported results of Ref. [112] for the large $s$ behaviour of $\pi_s$.

Then, we considered the method of Cohen et al. [31] that gives a solution for networks with any given degree distribution, in particular scale-free degree distributions. We showed that this method predicts correctly the shape (i.e., the functional form) of $\pi_s$ for the (percolation) critical regime of scale-free networks with degree distribution exponent $\gamma > 4$. Additionally, we performed the calculations for the multiplicative prefactor of $\pi_s$ and showed that the method of Cohen et al. does not give a correct prediction for the prefactor. Moreover, we performed further calculations for noncritical regimes and demonstrated that the functional form that Cohen et al. assumed $\pi_s$ will have away from the critical point is not always correct and holds only for $\gamma > 4$. We obtained the interesting result that for $3 < \gamma < 4$, at the subcritical regime, $\pi_s$ has a power-law form, which is in stark contrast to the assumption of Cohen et al. Also, we showed that their method cannot provide a valid result for the supercritical regime of $3 < \gamma < 4$. 

103
6. CONCLUSION

In addition to extending the method of Cohen et al., we provided three other methods that had certain differences compared to their method. Consequently, we developed methods that can predict accurately the full form (i.e., both the shape and the multiplicative prefactor) of $\pi_s$ at all the three regimes for $\gamma > 4$ or $3 < \gamma < 4$. We also described how percolation can be used as a simplified model of neuronal dynamics (Sec. 5.1) and mapped the size of neuronal avalanches to the size of percolation cluster sizes. Accordingly, in Sec. 5.7, we discussed the implications of our results and their potential application in investigating the behaviour of neuronal avalanches. We argued that a power-law distribution of neuronal avalanches can be observed even away from the critical point of the model. This, as mentioned above, occurs in the subcritical regime of scale-free networks with $3 < \gamma < 4$. Also, for networks with $\gamma > 4$, we found that if $\langle k \rangle \langle k^3 \rangle / \langle k^2 \rangle^2 \gg 1$ (where $\langle \Box \rangle = \sum_{k=0}^{\infty} \Box P(k)$ and $P(k)$ is the degree distribution) the avalanche size distribution takes the form of a power-law for a relatively broad range of avalanche sizes. These results can help to shed more light on the dynamical behaviour of neural systems.

In summary, we investigated the existing theoretical approaches for percolation on networks and the relation of these results with several other network processes. We first considered a class of theories developed for several binary-state dynamics and focused in particular on the theoretical predictions for the size of the network percolating cluster. We proposed the novel $L$-cloning methods that help to shed more light on the accuracy of these theories, especially on clustered networks. Moreover, we introduced the modular message passing theory that illustrates the emergence of a percolation phenomenon that had not been captured before. Next, we developed methods for describing the distribution of percolation cluster sizes and discussed the potential applications of these methods in understanding the statistics of neuronal avalanches. These novel results provide a better understanding of the percolation properties of networks and open up new avenues for research on other processes that operate on networks.
Appendix A: Relation between $L$-cloned and multilayer networks

To describe an $L$-cloned network as a multilayer network, we use the terminology of Ref. [83]. Accordingly, an $L$-cloned network has one aspect on which there exist $L$ layers of the network and all the layers have an equal size. The inter-layer connections are not couplings, hence $L$-cloned networks does not have any of the diagonal, layer-coupled or categorial couplings. If the clones of each node are assumed to be the same node then $L$-cloned networks are node-aligned; otherwise if different identities are assumed for clones of each node, then $L$-cloned networks are layer-disjoint networks. The former can be argued according to the fact that all clones of a certain node have the same degree and degree-degree correlation between and beyond nearest neighbours (i.e., the degree sequence $\{K\}_{i,d}$ of nodes at any distance $d$ from any node $i$ in an $L$-cloned network is identical to that of all other clones of node $i$). However, as other structural characteristics of clones of a node (e.g., closeness and betweenness [116]) are not exactly the same, assuming $L$-cloned networks to be the latter case, i.e., layer-disjoint networks, is also legitimate and informative.
A. APPENDIX A: RELATION BETWEEN L-CLONED AND MULTILAYER NETWORKS
Appendix B: Calculation of the probability that two CPCs are connected

