Robustness of Networks

Robustness of Networks

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof.dr.ir. J.T. Fokkema,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op dinsdag 8 september 2009 om 10.00 uur

door

Huijuan WANG

elektrotechnisch ingenieur
geboren te Harbin, Heilongjiang Province, China.

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. I.G.M.M. Niemegeers,	Technische Universiteit Delft
Prof.dr.ir. K.I. Aardal,	Technische Universiteit Delft
Prof.dr.ir. W.H. Haemers,	Universiteit van Tilburg
Prof.dr. C.J. Stam,	Vrije Universiteit Amsterdam
Dr.Eng. C.M. Scoglio,	Kansas State University
Dr.ir. R.E. Kooij,	Technische Universiteit Delft

Printed in The Netherlands

to my parent

Contents

Summary

Our society depends more strongly than ever on large networks such as transportation networks, the Internet and power grids. Engineers are confronted with fundamental questions such as "how to evaluate the robustness of networks for a given service?", "how to design a robust network?", because networks always affect the functioning of a service. Robustness is an important issue for many complex networks, on which various dynamic processes or services take place. In this work, we define robustness as follows: a network is more robust if the service on the network performs better, where performance of the service is assessed when the network is either (a) in a conventional state or (b) under perturbations, e.g. failures, virus spreadings etc. In this thesis, we survey a particular line of network robustness research within our general framework: robustness quantification, optimization and the interplay between service and network.

Significant progress has been made in understanding the relationship between the structural properties of networks and the performance of the dynamics or services taking place on these networks. We assume that network robustness can be quantified by a topological measure of the network. A brief overview of the topological measures is presented. Each measure may represent the robustness of a network with respect to a certain performance aspect of a service. We focus on the measure known as algebraic connectivity. Evidence collected from literature shows that the algebraic connectivity characterizes network robustness with respect to synchronization of dynamic processes at nodes, random walks on graphs and the connectivity of a network. Moreover, we illustrate that, on a given diameter, graphs with large algebraic connectivity tend to be dense in the core and sparse at the border. Such structures distribute traffic homogeneously and are thus robust in terms of traffic engineering.

How do we design a robust network with respect to the metric algebraic connectivity? First, the complete graph has the maximal algebraic connectivity, while its high link density makes it impractical to use due to the cost of constructing links. Constraints on other network features are usually set up to incorporate realistic requirements. For example, constraint on the diameter may guarantee certain end-to-end quality of service levels such as the delay. We propose a class of clique chain structures which optimize the algebraic connectivity and many other robust features among all graphs with diameter D and size N. The optimal graph within the class can be determined either

analytically or numerically. Second, complete replacement of an existing infrastructure is expensive. Thus, we design strategies for robustness optimization using minor topological modifications. These strategies are evaluated in various classes of graphs.

The robustness quantification, or equivalently, the association of the performance of a service with a topological measure, may be implicit. In this case, we explore the interplay between topology and service in determining the overall performance. Many services on communications and transportation networks are based on shortest path routing. The weight of a link, such as delay or bandwidth, is generally a metric optimized via shortest path routing. Thus, link weight tuning, a mechanism to control traffic, is also considered as part of the service. The interplay between service (shortest path routing and link weight tuning) and topology is investigated for the following performance aspects: (a) the structure of the transport overlay network, which is the union of shortest paths between all node pairs and (b) the traffic distribution in the overlay network. Important new findings are (i) the universal phase transition in overlay structures as we tune the link weight structure over different classes of networks and (ii) the power law traffic distribution in the overlay networks when link weights vary strongly in various classes of networks. Furthermore, we consider the service that measures a network topology as the union of shortest paths among a set of testboxes (nodes). The measured topology is a subgraph of the overlay network, which is again a subgraph of the actual network. The performance in terms of the sampling bias of measuring a network topology is investigated. Our work contributes substantially to a better understanding of the effect of the service (testbox selection) and the actual network structure on the performance with respect to sampling bias. Our investigations on the interplay between service and network reveal again the association between the performance of a service and certain topological feature, and thus, contribute to the quantification of network robustness.

The multidisciplinary nature of this research lies not only in the presence of robustness issues in many complex networks, but also in that advances in other disciplines such as graph theory, combinatorics, linear algebra and statistical physics are widely applied throughout the thesis to study optimization problems and the performance of large networks.

Chapter 1

Introduction

1.1 Robustness of complex networks

A network specifies how items, called nodes, are interconnected or related to other nodes by links. Many complex systems can be modeled by networks to capture the possibly inhomogeneous patterns of interactions within complex systems. The brain, where billions (10^{11}) of neurons are interconnected by synapses, is perhaps the most challenging network [85]. In a metabolic reaction network, nodes are molecular compounds, which play a role as educts or products in metabolic reactions. Directed links connect educts and products in a metabolic reaction. The Internet, one of the most booming communication networks, is generally a network of interconnected computers. Upon the Internet, The World Wide Web, the largest artificial network, has developed as billions[1] (2005) of documents connected by hyperlinks, which are mutual references in these documents. Humans have the inborn desire to be sociable. Collaborator networks range widely from acquaintance networks, movie actor networks to coauthorship networks, where people are connected through the acts of being friends, playing in a same movie, and writing a paper together, respectively. Topologies of networks [34] range from biological networks such as gene regulatory networks, metabolic networks, artificial networks like the Internet, the WWW to social networks, such as paper citations, collaboration networks, etc. Correspondingly, the study of networks pervades many fields of science.

