Measuring Robustness
of Complex Networks
Measuring Robustness
of Complex Networks

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op maandag 28 oktober 2013 om 10:00 uur

doors

Javier MARTÍN HERNÁNDEZ

Enginyer de Telecomunicació
van Universitat Politècnica de Catalunya, Spanje
geboren te Barcelona, Spanje.
Dit proefschrift is goedgekeurd door de promotor:

Prof. dr. ir. P.F.A. Van Mieghem

Samenstelling promotiecommissie:

Rector Magnificus voorzitter
Prof. dr. ir. P.F.A. Van Mieghem Technische Universiteit Delft, promotor
Prof. dr. ir. R.E. Kooij Technische Universiteit Delft
Prof. dr. ir. D. Epema Technische Universiteit Delft
en Technische Universiteit Eindhoven
Prof. dr. ir. C.M. Scoglio Kansas State University
Prof. dr. ir. D. Hutchison Lancaster University
Dr. ir. G. D’Agostino Italian National Agency for New Technologies, Energy and Sustainable Economic Development
Dr. ir. H. Wang Technische Universiteit Delft

Alma mater: Technische Universiteit Delft

Printed in the Netherlands by SIECA REPRO.
Front and back cover: The Link is Down. Picture by Cong Li. Original idea by Javier Martín Hernández.

Copyright © 2013 by J. Martín Hernández
All rights reserved. No part of the material protected by this copyright notice may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without the prior permission of the author.


Typeset by the author with the \LaTeX Documentation System.
Author email: javier.martin.hernandez@gmail.com
For mom and dad,
the stalwart supporters.
Summary

Measuring Robustness of Complex Networks

Networks are everywhere. Imagine for a minute two sisters, Julia and Maria, going through their daily chores. After switching off their alarm clocks and brushing their teeth, they devour a Mediterranean breakfast before heading to school by train, where they gossip about their latest Twitter posts. What Julia and Maria do not realize is that their lives fully depend on human-made networks. The power grid network supplied the electricity to their alarm clock, a water supply network together with the gas network provided the hot water to rinse their mouths and cook their breakfasts, a public transportation network drove them to school. All along, the train’s Wi-Fi provided Internet connection for them to access their online social hub through the cell-phone network. With the proliferation and penetration of technology into all areas of human life, networks have become a normality and even a necessity, contributing and enabling to a large extent our current quality of life. Large scale power outages like the great Northeastern blackout of 2003, the AS7007 black hole, the Pakistan Telecom-YouTube, or most recent Google incidents, show that even small or locally contained causes may have a large scale regional and in some instances even global impact on infrastructures, creating significant damages to a company, country or society in general.

An essential question arises: “How can we make our networks stronger, or more robust?” For example, the network manager of TUDelft would wish to know whether his current network is good, or a civil servant may want to know whether the railway network of The Netherlands will work normally. Unfortunately, these simple questions are ill-posed, and not clearly stated because “normal” or “good” need to be defined. Furthermore, contrasting definitions of robustness may arise from the use of the very same network. For example, let us imagine a communication network built to propagate data between networked computers. From a data transfer point of view (e.g. downloading a YouTube video), the given network is robust only if it transfers the entirety of the video in the presence of failures, regardless of delays. However, from a real-time application point of view (e.g. a Skype call), a robust communication network enables data to be
transferred without delays, even in the presence of losses. Provided a notion of robustness is settled upon, this thesis successfully addresses the issue of finding a unique number (or set of numbers) that unequivocally quantify robustness.

My approach to quantifying robustness works as follows. Since robustness expresses the quality of a network, I assume that it is bound to be a function of a set of network properties. By expressing a network as a graph, i.e. a set of nodes interconnected by a set of links, various properties of real networks can then be captured by means of linear algebra, and condensed as graph metrics, which are typically real-valued functions of a network model. The first half of this thesis (Chapters 2 and 3) is devoted to exploring algebraic ways of characterizing a network, which results in a metric taxonomy. The proposed taxonomy allows any existing metric to be categorized into three classes: distance, connection and spectra. Such a taxonomy will aid future researchers to better understand their networks, by offering them a complete and organized classification of network descriptors. Moreover, based on three independent graph metrics, Chapter 3 introduces three robustness definitions. These case-studies illustrate how robustness can be arbitrarily defined in any way that suits a given network purpose, and will expose the challenges that arise when attempting to do so, such as metric correlations (Chapter 4). In parallel, this thesis provides three major contributions: a novel challenge-aware protection scheme, a better understanding of modularity-based quarantine procedures, and a the discovery of a new phase transition in the context of interdependent networks.

In addition to the previous contributions, I believe that no definition of robustness is complete without bringing the external or hostile factors into the picture, i.e. events that cause the network to malfunction. In other words, robustness should not be defined as a static picture of an unmovable network, but as the ability of a network to adapt to adverse effects. For example, it has been recently discovered that the family of power-law graphs is very robust against random component failures. On the other side, this same type of network is extremely vulnerable to the manual extraction of carefully chosen nodes. This effect underlines the necessity of adding a new input variable to our robustness function; we named this variable the challenge dimension. In the second half of this thesis, Chapters 4 and 5 account for these external influences, and propose a general robustness framework that includes the mentioned challenge dimension, by means of studying robustness envelopes of impaired networks. This framework defines, characterizes, and computes network robustness. By being able to map a network topology to a unique number, we made possible the design of robust networks through the use of optimization algorithms. In addition, robustness benchmarks can be standardized so that networks are easily classified into robustness hierarchies.

Javier Martín Hernández
Delft, July 2013
Samenvatting

Measuring Robustness of Complex Networks


Een cruciale vraag die opkomt is: “Hoe kunnen we netwerken sterker maken, of robuuster?” De netwerkbeheerder van de TUDelft zou zich bijvoorbeeld af kunnen vragen of zijn netwerk “goed” is, of een ambtenaar zou kunnen willen weten of het Nederlandse spoorwegnet “normaal” werkt. Helaas zijn deze simpele vragen niet duidelijk gesteld omdat “normaal” en “goed” niet gedefinieerd zijn. Sterker nog, tegenstrijdige definities van robuustheid kunnen voortkomen uit verschillend gebruik van hetzelfde netwerk. Vanuit het perspectief van gegevensoverdracht (bijvoorbeeld het openen van een e-mail) is een netwerk robuust als het volledige bericht, eventueel met vertraging, verzonden wordt; zelfs wanneer zich netwerkfouten voordoen. Vanuit het perspectief van een real-time applicatie (bijvoorbeeld een Skype gesprek), daarentegen, is een netwerk pas robuust als...
er data verzonden kan worden, ongeacht netwerk fouten, zonder vertraging. Er van uitgaande dat er een overeenkomst is van wat robuustheid betekent, behandelt dit proefschrift met succes de problemen rond het vinden van een uniek getal (of groep getallen) dat onomstotelijk de robuustheid kwantificeert.

Mijn manier om netwerk robuustheid te kwantificeren werkt als volgt. Omdat robuustheid een mate van kwaliteit van een netwerk uitdrukt, neem ik aan dat het een functie van een aantal netwerk eigenschappen moet zijn. Door een netwerk te representeren als een graaf, dus een verzameling knooppunten en lijnen, kunnen verschillende eigenschappen van netwerken beschreven worden met behulp van lineaire algebra en teruggebracht worden tot kengetallen. Dit zijn typisch rele functies van een graaf model. De eerste helft van dit proefschrift (hoofdstukken 2 en 3) is gewijd aan het algebrasch beschrijven van een netwerk, wat resulteert in een taxonomie van kengetallen. Ieder kengetal kan in de taxonomie gecategoriseerd worden in n van drie klassen: afstand, verbinding en spectrum. Een dergelijke taxonomie helpt toekomstige onderzoekers om hun netwerken beter te begrijpen aan de hand van een allesomvattende en geordende classificatie van hun netwerk beschrijvingen. Vervolgens wordt in hoofdstuk 3 met behulp van drie onafhankelijke kengetallen drie definities van robuustheid geformuleerd. Deze gevalsanalyses tonen aan hoe robuustheid arbitrair gedefinieerd kan worden. Ook onthullen deze hoofdstukken de problemen die zich voordoen, zoals correlaties tussen kengetallen (hoofdstuk 4). Tegelijkertijd bevat dit proefschrift de volgende drie toevoegingen aan het onderzoeksgebied: een nieuw aanvalsbewust beveiligingsbeleid, een beter begrip van quarantaine procedures gebaseerd op modulariteit, en de ontdekking van een nieuwe fase verandering in de context van wederzijds afhankelijke netwerken.

Bovenop de eerder genoemde onderwerpen denk ik dat een definitie van robuustheid niet compleet is zonder de externe of vijandige factoren in het plaatje te betrekken. In andere woorden, robuustheid zou niet gedefinieerd moeten worden als een statische eigenschap van een vast netwerk, maar als het vermogen van een netwerk om zich aan te passen aan negatieve invloeden. Zo heeft men, bijvoorbeeld, recentelijk ontdekt dat de familie van power-law grafen zeer robuust is tegen het willekeurig uitvallen van knooppunten. Daarentegen is een dergelijk netwerk juist zeer gevoelig voor het handmatig verwijderen van zorgvuldig geselecteerde knooppunten. Dit effect maakt het extra duidelijk dat het nodig is om nog een inputparameter toe te voegen aan onze robuustheid functie. Deze laatste parameter noemen we de aanvalscomponent. In de tweede helft van dit proefschrift, in hoofdstukken 4 en 5, worden de externe invloeden op een netwerk behandeld. Hierin wordt de robuustheidsomhullende (robustness envelope) ontwikkelt, een algemene robuustheidsbeschrijving in termen van de aanvalscomponent. Nu het mogelijk is om een netwerk topologie te vertalen naar een uniek getal, is het mogelijk om robuuste netwerken te ontwerpen aan de hand van optimalisatie algoritmen. Ook kunnen robuustheidstests gestandaardiseerd worden zodat netwerken makkelijk in robuustheidshierarchien geclassificeerd kunnen worden.

*Javier Martín Hernández*

*Delft, juli 2013*
## CONTENTS

1 Introduction
   1.1 History of Graph Theory ................................................. 1
   1.2 Modern Times ......................................................... 3
      1.2.1 An Emerging Necessity ........................................... 3
      1.2.2 Previous Work ..................................................... 4
   1.3 Problem Statement .................................................... 6
   1.4 Thesis Outline ....................................................... 7

2 Graph Metrics
   2.1 Introduction to Graph Theory ......................................... 10
   2.2 Taxonomy of topological metrics ..................................... 10
   2.3 Chapter Summary and Conclusions ................................... 34

3 Robustness Measured as a Single Metric
   3.1 Three Approaches to Robustness .................................... 37
   3.2 Measuring Robustness as Hopcount ................................ 38
      3.2.1 Introduction ..................................................... 38
      3.2.2 Related Work ..................................................... 41
      3.2.3 Constructing Challenge-Aware Protection Schemes .......... 42
      3.2.4 Simulation Results .............................................. 45
      3.2.5 Section Summary and Conclusions ............................. 49
   3.3 Measuring Robustness as Modularity ............................... 50
      3.3.1 Introduction ..................................................... 50
      3.3.2 Quarantine Model ............................................... 52
      3.3.3 N-intertwined SIS Model ....................................... 53
      3.3.4 Simulation Results .............................................. 56
      3.3.5 Section Summary and Conclusions ............................. 61
   3.4 Measuring Robustness as Algebraic Connectivity .................. 63
      3.4.1 Introduction ..................................................... 63
      3.4.2 Definitions ...................................................... 65
CHAPTER 1

Introduction

“If I have seen further it is by standing on the shoulders of giants.”

Isaac Newton, 1642 - 1727.

1.1 History of Graph Theory

We see networks all around us: electric power grids, the Internet, transportation highways, underground transportation systems, brain neural networks, etc., all composed of real-life objects. In conjunction, one may also define networks composed of abstract elements, such as networks of acquaintances, protein interaction, or even relations among English words. In fact, we are ourselves, as organic systems, the end result of a network of biological reactions, and as individuals, we are the elements of a network of social relationships. Networks are everywhere.

Leonhard Euler is often credited with conceiving graph theory. In 1736, Euler published the solution to the Königsberg bridge problem, as illustrated in Fig. 1.1. The problem consisted on finding a trip that traversed exactly once, and only one, each of the bridges of the Prussian city. Since then, graph theory has witnessed substantial developments and has provided answers to a series of utilitarian questions, often posed as recreational puzzles like the Knight’s Tour problem. The latter consists of determining the sequence of moves of a knight that visits each square of a chessboard. These 19th century riddles sparked the study of cycles on polyhedra by W. Hamilton, which led to the concept of a Hamiltonian path, i.e. a path that visits each node exactly once. More importantly, the idea of a tree graph (a graph that contains no cycles) was implicitly discovered by G. Kirchhoff, who employed graph theory to understand electrical networks or circuits, leading to the well known Kirchhoff’s circuit laws (as a curiosity, Gustav Kirchhoff was born in Königsberg, Prussia). Around that time, graph theory could already determine the maximum flow per unit time from source to a sink in a network of pipes.
INTRODUCTION

1.1

Figure 1.1: Colored illustration of a Königsberg’s map back in 1726, with seven bridges joining the mainlands. The seven bridges of Königsberg problem consists on finding a walk through the city that crosses each of the seven red bridges once, and only once. If the starting point can be any mainland, can you find such a walk? Assuming that the mainlands cannot be reached by any route other than the bridges, the problem has no solution! The adversity Euler had to solve was not to realize that there is indeed no solution, but the subsequent development of an analysis that stated this assertion with mathematical rigor, i.e. graph theory.

Later in 1852, F. Guthrie posed another memorable problem, which remained unsolved until 1976: the four color riddle, which asks how should we color the regions of a map by using the minimum number of colors, so that neighboring regions receive different colors. Historically, the study of networks has occurred in a branch of mathematics, known as discrete graph theory.

In addition to the developments in mathematical graph theory, the study of networks has also matured in the social sciences context. Social network analysis started to develop in the early 1920s and focuses on the study of properties of individual nodes or groups of nodes. This translates into the real-world as relationships among social entities, communication between members of a group, trades among nations, or economic transactions between corporations. A thrilling result born from social studies is the so called Six Degrees of Separation, discovered in the early 1920s. This conjecture states that every person in this planet is, on average, less than seven steps away from any other person, if they were to be sequentially introduced through the network of personal acquaintances. This theory reveals the surprising small-world effect, which states that even in massive systems composed of billions of elements, two elements can find their way across the network in only six steps.

Later, in 1959, P. Erdős and A. Rényi successfully mixed the concepts of probabilistic methods with graph theory, giving birth to a fruitful branch of graph theory, known as random graph theory. This branch concentrates on the asymptotic behavior of connectivity and percolation in random graphs, which are intimately bound to the concept of robustness. Unexpectedly, interest in networks has seen an astonishing growth in the last decade, mainly due technological advances, in particular the Internet. The latter has
fueled an ever increasing availability of accurate network data sets, leading to the rise of new theories, techniques, algorithms, and models for complex networks, i.e. networks that cannot be understood by the sum of their components’ properties. Thus, the main focus shifted from the analysis of small hundred node networks to that of systems with thousands or even billions of elements. This increase in popularity was probably triggered by two seminal papers, largely inspired by empirical observations. First, that by D. Watts and S. Strogatz on small-world networks, which appeared in Nature in 1998, and second that by Barabási and Albert on scale-free networks, which appeared one year later in Science, breaking traditional assumptions of network theory. These two papers have been certainly product of an escalating computing power, which offers the possibility to study the properties of large databases of real networks.

1.2 Modern Times

1.2.1 An Emerging Necessity

The ever growing theory of complex networks has expanded into many fields of modern science. Some examples of successfully studied networks include:

- transportation networks, where nodes represent cities, and links commercial plane flights among them. These networks are commonly used in epidemiology studies, which model the spread of infectious viruses.

- phone call networks, where nodes represent persons, and links phone calls between them. Anonymized data is often used to detect user behavioral patterns, in order to minimize the rate of dropped calls or to improve the quality of added services.

- the Internet, where nodes represent Autonomous Systems (AS), and links exchanged traffic. A better understanding of the inter-domain topology enables the optimization of routing algorithms. But more importantly, studies of the AS graph led to the discovery of the power-law nature of the Internet topology.

- the World Wide Web, where nodes represent web pages, and links URLs.

- collaboration networks, where nodes represent actors (researchers), and links join two actors that appeared in a movie (scientific publication) together. These type of networks have been the main subject of community detection algorithms, which aim to organize nodes in within clusters with similar preferences.

- protein and metabolic interactions, where nodes represent proteins, and links interactions among them.

- brain networks, where nodes represent areas of the human brain, and links the correlation among clustered neurons. Current research in this field has unveiled the small-world nature of the human brain network, which brings researchers one step closer to understanding the inner workings of our brain and the diseases that affect it.

- online social communities such as Facebook or Twitter, where nodes represent users, and links social bonds. Community detection algorithms are currently used on such
INTRODUCTION

1.2

online networks to refine targeted advertisement algorithms, resulting in revenue increases.

Graph theory has become an essential part of modern science. The application of graph theory to real-world networks has been pushed much further than to cross bridges in Königsberg. To date, network designers search for inspiration in biological graphs in order to build more efficient data networks, stockbrokers use graph theory techniques to predict the trends of the stock market, and even neurophysiologists have realized that the human brain holds small-world properties, which gives important insights on the prevention, treatment, and cure for brain illness, such as Alzheimer’s disease.

As we have just seen, in the last five decades our society has grown more dependent on networks than ever before, in particular the Internet. Every modern society increasingly depends on networks for just about every aspect of daily lives. Consumers use the Internet to access information, obtain products and services, manage finances, and communicate with one another. Companies take advantage of the Internet to conduct business with consumers and other businesses. Even nations rely on the Internet to conduct the affairs of government, deliver services to their citizens, and, to some extent, manage homeland security and conduct military operations.

As the Internet increases its reach in global scope, services traditionally implemented on separate networks are increasingly subsumed by the Internet, either as overlays, gateway access, or replacement for the legacy networks. These include the PSTN (public switched telephone network), SCADA (supervisory control and data acquisition) networks for managing the power grid and other critical infrastructures, and even military networks. In the year 2001, after the World Trade Center incidents, the world shockingly realized that any system, anywhere, can be the subject of an attack. Later, large scale outages like the great Northeastern blackout of 2003, the AS7007 black hole, the Pakistan Telecom-YouTube, or most recent Google incidents confirmed the fact that even small or locally contained causes may have a large scale impact on any infrastructure. In other words, our quality of life, the economic viability of businesses and organizations, and the security of nations are directly linked to the resilience, survivability, and dependability of everyday networks, and in particular the Internet. The US Department of Homeland Security adopted the concept of Critical Infrastructure in 2002, as “systems and assets (whether physical or virtual) so vital to the country that the incapacity or destruction of such systems and assets would have a debilitating impact on national security, economy, public health, safety, or any combination of the previous”. Not surprisingly, networks such as transportation networks, power grids, telephone networks, and the Internet fall under the umbrella of Critical Infrastructures. The world we live in requires networks that resiliently conform and fulfill service level requirements, even under perturbations such as equipment failures, natural disasters and even intentional attacks. In other words, we require robust networks.

1.2.2 Previous Work

Fortunately, we have a body of graph theory to back on. The study of networked infrastructures from a robustness point of view has already been funded for half a century. A wealth of procedures to evaluate and improve network robustness though graph theory has been proposed since the 1940s (arguably, triggered by military purposes). The simplest way to achieve robustness was achieved by duplicating the most vulnerable elements in
Figure 1.2: The percolation problem consists of finding the minimal amount of resources (i.e. the probability \( p \) for any link not to malfunction), such that we can find a walk between two nodes in opposite ends of the boards. The left image illustrates the ideal scenario with infinite resources \( p = 1.0 \), where no link ever malfunctions. Lowering our resources to \( p = 0.5 \) makes the network 50\% cheaper, while still providing side-to-side connectivity. However, if we are too greedy and cut too many resources as in the case of \( p = 0.3 \) (right image), we will never find a walk that crosses the network. Percolation theory accurately predicts the threshold \( p_{th} \), such that the network is always connected for \( p > p_{th} \).

Besides reliability studies, one can also approach networks from a bottom-up perspective as a second means to optimize their robustness. That is, by defining specialized robustness frameworks heavily dependent on the network application purposes. Recent examples of these are the international projects MOTIA, GRID, AMBER, HIDENETS, ResiliNets, and the network of excellence ReSIST. These projects provide important advances in their respective fields (power grids, interdependent ICT systems, computing systems, etc.), proposing techniques and algorithms to improve system evolvability, assessability, usability or diversity. However, the majority of these studies focus on specific systems lacking the generality of a multidisciplinary framework. For instance, GRID studied power systems vulnerabilities, in view of the challenges driven by the transformation of the European power infrastructure, ReSIST led research activities to ensure that...
present and future computing systems, such as SQL databases, have the necessary resilience and survivability. However, the lack of a common vocabulary has made cooperation difficult, as a myriad of terms related to robustness have been proposed over the last fifty years, including reliability, safety, maintainability, dependability and degree-distribution entropy [4, 5]. Meyer [6] studied robustness in the context of his performability framework [7], whilst Cholda et al. [8] surveyed various optical-based robustness frameworks.

More generally, researchers have been studying robustness in the context of varying network types. Due to the behavior of topological metrics depending on the characteristics of the networks to which the metrics are applied, the resulting robustness profiles to also depend on network characteristics. Callaway et al. [9] and Holme et al. [10] have studied the robustness of random networks and power-law graphs. In particular, Cohen et al. have examined the robustness of the Internet and other power-law networks under random [11] and targeted [12] failures. Very recently, the robustness of time-evolving networks or temporal graphs [13, 14] has been researched in [15, 16]. A method based on the cumulative change of the giant component under targeted attacks has been proposed by Schneider et al. [17]. Cetinkaya et al. [18] developed a framework for analyzing packet loss relative to node and link failure. They consider packet loss under global targeted and random failure, as well as attacks contained within geographic regions. Our approach [19] is similar to theirs, although we consider not only average network performance under random attacks but the density function given the probability that a metric will assume a given value after a given fraction of node removals.

1.3 Problem Statement

All previously presented projects provide insights into specific infrastructure failures of real-world networks (e.g. the Sprint or Nobel data networks), with respect to a specialized metric. Such studies provide generated probabilistic risk and performance assessments. However, a deeper understanding into which topological features contribute to the overall resilience of infrastructures is still missing. If developed, such knowledge will enable network designers and operators to unequivocally create, develop and enhance communication infrastructures. This unsatisfactory situation may be attributed to the fact that current work on robustness is driven in an isolated and uncoordinated manner. Moreover, communications networks are inherently complex, due to multi-layered protocol suites, different aggregation levels, lack of missing metrics that adequately define quality of service, dynamic topologies, etc. This complexity illustrates why a framework to compute network robustness is still lacking. The goal of this thesis is to present a framework intended to help address the cited issues. Such a framework would provide a publicly accessible platform that combines and merges previous and ongoing work in the area of resilient networking.

The pooling of available proposals and tools has proved to be a fruitful route to get to an understanding of network robustness. This thesis incorporates an already significant body of previously published work in the areas of structural network analysis, topology generation and network metrics.

The current work brings forth a formal, stochastic framework for quantifying robustness in networked systems. The framework aims at monitoring and studying the behavior of networks, both across a number of different metrics as well as at different layers of the protocol stack and service architecture. Ideally, we wish to express the robustness of a
network as a single metric $R$ between 0 and 1. In this context, $R = 0$ corresponds to absence of network robustness and $R = 1$ reflects perfect robustness. This would enable network designers to measure robustness, benchmark robustness classes, and employ optimization algorithms.

1.4 Thesis Outline

The outline of this thesis can be split into six self-contained chapters, as schematically illustrated in Fig. 1.3.

Figure 1.3: Schematic overview of the present thesis, which is divided in six independent chapters.

- Chapter 1 (current) briefly illustrated the history of graph theory and its evolution over the last 300 years. Specially over the last ten years, societies depend more strongly than ever on networked systems. This dependency has grown to the point to which our well-being depends on these networked critical infrastructures. Bettering our understanding of such complex networks and their robustness is on the bleeding edge of graph theory, of which this thesis is part of.

- Chapter 2 is meant as an introductory chapter to algebraic graph theory. It exposes the necessary background knowledge to understand graph theory, together with an exhaustive survey of existing graph metrics. After examining this chapter, the reader will be familiar with key concepts such as adjacency matrix, graph, and degree distribution. At this point, we will be ready to tackle the concept of robustness.

- Chapter 3 further explores three of the presented metrics: the hopcount, spectra, and modularity. Each of these metrics is presented as a means to measure network robustness, thus illustrating the multi-dimensional nature of robustness. While exploring these three metrics, we additionally unveil novel network properties, such as a phase transition in interdependent networks.
• **Chapter 4** focuses on understanding the existing correlations among graph metrics, in particular we will focus on the betweenness, algebraic connectivity, and degree distribution of power-law graphs. By comparing a large set of distinct graphs with the same degree distribution, we will show that a single graph metric does not provide enough information to characterize a graph. Thus, I argue that graph robustness cannot be characterized by only one metric.

• **Chapter 5** introduces a novel multi-dimensional framework together with some real-world examples. Our main contribution, the R-model, is proposed, fueled by the concept of robustness envelopes. We believe that these tools will power the design and optimization of critical infrastructures, if implemented meticulously.

Finally, Chapter 6 concludes the present thesis by providing conclusive remarks, together with ideas for future work.
PRINCIPLE OF MANAGEMENT: ORDER; A PLACE FOR EVERYTHING AND EVERYTHING IN ITS PLACE (i.e. MATERIALS SHOULD BE IN THE RIGHT PLACE AT THE RIGHT TIME).

HENRI FAYOL, 1841 - 1925

PRIOR to 1732, it was already possible to reach the conclusion that the seven bridges of Königsberg cannot be crossed by a walking salesman, by exhausting all possible combinations. However, a formal proof would require L. Euler’s additional training, hard work, and a pinch of creativity. Similarly, nowadays everyone intuitively understands the concept of “robustness”. For example, we can judge whether a bookcase is robust based on its structural integrity, or we can determine whether an Internet connection is robust based on its responsiveness. However, finding a rigorous formulation of network “robustness” remains an open problem.

The first step towards quantifying a network property (e.g. robustness) is to find a model that accurately represents such networks. Over the last century, the research community accepted graph theory as a representative model of complex networks, and as such it helps us understand surprising network properties otherwise obscured by the network’s intrinsic complexity. This chapter is devoted to the exploration of such graph metrics, by conducting an exhaustive survey of existing metrics. Understanding the taxonomy of graph metrics will bring us one step closer to quantifying robustness.

The first section of this chapter briefly introduces the basics of graph theory. Afterwards, we present a taxonomy of topological metrics together with definitions and practical pointers.
2.1 Introduction to Graph Theory

Graph theory is our framework of choice for the mathematical treatment of complex networks, which can be represented as graphs. In its simplest form, a graph $G(N, L)$ is defined as a set of $N$ nodes interconnected by a set of $L$ links, where a link is defined as a pair of nodes in $N$. We denote the set of nodes by $N$ and its cardinality as $N$. Similarly, the set of links is denoted by $L$ and its cardinality as $L$. A graph example graph $G_1$ with five nodes and six links can be specified in the following way: $G_1(N, L)$, where $N = \{1, 2, 3, 4, 5\}$, and $L = \{\{1, 2\}, \{1, 3\}, \{2, 4\}, \{3, 4\}, \{2, 5\}, \{4, 5\}\}$. As a convention, nodes are usually referred to by their rank $i$ in the set $N$. In addition, two nodes are called adjacent if they are joined by a link.

Resorting to linear algebra in order to represent graphs is a common practice, as algebra provides us with powerful mathematical tools. A useful way to handle a graph $G(N, L)$ is by mapping it into an adjacency matrix $A$, a square matrix with order $N$ where the component $a_{i,j} = 1$ if $\{i, j\} \in L$, and $a_{i,j} = 0$ otherwise. Although such a simple model is already a fairly accurate representation of reality, one may take a step further by introducing the concept of directional links, such that links $(i, j) \neq (j, i)$ represent independent interactions. Alternatively, one may also define a map $L \rightarrow R$, such that each link is attributed a so called weight. Such extended graph models are called directed and weighted, respectively, and are illustrated in Fig. 2.1. Additional graph generalizations can lead to the definition of hypegraphs, where links can connect to any number of nodes (not necessarily two), or even dynamic graphs, where graph elements change properties as a function of time. For simplicity, in this document we will focus exclusively in simple static graphs, i.e. graphs without links connecting a node to itself, nor node pairs connected by more than one link.

Let me briefly illustrate the power of linear algebra with a practical example. Given an adjacency matrix $A$, one can easily prove that the elements of its $k$-th power, $(A^k)_{i,j}$, equal the total number of walks with length $k$ between nodes $i$ and $j$. This theorem also holds for $k = 1$: $(A^1)_{i,j}$ equals 1 if there exists a walk with length 1 between node $i$ and node $j$, i.e. $(A^1)_{i,j}$ is 1 if there exists a link between nodes $i$ and node $j$. This matrix multiplication trick is commonly used to evaluate how clustered a network is, by counting the number of closed walks with length $k = 3$ (also known as triangles). As we will show further in this section, computer networks tend to display higher clustering than randomly generated networks, thus rendering $A^3$ a powerful discriminator.

2.2 Taxonomy of topological metrics

We have shown that a graph can be expressed as a square adjacency matrix $A$, consisting of elements $a_{ij}$, normalized to the range $[0, 1]$. Based on this definition, we define as a topological metric any scalar (or vector of scalars) that can be computed by using exclusively the adjacency matrix $A$, such as the ones in Fig. 2.1. If any additional node, link, or hypergraph property is assumed in the calculations, the resultant metric is not a topological metric, but a service metric instead. The latter set of metrics will not be presented in this thesis, due to the excessive space it would take to cover. For the sake of brevity, this thesis focuses only on undirected, unweighted graphs, but most metric definitions can be easily generalized to directed and/or weighted graphs.
Given an undirected graph $G$ with $N$ nodes, we can build a space $S(N)$ containing $|S(N)| \approx 2^{N^2}$ different graphs. Additionally constraining the number of links to $L$ reduces the space to $|S(N, L)| \approx N^{2L}$ graphs. In order to split this graph space into classes of graphs with similar features, first we must be able to uniquely define graph features with a set of metrics. Experience tells us that a small set of topological measures provides enough information to characterize any given graph. Yet the thousand dollar question remains unanswered: how many metrics are needed?

Before delving into the metrics themselves, I would like to warn the reader that this classification is a living document, bound to change over time. Due to existing correlations among metrics, it is often the case that one metric can be interpreted in a variety of ways, all of them legitimately valid. Thus different scientists -with subjective insights- are bound to hold conflicting perspectives on the same metric. This thesis does not aim to provide the definitive classification, but instead an extensive survey, written as an introductory guide to graph theory. A variety of metric classifications have been proposed based on different criteria, based on:

- The nature of the variables used to compute the metric, i.e. the number of neighbors a random node has, or the distance between two randomly chosen nodes.
• The correlation between metrics. This criteria groups strongly correlated metrics into the same class. Given that correlations strongly depend on the graph under study, this classification may prove to be difficult.

• The local vs. global nature of the metric. This classification splits metrics in two subclasses: local metrics which can be computed by solely observing a node’s neighborhood up to \( h \) hops away, and global metrics which can only be computed when the entire graph is accessible.

• The computational complexity involved in computing the metric. Metrics may be classified based on the rate of the number of operations required to compute its value for a defined set of graphs. The complexity is usually a polynomial function of the number of nodes \( N \) and links \( L \) present in the network.

Depending on what criteria we choose, we have to strike a balance between a human friendly classification and a clear yet less intuitive classification. For example, the fact that a metric can be computed in \( O(N^2 \log N) \) seconds gives us little information (4th presented criteria) about it. On the other hand if we know that the metric was computed by using shortest paths, we can better identify its meaning and associations. The authors of this document opted for the first, human friendly classification. However all the presented criteria are equally valid and should be taken into consideration for future taxonomies. Alternative taxonomies are proposed by Boccaletti et al. [20] and DaCosta et al. [21].

We propose that topological metrics can be further split into three subclasses, these are: distance, connection, and spectra. The distance class gathers metrics making use of the hopcount random variable, which provides information about the number of nodes a message has to cross on its way to its destination. The connection class groups metrics related to the nodal degree random variable (i.e. the number of adjacent nodes), together with metrics that help grouping nodes into clusters or hierarchies, thus providing insights into the structure of the network. Finally, the spectra class includes metrics directly related to the eigenvalues and eigenvectors of a graph.

2.2.1 Distance Class

In communication networks, paths are basic entities that connect two communicating parties (i.e. nodes) in a graph \( G(N, L) \). We define a valid path \( P_{i \rightarrow j} \) between two nodes \( i \) and \( j \) as an ordered sequence of links, such that (a) two consecutive links are incident to the same node, and (b) no links are repeated. For example, in Figure 2.1, a valid path between nodes 1 and 4 is \( P_{1 \rightarrow 4} = \{(1,2),(2,5),(5,4)\} \), which can be alternatively expressed as a sequence of nodes \( \{1,2,5,4\} \). The cardinality of the link set is called the hopcount of the path. For the given example, the hopcount equals \( |P_{1 \rightarrow 4}| = 3 \).

**Hopcount**

Given a graph \( G \), the shortest hopcount \( H_{i \rightarrow j} \) between nodes \( i \) and \( j \) is the number of hops (i.e. links) in the shortest path connecting \( i \) and \( j \),

\[
H_{i \rightarrow j} = \min_{P_{i \rightarrow j} \in S_{P_{i \rightarrow j}}} (|P_{i \rightarrow j}|) \tag{2.1}
\]
2.2 TAXONOMY OF TOPOLOGICAL METRICS

where \( S_{P_{i \rightarrow j}} = \{P_{i \rightarrow j}^1, P_{i \rightarrow j}^2, \ldots P_{i \rightarrow j}^k\} \) is the set of all valid paths between nodes \( i \) and \( j \). For example, assume that \( P_{i \rightarrow j}^1 \) is a shortest path \( P_{i \rightarrow j}^1 = \{i \rightarrow n_2 \rightarrow n_3 \rightarrow ... \rightarrow j\} \), which contains \( k + 1 \) elements. Then the (shortest) hopcount between nodes \( i \) and \( j \) is the cardinality of the link sequence \( H_{i \rightarrow j} = |P_{i \rightarrow j}^1| \). As a convention, \( H_{i \rightarrow j} = \infty \) when there is no path exists between nodes \( i \) and \( j \). Additionally, the hopcount distribution \( \Pr[H = k] \) is the probability density function of the random variable \( H \), i.e. the probability a random pair of nodes to be \( k \) hops away \[22\], as illustrated in Fig 2.2. It can be expressed as

\[
\Pr[H = h] = \frac{\sum_{i \in N} \sum_{j \in N \setminus i} 1\{H_{i \rightarrow j} = h\}}{N(N - 1)} \tag{2.2}
\]

where \( 1_X \) is the indicator function, such that \( 1_X = 1 \) if the condition \( X \) is true, and \( 1_X = 0 \) otherwise. For example, given that \( \Pr[H = 2] = \frac{1}{2} \), then we have a 50% chance to reach a random node in 2 hops.

The hopcount of a path is often associated in physics to the distance, length, or geodesic of such path \[23\] \[24\]. The distance (or length) of a path \( \{i \rightarrow n_2 \rightarrow n_3 \rightarrow ... \rightarrow j\} \) is the sum of all the link weights that constitute such path. When all the link weights in the graph have link weight \( w_l = 1 \), hopcount and distance become equivalent terms.

We are now in position to rigorously define the Six Degrees of Separation experiment mentioned in Chapter 1.1 in the following way. The Six Degrees of Separation study claims that, in a network of human acquaintances, the number of hops between two random nodes \( (i, j) \) scales logarithmically with the total number of nodes in the network \( H_{i \rightarrow j} = \mathcal{O} = \log N \). Or, in words, the number of acquaintances we have to sequentially contact in order to reach a random node decreases exponentially with \( N \).

The hopcount distribution of a graph is crucial for the development of network applications. For example, the performance parameters of routing algorithms strongly depend on the hopcount distribution. Research suggests that paths should generally be as short as possible \[25\], in order to avoid stacking delays while minimizing points of failure. However, besides routing, hopcount also plays a vital role in robustness of the network to worms. Omic \textit{et al.} \[26\] showed that worms can quickly contaminate a network that has small distances between nodes. This result promotes the use of longer paths over short paths, which is in contradiction with the routing guidelines.

In random graph theory, a topology is assumed to be either completely regular or completely random. But many biological, technological and social networks lie somewhere between these two extremes. Systems can be highly clustered, like regular lattices, yet have small average hopcount \[27\], like random graphs. These are called small-world networks, by analogy with the small-world phenomenon \[28\]. The average hopcount distribution is an important tool to distinguish between small world and large diameter graphs, like lattices \[22\].

**Efficiency**

Efficiency \( E \) is an alternative measure of distance, introduced by Latora \textit{et al.} \[29\]. The efficiency of a graph is defined as the sum of the inverted shortest hopcounts, over all possible node pairs. Efficiency can be expressed as,
Figure 2.2: Hopcount probability density function of four graphs with $N = 1,000$ nodes and $L = 3,000$ links. The four graphs represent four different graph models, which mimic the properties of real-world networks (the models are Barabási-Albert, Erdős-Rényi, Watts-Strogatz, and Random Regular). Even in the presence of thousands of nodes, one may reach two random nodes in a very little number of hops, i.e. less than $\approx \log(N)$ hops. This counter-intuitive phenomena is known in the scientific community as small-world, or more informally, as the Six Degrees of Separation.

\[
E = \frac{1}{N(N-1)} \sum_{i \in N} \sum_{j \in N \setminus i} \frac{1}{H_{i \rightarrow j}} \tag{2.3}
\]

The efficiency ranges between $[0, 1]$, reaching its maximum for a fully connected graph, and its minimum for a disconnected graph (where all hopcounts are infinite).

Intuitively, the graph efficiency $E$ measures how many links are utilized to propagate information through a graph, i.e. the more links required to spread information in $G$, the less efficient $G$ is. The efficiency of an isolated node has also been studied as a centrality measure; Latora et al. [30] introduced a centrality metric based on efficiency for each node $i$ in a graph.

**Resistance**

The graph resistance $R$ (also called Kirchhoff index $Kf$) of a connected graph originates from electric circuit analysis [31]. Given a graph, we can treat each edge as a resistor, and compute the resistance of a graph as the sum over all effective resistances over all node pairs,

\[
R = \frac{1}{2} \sum_{i \in N} \sum_{j \in N} \Omega_{ij} \tag{2.4}
\]

where $\Omega_{ij}$ is the effective resistance between terminals $i$ and $j$. Recall Ohm’s law $V = IR$, i.e. voltage equals intensity times resistance between two points, hence effective resistance
equals the the voltage drop in \((i, j)\), over the flowing intensity. If the graph is unweighted, one generally assumes that all links have resistance 1Ω. If an edge has weight \(w_{i,j}\)Ω we will assign the corresponding link resistance \(\frac{1}{w_{i,j}}\)Ω. The reason for this is that we generally associate link weights as being inversely proportional to (intensity) flow. For example, having no link corresponds to having a resistor of infinite resistance.

The resistance can also be elegantly expressed in terms of the non-zero eigenvalues \(\mu_k\) of the Laplacian matrix \(Q\) as follows

\[
R = N \sum_{k=1}^{N-1} \frac{1}{\mu_k}
\]  

(2.5)

In practice, the resistance of a graph accounts for both the path diversity, and the length of the paths. In practice, it measures the quantity of back-up paths available [32]. As opposed to the efficiency \(E_G\), the resistance does not limit itself to the shortest paths, because intensity flows across all links in a resistive circuit.

**Communicability**

The communicability of a node pair \((i, j)\) is defined as the weighted sum

\[
C_{ij} = \left( \sum_{k=0}^{\infty} \frac{A^k}{k!} \right)_{ij}
\]

(2.6)

Intuitively, this metric accounts for all the possible routes connecting two nodes in a network, which provides information about how well communicated the two nodes are. Similarly to the graph resistance, the communicability [33] is not limited to the shortest paths, like the hopcount is. Instead, the communicability accounts for all possible paths between node pairs, assigning highest weights to the shortest ones. Resistance and communicability differ in the way they weight the different paths.

If the sum of all communicability distances in a network is significantly larger than the similar sum of shortest paths distances, the network shows high cliquishness [33]. On the other hand, we can find networks where the sum of all communicability distances is smaller than the sum of all shortest paths distances. These “unpacked” networks can be easily embedded into two-dimensional space.