As described in Eq. (4.5) of the main text, the probability that the CPCs of two neighbouring modules become connected at an occupation probability $p$ is:

$$\eta_{mn} = 1 - (1 - p v_m v_n)^I.$$  \hspace{1cm} \text{(B.1)}

Here, we describe the calculation of the probabilities $v_m$ and $v_n$ for $z$-regular modules of an SLN using the $p_k$ theory [95, 116]. These calculation can be further extended to more general cases including Eq. (4.7) in the main text which is for a given general structure. According to Eq. (16.2) of Ref. [116], in a random uncorrelated graph with degree distribution $P(k)$ we have:

$$u = 1 - p + p \sum_k k' P(k') \frac{u^{k'}}{z} u^{k' - 1},$$  \hspace{1cm} \text{(B.2)}

$$S_m = 1 - \sum_k P(k) u^k,$$  \hspace{1cm} \text{(B.3)}

where $u$ is the probability that a link from a node with degree $k$ leads to the percolating cluster (PC) of the graph, and $S_m$ is the size of the PC of the graph (or equivalently the probability that a randomly chosen node is in the PC). For a $z$-regular graph the above equations simplify to:

$$u = 1 - p + p u^{z-1},$$  \hspace{1cm} \text{(B.4)}

$$S_m = 1 - u^z.$$  \hspace{1cm} \text{(B.5)}
In an SLN with $z$-regular modules, for each link connecting the nodes in module $m$, Eq. (B.5) gives $u = (1 - S_m)^{1/z}$. It is safe to assume that the boundary nodes have only one link to the outside of the modules because in the creation of an SLN the number of links swapped with those of the neighbouring module is much smaller than the total number of links in each module. Accordingly, the boundary nodes have one connection less to the nodes inside their module and therefore the probability $v_m$ that a boundary node in module $m$ is in the CPC of that module is:

$$v_m = 1 - u^{z^{-1}} = 1 - (1 - S_m)^{(z^{-1})/z}.$$  \hfill (B.6)

Hence, for example, for the SLN of Figs. 4.3(b) and 4.3(c) of the main text we first calculate $u$ for each value of $p$ using Eq. (B.4) and then obtain $v_m$ (and similarly $v_n$) for the 4-regular module (3-regular module) by substituting $u$ and $z = 4$ (and $z = 3$) in Eq. (B.6). Then substituting $v_m$ and $v_n$ in Eq. (B.1) gives the probability $\eta_{mn}$ that the CPCs of the two modules are connected.

This calculation can be performed for each pair of modules in an uncorrelated network to obtain $\eta_{mn}$, and it can also be extended to the case of a correlated random network. In a more general case of a treelike network with a given adjacency matrix and modular structure, $\eta_{mn}$ is calculated using Eq. (4.7) in the main text.
Appendix C: Accuracy of the MMP approach for LFR networks with small modules