The terminology 'complex networks' was initially employed to address the non-trivial topological features of networks. One research area that studies the complexity of network topologies is the network modeling. Traditionally, networks have been modelled as random graphs [36], where nodes are linked in a random manner (Erdös-Rényi, 1959).

---

[1] In 2005 Yahoo! announced that its search engine index contained more than 19.2 billion documents. Given that Yahoo! does not cover all documents on the Web, it is reasonable to expect that the real number is higher.

The availability of powerful computers has made it possible to explore the structures of large-scale empirical networks. Two important discoveries of modern network theory are the small-world [99] and scale-free [8] characteristics, exhibited by many real-world networks. The corresponding small-world model by Watts and Strogatz (1998) and the scale-free network model by Barabási and Albert (1999) have promoted a tremendous amount of research activity in various disciplines. In a small-world network, any two nodes can be reached from each other via a small number of links despite the large size of the network, just as in a random graph. However, a small world network is much more highly clustered than a random graph, in the sense that direct neighbors of a node are more likely to be mutually connected. In a scale-free network, some nodes (hubs) are connected with a higher number of links than the others. Although the Erdös-Rényi random graph fails to capture the small-world and scale-free features displayed by empirical networks, it is still being extensively investigated due to its analytical beauty. Network modeling is, naturally, the first research issue that has been faced. It is also driven by our desire to further understand the dynamical or functional behavior of networks, because the network structure always affects its functioning.

Upon each network, various services, or more generally, dynamic processes are deployed. In communication networks, a dynamic process refers to a service provided by the network. For example, a communication service like email transports a message from a source to a destination node over the Internet. Synchronization is a process where systems (the nodes) adjust a given property of their motion due to a suitable coupling configuration, or to an external force. The synchronization phenomena that two pendulum clocks hanging at the same beam were able to perfectly synchronize their phase oscillations, was discovered by Christian Huygens in the 17th century [51]. This synchronization process is carried out over a network, that simply consists of two nodes (the pendulum clocks) connected by a link (the coupling between clocks). Other examples are neuron transport in the brain, financial transactions on a stock market, interactions in social networks, etc. Later, we will use the terminology 'service' for the general term 'dynamic process', to emphasize communications networks.

The topology has important influence on the service. A fundamental research question is "Given a network, is it robust for a given service?" Such question is, however, ill-posed, because, what is the robustness of a network? While humans intuitively have a notion of robustness, it remains vague and difficult to quantify. *Robustness* rooted in the Latin word *robur*, meaning "strength" or "hard wood" like oak, a symbol of strength. (a) When designing a network from scratch, robustness can be defined with respect to the final purpose of the network. For example, for a live stream service, the network should transport the data in real time (bearing small delay), and the loss of some data can be tolerated. (b) Furthermore, we can also define robustness as the capability of a network to withstand perturbations like failures, viruses and attacks. For biologists, robustness is the survivability of a cell under extreme conditions and frequent internal errors. In the view of a sociologist, it is the stability of the human

society in face of war and changes of policy. In this thesis, we generalize the definition of robustness to "a network is more robust if the service on the network performs better". Hence, the network robustness is evaluated by the performance of the service in (a) at a conventional condition and in (b) when the network is perturbed by all kinds of challenges, such as failures and malicious attacks.

The interplay between the network topology and service causes much more complexity than the topology itself. From this perspective, our study of network robustness for a service belongs to the broader research on complex systems. In fact, an accurate and complete description of complex systems is the greatest challenge today in all science [102]. Therefore, different fields have suppressed certain complications while emphasizing others [87]. For example, network modeling highlights the complication in structure while neglecting the dynamics on the network. In nonlinear dynamics, simplified structures, like geometrically regular networks, are favored, which allows us to sidestep the complication in topology and to concentrate on the complexity of the dynamics. The network robustness for a given service, incorporates both topology and service. Thus, our methodology is a compromising approach between network modeling and nonlinear dynamics. In the next section, we will explain our framework, which systematically investigates the network robustness.

1.2 Framework for network robustness

The goal of this framework is, for a given service, how to evaluate the robustness of a network and how to design a robust network. This framework, as shown in Figure 1.1, investigates three aspects of network robustness: robustness quantification, optimization and the interplay between network and service.

A network is more robust if the service performs better. Hence, the quantification of robustness is tightly coupled to the specific performance aspects of a particular service. We assume that the performance of a service can be characterized by or strongly related to a topology related metric $R(G)$ of the network G. Thus, the network robustness can be quantified by a topological metric $R(G)$. For example, communications networks require certain efficiency in resource usage. The resource consumed by a flow between a node pair is equal to the amount of the traffic times the number of links that the flow traverses. Accordingly, the average hopcount, the average number of links in a path, is able to characterize the network robustness in terms of the efficiency in resource usage. The service determines which topological metric R can quantify the network robustness.