**Closeness**

The closeness [34] of a given node \(i\) is the average hopcount obtained from node \(i\) to all the other nodes. It is defined as the reciprocal of the hopcount,

\[
C_i = \frac{1}{\sum_{j \in \mathcal{N} \setminus \{i\}} H_{i \rightarrow j}}
\]

(2.7)

The reciprocal of the node closeness is also known as the Wiener index \(W_i\) [35].

Closeness is often regarded as a participation metric, i.e. a measure to quantify the node centrality within a network. In data networks, nodes with low closeness scores tend to be close to other nodes, thus they disseminate information faster. Closeness has been used in biology to identify central metabolites in metabolic networks.
**Eccentricity, Diameter, Radius**

The eccentricity $\varepsilon_i$ of a node $i$ is defined as the longest hopcount between the node $i$ and any other node in $G$.

$$\varepsilon_i = \max_{j \in N} (H_{i \rightarrow j})$$  \hspace{1cm} (2.8)

The eccentricity of a graph $\varepsilon$ is the average eccentricity over all the nodes in $G$. It is closely related to the flooding time [22], which is the minimum time needed to inform the last node in a network. Intuitively nodes that play an important role in a topology should be easily reachable by the rest of the nodes in a graph. The diameter $D$ of a graph $G$ is the maximum node eccentricity over all the nodes in $G$.

$$D = \max_{i \in N} (\varepsilon_i)$$  \hspace{1cm} (2.9)

The diameter [36] can also be regarded as the longest shortest hopcount found in a graph. This measure gives an indication on how extended a graph is. The radius $R$ of a graph is the minimum node eccentricity over all the nodes in $G$.

$$R = \min_{i \in N} (\varepsilon_i)$$  \hspace{1cm} (2.10)

These three metrics measure worst-case scenarios, thus proving to be useful at computing upper bounds for graph properties such as eigenvalues [37]. As a last remark, notice that the diameter can be artificially inflated by long chains of nodes.

**Persistence**

The persistence of a graph of diameter $D$, as introduced by Boesch et al., is the minimum over all pairs of non-adjacent nodes of the maximum number of disjoint paths of length at most $D$ joining them. In words, it is the smallest number of links whose removal either increases the diameter, or disconnects the graph.

**Girth**

The girth $\gamma$ of a graph [38] (also known as chordality) is the hopcount of the shortest cycle contained in the graph. A cycle is a closed path $P_{i \rightarrow i}$, with no other repeated nodes than the starting and ending nodes. The girth of an acyclic graph, such as a tree, is defined to be infinite.

This measure has a limited use, as any graph with clustering coefficient larger than 0 will provide $\gamma = 3$. However, it can happen that routing algorithms may induce operation errors, which lead to data packets being endlessly routed in a closed loop. For this reason, graphs with high girth values are less prone to suffer from endless loops. Regardless, link-state routing protocols (e.g. OSPF) prevent self loops after a flooding, additionally distance-vector routing protocols (e.g. BGP) have built-in loop prevention algorithms.

**Expansion**

The expansion $e_h$ of a graph [22] is the average fraction of nodes in the graph that fall within a ball of radius $h$ (in hops) centered at a random node $i$. 
2.2 TAXONOMY OF TOPOLOGICAL METRICS

Figure 2.3: Contribution of a single node pair \((s, t)\) to the link betweenness of every link. There exist a total of three different paths between node \(s\) and \(t\): \(S_{P_{s \rightarrow t}} = \{\{a, b, c\}, \{d, b, c\}, \{e, f, c\}\}\). Freeman’s betweenness equally weights each path (upper, middle and bottom paths) by \(1/3\), contributing equally to every link that lies in any shortest path.

Betweenness

The betweenness of a node (or a link) \(B_k\) is defined as the total number of shortest paths that traverse a node (or link) \(k\),

\[
B_k = \sum_{i \in N} \sum_{j \in N} 1\{k \in P^*_{i \rightarrow j}\} 
\]

(2.12)

It is generally assumed the use of the shortest path length find the set \(P(i, j)\) in (2.12). However, in some real scenarios, routing protocols are subject to multiple service constraints, causing some datagrams to be forwarded through non-shortest paths. Thus alternative network-flow methods have been introduced [24] to compute the betweenness based in random walks. Consider the number of times a random traveling message passes through \(k\) along his journey averaged over a large number of trials. The full random walk betweenness of a node (or link) \(k\) will then be this value averaged over all possible source/target pairs \(i,j\).

Dozens of additional variants of betweenness have been introduced over the last 30 years [39] [40], in order to fit the metric to specific network applications. However, the predominant definition is the one suggested by Freeman [41] in 1977, which studies all possible paths running between a single node pair and weights them proportionally, as follows,
\[ B_k = \sum_{i \in N} \sum_{j \in N} \frac{1}{|S_{P_i \rightarrow j}|} \sum_{z \in S_{P_i \rightarrow j}} 1_{\{k \in z\}} \]  \hspace{1cm} (2.13)

In words, the contribution of a node pair \(i, j\) to link \(k\)'s betweenness is the proportion of shortest paths that crosses link \(k\) over the total number of existing shortest paths \(|S_{P_i \rightarrow j}|\), as illustrated in Fig. 2.3.

Betweenness has been heavily focused in the past as a measure of the centrality (or influence) of nodes in social networks. First, proposed by Freeman [41], \(B_k\) measures the influence of a node (or link) over the global flow of information. In communication networks, betweenness measures the potential amount of traffic that crosses a network element. This potential traffic will be affected when the node/link’s fails. Betweenness can also be extended to quantify the importance of a group of nodes. In addition, Huijuan et al. [42] show that in overlay trees of real world complex networks with exponential link weight distribution, the probability distribution function of \(B_k\) follows a power law \(\Pr[\text{B}_k = j] = c_0 j^{-c}\).

**Central Point of Dominance**

The central point of dominance [41] is defined as

\[ \text{CPD} = \frac{1}{N - 1} \sum_{i} (B_{\text{max}} - B_i) \]  \hspace{1cm} (2.14)

where \(B_{\text{max}} = \max_{k \in N} B_k\) is the largest value of betweenness centrality in the network. CPD is a measure of the maximum betweenness of any point in the graph: it equals 0 for complete graphs and 1 for star graphs (in which there is a central node that all paths include).

**Distortion**

Consider any spanning tree \(T\) on a graph \(G(N, L)\), and define \(H_{i,j}(T)\) as the shortest distance between nodes \(i, j\) in the graph \(T\). The distortion is defined as

\[ \text{dist} = \min_{T \in T^*} \left( \frac{1}{L} \sum_{i \in L} H_{i,j}(T) \right) \]  \hspace{1cm} (2.15)

where \(T^*\) is the set of all possible spanning trees of \(G\), and \(L\) is the set of links of graph \(G\). Distortion measures how much \(T\) distorts links in \(G\), i.e. it measures how many extra hops are required to go from one side of a link in \(G\) to the other, when restricted to using \(T\). The distortion is defined [43] to be the smallest such average over all possible trees \(T^*\).

Intuitively distortion measures how tree-like a graph is. The closest the distortion is to 1, the more tree-like a graph is.

**2.2.2 Connection Class**

One of the major concerns of network analysis is the identification of cohesive subgroups of actors within a network. However, a cohesive subgroup is an arbitrarily defined term,
which can be interpreted as a subset of nodes among whom there are relatively strong, direct, intense, frequent, or positive ties. In this document, we will stick to the following cohesion axiom: two nodes joined by a link behave more cohesively than two nodes not joined by a link.

In order to rigorously define cohesion, we introduce the degree \( d_i \) of a node \( i \), as the number of other nodes to which \( i \) is connected to,

\[
d_i = \sum_{j=1}^{N} a_{ij}
\]  

(2.16)

The nodal degree is a powerful tool to define and spot cohesive subgroups in a network. Often, nodes belonging to the same group are expected to share structural properties, which can be exploited to our advantage. For example, persons "Liking" the same Facebook page, or or P2P clients exchanging gigabytes of data are likely to share interests [44].

**Degree**

Let \( d(k) \) be the number of nodes with degree \( k \) of a given graph \( G \). The *node degree distribution* is the probability that a randomly selected node in \( G \) has degree \( k \),

\[
\Pr[D = k] = \frac{d(k)}{N}
\]  

(2.17)

where \( D \) is the random variable representing degree of a randomly chosen node. The average value of this distribution is called the *average degree*\(^1\) \( E[D] \), and obeys the basic law,

\[
E[D] = \frac{2L}{N} = k \Pr[d_X = k]
\]  

(2.18)

where the minimum and maximum node degree are denoted as \( d_{\text{min}} \) and \( d_{\text{max}} \), respectively. The degree distribution of *randomly built* graphs follows a binomial distribution [22]. However, some recent empirical results [36] show that the degree distribution of some real-world networks significantly deviates from a binomial distribution. In particular, some large networks such as the World-Wide Web [36], Internet AS [45] level or metabolic networks [46] [47] [48], show a skewed degree distribution, with a power-law tail as illustrated in Fig. 2.4.

It is well known that the topology of a network has a major impact on the performance of network protocols. For this reason, network researchers often use topology generators to generate realistic graphs for their simulations. These topology generators attempt to create network topologies that capture the fundamental characteristics of real networks, being the degree distribution one of the simplest metrics to mimic [43]. Chapter 4 extends on the concept of power-law generators, as a stepping stone to define network robustness.

As we will see in Chapters 3 and 5, robustness can be measured in several ways, but one of the most common indicators of network resilience is the variation on the fraction of nodes in the largest connected component upon link removals. In the context

\(^{1}E[D]\) it is often represented as \( \bar{k} \) in physics.
of communication networks the nodes in the giant component can communicate with an extensive fraction of the entire network, whereas nodes in the small components can only communicate with a few others. Studies performed on the Internet AS topology [12] [49] show that networks with power law degree distributions are relatively robust with respect to a random failures only a failure of central nodes is likely to cause the network to fragment. On the other hand this type of hub-based networks is extremely vulnerable to a targeted attack an attack in which the most highly connected nodes are removed first. These results lead to a popular graph feature, known as \textit{robust yet-fragile}.

\textbf{Joint Degree (JDD)}

The joint degree distribution (JDD) is the probability that a random selected link connects $d_1$ and $d_2$-degree nodes [46] [47]. Let $m(k_1, k_2)$ be the total number of links connecting nodes of degrees $d_1$ and $d_2$,

$$\text{Pr}[X = k_1, Y = k_2] = (2 - 1_{k_1 = k_2}) \frac{m(k_1, k_2)}{2L}$$

Note that $\text{Pr}[X = k_1, Y = k_2]$ is different from the conditional probability $\text{Pr}[X = k_1 | Y = k_2]$ that a given $k_1$-degree node is connected to a $k_2$-degree node.

While the node degree distribution tells us how many nodes of a given degree are found in a network, the JDD provides information on the interconnection between these nodes, by describing correlations of degrees of nodes located at distance 1. Given a specific form of $\text{Pr}[X = k_1, Y = k_2]$ we can always restore both the degree distribution $\text{Pr}[d_X = k]$, and average degree $E[d]$.
Assortativity

A direct way to determine the degree correlations is by considering the Pearson correlation coefficient \([50]\) of the degrees at either ends of each link. This normalized value is called the assortativity coefficient \(r\) of a graph \(G\) [51], which summarizes the joint degree distribution (JDD) as a single scalar,

\[
r = \frac{\text{Cov}[X,Y]}{\sigma_X \sigma_Y} = \frac{1}{\sigma_X \sigma_Y} \sum_{(j,k) \in \mathcal{L}} jk \left( \Pr[X = j, Y = k] - \Pr[X = k] \Pr[Y = j] \right) / \mathbb{E}[d]^2 = 
\]

\[
= \frac{1}{L} \sum_{(i,j) \in \mathcal{L}} d_i d_j - \left( \sum_{(i,j) \in \mathcal{L}} \frac{1}{L} (d_i + d_j) \right)^2 
\]

\[
= \frac{1}{L} \sum_{(i,j) \in \mathcal{L}} \frac{1}{2} (d_i^2 + d_j^2) - \left( \sum_{(i,j) \in \mathcal{L}} \frac{1}{L} (d_i + d_j) \right)^2
\]

where \(X\) and \(Y\) are the degrees of the nodes reached by following a randomly chosen link on the graph. Van Mieghem et al [52] have reformulated assortativity as a function of graph walks,

\[
r = \frac{N_1 N_3 - N_2^2}{N_1 \sum_{i=1}^{N} d_i^3 - N_2^2}
\]

where \(N_k = u^T A^k u\) is the total number of walks with \(k\) hops.

The assortativity coefficient lies in the range \([-1, 1]\). Assortative mixing \((r > 0)\) is defined as a preference for high-degree nodes to attach to other high-degree nodes, whereas disassortative mixing \((r < 0)\) as the converse, where high-degree nodes attach to low-degree ones. Assortative and disassortative mixing patterns indicate a generic tendency to connect to similar or dissimilar peers respectively.

Highly connected nodes tend to be connected with other high degree nodes in social networks [51]. On the other hand, technological and biological networks typically show disassortative mixing, as high degree nodes tend to attach to low degree nodes. Functional brain networks determined from EEG analysis have also been found to be assortative [53].

Degree diversity

The degree diversity is defined as [54]

\[
\delta = \frac{\mathbb{E}[D]^2}{\mathbb{E}[D]}
\]

Chung et al [55] found that the degree diversity approximates the largest eigenvalue of the adjacency matrix for Erdős-Rényi random graphs if \(\delta > \sqrt{d_{\text{max}} \log N}\). Scale-free graphs where \(\text{Var}[D] \rightarrow \infty\) as \(N \rightarrow \infty\), have \(\delta \rightarrow \infty\). Regular networks where \(\text{Var}[D] = 0\) have \(\delta = 0\).

Several properties of dynamic processes on networks, such as synchronization threshold in mean-field theory of coupled oscillators [56], the network percolation and epidemic thresholds [57] are related to the degree diversity of a graph.
Subgraph centrality

The subgraph centrality of a node $S(i)$ is a centrality measure that characterizes the local cliquishness around nodes [58]. It is defined as

$$S_i = \sum_{k=0}^{\infty} \frac{W_{i\rightarrow i}(k)}{k!}$$

where $W_{i\rightarrow i}(k) = (A^k)_{ii}$ is the set of closed walks of length $k$ starting and ending at node $i$. The sum of the subgraph centralities of all nodes in the graph is known as the Estrada index

$$S = \sum_{i=1}^{N} S_i$$

Estrada et al. [58] proved that $S$ can be obtained mathematically from the spectra of the adjacency matrix of the network through the following expression

$$S = \sum_{i=1}^{N} e^{\lambda_i} = \text{tr} (e^A)$$

Because smaller subgraphs are given more weight than larger ones, this measure is appropriate for characterizing network motifs. This rule is based on the observation that motifs in real-world networks tend to appear as small subgraphs. Additional mathematical results can be found in the following references [59][60].

Modularity

Modularity is a quality function $Q$ that quantitatively defines the quality of a node partition\(^2\). First defined by Newman et al. [61], it is based on the axiom that random networks do not exhibit community structure.

Let us define a graph partition of $\mathcal{N}$ into $c$ non-intersecting communities. In addition, define an $E$ matrix, where the elements $e_{ij}$ represent the fraction of links in $\mathcal{L}$ starting at a node in partition $i$ and ending at a node in partition $j$. By definition, random networks do not exhibit community structure. Thus if we generate a random graph, and we partition it into $c$ random communities, then the expected value of links within partitions is expected to be also random. In this community-less scenario, the expected fraction of links within community $i$ equals $\sum_{j=1}^{N} e_{ij} \cdot \sum_{j=1}^{N} e_{ij}$. Given network $G$ and a given partition, the modularity measures the difference between the existing number of links and the expected number of links within communities, expressed as

$$Q = \sum_{i=1}^{c} \left( e_{ii} - \left( \sum_{j=1}^{c} e_{ij} \right)^2 \right)$$

(2.24)

where $Q$ lies in the range $[-1, 1 - 1/c]$, for a disconnected network and for perfectly modular networks, respectively.

\(^2\)Not to be confused with the Laplacian matrix $Q$
2.2 TAXONOMY OF TOPOLOGICAL METRICS

Figure 2.5: This image exemplifies community detection by means of modularity optimization. If we group the nodes in the original graph (left image) into three arbitrary chosen communities (middle image), then we expect the modularity score of such partition to be low $Q = 0.27$. The node partition for which modularity is maximized corresponds to the perceived node partition (right image), for which modularity reaches its maximum $Q = 0.57$.

Even though modularity is a quality function designed to quantify the goodness of a partition, its most extensive use has been community detection. By maximizing the value of $Q$, one can find the graph partition that least resembles a random network, i.e. the partition with the highest community score [62] as illustrated in Fig. 2.5. Modularity has been found to be a good indicator of functional network divisions in many cases [63][64]. In particular, sociology studies may detect behavioral patterns among large groups of humans by identifying communities within them.

**Leverage centrality**

Leverage centrality [65] measures the relation between the degree of a given node and the degree of each of its neighbors, averaged over all neighbors,

$$L_i = \frac{1}{d_i} \sum_{k=1}^{N} a_{ik} \frac{d_i - d_k}{d_i + d_k} = \frac{1}{d_i} \sum_{k \in C_i(h)} \frac{d_i - d_k}{d_i + d_k}$$

A node with negative leverage centrality is considered to be influenced by its neighbors, as the neighbors connect and interact with far more nodes. A node with positive leverage centrality, on the other hand, influences its neighbors since the neighbors tend to have fewer connections.

Leverage centrality determines the connectivity of a node as relative to the connectivity of its neighbors. Although similar in concept, there are essential differences between eigenvector and leverage centrality that are discussed in [65].

**Coreness**

The $k - core$ is the subgraph obtained from the original graph by the recursive removal of all nodes of degree less than or equal than $k$ [66][67]. Hence, in a $k - core$ subgraph all nodes have at least degree $k$ as illustrated in Figure 2.6.
The node coreness $k_i$ of a given node $n_i$ is the maximum $k$ such that this node is present in the $k$–core graph, but removed from the $(k+1)$–core. This measure can be regarded as an indicator of node centrality, since it measures how deep within the network a node is located.

**Cliques and n-cliques**

A *clique* [67] of a given graph $G(N,L)$ is a subset of nodes such that all elements in the clique $S(N_S, L_S)$, where $N_S \leq N$, $L_S \leq L$, are fully connected, hence, forming a full mesh. The clique number of a graph is the number of nodes belonging to the largest clique $S$,

$$S : \max \left| \bigcap_{i \in S} e_i(1) \right|$$  \hspace{1cm} (2.25)

where $e_i(h)$ is the set of nodes that can be reached in $h$ hops from node $i$.

The definition of a clique is, however, too restrictive for practical purposes. Due to most real-world networks being sparse, it is relatively hard to find subsets of nodes fully connected between them. We may relax the definition by defining an *n-clique*. An *n-clique* $S'$ of a graph is a (maximal) set of nodes for which for all its elements $u,v \in S'$, are at a distance $H_{u \rightarrow v} \leq n$. In other words, an *n-clique* is a set of nodes in which every node can reach every other node in $n$ or fewer steps. The set $S'$ is maximal in the sense that no other node in the graph is at distance $n$ or less from every other node in the subgraph.

By definition, a *1-clique* and a *clique* are equivalent terms.

The knowledge about subgraphs with clique features within a network can decrease the complexity of algorithms designed for such network. For example, if a network is composed of trees hanging off a densely connected component, then an algorithm can run in the center component and a second algorithm tailored specifically for trees can run in the trees.
2.2 TAXONOMY OF TOPOLOGICAL METRICS

Figure 2.7: This figure illustrates two networks containing the same number of nodes and links ($N = 20$, $L = 40$), yet severely different clustering coefficients. The clustering coefficient of a random graph and a circular lattice are 0.09 and 0.50, as shown in the left and right images, respectively. In words, when a small number of $L$ links are placed at random, only 9% of a randomly picked node’s neighbors are connected. However, when the same number of $L$ links are carefully laid out in a lattice, up to 50% of any node’s neighbors are connected, a substantial ∼550% increase.

Clustering coefficient

The local clustering coefficient of a node $i$ in a graph $G$ measures the cliquishness its neighborhood

$$c_i = \frac{1}{d_i} \sum_{j=1}^{N} 1\{a_{ji}=1\} \binom{d_i}{2}$$  \hspace{1cm} (2.26)

where the numerator counts the number of links between neighbors of $i$.

The clustering coefficient $C$ of the whole graph $G$ is the average of the local clustering coefficients for all the nodes in $G$,

$$C = \frac{1}{N} \sum_{i \in \mathcal{N}} c_i$$  \hspace{1cm} (2.27)

Alternatively we can also express the clustering coefficient as the number of walks with 3 hops over the total number of triples that are connected minus the walks with 2 hops,

$$C = \frac{\text{Trace}(A^3)}{u^2 Au - 2L}$$  \hspace{1cm} (2.28)
where the denominator represents the number of links in the line graph \( l(G) \) of \( G \) [68]. This number is precisely the probability that two random neighbors of a random node are neighbors themselves. An alternative and less often used definition of clustering coefficient is the following

\[
C_\Delta = \frac{3\Delta}{N_A}
\]  

(2.29)

where \( \Delta \) is the number of triangles in the graph, and \( N_A \) is the number of connected triples. \( C_\Delta \) differs from \( C \), in that \( C_\Delta \) is the probability that a triangle is formed upon each triple in the network, as \( C \) is the average of the connection density over all the neighbors of each node.

The clustering coefficient can be interpreted as a measure of how close a node’s neighbors are to forming a 1-clique. The clustering coefficient has been extensively used in network topology studies, since it is a low complexity cliquishness indicator. It reaches its maximum 1 on a fully connected graph (everyone knows everyone else) and has typical values in the range of [0.1, 0.5] in many real-world networks [46] [22].

In random Erdős-Rényi graphs, the probability of loops involving a small number of vertices goes to 0 in the large network size limit [22], hence \( C \) tends to be really small. This effect presents a remarkable contrast with the abundance of short loops observed in many real world networks [69] [43].

**Cyclic coefficient**

The cyclic coefficient \( \Theta_i \) of a node \( i \) in a graph \( G \) measures the average length of the cycles crossing node \( i \). It is defined as

\[
\Theta_i = \frac{2}{d_i(d_i - 1)} \sum_{j,k} \frac{1}{S_{ijk}} a_{ij} a_{ik}
\]  

(2.30)

where \( S_{ijk} \) is the number of shortest paths that cross the node set \( \{i, j, k\} \), i.e. \( i \) and its two neighbor nodes \( j \) and \( k \). Note that if nodes \( j \) and \( k \) are connected, the smallest cycle is a triangle and \( S_{ijk} = 3 \). On the other hand, if there is no path crossing the three nodes, then these nodes are tree-connected and \( S_{ijk} = \infty \).

The cyclic coefficient of a network \( \Theta \) is defined as the average of the cyclic coefficient of all its vertices

\[
\Theta = \sum_{i \in \mathcal{N}} \Theta_i
\]  

(2.31)

Kim et al. [70] show that \( \Theta \) can be used to understand the structure of complex networks. For example, the larger the cyclic coefficient is, the less cyclic the network becomes. The cyclic coefficient ranges between 0 and 1/3, for tree graphs and fully connected graphs, respectively.

**Rich Club coefficient**

Given a graph \( G \), define \( \mathcal{S}_k \) as the subset of nodes with degree greater than \( k \), \( \mathcal{S}_k : \{ n \in \mathcal{N} | d_n > k \} \). We call this subset \( \mathcal{S}_k \) the \( k \)-club members. The rich-club coefficient \( \Phi_k \) is defined as the ratio of the number of links connecting the club members over the
maximum number of allowable links in $D_k$, which measures how well the rich nodes know each other,

$$\Phi_k = \frac{1}{|S_k|(|S_k| - 1)} \sum_{i,j \in S_k} a_{ij}$$

(2.32)

The rich club phenomenon in complex networks depicts the observation that the nodes with high degrees (called rich nodes) are inclined to intensely connect with each other [71]. In other words, the rich club connectivity is a measure of how close induced graphs are to cliques. In a social context, a strong rich-club phenomenon indicates the dominance of an elite of highly connected and mutually communicating entities, as opposed to a structure comprised of many loosely connected communities. In the Internet, such a feature would point to an architecture in which important hubs are much more densely interconnected than peripheral nodes in order to provide the transit backbone of the network.

A monotonic increase of $\Phi_k$ is not condition enough to infer the presence of a rich club phenomenon [71]. Even in random Erdős-Rényi model [72], the Molloy-Reed [73] model and the Barabási-Albert model [74], $\Phi_k$ decreases linearly with $k$.

**Giant component**

A strongly connected component of a directed graph, or a connected component of an undirected graph, is a maximal subgraph $G_{C}$, such that for every pair of nodes $i, j \in G_{C}$, there exist both a path $P_{i \rightarrow j}$ and a path $P_{j \rightarrow i}$. This definition can be extended to undirected graphs by ignoring link directionality. The concept of strongly connected components was first introduced by Tarjan [75], who also presented an efficient algorithm to compute its value.

The giant component of $G$ is defined as its largest strongly connected component, and its size denoted as $m(G)$. The node cardinality of a graph’s giant component is often linked to network robustness. Metric studies often dictate that a network is considered robust if the size of the giant component $m(G)$ remains constant as nodes or links are randomly removed from the network. Specially in computer science, it is often assumed that two nodes cannot exchange information when there exists no path among them, thus $m(G)$ should always be maximized.

Many of the presented topological metrics cannot be computed if the network under study contains more than one strongly connected component. For example, the hopcount between two nodes belonging to different disconnected components is not defined (and often regarded to be $\infty$). The usual way to deal with graph disconnectivity is by picking giant component, which in many real cases includes a high percentage nodes. The metrics computed for the giant component are afterward generalized to the totality of the network.

**Reliability**

Over the last century, a large body of metrics have been proposed under the term reliability. In general terms, the reliability of a given graph $G$ encompasses a wealth of metrics that measure the number of removed elements that lead to disconnected components, i.e. $m(G) < N$. Reliability is a super-set of the following metrics
• The vertex connectivity $\kappa$ and edge connectivity $\lambda(G)$ of a connected graph $G$ are the respective smallest number of nodes and links whose removal disconnects $G$. For any connected graph $G$ it holds that

$$\kappa(G) \leq \lambda(G) \leq d_{\text{min}}$$  (2.33)

where $d_{\text{min}}$ is the minimum node degree in the graph $G$, and it sets a higher bound for the vertex connectivity. A graph is called $k$-connected or $k$-vertex-connected if its vertex connectivity is $k$ or greater.

In addition, the connectivity function [76] specifies the minimum edge connectivity that can be achieved after first removing $k$ nodes.

• The cohesion $\mu(G)$ of a graph measures the minimum number of links whose removal creates a cut node in the network [77]. Thus removal of these $\mu(G)$ links, together with a single node, will disconnect the network.

• The $i$th order edge connectivity [78] is defined as the minimum $|S|$ such that $c(G\setminus S) = i + 1$. When a network becomes disconnected it is desirable to capture the extent of disruption by measuring the size and number of the remaining connected components (defined in Section 2.2.2). After all, a system that has been split into many parts represents a more severe degradation than a system that has been split into a few large parts. To aid in describing such measures suppose a set of links $S \subseteq L$ are removed from the graph $G$ yielding the network $G - S$ with $c(G - S)$ connected components and maximum component size $m(G\setminus S)$

• The $i$th edge separation value [missing reference] is defined as the minimum $|S|$ such that $m(G\setminus S) \leq i$

• The edge integrity [79] of $G$ is the minimum value of the sum $\{|S| + m(G\setminus S)|\}$ over all $S \subseteq L$.

• The edge toughness [80] is the minimum value of the ratio $\{|S| / c(G\setminus S)|\}$ over all disconnecting sets $S$. However, edge toughness always equals $\lambda(G)$, so a more appealing measure is the node toughness: the minimum ratio of the size of a node disconnecting set to the resulting number of components(notice that most of the link-based measures presented have appropriate node-based analogues).

Reliability metrics also consider the case where nodes and/or links fail randomly and independently with known probabilities $1 - p$, i.e. the probability of any node or link to be operative is $p$. Examples of such metrics follow

• The two-terminal reliability $R_{ij}(G)$ between the nodes $i$ and $j$ is the probability that $i$ and $j$ are connected by a path of operating nodes and links.

• The two-to-$K$-terminal reliability $R_{iK}(G)$ is the probability that there is an operative path from node $i$ to all nodes in a specified set $K \subseteq N$.

• The source-to-all-terminal reliability or reliability polynomial $R(G)$ is the probability that there is an operative path from any node $i$ to all other nodes in the network. Notice that $R(G)$ simply expresses the probability that the graph remains connected.
• The pair-connectivity or pair connected reliability \[81\] is the average number of node pairs able to communicate, taken over all possible node and link failures. This measure takes into account the fact that different ways of disconnecting the network are of different severity. For example, a certain number of link failures could separate \( G \) into several connected components \( G_1, ..., G_r \). All communication is then disrupted between nodes in different components, and the resulting communication capacity can be measured by the number of pairs of nodes now able to communicate

\[
\sum_{i=1}^{r} \left( \frac{n_i^2}{2} \right)
\]

where \( n_i \) is the number of nodes in component \( G_i \).

Communication between all pairs of nodes in a graph is usually regarded as a necessary condition for robustness. For example, when a graph becomes disconnected, potentially vital information (or services) stored in the disconnected nodes becomes unreachable. For example, a computing circuit may short circuit due to a failure leading to a disconnected component. Historically \[82\], reliability has been the classical way to quantify network robustness.

Numerous algorithms have been developed for computing the performance measures discussed previously. The effective computation of virtually all probabilistic measures exhibit a worst-case behavior that is exponential in the size of the network \[83\]. For this reason, the exact computation of reliability metrics has historically been confined to networks of small sizes.

**Chromatic number**

The coloring \[84\] of a given graph \( G \) is a map \( c : \mathcal{N} \to \mathcal{S} \) such that \( c_n \neq c_m \) whenever nodes \( n \) and \( m \) are adjacent. The elements of the set \( \mathcal{S} \) are called the available colors. The key property of \( \mathcal{S} \) is its size: typically, we shall be asking for the smallest integer \( k \) such that \( G \) has a \( k \)-coloring, a node coloring \( c : \mathcal{N} \to \{1, ..., k\} \). This \( k \) is the chromatic number of \( G \), denoted by \( \chi \). A graph \( G \) with \( \chi = k \) is called \( k \)-chromatic.

Note that a \( k \)-coloring of a graph is a node partition into \( k \) independent sets, called color classes. The non-trivial 2-colorable graphs, for example, are the bipartite graphs. The non-trivial four color theorem \[85\], and five color theorem \[84\] imply that every planar graph is 4-colorable and 5-colorable respectively.

To avoid signal interference in frequency-division multiplexing communications the channels used by the antennas are chosen so that the same channel is never concurrently used by two neighboring antennas. This problem is named channel allocation and is usually modeled as a generalized list coloring problem.

**Node cover**

Let \( i \) and \( j \) be two nodes in \( \mathcal{N} \). Then \( i \) covers \( j \) in \( G(\mathcal{N}, \mathcal{L}) \) if

1. for all \( k \in \mathcal{N} \setminus \{i\} \), the link \((j, k) \in \mathcal{L} \) implies the existence of link \((i, k) \in \mathcal{L} \), and
2. there exists at least one node \( k \notin \{i, j\} \subset \mathcal{N} \) such that link \((i, k) \in \mathcal{L} \) while \((n_j, n_k) \notin \mathcal{L} \).

In words, node \( i \) covers node \( j \) if all nodes linked to \( j \) are also linked to \( i \), and node \( n_i \) has at least one extra link \[86\][87]. The uncovered set \( U \) is the set of nodes in \( G \) that are not covered by any other node in \( G \), i.e. \( U = \{n \in \mathcal{N} : \text{no node } n_i \text{ covers } n \text{ in } G\} \). The set \( U \) has a handful of interesting properties \[88\], including the mediator property, such
that for each pair of distinct nodes $i$ and $j$ there exists a shortest path $P_{i\rightarrow j}$ that path crosses $U$.

In the field of data communications, an uncovered node can be regarded as possessing an advantage over a covered node. This is so because the uncovered node has access to more information (i.e. neighboring nodes) than the covered node.

### 2.2.3 Spectra Class

The algebraic eigenproblem consists in the determination of the scalars $\lambda$ and the corresponding vectors $x_{N \times 1} \neq 0$ of any matrix $A$ that satisfy the equation

$$Ax = \lambda x$$  \hspace{1cm} (2.34)

The set of scalars $\lambda$ and vectors $x$ are called eigenvalues and eigenvectors of $A$, respectively. Additionally, the set of distinct eigenvalues $\lambda$ is known as the spectrum of $A$. Many useful network properties, such as epidemic thresholds, are related to the spectrum of a graph [89] [22] [90].

The admittance matrix or Laplacian $Q$ is defined as

$$Q = \Delta - A$$  \hspace{1cm} (2.35)

where $\Delta = diag(d_1, d_2, ..., d_N)$ is the degree matrix. The eigenvalues of the Laplacian matrix are represented with the symbol $\mu$. The Laplacian matrix, or admittance matrix, appears in many contexts in the theory of networks, such as the analysis of diffusion and random walks on networks, Kirchhoff’s theorem for the number of spanning trees, etc.

#### Algebraic connectivity

Let $\mu_1 \geq \mu_2 \geq ... \geq \mu_N$ be the ordered set of eigenvalues of the Laplacian matrix $Q$. The algebraic connectivity $\mu_{N-1}$ is the second smallest eigenvalue of the Laplacian matrix [91].

The algebraic connectivity is related to the speed of solving consensus problems in networks (solve distributed decision-making problems with interacting groups of agents), and it is a lower bound for the vertex connectivity [89] [92],

$$\mu_{N-1} \leq \kappa(G)$$  \hspace{1cm} (2.36)

Graphs with certain high connectivity properties (such as concentrators) have been used in the construction of switching networks that exhibit high connectivity. Tanner [93] proved that connectivity properties can be analyzed by its eigenvalues. He observed that a small ratio of the algebraic connectivity to the dominant eigenvalue implies good expansion properties. Additional properties of the algebraic connectivity are described in Chapter 3.4 and Appendix A.

#### Spectral radius

Let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N$ be the ordered set of eigenvalues of the matrix $A$. The spectral radius $\rho$ of $A$ is

$$\rho = \max_{1 \leq i \leq N} |\lambda_i|$$  \hspace{1cm} (2.37)

$^3$Not to be confused with modularity $Q$. 
2.2 TAXONOMY OF TOPOLOGICAL METRICS

Figure 2.8: Fraction of infected nodes (ordinate axis) as a function of a virus infection rate $\tau$ (abscissa axis). Given a susceptible network and an infectious virus, if the virus spreads at a slower rate than an epidemic threshold $\tau_c$, then the virus will eventually die out. However, if the same epidemic spreads with a rate greater than $\tau_c$, then a faction of node in our graph will remain infected. The N-intertwined model (further explained in Chapter 3.3) predicts the epidemic threshold as a function of the largest eigenvalue of the adjacency matrix, as $\tau_c = \frac{1}{\lambda_1}$.

The epidemic threshold [94] is defined as follows: for effective virus spreading rates below $\tau$, the contamination in the network dies out, while for effective spreading rates above $\tau$, the virus is prevalent, i.e. a persisting fraction of nodes remains infected as displayed in Figure 2.8. Piet et al. [95] found the relation $\tau = \frac{1}{\rho}$. It follows from this result that the smaller the spectral radius, the higher the robustness of any network against the spread of viruses.

**Ratio $\mu_1/\mu_{N-1}$**

The ratio of the largest eigenvalue $\mu_1$ and the second smallest eigenvalue $\mu_{N-1}$ of the Laplacian matrix $Q$ is an index of the synchronizability of a graph. The synchronizability indicates whether the synchronized state of a graph dynamic will remain stable, for a sufficiently large time [96].

This ratio is also referred to as the paradox of heterogeneity, because networks with a homogeneous degree distribution synchronize faster than networks with a heterogeneous degree distribution [37]. The larger this ratio $\mu_1/\mu_{N-1}$ is, the more difficult it is to synchronize a network of oscillators, and vice versa [97].

**Fiedler vector**

The eigenvector corresponding to the second smallest eigenvalue (i.e., the algebraic connectivity $\mu_{N-1}$) of the Laplacian matrix is also called the Fiedler vector. Spectral partitioning methods have been developed to split the nodes of a graph into two groups, such that the number of links between the groups is minimized [98] [99]. For the specific case where the network is split into two non-intersecting groups, we can define an index vector
s with N elements:

\[
s_i = \begin{cases} 
+1 & \text{if the node } n_i \text{ belongs to group 1} \\
-1 & \text{if the node } n_i \text{ belongs to group 2}
\end{cases}
\] (2.38)

The group selection that minimizes the number of links between such groups follows

\[
s_i = \begin{cases} 
+1 & \text{if } x_{N-1}(i) \geq 0 \\
-1 & \text{if } x_{N-1}(i) < 0
\end{cases}
\] (2.39)

where \( x_{N-1} \) is the Fiedler vector.

Community detection in large networks might prove a very useful tool. Nodes belonging to a tight-knit community are more than likely to have other properties in common. For instance, in the world wide web, community analysis has uncovered thematic clusters, in neural networks communities may be functional groups, etc.

**Principal eigenvector**

The Principal eigenvector of an adjacency matrix \( A \) is the eigenvector corresponding to the largest eigenvalue \( \lambda_{\text{max}} \) (i.e., the spectral radius \( \rho \)). The principal eigenvector maximizes the eigen equation (2.34) such that

\[
Ax_1 = \rho x_1
\] (2.40)

which attributes useful properties related to graph theory [100] [92] and data processing [49] [101] to \( x_1 \).

First, the components of the principal eigenvector are directly related to node centrality or relative importance of nodes in a graph. Let me illustrate this idea with an example: the hyperlink structure of the World Wide Web can be modeled as a directed graph with \( N \) nodes where each node in the graph represents a web page and the directed links represent hyperlinks. The corresponding adjacency matrix \( A \) is called transition matrix. Now imagine a web surfer who at each time step visits a random web page in Figure 2.9. The user randomly picks a hyperlink on the current page \( i \) and jumps to a random page \( j \) it links to with a given Markovian probability \( p = a_{ij} / \sum_{j=1}^{N} a_{ij} \). The stationary vector \( \Pi \) (i.e. \( \Pi_i \) equals the chance that the user is viewing the web page \( i \)) is then defined to be the stationary distribution of the Markov chain, or equivalently the principal eigenvector \( x_1 \) of the transition matrix. Google’s PageRank algorithm [100] uses a variation of the principal eigenvector to assign authority weights to web pages as illustrated in Figure 2.9. While originally designed in the context of link analysis on the web, it can be readily applied to citation patterns in academic papers and other citation graphs.

The Principal eigenvector is tightly related to Principal Component Analysis (PCA) [101], both aiming to find patterns in high dimensional data. Where the luxury of graphical representation is not available, PCA is a way of identifying patterns in data and expressing the data in such a way as to highlight intrinsic similarities. Given a data set of limited elements, where each element is described with scalar measurements, the eigenvectors of the respective covariance matrix lie in the axis that maximizes the variance of the data set. In other words, the Principal eigenvector \( x_1 \) is a projection that accounts for as much of the variability in the data as possible. Although there are many ways to
2.2 TAXONOMY OF TOPOLOGICAL METRICS

Figure 2.9: Numeric example of PageRank values in a small graph with 11 related pages. The size of the node \( i \) is proportional to the \( i \)-th elements of the principal eigenvector. The image shows that Page B is the most important, therefore Page B is the first page displayed in a Google search. Often, the eigenvector centrality is positively correlated to the nodal degree, but this is not always the case. As illustrated, Page C has a higher PageRank than Page E, even though it has a lower degree.

apply PCA (e.g. image compression techniques [49]), the original usage as a descriptive, dimension reducing technique is probably still the most prevalent single application.

Eigenvector and Eigenvalue Set

Generally, researchers focus on the Fiedler vector of the matrix \( Q \) and the Principal eigenvector of matrix \( A \), i.e. the eigenvectors and eigenvalues that maximize and minimize the eigen equation (2.34), and rightfully so. The high-order parameters of a graph have the highest impact on nonlinear dimensionality reduction methods and the related diffusion maps, because of their maximization properties. An example of this is Google’s PageRank, which is exclusively based on the largest eigenvector of the World Wide Web adjacency matrix. For an in-depth analysis of the Fiedler vector, we refer the reader to Chapter 3.4, where we study the effect of the Fiedler eigenvector on synchronization processes.