As discussed in the main text, the accuracy of the MMP theory can be affected if there are very small modules in a network. An example for this case was shown in Fig. 4.5(c). Figure C.1(a) shows similar results for an LFR network with different parameters. In the LFR network of Fig. C.1(a) there exist many small modules; the minimum module size in this network is 10 and the module sizes are power-law distributed with the exponent of $\beta = 2$. The slight deviation of the MMP results from the numerics in Fig. C.1(a) can be associated with the presence of such small modules. In Figs C.2(a)-C.2(c) we illustrate results for the dependency of the performance of theoretical methods on the size of the modules as well as some other parameters in LFR networks with considerable number of small modules. The performance of the MMP and MP theories is measured by the mean absolute error $E$ between the theoretical predictions and numerics. Despite the fact that there exist many small modules in the LFR networks of Fig. C.2(a), the MMP theory performs well, showing an error much smaller than that of the MP theory. The small mixing of $\mu = 0.005$ between the modules facilitates the formation of (independent) CPCs and helps MMP to perform well. For lower average degree of 1.9, MMP is slightly less accurate (Fig. C.2(a)), as some modules resemble small trees that can not percolate independently. On the other hand, the LFR networks of Figs. C.2(b) and C.2(c) have $\beta = 2$ and hence a larger number of small modules; they also have a larger $\mu = 0.1$. On such networks, MMP theory still performs very well when the minimum size of modules, $N_{m}^{\text{min}}$, is sufficiently large; however for $N_{m}^{\text{min}} \leq 50$ its accuracy starts to decrease, but it still outperforms the
C. APPENDIX C: ACCURACY OF THE MMP APPROACH FOR LFR NETWORKS WITH SMALL MODULES

**Figure C.1:** Bond percolation results for two LFR networks with small module sizes. The exponents for the degree and module size distributions are respectively $\alpha = 3$, $\beta = 2$ for the network of panel (a) and $\alpha = 3$, $\beta = 1$ for the network of panel (b). The circles represent $\bar{S}$ for numerical simulations. The colored region denotes the interval between the percentiles of the distribution of $\bar{S}$ for single realizations of bond percolation obtained from the numerical simulations. The blue solid and dashed lines denote the results obtained from the MMP theory.

**Figure C.2:** The mean absolute error for the MMP and the MP theories versus $N_m^{(\text{min})}$ for LFR networks with $N = 5000$, average degree $k_{av}$, $N_m^{(\text{max})} = 1000$. In (a) $\mu = 0.005$, $\alpha = 2$, $\beta = 1$, and in (b) and (c) $\mu = 0.1$, $\alpha = 2.5$, and $\beta = 2$. The error bars show the standard deviation for 50 realizations of the LFR networks with the same specified parameters.
MP theory. Figures 4.5(d) and C.1(b) show percolation results and the accuracy of the MMP theory on examples of LFR networks with modules that have as few as 50 nodes.
C. APPENDIX C: ACCURACY OF THE MMP APPROACH FOR LFR NETWORKS WITH SMALL MODULES
Appendix D: Application of the MMP theory to real-world networks

The results of employing the MMP theory for bond percolation on several real-world networks are shown in Fig. D.1. We first used the Ronhovde-Nussinov (RN) community detection algorithm [131] to identify different representations for the network modular structure. Each representation corresponds to a realization of the detection algorithm for a given value of the input parameter of the algorithm and has modules of specific sizes, possibly different from those in the other representations. Accordingly, we find the representation that maximises $\frac{M-M_s}{M}Q$. Figure D.2 shows that for real-world networks, the representation with maximum $\frac{M-M_s}{M}Q$ has a modularity $Q$ close to the maximum possible $Q$. Hence, optimising for $\frac{M-M_s}{M}Q$ yields a representation with small number of modules of very small sizes yet with high modularity, capturing the modular structure of the network.

The results for bond percolation—shown in Fig. D.1—demonstrate that MMP improves significantly over the predictions of the MP theory for the real-world networks. The MMP results on the Facebook [94] and protein folding [129] networks matches very well the numerical results. For the rest of the real-world networks considered, the MMP predictions are close to the numerical results on the $P_{k,k'}^{i,i'}$ rewired version of the corresponding real-world network [96] (Fig. D.1), in which the density of short loops is significantly reduced, but the modular structure and degree-degree correlations inside and between modules are preserved. This confirms that in these real-world networks a significant cause of deviation of numerical results from the MP theory is the emergence
of the CPCs. The MMP theory also provides a prediction of how $S$ varies for single realizations of the percolation process according to the percentiles for the distribution of $S$ (Fig. D.1).