Given the metric R that characterizes the network robustness, we investigate the following optimization problems: (a) design a robust network that maximizes the metric R, or furthermore, determine the network that maximizes the metric R given the constraints on other metrics; (b) how to optimize the robustness of an existing network via a minor topology modification, e.g. the addition of links? Practical network design
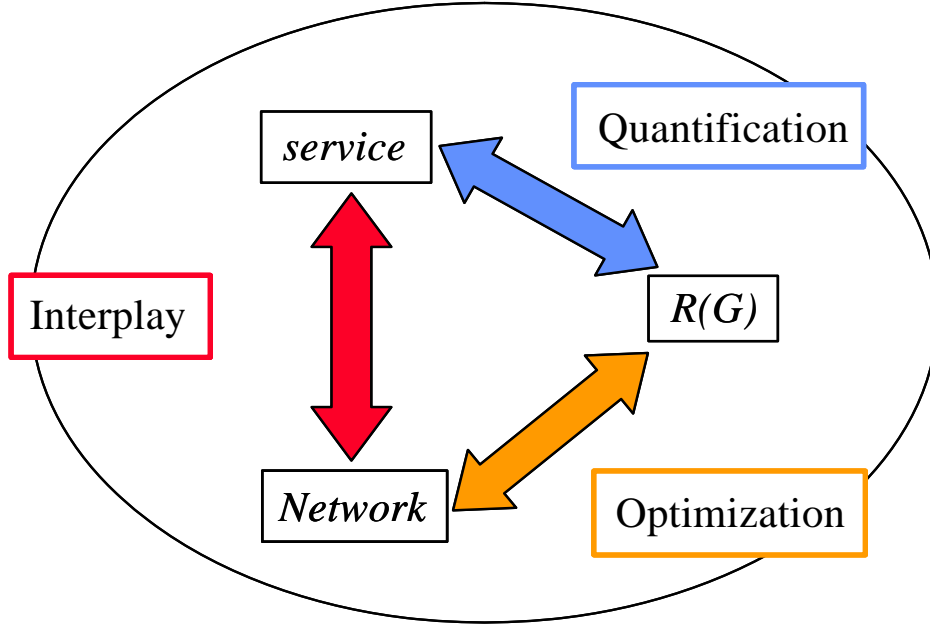
Figure 1.1: The framework to investigate network robustness.

requirements, such as the economical concerns on the number of links, can be taken into account via the constraints on other metrics in (a) and on the minor topology modification in (b).

For complicated services, however, the association of the performance of the service with a topology measure $R(G)$ may be non-trivial. Thus, we investigate the interplay between network and service in determining the performance of the service. This interplay study, in turn, may reveal the relation between the performance of the service and a topological metric R. Thus, the robustness quantification can be better understood.

1.3 Positioning of the framework

We first generally define network robustness in relation to the performance of a service, when the network is either in a normal condition or under disruptions. Many criteria to compare network robustness exist in literature. Complex ones take into account the performance of a service under different levels of disruptions. They are, however, not discussed in this thesis, mainly because analytic computations are basically intractable [94].

Second, we assume that the performance of a service can be approximated by a certain topological metric R, which, thus, quantifies network robustness. The robustness of networks can be, actually, more precisely evaluated by directly computing a

performance measure[2] such as a stochastic measure (average, minimum, 95% quantile, variance, etc.) of the delay, available capacity and so on. For example, a real-time communication service can tolerate only a short end-to-end delay. The performance measure, the average delay, represents the network robustness more precisely than its approximation, the topological measure, the average hopcount. We prefer a topological metric R to quantify the network robustness, because of the following reasons. (a) It is difficult to optimize the performance of a service via system modifications, if both the network topology and the service aspects can be modified. We emphasize the role of the network topology in the performance of a service. The quantification via $R(G)$ provides a direct criterion of network structure optimization. (b) Stochastic measures of the performance, like the average delay, are difficult to obtain due to the highly dynamic nature of the service. For example, the delay over each link, is unknown and widely varying over both time and the links in the network. (c) Significant progress has been made in understanding the relationship between the structural properties of networks and the nature of the dynamics or services taking place on these networks. For instance, the 'synchronizability' of complex networks of coupled oscillators is shown to be determined by a topological measure in graph spectral analysis.