However, low-order eigenvectors and eigenvalues should not be ignored. They provide a unique graph signature, much like the Fielder transform provides a unique representation of a time series. For example, Dajie et al. showed that the graph spectra can be used to measure the randomness of a graph [102]. Another recent study shows that even low-order eigenvectors can be meaningfully interpreted in terms of structural heterogeneity in the data [103]. In this publication, the USA country-to-country migration network is studied, focusing on the low-order eigenvectors. Cucuringu et al. show that low-order eigenvectors display high localization properties, which closely match real migration data. To further illustrate the relevance of low-order eigenvectors, Fig. 2.10 illustrates the heatmaps of four different tree-like graphs, where the symmetrical spectral signature of tree graphs can be observed [22].
After reading this chapter, the reader should be familiar with the basic concepts of graph theory. The notions of node, link, hopcount, degree, and spectra proved to be powerful tools to investigate social, technological and biological networks in the past, present and probably in the future. With the new tools at our disposal, we are now able to state

Regarding the set of eigenvalues, Faloutsos et al. [104] showed that the higher eigenvalues of the Internet AS-level correspondent adjacency matrix follow an empirical power law, such that:

$$ f_\lambda(x) \simeq x^{-\delta} \quad (2.41) $$

Separately, Dogorostev et al. [105] showed that the tail parameter $\beta$ of the degree distribution, and the tail parameter $\delta$ for the highest eigenvalues are directly related through

$$ f_\lambda(x) \simeq x^{1-2\beta} \quad (2.42) $$

Eq. (2.41) and (2.42) provide a relation between the tail exponent of the degree distribution, and that of the spectrum. The reader is referred to Cvetkovic’s work [106], for a list of additional graph spectra properties.

2.3 Chapter Summary and Conclusions

The concepts of node, link, hopcount, degree, and spectra proved to be powerful tools to investigate social, technological and biological networks in the past, present and probably in the future. With the new tools at our disposal, we are now able to state
the famous Königsberg bridge problem in the following, rigorous way: define a graph $G(N, L)$, where nodes represent the four landmasses in Fig. 1.1, and two nodes $(i, j)$ are connected if there exists a bridge joining landmasses $i$ and $j$. The problem lies on answering the following question: does a node $i \in N$ exist for which we can find a path $P_{i \rightarrow i} = \{(i, k_1), (k_1, k_2), \ldots, (k_L, i)\}$, such that $P_{i \rightarrow i} \cap L = L$?

This chapter also exposes an extensive list of thriving, commonly used metrics, as they have been extensively used by the scientific community over the last decades. Even though graph theory has been around for centuries, I believe that the latest technological advancements and overwhelming data availability are certain to restructure the landscape of graph metrics. New, challenging observations are published continuously, and novel metrics are born as a consequence. A continuous, collective, effort is required to reach a consensus on a taxonomy of graph metrics.
CHAPTER 3

Robustness Measured as a Single Metric

“God does arithmetic.”

Carl Friedrich Gauss, 1777 - 1855

3.1 Three Approaches to Robustness

To date, there does not exist a quantitative definition of network robustness, which, depending on the context, covers contrasting interests and points of view. Single metric definitions appear on a daily basis to solve practical case-scenarios. These measurements are often related to node up-time, or mean number of failures per time unit, however, these specialized metrics cannot be generalized to any network scenario. Attaining a general definition of robustness remains an open question, barred by our limited understanding of metric correlations, normalization, targeted attacks, and network interdependencies. These complex issues have been partially addressed by different initiatives, causing works and frameworks related to resilience metrics to appear over the last years [107]. However, different underlying assumptions result in a lack of standardization.

Based on three real world case scenarios, this chapter introduces three examples of independent single-metric robustness definitions. Our three examples will demonstrate how robustness can be arbitrarily defined in any way that suits a given network purpose, and will help us understand the challenges that arise when attempting to define it.

First, Section 3.2 approaches robustness of voice traffic exchanged among base stations in a cellphone network, by minimizing the hopcount between all node pairs. In order to protect active traffic against link or node failures in multi-hop communications networks, several so-called protection schemes have been introduced. The most established ones are path, segment, node and link protection. However, these schemes are limited as
challenges are modeled at random. Challenges in real networks can have very different characteristics, e.g. unusual high load can temporarily overload the network in a particular area, whereas a heavy rain thunderstorm cell affects all wireless links while moving across the network area, and last attacks exploiting software bugs affect systems spread all over the network. Thus, we propose to explicitly take the high impact challenges of the deployment into account as they are perceived by a risk assessment procedure. We do this by introducing a risk-group concept into the multi-path placement scheme, which provides an evaluation of the likelihood of a challenge to simultaneously affect two network elements. In such a real-time scenario, the hopcount of the backup path becomes the critical robustness metric, which directly relates to the quality of the phone calls.

Second, in Section 3.3 we delve into the topic of epidemiology, where nodes represent vulnerable entities, susceptible to a disease. We know that communication networks are prone to virus and worms spreading and cascading failures. Recently, a number of social networking worms have spread over public web sites. Another example is error propagation in routing tables, such as in BGP tables. The immunization and error curing applied to these scenarios are not fast enough. There have been studies on the effect of isolating and curing network elements, however, the proposed strategies are limited to node removals. This section proposes a link isolation strategy based on the quarantining of susceptible clusters in the network. This strategy aims to maximize the robustness towards epidemic control while minimizing the impact on the clusters performance. We empirically study the influence of clustering on robustness against epidemics in several real-world and artificial networks. Our results show an average curing rate improvement for the studied real-world networks under analysis.

Last, Section 3.4 equates robustness to network synchronizability by studying the algebraic connectivity of interdependent networks. Synchronization phenomena in large communities of interacting systems appears in physical, biological, chemical, and social systems. A fruitful approach to the problem of synchronization consists of modeling nodes in a graph as oscillators within a topology describing the interdependence among them. In this section, synchronization (or robustness) is placed in the context of networks-of-networks, where single network models are replaced by a more realistic hierarchy of interdependent networks. We show analytically and numerically how interdependent networks experience a sharp transition after the addition of sufficient interlinks among them. The location of the transition is established by the algebraic connectivity ($\mu_{N-1}$) of the network.

Further, normalization and correlation issues will arise if we attempt to define robustness as a metric aggregate, as opposed to a single-metric. The later issues will both be tackled later in Chapter 4, leading to the general picture of robustness exposed in Chapter 5.

### 3.2 Measuring Robustness as Hopcount

#### 3.2.1 Introduction

In the face of exploding user traffic in cellular networks, increasing the spectral efficiency to achieve higher access capacities is widely considered to be only possible via SDMA (space division multiple access). To this end, very large numbers of small cell base stations will be brought up in the near future. As an example, Picochip, a femtocell maker, claims
that London needs to install 70,000 femtocells by 2015 to provide decent 4G LTE mobile services [108]. Since wired backhaul will not be generally available for all small cells, wireless backhaul networks will gain importance. An example network layout is depicted in Figure 3.1. Small cells provide access to the users via their LTE interface and use point to multi-point or point to point links (e.g. in the millimeter wave spectrum) for backhauling building a mesh network for redundancy reasons. Uplink data is forwarded to the wireless aggregation network consisting of relay nodes which do not provide access to users. Finally, the data gets forwarded to the wired part of the aggregation network (often called metro network) which is connected to the mobile operator’s core network. In areas where huge numbers of users are expected, e.g., train stations, or an Olympic stadium, the small cell density can be very high.

A drawback of a wireless backhaul network is the instability of its links. Even highly directed microwave links such as used for carrier-grade mobile backhaul networks are affected by bad weather conditions and link quality may degrade or a link might fail completely [109]. To avoid re-routing latency, a common approach is to proactively establish backup routes which are used in case the primary route fails. In the routing community, this is called multi-path routing, in the telecom world, this is known as protection.

Given a source node, a destination node and a primary path between them, protection can happen at multiple levels of granularity. Figure 3.2 shows different widely established protection schemes. Clearly, more fine-grained protection comes at the expense of more signaling messages. On the other hand, the problem of less connectivity between primary and backup path is that, in cases of failures, it might be difficult to salvage so-called in-flight packets, i.e. packets which are already on their way when a failure occurs. This leads to larger gaps (i.e. consecutively lost packets), which is particularly detrimental to voice traffic where clipping is audible already with two or three lost packets in a row.

Rope-ladder protection (RLP) as introduced in [110] (cf. also Figure 3.2e)) combines the advantages of path, node and link protection by constructing two node-disjoint paths between \( s \) and \( d \) (the “ropes”) and connecting each node on the primary path with a node on the backup path via so-called “rungs”. As shown in [110], this increases path diversity and path lifetime while reducing loss gaps. However, [110] focused on rope-ladders, where primary and backup path are placed close to each other to minimize the switch-over delay. Moreover, it assumed that each node of the primary node is connected to the backup path by a rung. In this section, we will focus on this first requirement and the impact the limited freedom of path placement has on the resilience of the resulting
Figure 3.2: Different protection schemes. a) path protection. b) segment protection. c) node protection. d) link protection. e) rope-ladder protection.

The major novelty of this contribution over [110] is that it introduces the notion of robustness, by means of challenge-aware protection schemes. For example, a challenge to a wireless backhaul communication network could be a thunderstorm, a congestion hotspot or a virus attack. Many protection schemes are either completely challenge-unaware or designed to meet only one particular challenge. In this section, a path construction algorithm is proposed that can optimize the protection structure with respect to multiple high-impact challenges [111]. We consider a challenge to be of high impact if its effect on the protected assets is high, i.e. the probability of the challenge occurring and the imposed loss is high. The risk assessment must be performed by experts of the network operator and are deployment and business specific. Based on this assessment, we apply the Shared-Risk Link Group concept known from the optical networks space [112] to the used protection scheme. In a nutshell, the optimization process minimizes the impact of the challenges by minimizing the weighted sum of the backup links, i.e. the hopcount.

A special entity of the network management system translates a challenge into a so-called \textit{risk group}. Based on this, it steers the rope-ladder construction process such as to be maximally robust towards the high impact challenges, by minimizing the crossed link weights. Our simulations show that a focus on a small challenge set is necessary for high resilience, as with every additional challenge considered the resilience degrades (in some cases even worse than the challenge agnostic scheme).

The mentioned special entity is the Graph Explorer, introduced in [113]. The Graph Explorer (GE) is a general tool that can explore a large set of properties in multi-hop networks. One example would be to determine a “link criticality” value for each link in the network, i.e., the severity of the impact which the failure of a link has on the rest of the network in terms of traffic load distribution, jitter, etc. For the sake of this section, we will use the capabilities of GE to compute the impact of challenges on the network and to steer the rope-ladder construction process accordingly. This will be described in detail in Section 3.2.3.

This section is further split into four \textit{subsections} (hereafter referred to as \textit{segments}). Segment 3.2.2 presents previous research in the domain of multipath QoS routing or protection and its relation to our proposition. Segment 3.2.3 introduces the rope-ladder construction with the help of GE, which find the protection scheme that minimizes the hopcount. Segment 3.2.4 presents a performance evaluation based on simulations. Seg-
3.2 MEASURING ROBUSTNESS AS HOPCOUNT

3.2.5 concludes with a short summary.

3.2.2 Related Work

As mentioned previously, work in the context of this section is discussed as multi-path routing as well as protection schemes. There is quite a number of contributions in the domain of multi-path routing protocols. An overview of this diverse field can be found in [114]. Different multi-path routing protocols focus on a number of aspects like load balancing, bandwidth bundling, security, congestion control or even security (sending packets of a sensitive flow via different paths to make eavesdropping more difficult). One goal is obviously resilience to failures.

Multi-path protocols can be divided into proactive, reactive or hybrid protocols. In proactive strategies, topological information needed for route discovery is exchanged regardless of whether or not it is needed at the moment, in reactive strategies, routes are only discovered on demand. Proactive protocols have the advantage that alternate paths are already available when needed, while reactive protocols must discover them first. On the other hand, proactive protocols typically require more messages in scenarios with mobility and moderate traffic, since in these cases a lot of information is propagated prophylactically even if never used at all. There are hybrid strategies which use a proactive protocol within a limited zone (e.g. in a cluster) and a reactive protocol for larger distances.

Most works on multi-path routing try to establish paths in a decentralized way. In this section, however, we will present a rope-ladder construction process which assumes a centralized entity with global topological knowledge (which hosts the GE). While this might be unrealistic for MANETS or sensor networks, it is realistic for operator-owned networks like metro- or backhaul networks.

A further classification refers to the independence of the individual paths of the multipath. To increase security or robustness, the paths should be as independent as possible which gives rise to node-disjoint or link-disjoint multipaths. Some protocols do not make any statement or assumption about the path independence. Another criterion of multi-path routing protocols is the path count, i.e. the number of alternate paths. In some protocols, this is fixed [115], in others, the number can be parametrized [116].

Among the most well-known multi-path protocols are ad hoc on-demand distance vector multi-path (AODVM) [117] or split multipath routing (SMR) [115], two path protection schemes. Segment and node protection schemes are sometimes referred to as braided multi-path. A known scheme is proposed in [118]. The advantages of rope-ladder protection over these schemes have been shown in [110].

In the telecom world, particularly in the area of optical networks, related work is known as protection schemes. Terms like path, node or link protection are commonly used in traffic engineering technologies such as MPLS and do not need further discussion here. Interesting to mention, however, is that the optical networks community has introduced a concept called shared-risk link group (SRLG) [112]. A SRLG contains all links in a network that are susceptible to the same risk. The typical use case would be two optical fibers which share a common duct. If the duct is destroyed, likely not only one but both fibers share its fate, leading to disruption of traffic through both fibers. The concept of SRLG is very generic, allowing to capture arbitrary risks. Similarly, shared-risk node groups capture risks impacting one or multiple nodes. In [119], SRLG and SRNG are
combined into shared-risk resource groups.

Probably most of the SRLG related propositions occupy themselves with finding SRLG (SRNG/SRRG) diverse paths (path protection). However, they do not compare different protection schemes. In this section, we will compare path and rope-ladder protection in the face of SRLGs.

3.2.3 Constructing Challenge-Aware Protection Schemes

In this segment, we describe how risk-aware protection schemes between a source and a destination node are constructed. First, we will introduce risk group models of three different challenges. Then we describe protection construction process based on the challenges which are to be considered.

**Challenge Model and Risk Groups**

A challenge is an event which occurs in the network and which threatens the network’s normal operation. Examples for such challenges in wireless networks include for example adverse weather conditions, virus attacks, failures of software components, equipment theft or network overload. An exhaustive classification of challenges can be found in [120]. The network is defined by its interconnection pattern, i.e. network topology, which can be represented as a graph consisting of sets of nodes $N$ and links $L$.

As indicated previously, the optical networks domain has introduced the idea of shared-risk link groups (SRLG). Here, we apply this concept to multi-path protection. A challenge $C$ is modeled in terms of such a risk group as a set of network elements failing simultaneously. The risk group of a challenge $C$, denoted as $RG_C$, is defined in terms of the logical vicinity of the protected elements (i.e. nodes and links): given that node $n_1$ is in risk group $RG_C$, node $n_2$ is also in $RG_C$ if its logical vicinity to $n_1$ – denoted $v(RG_C,n_1,n_2)$ – is above a certain threshold $\tau$. Formally,

$$n_1 \in RG_C \Rightarrow \forall n_2, v(RG_C,n_1,n_2) > \tau : n_2 \in RG_C$$

and similarly

$$l_1 \in RG_C \Rightarrow \forall l_2, v(RG_C,l_1,l_2) > \tau : l_2 \in RG_C$$

for the links. The impact of all challenges has been modeled with a step function (corresponds with setting $\tau = 0$) in our simulations. In other words, any link or node affected by the challenge $C$ fails reducing its bandwidth to zero hence $RG_C = C$. The impact function $\tau$ can easily be extended to complex and realistic scenarios.

The logical vicinity function $v(RG_C,n_i,n_j)$ of all considered challenges needs to be defined manually by a network expert and can span one or more arbitrary dimensions. For instance, in areal challenges (e.g. a storm cell) the logical vicinity of two elements correlates with the geographical vicinity of the elements, whereas logical vicinity of a challenge exploiting a software bug correlates with the vendor ID. In this segment, we have modeled three different classes of challenges: (i) a flash crowd event at a congestion hotspot, (ii) a heavy rain cell, and (iii) a virus targeting a firmware bug.

**Congestion hotspot.** A single congestion hotspot is defined by a static area, e.g., a train station or a stadium, where huge numbers of users can cause overload situations. All received connection requests are legitimate but cannot be satisfied by the system simultaneously; such events are called flash crowd events in contrast to denial of service
attacks which are of malicious nature. The logical vicinity function is defined by the area the \( C \) affected nodes are located at:

\[
n_j \in RG_{C_i} \Rightarrow Hotspot(n_j) = Hotspot(n_i)
\]

**Thunderstorm cell.** A different type of areal disruption is a thunderstorm cell moving randomly across the graph, producing a large set of independent thunderstorm challenges \( \mathcal{C} = \{C_1, C_2, \ldots, C_k\} \). The logical vicinity function of a single thunderstorm challenge \( C_i \) (for \( i \leq k \)) provides that links \( l_1 \) and \( l_2 \) appear within the same risk group if the rain cell area with a radius \( r(C_i) \) and an given epicenter \( e(C_i) \) overlaps both links. Formally,

\[
\forall l_j \in RG_{C_i} \Rightarrow d(l_j, e(C_i)) < r(C_i)
\]

where \( d(a, b) \) is the distance function between points \( a \) and \( b \). We consider this a reasonable first order approach, since environmental studies model storms as concentric ellipses [109]. Often, logical vicinity will be related to geographic positioning (e.g. distance to the epicenter) but other environmental characteristics may define this function, too.

**Firmware vulnerability.** The last attack we considered is a generic virus attacking one firmware vulnerability of exposed graph elements, producing a set of firmware challenges \( \mathcal{C} \). If a single firmware challenge \( C_i \) is launched against the mesh network, all nodes using the targeted firmware version \( i \) are threatened and hence share the same risk group. Formally,

\[
\forall n_j \in RG_{C_i} \Rightarrow firmware(n_j) = firmware(n_i)
\]

where \( firmware(e) \) is the firmware type on a mesh element vulnerable to a predefined virus type. Without loss of generality, we applied step functions to model firmware impacts, such that every node featuring \( firmware(e) \) will have their forwarding capacity reduced to zero.

Before going into the details, consider some principal advantages of challenge aware protection schemes over challenge agnostic protection schemes. If the challenge would be, for example, a thunderstorm cell, then a multi-path protection scheme could be optimized by providing two paths which are separated by more than the diameter of the expected thunderstorm cell, which means that the latter would normally not affect both paths at the same time.

**The Construction Process**

In order to construct challenge-aware rope-ladders, we have combined the RLP scheme with the Graph Explorer as introduced in [113]. The Graph Explorer is a tool to assess various metrics of a network in the face of an arbitrary number of challenges occurring simultaneously in this network. We use this tool to calculate the risk groups before starting to place the rope-ladder structure. The required calculation can already be done during the network planning phase and the results stored for the actual placing of the protection structure. In Segment 3.2.4 we will compare the original protection scheme with an improved, now challenge-aware protection scheme for two different types, namely path protection (PP) and rope-ladder protection (RLP). The construction of a rope-ladder is divided in three sequential steps as follows.

**Step A: Placement of the Primary path.** As the risk groups depend on the links and nodes of the primary path, the choice of the primary path is a crucial one in our
process. An intuitive approach would choose the primary path to be the shortest path from source to destination. However this can lead to fragile backup paths crossing high risk elements, as illustrated in Figure 3.3.

We propose that the process should iterate over all paths up to a maximum stretch with respect to the shortest path. thus providing providing a set of $S_P$ paths, such that for all $|P|$ in $S_P$, $|P|$ is smaller than the shortest path times an arbitrary stretch factor. Eventually, the primary path which leads to the shortest (weighted) hopcount backup path will be selected as the primary path for the connection (as we will describe in Step C).

**Step B: Calculation of Link Weights.** The input to this step are all the primary paths $S_P$ provided by step A, and the high-impact set of challenges $C$ from which the network should be protected. Depending on the vicinity function $v$, additional information must be made available such as firmware ID, the frequency allocation plan, etc. Multiple risk groups can be added into a unified risk group (URG) by merging the link weights of different challenge types. This merging function must be determined during the network manager’s risk assessment process, and it should account for the respective occurrence probability of the different challenges types. In our simulations we assumed that all challenge types are both independent and equally probable. Hence all weights belonging to the same challenge type were further normalized to sum the complete probability.

The output of step B are multiple weight clouds $W$, i.e. sets of node and link weights representing the risk group memberships with respect to each primary path $P$ in $S_P$. This weight cloud calculation process is described in Algorithm 1. In words, the weight
of element $y$ will increment for every time that $y$ shared a challenge in $C$ with any of the elements in $P$. An intuitive visual representation of each primary path’s risk group is the union of all the challenge instances $C$ (e.g. thunderstorms in this example) that intersect with the primary path by at least one link or node. The storm’s link weights associated with the shortest path can be illustrated as the cloud shown in Figure 3.4.

**Algorithm 1:** Compute a primary path’s weight cloud.

**Data:** Primary path $P$

**Data:** Set of challenges $C = \{C_1, C_2, \ldots, C_k\}$

**Result:** Weight cloud $W$

begin

\[
W[] \leftarrow 0
\]

for $x \in P$ do

for $C \in C$ do

if $x \in C$ then

for $y \in C, y \neq x$ do

\[
W[y] = W[y] + v(RG_C, n_x, n_y)
\]

normalize($W$)

end

**Step C: Placement of the Backup Structure.** A set of backup paths $B$ is found by iterating over all URG pairs $\{P, W\}$ offered by step B. A shortest path algorithm determines the backup path with the least weight which does not exceed an arbitrary stretch limit. The basic idea of our approach is that the backup path circumvents the cloud in Figure 3.4 and stays out of it for as long as possible, hence minimizing the link weights that will be crossed.

Finally, once all $\{P, W\}$ pairs have been processed, the backup path with the lowest weight in $B_{optimal}$ is selected to form the rope-ladder; the rungs are deemed secondary, and determined after $B_{optimal}$ is set (the rungs are cross-connects from the primary path to the backup path to minimize the loss gap for real time application flows - see [110] for details).

### 3.2.4 Simulation Results

This segment describes the simulation scenario built to evaluate the performance of challenge-aware RLP, as introduced in Segment 3.2.3, and presents a qualitative analysis of the simulation results.

**Scenario description**

Our simulation scenario consists of the wireless backhaul network topology $G_B$ introduced in Segment 3.2.1. We modeled the wireless part of the network between Access Points as a meshed lattice with 61 Access Point nodes, 3 Relay nodes, and 197 links as depicted in Figure 3.4. We chose a meshed lattice as a testbed because it allows an easy construction of rope ladders even when multiple links are under the effects of a challenge. This enables
us to progressively measure the effect of evolving challenges while being certain that backup paths will exist for most node pairs.

The three challenges proposed in the previous segment are configured as follows. For the thunderstorm cell challenge, the storm has a radius in the range $[50, 120]$ m. Its epicenter was shifted in steps of 14 m, providing a total of 5,000 storm instances. For the firmware virus challenge we assumed that each link uses one firmware version, and randomly scattered 4 different firmwares among all 197 links. Finally the hotspot challenge expands the size of an arbitrary train station 150x150m in a static area next to a Relay, as roughly illustrated by a shaded gray ellipse in Figure 3.1.

The selected application is a unicast VoIP application. A source node communicates with a destination node by generating an uninterrupted unicast stream of 64 kbps CBR with 160 bytes data frames, which simulates a G.711 VoIP codec. All the links in the graph have a capacity of 1 Mbps duplex connections with 10 ms delay. For each non adjacent AccessPoint-Relay pair in $G_B$ a RLP scheme was constructed as described in the previous segment, consisting of a primary path, a backup path and the respective rungs. The structure is built via the Graph Explorer such that the backup path is maximally disjoint from the primary path’s risk group, while favoring the shortest distance between source and destination. Figure 3.4 illustrates this process for a single source-destination pair by assigning a color to each link in the graph.

To simulate the effect of a challenge on a voice stream, a voice call is held between two random nodes for an arbitrary time span of 3 minutes. This data flow is established
via a primary path following a RLP scheme. One minute into the call, an instance of a challenge occurs; if the challenge is an instance of a storm cell risk group, the affected links represent the area of the storm; for an instance of a firmware virus attack risk group, the affected links are sharing the same firmware. In both cases, the bandwidth of the affected links is reduced to zero for the duration of the challenge, virtually disconnecting them. As soon as the links become unavailable, the central routing protocol will divert in-flight packets and adapt routing tables to the RLP scheme backup path through the rung which is closest to the challenge. The challenge remains in place for one minute, after which all the links in the network are restored to their initial state, and the data stream is reverted to its default primary route for the remaining minute. We repeated this setup for the three considered challenge types.

Simulation results

The main outcome of the simulations is illustrated in Figures 3.5-3.7 and can be interpreted as follows.

First, we measured the packet loss that different protection schemes suffered by a storm cell occurring. The percentage of packets lost by an oblivious RLP scheme (i.e. a rope ladder uninformed about possible challenges during construction) is 10.3%, approximately two times the percentage of packets lost by the $RG_{Storm}$ aware RLP scheme, which lost 4.8% of packets as shown in the leftmost chart of Figure 3.5. Similarly, for PP schemes the inclusion of the risk groups also reduced the percentage of lost packets. These numbers illustrate a significant performance increase with respect to the experienced packet loss at a cost of no additional resources, since the length of the primary path remains almost constant.

Secondly we evaluate the gap size, measured as the the maximum difference in sequence numbers between two consecutive received frames. Given that the routing is controlled by a central authority, the delay induced by a challenge message propagation is dismissed, i.e., the routing tables are instantly updated across the network. This effect works in favor of PP schemes, by ignoring the propagation delay involved in route table synchronization. Nevertheless the flows’ gap size effect of rerouted in-flight packets is still noticeable. For example, suppose that the last link in the primary path fails. The PP scheme virtually loses all in-flight packets along the main path, because they are to be routed all the way back to the source, causing them to arrive to destination with an expired sequence number. Gap-sensitive applications such as voice traffic will suffer clipping when exposed to this sort of gap losses. RLP schemes avoid the generation of gaps by instantly re-routing in-flight packets to the backup path via the closest rung, avoiding packet delays. All three simulation scenarios displayed in Figure 3.6 illustrate RLP challenge aware schemes suffering smaller gap sizes than PP oblivious schemes.

In order to test the behavior of RLP schemes subject to challenges not included in the planned risk group $RG_C$ we also measured the performance of a challenge aware RLP scheme against unexpected sets of challenges. The blue striped bars in Figure 3.6 illustrate this effect: challenge aware RLP schemes’ performance degrades under the attack of unexpected challenges. The performance of schemes under unexpected attacks may even degrade beyond their oblivious counterparts. Such is the case for RLR $RG_{Storm}$ schemes under Firmware challenges, which lost 4.4% of the voice packets, as opposed to only 4.0% for the oblivious scheme.
Robustness Measured as a Single Metric

3.2

Figure 3.5: Percentage of packets lost by seven different protection schemes. The left, middle, and right sets of charts display the packet loss effects of Storm, Firmware, and Hostspot challenges on $G_B$, respectively. Each set displays seven bars, corresponding to the protection schemes: oblivious PP, oblivious RLP, $RG_{\text{Storm}}$-aware, $RG_{\text{Firmware}}$-aware, $RG_{\text{Hotspot}}$-aware, and $URG_2$-aware and $URG_3$-aware. As expected, the protection schemes built to withstand a specific challenge are always the most robust. For example, $RLP_{\text{storm}}$ performs best against storm challenges.

This adverse effect motivates us to study multi-challenge aware protection schemes through the use of URG. Ideally a multi-challenge aware rope ladder structure can withstand different non-simultaneous challenge attacks without a significant drop in performance. First we define $URG_2$ as the unified risk group formed by adding the two risk groups with the highest impact out of the storm cell, firmware virus and hotspot. Additionally we define $URG_3$ as the risk group resulting from the addition of all three risk groups. Weights are consequently normalized, as specified in segment 3.2.3.

Simulations show that the percentage of packets lost by $URG_2$ and $URG_3$ schemes under a storm cell challenge are 5.4% and 5.2% respectively (as illustrated Figure 3.5), representing a significant improvement over the oblivious RLP scheme (10.3%), yet not performing as good as a $RG_{\text{Storm}}$ aware scheme (4.8%). Overall, in terms of packet loss all URG protection schemes’ outperform both their oblivious and challenge-aware schemes under the effects of unexpected challenges.

On the other hand Figure 3.6 illustrates URG schemes suffering large gap sizes. The protection scheme with the lowest gap size is the one tailored to the challenge, i.e. $RG_{\text{Hotspot}}$ with a gap of 1.6 packets, however the gap sizes of $URG_2$ and $URG_3$ schemes under flash crowd challenges are 8.6 packets and 3.9 packets respectively, whereas the oblivious RLP scheme only lost 8.3 packets. In conclusion, URG schemes’ gap sizes are highly dependent on the type of challenge and considered risk groups.

In addition to the packet loss ratio and gap size metrics, we also counted the number of cases where a scheme’s source and destination nodes became disconnected, i.e. cases where the challenge split the protection scheme in two disconnected components. In the event that a RLP or PP scheme becomes disconnected the voice call is interrupted, due to the data flow not being able to reach its destination. This effect contributes to the increase
Figure 3.6: Average gap size of protection schemes. The three sets of charts display the gap size effects of Storm, Firmware, and Hostspot challenges on $G_{B}$. Each set displays seven bars, corresponding to the protection schemes: oblivious PP, oblivious RLP, RG$_{\text{Storm}}$-aware, RG$_{\text{Firmware}}$-aware, RG$_{\text{Hostspot}}$-aware, and URG$_2$-aware and URG$_3$-aware. In general, rope-ladder protection schemes perform better than path protection schemes.

of packet loss for the duration of the challenge, as we currently do not consider creation of a new multi-path (RLP or PP) once a voice call was interrupted. Figure 3.7 displays the percentage of schemes maintaining the connectivity between source and destination nodes, for different storm sizes. The challenge occurring during each simulation run had exactly the radius which was assumed during risk assessment and thus the protection scheme was optimized for. For the network under study, schemes constructed with risk group information are more resilient to disconnection than oblivious schemes, as reflected in Figure 3.7. In addition the percentage of surviving protection schemes decreases with the size of the thunderstorm cell.

To summarize, challenge aware RLP and PP schemes outperform their oblivious counterparts when faced to an expected sets of challenges. However these optimizations can lead to performance degradations in face of unexpected challenges. Overall all presented protection schemes performed surprisingly well due to the regular structure of a full meshed lattice graph, which offers high path diversity even under the effect of geographical challenges.

3.2.5 Section Summary and Conclusions

In this section, we have presented an algorithm to improve the robustness of rope-ladder protection schemes in multi-hop wireless networks. We equated the robustness of a path to its length (or hopcount), and evaluated the total network performance by measuring the packet loss and gap size of a simulated VoIP application.

During the network planning phase, we have employed a unified risk group approach which makes use of a logical vicinity function, which through link weights, relates the network elements to risk groups. During network operation the Graph Explorer assesses possible placements of rope-ladders such as to minimize the hopcount, i.e. be maximally
robust towards certain challenges. Using simulations, we have evaluated the robustness of this challenge aware rope-ladder scheme with the original rope-ladder scheme and with path protection. The packet loss rate was reduced by up to 80.4% compared to the oblivious scheme, but more remarkably the number of protection schemes surviving the challenge onset was increased by up to 25%. The focus on the high impact challenges during the network design stage has proven to be critical.

Within the context of backup schemes, two important factors remain as future work. First, there will always be unforeseen challenges which were not taken into account while establishing the protection. We showed that facing the rope-ladder scheme to unexpected challenges can lead to a decrease in performance, which may be alleviated by the addition of unified risk groups. Second, multiple challenges may occur at the same time or at least overlap, so that (protected) paths are affected by a number of simultaneous challenges, rather than a single challenge at a time. This means that robustness of protection schemes must actually also be tested with respect to unforeseen challenges as well as an overlay of risk groups stemming from different challenges which may happen simultaneously.

### 3.3 Measuring Robustness as Modularity

#### 3.3.1 Introduction

From worm epidemics in computer networks to information spread in P2P and ad-hoc networks [121], [122] epidemics have recently attracted a lot of attention. After the scanning worms, a new challenge for network security is posed by the strain of worms that use social networking Websites to spread. Web applications for exchange of information and data introduced new vectors of spread. Many social network worms use AJAX\(^1\) scripts

---

\(^1\)Asynchronous JavaScript and XML
like Samy [123], Yamanner [124] and Mikeyy [125]. Worm spreading usually involves user interaction in order to download worm payload on the local machine as Koobface [126], but recently Web clients are infected simply by visiting a Web page; no user interaction is necessary [123]. The infection risk increases since social networks are not restricted only to Facebook and Twitter, but are becoming embedded in other not strictly social websites like Digg and YouTube. Additionally, social networks have power law network structure which makes them prone to epidemic spreading [122], [127], [128] and [129].

The epidemic algorithms for information dissemination in unreliable distributed networks such as P2P and ad-hoc networks show similar epidemic dynamics on networks [121], [130]. Finally, the propagation of faults and failures can be modeled as an epidemic. Coffman et al. [131] models cascading BGP failures on a fully connected topology. In this section, we concentrate on protection against worms and error propagation in communication networks. By our standards, a robust network is a network that debilitates the spread of epidemics. The protection of important networks in the above mentioned cases is in practice not fast enough, and the infection easily reaches all the segments of the network.

This section proposes and analyzes a fast method to stop or reduce epidemic spreading on networks, thus increasing their robustness. When an epidemic is detected, a network cut is performed by removing links leading to several disconnected clusters of nodes. This separation process allows limited intercommunity communication between nodes to continue, while possibly quarantining the rest of the network. Depending on the speed of the epidemic reaction, it is possible to totally prevent any risk of infection for a number of disconnected clusters. Even with very delayed reaction, the amount of protection, that has to be applied in the network in order to stop the spreading, can be reduced. Thus, clustering can be used in addition to other protection methods. In addition, many real-world networks from on-line social networks to airline transport networks and Internet ASes network typically show a strong community structure [132] [133], which community based provides network cuts.

The removal of links as protection against epidemics was proposed in mathematical epidemiology. The Equal Graph Partitioning (EGP) method uses immunization to remove specific nodes that cut the graph into clusters [134]. However, the immunization takes time, while individual nodes can stop communicating with other nodes immediately after receiving the news about the epidemic. Several authors have studied the reduction of disease spreading using air line restrictions. Goedecke et al. [135] and Epstein et al. [136] used the Susceptible Exposed Infected Recovered (SIER) model and dynamic time travel restrictions. Marcelino et al. [133] used the Susceptible Infected (SI) model together with edge betweenness and the Jaccard coefficient to increase the spreading time [133] by 81% by removing 25% of the links. Due to the multi-community structure of the network with most connected nodes not being the most central, the optimal strategy for flight cancellation is not the removal of nodes (cities), but the removal of intercommunity flights, which introduced an increase in spreading time [133]. We are interested in specific link removal such that intra-community communication is preserved. We are not interested in optimizing existing clustering algorithms, but instead in the general improvement of protection that is possible by using well-defined clustering techniques.

Several algorithms have been proposed to find network communities, modularity maximization being the most popular. The modularity Q, introduced in Chapter 2 of this thesis, is a quantitative criterion that evaluates how good a graph partition is [137]. It
maximizes links within communities, while minimizing the links between them. However, given the exponential number of possible partitions, modularity maximization is an NP problem [52]. In this section, we use a greedy heuristic proposed by Clauset et al. [138] to find an optimal modularity clustering.

In order to quantify the improvements of the network clustering in terms of epidemics, we use the epidemic threshold concept and the $N$-intertwined Susceptible Infected Susceptible ($SIS$) epidemic model [95] on a large set of networks. In a $SIS$ epidemic model, a given epidemic can be stopped provided the network protection functionalities against the virus perform faster than the reproduction of the virus. The epidemic thus exhibits threshold behavior.

This section is structured as follows. In segment 3.3.2, we explain the protection algorithm and describe our examined networks. The epidemic theory used to estimate the protection is explained in segment 3.3.3. Numerical results of our protection algorithm are presented in segment 3.3.4, where we compare early clustering with delayed clustering, and our quarantine model with random link removal. Segment 3.3.5 concludes this section.

### 3.3.2 Quarantine Model

A network splitting protection method based on link removal will be referred to as clustering or quarantining. The moment when a network is quarantined determines how many nodes are completely protected, since the virus is not able to infect nodes outside its cluster. In the first case, if we are able to quarantine a network into clusters faster than the virus is spreading, only a single cluster will contain infected nodes. On the other hand, if the virus infects all the clusters before a quarantine takes place there are still benefits, which are discussed in more details in segment 3.3.4. Usually, the effective speed of clustering the network will be somewhere in between.

We discuss the two boundary cases separately. In the first case we determine the size of the clusters, which provides an estimate of how many nodes will never get infected. The size of the clusters also affects the performance of the network. Larger clusters mean that a larger part of the network can continue exchanging information. Second, we show that the epidemic threshold that divides non-infected from infected networks favorably increases in networks that display clustering features.

To illustrate the influence of clustering on epidemic spreading, we use the following four real-world networks, also summarized in Table 3.1.

- **Euro**: The European direct airport-to-airport traffic network obtained from the European commission for statistics Eurostat.
- **USA**: The direct airport-to-airport American traffic network maintained by the U.S. Bureau of Transportation Statistics and the European direct airport-to-airport traffic network obtained from the European commission for statistics Eurostat.
- **Pol. Blog**: A social network between weblogs on US politics recorded in 2005 by Adamic and Glance [139]. The political blog network is shown in Fig. 3.8, with nodes belonging to different clusters colored in different colors.
- **AS’06**: The Internet AS level topology obtained by Route View in 2006 and posted by the University of Oregon is used to demonstrate the effect of clustering on the virus spread in large infrastructural networks.
• **Digg**: Online social network of friends from www.digg.com, collected by the NAS group at Delft University of Technology. In disease modeling, transport networks are frequently used.

<table>
<thead>
<tr>
<th>Network</th>
<th>$N$</th>
<th>$L_{\text{tot}}$</th>
<th>$L_{\text{removed}}%$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euro</td>
<td>1,247</td>
<td>14,952</td>
<td>47.27%</td>
</tr>
<tr>
<td>USA</td>
<td>2,179</td>
<td>31,326</td>
<td>18.11%</td>
</tr>
<tr>
<td>Pol. Blog</td>
<td>1,222</td>
<td>19,021</td>
<td>7.16%</td>
</tr>
<tr>
<td>AS ’06</td>
<td>22,963</td>
<td>48,436</td>
<td>20.62%</td>
</tr>
<tr>
<td>Digg</td>
<td>281,471</td>
<td>4,354,174</td>
<td>25.02%</td>
</tr>
<tr>
<td>BA 2m</td>
<td>1000</td>
<td>1,971</td>
<td>42.46%</td>
</tr>
<tr>
<td>BA 3m</td>
<td>1000</td>
<td>2,673</td>
<td>58.88%</td>
</tr>
<tr>
<td>ER 0.002</td>
<td>808</td>
<td>980</td>
<td>17.34%</td>
</tr>
<tr>
<td>ER 0.006</td>
<td>1000</td>
<td>3,054</td>
<td>51.27%</td>
</tr>
<tr>
<td>ER 0.02</td>
<td>1000</td>
<td>9,938</td>
<td>55.02%</td>
</tr>
</tbody>
</table>

**Table 3.1**: For each real-world network and graph models, we display the number of nodes $N$, the number of links $L$, and the percentage of links that our modularity-based quarantine algorithm would remove $L_{\text{removed}}\%$ (see segment 3.3.4 for further details).

In order to better understand of the effects of clustering on the network robustness against virus spread, we include several artificial networks with $N = 1,000$ nodes:

- **Erdös-Rényi (ER)**: each node in the ER random graph is connected to every other node with probability $p$. The probability $p$ determines the number of links in the network [140].

- **Barabási-Albert (BA)**: we model power law networks using the Barabási-Albert model (BA) of preferential attachment for different number of links [141].

- Newman’s artificial model of clustered networks as proposed in [61]. The network is constructed in a similar manner as the ER random graph, with two probabilities of link existence: one for inter-community connections and the other for intra-community connections. We have generated several different networks with $N = 1,000$. We have considered networks with 2, 4, 6, 8, 10 clusters. We decided to generate a greater number of networks with two clusters because most of our studied real-world networks consist of mainly two big clusters.

- Additionally, we consider the square lattice, path, ring and tree topologies.

None of the links in any of our networks are weighted; however, the $N$-intertwined model is extendable to a weighted setting [142]. In this section, all networks are modeled as a connected, bidirectional graph $G(N, L)$.

### 3.3.3 $N$-intertwined SIS Model

To model epidemic spread, we use the $N$-intertwined $SIS$ model, which was introduced and discussed by Van Mieghem et al. [95]. $SIS$ models have become a standard to model
Figure 3.8: Network of weblogs on US politics clustered network. Nodes (i.e. weblogs) affiliated to the democrat and republican parties are colored in white and black, respectively. Assuming that a spring embedding layout is a good representation of node affiliation, then modularity does a good job at telling political parties apart.

epidemics, where a node is susceptible to infection (S), then it becomes infected (I) and, after curing, it is susceptible to infection (S) again.