Figure D.1: Results for the bond percolation process for several real-world networks. The networks are (a) Facebook [94], (b) protein folding [129], (c) the Polish power grid [47,156], (d) the network science coauthorships [110], (e) the international E-road network [146], and (f) the s_838 electronic circuit [99] network. In all examples our MMP theory outperforms the standard MP theory.
Figure D.2: Measurements of the modular structure for the real-world networks used in this chapter (see also Table 4.1). For each network, the values of $Q$ and $\frac{M-M_s}{M}Q$ are plotted with respect to $M$ for the different representations of the modular structure obtained from at least 30 realizations of the RN algorithm. Results of the MMP theory in Figs. 4.6 and D.1 correspond to the modular structure with the maximum $\frac{M-M_s}{M}Q$. The networks are (a) Facebook [94], (b) protein folding [129], (c) the Polish power grid [47, 156], (d) the network science coauthorships [110], (e) the international E-road network [146], (f) the s_838 electronic circuit [99], (g) the AS Internet [92], and (h) the western US power grid [148] network.
D. APPENDIX D: APPLICATION OF THE MMP THEORY TO REAL-WORLD NETWORKS
Appendix E: Derivation of Eq. (5.40)

Using the Cauchy theorem (Eq. (5.12)), Eq. (5.10) can be written as

\[ \pi_s = \frac{1}{2\pi i} \oint_C \frac{H_0(z)}{z^s} \, dz, \]  

(E.1)

where \( C \) can be any closed contour around the origin that does not enclose any poles in \( H_0 \). We choose a unit circle and deform it into the keyhole contour \( C_R \cup l_1 \cup C_\epsilon \cup l_2 \) shown in Fig. E.1. In Fig. E.1, \( \alpha \) is a branch point, and the branch cut is located on the real axis from \( \alpha \) to \( \infty \). It is straightforward to show that the integrals along the circular contours \( C_R \) and \( C_\epsilon \) limit to zero [67, 81]. Hence, for \( H_0(1 - w) \sim \) analytic part \( -\sqrt{a + bw} \) as

**Figure E.1:** The keyhole contour \( C \) in the complex \( z \)-plane for the integral of Eq. (E.1). A branch cut extends from \( \alpha = 1 + a/b \) (see Eq. (E.2b)) to \( \infty \). The circular arcs have radii \( \epsilon \) and \( R \).
w \to 0, we will have:

\[ \pi = -\frac{1}{2\pi i} \int_{l_1 \cup l_2} \frac{\sqrt{a + b} (1 - z)}{z^s} \, dz \quad (E.2a) \]

\[ = \frac{1}{2\pi i} \int_{l_1 \cup l_2} \frac{\sqrt{a + b w}}{(1 - w)^s} \, dw \quad (E.2b) \]

\[ = -\frac{1}{2\pi i} \int_{l_1 \cup l_2} \frac{\sqrt{a - b \rho}}{e^\rho} \, d\rho \quad (E.2c) \]

\[ = \frac{e^{-(a/b)\pi}}{2\pi b} \int_{l_1 \cup l_2} u^{1/2} e^{(a/b)u} \, du, \quad (E.2d) \]

where in Eq. (E.2a), we used the fact that the net contribution from the analytic part of \( H_0 \) on \( l_1 \cup l_2 \) is zero. In Eq. (E.2c), we made the substitution \( z = e^\rho \), and \( w = 1 - e^\rho \sim -\rho \) as \( \rho \to 0 \). To obtain Eq. (E.2d), we employed the substitution \( u = a - b\rho \). Next, we make another change of variable: \( u = re^{i\theta} \), where

\[
\begin{cases}
\text{along } l_1 : & \theta = \pi, \quad u = -r, \quad \text{and } \quad u^{1/2} = r^{1/2} e^{i\pi/2} \\
\text{along } l_2 : & \theta = -\pi, \quad u = -r, \quad \text{and } \quad u^{1/2} = r^{1/2} e^{-i\pi/2}.
\end{cases}
\]

(E.3)