Finally, we relate the performance of a service to only one topological metric R in the quantification of network robustness. However, each metric reflects partially the feature of a network. It is more precise to characterize network robustness by $\sum_{k=1}^{m} s_k R_k$, a set of topological metrics $[R_1, R_2, ..., R_m]$ coupled by a weighing vector $[s_1, s_2, ..., s_m]$. The components of the weighing vector s reflect the importance of the corresponding topology metrics for the service. The weighing vector has to be unique and stable such that the comparisons of networks in terms of robustness are fair. However, the correlations of topological metrics are topology dependent, which may potentially modulate the weighing vector when different networks are evaluated. For example, we consider three metrics $[R_1, R_2, R_3]$ and a weighing vector $[s_1, s_2, s_3]$. Suppose that the three metrics are independent in graph G_1. Thus, the robustness can be quantified by $s_1 R_1 + s_2 R_2 + s_3 R_3$. However, in graph G_2, the metric R_3 is dependent on R_1 and R_2, $R_3 = a R_1 + b R_2$. The robustness of G_2 becomes $(s_1 + a) R_1 + (s_2 + b) R_2$. This shows that m is effectively 2, instead of 3, and that the weighing vector s is modified by the topology. The dependence between metrics, which is topology dependent, seems a hard, inherent challenge of the robustness problem. Our framework chooses an elementary start, only one metric R. This limitation is partially compensated (see Section 1.2), by the optimization (a), where we are looking for a network maximizing metric R, subject to constraints on other metrics R_1, R_2... The correlation of metrics is also addressed in the optimization (b), where the strategy of topology modification to optimize robustness R is based on certain topology features. The correlation between the metric R and these topology features of a network are dependent on topology. Thus, a strategy may perform differently

---

[2]In this thesis, the word "measure" and "metric" are used interchangeably.

for distinct networks.

1.4   Thesis outline

We first introduce in Chapter 1 our framework to study the network robustness for a given service.

The main body of this thesis consists of 7 chapters and is structured into 3 parts: robustness quantification, optimization and the interplay between network and service. The last chapter of this thesis, Chapter 9 highlights the main conclusions drain from this thesis.

1.4.1   Part I: Quantification of Network Robustness

Chapter 2 introduces the essential aspects that describe a network: topology, link weight, and service. Network models, link weight structures and services, that will be investigated throughout this thesis, especially in Part III, are elaborated in this chapter. **Chapter 3** first presents a brief overview of practically important topological measures. Without losing generality, any of the topological measures may possibly represent certain network robustness for a service. Thus, the robustness represented by each metric is not discussed in detail. Instead, we exemplify what kind of network robustness can be represented by two metrics: the hopcount and the algebraic connectivity.

In addition, the robustness quantification with respect to specific performance aspects of a service is also revealed in Part III.

1.4.2   Part II: Robustness Optimization

In this part, we focus on the case where network robustness is quantified by the algebraic connectivity.
Chapter 4 examines the graph maximizing the algebraic connectivity, subject to the constraint on the diameter. We propose a class of graphs, which can achieve many robust features: the maximal number of links, the minimum average hopcount, and more interestingly, the maximal of any Laplacian eigenvalue (including the algebraic connectivity) among all graphs with N nodes and diameter D. The graph optimizing a given robustness measure, e.g. algebraic connectivity, can be either analytically determined or be searched out of the class of graphs. Features of these optimal graphs as well as of the extreme (maximal or minimal) values of the robust measures are further explored.
Chapter 5 investigates how to refine a network to optimize its robustness. In particular, where should we add a link to a network such that the algebraic connectivity can be increased the most? Exhaustive searching for the optimal link addition is computationally infeasible in large networks. Hence, we propose two strategies, which are

compared with random link addition and with the optimal link addition if the network is not too large. The investigation on adding one link will provide insights on how to dynamically add a set of links one by one for robustness optimization.

1.4.3 Part III: Interplay between network and service

When the quantification of network robustness by a topological metric is not straightforward, we investigate the interplay between network structure and service in determining the overall performance. We focus on the fundamental service in communications networks, where traffic is routed along the shortest paths. The link weight structure, which is also regarded as part of the service, can be tuned. Our work is motivated by three potential applications: overlay networks, Internet topology interference and traffic engineering.

Chapter 6 examines the structure of the overlay network, the union of shortest paths between all node pairs, which is the maximally observable part of a network. As we tune the link weight structure, a universal phase transition in the overlay structure has been observed for various networks.

Chapter 7 inspects the bias phenomenon of sampling a network by shortest paths among a set of m testboxes. The sampling bias is shown to depend on both the network characteristics like the link density as well as the service (the measurement), such as the selection of the m testboxes.

Chapter 8 investigates the traffic distribution as we tune the link weight structure over different networks.

Therefore, the performance explored in this part is the structure of the overlay network, the sampling bias of measuring a network topology and the traffic distribution.

1.5 List of publications

In this section, we give the list of papers presented in this thesis.