By separately observing each node, a node $i$ at time $t$ can be in one of the two states: infected, with probability $v_i(t) = \Pr[X_i = 1]$ or susceptible, with probability $1 - v_i(t)$. The sum of the probabilities of being infected and susceptible are equal to 1 because a node can only be in one of these two states. The state of a node $i$ is specified by a Bernoulli random variable $X_i \in \{0,1\}$: $X_i = 0$ for a susceptible node and $X_i = 1$ for an infected node. We assume that the protection process per node $i$ is a Poisson process with rate $\delta$, and that the infection per link is a Poisson process with rate $\beta$ which is imminent for all nodes and thus constant in the network. For a node $i$, we can formulate the mean-field approach differential equation

$$\frac{dv_i(t)}{dt} = \beta(1 - v_i(t))\sum_{j=1}^{N} a_{ij}v_j(t) - \delta v_i(t)$$

where $a_{ij}$ is the element of the adjacency matrix $A$ and it is equal to 1 if the nodes $i$ and $j$ are connected, otherwise it is 0. A node is not considered connected to itself, i.e. $a_{ii} = 0$. The probability of a node being infected depends on the probability that it is not infected $(1 - v_i(t))$ multiplied with the probability that a neighbor $j$ is infected $a_{ij}v_j(t)$ and that it tries to infect the node $i$ with the rate $\beta$. Detailed derivations are given in [95] and [142].

In the steady-state, where it holds that $\frac{dv_i(t)}{dt} = 0$, and $\lim_{t \rightarrow \infty} v_i(t) = v_{i\infty}$ for each
Figure 3.9: Fraction of infected nodes in the meta-stable state, as a function of the effective infection rate $\tau$ (rate of spread divided by rate of protection). The figure illustrates the existence of an epidemic threshold $\tau_c$, below which the infection perishes, and above which the infection never dies. This threshold precisely equals the inverse of the largest eigenvalue of the $A$ matrix.

For a fixed curing rate and spreading rate, the fraction of infected nodes as a function of the effective spreading rate $\tau$ is given in Fig. 3.9. The model as well as the real epidemic process have a threshold value at $\tau_c$, as illustrated in Figure 3.9. The threshold can be defined as follows: for effective spreading rates (rate of spread divided by rate of protection) below some critical value the virus in the network with $N$ nodes dies out before...
a large population is infected with a mean epidemic lifetime of the order of \(O(\log N)\). For effective spreading rates above the critical value \(\tau_c\), the epidemic persists and the number of infected nodes is large, with a mean epidemic lifetime \([143]\) of the order of \(O(e^{N^\alpha})\) for a SIS model. The state above the epidemic threshold is referred to as the metastable state. In the metastable state, some constant mean portion of nodes is infected \([95]\).

The epidemic threshold is equal to \(\tau_c = \frac{1}{\lambda_1(A)}\), where \(\lambda_1(A)\) is the largest eigenvalue of the matrix \(A\) \([143]\) \([144]\), as introduced in Chapter 1 of this thesis. Similar bounds exist for Susceptible Infected Removed SIR model \([145]\), \([146]\). In this section we use the equivalent terms \(\lambda_1(A)\) and \(\lambda_{\text{max}}G\) interchangeably.

If the effective spreading rate \(\tau < \tau_c\), the infection will eventually be cured, and for \(\tau > \tau_c\) the infection persists with the average number of infected nodes equal to \(y_\infty\). For example, the largest eigenvalue of a path graph is \(\lambda_{\text{max}}G_{\text{path}} \approx 2\), while that of a star topology is \(\lambda_{\text{max}}G_{\text{star}} = \sqrt{N-1}\). Note that, even though both graphs have the same number of nodes and links \(L = N - 1\), the fraction of infected nodes in a star topology is significantly higher than in a star topology, due to \(\lambda_{\text{max}}G_{\text{star}} > \lambda_{\text{max}}G_{\text{path}}\).

### 3.3.4 Simulation Results

In this segment, we examine two extreme cases. First we study the case of \textit{early} clustering, where a network is clustered faster than the worm is spreading, resulting in a single infected cluster. Secondly, we study delayed clustering, where all the clusters are infected before the quarantine process had time to react. Finally, we compare our quarantined models with random disconnection models, where the same number of links has been removed at random.

#### Early clustering

Defending a network from epidemics by performing quarantines leads to important advantages. First of all, if a network is cut on time and the infection is limited to one cluster, only a percentage of nodes will eventually be exposed to infection. In addition, from the interlacing theorem of graph theory \([106]\), the largest eigenvalue of a subgraph is always smaller than that of the graph. Thus, the thresholds \(\tau_c = 1/\lambda_{\text{max}}\) will always increase for any subgraph, making the quarantined subgraphs more robust against epidemic spreading. The case in which all the clusters are initially infected is later discussed in the next segment. Finally, the lifetime of the metastable state depends on the number of nodes \([143]\) as \(\Omega(e^{N^\alpha})\), for \(\alpha > 0\). The number of removed links using the modularity algorithm ranges from 7% to 58% of the links. The values for different networks are given in Table 3.1.

As illustrated in Figure 3.10, the behavior of \(\lambda_{\text{max}}\text{Cluster}\) for the different network is diverse. For networks with high modularity, such as the lattice and tree topologies, the improvement, measured a lowering of \(\frac{\Delta_{\text{max}}\text{Cluster}}{\lambda_{\text{max}}G}\), is not so significant. For the same type of networks e.g. BA or ER with different number of links, a reduced modularity results in a reduced \(\lambda_{\text{max}}\), which is an improvement. For both cases, the modularity is reduced by generating topologies with a larger number of links (by respectively increasing the parameter \(m\) in the BA model and the parameter \(p\) in the ER model). In addition, the difference between the two largest eigenvalues of different clusters is greater for BA than for ER. The effect can be caused by the homogeneity of the degree distribution of clusters.
3.3 MEASURING ROBUSTNESS AS MODULARITY

Figure 3.10: Early clustering’s relative largest eigenvalue of each cluster $\lambda_{\text{max Cluster}}/\lambda_{\text{max } G}$ as a function of the modularity $Q$ for real-world networks, real-world models (left image), and cluster network models (right image). The lower the eigenvalue ratio, the higher the epidemic threshold, thus lowering the chance we will have infected nodes.

in the ER case, while BA shows a significantly heterogeneous cluster degree distribution.

The threshold $\tau_{c,\text{cluster}} = \frac{1}{\lambda_{\text{max Cluster}}}$ increases as a function of the number of links removed between a cluster and the rest of the network, as shown in Fig. 3.11. In order to preserve as much network communication as possible upon link removal, a small number of links should be removed during the quarantine. On the other hand, $\tau_c$ is inversely proportional to $\lambda_{\text{max Cluster}}$. Hence the networks with best performance show clusters with both low $\lambda_{\text{max Cluster}}$ and low $L_{\text{out}}$, close to the point $(0, 0)$ in the figures. Real-world networks such as the airline networks and AS network perform well, while artificial networks perform much better the smaller the number of clusters in the graph is.

For graphs of the same type, the dependency between the threshold improvement and the number of links removed is close to linear. This is indicated by a constant decrease rate in the largest eigenvalues, as seen in Figure 3.11. In particular, sparse ER graphs, depicted with asterisks in Figure 3.10, required a small number of links to be removed. Thus ER graphs are easily clustered, although they show no significant improvement in $\tau_c$. As expected, the artificial graphs generated with low-modularity generated show the worst (i.e. highest) number of removed links, as seen in the right hand side of Figure 3.10. In other words, graphs with low modularity scores require a high number of removed links, in order to be effectively quarantined.

The size of the clusters after cutting is also an important variable for the performance of the network. Large clusters will allow for node communication after a quarantine. But on the other hand smaller clusters will be more robust to virus spread. The size of the clusters is decided by the modularity algorithm. The distribution of cluster sizes $N_c/N$ is shown in Fig. 3.12. Since we are clustering at an early detection stage, the network will be cut into clusters before the virus can reach any other cluster, except for the one it starts to spread in. This means that the worst case scenario happens when the virus appears in the largest cluster.

Most of the real-world networks have one cluster that contains half of the nodes. In
the case of the European air network, the three clusters pop up, thus leaving more than two thirds of network protected. A BA graph has many small clusters of the size one fifth of network, which leaves four fifths of network protected, as shown in Fig. 3.12. The Digg network has one large cluster which covers half of the network and many significantly smaller ones. The USA air network and the political blog network have 2 large clusters, while the European air network has 3 large clusters and several small ones. The Internet AS topology is more differentiated. There are 8 clusters with 1,000 – 1,500 nodes and two larger ones with 3,000 and 6,000. At the other end of the spectrum, artificial networks can be clustered into small components, e.g. the ER and BA network have a lot of smaller clusters, all comparable in size.

In the right hand side of Fig. 3.12, the number of nodes in the cluster is given as a function of the number of removed links between the cluster and the rest of the network. From our data, we observe that the air network of USA airports has the largest cluster with the smallest number of deleted links.

For the same network, larger clusters tend to have a larger \( \lambda_{\text{max}} \) than the smaller clusters, as seen in Fig. 3.13. This is, however, not true for any graph: compare the path graph of any size with the complete graph of any smaller size.

### Delayed clustering

As opposed to early clustering, where a virus is immediately detected, we now assume a delayed clustering, where an epidemic has spread over the entire network before we are even aware of its existence. If we examine the number of infected nodes using the \( N \)-intertwined model, then it is necessary to apply a protection rate \( \delta \) such that the effective spreading rate \( \tau = \beta \delta \) lies below the threshold \( \frac{1}{\lambda_{\text{max}}} \). For example, given the extreme case where every cluster in the network is infected, the necessary amount of cleaning required to eradicate an epidemic is significantly reduced after partitioning the full network. This is due to \( \lambda_{\text{max}} \leq \lambda_{\text{max}G} \), therefore \( \tau_c(G) \leq \tau_c(\text{Cluster}) \). In other words, we follow...
3.3 MEASURING ROBUSTNESS AS MODULARITY

Figure 3.12: Relative number of nodes in the cluster \( N_{\text{Cluster}}/N \) as a function of the modularity \( Q \) (left image), and the relative number of links leaving the cluster \( L_{\text{out Cluster}}/L \) (right image), of real-world networks and real-world models.

a *divide an conquer* strategy, in which protecting a network split in clusters becomes cheaper than protecting the network as a whole.

Fig. 3.14 presents the percentage of infected nodes as a function of the effective spreading rate \( \tau \) for different clusters, given an artificial network with low modularity \( Q \).

We evaluate the quality of the quarantine by calculating the fraction of infected nodes in the clustered network \( y_{\text{clust}} \) for the effective spreading rate \( \tau \) for which the number of infected nodes in the original network \( y_{\text{tot}} \) reaches 50% and 80%. Then, we calculate the difference between the original value and the improved one:

\[
i_{50\%} = y_{\text{tot},50\%} - y_{\text{clust}}; \quad i_{80\%} = y_{\text{tot},80\%} - y_{\text{clust}}
\]

In Fig. 3.15, the upper bound on the reduction of infected nodes exhibits the tendency to decrease with the modularity of the graph. The improvement is different when there are 50% and 80% of infected nodes in the original network. Air travel networks and ER networks with small average degree do not show significant difference between improvements and have generally small improvements.

The number of infected nodes decreases as the number of removed links (in the whole network) increases, as shown in the right hand side of Fig. 3.15. This is not surprising because the power of spreading in a network decreases with link density. Real-world networks do not show a significant reduction in number of infected nodes.

**Random Link Removal**

In this segment, we compare the threshold \( \tau_c \) between quarantined networks with networks where the same number of links has been randomly removed. Links are removed at random and the average over many simulations of the largest eigenvalue of the largest connected component is calculated together with the variance of the largest eigenvalue. We measure the difference as the largest eigenvalue of the original graph \( \lambda_{\text{max}} G \), the size of the giant
connected component $\frac{N_{\text{rand.big.comp}}}{N_G}$, its largest eigenvalue $\lambda_{\text{max.rand}}$, the size $\frac{N_{\text{big.clust}}}{N_G}$, and the largest eigenvalue $\lambda_{\text{max.L clustered}}$ of the largest cluster in the clustered network.

The results are presented in Table 3.2. A large part of the network remains connected and can transmit infection, which is an expected result of random link removal. Between 80% and 90% of the network can be affected, compared with at most 50% in case of clustering. Further, the largest eigenvalue of the largest cluster is still smaller than that of the large component in the case of random link removal.

In the USA airline network, ER graphs with $p = 0.002$ and $p = 0.006$, there exists a smaller cluster with a $\lambda_{\text{max.Cluster}}$ larger than the giant component’s. In ER graphs and the political blog, two or more components similar in size have almost the same largest eigenvalue. In the case of the political blog the advantage of clustering over random link removal lies in the fact that the other half of the nodes will not get infected if the clustering is performed before the virus has spread. In the case of the AS Internet topology, the smaller cluster of $N = 3,600$ nodes also has a larger $\lambda_{\text{max.Cluster}}$ than the largest cluster of 6,200 nodes. The Digg network also has smaller cluster of $N = 36,491$ nodes with the largest eigenvalue $\lambda_{\text{max.Cluster}} = 701.61$, while all the rest of the network has significantly smaller largest eigenvalue. In the case of cluster 28s and 49s, two disconnected components have the same largest eigenvalue, which is the same as for random removal.

The variance of the largest eigenvalue for different simulations of random link removal is less than 0.2 in all cases.

**Discussion**

When dividing the network into clusters, a virus can be stopped and annihilated faster than in the original network. However, protection comes at a cost. Shutting down links from the network reduces the communication and reachability of nodes in the network. Assuming that the graph is disconnected only temporally, we calculate the price of quarantine as the number of links that are removed from the graph as a result of a modularity
3.3.5 Section Summary and Conclusions

We have found that, after clustering, real-world networks tend to show a better epidemic threshold \( \tau_c \) than artificially generated graphs. Overall, network protection against epidemics can be improved for any kind of graph. Not surprisingly, the epidemic threshold highly depends on the network topology, at least for all the networks under study. Regarding the network clustering features, an easily clustered graph does not guarantee a slower epidemic threshold, but the way the links intertwine between inter- and intra-

Figure 3.14: Percentage of infected nodes \( y_\infty \) as a function of the effective spreading rate \( \tau \) for the original network and the corresponding clustered network. The epidemic threshold increases as a consequence of having smaller networks, thus reducing the required curing resources.

clustering. As a result of the quarantine, the number of removed links varies from 0.4% to 60%. Most of the considered networks have around 50% of removed links which is a significant fraction. In networks where a small number of links is removed, no significant improvement in the largest eigenvalue and the number of infected nodes is found in the steady-state.

Although the modularity maximization algorithm is on the rise [137], it has not passed a rigorous theoretical examination. The question is also how good the resulting clustering is, as we have not examined other algorithms that may perform differently, because we have concentrated on keeping the communities intact.

The largest eigenvalue improvement using the modularity algorithm is comparable with random links removal for several networks; however, in this case the worm can spread up to 90% of the network.
communities play a major role. However, the number of removed links is, in practice, unacceptably high. The advantages of early quarantine are shadowed by the fact that up to half of the links must be shut down for the quarantine to take effect.

The real-world networks have typically two or three big clusters and several smaller ones, while BA and ER graphs, assumed to model the real-world complex networks, have several smaller ones comparable in size. Thus, in respect of the size of the clusters, BA and ER fail to match real-world networks. This diversity in results appears valuable to create a general classification of types of networks. Up to date, the degree distribution of a given graph has been widely used for this purpose. One could come up with a new network classification, generated by taking the largest eigenvalue of the adjacency matrix of clusters $\lambda_{\text{max, Cluster}}$ vs. links that are removed $L_{\text{out}}$ as an input.

The effects of random link removal led us to conclude that the largest eigenvalue of the largest cluster can be less or comparable to the largest eigenvalue of the biggest component generated by random links removal. However, secondary clusters have a significantly smaller largest eigenvalue, which leads to a smaller amount of cleaning necessary to completely remove the worm from the network. Furthermore, if only the largest cluster is infected, then only up to 50% of the network will need cleaning.

This section combines the concepts of modularity clustering, graph spectra, and epidemic spread, aiming to improve the robustness against the spread of malware. We showed that the interaction between these three variables is not trivial, and it highly depends on the topology at hand. Furthermore, there exists a large number of partitioning algorithms besides modularity, accounting for different definitions of community. The investigation of how different clustering algorithms affect the epidemic dynamics stands on the agenda for future work.
3.4 Measuring Robustness as Algebraic Connectivity

3.4.1 Introduction

In the last decades, there has been a significant advance in understanding the structure and functioning of complex networks [147, 27]. Statistical models of networks are now widely used to describe a broad range of complex systems, from networks of human contacts to interactions amongst proteins. In particular, emerging phenomena of a population of dynamically interacting units has always fascinated humans. Dynamic phenomena are ubiquitous in nature and play a key role within various contexts in sociology [148], and technology [149]. To date, the problem of how the structural properties of a network influences the convergence and stability of its synchronized states has been extensively investigated and discussed, both numerically and theoretically [150, 151, 152, 153, 154], with special attention given to networks of coupled oscillators [155, 156, 157, 158].

In the present work, we focus on the second smallest eigenvalue $\mu_{N-1}$ of a graph’s Laplacian matrix, also called algebraic connectivity. This metric plays an important role on, among others, synchronization of coupled oscillators, network robustness, consensus problems, belief propagation, graph partitioning, and distributed filtering in sensor networks [159, 160, 161, 162, 163]. For example, the time it takes to synchronize Kuramoto oscillators upon any network scales with the inverse of $\mu_{N-1}$ [164, 165, 166, 167]. In other words, larger values of $\mu_{N-1}$ enable synchronization in both discrete and continuous-time systems, even in the presence of transmission delays [168, 169]. As a second application, graphs with “small” algebraic connectivity have a relatively clean bisection, i.e. the smaller $\mu_{N-1}$, the fewer links must be removed to generate a bipartition [170]. Furthermore,

<table>
<thead>
<tr>
<th>Network</th>
<th>$\lambda_{\text{max}G}$</th>
<th>$\frac{N_{\text{rand.bigg.comp}}}{N_G}$ %</th>
<th>$\lambda_{\text{max}\text{rand}}$</th>
<th>$\frac{N_{\text{big.clust}}}{N_G}$ %</th>
<th>$\lambda_{\text{max}1\text{.Clust}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euro</td>
<td>80.92</td>
<td>83.23</td>
<td>53.48</td>
<td>31.99</td>
<td>43.07</td>
</tr>
<tr>
<td>USA</td>
<td>144.61</td>
<td>96.51</td>
<td>118.67</td>
<td>47.54</td>
<td>42.36</td>
</tr>
<tr>
<td>Pol. Blog</td>
<td>74.08</td>
<td>99.01</td>
<td>69.88</td>
<td>51.88</td>
<td>62.11</td>
</tr>
<tr>
<td>AS ’06</td>
<td>71.61</td>
<td>90.59</td>
<td>58.49</td>
<td>27.27</td>
<td>51.22</td>
</tr>
<tr>
<td>Digg</td>
<td>775.33</td>
<td>92.7</td>
<td>582.11</td>
<td>48.13</td>
<td>317.32</td>
</tr>
<tr>
<td>BA 2m</td>
<td>16.09</td>
<td>85.50</td>
<td>12.01</td>
<td>10.08</td>
<td>8.22</td>
</tr>
<tr>
<td>BA 3m</td>
<td>28.11</td>
<td>88.40</td>
<td>20.41</td>
<td>16.30</td>
<td>13.24</td>
</tr>
<tr>
<td>Cluster 0.17s</td>
<td>22.88</td>
<td>100</td>
<td>11.77</td>
<td>37.9</td>
<td>13.17</td>
</tr>
<tr>
<td>Cluster 0.28s</td>
<td>23.51</td>
<td>100</td>
<td>18.41</td>
<td>50.0</td>
<td>18.77</td>
</tr>
<tr>
<td>Cluster 0.49s</td>
<td>25.32</td>
<td>100</td>
<td>25.23</td>
<td>50.00</td>
<td>25.26</td>
</tr>
<tr>
<td>ER 0.002</td>
<td>3.59</td>
<td>83.41</td>
<td>3.29</td>
<td>6.68</td>
<td>2.67</td>
</tr>
<tr>
<td>ER 0.006</td>
<td>7.23</td>
<td>93.2</td>
<td>4.29</td>
<td>13.7</td>
<td>4.03</td>
</tr>
<tr>
<td>ER 0.02</td>
<td>20.93</td>
<td>100</td>
<td>10.05</td>
<td>30.4</td>
<td>10.77</td>
</tr>
</tbody>
</table>

Table 3.2: Comparison of the random links removal strategy with clustering strategy - largest eigenvalue of largest connected component and largest cluster. From left to right, we display the largest eigenvalue of the original graph $\lambda_{\text{max}G}$, the size of the giant connected component $\frac{N_{\text{rand.bigg.comp}}}{N_G}$, its largest eigenvalue $\lambda_{\text{max}\text{rand}}$, the size of the largest cluster $\frac{N_{\text{big.clust}}}{N_G}$, and the largest eigenvalue $\lambda_{\text{max}1\text{.Clust}}$ of the largest cluster.
we illustrate the role of the algebraic connectivity in the diffusion dynamic process. For the sake of simplicity, we model the diffusive dynamics as a commodity exchange governed by the following differential equation:

$$\frac{ds_i}{dt} \approx \sum_{j \in N_i} (s_i(t) - s_j(t)) \approx \sum_{j=1}^{N} Q_{ij} s_j(t);$$ (3.2)

where $s_i$ represents the commodity or the state of the $i$-th component, $N_i$ its neighbors, and $Q$ the Laplacian matrix, as further defined in segment 3.4.2. The equilibrium state is that in which all gradients in (3.2) reach zero, thus the rate of the exponential decay (of the deviation from the equilibrium) is proportional to the algebraic connectivity [171]. Hence, the higher the algebraic connectivity of the $Q$ matrix, the smaller the “proper time”.

Despite the latest advances in the research on synchronization and graph spectra, current research methods mostly focus on individual networks treated as isolated systems. In reality, complex systems are seldom isolated. For example, a power grid and a communication network may strongly depend on each other. A power station depends on a communication node for information, whereas a communication node depends on a power station for electricity [165]; similarly, a pathogen may spread from one species to another. Much effort has been devoted to predict cascading effects in such interdependent networks [172, 173, 174]: the largest connected component has been shown to exhibit a spectacular phase transition after a critical number of faults is reached. Quite recently, a novel approach has been introduced by resorting to the spectral analysis of interdependent networks. By means of the graph spectra, the epidemic thresholds of interdependent networks have been estimated, and absolute boundaries have been provided [172]. These scenarios motivated us to study the influence of interdependent networks on diffusive processes via their spectral properties.

In this work, we show analytically and numerically how the algebraic connectivity of interdependent networks experiences a phase transition upon the addition (or removal) of a sufficient number of interlinks between two identical networks. As a direct consequence,
the proper time of a diffusion process on top of the NoN system is not affected by interlink additions, as long as the number of interlinks is higher than a critical threshold. The location of the described transition depends on the link addition strategy, as well as on the algebraic connectivity of the single networks. Gomez et al. [175] used perturbation theory to approximate lower bounds for $\mu_{N-1}$ in a NoN scenario. We contribute further by characterizing the phase transition of the $\mu_{N-1}$ via mean-field theory. Moreover, we further investigate additional spectral properties, all of which provide evidence for the existence of a transition. In particular, we observe that the transition is reflected in spectral partitioning algorithms, as illustrated in Figure 3.16.

This section is structured as follows. Segment 3.4.2 introduces some required terminology, the Laplacian matrix, and its corresponding spectra. Segments 3.4.3 and 3.4.3 provide some analytical results for the algebraic connectivity of interdependent networks, based on both mean-field approach and perturbation theory, respectively. Our models are able to predict the fraction of links that will cause the algebraic connectivity transition. Finally, Segment 3.4.4 validates our previous results through extensive numerical results. This segment also exposes results on regular, random, small-world, and scale-free networks. Conclusions are drawn in Segment 3.4.5.

3.4.2 Definitions

Graph Theory Basics

A graph $G$ is composed by a set of nodes interconnected by a set of links $G(N, L)$. Suppose one has two networks $G_1 = (N_1, L_1)$ and $G_2 = (N_2, L_2)$, each with a set of nodes $(N_1, N_2)$ and a set of links $(L_1, L_2)$ respectively. For simplicity, in the following we will suppose any dependence relation to be symmetric, i.e. all networks are undirected.

The global system resulting from the connection of the two networks is a network $G$ with $N_1 \cup N_2$ nodes and $L_1 \cup L_2$ ”intralinks” plus a number of ”interlinks” $L_{12}$ joining the two networks; that is $N = N_1 \cup N_2$ and $L = L_1 \cup L_2 \cup L_{12}$, thus $(N, L) = G \overset{def}{=} (N_1 \cup N_2, L_1 \cup L_2 \cup L_{12})$.

Let us denote $N_i$ as the number of nodes in $|N_i|$, and $L_i$ as the number of links in as $|L_i|$, also $N = N_1 + N_2$, and $L = L_1 + L_2$; let $A_1$ and $A_2$ be the adjacency matrices of the two networks $G_1$ and $G_2$, and $A$ that of the whole system $G$, whose entries or elements are $a_{ij} = 1$ if node $i$ is connected to node $j$, otherwise $a_{ij} = 0$. When the two networks are disconnected ($L_{12} = \emptyset$), the matrix $A$ is defined as the $N \times N$ matrix:

$$A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}.$$  \hspace{1cm} (3.3)

When an interaction is introduced ($L_{12} \neq \emptyset$), the adjacency matrix acquires non-trivial off-block terms denoted by $B_{ij}$, defined as the $N_i \times N_j$ interconnection matrix representing the interlinks between $G_1$ and $G_2$. The interdependency matrix $B$ is then

$$B = \begin{bmatrix} 0 & B_{12} \\ B_{12}^T & 0 \end{bmatrix},$$

and adjacency matrix of the total system can be written as:

$$A + \alpha B = \begin{bmatrix} A_1 & \alpha B_{12} \\ \alpha B_{12}^T & A_2 \end{bmatrix}.$$  \hspace{1cm} (3.3)
ROBUSTNESS MEASURED AS A SINGLE METRIC

where $\alpha$ represents the coupling strength of the interaction.

Similarly to the adjacency matrix, one may introduce the Laplacian matrix $Q = D - A$; where $D$ is the diagonal matrix of the degrees, where the degree of the $i$-th node is $d_i \triangleq \sum_j a_{ij}$. In the same vein, one may define the diagonal matrices:

$$
\begin{align*}
(D_1)_{ii} & \triangleq \sum_j (B_{12})_{ij}, \\
(D_2)_{ii} & \triangleq \sum_j (B_{21})_{ij} = \sum_j (B_{12}^T)_{ij};
\end{align*}
$$

and the Laplacian $Q$ of the total system $G$ reads:

$$Q = Q_A + \alpha Q_B = \begin{bmatrix} Q_1 + \alpha D_1 & -\alpha B_{12} \\ -\alpha B_{12}^T & Q_2 + \alpha D_2 \end{bmatrix}. \tag{3.4}
$$

where $Q_1 = Q_2$ is the Laplacian matrix of $A_1 = A_2$, and $Q_B$ is the Laplacian only representing the interlinks:

$$Q_B = D - B = \begin{bmatrix} D_1 & -B_{12} \\ -B_{12}^T & D_2 \end{bmatrix}. \tag{3.5}
$$

**Fiedler Partitioning**

A graph bipartition is defined as two disjoint sets of nodes $\{R, S\}$, where $R \cup S = \mathcal{N}$. In addition, we define the natural partition of $G$ as the bipartition that coincides with the two original node sets: $R = \mathcal{N}_1$, and $S = \mathcal{N}_2$, as illustrated in Fig. 3.17. The number of nodes in $R$ and $S$ is counted by their cardinality $|R|$ and $|S|$, respectively.

Since $Q$ is a non-negative real symmetric matrix, it has $N$ real eigenvalues [171], which we order non-decreasingly $0 = \mu_N \leq \mu_{N-1} \leq \cdots \leq \mu_1$. The eigenvector $x_{N-1}$ corresponding to the first non-zero eigenvalue $\mu_{N-1}$ provides a spectral bipartition named after Fiedler, who derived the majority of its properties [162, 176]. Since this segment only deals with the Fiedler eigenvector, we will simplify the notation of the eigenpair $(\mu_{N-1}, x_{N-1})$ by simply writing $(\mu, x)$. Fiedler partitioning bisects the nodes of $\mathcal{N}$ into two clusters, such that two nodes $i$ and $j$ belong to the same cluster if $x_i x_j > 0$, i.e. the corresponding components of the Fiedler eigenvector $x$ have the same sign. For example, if the coupling strength $\alpha$ in (3.4) is zero, the bipartition resulting from Fiedler partitioning is equivalent to the two natural clusters, i.e. $R = G_1$ and $S = G_2$.

In order to quantify the effect of the Fiedler partition on interdependent networks, we will study the three following spectral partitioning metrics:

- **Fiedler cut-size** $\triangleq \frac{l(R,S)}{L_1 + L_2}$, where $l(R,S) = l(S,R)$ equals the number of links with one end node in $R$ another end node in $S$. The Fiedler cut-size represents the fraction of links with one end in $R$ and another end in $S$ (irrespective of the directionality of the link) over the starting number of links.

- **interdependence angle**, defined as the angle between the normalized Fielder vector $x$ and the vector $x^{(0)} = \frac{1}{\sqrt{N}}(1, \ldots, 1, -1, \ldots, -1)$, which we explain in detail in Appendix B. The interdependence angle is minimized when the Fiedler vector is parallel to $x^{(0)}$, i.e. when the Fiedler partition matches the natural partition.
3.4 MEASURING ROBUSTNESS AS ALGEBRAIC CONNECTIVITY

Figure 3.17: The four main partition sets are displayed: \( N_1 \) (set of black nodes), \( N_2 \) (set of white nodes), \( R \) (set of nodes within the gray rectangle), and \( S \) (set of nodes within the white rectangle). Both, the partition sets and the interlinks (dashed lines) were arbitrary chosen for illustration purposes and do not represent the corresponding Fiedler partition.

- **entropy** \( H \) of the squared Fiedler vector components \( H \overset{def}{=} - \sum_{i=1}^{N} x_i^2 \log x_i^2 \). Based on Shannon’s information theory metric, the entropy indicates how homogeneous the values in \( x \) are, similarly to the participation ratio or vector localization. The higher the entropy, the lower the dispersion among the values in \( x \).

Some partition quality metrics may be undefined if the Laplacian matrix \( Q \) is defective [177]. In particular, if the second and third smallest eigenvalues of \( Q_A + \alpha Q_B \) are equal \( \mu_{N-1} = \mu_{N-2} \) then any linear combination \( x' = ax_{N-1} + bx_{N-2} \) is also an eigenvector of \( Q \) with eigenvalue \( \mu_{N-1} \), thus the Fiedler vector is not uniquely defined. However we will ignore these cases, which tend to occur only in graphs with deterministic structures (e.g. the cycle graph [171]).

3.4.3 Analytical Results

This segment introduces two independent analytical approaches to compute \( \mu \) for the interdependent graph setup described in the previous segment. The two approaches are based on mean-field theory and perturbation theory, which span Segment 3.4.3 and Segment 3.4.3, respectively. For a small number of added interlinks, both of the proposed theories are in agreement with each other, which validates our analysis.

**Exact results for mean-field theory**

**Diagonal interlinking**

Let us start with the case of two identical networks connected by \( L_{12} \) corresponding interlinks, which we refer to as “diagonal interlinking”. We can add as little as 1 link and as many as \( N \) links. This strategy was chosen to achieve the maximum effect by meticulously adding a small number of interlinks.

The mean-field approach to such a system consists in studying a graph of two identical networks interacting via \( N_1 \) weighted connections among all corresponding nodes. The
weight of each link, represented by $\alpha = \frac{L_{12}}{N_1}$, equals the fraction of nodes linked to their corresponding neighbour in the exact network. In other words $B_{12} = I$, such that the synchronization interdependence is modulated by the parameter $\alpha$:

$$Q_A + \alpha Q_B = \begin{bmatrix} Q_1 + \alpha I & -\alpha I \\ -\alpha I & Q_2 + \alpha I \end{bmatrix}. \tag{3.6}$$

In the language of physics, $\alpha$ represents the coupling constant of the interaction between the networks. Consistently with the rest of the segment, this system will also be referred to as the mean-field model of the diagonal interlinking strategy. Regardless of its origin, this system exhibits some interesting properties worth discussing.

Let $\xi_{N_1}, \xi_{N_1-1}, ..., \xi_1$ be the set of eigenvectors for the Laplacian of the single network $A_1$, and $\omega_{N_1}, \omega_{N_1-1}, ..., \omega_1$ be their relative eigenvalues. Since the perturbation $Q_B$ commutes with $Q_A$, all the eigenvectors of the interdependent graph are kept unchanged [171]. All the (unperturbed) eigenvalues are degenerate in pairs and, hence, one may define a set of eigenvectors based on those of the single networks:

$$\begin{cases} x_{2i} = \begin{bmatrix} \xi_i \\ \xi_i \end{bmatrix}, \\ x_{2i+1} = \begin{bmatrix} \xi_i \\ -\xi_i \end{bmatrix}. \end{cases} \tag{3.7}$$

The eigenvalues for the total non-interacting system (i.e. $\alpha = 0$) are the same as those of the unperturbed system $\mu_{2i} = \mu_{2i+1} = \omega_i$, hence, the ascending sequence of eigenvalues for the non-interactive system is $\omega_{N_1} = 0, 0, \omega_{N_1-1}, \omega_{N_1-1}, ..., \omega_1, \omega_1$. When the interaction is switched on (i.e. $\alpha \neq 0$), assuming that $A_1 = A_2$, a trivial proof shows that the even eigenvalues are kept unaltered, while the odd ones increase linearly by the same amount $2\alpha$,

$$\begin{cases} \mu_{2i} = \omega_i, \\ \mu_{2i+1} = \omega_i + 2\alpha. \end{cases} \tag{3.8}$$

For $\alpha$ close to zero, the eigenvector ranking is kept unchanged $\mu_N = \omega_{N_1} = 0, \mu_{N-2} = \omega_{N_1-1}, \mu_{N-3} = \omega_{N_1-1} + 2\alpha, ..., \omega_1, \omega_1 + 2\alpha$. However, when $\alpha > \frac{\omega_{N_1-1}}{2}$ the second and third smallest eigenvalues of the interdependent network ($\mu_{N-1}$ and $\mu_{N-2}$) swap. Therefore, the first non-zero eigenvalue increases linearly with $2\alpha$ up to the value of the isolated networks $\omega_{N_1-1}$ at which it reaches a plateau. In terms of equation (3.2), when $\alpha$ is greater than the threshold $\alpha_I = \frac{\omega_{N_1-1}}{2}$, the interactive system is capable of synchronizing with the same promptness as the single isolated network. Thus when the system coupling channel is quicker than the proper time, the time it takes the interactive system to reach equilibrium equals that of the single network. The critical value $\alpha_I$ for the exact model corresponds to a critical value of links $l_I$ to be included to achieve the promptness of the single network:

$$l_I = \alpha_I N_1 = \frac{\omega_{N_1-1} \cdot \tilde{N}_1}{2}. \tag{3.9}$$

**General interlinking**

As a variation of the localized diagonal interlinking, we have introduced a second strategy that may be treated algebraically; it corresponds to the mean field approximation.
of two identical networks interacting via $L_{12}$ random connections, which we named general interlinking strategy.

The mean-field approach leads to an interdependence matrix with all unitary components: $B_{12} = J$, where $J$ is the all ones matrix; the weight of each interlink is $\alpha = \frac{L_{12}}{N_1^2}$, and

$$Q = Q_A + \alpha Q_B = \begin{bmatrix} Q_1 + \alpha N_1 I & -\alpha J \\ -\alpha J & Q_2 + \alpha N_1 I \end{bmatrix}. \quad (3.10)$$

As in the previous case, the $Q_B$ matrix commutes with $Q_A$, hence a common set of eigenvectors can be chosen (3.7). The null eigenvalue $\mu_N$ is always present, while all the others experience some increase for a non-trivial $\alpha$: all eigenvalues $\mu_i$ for $i$ smaller than $N - 1$, increase for a fixed amount $\alpha N_1$, while $\mu_{N-1}$ increases by twice that quantity,

$$\begin{cases} 
\mu_N = 0, \\
\mu_{N-1} = 2\alpha N_1, \\
\mu_i = \omega_i + \alpha N_1, \text{ for } i \leq N - 1.
\end{cases} \quad (3.11)$$

This different rate of growth again implies that there exists a critical value $\alpha_J$ beyond which the second and third eigenvectors ($\mu_{N-1}$ and $\mu_{N-2}$) swap. The threshold $\alpha_J$ can be obtained by imposing the crossing condition $\mu_{N-1} = \mu_{N-2}$,

$$\alpha_J = \frac{\omega_{N-1}}{N_1}. \quad (3.12)$$

Knowing that $\alpha = \frac{L_{12}}{N_1^2}$, we can estimate the critical number of links for the general interlinking strategy as

$$l_J = \alpha_J N_1^2 = \omega_{N-1} \cdot N_1. \quad (3.13)$$

It is worth noting that, the critical number of interlinks corresponding to the mean-field theory of the diagonal (3.9), and general (3.13) interlink strategies, differ simply by a factor of two.

**Physical interpretation**

If we interpret network robustness as the ability of a system to perform its function upon damage or attacks, then it is worth discussing what happens when two networks, $A_1$ and $A_2$, originally fully connected by diagonal interlinking $B$, are subject to some interlink loss or intentional attacks. Our simple, exact model shows that when these two fully connected networks are subject to minor interlink loss, the response of the total interacting system $A + \alpha B$ takes place at the same speed as the single component network $A_1$. In other words, when the operability of the control channel via $\alpha$ is mildly reduced, the global system synchronizability does not decrease. However if the operability of the connection devices degrades below the critical value $\alpha_J$, the synchronization process starts to slow down. From the mean-field approach point of view, this means that the system may lose a fraction of interlinks while keeping its synchronization time unchanged.

Following the statistical variant, the parameter $\alpha$ can be regarded as a coupling constant or inverse temperature. If one identifies the Fiedler eigenvalue $\mu_{N-1}$ with the internal energy of a thermodynamical system, then its first derivative exhibits a jump
from zero to a finite value. Nevertheless, this derivative does not diverge as expected for a second order\(^2\) transition \cite{178}. On the other hand, if one employs the Fiedler eigenvalue as a metric for the synchronizability and regards it as a thermodynamical potential such as the free enthalpy, its Legendre transform corresponds to the internal energy and exhibits a discontinuity at \(\alpha = \alpha_I\). In this perspective, one may interpret the observed abrupt change as a first order phase transition. Despite this interesting parallel, it is worth noting that the Fiedler eigenvalue and its Legendre transform are not extensive quantities and, hence, they cannot be properly regarded as thermodynamical potentials. However, the behavior of the system closely resembles a phase transition.

To understand the intimate nature of the phase transition, one may inspect the topological properties of the eigenvectors. Below the critical value \(\alpha_I\), the cut links associated to the Fiedler partition are localized outside the original networks (i.e. interlinks are cut), whereas just above the critical value, all cut links lay inside the originally isolated networks (i.e. intralinks are cut). This means that, below \(\alpha_I\), the synchronization is dominated by the intralinks in \(\alpha B\), while beyond \(\alpha_I\) the synchronization involves the whole system, \(A + \alpha B\).

### Approximating \(\mu\) using perturbation theory

In order to validate our mean-field analysis, we conducted a perturbative analysis of the Laplacian spectra. Perturbation theory provides insights on diffusion processes by analytically describing \(\mu_{N-1}\) when the diffusive coupling between \(G_1\) and \(G_2\) is asymptotically small or large.

The problem consists in finding the minimum of the associated quadratic form in the unitary sphere \((x^T x = 1)\), with the constraint \(u^T x = 0\).

\[
\mu = \mu_{N-1} = \inf_{x \neq 0, u^T x = 0} \frac{x^T Q x}{x^T x}; \quad (3.14)
\]

In our case, the matrix \(Q\) is the sum of a matrix \(Q_A\) linking only nodes inside the same net, and a “perturbation” \(\alpha Q_B\) that only connects nodes in different networks \((Q_A + \alpha Q_B)\). Therefore, we want to find the minimum that satisfies the spectral equations:

\[
\begin{cases}
(Q_A + \alpha Q_B - \mu I)x = 0, \\
x^T x = 1, \\
u^T x = 0.
\end{cases} \quad (3.15)
\]

When the solution is analytical in \(\alpha\), one may express \(\mu\) and \(x\) by Taylor expansion as

\[
\mu = \sum_{k=0}^{\infty} \mu^{(k)} \alpha^k \quad (3.16)
\]

\[
x = \sum_{k=0}^{\infty} x^{(k)} \alpha^k \quad (3.17)
\]

Substituting the expansion in the eigenvalue equation (3.14) gives a hierarchy of equations. Solving the latter up to the second order yields the following approximations,

\(^2\)The order of a phase transition is the order of the lowest differential which shows a discontinuity.
3.4.4 Simulations

The previous segment provided two analytical means to estimate the dependence of $\mu_{N-1}$ on the topology of the interdependence links, namely mean field and perturbation theory. In this segment we will introduce four model networks to test the predictability and the limits of our predictions.