To write the limits of integrals note that, along \( l_1 \), \( w \) in Eq. (E.2b) varies from \(-\infty\) to \(-a/b\); this corresponds to a variation in \( z \) from \( \infty \) to \( 1 + a/b \). Hence, \( \rho \) in Eq. (E.2c) varies from \( \infty \) to \( a/b \), and \( u \) varies from \(-\infty\) to 0. Similarly, along \( l_2 \), \( u \) varies from 0 to \(-\infty\). branch cut lies from 0 to \(-\infty\); hence, on the interval \([-a, 0)\), integrals on \( l_1 \) and

![Figure E.2: The contour of Fig. E.1 sketched in the complex u-plane.](image)
\( l_2 \) give opposite contributions that are canceled out. Thus, along \( l_1 \), \( r \) varies from \( \infty \) to 0, and along \( l_2 \), it varies from 0 to \( \infty \). Therefore,

\[
\pi_s = \frac{e^{-a/b_s}}{2\pi b} \left[ \int_\infty^0 e^{i\pi/2} + \int_0^\infty e^{-i\pi/2} \right] r^{1/2} e^{-(s/b)r} (-dr) \quad (E.4a)
\]

\[
= \frac{e^{-(a/b_s)}}{\pi b} \left[ e^{i\pi/2} - e^{-i\pi/2} \right] \int_0^\infty r^{1/2} e^{-(s/b)r} dr \quad (E.4b)
\]

\[
= \frac{e^{-(a/b_s)}}{\pi b} (b/s)^{3/2} \int_0^\infty t^{3/2-1} e^{-t} dt \quad (E.4c)
\]

where the change of variable \( r = (b/s)t \) is performed in Eq. (E.4c). Now, the integral in Eq. (E.4c) equals \( \Gamma(3/2) = \sqrt{\pi}/2 \). Hence,

\[
\pi_s \sim \frac{b}{4\pi} s^{-3/2} e^{-(a/b_s)} \quad (E.5)
\]
Appendix F: \((a - z)\beta\) non-analyticity

Consider a function \(H(z) = \sum_{n}^{\infty} \pi_n z^n\) that has a branch point at \(z^* = a\); hence its expansion around \(a\) has a leading order behaviour in the form

\[
H(z) \sim \text{analytic part} + R(a - z)^\beta \quad \text{as } z \to a. \tag{F.1}
\]

Then sequence \(\pi_n\), for large \(n\), is determined by the leading order nonanalytic term\(^1\) \(R(a - z)^\beta\) \([151]\). Hence, to obtain \(\pi_n\), we need to find the coefficients of \(z^n\) in the expansion of \((a - z)^\beta = \sum_{n}^{\infty} \eta_n z^n\). To do so, we can use binomial expansion \([151]\):

\[
(a - z)^\beta = \sum_{n=0}^{\infty} \binom{\beta}{n} (-z)^n a^{\beta-n} \tag{F.2a}
\]

\[
= \sum_{n=0}^{\infty} \binom{n - \beta - 1}{n} (\ln a)^{(\beta-n)} z^n. \tag{F.2b}
\]

\(^1\)If the function has more than one singular point, for example two singularities, the behaviour of \(\pi_n\) for the largest values of \(n\) is determined by the leading order term for the branch point closest to the origin; but for some smaller \(n\) (which are still sufficiently large) the behaviour may be determined by the leading order term for the branch point further from the origin \([91]\). The accuracy of approximations for \(\pi_n\) is increased by the number of singularities considered according to a non-increasing order of the modulus of the singular points \([151]\).
F. APPENDIX F: \((A - Z)^\beta\) NON-ANALYTICITY