- H. Wang and P. Van Mieghem, *Graphs with given diameter maximizing the algebraic connectivity*, submitted to COMBINATORICA, 2008. (Chapter 3, 4)

- P. Van Mieghem and H. Wang, *Spectra of a new class of graphs with extremal properties*, submitted to SIAM Journal on Discrete Mathematics, 2008. (Chapter 4)

- H. Wang and P. Van Mieghem, *Algebraic connectivity optimization via link addition*, Bionetics 2008, Japan, November, 2008. (Chapter 5)

- P. Van Mieghem and H. Wang, *The observable part of a network*, IEEE/ACM Transaction on Networking, vol. 17, No. 1, pp. 93-105, 2009. (Chapter 6)

- H. Wang and P. Van Mieghem, *Sampling networks by the union of m shortest path trees*, submitted to Computer Networks, 2008. (Chapter 7)

- H. Wang, J. Martin Hernandez and P. Van Mieghem, *Betweenness centrality in weighted networks*, Physical Review E, vol. 77 (046105), April, 2008. (Chapter 8)

Part I

Quantification of Network Robustness

Chapter 2

Networks: topology, link weight and service

Before examining the set of topological measures that may characterize the network robustness in Chapter 3, we firstly introduce the essential aspects that describe a network: the topology, the link weight structure, and the service. Notations and terminologies in this thesis are based on graph theory and communications networking. We mainly elaborate the network models, the link weight structures and the services that will be investigated throughout this thesis. For a more general view about topology and dynamics, we refer to the review of Boccaletti et al. [12].

2.1 Network topology

The topology, structure or interconnection pattern of a network can be represented by a graph $G(N, L)$, in short G, which consists of a set $\mathcal{N}$ of N nodes interconnected by a set $\mathcal{L}$ of L links. For example, a complete graph K_N, or a clique consists of N nodes and $L = L_{\max} = \frac{N(N-1)}{2}$ links, where every node has a link to every other node. The topology of a network with N nodes can be represented by an *adjacency matrix* A, a $N \times N$ matrix consisting of elements a_{ij} that are either one or zero depending on whether there is a link between node i and j or not. We consider only undirected networks. Hence, the adjacency matrix is symmetric, i.e. $a_{ij} = a_{ji}$. Throughout this thesis, we will consider both network models which capture certain topological features observed in empirical networks as well as a set of real-world networks which represent the topology of various complex systems.

2.1.1   Erdös-Rényi random graphs

Traditionally, complex networks have been modeled as Erdös-Rényi random graphs $G_p(N)$, which can be generated from a set of N nodes by randomly assigning a link with probability p to each pair of nodes. The Erdös-Rényi random graph was initiated by Erdös and Rényi [36][37] in 1959. Due to their pioneering work, the random graph has become one of the most studied graphs [13]. Besides their analytic tractability, the Erdös-Rényi random graphs have also served as idealized structures for peer-to-peer networks [17], ad-hoc networks [50], gene networks, ecosystems [63] and the spread of disease or computer viruses [55].

The degree d_j of a node j in a graph is the number of links that connect with that node. The probability density function (pdf) of the degree D_{rg} of an arbitrary node in an Erdös-Rényi random graph $G_p(N)$ equals

$$\Pr[D_{rg} = k] = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

When the link density is smaller than the disconnectivity threshold, i.e. $p < p_c \sim \frac{\log N}{N}$ for large graph size N, the random graph $G_p(N)$ is almost surely disconnected, whereas $G_p(N)$ is almost surely connected if $p > p_c$. The transition around $p_c \sim \frac{\log N}{N}$ has a width of $O(\frac{1}{N})$.

2.1.2   Lattices

In D-dimensional lattices, all interior nodes have the same degree 2D, where D is the dimension. Here, we confine ourselves to the hyper-cube D-lattices in which each edge is of equal size. In this case, a 2D-lattice becomes a square lattice and a 3D-lattice equals a cubic lattice. Examples of a square lattice and a cubic lattice are shown in Figure 2.1. The lattice is the basic model of a transport network (Manhattan grid) and is crucial in percolation theory [84]. Moreover, it is frequently used to study the network traffic [79].

2.1.3   Small-world graphs

The small-world model proposed by Watts and Strogatz [99] encompasses the following two structural features as observed in real-world networks. Any two nodes can be reached within a small number of links despite the large size of networks. Nodes are well clustered in the sense that two direct neighbors of a node are more likely to be connected compared to those in random graphs. The small-world model starts by building a ring with N nodes and by joining each node with $2s$ nearest neighbors (s on either side of the ring). Upon the resulted ring lattice, each link connected to a clockwise neighbor is rewired to a randomly chosen node with a probability p_r, and
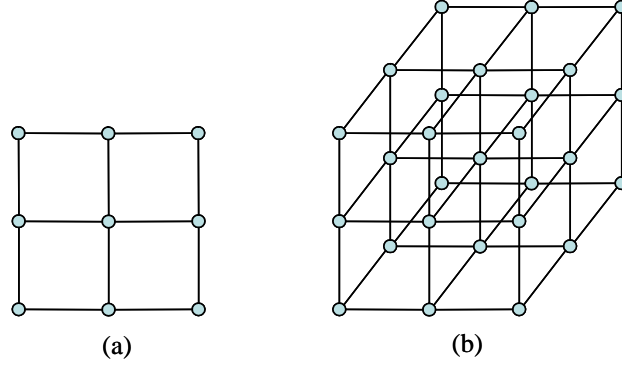
Figure 2.1: (a) The square lattice with $N = 9$ and (b) the cubic lattice with $N = 27$.

is preserved with a probability $1 - p_r$. The small-world graph interpolates between a ring or lattice ($p_r = 0$) and a random graph with the constraint that each node has the minimal degree s ($p_r = 1$). The small-world model is not discussed in this thesis. Instead, we consider the two extremes: the lattices and the random graphs.