Interdependent networks model

Our interdependent network model consists of two main components: a network model for the single networks, and the rules by which the two networks are linked as defined in Segment 3.4.2. In other words, to model two interdependent networks one needs to select two model networks and one interlinking strategy.

In the numerical simulations discussed here, we considered four different graph models for our coupled networks:

Figure 3.18: Algebraic connectivity $\mu_{N-1}$ of four graph models with $N = 1000$ nodes, as simulated (solid lines) and approximated by second order perturbation theory (dashed lines). Interlinks are added between single networks following two strategies: diagonal interlinks (left image) and general interlinks (right image). Perturbation theory best approaches $\mu_{N-1}$ for the parabolic region of the diagonal interlinks strategy, which saturates after adding $\frac{\alpha T N^2}{2}$ links, as we detailed in segment 3.4.3.

$$\mu^{(1)} = \left( x^{(0)} \right)^T Q_B x^{(0)}$$

$$\mu^{(2)} = \left( x^{(0)} \right)^T Q_B \left( x^{(1)} \right) = - \left( x^{(1)} \right)^T Q_A \left( x^{(1)} \right) \leq 0.$$
• **Random Regular (RR):** random configuration model introduced by Bollobás [179]. All nodes are initially assigned a fixed degree \( d_i = k, i \in N \). The \( k \) degree stubs are then randomly interconnected while avoiding self-loops and multiple links.

• **Barabási-Albert (BA):** growth model proposed by Barabási et al. [147] whereby new nodes are attached to \( m \) already existing nodes in a preferential attachment fashion. For large enough values of \( N \), this method ensures the emergence of power-law behavior observed in many real-world networks.

• **Watts-Strogatz (WS):** randomized circular lattice proposed by Watts et al. [27] where all nodes start with a fixed degree \( k \) and are connected to their \( \frac{k}{2} \) immediate neighbors. In a second stage, all existing links are rewired with a small probability \( p \), which produces graphs with low average hopcount yet high clustering coefficient, which mimics the small-world property found in real-world networks.

• **Lattice (LA):** a deterministic three-dimensional grid which loops around its boundaries (i.e. a geometrical torus).

These models exhibit a wide variety of topological features and represent the four different building blocks of \( A_1 \) and \( A_2 \). The input parameters for each model are set such that all graphs have the same number of nodes and links. In addition, all simulated graphs consist of a single connected component, i.e. random graphs containing more than one connected component were discarded.

In addition to the four model networks, we define two strategies to generate the interdependency matrix \( B \), which we solved analytically in Segment 3.4.3:

• **diagonal interlinking** strategy: links are randomly added to the diagonal elements of \( B \), thus linking single network’s analogous nodes.

• **general interlinking** strategy: random links are uniformly added to \( B \) without restrictions, generating a random interconnection pattern.

In the next segments, we explore the effects of the two interlinking strategies on the Fiedler partition of our interdependent model, which are testimony of the phase transition experienced by \( \mu_{N-1} \).

**Diagonal Interlinking Strategy**

*Initial and final states*

We will refer to the natural, initial or unperturbed state as the scenario where there exist no interlink connecting the two networks \( G_1 \) and \( G_2 \). In this scenario, the algebraic connectivity dips to a null value due to the networks being disconnected [171]; the Fiedler partition then becomes undetermined. Nevertheless, the sign of the \( x^{(0)} \) components allows splitting the network into two clusters \( P = G_1 \) and \( Q = G_2 \), corresponding to the two unperturbed networks.

The final state of the diagonal interlink strategy corresponds to \( N_1 \) added interlinks, thus \( B = I \) and the Fiedler vector becomes the vector \( (x(Q_1), x(Q_1)) \), as proved in segment 3.4.3. Therefore, the final partition depends exclusively on \( G_1 \) and \( G_2 \), independently of \( B \). Assuming that both networks are equal \( G_1 = G_2 \), the final cut consists of a subset of intralinks of \( A_1 \), as illustrated in Fig. 3.16c. The critical point for the
3.4 MEASURING ROBUSTNESS AS ALGEBRAIC CONNECTIVITY

Figure 3.19: (Color online) Four metrics’ averages are displayed to evaluate the effect of adding interlinks following the diagonal strategy: algebraic connectivity ($\mu$), Fiedler cut ($l(R, S) / (L_1 + L_2)$), interdependence angle ($\text{acos} \left( \frac{x^T x(0)}{\|x\| \|x(0)\|} \right)$), and entropy ($- \sum_{i=1}^{N} x_i^2 \log x_i^2$). All metrics experience a transition that sharpens for increasing $N$. BA and RR graphs transition around 80% added interlinks, whereas WS and LA graphs transition around 20%. The size of the network $N_1$ has a relatively little impact on BA and RR curves, which suggests that the transition is independent of the network size $N_1$. The flat lines signaled with arrows in the top left plot benchmark the average algebraic connectivity of the $N_1 = 1,000$ respective single networks.

The diagonal strategy is characterized by a transition from an interlink cut to an intralink cut as illustrated in Fig. 64. The transition takes place after adding a set number of links, beyond which $\mu(Q)$ remains constant. In other words, a phase transition occurs when the algebraic connectivity of the interdependent network $\mu(Q)$ reaches the algebraic connectivity of the single networks $\mu(Q_1)$. This transitional effect is clearly visible from the spectral partitioning metrics displayed in Fig. 3.19 and Fig. 3.20. This image shows how the transitional region narrows upon increasing $N$, as it approaches an asymptotic value. Our mean field theory predicts this critical value to be $l_I = \frac{\mu(Q_1) N_1}{2}$, as exposed in Segment 3.4.3.

The precise location of the jump in the simulated experiment, i.e. the critical value
The fluctuation $\sigma^2$ of the Fiedler cut and the interdependence angle are displayed to evaluate the effect of adding interlinks following the diagonal strategy. The narrowing peaks illustrate the sharpening of the transition observed in Fig. 3.19. The narrowing effect is specially noticeable in the Watts-Strogatz model (blue curves).

of interlinks per node $l_I$, depends on the graph model. Regardless, the phase transition is a general phenomenon, which occurrence only depends on the fact that there exists a Fiedler cut for the single networks.

Spectral partitioning metrics

We further investigate the properties of the phase transition, by looking at how the partition metrics displayed Fig. 3.19 evolve as interlinks are added to the $B$ matrix.

The algebraic connectivity (Fig. 3.19a) starts at its minimum value $\approx \frac{1}{2N_1}$ as predicted, which increases almost linearly until it reaches its maximum value $\mu(Q_1)$ when sufficient interlinks are added. This means that a network with 100% diagonal interlinks and the same network with 90% interlinks synchronize virtually at the same speed. Comparing the final values of the algebraic connectivity, it is remarkable that random networks synchronize faster than lattice networks. This is reasonably due to the longer average distance in the latter.

The Fiedler cut (Fig. 3.19b) starts at $\frac{1}{2L_1}$ for a single added interlink. Notice that it increases linearly with the percentage of interlinks, because all added interlinks directly become part of the Fiedler cut. For all networks, we observe a tipping point (which depends on the network type) upon which adding a single link abruptly readjusts the partition: the Fiedler cut switches from pure interlink cutting to a cutting of an invariable set of intralinks.

The interdependence angle (Fig. 3.19c) tells us that the Fiedler vector starts being parallel to the first order approximation $x^{(0)}$ for 1 added interlink. Progressively, the Fiedler vector crawls the $N$-dimensional space up to the transition point, where it abruptly jumps to the final (orthogonal) state $(x(Q_1), x(Q_1))$. Similarly to the interdependence angle, the high values of entropy (Fig. 3.19c) reflect the flatness of $x^{(0)}$, where all components have (almost) the same absolute value. At this initial point, entropy is maximum and almost equal to $\log(2N_1)$, which tells us that the initial partition consists purely of interlinks.
When the partition turns to the final state, the entropy is instantly shaped by the network topologies of $A_1$ thus dropping to relatively much lower values. Notice that, for all values of $N$, the highest final entropy is attained by the lattice graph due to its homogeneous structure, as seen in Fig. 3.19.

**General Interlinks Strategy**

![Graphs showing algebraic connectivity, Fiedler cut, interdependence angle, and entropy](image)

**Figure 3.21:** (Color online) Four metrics’ averages are displayed to qualitatively evaluate the effect of adding interlinks following the general strategy: algebraic connectivity ($\mu$), Fiedler cut ($l(R, S)/L_1 + L_2$), interdependence angle ($\cos(\frac{x^T x^{(0)}}{\|x\|\|x^{(0)}\|}$), and entropy ($-\sum_{i=1}^{N} x_i^2 \log x_i^2$). The four metrics indicate the existence of up to three regimes, but the regime transitions are not as sharp as in the diagonal strategy scenario. The flat lines signaled with arrows in the top left plot represent the average algebraic connectivity of the $N_1 = 1,000$ respective single networks.

Upon the progressive addition of general interlinks, the algebraic connectivity of all models experiences two regimes, as observed in Fig. 3.21a. Initially, for a weak coupling, the algebraic connectivity reaches a minimum as is the case for the diagonal strategy and represents a good starting point for the perturbation theory. If we slightly increase the number of interlinks, the average algebraic connectivity and Fiedler cut curves show a
linear increase. When we reach the critical number of links \( l_J = \mu \cdot N \), the algebraic connectivity \( \mu \) swaps with \( \mu_{N-2} \), causing the slope to decrease by a factor of two, as can be seen from Fig. 3.21a and Fig. 3.21b. These observations are in perfect agreement with our theoretical prediction (3.11). However not only the average, but also the fluctuations steadily increase, which is expected due to the large set of available graph configurations.

As we can see from the interdependence angle in Fig. 3.21c, in the first regime the natural partition is partially preserved up to \( l_J \). At this turning point, the interdependence angle experiences a sharp increase. This is due to the fact that the Fiedler cut in all our isolated model networks scales less than linearly with the network size, which is consistent with the picture of a phase transition between a Fiedler cut dominated by interlinks. As opposed to the diagonal strategy, the final eigenvector of the general strategy is not strictly identical to the Fiedler eigenvector of the isolated networks \( x \), but it also involves interlink cuts. This is due to the fact that in the general case \( x \) does not belong to the kernel of \( Q_B \), as opposed to the diagonal case.

The exact location of the phase transition can also be predicted employing perturbation theory, by imposing the perturbed value \( x(Q) \) of the configuration to be equal to that achieved starting from the \( x(Q_1) \) initial state. However, the resulting formulas are not particularly simple and their numerical calculation requires a time comparable with the Fiedler eigenvalue evaluation of the sparse metrics. For this reasons such estimates are not reported here.

**Network Model Differences**

In this segment we briefly discuss the differences that arise between the four network models. For the diagonal strategy, we observed that RR and BA synchronize relatively faster than deterministic networks. This is due to random interconnections shortening the average hopcount, thus the so called small-world effect [27]) is exhibited. For the particular case of BA, we observe the emergence of a dominant partition which contains approximately 90% of the total number of nodes.

The difference between random and grid networks still exists for the general strategy, but it is not as dominant as in the diagonal case. This effect is expected due to the randomization resulting from the random addition of links to regular structures, which is the conceptual basis of the WS model. In general, we observe that the optimal link addition strategy depends on the network topology.

We observe from Fig. 3.21a that the general strategy synchronizes more slowly than the general strategy in the case of RR and BA. On the other hand, the general strategy synchronizes faster in WS and LA models. Thus if we were to add precisely \( N_1 \) general interlinks between two identical networks, regular structures benefit more than random ones.

To test whether the phase transition is merely an artifact of our synthetic models, additional simulations were carried out using real topologies from a public dataset. Simulation results verify that the transition from the natural partition to the final orthogonal partition also occurs in real networks in the Koblenz Network Collection. The transition in real networks takes place very early in the link addition process, due to the poor synchronization capabilities of networks not designed for such purpose. The positive interpretation of such result is that, to provide a real network with a complete backup mirror without adding diffusive delays, a small number of interlinks are required.
3.4 MEASURING ROBUSTNESS AS ALGEBRAIC CONNECTIVITY

3.4.5 Section Summary and Conclusions

Our contribution beacons a significant starting point to the understanding of diffusion driven dynamics on interdependent networks. Having in mind synchronization applications, this work focuses on the algebraic connectivity of interdependent networks. We provided evidence that upon increasing the number of interlinks between two originally isolated networks, the algebraic connectivity experiences a phase transition. That is, there exists a critical number of diagonal interlinks beyond which any further inclusion does not enhance the algebraic connectivity $\mu_{N-1}$ at all. Similarly, there exists a critical number of general interlinks beyond which algebraic connectivity increments at half of the original rate.

This phase transition of the algebraic connectivity has been observed in diverse interdependent graph models, although the critical transition threshold depends on the topology of the graph models. We analytically derived the critical transition threshold for both, the diagonal and general interlink strategies, by mean field approximations. The transitions occur upon adding: $\frac{\mu(Q_1)\cdot N_1}{2}$ links for the diagonal interlinks strategy, and $\mu(Q_1)\cdot N_1$ links for the general interlinks strategy. Moreover, we show that the critical number of interlinks that trigger the transition can also be estimated via the properties of the NoN’s Fiedler vector, e.g. the entropy.

The study of heterogeneous interdependent networks where the individual networks with different topologies are paired should be considered as future work. Other extensions could encompass additional graph models and broader collections of interlinking strategies.
CHAPTER 4

Metric Correlations

“Everything should be made as simple as possible, but not simpler.”

Albert Einstein, 1879 - 1955

4.1 Introduction

In the last chapter we demonstrated that, depending on the context, robustness can be arbitrarily defined as a function of an arbitrary topological metric, e.g. as hopcount, modularity, or algebraic connectivity. For example, in Section 3.2 we assumed that, in cellphone networks, the average weighted hopcount is strongly related to the packet delay. Thus if we are to minimize the delay, we can minimize the hopcount instead. Similarly, one may conveniently draw parallels between topological metrics and robustness, for as long as one can map a subjective perception of robustness into formulas. Thus, at this point, a definition of network robustness would be complete if I claimed that “robustness can be defined as the (topological) metric that correlates best with the services at hand”.

However, the purpose of this work is not to limit itself to hand-picked scenarios, but to define a generalized framework applicable to any graph. To this end, we note that the simple metric robustness approach presented in the previous chapter suffers two major shortcomings, namely incompleteness and service aggregation.

First, let me remind that a single metric does not provide enough information about a complex structure such as a graph, thus any robustness definition based on a single metric is bound to be insufficient. Traditionally, network types are classified based on only one metric. For instance, the set of scale-free graphs is the set of graphs with a power-law degree distribution, i.e. a degree distribution which asymptotically obeys Pr[X = k] = ck^−β. However, scale-free graphs span a wide variety of heterogeneous graphs, featuring a myriad of topologies with substantially distinct properties. As a consequence, we believe that this limited single-metric approach cannot be generalized
Figure 4.1: Correlation diagram between a small set of metrics \((10^3\) realizations): synchronizability \(\mu_1/\mu_{N-1}\), algebraic connectivity \(\mu_{N-1}\), spectral radius \(\lambda_1\), clustering coefficient \(C\), resistance \(R\), and average hopcount \(E[H]\) of Erdös-Rényi random graphs. Solid lines, dashed lines, and dotted lines express strong correlation, moderate and weak correlation, respectively. For this particular class of graphs, the average hopcount, the spectral radius and the graph resistance are strongly correlated and, thus, redundant. I am very thankful to Li et al.\cite{li2018} for allowing me to publish this diagram of theirs.

To the definition of robustness. The first half of this section addresses this problem by showing the sharp contrast between networks sharing a power-law degree distribution, thus proving that a single metric approaches are robustness oversimplifications.

Second, single-metric approaches cannot be applied to networks supporting aggregated services. Traditionally, data communication networks were built to transport only one type of traffic, namely telephony traffic. But with the latest advances of microelectronics leading to fast computers and high speed networks, new types of traffic were born over the last decade: web traffic, e-mail, P2P, video on demand, and most recently video streaming, which in a sense, fight against each other for the control of the network resources. These dissimilar traffic profiles are all based on different services, each one featuring its own constraints and independent requirements while largely running over the same physical platform. In these aggregated scenarios, we require multiple metrics to strike the combination that accounts for all services. This problem will be addressed later in Chapter 5.

Keeping apart from single-metric approaches and suggesting a multidimensional approach we also bring new problems to the table. One such problem are metric correlations, a non trivial topic still in its infancy. From a geometrical point of view, the ideal scenario is one such that every topological metric represents an orthogonal dimension in a high dimensional space. Thus if we picked a large enough set of \(k\) random metrics, any graph could be uniquely represented by \(k\) metrics. On the other hand, if \(k\) is small, more graphs may possess the same topology vector and the point within the \(k\)–dimensional unit-norm ellipsoid is not necessarily representing a graph in a unique way. In addition, a small \(k\) may “color” the physical meaning of robustness. For example, if we choose the average hopcount and the maximum hopcount, the network robustness basically measures the distances of a graph, independently of other topological features that may impact the service
such as, for example, the connectivity or the algebraic connectivity (see Fig. 4.1). These arguments underline the necessity to choose a sufficiently large set of profiling metrics.

The remainder of the chapter is split into two sections. The first section motivates the need for a multidimensional definition of robustness, by exposing the contrasting properties of Internet topology generators. The second section studies general metric correlations, along with the relation between two prominent metrics: the algebraic connectivity and the link betweenness. Finally we close the chapter with a brief conclusion on metric correlations.

4.2 A Comparison of Degree Generators

4.2.1 The Internet Topology

During the design phase of any networking protocol, networks designers are advised to perform simulations in order to assess the protocol feasibility in terms of performance. If such simulations are to give precise guidance to the protocols and algorithms being studied, then the chosen topologies must model fundamental properties empirically found in the actual existent structure of the network the protocols are designed for, e.g. the Internet. In order to generate accurate artificial topologies, when studying a network one has to assume certain topological characteristics. Probably because its simplicity and efficacy, the most commonly studied metric has been degree distribution, which has been used in Internet protocol simulations since the early 80s [181].

One of the first widely used Internet topology generators was developed by Waxman [181]. The Waxman algorithm is a variant of the Erdős-Rényi random graph [182] based on Euclidean distances. Later research claimed that the real Internet topology does not obey a random-structure, but instead possesses some kind of hierarchy, for instance a differentiation between transit and stub nodes. As a consequence, the structural generators as Transit-Stub [183], Tiers [183], and GT-ITM\(^1\) appeared. These structural generators were considered valid until the appearance of a seminal paper by Faloutsos et al. [104] in 1999. In that paper, the nodal degree of the Internet AS-level topology was shown to closely obey a power law. In particular, the Internet topology at the granularity of Autonomous Systems (AS) is a heavy-tailed scale-free small-world network [184] [185] [104]. However, the question of how closely the Internet topology follows power laws remains open to debate [186] [187]. The graphs generated by the structural generators do not exhibit this power law behavior, turning them into deficient Internet topology models [43]. Thus the work by Faloutsos et al. [104] fueled the development of a new family of generators, such as the Barabási-Albert (BA [74]) and Power Law Random Graph (PLRG [188]). In this section, we focus on a comparison between this last generation of power law algorithms, all of which produce scale-free graphs with power-law degree distributions.

Power-Laws and Scale-freeness are two very important concepts to clarify. A random variable \(X \geq 0\) is said to possess a power law if the probability density function (pdf) obeys

\[
\Pr[X = k] = ck^{-\beta} \quad (k \geq 1, \beta > 1)
\]

\(^1\)http://www3.cc.gatech.edu/projects/gtitm/
where $c$ is a normalization constant, and $\beta$ is the power law exponent. The corresponding power law density function is scale-free because it satisfies

$$f_X(ak) = g(a)f_X(k)$$

(4.2)

An increase by a factor $a$ in the scale or units by which one measures $x$ does not change the overall density $f_X(k)$, except for a multiplicative scaling factor.

Scale-free can be interpreted as a synonym for power law, but most actual usages of "scale-free" appear to have a richer notion. Sometimes additional features are considered such as underlying self-similarity or fractal geometry. In the remainder of this section, we will be using below the notions of scaling and power law distribution interchangeably and only insist that the right tail of the distribution satisfies property (4.1).

A large number of algorithms have appeared trying to emulate the Internet AS-level graph. All the algorithms have an evocative approach to the problem [189] in common: the algorithms can reproduce a metric of interest, but do not capture the underlying causal mechanisms. Usually a well-understood network metric (in most cases the nodal degree distribution) is chosen and an algorithm that matches the metric is developed. This approach presents several problems. First, it is hard to choose the single metric. Second, a method that matches the chosen metric, often does not fit other metrics of interest.

In heavy-tail distributions such as the Lognormal, Weibull and Pareto, the (right) tail decreases sub-exponentially. Thus, the deviation from the mean can vary by orders of magnitude, turning the mean into an uninformative and not representative measure. In general, the higher moments $E[X^k]$ of the distribution function (4.1) only exist if $\beta > k$ [22].

### 4.2.2 Topology Generators

We can classify the topology generators into two families: the curve fitting family and the preferential attachment family. The curve fitting family generators use an explicit scale-free degree distribution $\mathcal{D} = \{d_1, d_2, ..., d_N\}$ (the curve). Given $\mathcal{D}$, the algorithms interconnect the set of $N$ nodes such that the resultant graph $G(N, L)$ with $N$ nodes and $L$ links has degree distribution $D$, as possibly further properties. Generators of this family are PLRG, Havel-Hakimi and Takao. The growth family combines the ideas of network growth and preferential attachment. The graph starts with a small fully connected graph, and divides its growth into time-steps. Every time-step adds one node and $m$ edges. For the added edges, the probability to attach to an existing node is proportional to the degree of the latter. Generators of this family we consider are Barabási-Albert, Barabási-Albert rewire and GLP. Finally, we also consider hybrid generators as inet3 that use both curve fitting and preferential attachment

#### Degree distribution generation

Applying logarithms to the both sides of equation (4.1) we obtain the equation of a line:

$$\log(y) = \log c - \beta \log x$$

(4.3)

The simplest way to obtain $\beta$ is performing a linear regression of $f_X(x)$ when plotted in a log-log scale. To illustrate Eq. (4.3), we plot in Figure 4.2 the degree distribution of PLRG in log-log scale. The data set has been obtained from simulations.
Figure 4.2: Probability density function of PLRG’s degree distribution, $N = 3050$ nodes. The right-hand of the distribution is noisy because of sampling errors. This can be partially solved by increasing the number of simulations.

The frequency distribution plot systematically underestimates $\beta$ [189]. Even worse, the plotted pdfs have a tendency to suggest falsely that a scaling behavior exists. Figure 4.3, and Figures 1a-1d on [189] demonstrate how an exponential distribution can be wrongly interpreted as scale-free.

Figure 4.3 illustrates simulation results for BA-r. The frequency degree data (pdf) appears linear on log-log scale, which leads us to think a priori that the degree distribution behaves like a power law. But if we plot the same data as a ccdf, we can clearly observe how the power law relation fails for high degrees. In conclusion, the rank degree plots are more reliable than the frequency degree plots. Hereforth, we only use rank degree plots.

The performance of the PLRG, Havel-Hakimi and Takao algorithms highly depends on the provided input degree sequence $D = \{d_1, d_2, ..., d_N\}$. Each nodal degree $d_j$ is a random variable with power law distribution (4.1). All the degrees are assumed to be independent, thus ignoring the basic law [22] of the degree $\sum d_j = 2L$ that eventually correlates all nodal degrees. Since the process followed to generate degree sequences is identical for PLRG, Havel and Takao, their probability density functions present the same behavior (while others parameters may differ).

**PLRG**

PLRG stands for Power Law Random Graph [188]. The algorithm first assigns the calculated degree sequence $D$ to the $N$ nodes in the graph. It then randomly matches degrees among all the nodes. The produced graphs may not be connected and may contain self loops and duplicated links. In our simulations, we delete self-loops, merge duplicated links, and extract the giant component. The connected components theory [190][191][192] states that for values of $\beta$ between $2 < \beta < \beta_0$ ($\beta_0 = 3.478$ as derived in [188]) the ran-
Figure 4.3: pdf of BA-r in log-log scale (left) and cpdf of the same distribution in log-log scale (right). The pdf apparently behaves like a power law. The cpdf clearly demonstrates that the sequence is not obeying (4.1), at least for high degrees. Thus the degree sequence is not scale-free.

dom graph has a.s. a giant component, and the size of the second largest component is \(O(\log N)\). For \(1 < \beta < 2\) the second largest component is a.s. of size \(O(1)\).

We use the notation from Cormen et al. [193] to display the algorithm.

**Algorithm 2:** PLRG algorithm.

| Data: Degree vector \(D\) |
| Result: Link set \(L\) |
| \(N \leftarrow \text{length}(D)\) |
| \(L \leftarrow \emptyset\) |
| \(Q \leftarrow \text{empty queue}\) |
| for \(i=1\) to \(N\) do |
| \(\text{for copies}=1\) to \(i\) do |
| push \(D[i] \rightarrow Q\) |
| while \(Q\) not empty do |
| dequeue \(a \leftarrow Q\) |
| dequeue \(b \leftarrow Q\) |
| put \((a, b) \rightarrow L\) |

Our topologies have a size of around 3,000 nodes, meaning that the second largest component size is \(\leq 3\). Empirical results [194] show that 20\% of the total number of nodes do not belong to the giant component.

**Havel-Hakimi**

Similar to PLRG, Havel-Hakimi requires as input a given degree sequence. An important feature of Havel is that its behavior is deterministic: given a degree sequence \(D = \{d_1, d_2, ..., d_N\}\), the resultant graph \(G\) is always the same. The original Havel
algorithm was designed to check whether a degree sequence is graphical\textsuperscript{2} or not. By reversing the original Havel algorithm, we have a deterministic and efficient algorithm to generate a connected graph. Havel algorithm is based on the following theorem.

**Theorem 4.2.1.** (Havel-Hakimi theorem) Let $D = \{d_1, d_2, ..., d_N\}$ be a sequence of non-negative integers with $d_1 \geq d_2 \geq ... \geq d_N$. Let $D'$ be a sequence $\{d'_1, d'_2, ..., d'_N\}$ obtained from $D$ by setting $d'_1 = 0$, $d'_i = d_i - 1$ ($i = 2, ..., d_1 + 1$), and $d'_i = d_i$ ($i = d_1 + 2, ..., N$). Then $D$ is graphical if $D'$ is graphical.

<table>
<thead>
<tr>
<th>Algorithm 3: Determine whether a degree sequence is graphical.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data:</strong> Degree vector $D$</td>
</tr>
<tr>
<td><strong>Result:</strong> Boolean $\text{graphical}$</td>
</tr>
<tr>
<td><strong>SORT</strong>(D)</td>
</tr>
<tr>
<td><strong>while</strong> $\text{POSITIVES}(D) &gt; 0$ <strong>do</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>if</strong> $D = 0$ <strong>then</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>else</strong></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

In words, $D$ is graphical if the following sequence is graphical: replace $d_1$ by 0 and subtract 1 from the next $d_1$ degrees:$\{d_2, d_3, ..., d_{d_1+1}\}$. The recursion ends when the sequence consists of all zeros, or when the sequence has only one non-zero element. The sequence is then classified graphical and non-graphical respectively. The pseudo-code can be found in Algorithm 3, where the function $\text{POSITIVES}(D)$ returns the number of positive elements in the numerical vector $D$.

When a non-graphical sequence is detected the degree sequence is ignored (terminate), and a new one is requested. The Havel algorithm meta code can be found in Algorithm 4, where the function $\text{SORT}(D, M)$ sorts the elements of vector $D$ in non-increasing order, and permutes the elements of vector $M$ accordingly. For example, $\text{SORT}(\{1, 3, 2\}, \{\alpha, \gamma, \beta\})$ returns the two vectors and $\{3, 2, 1\}$ and $\{\gamma, \beta, \alpha\}$.

Algorithm 4 (Havel-Hakimi) exploits the fact that $2D[i]$ degrees are subtracted at every step of Algorithm 3. By adding a link between corresponding subtracted degrees, we obtain a connected graph with degree sequence $D$. In words, at each step we are removing from $D$ the node with the highest degree $h$, and subtracting $d_h$ units from the remaining set of highest degrees.

**Takao**

The Takao algorithm was introduced by Takao Asano in [195]. Similarly to PLRG and Havel, the Takao algorithm requires an explicit degree sequence as input. Takao is a

\textsuperscript{2}A degree sequence $D$ is called graphical if is possible to draw at least one graph with degree sequence $D$. 
deterministic algorithm, such that, for degree sequence $\mathcal{D}$ it will always produce the same output graph.

The main feature of the Takao and Havel algorithms is their speed, in the order of $O(N)$ provided a graphical $\mathcal{D}$. The Takao algorithm first checks whether the given degree sequence is graphical by applying Theorem 4.2.2. Similarly to Havel, the connectivity information is obtained from the processed matrix $M'$, but the connectivity processes are completely different. Takao theorem is defined as follows

**Theorem 4.2.2.** (Takao theorem) Given a degree sequence $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$ of positive integers where $n > d_1 \geq d_2 \geq \ldots \geq d_n > 0$. Let $h = d_n$, $x = \min\{j|d_j = d_h\}$, $y = \max\{j|j \leq n - 1, d_j = d_h\}$. Describe $\mathcal{C} = \{c_1, c_2, \ldots, c_{N-1}\}$ as a sequence of positive integers where $c_1 \geq c_2 \geq \ldots \geq c_{n-1}$, and

$$
c_i = \begin{cases} 
d_i - 1, & \text{if } 1 \leq i \leq x - 1 \text{ or } y - h + x \leq i \leq y \\
d_i, & \text{if } x \leq i \leq y \text{ or } y + 1 \leq i \leq n - 1
\end{cases}
$$

$\mathcal{D}$ is graphical if $\mathcal{C}$ is graphical

In words, at each step we are removing from $\mathcal{D}$ the node with the lowest degree $h$, and subtracting $d_h$ units from the set of highest degrees. This process is the exact opposite of Havel-Hakimi, where the node with the highest degree is removed instead.

**Barabási-Albert**

The two main concepts behind BA algorithm are preferential attachment and Yule’s process. The combination of a growth process with preferential attachment has been proved theoretically [196] and empirically [74] to lead to scale invariant distributions.

The algorithm starts creating a small fully connected core with $m_0$ nodes. Then, BA incrementally constructs a topology by continuously adding nodes. At each time step one node is added with $m$ links. After $t$ time steps, the model leads to a random network with $m_0 + t$ nodes and $mt$ links. The probability that a new node will be connected to the node $i$ is given by

---

**Algorithm 4:** Havel-Hakimi algorithm.

- **Data:** Degree vector $\mathcal{D}$
- **Result:** Link set $\mathcal{L}$
- **for** $i=1$ to $\text{length}(\mathcal{D})$ **do**
  - $M[i] \leftarrow i$
- $\text{SORT}(\mathcal{D})$
- **while** $\text{POSITIVES}(\mathcal{D}) > 0$ **do**
  - **for** $j=1$ to $\mathcal{D}[i]$ **do**
    - $D[j + 1] \leftarrow D[j + 1] - 1$
    - put $(M[1], M[j + 1]) \rightarrow \mathcal{L}$
  - $D[1] \leftarrow 0$
- $\text{SORT}(\mathcal{D}, M)$
4.2 A COMPARISON OF DEGREE GENERATORS

\[ \text{Pr}[X = i] = \frac{d_i(t)}{\sum_{j \neq i} d_j(t)} \] (4.4)

The linear preferential attachment via (4.4) represents the rich-get-richer idea. The probability of a new node attaching to a node with high degree is high, while the probability of a new node attaching to a node with low degree is low. The BA algorithm is described in Algorithm 5.

**Algorithm 5: Barabási-Albert algorithm.**

<table>
<thead>
<tr>
<th>Data:</th>
<th>$N, m_0, m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result:</td>
<td>Link set $\mathcal{L}$</td>
</tr>
<tr>
<td>GENERATE_CLIQUE($m_0$)</td>
<td></td>
</tr>
<tr>
<td>$n \leftarrow m_0$</td>
<td></td>
</tr>
<tr>
<td>while $n &lt; N$ do</td>
<td></td>
</tr>
<tr>
<td>$D \leftarrow$ degree sequence $\mathcal{L}$</td>
<td></td>
</tr>
<tr>
<td>for $j = 1$ to $m$ do</td>
<td></td>
</tr>
<tr>
<td>$r \leftarrow \text{RANDOM_PA}(D)$</td>
<td></td>
</tr>
<tr>
<td>put $(n, r) \rightarrow \mathcal{L}$</td>
<td></td>
</tr>
<tr>
<td>$n \leftarrow n + 1$</td>
<td></td>
</tr>
</tbody>
</table>

The function RANDOM_PA() returns a random node selected through preferential attachment, using Eq.(4.4).

For graphs built using this algorithm, the probability that a node $i$ has $d_i$ edges follows a power law with tail exponent $\beta = 3$, hence the BA model predicts correctly the emergence of power laws. However, for many real systems such as the Internet, $\beta$ is somewhere between 2 and 3. The static $\beta = 3$ provided by BA algorithm is therefore not satisfactory.

The rewiring variant Barabási-Albert rewire (BA-r) [197] separates the growing process into three main events: addition of node, addition of $m$ links, or rewire of $m$ links. Each of these operations is randomly chosen with fixed probabilities $p$ and $q$, where $p + q < 1$. BA-r algorithm is defined in Algorithm 6.

**GLP**

The Generalized Linear Preference (GLP) is a variant of the BA model. GLP adds more flexibility than BA in specifying how nodes connect to other nodes [194]. It has been reported [186] that in the real Internet, new ASes have a much stronger preference to connect to high degree ASs than predicted by the linear preferential model. To achieve a stronger preference for high degree nodes, the probability that a new node will be connected to the node $i$ is adapted as

\[ \text{Pr}[X = i] = \frac{d_i - \theta}{\sum_{j \neq i}(d_j - \theta)} \] (4.5)

where $\theta \in (-\infty, 1)$ is a tunable parameter. When $\theta$ is smaller, less preference is given to high degree nodes. The GLP algorithm is a mix between BA and BA-r. GLP decides at
Algorithm 6: Barabási-Albert rewiring algorithm.

**Data:** $N, m_0, m, p, q$

**Result:** Link set $L$

GENERATE_CLIQUE($m_0$)

$n ← m_0$

while $n < N$ do

$D ←$ degree sequence $L$

rand $←$ RANDOM(0, 1)

// Add a link

if $rand < p$ then

$a ←$ RANDOM_PA($D$)
$b ←$ RANDOM_PA($D$)

put $(a, b) → L$

// Rewire a random link

if $p ≤ rand < p + q$ then

pull $(a, b) ← L$

if $p ≤ rand < p + \frac{q}{2}$ then

put $(a, \text{RANDOM_PA}(D)) → L$

else

put $(b, \text{RANDOM_PA}(D)) → L$

// Add a node with $m$ links

if $p + q ≤ rand$ then

for $j = 1$ to $m$ do

$r ←$ RANDOM_PA($D$)

put $(n, r) → L$

$n ← n + 1$

Each time step between two operations with probability $p$: addition of a node, or addition of $m$ links. Given the similarity between GLP and Algorithm 6, we decided to skip the pseudo-code.

GLP is shown to yield graphs with power law degree distributions. In addition, Bu et al. [194] demonstrate how to choose parameters so as to produce a desired power law exponent. The implementation of BA, BA-r and GLP algorithms have been retrieved from the BRITE project\(^3\).

**Inet3**

Inet3 [198] is based on empirical data extracted from the Oregon Route-Views project\(^4\). Thus the algorithm is based on the BGP Autonomous System (AS) topology.

Inet3 follows the next sequential steps to generate the final topology: first compute the number of months $t$ that would take the 1997’s Internet to reach $N$ nodes (exponential growth of the number of nodes is assumed). Second, compute the new frequency (and

---

\(^3\)http://www.cs.bu.edu/brite/

\(^4\)http://www.routeviews.org
A COMPARISON OF DEGREE GENERATORS

4.2

The degree distribution is calculated through the pdf \( f_X(x) = e^{at+b}x^{-S} \) where \( a \) and \( b \) are known constants (extracted from Oregon Route-Views). For the 2\% higher degree nodes, apply instead the ccdf formula \( F_X(x) = e^{ct+d}x^{-R} \). Third, assign degree 1 to the \( m\% \) of nodes. Fourth, form a spanning tree with nodes of degree higher than 1, creating \( G \). Fifth, attach nodes with degree 1 to \( G \) using linear preferential attachment (4.4). Finally match the remaining nodes with \( G \) using linear preferential attachment.

Inet3 depends on the empirical constants \( a, b, c, d, S, R \) and \( m \). The values for the constants have been extracted from November 1997’s Internet snapshot\(^5\).

4.2.3 Comparing Generators by Simulation

Methodology

All generated networks consist of \( N = 3050 \) nodes. The main reason for this number is that Inet3 requires at least 3037 nodes, which is the number of nodes in 1997 Internet AS topology.

To generate the degree sequences, we choose \( \beta = 2.18 \) as tail exponent. This value of \( \beta \) is in agreement with results for the Internet AS-level topology [194] and [199]. Even though the Internet topology is constantly evolving, the tail exponent \( \beta \) appears to be constant over time [198].

Degree Distribution

Figure 4.4 plots the ccdf in log-log scale of all the degree distributions obtained through simulation. The linear slope in the ccdf plot indicates the obtained distributions follow indeed a power law.

The linear regression has been calculated for both the pdf and ccdf data sets. Additionally, we calculate the Pearson’s \( r \) linear correlation coefficient and the coefficient of variation \( (C_v = \sigma^2/\mu) \). The first estimates the goodness of the linear fits, the second quantifies the grade of variability. All these analytic results are summarized in Table 4.1.

<table>
<thead>
<tr>
<th></th>
<th>pdf ( \beta )</th>
<th>ccdf ( \beta )</th>
<th>PCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLRG</td>
<td>2.07</td>
<td>2.47</td>
<td>0.99</td>
</tr>
<tr>
<td>Havel</td>
<td>2.22</td>
<td>2.05</td>
<td>0.97</td>
</tr>
<tr>
<td>Takao</td>
<td>2.10</td>
<td>2.27</td>
<td>0.99</td>
</tr>
<tr>
<td>Inet3</td>
<td>2.27</td>
<td>2.21</td>
<td>0.99</td>
</tr>
<tr>
<td>BA</td>
<td>2.58</td>
<td>2.96</td>
<td>0.99</td>
</tr>
<tr>
<td>BA-r</td>
<td>1.76</td>
<td>2.40</td>
<td>0.89</td>
</tr>
<tr>
<td>GLP</td>
<td>2.12</td>
<td>2.34</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Table 4.1: Linear fit of the Betas for the pdf and the ccdf graphics. We only show the Pearson’s correlation coefficient (PCC) for the ccdf fit. The correlation coefficient for the pdf is greater than 0.97 in all the cases, which indicates a strong fit.

Figure 4.5 shows the difference between the \( \beta \) extracted from the pdf, and the \( \beta \) extracted from the ccdf. This error decreases if we consider the degrees with small probability

---

\(^5\)So far, the last version of Inet generator is Inet-3.0.
as noise, thus ignoring them (i.e. degrees with $f_X(x) < 10^{-4}$). To avoid this estimation error, the nodes with $f_X(x) < 1\%$ have not been taken in account in the linear fit process of $\beta$ for the pdf.

One may reasonably consider nodes with low-probability degrees to be noise (e.g. $10^{-4}$). If we decide to remove these low-probability nodes, the estimated tail exponent $\beta$ will change accordingly. Table 4.2 assesses the effect of removing high degree nodes (the degrees of which which appear with low probability) on the estimation of $\beta$.

<table>
<thead>
<tr>
<th></th>
<th>$f_X(x)$</th>
<th>$f_X(x) &gt; 10^{-4}$</th>
<th>$f_X(x) &gt; 10^{-3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>1.12</td>
<td>1.71</td>
<td>2.22</td>
</tr>
</tbody>
</table>

Table 4.2: Havel measured tail exponents. The first column shows the tail exponent calculated for the full set of data. The second and third columns show the tail exponent calculated ignoring the degrees with probability less than 1e-4 and 1e-3 respectively.