Accordingly, \(\eta_n\), which are the coefficients of \(z^n\) in the above expansion, are equal to

\[
\eta_n = \binom{n - \beta - 1}{n} e^{(\ln a)(\beta-n)}
\]  
(F.3a)

\[
= \frac{\Gamma(n - \beta)}{\Gamma(-\beta)\Gamma(n + 1)} e^{(\ln a)(\beta-n)}
\]  
(F.3b)

\[
\sim \frac{e^{(\ln a)\beta}}{\Gamma(-\beta)} n^{-(\beta+1)} e^{-(\ln a)n} \quad \text{as } n \to \infty,
\]  
(F.3c)

where in F.3c we used Stirling’s approximation \(\Gamma(n + 1) = n! = \sqrt{2\pi n} \left(\frac{n}{e}\right)^n\). Hence,

\[
\pi_s \sim \frac{Re^{(\ln a)\beta}}{\Gamma(-\beta)} n^{-(\beta+1)} e^{-(\ln a)n} \quad \text{as } n \to \infty.
\]  
(F.4)

For the generating function \(H(z)\) in the form of Eq. (5.24), \(a = 1\) and for that in the form of Eq. (5.65), \(a = z^*\). For \(a = 1\), Eq. (5.25) is obtained. For \(a \sim 1 + \xi\) (which is the case for Eq. (5.65)), and \(\xi << 1\) we get

\[
\pi_s \sim \frac{Re^{\xi\beta}}{\Gamma(-\beta)} n^{-(\beta+1)} e^{-\xi n} \quad \text{as } n \to \infty.
\]  
(F.5)
**URL:** [http://dx.doi.org/10.1023/A:1024474819512](http://dx.doi.org/10.1023/A:1024474819512)

**URL:** [http://doi.acm.org/10.1145/335305.335326](http://doi.acm.org/10.1145/335305.335326)

**URL:** [http://dx.doi.org/10.1007/978-1-4615-0005-6_4](http://dx.doi.org/10.1007/978-1-4615-0005-6_4)


**URL:** [http://www.nature.com/nature/journal/v397/n6715/abs/397168a0.html](http://www.nature.com/nature/journal/v397/n6715/abs/397168a0.html)

**URL:** [http://dx.doi.org/10.1140/epjst/e2014-02266-y](http://dx.doi.org/10.1140/epjst/e2014-02266-y)
REFERENCES


URL: http://dx.doi.org/10.1371%2Fjournal.pone.0017680 4

URL: http://www.sciencedirect.com/science/article/pii/S0378437102007367 3


URL: http://dx.doi.org/10.1140/epjb/e2004-00127-8 3

URL: http://www.jneurosci.org/content/23/35/11167.abstract 13


124


[26] Caldarelli, G. [2007], *Scale-free networks: complex webs in nature and technology*, Oxford University Press. 2, 3, 9, 10

[27] Caldarelli, G. and Vespignani, A. [2007], *Large scale structure and dynamics of complex networks*, World Scientific, Toh Tuck Link, Singapore. 23
REFERENCES


**URL:** http://dx.doi.org/10.1371%2Fjournal.pcbi.1000807  


**URL:** http://www.pnas.org/content/100/21/12516.abstract


**URL:** http://arxiv.org/abs/0705.0010 30, 31

**URL:** http://link.aps.org/doi/10.1103/PhysRevE.64.025101

**URL:** http://www.sciencedirect.com/science/article/pii/S0378873315000313

**URL:** http://link.aps.org/doi/10.1103/PhysRevE.66.035103

**URL:** http://www.sciencedirect.com/science/article/pii/S0370157307001287


23
REFERENCES


[54] *Figure 1.1 is taken from Wikimedia commons at the specified url [See below].* URL: http://www.fermentas.com/techinfo/nucleicacids/maplambda.htm

[55] *Figure 1.2 is taken from Wikimedia commons at the specified url [See below].* URL: https://commons.wikimedia.org/wiki/File:Dendrite_(PSF).svg