2.1.4 Power law graphs

Power law graphs are random graphs specified by a power law degree distribution $\Pr[D = i] = ci^{-\tau}$, where c is a constant

$$c = \frac{1}{\sum_{i=1}^{N-1} i^{-\tau}}$$

The power law degree distribution is followed by many natural and artificial networks such as the scientific collaborations, the world-wide web and the Internet with exponents varying within the range $2 < \tau < 3$. Power law graphs are also called scale-free graphs, because the power law function $f(i) = ci^{-\tau}$ has the same functional form at all scales of i, namely, $f(\alpha i) = \beta f(i)$. We introduce two types of power law graphs: the Barabási-Albert (BA) [4] and Havel-Hakimi power law graphs [19][18].

Barabási-Albert power law graphs

The BA power law graph, we generated, starts with m nodes. At every time step, we add a new node with m links that connect the new node to m different nodes already present in the graph. The probability that a new node will be connected to node i in step t is proportional to the degree $d_i(t)$ of that node

$$P_i(t) = \frac{d_i(t)}{\sum_j d_j(t)} = \frac{d_i(t)}{2L_t}$$

where L_t is the number of links in step t. Figure 2.2 displays a BA power law graph with $m = 1$ and $N = 100$. The degree of a node is reflected by the size of the node. Hence, a BA graph contains a few highly connected nodes (hubs) compared to the others. The
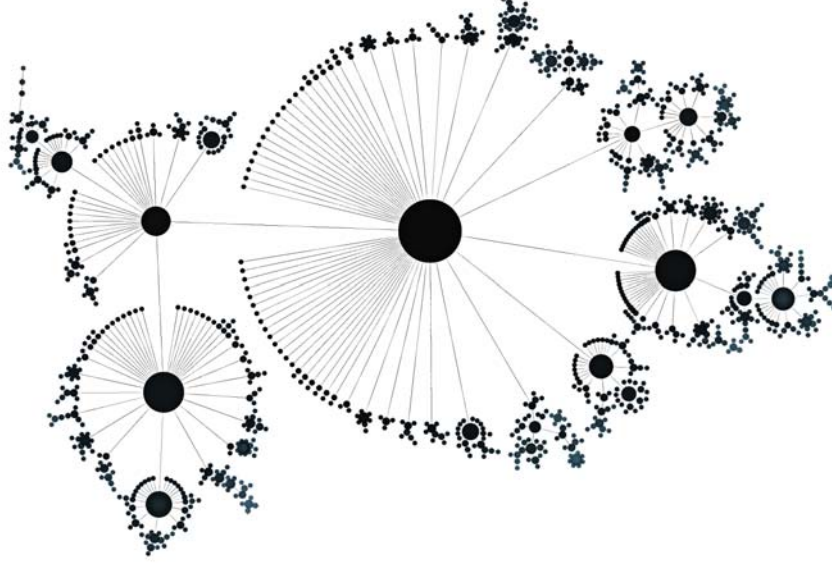
Figure 2.2: A Barabási-Albert power law graph with $m = 1$ and $N = 1000$.

degree distribution of a BA power law graph follows $\Pr[D = i] = ci^{-\tau}$, where $\tau_{BA} = 3$. This power law behavior in the Barabási-Albert model is observed for large N.

Havel-Hakimi power law graphs

In a Havel-Hakimi graph, the degree of all the nodes are assigned according to the power law distribution. A Havel-Hakimi graph is constructed by successively connecting the node of highest degree to other nodes of highest degree, resorting nodes by remaining degree, and repeating the process. The resulting graph has a high degree-associativity. Hence, the Havel-Hakimi power law graph has a relatively "dense core". Different from the BA power law graphs where $\tau_{BA} = 3$, the Havel-Hakimi graphs can be built according to power law degree distributions with various exponent τ. Moreover, it shows already power law behavior for small N.

2.1.5   Real-world complex networks

We also consider the real-world networks which represent the topology of various complex systems. Most of the data sets we have used are available publicly. They are complex networks from a wide range of systems in nature and society:

- the file sharing network Gnutella [10] snapshots (Crawl2) retrieved from firewire.com;

- the air transportation network representing the world-wide airport connections, documented at the Bureau of Transportation Statistics (http://www.bts.gov) database, and the connection between United States airports [24];

- the Western States Power Grid of the United States[100];

- the coauthorship network [77] between scientists posting preprints on the High-Energy Theory E-Print Archive between Jan 1, 1995 and December 31, 1999;

- the citation network[1] created using the Web of Science database: Kohonen and SciMet;

- the coauthorship network [78] of scientists working on network theory and experiment;

- the network representing soccer players association to Dutch soccer team [52], where a link exists between two players if they used to play in a same match;

- the adjacency network [78] of common adjectives and nouns in the novel David Copperfield by Charles Dickens.