BA-rewire⁶ generator has a remarkably small $C_V$. Figures 4.3 and 4.4 illustrate that BA-r follows a power law distribution up to $d_i \approx 100$. However, the degree distribution becomes exponential for higher degrees, which explains the relatively small coefficient of variation. This critical exponential behavior cannot be easily observed from the pdf graphics, but it is clearly exposed in the ccdf as an abrupt bend (Figure 4.4).

As mentioned in Section 4.2.2, the PLRG algorithm generates topologies with isolated nodes, duplicated links and self-loops. After these irregularities have been solved (delete isolated nodes, merge duplicated links, delete self-loops) the remaining topology has an

⁶The family of Barabási-Albert generators does not have nodes with degree 1, since during the generation process each new node starts with initial degree $m = m_0 = 2$. 

---

Figure 4.4: Complementary cumulative density plots (ccdf) for the 7 studied algorithms. The behavior of all the algorithms (except BA-rewire) closely follows a power-law, i.e. a straight line.
obvious deficiency in the degree distribution, as several low degree nodes are removed from the graph. Consequently the tail exponent \( \beta \) is affected.

**Assortativity**

Figure 4.6 plots the average degree of the neighbors of a node with degree \( d \) for BA and PLRG. The top of Figure 4.6 illustrates the average degree of the neighbors for BA. The degree of the neighbors is fairly constant, which means that there is a weak correlation between the degrees of a node and that of its neighbors. As observed in Table 4.4, the assortativity of inet3, PLRG, Havel, and Takao is clearly negative, which paints to disassortative behavior. This effect is represented as a negative slope in the down side Figure 4.6, meaning that low degree nodes tend to connect to high degree nodes, and conversely, high degree nodes tend to interconnect with lower degree nodes.

The assortativity coefficient for the Internet AS level graph is found to be \( r_{\text{Internet}} = -0.189 \) [51], showing a significant disassortative behavior.

As an expansion of BA, GLP improves the assortativity results: GLP introduces a significant disassortative behavior.

Empirical data shows that 20% of the PLRG nodes do not belong to the Giant Component, and thus they are erased. As these erased nodes represent a significant fraction of nodes with low degree, subtracting them from the graph implies that the higher degree nodes are losing low degree connectivity. This produces an increase on the average degree of neighbors for the high degree nodes.

Intuitively we might infer that that the assortativity of Havel is be positive, as the algorithm uses the lemma ”Higher degree connect to higher degree”. But the empirical results \( r_{\text{Havel}} = -0.38 \) contradict this interpretation. The essential reason for this
Figure 4.6: Average degree of the neighbors (y-axis) given a random node with degree $d$ (x-axis). The flatter the cloud is, the closer to zero the assortativity is, and the smaller is the relation between the degree of a node and the degree of its neighbors, in which case the degree of a node does not contain any information about the degree of its immediate neighbors. The steeper the cloud is, the higher assortativity, which is an evidence of high degree nodes connecting to low degree nodes.

An apparent conflict is that $r$ is normalized against a set of graphs containing self-loops and isolated nodes. So Havel could be thought of as disassortative when compared with all graphs [189].

**Hopcount**

Figure 4.7 shows the histogram of the hopcount distribution for all the proposed generators. The average hopcount of all the generators lies under 5 hops: the networks have tendency to present a small world graph. Takao is a notable exception, it interconnects long chains of nodes making inappropriate use of the input degree sequence. Table 4.3 presents the simulation results.

Theory [192] [200] states that, for large $N$, that the average hopcount of power law graphs is

$$E[H_N] = 2 \frac{\log \log N}{\log(\beta - 2)}(1 + o(1))$$

yielding for $N = 3050$ and $\beta = 2.18$ that $E[H_N] \approx 2.42$.

The hopcount of the algorithms belonging to the same family of generators present remarkable differences. For instance, BA with mean value of 4.5 lies far from BA-r with mean of 2.4. These difference between generators of the same family implies that the hopcount distribution can be easily altered with simple algorithm modifications.

Havel and BA-rewire show a very narrow distribution that centers around two hops. Havel’s algorithm (Section 4.2.2) systematically interconnects high-degree nodes with...
4.2 A COMPARISON OF DEGREE GENERATORS

Figure 4.7: Hopcount histogram for all the algorithms considered (top, laying horizontally), and histograms of clustering coefficients (bottom, laying vertically); from left to right: BA, BA-r, GLP, inet3, PLRG, Havel and Takao). The top figure gives away the fact that most generators display the small-world phenomena, due to the low average hopcounts given networks with thousands of nodes. From the bottom figure we can see that the BA graph family has relatively low clustering and low variance (left hand side), whereas PLRG and deterministic generators span a wider range of values.
Table 4.3: Theoretical mean, empirical mean, variance, and maximum value of the hopcount for the studied generators (left); mean and standard deviation for the clustering coefficient (right). Takao’s maximum hopcount clearly stands out of the rest. BA theoretical value is not correctly approximated, as the formula is only valid in the range $2 < \beta < 3$.

<table>
<thead>
<tr>
<th></th>
<th>Hopcount</th>
<th>Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>theory $\mu$ $\sigma^2$ max</td>
<td>$\mu$ $\sigma^2$</td>
</tr>
<tr>
<td>PLRG</td>
<td>5.51 3.8 0.83 9</td>
<td>0.139 0.070</td>
</tr>
<tr>
<td>Havel</td>
<td>1.39 2.4 0.54 4</td>
<td>0.698 0.190</td>
</tr>
<tr>
<td>Takao</td>
<td>3.18 7.3 5.19 243</td>
<td>0.687 0.150</td>
</tr>
<tr>
<td>inet3</td>
<td>2.66 3.6 0.72 7</td>
<td>0.126 0.0</td>
</tr>
<tr>
<td>BA</td>
<td>102 4.5 0.77 7</td>
<td>0.012 0.002</td>
</tr>
<tr>
<td>BA-r</td>
<td>4.54 2.9 0.36 5</td>
<td>0.123 0.005</td>
</tr>
<tr>
<td>GLP</td>
<td>3.86 3.6 0.72 7</td>
<td>0.127 0.018</td>
</tr>
</tbody>
</table>

As the average Takao hopcount is too large, this algorithm does not present the small world graph properties. As remarked in [189] a power law degree sequence is not warranty for a small world graph. BA presents a very small clustering coefficient, neither representing a small world. PLRG, BA-r, GLP and Havel follow the small world property, as they combine low average hopcount with high clustering coefficients when compared [22] to the random graph $C_{p(N)} \approx 10^{-3}$.

Spectrum

As pointed by Faloutsos et al. [104], there exists an eigen exponent $\delta$, such that the spectral density decays as a power law (2.41) for large eigenvalues. However, the linear regression is not trivial, as it depends highly on what we consider a large eigenvalue. This problem is illustrated in Figure 4.9.
4.2 A COMPARISON OF DEGREE GENERATORS

There is a rich literature on the eigenvalues of graphs and their relation with topological properties. The eigenvalues of random graphs tend to the semicircle law \cite{105,22}, whereas for scale-free networks the spectrum resembles a bell shape representing a tree-like graph \cite{201}. Moreover, the shape of the spectrum at lower eigenvalues is directly related to the randomness used by the algorithms to generate the topologies. Stochastic topology generating algorithms create graphs that present bulk-shape forms for small eigenvalues \cite{201}. This behavior can be observed for BA (top of Figure 4.8), BA-rewire, GLP and PLRG. On the other hand, the deterministic algorithms Havel (bottom of Figure 4.8) and Takao concentrate their densities in the zero-eigenvalue $\lambda_0$. The Inet3 spectrum is qualitatively between the deterministic and random behavior.

Eq. (2.41) and (2.42), indicate that real Internet AS-level graphs present power law features in spectrum. The generators under study correctly emulate this scaling behavior: the spectrum decays as a power law for large eigenvalues. Figure 4.9 illustrates an example of this fitting process with GLP.

Table 4.4 shows the relation between the tail exponent of the spectrum $\delta$ and the tail exponent of the degree distribution $\beta$ stated in Section 4.2.2. The difference is expressed with $|\Delta| = \delta - 2\beta + 1$.

In the experiments made by Dorogovstev et al. \cite{105}, the slope of the spectrum for BA model is $\delta_{BA} \approx 5$, while in our experiments we obtain $\delta_{BA} \approx 4.8$. The main difference between \cite{105} and ours is the size of the networks.

Dorogovstev et al. \cite{105} used $N = 2 \cdot 10^4$ nodes, whereas we use only $N = 3050$ nodes. We have performed simulations for BA with increasing $N$. Figure 4.10 shows the evolution in the spectrum as we increase $N$ from 200 to 3050. As depicted in Figure 4.10, the large eigenvalues tend to a straight line as $N$ increases. The value where $\delta$ tends to can not be empirically obtained because of network size limitations, but we may expect $\delta_{BA,N\to\infty} \approx 5$ as shown in \cite{105}. This demonstrates that the topologies we are using need a higher $N$ to obtain more precise results, but we still can extract qualitative conclusions.

Table 4.4 shows that BA-r, inet3 and Havel do not behave like scale-free graphs, as the measured $\delta$ do not match Eq. (2.42). The slope of the higher eigenvalues of Takao

Figure 4.8: Left: averaged BA spectrum. The bell shape of the lower eigenvalues reveal a stochastic topology generator. Right: averaged Havel-Hakimi spectrum, which presents very sharp shapes, grouping the 87% of the total density in only 3 eigenvalues.
4.2 Evaluation of Power-law Generators

The topological metrics analyzed here are a grasp of the full range of characteristics that can be computed, but they seem sufficient to differentiate between all the proposed algorithms. The experimental results can be summarized as follows:

- **inet3**, is not reliable. Trying to mimic the power law distribution (using constant parameters), inet3 provides undesirable secondary effects, for instance, the static clustering coefficient.

- **Barabási-Albert**, exemplifies how preferential attachment processes successfully yields power laws. But the lack of input parameters renders it lackluster to generate real-Internet topologies.

- **Barabási-Albert rewire**, the simulations provide a defect in the tail of the degree distribution as it decays clearly exponentially. This invalidates the model.

- **GLP** was designed to improve Barabási-Albert by softening the hopcount distribution, introducing weak disassortative behavior and closer tail exponent in the degree distribution. However, there exist many input parameters, which highly depend on real Internet AS-level data.

- **PLRG** cannot explicitly decide the number of nodes in the topology, due to the randomness of interconnections. However, PLRG has demonstrated to be one of...
A COMPARISON OF DEGREE GENERATORS

Table 4.4: Comparison between the tail exponents of the spectrum ($\delta$) and that of the degree distribution ($\beta$) (left); average assortativity coefficient $r$ of the considered algorithms (right). Dogorostev et al. found that the empirical relation $\delta = 2\beta - 1$ holds for the real Internet AS graph.

<table>
<thead>
<tr>
<th></th>
<th>Tail exponents</th>
<th>Assortativity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta$</td>
<td>$\delta$</td>
</tr>
<tr>
<td>PLRG</td>
<td>2.47</td>
<td>3.95</td>
</tr>
<tr>
<td>Havel</td>
<td>2.05</td>
<td>1.88</td>
</tr>
<tr>
<td>Takao</td>
<td>2.27</td>
<td>3.55</td>
</tr>
<tr>
<td>inet3</td>
<td>2.20</td>
<td>2.39</td>
</tr>
<tr>
<td>BA</td>
<td>2.95</td>
<td>4.78</td>
</tr>
<tr>
<td>BA-r</td>
<td>2.40</td>
<td>1.69</td>
</tr>
<tr>
<td>GLP</td>
<td>2.34</td>
<td>3.55</td>
</tr>
</tbody>
</table>

Figure 4.10: The different lines represent the positive eigenvalues for $N = 200, 400, 800, 1600$ and 3050 respectively. The spectrum tends to a straight line with slope -5, for increasing $N$.  

Dogorostev et al. found that the empirical relation $\delta = 2\beta - 1$ holds for the real Internet AS graph.
the most stable algorithms: none of its features deviates considerably from the (estimated) real Internet AS results.

- **Havel**, its design purpose is not the interconnection of power law degree distributions, but to deterministically interconnect a given degree sequence. Although the results are consistent, the high clustering coefficients and the non-fitting spectrum indicate that the Havel topologies remain still far from real-world networks.

- **Takao**, similarly to Havel, its purpose is not the interconnection of power law degree distributions but only interconnect a given degree sequence. The extremely high hopcount renders this model as extremely unrealistic.

To conclude, all presented Internet topology generators aim to generate an Internet-like topology by mimicking its recently discovered power-law degree distribution. However, extreme differences arise among the generated topologies, thus proving that more than one metric is required to accurately characterize the class of power-law graphs.

### 4.3 Map of Metric Correlations

#### 4.3.1 General Correlations

The correlation among topological metrics strongly depends on the graph under study [202, 180]. Furthermore, there exist extreme scenarios where two metrics always exhibit a high level of correlation regardless of the underlying networks, i.e. the metrics express the same information for all graph types. For example, the average hopcount and the average betweenness (introduced in Section 2.2.1) are always tightly correlated. If \( H_{i \rightarrow j} \) denotes the number of hops in the shortest path from node \( i \) to node \( j \), then the total number of hops \( H_G \) in all shortest paths in \( G \) equals

\[
H_G = \sum_{i=1}^{N} \sum_{j=i+1}^{N} H_{i \rightarrow j}
\]

which also equals

\[
H_G = \sum_{l=1}^{L} B_l
\]

where \( B_l \) is the betweenness the link \( l \) in \( G \). Taking the expectations both expressions and merging them yields the average hopcount in terms of the average link betweenness

\[
E[B_l] = \frac{(N^2)}{L} E[H]
\]

The same reasoning can be followed for the average node betweenness, resulting in

\[
E[B_n] = \frac{N-1}{2} (E[H] - 1)
\]

Formulas (4.6) and (4.7), together with Fig. 4.12 prove that the hopcount and the betweenness are always proportional. However, we shall be careful with our judgment.
The fact that the averages are proportional does not mean that both metrics are always proportional. The correlation is limited to the first moment, i.e. higher moments of $H_i \rightarrow j$ and $B_l$ may (and do) show different behavior.

In addition to the hopcount, the correlation between sets of metrics in the connection metric class was studied by Jamakovic et al. [202, 180]. Using a large sample of real world networks and toy models, a set of topological metrics was computed and correlation coefficients were drawn for each metric combination. Real world networks in particular show a correlation coefficient of 0.81 (or higher) between average node degree, average coreness, and clustering coefficient, and a correlation of 0.72 appears between the algebraic connectivity and the rich club coefficient. The studies conclude that topological metrics tend to be correlated in unpredictable ways, depending on the graph at hand.

We can also find strong correlations among centrality measures, which are those metrics expressing the relative importance of nodes within a graph. Different centrality measures rank nodes differently in order of “importance”. To quantify the similarity of centrality rankings, we define a centrality similarity metric

**Definition.** For two node rankings $A = [a(1), a(2), \ldots, a(N)]$ and $B = [b(1), b(2), \ldots, b(N)]$, $M_{A,B}(k)$ is the percentage of nodes in $\{a(1), a(2), \ldots, a(\lfloor kN \rfloor)\}$ that also appear in $\{b(1), b(2), \ldots, b(\lfloor kN \rfloor)\}$. which is different from the scalar correlation of topological metrics [180]. In words, $M_{A,B}(k)$ gives the percentage of overlapping nodes from the top $k\%$ of nodes in the rankings $A$ and $B$. For instance, it reveals whether the nodes with the highest betweenness values are also those with the highest degrees. The measure $M_{A,B}(k)$ is small when the rankings $A$ and $B$ differ in the nodes that are deemed central. In such cases, both centrality measures should be considered as independent attack strategies, since each draw a relatively different picture of the network centrality map.

The results of $M_{A,B}(k)$ for real-world networks are given in Fig. 4.11. From the previous figure, we observe that $M_{\text{closeness, eigenvector}}(k)$ generally has the highest value and that it is closely followed by $M_{\text{degree, betweenness}}(k)$. On the other hand, $M_{\text{betweenness, eigenvector}}(k)$ shows that there is little overlap between the node rankings derived from the betweenness and eigenvector centrality measures. In both the US and the European power grid networks (Figs. 4.11c and 4.11d), $M_{\text{degree, betweenness}}(k)$ attains large values. On the other hand, in the citation and railway networks (Figs. 4.11a and 4.11b), $M_{\text{closeness, eigenvector}}(k)$ attains large values (i.e. they are strongly correlated).

The voice of experience tells us that, regardless of the topological metrics of choice, we should always expect correlations to be present. As further illustrated by scatter plots in Fig. 4.12, the majority of topological metrics are entangled in a web of correlations. This adverse phenomena arises from the fact that modern metrics were never targeted to build an orthogonal metric space, i.e. a space where each metric independently captures a dimension of the graph. Spectral analysis shines in this regard: all the eigenvectors of a matrix are orthogonal to each other, hence forming the perfect metric space. However, the practical implications of a matrix’s eigenvectors are far from being fully understood. The number of graph metrics has experienced a boom over the last decades [51] [46] [89], as new metrics appear after the need to measure newly observed graph properties.

This tumultuous originating process leads to correlations, and even to the overloading of common terms. For example, clustering coefficient is also known as transitivity in the physics community.
Albeit the algebraic connectivity and the betweenness appear to tell different stories, they are closely related. From a data communication point of view, link betweenness can be seen as a measure of congestion. When a data link with high betweenness is subject to failure, a high percentage of routes are bound to be disrupted, and parts of the network may even become isolated. In the same way, networks with high values of algebraic connectivity become harder to disconnect. At a first glance, the issues of congestion and disconnectivity may seem distant. However, one could argue that betweenness and algebraic connectivity are inversely proportional, because the more bottlenecks a network has, the more likely it is to become disconnected. This intuition proves to be right: Comellas and Rad et al. [203, 40] proved that the algebraic connectivity is lower bounded by the (maximum) link betweenness. This novel bound not only underlines the tight relation between these two metrics, but also proves to be the most accurate lower bound for algebraic connectivity to date.
Figure 4.12: This scatter matrix illustrates all-to-all metric scatter plots, for every existing connected graph with $N = 8$ nodes (a total of 11,117 graphs). The computed metrics are (from top-right to bottom-left): dominance, algebraic connectivity, spectral radius, coreness, clustering coefficient, weighted link betweenness, weighted node betweenness, node betweenness, hopcount and assortativity. Some metrics are strongly correlated, hence the scatter plot resembles a straight line (e.g. hopcount vs. node betweenness); some metrics are weakly correlated (e.g. hopcount vs. clustering coefficient); some metrics are uncorrelated thus the scatter plots resemble random noise (e.g. hopcount vs. dominance).
Alternative Definitions of Betweenness

The betweenness of a link (or node) $B_l$ is defined as the total number of shortest paths that traverse such link (or node) $l$,

$$B_l = \sum_{i \in N} \sum_{j \in N} 1_{\{l \in P(i,j)\}}$$

(4.8)

where the shortest hop path $P(i,j)$ between two nodes $i$ and $j$ is an ordered sequence of links, such that (a) two consecutive links are incident to the same node, and (b) no links are repeated. The node betweenness proposed by Freeman [41] has been heavily used in the past as a centrality measure in social networks.

Although the subset of links $P(i,j)$ is generally defined as the shortest path between the node pair $(i,j)$, the shortest path is not necessarily unique, for there may exist a secondary path with equal hop count $h(P(i,j))$. Hence, the path $P(i,j)$ in (4.8) may not be uniquely defined, causing $B_l$ to be undetermined. We solve this indeterminacy by introducing the set of all shortest paths $S_P(i,j)$

$$S_P(i,j) = \{P_1(i,j), P_2(i,j), \ldots, P_{\varphi(i,j)}(i,j)\}$$

(4.9)

where $\varphi(i,j)$ is the number of different shortest paths between the node pair $(i,j)$. The set $S_P$ allows us to conceive three variations of link betweenness, which we name Full link betweenness $B_l^{(F)}$, Empty link betweenness $B_l^{(E)}$, and Brandes link betweenness $B_l^{(B)}$. The latter is defined as

$$B_l^{(B)} = \frac{1}{2} \sum_{i \in N} \sum_{j \in N} \frac{1}{\varphi(i,j)} \sum_{Z \in S_P(i,j)} 1_{\{l \in Z\}}$$

where $\frac{1}{\varphi(i,j)} \sum_{Z \in S_P(i,j)} 1_{\{l \in Z\}}$ represents the fraction of shortest paths between nodes $i$ and $j$ that cross link $l$. The Brandes Betweenness $B_l^{(B)}$ is the de facto standard for betweenness in most scientific publications. It was first introduced by Freeman [41] and further studied extensively as a node centrality metric in the last decades [39]. The contribution to link $l$’s betweenness of a node pair $(i,j)$ is the proportion of shortest paths that cross the link $l$ over the total number of existing shortest paths $\varphi(i,j)$. This value is then added over all $\left(\begin{array}{c} N \\ 2 \end{array}\right)$ node pairs to find the betweenness $B_l^{(B)}$ of link $l$. The Brandes betweenness $B_l^{(B)}$ can be interpreted as an automatic load balancing, where a communicating node pair $(i,j)$ splits its bandwidth fairly over all shortest paths between $i$ and $j$.

As opposed to the load balancing policy assumed in Brandes betweenness, any node may also choose to transmit their full load through a single path. Given a link $l$, a random path selection may lead to a worst case scenario where all node pairs $\left(\begin{array}{c} N \\ 2 \end{array}\right)$ transmit their full load over $l$. This maximum offered load is quantified by the Full betweenness of link $l$, or $B_l^{(F)}$, which represents the highest possible load that link $l$ may offer (assuming all-to-all node transmission). On the contrary, one may contemplate the opposite scenario for link $l$, where every node pair has chosen an alternate route not including $l$, whenever possible. The Empty betweenness $B_l^{(E)}$ of link $l$ represents the lowest possible load that link $l$ may offer (again, assuming all-to-all node transmission). The Full and Empty betweenness are defined as,
respectively. Where \( \max_{Z \in S_P(i,j)} (1_{\{l \in Z\}}) \) equals 1 if there is at least one path \( Z \) in the set \( S_P(i,j) \) that crosses link \( l \), and \( \min_{Z \in S_P(i,j)} (1_{\{l \in Z\}}) \) equals 1 only if every path in the set \( S_P(i,j) \) crosses link \( l \).

We additionally define weighted-scaled versions of the Full link betweenness \( r_l^{(F)} \), Brandes link betweenness \( r_l^{(B)} \), and Empty link betweenness \( r_l^{(E)} \), as

\[
\begin{align*}
    r_l^{(F)} &= \frac{1}{2} \sum_{i \in N} \sum_{j \in N} h(P(i,j)) \max_{Z \in S_P(i,j)} (1_{\{l \in Z\}}) \\
    r_l^{(B)} &= \frac{1}{2} \sum_{i \in N} \sum_{j \in N} h(P(i,j)) \frac{1}{\varphi(i,j)} \sum_{Z \in S_P(i,j)} (1_{\{l \in Z\}}) \\
    r_l^{(E)} &= \frac{1}{2} \sum_{i \in N} \sum_{j \in N} h(P(i,j)) \min_{Z \in S_P(i,j)} (1_{\{l \in Z\}})
\end{align*}
\]

The weighted link betweenness is closely related to the algebraic connectivity, as we will prove in the next section. By definition, \( r_l^{(F)} \geq r_l^{(B)} \geq r_l^{(E)} \), which are lower bounded by 1, because the node pair at both ends of link \( l \) (i.e. \( (l^+, l^-) \)) always contributes to the Empty link betweenness of \( l \).

**Bounds for the Algebraic Connectivity**

If \( x \) is the eigenvector of the Laplacian matrix \( Q \) belonging to \( \mu_{N-1} \), and \( x(i) \) is the component of \( x \) corresponding to the \( i \)-th node, then [204, p. 81]

\[
\mu_{N-1} = \frac{2N \sum_{l \in \mathcal{L}} (x(l^+) - x(l^-))^2}{\sum_{u \in N} \sum_{v \in N} (x(u) - x(v))^2}
\]  

(4.11)

The bound \( \mu_{N-1} \geq \frac{1}{\rho_N} \), derived in [204, p. 84], can be improved by considering the Cauchy-Schwarz bound [204, p. 84]

\[
(x(v) - x(u))^2 \leq h(P(u,v)) \sum_{l \in P(u,v)} (x(l^+) - x(l^-))^2
\]  

(4.12)

which holds for any randomly chosen shortest path \( P(u,v) \in S_P(u,v) \). However, if we were to sum (4.12) over all node pairs, there will exist a set of particular paths that minimizes (or maximizes) the right hand side of the equation. To avoid this indeterminacy, we generalize the previous expression by considering all \( \varphi(i,j) \) shortest paths joining the node pair \( (u, v) \) by extending (4.12) into

\[
(x(v) - x(u))^2 \leq \frac{1}{\varphi(u,v)} \sum_{Z \in S_P(u,v)} h(Z) \sum_{l \in Z} (x(l^+) - x(l^-))^2
\]  

(4.13)

Summing the previous expression over all node pairs yields
\[
\sum_{u \in N} \sum_{v \in N} (x(u) - x(v))^2 \leq \sum_{u \in N} \sum_{v \in N} \sum_{Z \in S_P(u,v)} h(Z) \sum_{l \in L} (x(l^+) - x(l^-))^2
\]

\[
= \sum_{u \in N} \sum_{v \in N} \frac{h(Z)}{\varphi(u,v)} \sum_{Z \in S_P(u,v)} \sum_{l \in L} (x(l^+) - x(l^-))^2 1_{\{l \in Z\}}
\]

\[
= \sum_{l \in L} (x(l^+) - x(l^-))^2 \sum_{u \in N} \sum_{v \in N} \frac{h(P(u,v))}{\varphi(u,v)} \sum_{Z \in S_P(u,v)} \sum_{l \in Z} (1_{\{l \in Z\}})
\]

Using the definition (4.10) of the weighted Brandes betweenness \( r_i^{(B)} \) we obtain and

\[
\sum_{u \in N} \sum_{v \in N} (x(u) - x(v))^2 \leq 2 \sum_{l \in L} (x(l^+) - x(l^-))^2 r_i^{(B)} \quad (4.14)
\]

Comparing the last equation with Fiedler’s lower bound (4.11), we obtain the final expression,

\[
\mu_{N-1} \geq \frac{N \sum_{l \in L} (x(l^+) - x(l^-))^2}{\sum_{l \in L} (x(l^+) - x(l^-))^2 r_i^{(B)}} \quad (4.15)
\]

which we can simplify (4.15) down to

\[
\mu_{N-1} \geq \frac{N}{\max_{l \in L} r_i^{(B)}} \quad (4.16)
\]

which has already been proved in [40]. In general, the right-hand side in (4.14) can be bounded as

\[
\min_{l \in L} r_i^{(B)} \sum_{l \in L} (x(l^+) - x(l^-))^2 \leq \sum_{l \in L} (x(l^+) - x(l^-))^2 r_i^{(B)} \leq \max_{l \in L} r_i \sum_{l \in L} (x(l^+) - x(l^-))^2 \quad (4.17)
\]

While the minimum \( \sum_{l \in L} (x(l^+) - x(l^-))^2 r_i^{(B)} \) would yield the largest lowerbound for \( \mu_{N-1} \), we cannot guarantee that it equals \( \min_{l \in L} r_i \sum_{l \in L} (x(l^+) - x(l^-))^2 \), nor that \( \frac{N}{\min_{l \in L} r_i^{(B)}} \) will still lower bound \( \mu_{N-1} \). Intuitively though, we may expect that the algebraic connectivity can be expressed as a function of the weighted betweenness,

\[
\mu_{N-1} \approx \frac{N}{f(r_i^{(B)})} \quad (4.18)
\]

The purpose of the present segment is to study the validity of (4.18), and more generally, the relation between \( \mu_{N-1} \) and various measures of the weighted betweenness, such as \( \min_{l \in L} r_i^{(F)} \), \( \max_{l \in L} r_i^{(E)} \), and the average \( \frac{1}{L} \sum_{l=1}^L r_i^{(B)} \).
Simulation Results for Small graphs

In order to test whether our approximation (4.18) holds for the three variations of weighted betweenness, we exhaustively computed $\mu_{N-1}$, $r_i^{(E)}$, $r_i^{(B)}$, and $r_i^{(F)}$ for small graphs, i.e. every existing connected graph up to 10 nodes. The full set of non-isomorphic connected graphs $C(N)$ with $N$ nodes can be swiftly generated through McKay’s algorithm [205]. We chose to exclude disconnected graphs because the algebraic connectivity of such graphs is zero ($\mu_{N-1} = 0$), which is an uninteresting case study. Figure 4.13 and Table 1 display the numerical results obtained by comparing $\mu_{N-1}$ to the following nine inequalities:

- **Empty Betweenness:**
  \[ \mu_{N-1} \geq \frac{N}{\min r_i^{(E)}} \]
  \[ \mu_{N-1} \geq \frac{N}{E[r_i^{(E)}]} \]
  \[ \mu_{N-1} \geq \frac{N}{\max r_i^{(E)}} \]

- **Brandes Betweenness:**
  \[ \mu_{N-1} \geq \frac{N}{\min r_i^{(B)}} \]
  \[ \mu_{N-1} \geq \frac{N}{E[r_i^{(B)}]} \]
  \[ \mu_{N-1} \geq \frac{N}{\max r_i^{(B)}} \]

- **Full Betweenness:**
  \[ \mu_{N-1} \geq \frac{N}{\min r_i^{(F)}} \]
  \[ \mu_{N-1} \geq \frac{N}{E[r_i^{(F)}]} \]
  \[ \mu_{N-1} \geq \frac{N}{\max r_i^{(F)}} \]

where we know that the framed inequalities, i.e. the maximum Brandes and maximum Full betweenness provide inequalities that hold for all graphs in $C(N)$, for any $N$. Thus we will only focus in the remaining seven inequalities. The results of our exhaustive exploration of the graph are summarized in Figure 4.13 and Table 4.5.

The three average values of the three betweenness types provide bounds that only hold up to 70% of the graph space. Interestingly, we see three different behaviors happening. First, the accuracy of the average Empty betweenness declines as $N$ increases, reaching 0.0001% for $N = 10$. On the other hand, the accuracy of the average Full betweenness improves as $N$ increases: it starts at 66% accuracy for all graphs with $N = 5$ nodes, and reaches 94% accuracy for $N = 10$. Lastly, Brandes betweenness shows a double regime as illustrated in Figure 4.13a: for $N < 7$, the accuracy declines for increasing $N$; however, for $N \geq 8$ the accuracy rises for increasing $N$.

Similarly to the average betweenness, the minimum betweenness shows varied behaviors. The accuracy of both the minimum Empty and Brandes betweenness steadily decline, whereas the minimum Full betweenness shows a double regime, i.e decreasing for $N < 8$, and increasing for $N \geq 8$. Surprisingly, the maximum Empty betweenness is outperformed by the average Full betweenness even for small networks. This means that even for small values of $N$, an average-based estimator outperforms a worst-case estimator.

To sum up, the weighted betweenness is not a good approximation for the algebraic connectivity of small graphs. As we have seen, only two out of the seven presented inequalities (the average Full and the maximum empty betweenness) hold for more than 50% of the graph space. Little can be concluded from our simulations; given the limited size of the networks at hand ($N \leq 10$), combinatorial effects may dominate the network dynamics.
Figure 4.13: Number of graphs that obey the indicated inequalities for networks sizes in the range [4,10], both as a percentage and as an absolute number (left image and right image, respectively). The top dashed curve represents $C(N)$, i.e. the total number of non-isomorphic connected graphs with $N$ nodes.
### Absolute values

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td></td>
<td>min</td>
<td>avg</td>
<td>max</td>
<td>min</td>
<td>avg</td>
<td>max</td>
<td>min</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>3</td>
<td>4</td>
<td>13</td>
<td>5</td>
<td>9</td>
<td>21</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>112</td>
<td>2</td>
<td>5</td>
<td>69</td>
<td>10</td>
<td>38</td>
<td>112</td>
<td>24</td>
</tr>
<tr>
<td>7</td>
<td>853</td>
<td>4</td>
<td>14</td>
<td>563</td>
<td>25</td>
<td>202</td>
<td>853</td>
<td>143</td>
</tr>
<tr>
<td>8</td>
<td>11117</td>
<td>5</td>
<td>25</td>
<td>7098</td>
<td>76</td>
<td>2625</td>
<td>11117</td>
<td>1655</td>
</tr>
<tr>
<td>9</td>
<td>261080</td>
<td>6</td>
<td>79</td>
<td>160615</td>
<td>426</td>
<td>68487</td>
<td>261080</td>
<td>43901</td>
</tr>
<tr>
<td>10</td>
<td>11.71e6</td>
<td>13</td>
<td>310</td>
<td>6.79e6</td>
<td>4831</td>
<td>3.70e6</td>
<td>11.71e6</td>
<td>2.76e6</td>
</tr>
</tbody>
</table>

### Relative values

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td></td>
<td>min</td>
<td>avg</td>
<td>max</td>
<td>min</td>
<td>avg</td>
<td>max</td>
<td>min</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>4</td>
<td>100%</td>
<td>33.33%</td>
<td>33.33%</td>
<td>66.66%</td>
<td>50%</td>
<td>50%</td>
<td>100%</td>
<td>50%</td>
</tr>
<tr>
<td>5</td>
<td>100%</td>
<td>14.28%</td>
<td>19.04%</td>
<td>61.90%</td>
<td>23.80%</td>
<td>42.85%</td>
<td>100%</td>
<td>28.57%</td>
</tr>
<tr>
<td>6</td>
<td>100%</td>
<td>1.78%</td>
<td>4.46%</td>
<td>61.60%</td>
<td>8.92%</td>
<td>33.92%</td>
<td>100%</td>
<td>21.42%</td>
</tr>
<tr>
<td>7</td>
<td>100%</td>
<td>0.46%</td>
<td>1.64%</td>
<td>66.00%</td>
<td>2.93%</td>
<td>23.68%</td>
<td>100%</td>
<td>16.76%</td>
</tr>
<tr>
<td>8</td>
<td>100%</td>
<td>0.044%</td>
<td>0.22%</td>
<td>63.84%</td>
<td>0.68%</td>
<td>23.61%</td>
<td>100%</td>
<td>14.88%</td>
</tr>
<tr>
<td>9</td>
<td>100%</td>
<td>0.0022%</td>
<td>0.030%</td>
<td>61.51%</td>
<td>0.16%</td>
<td>26.23%</td>
<td>100%</td>
<td>16.81%</td>
</tr>
<tr>
<td>10</td>
<td>100%</td>
<td>0.00011%</td>
<td>0.0026%</td>
<td>58.03%</td>
<td>0.041%</td>
<td>31.64%</td>
<td>100%</td>
<td>23.60%</td>
</tr>
</tbody>
</table>

**Table 4.5:** For an increasing number of nodes, we display the number of graphs that obey the five inequalities as absolute (top) and fractions (bottom) over the total graph space —$C(N)$—. The percentages of complying graphs show a decrease for initial values of $N$. However, after $N$ grows larger than 9 nodes the percentage of complying graph starts increasing, hinting the existence of two regimes: small ($N<9$) and large ($N\geq 9$) graphs. The causes behind such behavior are under investigation.
4.4 Chapter Summary and Conclusions

In this chapter we have discovered that the minimum degree $d_{\text{min}}$ and the algebraic connectivity $\mu_{N-1}$ are correlated (via $0 \leq \mu_{N-1} \leq d_{\text{min}}$), the average hopcount equals the average betweenness, and the weighted betweenness can be lower bounded by the algebraic connectivity $N/\mu_{N-1}$. The list goes on, thus showing that random pairs of topological metrics are likely to be correlated, even when their definitions are distant. The higher the correlation between two metrics is, the more their topology vectors are aligned, and the less information is reflected causing a loss of information. Defining a minimum set of non-correlated metrics in order to quantify robustness proves to be a complex task, which requires a full understanding of the metrics at hand.

The stated multidimensional problem begs the question of how many metrics are necessary to fully characterize a graph. Or equivalently, how many metrics are necessary to fully characterize robustness: one metric? two? three? $N/2$? $N$? Although the number of necessary metrics is not yet known, we hope this chapter illustrated that the answer is not “one”, by exposing non-trivial correlations among randomly chosen metrics: the degree distribution, centrality metrics, algebraic connectivity and betweenness.
CHAPTER 5

A Robustness Framework

“But to us [finite beings], probability is the very guide to life.”

Joseph Butler, 1692 - 1752

5.1 Introduction

5.1.1 A Robustness Framework

The main purpose of a robustness framework is to propose a methodology to uniquely quantify the robustness level of a network, ideally as a single value. This robustness value allows the production of benchmarks and robustness classes, which can be used by network designers to enhance the quality of their networks, as shown in Fig. 5.1. The main benefit of defining robustness as a number (referred to as R-value for the remainder of this thesis) is that the current R-value of a network can be compared to a minimal desired R-value, $R_{\text{thresh}}$. Either the R-value is sufficient, in which case the network conforms to the benchmark, or the R-value is too low, in which case a modification to improve the graph is required. By reiterating the previous process, we can design efficient “possibly optimal” strategies how a graph can be modified to increase its R-value, subject to some cost criterion. Alas, to date, there does not seem to exist a commonly agreed definition of network robustness, nor a framework to modify a network in order to meet some desired level of robustness. As mentioned, the goal of this thesis is to present such a definition, though a framework that establishes a definition of robustness by means of graph theory. Thus far, this thesis has already introduced the concepts of graph theory, single-metric approaches, and metric correlations. All these tools allow us to tackle the concept of robustness rigorously.

This framework approach crucially assumes that we have the possibility to alter the topology of the network. When the network topology cannot be changed, the robustness could be improved by installing proper functionality at network nodes as discussed in
Figure 5.1: Organigram of the high level goal to achieve network robustness. The desired robustness level (specified by a threshold) is achieved via a feedback loop, which is able to quantify robustness via the so called R-value. Note that a network is fully specified by both, its topology and its services.

[120] by implementing, for example, a $D^2R^2 + DR$ methodology. Some networks, such as ad-hoc or adaptive and growing networks (e.g. social interaction, biological living), have a flexible topology that varies over time, which can be freely modified.

Network robustness, or resilience, has already been approached by a number of researchers in the past, i.e. this thesis is not aiming to reinvent the wheel. Yet, the lack of a common vocabulary has made cooperation difficult, as several terms related to robustness have been proposed over the last fifty years, including reliability, safety, maintainability, dependability and degree-distribution entropy [4, 5]. Meyer [6] studied robustness in the context of his performability framework [7], whilst Cholda et al. [8] surveyed various robustness frameworks. In previous research [206, 207, 208], maintenance of connectivity under failure has typically been used to characterize network robustness. Connectivity has been studied from a probabilistic point of view in the context of graph percolation [1, 2] and reliability polynomials [3]. Most probabilistic studies assume that link failures are independent and that failures occur with the same, fixed probability.

Since the behaviors of topological metrics depend on the characteristics of the networks to which they are applied, robustness profiles based on these metrics also depend on the network under study. In the wake of this observation, researchers have studied robustness in the context of various network types. For example, Callaway et al. [9] and Holme et al. [10] have studied the robustness of random networks and power-law graphs. In particular, Cohen et al. have examined the robustness of the Internet and other power-law networks under random [11] and targeted [12] failures. Recently, the robustness of time-evolving networks or temporal graphs [13, 14] has been researched in [15, 16]. A method based on the cumulative change of the giant component under targeted attacks has been proposed by Schneider et al. [17]. Cetinkaya et al. [18] developed a framework for analyzing packet loss relative to node and link failure. They consider packet loss under global targeted and random failure, as well as attacks contained within geographic regions. Our approach is similar to theirs, although we consider not only average network
performance under random attacks but the density function given the probability that a metric will assume a given value after a given fraction of node removals.

5.1.2 A Novel Framework

Before resorting to formulas, let us start by introducing a definition of what robustness stands for. In this thesis, we understand robustness to be the degree of the network’s capability to withstand perturbations during a given time interval. First and foremost, this definition of robustness mentions an overlooked concept, i.e. the concept of failures, or challenges. A challenge is an event occurring at time $t$ that changes the network $G$ by either

- adding a node to $G$
- removing a node from $G$
- adding a link to $G$
- removing a link from $G$
- in weighted graphs, changing the link (or/and node) weight

or any combination of them. Network failures are caused by unintentional failures or intentional attacks. Unintentional failures include human error, manufacturing defects and worn-out mechanical parts. These kinds of failures appear randomly and are characterized as random breakdowns [11, 15]. Intentional attacks, on the other hand, are not random and are aimed at maximizing damage. In the literature, they are known as targeted attacks [10, 209, 12]. Our framework takes into account this challenge dimension, by modeling challenges as the so called robustness envelopes. In short, robustness envelopes are probability density functions of a set of two variables: a topological metric, and a set of challenges.