REFERENCES


REFERENCES

URL: http://doi.acm.org/10.1145/1298306.1298334


URL: http://dx.doi.org/10.1023/A:1015227414929


URL: http://www.nature.com/nature/journal/v407/n6804/full/407651a0.html

URL: http://www.pnas.org/content/93/1/13.abstract

URL: http://link.aps.org/doi/10.1103/PhysRevE.82.016101

URL: http://link.aps.org/doi/10.1103/PhysRevE.82.066118

REFERENCES


REFERENCES

URL: http://stacks.iop.org/1751-8121/40/i=26/a=002 81, 84, 88, 91, 92, 98, 100, 121

URL: http://doi.acm.org/10.1145/1081870.1081893 63, 68, 115

URL: http://www.jstor.org/stable/2083277 4


URL: http://scitation.aip.org/content/aip/journal/chaos/24/2/10.1063/1.4869983 24, 45, 51, 59, 69, 103, 113


URL: http://www.ams.org/publications/authors/books/postpub/gsm-75 86

URL: http://www.sciencemag.org/content/303/5663/1538.abstract 63, 114, 115

132
**URL:** http://dx.doi.org/10.1002/rssa.3240060204

**URL:** http://link.aps.org/doi/10.1103/PhysRevLett.105.128701

**URL:** http://link.aps.org/doi/10.1103/PhysRevE.64.016131


**URL:** http://link.aps.org/doi/10.1103/PhysRevE.66.016128

**URL:** http://epubs.siam.org/doi/abs/10.1137/s003614450342480


**URL:** http://dx.doi.org/10.1080/00107510500052444

**URL:** http://link.aps.org/doi/10.1103/PhysRevE.74.036104
**URL:** http://www.pnas.org/content/103/23/8577.abstract

**URL:** http://journals.aps.org/pre/abstract/10.1103/PhysRevE.76.045101


**URL:** http://www-personal.umich.edu/~mejn/networks-an-introduction/

**URL:** http://www-personal.umich.edu/~mejn/networks-an-introduction/

**URL:** http://www-personal.umich.edu/~mejn/networks-an-introduction/

**URL:** http://link.aps.org/doi/10.1103/PhysRevE.69.026113

**URL:** dx.doi.org/10.1103/PhysRevE.64.026118

**URL:** http://link.aps.org/doi/10.1103/PhysRevE.64.016706

**URL:** http://www.sciencedirect.com/science/article/pii/S0378437113000575

134

**URL:** [http://science.sciencemag.org/content/342/6158/1238411](http://science.sciencemag.org/content/342/6158/1238411)


[26, 41]


**URL:** [http://www.ams.org/notices/200909/21](http://www.ams.org/notices/200909/21)


[3]


**URL:** [http://www.nature.com/ncomms/2015/151215/ncomms10196/full/ncomms10196.html](http://www.nature.com/ncomms/2015/151215/ncomms10196/full/ncomms10196.html)


[63, 113, 114, 115]


[34]
URL: http://link.aps.org/doi/10.1103/PhysRevE.80.016109  68, 113

URL: http://www.pnas.org/content/105/4/1118.abstract 21, 68

URL: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002038 72, 99

URL: http://bioinformatics.oxfordjournals.org/content/17/10/988.abstract 10

URL: www.sciencedirect.com/science/article/pii/S0370157315002008 51

URL: http://link.aps.org/doi/10.1103/PhysRevE.66.015104 81, 100

URL: https://books.google.ie/books?id=MJoIGBjYDGEC 2, 3


URL: http://cercor.oxfordjournals.org/content/early/2014/10/21/cercor.bhu252.abstract 13, 73


[URL: http://dx.doi.org/10.1080/01972240590925348]


[URL: http://rspb.royalsocietypublishing.org/content/268/1478/1803]

[URL: http://www.nature.com/nature/journal/v393/n6684/abs/393440a0.html]

[URL: http://link.springer.com/article/10.1007%2FBF02179402]
REFERENCES


URL: http://dx.doi.org/10.1080/00220671.1926.10879675


URL: https://www.math.upenn.edu/~wilf/DownldGF.html


URL: http://www.sciencedirect.com/science/article/pii/S0378437109000338


URL: http://dl.acm.org/citation.cfm?id=1540618