- the Internet network at the level of autonomous systems [80];

- the network of American football games between Division IA colleges during regular season Fall 2000 [44];

A network is connected if there exists a path between each pair of nodes. We consider only the networks formed by the largest connected component of our real-world networks.

2.2 Link weight structure

Apart from the topological structure specified via the adjacency matrix A, the link between node i and j is further characterized by a link weight $w(i \rightarrow j)$, a non-negative real number, which quantifies a property of that link such as the delay incurred when traveling over that link, the distance, the capacity, etc. The link weight structure refers to the characteristics of link weights in a network. On one hand, the link weight can be considered as the intrinsic feature of the network, like the distance between two airports in the air transportation network. On the other hand, the link weights can be set up by a service provider, as a means to control or steer transport in man-made

---

[1]V. Batagelj and A. Mrvar (2006): Pajek datasets.

infrastructures such as the Internet. For example, from a traffic engineer's perspective, an Internet Service Provider (ISP) may want to tune the weight of each link such that the resulting path between a particular set of in- and egresses follow the desirable routes in its network. Here, we consider the latter, where the link weight structure is regarded as part of the services.

A *path* from node A to node B with $k-1$ hops or links is the node list $P_{A \to B}$ $= n_1 \to n_2 \to \cdots n_{k-1} \to n_k$ where $n_1 = A$ and $n_k = B$ and where $n_j \neq n_i \in \mathcal{N}$ for each index i and j. We confine ourselves to additive and strict positive link weight measures such that the weight of a path P is $w(P) = \sum_{(i \to j) \in P} w(i \to j)$. Thus, $w(P)$ equals the sum of the weights of the constituent links of P. Multiplicative measures (e.g. packet loss) can be transformed into additive weights by using the logarithm. The link weights are chosen *independently* of the topology. Although in some biological networks, the link weight or strength of a link is coupled to the structure of the underlying topology, in many man-made large infrastructures such as the Internet and WWW, the link weight structure can be chosen independently. The latter allows us to control or steer transport in the network. Therefore, we deem the undirected, independent identically distributed (i.i.d.) and additive link weights as a reasonable approximation in many large networks, with the exceptions of wireless networks[2]. Two types of link weights or two ways of link weight tuning are investigated: one-dimensional and multi-dimensional link weights.

2.2.1   One-dimensional link weights

Current best-effort routing simply computes appropriate paths based on a single, relatively static measure (e.g. the delay, the monetary cost, etc.). Each link is specified by a single weight measure, an i.i.d. random variable. The shortest path (SP) routing, which appears in many communications networks and is discussed in Section 2.3, is mainly sensitive to the smaller, non-negative link weights.

A *regular link weight distribution* $F_w(x) = \Pr[w \leq x]$ has a Taylor series expansion around $x = 0$,

$$F_w(x) = f_w(0)x + O(x^2)$$

since $F_w(0) = 0$ and $F'_w(0) = f_w(0)$ exists. A regular link weight distribution is thus linear around zero. Both the uniform distribution within $[0, 1]$, $F_w(x) = x 1_{x \in [0,1)} + 1_{x \in [1, \infty)}$ and the exponential distribution, $F_w(x) = 1 - \exp(-\lambda x)$, are regular link weight distributions. Apart from being attractive in a theoretical analysis, the uniform distribution on $[0, 1]$ is the underlying distribution to generate an arbitrary other distribution and is especially interesting for computer simulations. Hence, this distribution appears most often in network simulations and deserves – for this reason alone perhaps – to be studied. The uniformly distributed link weights will be frequently investigated in Part III.

---

[2] All nodes in the radio-range of some sending node (or base-station) are correlated by (a) the nature of electromagnetic waves and (b) wireless MAC protocols.

The simplest distribution of the link weight w with a distinct different behavior for small values is the polynomial distribution,

$$F_w(x) = x^\alpha 1_{x \in [0,1)} + 1_{x \in [1,\infty)}, \quad \alpha > 0, \tag{2.1}$$

where the indicator function 1_x is one if x is true else it is zero. More motivations to select a polynomial distribution is given earlier [72]. The corresponding density is $f_w(x) = \alpha x^{\alpha-1}$, $0 < x < 1$. The exponent

$$\alpha = \lim_{x \downarrow 0} \frac{\log F_w(x)}{\log x}$$

is called the *extreme value index* of the probability distribution. As shown in Figure 2.3,
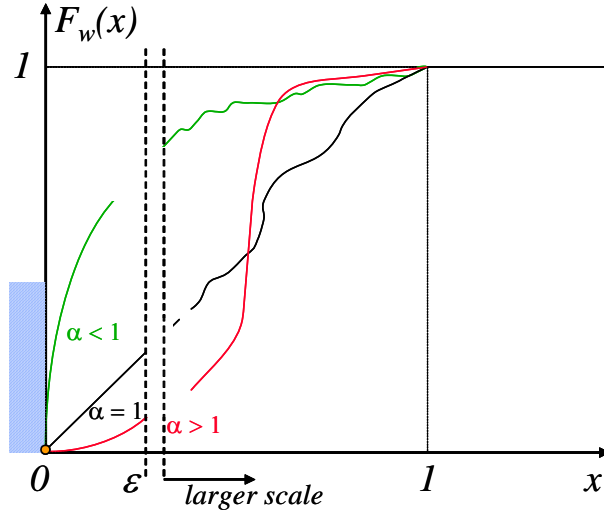
Figure 2.3: A schematic drawing of the polynomial distribution for three different α regimes. The scaling invariant property of the shortest path allows us to divide all link weights by the largest possible such that $F_w(1) = 1$ holds for all link weight distributions.