As we already exposed in Chapter 4, the two main problems that arise from using a multi-metric approach are metric correlations and service aggregation. Metric correlations constitute a persistent difficulty, and should always be taken into consideration by further studying metric correlations. To this end, Chapter 4 offered a glimpse into well known correlations. We solve the second problem by modeling service aggregation as a network possessing two crucial “features”: a network topology or infrastructure and a service for which the network is designed or created. These two features are explained as follows.

The network topology solely specifies how items, called nodes, are interconnected or related to other nodes by links. The interconnection pattern can be represented by a graph $G$, consisting of $N$ nodes and $L$ links. Each link in $G$ can be further specified by a set of link weights that reflect attributes such as delay, packet loss, available capacity, monetary cost, distance, etc. On top of this topology there exists at least one network service, which is more abstract and less clearly defined. In general, a service uses the network infrastructure to transport items between a group of nodes, possibly subject to some constraints. Secondly, a service is specified by a protocol, consisting of the application (software executed at source and destination nodes) and of a network communication “engine”. For example, in the Internet, a communication service such as email transports a message from a source node to a destination over the network topology. Other examples
of services in complex networks are road transport, neuron transport in the brain, financial transactions on a stock market, news spreading in social networks, etc. Both the topology and the service are usually time-variant and may have their own specific properties and requirements. Although a service is often designed independently of the topology, the end-to-end behavior of a service is influenced by the topology and a topology is most often designed to offer a certain service. In addition, many services can share the same underlying topology, as in the Internet (file transfer, email, web services, etc.). Thus, in general, the duo of service and topology are not necessarily operating in some optimal way. Our definition of network robustness takes both “planes” - topology and service - into account, as illustrated in Fig. 5.1.

The remainder of this chapter is structured as follows. Section 5.2 presents a formal definition for the concepts of service aggregation and the challenge dimension, by means of linear combinations and robustness envelopes, respectively. Section 5.3 applies our framework to a small set of real networks and toy models, and proposes a possible way to optimize robustness via assortativity optimization in Section 5.4. Finally, Section 5.5 closes the chapter.

5.2 Envelope Computation and Comparison

5.2.1 Robustness and the $R$-value

We believe that a network is only fully defined when, in addition to the topological structure, multiple protocol dynamics and constraints are set on top of the graph. This multilayer nature of a network we introduced in Chapter 4 leads to a natural metric classification split in topological metrics, and service metrics. The first class groups network properties obtainable by processing topological information, leading to an understanding on the connectivity of the network elements. Further, service metrics attempt to model the dynamic processes (such as end-to-end delivery, virus spread, etc.) present in the network. This differentiation between low and high level metric introduces the concepts of service and decomposability, which we further discuss in Section 5.

In order to bring the service dimension into our definition of robustness, we propose to compute the R-value of the network robustness by the following weighted, linear norm

$$ R = \sum_{k=1}^{m} s_k t_k $$

where $s$ and $t$ are the $m \times 1$ weight and topology vectors, respectively. The components of the topology vector $t$ are $m$ graph metrics that characterize the topology/graph. For example, $t_1$ may represent the average hopcount, $t_2$ the minimum degree, $t_3$ the assortativity, $t_4$ the algebraic connectivity, and so on. The components of the weight vector $s$ reflect the importance of the corresponding topological metrics for the service. For example, a real-time communication requires certain end-to-end delay bounds. The amount to which metrics influence the end-to-end delay, such as e.g. the average hopcount and the betweenness, is reflected by the value of the corresponding component of $s$. In most cases, it is difficult to determine the numerical values of the components $s_k$ based on a service. A simple example, that circumvents the complications induced by the service, considers the weight vector as a zero-one vector: if $s_k = 0$, then the topology metric $t_k$
is not relevant for the service, while the opposite holds if $s_k = 1$. By this confinement, the R-value computation is greatly simplified and (5.1) can be computed for any graph, provided that the association of the topology metric $t_k$ with the service can be made.

In addition, we normalize $R$ to the interval $[0, 1]$. Thus, $R = 0$ corresponds to absence of network robustness and $R = 1$ reflects complete robustness. If $\| \cdot \|_q$ denotes a $q$-norm [22, p.445] defined as $\| x \|_q^q = \sum_{k=1}^{m} x_k^q$ then the unnormalized R-value, denoted by $\tilde{R}$ is

$$ \tilde{R} \leq \| s \|_q \cdot \| t \|_q $$

from which normalization follows as

$$ 0 \leq R = \frac{|s^T t|}{\| s \|_q \cdot \| t \|_q} \leq 1 $$

Since every components of the topology vector reflects a different topological metric, the units are different as well as the range of the possible values. This will necessitate additional per component (per metric) normalization.

We have defined robustness as the maintenance of function under node or link removal. In this context, function is measured by one or more graph metrics. As in [210], we express robustness as a real-valued function $R$ of graph metrics\(^1\), normalized to the range $[0, 1]$. A value of $R = 0$ means that the network is completely non-functional, whereas $R = 1$ means that the network is fully functional.

In this chapter, we consider two different $R$-values, computed equating $R$ to the 1) size of the giant component and 2) efficiency. This choice of metrics is arbitrary and it boils down to the network function we are optimizing. The presented method translates easily to different sets of metrics.

1) **Size of the giant component.** The number of nodes in the largest connected component of a network. This metric is a measure of the global connectivity of the network.

2) **Efficiency.** The efficiency [211] of a given network $G$ is the mean of the reciprocals of all the hopcounts in a network

$$ E[1/H] = \sum_{1 \leq i < j \leq N} \frac{1}{H_{i,j}}. $$

The hopcount $H_{i,j}$ is the number of links in the shortest path from node $i$ to node $j$. If there is no path from $i$ to $j$, $H_{i,j} = \infty$ and $1/H_{i,j} = 0$. We consider this metric to give an indication of how quickly information spreads through a network. When $E[1/H] = 0$, the network is completely disconnected and when $E[1/H] = 1$, it is fully connected.

5.2.2 **Network perturbations or challenges**

A perturbation or challenge $P$ is defined as a set of elementary changes [210]. Elementary changes include: (1) addition of a node, (2) removal of a node, (3) addition of a link, (4)

\(^1\)Not to be confused with the graph resistance $R(G)$. By the context, it should be clear whether $R$ is referencing Robustness or Resistance.
removal of a link and, (5) in weighted networks, a change in the weight of a link (or node). We consider only node removals, but our analysis can be extended to all five perturbation types. A realization is a vector \([P_1, P_2, \ldots, P_N]\) of perturbations, where \(P_i\) is a subset of \(i\) nodes. In addition, a realization is called successive iff \(P_1 \subset P_2 \subset \ldots \subset P_N\). Since every perturbation has an associated \(R\)-value, any realization can also be expressed as a sequence of \(R\)-values denoted \(\{R[k]\}_{0 \leq k \leq 1}\), where \(k\) is the fraction of removed nodes.

### 5.2.3 Random failures and targeted attacks

Network perturbations are classified either as random (un-intentional) failures [11] or as targeted attacks [10, 209].

**Random attacks**

Assuming that the nature of the attacks is unknown and attacks occur independently, \(R[k]\) is a random variable. We employ probability density function (PDF), which is the probability of a random variable to fall within a particular region. The PDF of this \(R[k]\) is computed using all subsets of \([kN]\) nodes of the set \(P_r\) of all possible perturbations.

The envelope for a graph \(G\) is constructed using all \(R[k]\) for \(k \in \{\frac{1}{N}, \frac{2}{N}, \ldots, 1\}\), where boundaries are given by the extreme \(R\)-values

\[
R_{\text{min}}^{(P_r)}[k] = [\min(R[\frac{1}{N}]), \min(R[\frac{2}{N}]), \ldots, \min(R[1])] \\
R_{\text{max}}^{(P_r)}[k] = [\max(R[\frac{1}{N}]), \max(R[\frac{2}{N}]), \ldots, \max(R[1])]
\]

Such boundaries can be seen in Fig. 5.2. Although extreme \(R\)-values give the best- and worst-case metrics for a network after a given number of perturbations, we are just as often interested in the expected \(R\)-value resulting from \(k\) perturbations

\[
R_{\text{avg}}^{(P_r)}[k] = [E[R[\frac{1}{N}]], E[R[\frac{2}{N}]], \ldots, E[R[1]]]
\]

Finally, since \(R[k]\) defines a PDF, we are also interested in the percentile lines of \(R[k]\), since they enable one to calculate contours that describe the robustness for a given percentage of perturbations

\[
R_{m\%}^{(P_r)}[k] = [R_{m\%}[\frac{1}{N}], R_{m\%}[\frac{2}{N}], \ldots, R_{m\%}[1]]
\]

where \(R_{m\%}[k]\) are the points at which the cumulative distribution of \(R[k]\) crosses \(m/100\), namely

\[
R_{m\%}[k] = t \Leftrightarrow \Pr[R[k] \leq t] = \frac{m}{100}
\]

We refer to \(R_{m\%}[k]\) as an \(m\)-percentile. By definition \(R_{0\%}[k] = R_{\text{min}}[k]\), and \(R_{100\%}[k] = R_{\text{max}}[k]\). The dark-gray areas in Fig. 5.2a are bounded by low-percentile lines whereas the lighter-gray areas correspond to higher-percentile lines.

In the case where \([kN]\) nodes in the network are attacked, \((\frac{N}{kN})\) \(R\)-values need to be computed. It has been shown that the problem of finding a set of nodes minimizing \(R[k]\) is NP-complete [212]. For this reason, we perform random sampling to approximate the PDF of \(R[k]\) and targeted attacks to approximate the maxima and minima of the PDFs.
Targeted attacks

Targeted attacks are perturbations involving vulnerable nodes or links. In order to determine node vulnerability, the attacker must have some knowledge of the topology of the network under attack. For simplicity, we assume that the nodes are ranked once by the attacker in order from most vulnerable (most important) to least vulnerable (least important) and are attacked in that order.

Centrality measures may provide a set of such rankings. We consider five different measures: (a) node degree; (b) betweenness \[213\]; (c) closeness \[214\] and (d) eigenvector centrality \[37\]. In Chapter 4 we studied the extent to which these rankings overlap.

For each of the five centrality measures and for each graph \(G\), we may obtain two successive realizations: a top realization \(\{R_G^{(P_{\text{top}})}[k]\}_{0 \leq k \leq 1}\) resulting from a perturbation \(P_{\text{top}}\) targeting the highest ranking \(k\) nodes of centrality ordered list, and a bottom realization \(\{R_G^{(P_{\text{bot}})}[k]\}_{0 \leq k \leq 1}\) resulting from a perturbation \(P_{\text{bot}}\) targeting the lowest \([kN]\) ranked nodes.

**Figure 5.2:** Depictions of robustness envelopes. The y-axis represents the topological metric, and the x-axis represents the fraction of attacked elements. The colored areas represent regions where random events fall with equal probability. Darker zones represent high event density, and bright zones represent low even density. In other words, the envelopes are sets of contour lines, similar to the ones we see in topographical elevation maps.
5.2.4 Comparison of networks via envelopes

Suppose that the same perturbation sequence $\mathcal{P}$ is applied to two graphs $G_1$ and $G_2$ and that the impact of a single perturbation is measured via the metric $R$. The $R$-values at step $k$ are denoted $R_{G_1}^{(P)}[k]$ and $R_{G_2}^{(P)}[k]$ respectively. In the simple case where $G_1$ and $G_2$ have the same number of nodes and $R_{G_1}^{(P)}[k] > R_{G_2}^{(P)}[k]$ for all $k$, it is clear that $G_1$ is more robust than $G_2$ with respect to $\mathcal{P}$. But such cases are rare and we propose two simple metrics for comparing the robustness of different sized networks: the energy $\mathcal{E}$, and the sensitivity $\mathcal{S}$.

The energy $\mathcal{E}$ of a graph is the normalized sum of the average $R$-values over all random perturbations or in the case of targeted attacks, the normalized sum of the $R$-values

$$\mathcal{E}^{(P)} = \frac{1}{K} \sum_{k=1}^{K} R^{(P)}[k]$$

(5.2)

where $K = |\mathcal{P}|$. Energy expresses how robust, on average, a graph is against a given type of attack. For instance, if $\mathcal{E}^{(P)}_{G_1} > \mathcal{E}^{(P)}_{G_2}$, $G_1$ has higher energy than $G_2$ with respect to the perturbation $\mathcal{P}$. Other examples of energy include those computed from the maximal realization $\mathcal{E}^{(P)}_{\text{max}}$, minimal realization $\mathcal{E}^{(P)}_{\text{min}}$, expected realization $\mathcal{E}^{(P)}_{\text{avg}}$, and $m$-percentile realization $\mathcal{E}^{(P)}_{m\%}$, as illustrated in Figs. 5.2b-5.2c.

The sensitivity $\mathcal{S}$ is defined as the energy increment between the 80-percentile and 20-percentile realizations

$$\mathcal{S}^{(P)} = \mathcal{E}^{(P)}_{80\%} - \mathcal{E}^{(P)}_{20\%}$$

(5.3)

The sensitivity $\mathcal{S}$ indicates how likely the $R$-value is to shift upon random removals, as illustrated in Fig. 5.2d. The smaller the sensitivity, the narrower the uncertainty of the $R$-value, thus the better the robustness. The sensitivity together with the percentiles of $R$-values express the variability of different random attacks in a given network.

5.3 Robustness of Random and Real Networks

In this section, we study the properties of a variety of random network models and real-world networks under random and targeted attacks. We expect different behaviors for different types of networks, leading to a classification of networks based on their energy and sensitivity characteristics.

We consider four network models with different structural properties: Erdős-Rényi networks, Watts-Strogatz networks, Barabási-Albert networks, and lattices. Erdős-Rényi networks [215, 216] are a 2-parameter family of random networks denoted $G_p(N)$. The parameter $N$ is the number of nodes in the network whilst the parameter $p$ is the probability that two nodes are connected by a link. Watts-Strogatz $W(N, q, p)$ networks [27] are a family of networks with small-world properties, whose main features are small average shortest paths and high clustering coefficients. Initially, a Watts-Strogatz instance is a regular ring lattice in which each node is connected to $q$ neighbors. The topology is then randomized by replacing, with a probability $p$, an incident node of each link with a random node, provided that no self-loops or multiple links between nodes are introduced. Barabási-Albert networks [217] are a family of scale-free networks whose architectures emerge from preferential attachment. Initially a Barabási-Albert network instance has
\(m_0\) nodes. The remaining \(N - m_0\) nodes are added one at a time, each one connected by \(m\) links to already-placed nodes with probabilities proportional to the degrees of those nodes. We also consider rectangular lattice networks. A lattice \(L_{N \times M}\) has \(NM\) nodes; the central \((N-2)(M-2)\) nodes have degree 4; the \(2(N + M - 2)\) non-corner nodes have degree 3 and the 4 corner nodes have degree 2.

The instances of the network models considered in this chapter all have \(N = 100\) nodes, except for the lattices where the number of nodes is defined the “width” and the “height” of them. We consider (sparse) networks with \(L \approx 500\) links, as well as (relatively dense) networks with \(L \approx 3200\) links. The parameter choices of our network models are therefore chosen to generate networks with (approximately) these link counts. The rewiring probability for the Watts-Strogatz instances is chosen to be \(p = 0.1\), leading to networks with high clustering coefficients and low average hop-counts (this is called the \textit{small-world} regime). The lattice network does not accept any input parameters, hence we displayed two arbitrarily chosen lattices: a square-like with 20 by 20 nodes, and a stretched lattice with 100 by 10 nodes.

<table>
<thead>
<tr>
<th>Network</th>
<th>(N)</th>
<th>(L)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>USp</td>
<td>4941</td>
<td>6594</td>
<td>Western US power grid network [27]</td>
</tr>
<tr>
<td>CA</td>
<td>5242</td>
<td>14484</td>
<td>Co-authorship network [218]</td>
</tr>
<tr>
<td>EUr</td>
<td>8730</td>
<td>11350</td>
<td>Western Europe railway network</td>
</tr>
<tr>
<td>EUp</td>
<td>9168</td>
<td>10417</td>
<td>Western Europe power grid network</td>
</tr>
</tbody>
</table>

Table 5.1: Real networks used in this chapter, ordered by size.

In addition to instances of random network models, we consider four real-world networks. First, are the high-voltage power grids of the Western United States [27] and of Western Europe [202]. In the remainder of the chapter, we refer to these two networks as USp and EUp respectively. Nodes represent power stations, transformers and generators and links represent high-voltage connections between nodes. Second, we study a social collaboration network from ArXiv that covers papers joining authors in the field of Relativity and Quantum Cosmology [218] in the period January 1993 to April 2003. We refer to this network as CA. Here, two nodes are joined if the two authors appear as co-authors in at least one paper. Finally, we consider the Western European Railway network, referred to as EUr. The nodes in the network represent railway stations and links represent railway tracks between stations. The size of each real network is given in Table 5.1.

5.3.1 Theoretical preliminaries

Let us denote by \(G(N; k)\) a network with \(N\) nodes which has had a fraction \(k\) of its nodes attacked. Before any attacks, the network is thus denoted by \(G(N; 0)\). We are interested in calculating the change of the network metric \(R = R_{G(N; k)}^{(P)}\) as a function of the percentage of attacked nodes \(k\). Denote by \(\mathcal{T}\) the set of nodes that have been attacked and denote by \(\mathcal{N} \setminus \mathcal{T}\) the nodes that have not been attacked. Here, \(\mathcal{N}\) is the set of all nodes in the network. The number of attacked nodes in \(G(N, k)\) is \(m = |\mathcal{T}| = \lfloor kN \rfloor\) and therefore the number of nodes that have not been attacked is \(N - m = |\mathcal{N} \setminus \mathcal{T}| = N - \lfloor kN \rfloor\).
A metric, such as efficiency, whose value is the average over all node pairs is dealt with in a similar fashion. Denote by $R_{ij}$ the contribution of a pair of nodes $i$ and $j$ ($i \neq j$) to the $R$-value. If either node $i$ or $j$ has been removed (that is, $i \in T$ or $j \in T$), $R_{ij} = 0$. Thus,

$$R = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} R_{ij} = \frac{1}{N(N-1)} \sum_{i,j \in N\setminus T, i \neq j} R_{ij}$$

(5.4)

### 5.3.2 Analytical results for Erdős-Rényi networks

Here, we provide analytical results for the robustness of Erdős-Rényi random networks relative to the efficiency and size of the giant component. In the case of random removal, where $k\%$ of the nodes are discarded, the resulting network has $N'$ nodes of degree 0. The remaining nodes form an Erdős-Rényi random network $G_p (N - m)$ with the same link density $p$ because the link between two nodes from $N \setminus T$ appears with a fixed probability $p$. Targeted attacks afford no such easy analysis, making them much less analytically tractable.

**Efficiency.** The average efficiency is the reciprocal of the mean hopcount, which is approximately $h_{ij} \approx \frac{\ln(N)}{\ln(Np)}$ for an arbitrary pair of nodes $i$ and $j$ in a connected Erdős-Rényi network [200, 22]. Consequently, the efficiency $e_{ij}$ for the pair $i,j$ is $e_{ij} = \frac{1}{h_{ij}} \approx \frac{\ln(Np)}{\ln(N)}$. Consider the independent, random removal of $k\%$ of the nodes. The resulting network is an Erdős-Rényi network $G_p (N - \lfloor kN \rfloor)$ with $N' = \lfloor kN \rfloor$ isolated nodes. Thus, the efficiency $e_{ij}$ of an arbitrary pair of nodes $i$ and $j$ is approximately

$$e_{ij} = \begin{cases} \frac{\ln((N-\lfloor kN \rfloor)p)}{\ln(N-\lfloor kN \rfloor)}, & \text{for } i,j \in N\setminus N' \\ 0, & \text{otherwise} \end{cases}$$

(5.5)

Substituting (5.5) into (5.4), yields

$$E[1/H] = \sum_{i,j \in N\setminus N', i \neq j} e_{ij} \approx \sum_{i,j \in N\setminus N', i \neq j} \frac{\ln((N-\lfloor kN \rfloor)p)}{\ln(N-\lfloor kN \rfloor)} \approx \frac{\ln((1-k)Np)}{\ln((1-k)N)} \sum_{i,j \in N\setminus N', i \neq j} 1$$

$$= \frac{\ln((1-k)Np)}{\ln((1-k)N)} \frac{N'(N' - 1)}{N(N-1)} \approx \frac{\ln((1-k)Np)(N - kN)^2}{N^2} = (1-k)^2 \frac{\ln((1-k)Np)}{\ln((1-k)N)}$$

(5.6)

The shape of (5.6) is validated by Figs. 5.4a and 5.4d.

**The size of the giant component.** The size of the giant component decreases when the network is attacked, as attacked nodes are removed from the giant component. Thus,

$$S \leq 1 - k$$

(5.7)

where equality holds if and only if all nodes in $N\setminus T$ form a giant component. An Erdős-Rényi network $G_p (N)$ is almost certainly connected if $p > p_c = \frac{\ln N}{N}$, therefore:

$$S = 1 - k, \text{ if } p > \frac{\ln (N - \lfloor kN \rfloor)}{N - \lfloor kN \rfloor}$$

(5.8)

The function $\frac{\ln(N-\lfloor kN \rfloor)}{N-\lfloor kN \rfloor}$ increases with the percentage of attacked nodes $k$. Thus, for fixed values of $p$ and $N$ and large enough values of $k$, $p \leq \frac{\ln(N-\lfloor kN \rfloor)}{N-\lfloor kN \rfloor}$. As this is the
connectivity threshold for Erdős-Rényi networks, we find that \( S < 1 - k \). The “dips” in the lines \( R = 1 - k \) for large \( k \) in Figs. 5.3a and 5.3d are manifestations of disconnected giant components. As can be seen in Fig. 5.3a, when \( p \) is small, disconnection happens for smaller values of \( k \). The size of the giant component is approximately \[ S = 1 - e^{-p(N - \lfloor kN \rfloor)S} \]
which explains the “dip” in the linear line \( R = (1 - k) \). In the analysis for the size of the giant component, we consider \( R = S \), however a slightly similar approach is comparing the absolute values by taking \( R = S/S[0] \), where \( S[0] \) is the size of the giant component in the original network. Clearly, both approaches are identical if the original network does not have disconnected parts.

5.3.3 **Robustness of random network model instances**

In this section, we interpret simulation results of the random network model instances. The properties of the network models considered in the analysis are stated at the beginning of this section (Section 5.3). The simulations have been repeated 1,000 times to obtain the energy, the sensitivity and \( R \) values.

**Size of the giant component.**

*Energy analysis:* The maximum energies for all strategies and networks exceed 0.460 (0.5 is the maximum energy attainable for the giant component, as the slope of \( R \)-value cannot exceed \( (1 - k) \)). The \( R \)-values for the giant component are shown in Fig. 5.3 and Table 1 in the supplementary material [219]. For almost all networks, there are sequences of node removals that render large giant components. In addition, lattice networks show interesting behavior: there seems to be a phase transition around 50% as seen in Fig. 5.3g. After randomly removing more than 50% of the nodes, all the topologies lose energy at an increased rate, due to the loss of connectivity. This result is in accordance with percolation theory [220], where the critical probability of bond percolation equals 0.5\( N \).

<table>
<thead>
<tr>
<th></th>
<th>( G_{ER} )</th>
<th>( G_{WS} )</th>
<th>( G_{BA} )</th>
<th>Lattice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Betweenness</td>
<td>(- R^{(top)})</td>
<td>(- R^{(top)})</td>
<td>(- R^{(top)})</td>
<td>(- R^{(top)})</td>
</tr>
<tr>
<td>Closeness</td>
<td>(- R^{(top)})</td>
<td>(- R^{(top)})</td>
<td>(- R^{(top)})</td>
<td>(- R^{(top)})</td>
</tr>
<tr>
<td>Degree</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eigenvector</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 5.2:** *Summary of the most and least destructive targeted attack strategies on random networks relative to the sizes of their giant components.* Larger giant components are deemed more desirable. The symbol - means “most destructive” whilst + means “least destructive”. All considered attacks had approximately the same least effect on all networks. As we already mentioned, every attack’s maximum \( R \)-value is above 0.46.

*Sensitivity analysis:* Lattice networks display the highest sensitivity, followed by Watts-Strogatz networks \( G_{WS} \), Barabási-Albert networks \( G_{BA} \), and finally Erdős-Rényi networks \( G_{ER} \) (see Table 1 in [219]). Erdős-Rényi networks are the least sensitive to node
removals, suggesting that this topology is the most robust in terms the giant component’s sensitivity. However, when the link density is sufficiently high, sensitivity values are small for all topologies.

**Targeted versus random attacks:** Amongst the random network models, the ratio $\mathcal{E}_{\text{min}}/\mathcal{E}_{\text{avg}}$ attains the highest value for Barabási-Albert networks (an unfavorable condition), followed by Erdős-Rényi networks and finally Watts-Strogatz networks. As with efficiency, the lattice network has the highest ratio $\mathcal{E}_{\text{min}}/\mathcal{E}_{\text{avg}}$ for all targeted strategies, peaking at 1.42 for node-degree targeted attacks. Again, this means that, for grid networks, the targeted strategies perform worse (on average) than a random strategy. The ratio $(\mathcal{E}_{\text{max}} - \mathcal{E}_{\text{min}})/\mathcal{S}$ is the highest for Barabási-Albert networks, followed by Erdős-
Rényi networks, Watts-Strogatz networks and lattices. Targeted attacks have the largest impact on Barabási-Albert networks, whilst Erdös-Rényi networks are the least affected. The most destructive perturbations are those based on degree and betweenness centrality.

**Efficiency**

![Graphs showing efficiency](image)

**(a) Erdös-Rényi** $p \approx 2p_c$

**(b) Watts-Strogatz** $q = 10$, $p = 0.1$

**(c) Barabási-Albert** $m = 3$

**(d) Erdös-Rényi** $p \approx 15p_c$

**(e) Watts-Strogatz** $q = 64$, $p = 0.1$

**(f) Barabási-Albert** $m = 5$

**(g) Lattice** $20 \times 20$

**(h) Lattice** $100 \times 10$

**(i) Legend**

**Figure 5.4:** The $R$-values for the efficiency. The network model considered and its property (the link density $p$ for Erdös-Rényi, the number of neighbors $q$ per node and the rewiring probability $p$ in Watts-Strogatz and $m$ the number of links of a newly added node in Barabási-Albert model) is given in sub-captions (a) - (h). The x-axis is the percentage of removed nodes either at random or according to a centrality measure as it is shown in the legend.

As can be seen from Fig. 5.4, amongst the sparse networks, the lattice has the lowest average efficiency energy, followed by $G_{WS}$ (with $q = 10$). Both of these networks are fairly regular ($G_{WS}$ has a low rewiring probability in our chapter) leading us to conclude that regularity does not confer robustness in terms of efficiency. $G_{BA}$ networks are the
most robust to random attacks as well as being the most sensitive, making them the most vulnerable to targeted attacks. Again, $G_{ER}$ networks win in terms of energy and sensitivity, making them robust both to random and targeted attacks.

Table 5.3 reveals the effect of particular attack strategies on the network models. Again, node degree and betweenness attack strategies perturb non-lattice networks the most, in contrast to lattices where the eigenvector attack strategy is the most disruptive.

<table>
<thead>
<tr>
<th>Attack Strategy</th>
<th>$G_{ER}$</th>
<th>$G_{WS}$</th>
<th>$G_{BA}$</th>
<th>Lattice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Betweenness</td>
<td>$R^{(top)}$</td>
<td>$R^{(top)}$</td>
<td>$R^{(top)}$</td>
<td></td>
</tr>
<tr>
<td>Closeness</td>
<td>$R^{(bot)}$</td>
<td>$R^{(top)}$</td>
<td>$R^{(bot)}$</td>
<td></td>
</tr>
<tr>
<td>Degree</td>
<td>$R^{(top)}$</td>
<td>$R^{(bot)}$</td>
<td>$R^{(bot)}$</td>
<td></td>
</tr>
<tr>
<td>Eigenvector</td>
<td>$R^{(bot)}$</td>
<td>$R^{(bot)}$</td>
<td>$R^{(bot)}$</td>
<td>$R^{(top)}$</td>
</tr>
</tbody>
</table>

Table 5.3: Summary of the most and least destructive targeted attack strategies on random networks relative to efficiency. Higher efficiency values are deemed more desirable. The symbol - means “most destructive” whilst + means “least destructive”.

5.3.4 Robustness of real networks

In this section, we compare the robustness profiles of real-world networks to the robustness profiles of the network models presented in the previous section. Many numerical details regarding the energy and the sensitivity are given in Table 3 of the supplementary material [219].

The size of the giant component

Some of the real-world networks are composed of several disconnected components, leading to initial $R$-values that are smaller than 1.0.

The ratio $(E_{max} - E_{min})/S$ is the largest for the CA network (27.0), followed by EUR network (14.0), the EUp network (11.7) and finally the USp network (11.4). Targeted attacks have the biggest impact on the Western United States power grid and the smallest impact on the co-authorship network. In addition, the ratio $(E_{max} - E_{min})/S$ is in all cases higher than for model network ratios (which fall in the range [2.4, 9.6]). Real-world networks are more easily disconnected than the instances of the random models.

As before in Section 5.3.2, the most effective attack strategies are the node degree and node betweenness attacks. The least effective attack strategy is the node closeness attack (e.g. Fig. 5.5b), which leaves the size of the giant component nearly untouched for all real networks. The four real networks considered are: coauthorship network (CA), US power grid (USp), European power grid (EUp), and European railway (EUR).

Efficiency

The network with the highest absolute efficiency value is the co-authorship network. As before, this is due to the high link density and the presence of many cliques. Remarkably, all four real-world networks show rapid decreases in efficiency after only $\approx 10\%$ of their nodes are removed. This behavior is similar to that observed for the Barabási-Albert model: in this case, the removal of $\approx 20\%$ of the nodes causes a large drop in efficiency.
5.4 Robustness Optimization by Degree-preserving Rewiring

This section demonstrates the use of our robustness framework by studying changes in the metric envelope of a network as it is rewired (through degree-preserving transformations) in order to increase or decrease its degree assortativity [221, 52].
5.4.1 Degree assortativity

Allow me to refresh the definition of degree assortativity. Degree assortativity measures the tendency of links to connect nodes with similar degrees. Formally, it is defined \[ \rho = 1 - \frac{\sum_{i \sim j} (d_i - d_j)^2}{\sum_{i=1}^{N} d_i^3 - \frac{1}{2L} (\sum_{i=1}^{N} d_i^2)^2} \]
where \( i \sim j \) denotes a link between nodes \( n_i \) and \( n_j \), \( d_i \) the degree of node \( n_i \) and \( D = [d_1, d_2, \ldots, d_N] \) the degree-sequence of the network. The degree assortativity has been shown [222] to be an important indicator for the epidemic spread such that assortative networks spread are more prone to the propagation of epidemics. Moreover, the close relation between the degree assortativity and the modularity, which is an indicator for network clusterness, has been studied in [223].

5.4.2 Degree-preserving rewiring

Degree-preserving rewiring [52] allows for the modification of the link architecture of a network without changing its degree sequence. In a rewiring step, a pair of links \( \{u, v\}, \{w, x\} \) in a network \( G \) is selected such that \( u, v, w, x \) are distinct nodes. If \( \{u, w\} \notin \mathcal{L}(G) \) and \( \{v, x\} \notin \mathcal{L}(G) \), \( \{u, v\} \) and \( \{w, x\} \) can be rewired to (that is, replaced by) \( \{u, x\}, \{w, v\} \).

5.4.3 Rewiring algorithm for assortativity optimization

We used the greedy degree-preserving rewiring algorithm of [224] to optimize degree assortativity. In each iteration, the algorithm samples up to \( s \) pairs of links. If a sampled pair of links is rewirable and if the rewiring leads to a desired change in the degree assortativity (see Lemma 1 in [52]) of the network, the change is made. If, after \( s \) sampling attempts, no such pair of links is found, the algorithm terminates.

5.4.4 Experiment setup

Using our simple algorithm, we maximized and minimized the degree assortativity of an Erdős-Rényi graph as well as a Barabási-Albert graph. The number of rewirings needed to achieve high or low degree assortativity can number in the hundreds or even thousands. Therefore, it is impractical to study the robustness profiles of the networks associated with each rewiring step. For each network, we study five snapshots: (1) a rewired network whose assortativity is fully maximized; (2) a rewired network whose assortativity is halfway between the fully maximized value and that of the original network; (3) the original network; (4) a rewired network whose assortativity is halfway between the fully minimized assortativity value and that of the original network; and (5) a rewired network whose assortativity is fully minimized. Snapshots of the \( G_{ER} \), along with corresponding energy and sensitivity changes for the giant component and efficiency are shown in Fig. 5.6. The analogues for \( G_{BA} \) are shown in Fig. 5.7.
Figure 5.6: The influence of degree-preserving assortativity-optimization on the robustness of an Erdős-Rényi network. Robustness is measured relative to the giant component size (left) and the efficiency (right). In the first (top) row, a rewired network whose assortativity is fully maximized (up to 0.83); in the second row, a rewired network whose assortativity is halfway between the fully maximized value and that of the original network; in the third (middle) row, the original network; in the fourth row, a rewired network whose assortativity is halfway between the fully minimized assortativity value and that of the original network; and in the fifth (bottom) row, a rewired network whose assortativity is fully minimized (down to −0.91). The legend is the same as the ones in Figs. 5.3, 5.4 and 5.5.
Figure 5.7: The influence of degree-preserving assortativity-optimization on the robustness of a Barabási-Albert graph. Robustness is measured relative to the giant component size (left hand) and the efficiency (right). In the first (top) row, a rewired network whose assortativity is fully maximized (up to $0.47$); in the second row, a rewired network whose assortativity is halfway between the fully maximized value and that of the original network; in the third (middle) row, the original network; in the fourth row, a rewired network whose assortativity is halfway between the fully minimized assortativity value and that of the original network; and in the fifth (bottom) row, a rewired network whose assortativity is fully minimized (down to $-0.66$). The legend is the same as the ones in Figs. 5.3, 5.4 and 5.5.
5.4.5 Interpretation

As assortativity is maximized, the $E_{\text{avg}}$ of both the giant component and efficiency decrease (the black lines in Figs. 5.6 and 5.7). In the intermediate assortativity-maximized cases, the decrease is mild, and what these networks lose in $E_{\text{avg}}$ they gain by lowering the $(E_{\text{max}} - E_{\text{min}})/S$ ratio. In other words, intermediate assortativity-maximized networks become less robust against random attacks, but relatively stronger against targeted attacks. Finally, the assortativity-maximized networks display the lowest average energy $E_{\text{avg}}$ for both metrics. However, these maximized networks are relatively strong to targeted attacks, as depicted by low $(E_{\text{max}} - E_{\text{min}})/S$ ratios.

The situation is almost reversed when assortativity is minimized, where $E_{\text{avg}}$ remains high while $(E_{\text{max}} - E_{\text{min}})/S$ ratios dramatically increase: targeted attacks are more devastating for assortativity-minimized networks than random attacks are. In addition, these intermediate disassortative networks have slightly higher $E_{\text{avg}}$ than the original networks. Finally, $G_{\text{ER}}$, whose assortativity is fully minimized is fragile against targeted attacks and its average energy is not particularly good. In contrast, $G_{\text{BA}}$ with fully minimized assortativity is still more competitive than its less-rewired sibling.

Our observations suggest that networks whose assortativities are moderately maximized (through degree-preserving transformations) are more tolerant to targeted attacks whilst having worse average-case robustness. On the other hand, networks whose assortativities are moderately minimized are more tolerant to random attacks (and less tolerant to targeted attacks). These observations match those of Friedel and Zimmer [225], who researched the role of assortativity in protein interaction networks.

5.5 Chapter Summary and Conclusions

In this chapter, we studied the robustness of network topologies under various challenges, by applying our framework to random network models and real networks. Our framework extends the traditional approach for analyzing network robustness by considering not only average network performance but approximate probability density functions. The density functions reflect the probable performance characteristics of a network after a number of random node removals.

Within the topological robustness framework [210], we have extended and detailed the concept of robustness envelopes. We studied the robustness envelopes of sparse and dense instances of well-known random classes of networks, as well as four real-world networks. Our envelope approach shows that although networks may have similar average-case performance under attack, they may differ significantly in their sensitivities to certain attack sequences, as illustrated in Fig. 5.8. We also contrasted robustness envelopes of the studied networks to their responses when subjected to targeted attacks, based on node centrality measures.

We found that targeted attack strategies often lead to performance degradation beyond the limits of the robustness envelopes that we computed. This leads us to conclude that centrality-based targeted attacks are sufficient for studying the worst-case behavior of real-world networks. In this regard, our analysis suggests that real-world networks are susceptible to rapid degradation under targeted attacks. The overlap between centrality rankings reveals that attack strategies based on different centrality measures may have very similar results. We argue that degree centrality and eigenvector centrality strike
Figure 5.8: Two graphs’ robustness envelopes. The top envelope shows a higher average, but greater range of values. The bottom envelope spans a lower range of values, but has a lower average. Classification schemes come into play, to decide which graph is better depending on the network’s purpose, i.e. its offered services.

a good balance between differences in attack sequences and in computational power required.

Finally, we investigated envelopes and the effect of attack patterns on networks whose structures were modified, through degree-preserving rewiring. We found that by slightly increasing degree assortativity of those networks, our networks became more resilient against targeted attacks, if somewhat less resilient against random attacks. The converse was true when decreasing degree assortativity. How to modify a given graph to enhance its robustness level from $R_1$ to $R_2$ is, in general, difficult to determine. However, the problem is well-defined and the complication lies in the computational feasibility (even tractability), not in the concept, nor in the specification. The problem is thus more of an optimization nature: given a graph, each elementary change has an effect on $R$, and a “perturbation” may deteriorate $R$ below a minimum desired robustness level $R_{\text{thresh}}$.

Our framework strives to be both simple and general, through the use of a linear combination of topological and service metrics. However, this generality comes with a series of caveats. One of the main strengths of our framework comes from the linearized contribution of aggregated services. However, in most cases, we expect it will be difficult to determine the numerical values of the components $s_k$ based on a series of services. Furthermore, modeling networks as nodes and links overlooks the complexity of protocol interactions. For lack of time and space, I was not able to further investigate this subject in the span of this thesis. Yet, the effects of linearity on service aggregation needs further exploration.
5.5.1 Discussion

An obstacle left unmentioned is that of metric normalization. In general, metrics are normalized to very distinct ranges of values. For example, the average hopcount ranges from 1 to $N/3$, for a full connected graph and a path graph, respectively. Even if a graph contained millions of nodes, the average hopcount usually embodies the small-world phenomena, by which the average hopcount is usually in the order of $\log(N)$. This causes the average hopcount to take a very small range of values which usually range within $[1, 20]$, even in networks with billions of nodes. On the other hand, metrics such as the clustering coefficient range from 0 to 1, and it is common to find networks with values as low as 0.01 and as high as 0.9. If we are presented with the need to compute the R-value using both the hopcount and the clustering, we will face a difficult normalization problem. For instance, a linear max-min normalization will skew $R$ towards the clustering coefficient, because it spans a relatively wider range of values than the hopcount.

An additional issue is the comparison among networks with a different number of nodes and links. In other words, we still lack a good normalization of a graph matrix. In many complex networks, properties for small graphs (small $N$) differ from that in the asymptotic regime (large $N$). The difficulty of comparing two networks with different sizes $N$ or number of links $L$ lies in the fact that the scaling of topological metrics with respect to $N$ is network dependent. Indeed, consider an Erdős-Rényi graph $G_p(N)$ with $N$ nodes and where $p$ is the probability to have a link (independent of the existence of other links) between two nodes. If $N$ is small, then $G_p(N)$ is a random graph, whereas for large $N$, $G_p(N)$ tends to become a deterministic, regular graph (see [22, p. 42]). This example shows that, when a graph of a certain class grows in size, its properties may change. Another similar example appears in the class of power law graphs, that seem quite good models matching the degree distribution in complex networks such as the Internet. Power law graphs can be constructed by a stochastic growth rule such as preferential attachment. Only when the power law graph is sufficiently large (in practice $N > 500$), a power law for the degree is observed. For smaller graphs (about $N < 500$), the degree does not follow a clear power law [226] and most often an exponential fit is statistically equally significant as a power law fit. In summary, the way properties (measured via graph metrics) change with $N$ is generally different for each class of graphs. This means that scaling laws are graph dependent, which complicates proper comparison between graphs of different sizes. Comparison of properties in graphs with different number of nodes and links needs to be investigated.
6.1 General Conclusions

Back in 2007, I was confronted with a challenging task: seeking a general network robustness framework, a framework that would apply to any network, for any function, in any context. After a review of the related work done in the last century, I shortly realized that there is a lack of a precise definition of robustness, no standard practices, and an absence of knowledge (and awareness) of resilience metrics, which made it difficult to find a starting point to build a framework. In my attempt to craft a definition, I searched for inspiration in this collage of earlier studies. The existence of two common denominators rapidly jumped to my attention. These two major factors were: the type of stress the network is undergoing, and the goal the network is designed to accomplish. Indeed, the English Oxford dictionary defines “robust” as something “able to withstand force, pressure, or wear”, which emphasizes the role of external aggravating factors. In other words,

a system is not simply strong or robust, but instead it is strong against something;

or shamelessly borrowing Einstein’s words, everything is relative. I believe that this relative notion on robustness makes finding an all-encompassing definition “a bridge too
far”, tantamount to finding the Theory of Everything. Nevertheless, far from giving up, we decided to step down our goals for this thesis. Instead of looking for a general absolute framework, I decided to build a framework that accommodates to the current network necessities. Ideally, this framework is as general and as simple as possible.

This thesis can be interpreted as my captain’s ship log, full of carefully laid annotations delineating the difficulties I came across during my quest towards a robustness framework, which we successfully found. The first step towards quantifying a property of any system is to probe such a system. For that purpose, I started studying all possible ways one can measure a graph by means of the simplest tools, i.e. topological graph metrics. To my alarm, I came to the conclusion that

graph metrics are a constantly evolving body of knowledge.

For example, fifteen years ago power-laws were considered a curious rule of thumb, an amusing observation where 80% of the effects come from 20% of the causes. Nowadays, power-law degree distributions are studied on a daily basis while being spotted left and right, from human made networks to biological systems. Similarly, young metrics such as assortativity or modularity were born shortly after this thesis was initiated, yet they have become the de facto metrics used in epidemiology and community detection algorithms, respectively. My assessment is that any current taxonomy of graph metrics is bound to evolve, as new key network proprieties are found as years go by. Thus I do not suggest focusing on building a conclusive taxonomy, but instead an evolving platform that walks hand in hand with current developments. The taxonomy depicted in Chapter 2 is my proposed starting point, which offers an updated view of the state of the art in topological metric. However, this list requires constant updates in order to keep up with latest investigations. I am confident that, when rebuilt in ten years time, this list will look completely different.

Once our topological metrics were tidily classified in Chapter 2, me and my colleagues delved into three particular scenarios: the hopcount of cellphone networks, the modularity of epidemics, and the spectrum of interdependent networks, all included in Chapter 3. By conducting three independent robustness case studies we demonstrate how

robustness can be arbitrarily defined in any way that suits a given network purpose.

At this point in time, we realized that a single metric is not able to accurately measure any network’s robustness. In particular, the network service seriously complicates a computable framework. For instance, we observe that a network that is very efficient in propagating information is, on the other hand, also quite vulnerable to virus or malware spread (that is “undesirable information”). This illustrates that opposite services may exist over a same topology, at any rate, if “virus anti-spread” can be regarded as a service. Even tough we provided no formal proof, experiments led us to believe that for any metric there always exists a counterexample, in which the offered notion contrasts the initial intuition (excluding well understood deterministic structures, such as lattices or rings). In other words,
one metric does not provide enough insights into a graph structure so as to classify it or to evaluate its robustness.

Realizing that simple metric approaches fall short, we decided to approach the problem from a multidimensional perspective, i.e. using more than one topological metric at a time. Even while following this approach, we still aimed to define robustness as a single value as a combination of topological metrics. Ideally, we expect the combination of the right metrics to draw a complete figure of the graph, similarly to three orthogonal projections of a three dimensional figure describing its outer shape. However, a multidimensional approach poses an additional non trivial problem, i.e. the issue of metric correlations. By conducting correlation studies in Chapter 4, we realized that topological metrics are correlated in non-trivial ways.

The degree of correlation generally depends on the type of graph, but there exist cases where this correlation is graph independent. In order to come up with a set of interdependent metrics, we spent time and effort studying the correlation between the algebraic connectivity, the degree distribution, and the weighted link betweenness, among others.

The second limiting factor to approaching robustness is the service aggregation model, which we assumed to follow a linear behavior so as to induce a general framework. However, this thesis’ main focus has been the topology of networks, relegating the service layer to a second plane. We are aware the linear models may not be first order approximations, thus in order to better model the effects of services on network robustness, we need to further investigate the performance and response patterns of services under stress using our multi-dimensional approach.

Armed with a broad knowledge in topological metrics and their correlations, we can flash back to our original problem: defining robustness. Recall that the main ingredients of robustness are: the type of stress the network is undergoing, and the goal the network is designed to accomplish. With the help of existing statistical tools, we modeled network challenges with our robustness envelopes.

The area under a robustness envelope, which can be regarded as the variation of the impact of a certain perturbation on a graph, is a good robustness estimator.

More intuitively, the measured area allows us to quantify the uncertainty or the amount of risk due to a perturbation. If the type of challenge (such as storms, floods, earthquakes, DDOS, worms etc.) a network will face can be determined a priori, the envelopes may be refined resulting in better risk assessments. Secondly, we model the multi-level service nature of networks by means of a linear combination. This approach benefits from simplicity and generality, which follow from its linearity and the \( m \)-dimensionality of the topology vector, respectively. However, in most cases we expect it to be difficult to determine the numerical values attributed to the service vector, if even possible.

To sum up, by measuring the extent to which topological metrics degrade under random and targeted attacks, we simplified the concept of robustness to a single quantifiable
value, the R-value. Evaluating network robustness in this way unequivocally quantifies robustness as a measure of graph deterioration in the presence of challenges. Similarly to drawing a trajectory over an multi-dimensional space, the R-value defines a unique trajectory in the $m$-dimensional metric space, where $m$ is the number of metrics the R-value is composed of. I am confident that boundaries and classes can be built in this metric space, enabling the standardization of frameworks across sectors and countries across the globe. Alas, by the end of this thesis I ran short on time, so I had little time to explore the extent to which linearity holds, or the effect of complementary concepts such as defense, remediation or recovery mechanisms.

6.2 Main Contributions

The main contributions of this thesis can be summarized as follows:

- We proposed a taxonomy of topological graph metrics, which allows any existing metric to be categorized into three classes: distance, connection and spectra (Chapter 2). Such a taxonomy will help future researchers when probing their systems, by offering them a complete and organized classification. Nevertheless, the full understanding of metric correlations is still an open challenge, which clouds any classification’s integrity. For example, link betweenness can be easily interpreted as either a distance metric or a connection metric.

- We learned that the spread of viruses is highly dampened by quarantining infected regions through modularity cuts (Chapter 3). Although this is not a novel result from an epidemiology point of view, we find it remarkable that community detection metric can be used as a defense tool in real scenarios.

- A new phase transition was discovered and characterized on synchronization of interdependent networks (Chapter 3), characterized by sharp transitions in the Fiedler vector after the addition of sufficient links among interdependent networks. This result is critical for robustness studies related to synchronization, as it determines a unique critical degradation point above which synchronization is not affected.

- We designed a challenge-aware protection scheme (Chapter 3), based on NEC’s proposed Rope-Ladder. Our challenge awareness algorithm is able to circumvent concurrent challenges, by evaluating the conditional failure probabilities of adjacent elements. Furthermore, we learned how to rigorously deal with multiple challenges occurring at the same time.

- The quality of different Internet topology generators was evaluated, by measuring the produced graphs (Chapter 4). We conclude that GLP is the one which produces those topologies closest to reality.

- Extensive insights were gained in a variety of metric correlations (Chapter 4). More specifically, we discovered a tight bond between the algebraic connectivity and the weighted link betweenness. The latter proves to be a very good estimator for the first.
Lastly, we defined a general robustness framework through the study of robustness envelopes (Chapter 5). This framework defines, characterizes, and computes network robustness, allowing us to categorize networks into classes, or attain desired robustness values by means of network modifications. Our framework is both linear and general (i.e. can be applied to any network), however, its parameters must be defined beforehand and a profound knowledge of the working of the services is required.

6.3 Directions for Future Work

This thesis represents a mere drop, a minute contribution in the vastness of graph theory. After a brief period reading scientific journals, one comes to the unavoidable realization that humanity holds more open questions than it has answers for, and the field of network robustness poses no exception. Open questions branch out of every single page of this thesis, but this section will limit itself to the ideas that I consider most relevant, while adding a personal flavor to them.

In the last chapters of this book, the building blocks of a robustness framework were laid out, but the framework applicability is yet to be thoroughly investigated and validated. For this purpose, I propose following a bottom-up approach, in which we distance ourselves from analytical tools and delve into realistic world scenarios. I expect that new insights will be gained after getting our hands dirty with real data.

While this thesis has directed most of its efforts towards quantifying average scenarios via robustness envelopes, the overlooked the effect of particular realizations, such as best case and worst case scenarios. From a practical standpoint, one could argue that network failures are scarce and seldom happen, thus Fukushima Daiichi’s network manager may not be as interested in the expected degradation after 1,000 trivial failures as much as (s)he is in the worst case scenario, e.g. a tsunami. Further work on characterizing the behavior of individual non-trivial metric realizations is required, this research would yield tighter bounds and better characterizations of individual scenarios. However, focusing on single realizations instead of large statistical properties would entail a paradigm switch, in which robustness is no longer understood as a statistical response, but as a tailored case-by-case protection scheme.

Scattered within this thesis, we showed a number of figures that elegantly illustrate small lattices and simple topologies (e.g. Figure 1.2). I find it beautifully challenging to produce figures of small graphs, which is often readily done by the means of force-directed layouts. Yet, displaying graphs larger than 100 nodes is often as problematic as it is slow. To our advantage, it so happens that the human brain excels at pattern recognition (e.g face recognition), by finding non-trivial patterns in otherwise random structures. For example, it would be easy for a human to notice the existence of a power-law degree distribution if high-degree nodes are drawn in dark red colors, and low-degree nodes in bright blue colors. Displaying complex graphs in ways that emphasize their intrinsic properties remains an open issue, which has partially been addressed for small graphs by spring embedded algorithms.

At the time of writing this thesis, “interdependent networks” has become a hot topic. Mainly due to the critical (and unexpected) effects of cascading failures, the study of multiple interlinked networks has caught the attention of the scientific community. In this thesis we delved into synchronization properties by studying the spectral properties of the
corresponding graph. To date, many properties of simple interdependent networks remain unexplored, including percolation, the dynamics of epidemic processes, and cascading failures.

In general, graph analysts assume that real networks can be accurately modeled by simple graphs. However, as in any natural and social sciences field, the trick to understand the occurring phenomena is to describe it with models that are as simple as possible - but not too simple - by striking a compromise between simplicity and verisimilitude. Even though undirected unweighted graphs have proven to be a faithful representation of a wide spectrum of real networks, the latter tend to be dynamic, contain heterogeneous nodes, and often display links with varying capacities. We can further increase the complexity of our robustness model by adding the notions of link and node categories, at the cost of increased complexity.

Finally, we may pursue the slippery set of orthogonal metrics by further studying metric correlations. To my understanding, this enterprise is a necessary yet perpetual activity, due to the inability to reach a consensus under the constant flood of new metrics, which keep emerging from the latest scientific publications. Although finding this orthogonal set may take us additional decades, we should not forget that a set of orthogonal values was already proposed a century ago: the graph spectra. Probably due to the lack of a tangible meaning, graph spectra remained elusive for the last 60 years, yet the properties of largest eigenpairs have been thoroughly studied and offer a stable starting point. By definition, all the eigenvectors of a graph are orthogonal to each other, which leads me to believe that the spectra is a prime tool to find an orthogonal metric system. In particular, I strongly encourage the study of low-order eigenvectors as opposed the largest eigenpairs. Pattern detection may prove a useful tool for finding new structures within the graph spectra, where a wealth of patterns awaits its discovery.
BIBLIOGRAPHY


dynamics in a p2ptv network,” in Proceedings of the 8th International IFIP-TC
6 Networking Conference, NETWORKING ’09, (Berlin, Heidelberg), pp. 326–337,
Springer-Verlag, 2009.

the internet topology,” in SIGCOMM ’99: Proceedings of the conference on Applica-
tions, technologies, architectures, and protocols for computer communication, (New

[46] P. Mahadevan, D. V. Krioukov, M. Fomenkov, B. Huffaker, X. A. Dimitropoulos,

[47] P. Mahadevan, D. Krioukov, M. Fomenkov, X. Dimitropoulos, K. C. Claffy, and
A. Vahdat, “The internet as-level topology: three data sources and one definitive


[49] C. Clausen and H. Wechsler, “Color image compression using pca and backpropaga-

[50] J. L. Rodgers and A. W. Nicewander, “Thirteen ways to look at the correlation


ativity and degree-preserving rewiring on the spectra of networks,” The European
August 2010.

[53] W. De Haan, Y. A. Pijnenburg, R. L. Strijers, Y. Van Der Made, W. M. Van
Der Flier, P. Scheltens, and C. J. Stam, “Functional neural network analysis in
frontotemporal dementia and alzheimers disease using eeg and graph theory,” BMC

[54] A. Barrat, M. Barthlemy, and A. Vespignani, Dynamical Processes on Complex


Apr. 2002.


148 BIBLIOGRAPHY


APPENDIX A

Spectra of Isomorphic Graphs.

This appendix offers additional material to that of Section 2.2.3, by proving that isomorphic graphs have the same eigenvalues and eigenvectors (accordingly reshuffled by the corresponding permutation).

Let us start by laying out basic definitions. Two graphs are said to be isomorphic if they present the same topology, disregarding the node labels. For example, the three graphs displayed in Figure A.1 are isomorphic.

\[ \text{Figure A.1: Example of two isomorphic graphs.} \]

It has been proven that the eigenvalues of two isomorphic graphs are the same [37]. However, the eigenvectors of two isomorphic graphs may not be the same. In other words, the heatmaps (i.e. the eigenvectors) may be subject to change under the effect of both matrix shuffling and graph isomorphisms. But how exactly? A simple analysis of permutation matrices gives us the answer.

Let’s start with some preliminaries, graph isomorphisms can be expressed through permutation operations of the original adjacency matrix \( A \). These permutations use permutation matrices \( P_\pi \). Given a permutation \( \pi \) of \( m \) elements

\[ \pi : \{1, \ldots, m\} \rightarrow \{1, \ldots, m\} \]

given in two line form by
its permutation matrix is the \( m \times m \) matrix \( P_\pi \) whose entries are all 0 except that in row \( i \), the entry \( \pi(i) \) equals 1. Formally,

\[
P_\pi = \begin{pmatrix}
e_{\pi(1)} & e_{\pi(2)} & \cdots & e_{\pi(m)}
\end{pmatrix}
\]

where \( e_j \) denotes a row vector of length \( m \) with 1 in the \( j \)th position and 0 in every other position. The permutation matrix is orthogonal [37, art. 11, 12], i.e. \( P^{-1} = P^T \).

If two different graphs with adjacency matrices \( A \) and \( B \) are isomorphic, that means there exists a permutation of labels \( P_\pi \) such that

\[
B = PAP^T
\]

by applying the eigenvector decomposition of a matrix

\[
B = PAP^T = X_B \Lambda_B X_B^T = P (X_A \Lambda_A X_A^T) P^T
\]

\[
X_B \Lambda_B X_B^T = PX_A \Lambda_A (PX_A)^T
\]

Since it is proven in [37, art. 12] that isomorphic graphs have the same characteristic polynomial (i.e. \( \Lambda_B = \Lambda_A \)), we arrive to the conclusion

\[
X_B = PX_A \quad (A.1)
\]

which means that the eigenvectors of matrix \( B \) are a row-permutation of the eigenvectors of matrix \( A \). For example, if we relabel nodes \( i \) and \( j \) of any graph \( A \) (for \( N \geq \max(i,j) \)), the \( i \)th and \( j \)th components of all column vectors in \( X_A = \{x_1, x_2, \ldots, x_N\} \) will swap accordingly.

Simulations verify this observation, however we must exercise caution when encountering eigenvalue multiplicity. Wilkinson [177] proved that if \( x_1 \) and \( x_2 \) are eigenvectors of the same root \( \lambda_1 \), then any vector in the subspace spanned by \( x_1 \) and \( x_2 \) is also an eigenvector of \( A \). Formally,

\[
A(\alpha x_1 + \beta x_2) = \lambda_1 (\alpha x_1 + \beta x_2) \quad (A.2)
\]

where \( \alpha \) and \( \beta \) are constants. Hence, if an eigenvalue of \( A \) has multiplicity greater than one, there is an infinite number of vectors that fulfill the eigen equation \( Ax = \lambda x \). This adds a new level of complexity to the isomorphism analysis, because graphs containing root multiplicity have an infinite number of eigenvectors. In other words, if two graphs satisfy equation A.1 then the graphs are isomorphic. But if two graphs do not satisfy equation A.1, they may still be isomorphic if their characteristic polynomials had a repeated root.
APPENDIX B

Approximating the Algebraic Connectivity with Perturbation Theory.

First I would like to express my deepest thanks to Gregorio D’Agostino, for providing all the derivations included in this appendix, which offer additional material to that of Section 3.4. This appendix shows a summarized process of obtaining the first and second order perturbative analysis for $\mu_{N-1}$ of interdependent networks.

Given an interdependent network Laplacian matrix, $Q$, defined in Section 3.4.2, the problem consists in finding the minimum of the associated quadratic form in the unitary sphere ($x^T x = 1$), with the constraint $u^T x = 0$. We want to find the minimum that satisfies the spectral equations:

$$
\begin{align}
(Q_A + \alpha Q_B - \mu I)x &= 0, \\
x^T x &= 1, \\
u^T x &= 0.
\end{align}
$$

When the solution is analytical in $\alpha$, one may express $\mu$ and $x$ by Taylor expansion as

$$
\begin{align}
\mu &= \sum_{k=0}^{\infty} \mu^{(k)} \alpha^k \\
x &= \sum_{k=0}^{\infty} x^{(k)} \alpha^k
\end{align}
$$

Substituting the expansion in the eigenvalue equation (3.14) gives the hierarchy of equations:
\[
\begin{cases}
Q_A x^{(k)} + \alpha Q_B x^{(k-1)} = \sum_{i=0}^{k} \mu^{(k-i)} x^{(i)} & \text{for all } k, \\
\sum_{i=0}^{k} x^{(k-i)} x^{(i)} = 0 & \text{for } k \geq 1, \\
u^T x^{(k)} = 0 & \text{for all } k.
\end{cases}
\] (B.4)

### B.0.1 Explicit approximations up to the second order

The zero order expansion of (B.4) provides a simple set of equations:

\[
\begin{cases}
Q_A x^{(0)} = \mu^{(0)} x^{(0)}, \\
x^{(0)} x^{(0)} = 1, \\
u^T x^{(0)} = 0.
\end{cases}
\] (B.5)

Let \((\mu_{N-1})_{A1}, (\mu_{N-1})_{A2}\) and \((x_{N-1})_{A1}, (x_{N-1})_{A2}\) denote the smallest non-zero eigenvalue and the corresponding eigenvector of \(Q_1, Q_2\), respectively. Similarly

\[
\begin{cases}
(x_{N1})_{A1} = 1/\sqrt{N_1}(1, 1, \ldots, 1, 0, 0, \ldots, 0), \\
(x_{N2})_{A2} = 1/\sqrt{N_2}(0, 0, \ldots, 0, 1, 1, \ldots, 1).
\end{cases}
\] (B.6)

will represent the null eigenvectors of network \(G_1\) and \(G_2\), respectively. When the networks are put together, any combination of the former is a null eigenvector. Two special combinations are worth employing: the trivial solution corresponding to the constant vector:

\[
x_N = \frac{1}{\sqrt{N}}(1, \ldots, 1) = \sqrt{\frac{N_1}{N}}(x_{N1})_{A1} + \sqrt{\frac{N_2}{N}}(x_{N2})_{A2}.
\] (B.7)

and the other combination orthogonal to the former that represents a useful starting point for the perturbation theory:

\[
x_{N-1}^{(0)} = x^{(0)} = \frac{1}{\sqrt{N}}(1, \ldots, 1, -1, \ldots, -1) = \sqrt{\frac{N_1}{N}}(x_{N1})_{A1} - \sqrt{\frac{N_2}{N}}(x_{N2})_{A2}.
\] (B.8)

which satisfies the zero order approximation (B.5). The zero order approximation to the Fiedler eigenvalue is then null:

\[
\mu^{(0)} = 0.
\] (B.9)

The first order approximation equations follow from (B.4) as:

\[
\begin{cases}
Q_A x^{(1)} + \alpha Q_B x^{(0)} = \mu^{(1)} x^{(0)} \\
x^{(0)} x^{(1)} = 0 \\
u^T x^{(1)} = 0.
\end{cases}
\] (B.10)

Taking the projection over \(x^{(0)}\) of the first equation of (B.10), one obtains the first order correction \(\mu^{(1)}\) that depends on the zero order eigenvector only:

\[
\mu^{(1)} = (x^{(0)})^T \alpha Q_B x^{(0)}
\] (B.11)

A simple case to analyze is that where only one interlink joins \(A_1\) with \(A_2\): \((B_{12})_{ij} = \delta_{ik}\delta_{kj}\); in this case \((d_1)_{kk} = \delta_{ik}\) and \((d_2)_{ll} = \delta_{jl}\) and the perturbation estimate gives:
\[ \mu^{(1)} = \frac{1}{2}(1 + 1) + 1)\eta_i^2 = \frac{2}{N_1} \geq \mu_{N-1}(Q). \]  

where \( \eta \) is the single net \((N_1\) dimensional) unitary vector \( \eta \overset{\text{def}}{=} 1/\sqrt{N_1}(1, 1, \ldots, 1) \). When \( k \) interlinks are included, \( Q_B \) is just the sum of \( k \) contributions of the previous type thus \( \mu^{(1)} = \frac{2k}{N_1} \). That is, the first order correction to the Fiedler eigenvalue increases linearly with the number of interlinks. The first order correction to the eigenvector can be evaluated from (B.10) as a solution of the linear equation:

\[ Q_A x^{(1)} = - (Q_B - \mu^{(1)}) x^{(0)}. \]  

where the operator \( Q_A \) is invertible out of its kernel \((Q_A v = 0)\); since \((\alpha Q_B - \mu^{(1)}) x^{(0)} \) is orthogonal to the kernel, (B.13) is solvable.

The second order equations follow from (B.4) as

\[
\begin{aligned}
(Q_A x^{(2)} + \alpha Q_B x^{(1}) &= \mu^{(0)} x^{(2)} + \mu^{(1)} x^{(1)} + \mu^{(2)} x^{(0)} \\
(x^{(0)})^T x^{(2)} + (x^{(1)})^T x^{(1)} + (x^{(2)})^T x^{(0)} &= 0 \\
u x^{(2)} &= 0
\end{aligned}
\]  

that is, the second order correction is quadratic and equals:

\[ \mu^{(2)} = (x^{(0)})^T Q_B (x^{(1)}) = - (x^{(1)})^T Q_A (x^{(1)}) \leq 0. \]  

As expected \( \mu^{(2)} \) is negative, thus improving the estimate of the algebraic connectivity. The former perturbation estimates are illustrated in Fig. 3.18 together with numerical simulations.

Perturbation theory may also be applied to any initial eigenvector of the unperturbed networks. Different perturbations \( \alpha B \) will have different effects on the quadratic form of (3.14) associated with all initial eigenvectors. Therefore, it may happen that the perturbed value of \( \mu \) obtained starting from \( x^{(0)} \) is smaller than the quadratic form associated with the \( x_{N-1} \) (the unperturbed eigenvector in (3.7)) or some other educated guess. This is precisely the origin of the phase transition.

The estimates resulting from the second order perturbation theory are compared in Fig. 3.18 with the results of numerical calculations. As can be seen, for both the diagonal and the general strategies the agreement is good up to the phase transition where the starting point of the perturbation theory should be changed.

### B.0.2 Perturbative approximations and exact upper bounds

Since we are dealing with a constraint optimization problem, finding a minimum of a positive form, any test vector \( v \) provides an upper bound for the actual minimum value:

\[ \mu = \mu_{N-1} \leq \frac{v^T Q v}{v^T v}. \]  

The perturbation theory provides natural candidates as test vectors. The zero order solution provides the simplest inequality:
\[ \mu_{N-1}(Q) \leq \alpha \frac{(x^{(0)})^T Q x^{(0)}}{(x^{(0)})^T x^{(0)}} = \alpha \mu^{(1)}. \]  

(B.17)

The first order approximation provides a better (i.e. lower) upper bound:

\[ \mu_{N-1}(Q) \leq \frac{(x^{(0)} + \alpha x^{(1)})^T Q (x^{(0)} + \alpha x^{(1)})}{(x^{(0)} + \alpha x^{(1)})^T (x^{(0)} + \alpha x^{(1)})}. \]

that is:

\[ \mu_{N-1}(Q) \leq \frac{\alpha \mu^{(1)} + \alpha^2 \mu^{(2)} + \alpha^3 (x^{(1)})^T Q B x^{(1)}}{1 + \alpha^2 (x^{(1)})^2}. \]  

(B.18)

which for small enough \( \alpha \) is always lower than \( \alpha \mu^{(1)} \).
APPENDIX C
Algebraic Connectivity vs. Betweenness for Small Graphs

The following plots offer additional material to that of Section 4.3. The sequence of images depicts the relation between the algebraic connectivity and the Full, Brandes, and Empty link betweenness of small graphs (i.e. graphs with 4 to 9 nodes). As proved in Section 4.3 of Chapter 4, the only two lower bounds that always hold for any general graph are the ones based on maximum Brandes betweenness and maximum Full betweenness, i.e. \( \mu_{N-1} \geq \frac{N}{\max_{i \in E} r_i^{B_T}} \) and \( \mu_{N-1} \geq \frac{N}{\max_{i \in E} r_i^{F_T}} \), which correspond to the right hand images of the middle and bottom rows, respectively.

No sampling process was performed to explore the graph space, but the complete space was exhaustively processed, thanks to public algorithms by McKay [205]. Although we have no analytical expressions for all the illustrated values, the exhaustive nature of our exploration provides a first order approximation about their quality.
Figure C.1: \( N = 4 \) nodes. The cyan line illustrates the equality \( \mu_{N-1} = N/r_1 \), thus graphs lying below the cyan line obey the respective inequality, points lying above do not. The inset graph displays a contour plot of the data: blue contours high density, red contours low density.
Figure C.2: $N = 5$ nodes. The cyan line illustrates the equality $\mu_{N-1} = N/r_1$, thus graphs lying below the cyan line obey the respective inequality, points lying above do not. The inset graph displays a contour plot of the data: blue contours high density, red contours low density.
Figure C.3: $N = 6$ nodes. The cyan line illustrates the equality $\mu_{N-1} = N/r_1$, thus graphs lying below the cyan line obey the respective inequality, points lying above do not. The inset graph displays a contour plot of the data: blue contours high density, red contours low density.
Figure C.4: $N = 7$ nodes. The cyan line illustrates the equality $\mu_{N-1} = N/r_t$, thus graphs lying below the cyan line obey the respective inequality, points lying above do not. The inset graph displays a contour plot of the data: blue contours high density, red contours low density.
\textbf{Figure C.5:} $N = 8$ nodes. The cyan line illustrates the equality $\mu_{N-1} = N/r_1$, thus graphs lying below the cyan line obey the respective inequality, points lying above do not. The inset graph displays a contour plot of the data: blue contours high density, red contours low density.
Figure C.6: $N = 9$ nodes. The cyan line illustrates the equality $\mu_{N-1} = N/r_1$, thus graphs lying below the cyan line obey the respective inequality, points lying above do not. The inset graph displays a contour plot of the data: blue contours high density, red contours low density.
List of Abbreviations

BA  Barabási-Albert graph type
BA-r  Barabási-Albert rewired (graph generator)
Blog  Web log
CDF  Cumulative (probability) Density Function
CCDF  Complementary Cumulative (probability) Density Function
CPD  Central Point of Dominance
ENEA  Ente Nazionale per l’Energia Atomica (National Agency for Atomic Energy)
ENISA  European Network and Information Security Agency
ER  Erdős-Rényi graph type
GC  Giant component
GLP  Generalized Linear Preference (graph generator)
HH  Havel Hakimi
IP  Internet Protocol
ISP  Internet Service Provider
MOTIA  Modelling Tools for Interdependence Analysis for ICT Systems
NWO  The Netherlands Organisation for Scientific Research
PDF  Probability Density Function
PCC  Pearson product-moment Correlation Coefficient
PLRG  Power Law Random Graph (graph generator)
Pol  Political
PP  Path Protection
PS  Protection Scheme
QoS  Quality of Service
LA  Lattice
RG  Risk Group
RL  Rope-Ladder
RLP  Rope Ladder Protection
URG  Universal Risk Group
WS  Watts-Strogatz graph type
Only when explicitly mentioned, will we deviate from the standard notation and symbols outlined here.

In general, random variables and matrices are written with capital letters, while complex, real, integer, etc., variables are in lower case. For example, $X$ refers to a random variables, $A$ to a matrix, whereas $x$ is a real number and $z$ is a complex number. Usually, $i,j,k,l,m,n$ are integers. Operations on random variables are denoted by $[,]$, whereas $(,)$ is used for real or complex variables. A set of elements is embraced by $\{,\}$.

The reader should be aware that I highly value consistency, i.e. making sure every symbol is uniquely and unequivocally defined. However, the current landscape of uncoordinated and overwhelming nature of graph metrics inevitably leads to a symbol overload, e.g. $r$ was coined as both, assortativity and weighted betweenness. In order to stay true to the surveying nature of Chapter 2, I decided to retain the symbols that are commonly found in the scientific literature at the cost of reduced consistency. Nonetheless, I believe that retaining the original symbols as much as possible is a healthy practice, which avoids an excess of redundant redefinitions, and helps consolidating a standardized taxonomy. Nevertheless, I made sure that within this thesis there is no room for ambiguity, and the definition of any symbol is always made be evident by its context - or by footnotes.

**Linear Algebra**

\[ A \quad n \times m \text{ matrix} \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} \]

\[ \det A \quad \text{determinant of a square matrix } A; \text{ also denoted by } |A| \]
trace($A$) $\sum_{j=1}^{N} a_{jj}$: sum of diagonal elements of $A$

diag($a_k$) diag($a_1, a_2, \ldots, a_n$): a diagonal matrix with diagonal elements listed, while all off-diagonal elements are zero

$A^T$ transpose of a matrix, the rows of $A^T$ are the columns of $A$

$A^*$ matrix in which each element is the complex conjugate of the corresponding element in $A$

$A^H = (A^*)^T$: Hermitian of matrix $A$

$c_A(x) = \det(A - xI)$: characteristic polynomial of $A$

$\text{adj}A = A^{-1}\det A$: adjugate of $A$

e$_j$ basic vector, all components are zero, except for component $j$ that is 1

u all-one vector

$J$ all-one matrix

$I = \text{diag}(1,1,\ldots,1)$: identity matrix

$\delta_{kj}$ Kronecker delta, $\delta_{kj} = 1$ if $k = j$, else $\delta_{kj} = 0$

---

**Graph Theory**

\[ G = (\mathcal{N}, \mathcal{L}): \text{a graph} \]

\[ \mathcal{L} \] set of links of a graph

\[ \mathcal{N} \] set of nodes of a graph

\[ L = |\mathcal{L}|: \text{number of links in a graph} \]

\[ N = |\mathcal{N}|: \text{number of nodes in a graph} \]

\[ K_N \] the complete graph with $N$ nodes

\[ K_{n,m} \] the complete bi-partite graph with $N = n + m$ nodes

\[ A \] adjacency matrix of a graph

\[ B \] incidence matrix of a graph

\[ Q = BB^T \] laplacian matrix of a graph

\[ \Delta_G \] the number of triangles of a graph

\[ \Delta \] diag($d_1, d_2, \ldots, d_N$): diagonal matrix of the nodal degrees

\[ \{\lambda\}_{1 \leq k \leq N} \] set of eigenvalues of $A$, ordered as $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$

\[ \{\mu\}_{1 \leq k \leq N} \] set of eigenvalues of $Q$, ordered as $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_N$

\[ N_k \] total number of walks with length $k$

\[ W_k \] total number of closed walks with length $k$

\[ P_{i\to j} \] path between nodes $i$ and $j$, defined as a sequence of links

\[ P_{i\to j}^* \] shortest path between nodes $i$ and $j$, defined as a sequence of links

\[ S_{P_{i\to j}} \] set of all possible shortest paths between nodes $i$ and $j$

\[ v_i(t) \] Probability of node $i$ to be infected at time $t$
**Probability Theory**

- \( \Pr[X] \): probability of event \( X \)
- \( \mathbb{E}[X] \): expectation of the random variable \( X \)
- \( \text{Var}[X] = \sigma_X^2 \): variance of the random variable \( X \)
- \( f_X(x) = \frac{\partial F_X(x)}{\partial x} \): probability density function of \( X \)
- \( F_X(x) = \Pr[X \leq x] \): probability distribution function of \( X \)
- \( \varphi_X(z) \): probability generating function of \( X \)

\[ \varphi_X(z) = E[z^X] \] when \( X \) is a discrete random variable

\[ \varphi_X(z) = E[e^{-zX}] \] when \( X \) is a continuous random variable

- \( X_{k1 \leq k \leq m} \): \( k \)-th order statistics, \( k \)-th smallest value in the set \( X_{k1 \leq k \leq m} \)
- \( P \): transition probability matrix (Markov process)
- \( 1_{\{x\}} \): indicator function: \( 1_{\{x\}} = 1 \) if the event or condition \( \{x\} \) is true, else \( 1_{\{x\}} = 0 \).

**Topological Metrics**

- \( B_i \): betweenness of element \( i \)
- \( C \): average clustering coefficient
- \( c_i \): clustering coefficient of node \( i \)
- \( C_i \): closeness of node \( i \)
- \( C_{ij} \): communicability of a node pair \((i, j)\)
- \( D \): degree, or diameter
- \( d_i \): degree of node \( i \)
- \( d_{\text{max}} \): maximum degree in a graph
- \( d_{\text{min}} \): minimum degree in a graph
- \( d_{(j)} \): the \( j \)-th largest nodal degree in a graph
- \( \text{dist} \): distortion
- \( e_h \): average \( h \)-hops expansion
- \( e_i(h) \): \( h \)-hops expansion of element \( i \)
- \( E \): efficiency
- \( G_C \): giant component, or largest component
- \( k_i \): coreness of node \( i \)
- \( L_i \): leverage centrality of node \( i \)
- \( m(G) \): number of nodes in the giant component
- \( H_{i \rightarrow j} \): number of hops in \( P_{i \rightarrow j} \)
- \( H \): hopcount, or entropy
- \( \Pr[d_i, d_j] \): joint degree
- \( Q \): modularity
- \( R \): radius, resistance, or robustness value (i.e. R-value)
- \( r \): assortativity
- \( r_i \): weighted betweenness of element \( i \)
- \( S_i \): subgraph centrality of node \( i \)
- \( t \): distortion
- \( x_i \): \( i \)-th largest eigenvector (of \( A \) or \( Q \), depending on the context)
- \( \mathcal{X} \): chromatic number
Topological Metrics: Greek Symbols

εᵢ  eccentricity of node i
δ  degree diversity
γ  girth
κᵢ(G)  node connectivity (also known as vertex connectivity)
κₑ(G)  link connectivity (also known as edge connectivity)
λ₁ = maxₓ=1₂ₓᵀQₓ, spectral radius
µₙ₋₁ = minₓ=₁₉ₓ=₀ₓᵀQₓ, algebraic connectivity
ρ  = λ₁, spectral radius
Θ  average cyclic coefficient
Θᵢ  cyclic coefficient of node i
τ  effective spreading rate
Φₖ  rich club coefficient
Publications by the Author

Peer-Reviewed Conference Publications


Journal and Magazine Publications


JS.1 J. Martin Hernandez, Gregorio D’Agostino, H. Wang, and P. Van Mieghem “On Synchronization of Interdependent Networks”, *Currently under review*.

JS.2 Z. Li, J. Martin Hernandez, and P. Van Mieghem “Weighted Betweenness and the Algebraic Connectivity”, *Ongoing work*.

**TUDelft Technical Reports**


Relations to this Thesis

This thesis consists of previously published publications by the present author. Only the introductory and conclusive pieces were rewritten, so as to provide a cohesive scientific discourse. Table C.1 provides a relation between publications and chapters of this thesis.

<table>
<thead>
<tr>
<th>Publication</th>
<th>Intro.</th>
<th>Ch. 2</th>
<th>Ch. 3</th>
<th>Ch. 4</th>
<th>Ch. 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>C.1</td>
<td></td>
<td></td>
<td></td>
<td>⬤</td>
<td></td>
</tr>
<tr>
<td>C.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>⬤</td>
</tr>
<tr>
<td>C.3</td>
<td></td>
<td></td>
<td>⬤</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C.4</td>
<td></td>
<td>⬤</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C.5</td>
<td></td>
<td></td>
<td>⬤</td>
<td></td>
<td>⬤</td>
</tr>
<tr>
<td>J.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>⬤</td>
</tr>
<tr>
<td>J.2</td>
<td></td>
<td></td>
<td></td>
<td>⬤</td>
<td></td>
</tr>
<tr>
<td>J.3</td>
<td></td>
<td></td>
<td>⬤</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JS.1</td>
<td></td>
<td></td>
<td></td>
<td>⬤</td>
<td></td>
</tr>
<tr>
<td>JS.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>⬤</td>
</tr>
<tr>
<td>R.1</td>
<td></td>
<td>⬤</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R.2</td>
<td></td>
<td></td>
<td></td>
<td>⬤</td>
<td></td>
</tr>
<tr>
<td>R.3</td>
<td></td>
<td></td>
<td></td>
<td>⬤</td>
<td></td>
</tr>
<tr>
<td>R.4</td>
<td></td>
<td></td>
<td></td>
<td>⬤</td>
<td></td>
</tr>
</tbody>
</table>

Table C.1: Relation between chapters of the thesis and the list of publications by the author. The meaning of the bullets is: ⬤ major relation, ○ minor relation.
Curriculum Vitae

Javier Martín Hernández was born in Barcelona (Spain) on March 25, 1982. In October 2006, he graduated in Spain from Universitat Politècnica de Catalunya (UPC), with a MSc. degree in Telecommunications Engineering. Thanks to the European program ERASMUS, he invested the last year of his master studies abroad, in The Netherlands (2006). His MSc. dissertation was supervised by Dr. H. Wang and concerns the comparison of graph generators, which resulted in a conference publication.

Immediately after collecting his engineering degree, Javier moved to The Netherlands for six years (2007-2013) where he became a PhD candidate in the Networks Architectures and Services group at Delft University of Technology (TUDelft). Under the daily supervision of Prof. P. Van Mieghem, Javier published 9 papers in peer-reviewed international conferences and journals, worked in collaboration with a multinational IT provider (NEC), successfully guided MSc. student dissertations, assisted Prof. P. Van Mieghem at teaching lectures on Performance Analysis, and attended international seminars on state of the art topics. The bulk of Javier’s work consisted of approaching network robustness from a graph theoretical point of view, in order to increase the strength of critical infrastructures such as the Internet.

By the beginning of 2013, right before writing this very thesis, the European project MOTIA enabled his stay in Italy as an “international exchange researcher” at ENEA, the Italian National Agency for New Technologies. Mentored by dr. G. D’Agostino, he delved into the topic of interdependent networks. This stay provided Javier with invaluable connections, industry experience, and a handful of chocolates (hence the smiley face).
Acknowledgements

I would like to express my deepest gratitude to Piet Van Mieghem, for diligently helping me \textit{challenge my future}, and for never giving up on me. In no way I am exaggerating when I say that he has been a parenting figure over the last six years. I will miss you.

Although it sounds \textit{cliché}, this thesis would not have been possible without the help as well as presence of many people, including but not limited to: mom and dad for their infinite love; Gregorio D’Agostino for driving me to Rome in \textit{the world’s oldest car}; \textit{Gladiator} Junglist, for being there through thick and thin; my Spanish professors Roc Messegueur, Manel Gasulla, Sebastià Sallent, Juan Carlos Aguado, Cristina Cervelló, Jorge Olivella, Antoni Oller, Silvia Ruiz, Francesc Tarrés, but specially Salvatore Spadaro, for encouraging me to travel abroad; Laura Baumann, Marjon Verkaik and Wendy Murtinu for being the best secretaries in the world; Jingwei Wu for trusting my unseasoned skills; all committee members for their precious time and valuable feedback; and last but not least, my dearest office mates, Ruud van de Bovenkamp, Cong Li, Anteneh Ayalew Beshir, Siyu Tang, and Yue Lu for tolerating my calamities.

Before this page runs too long, I want to at least mention a special group of people: Christian, Norbert, Huijuan, Rob, Stojan, Przemek, Wynand, Dajie, Bo, Niels, Edgar, Evangelos, Rob, Fernando, Almerima, Jasmina, Tom, Ebisa, Farabi and Cheng.

To all of You, \textbf{thanks}!

\textit{Javier Martín Hernández}
Delft, June 2013