the link weight structure can be controlled by the *extreme value index* α. If $\alpha \to \infty$, it follows from (2.1) that $w = 1$ almost surely for all links. The $\alpha \to \infty$ regime is entirely determined by the topology of the graph because the link weight structure does not differentiate between links. The network can be considered as unweighted. When $\alpha \to 0$, all links will be close to 0, but, relatively, they differ significantly with each other. When $\alpha = 1$, the polynomial distribution becomes a uniform distribution where link weights are regular. Small link weights around zero, $w \in [0, \varepsilon]$ in Figure 2.3 dominantly influence the property of the shortest paths. The remainder link weights (denoted by the arrow with larger scale) only plays a second order role.

2.2.2   Multi-dimensional link weights

Several quality-of-service (QoS) based networking frameworks (e.g., IntServ, DiffServ, MPLS) have been extensively investigated. QoS routing takes into account multiple measures including both the applications' requirements and the availability of network resources. We confine ourselves to the case that each link is specified by a 2-dimensional link weight vector $\overrightarrow{w}(u \rightarrow v) = [w_1(u \rightarrow v), w_2(u \rightarrow v)]$, where the component w_i is a QoS measure such as delay, jitter, cost, etc. The importance of uniformly distributed link weights is discussed in Section 2.2.1. Furthermore, specific dependencies or correlations exist between QoS measures due to e.g. Weighted Fair Queueing scheduling. Hence, we investigate the link weight structure where the two vector components are correlated uniformly distributed random variables $\in [0, 1]$ with correlation coefficient ρ [58]. We tune the correlation coefficient $\rho \in [-1, 1]$ to vary the link weight structure.

2.3   Service

We investigate an elementary service, where the traffic is routed along the shortest path. The shortest path problem on networks is of importance since the purpose of many real networks is to provide efficient traffic route between nodes. Routing in communication networks is based on shortest paths (or the best approximation due to e.g. the distracting influence of BGP) between any two nodes of the network. Even for the Internet, it is a reasonable assumption, since roughly 80% of the routes seems to correspond to shortest paths.

The shortest path from a source to a destination is the path that minimizes the sum of the weights of its constituent links. For one-dimensional link weights, powerful shortest path algorithm like that of Dijkstra [30] exist. When 2-dimensional link weights are considered, we use SAMCRA [69], a Self-Adapting Multiple Constraints Routing Algorithm to find the shortest or optimal path that satisfies the constraint[3] $[L_1, L_2]$ such that $w_i(\mathcal{P}) = \sum_{(u \rightarrow v) \in \mathcal{P}} w_i(u \rightarrow v) \leq L_i$ and minimizes the path length function $l(\mathcal{P}) = \max_{1 \leq i \leq 2} \left[\frac{w_i(\mathcal{P})}{L_i} \right]$.

Both the topology and link weight structure seriously influence the path properties, and so the performance of the service. In Part III, we investigate the interplay between service (including link weight tuning) and topology in determining the performance of the shortest path routing.

---

[3]Actually, we choose loose constraints such that an optimal path can always be found via SAMCRA.

Chapter 3

Quantify Robustness by topological measure

The robustness of a network for a given service can usually be characterized by a topological measure. Without losing generality, any topological metric may possibly represent a certain network robustness for a service. Therefore, we will elaborate on the basic topological measures (graph metrics), that are frequently used throughout this thesis. Over the past several years, a variety of measures in both the structural domain and the spectral domain have been proposed to capture different features of a network topology as well as to classify graphs. We refer to [27] for a quite extensive survey of metrics. Actually, the 'network topology' is the same as 'network structure'. Here, structural measures refer to those measures such as degree and clustering coefficient, that represent topological properties more directly compared to spectral measures, which always involve in an eigenvalue computation. Finally, we exemplify what kind of network robustness can be represented by two metrics: the hopcount and the algebraic connectivity.

3.1   Structural measures

In general, topological measures are function of the topology or graph $G(N, L)$. The number of nodes N and the number of links L are mostly regarded as parameters of a graph, not metrics. Many measures are highly correlated with the size of the graph N and the number of links L.

The *degree* d_j of a node j in a graph is the number of links that incident with that node. The degree is an important characteristic of a node. For example, it may reflect the traffic capacity, the popularity of the node. The *degree distribution* of a graph, $\Pr[D = k]$ expresses the fraction of nodes in the graph with degree k. In other words, it is the probability that a randomly chosen node possesses degree k. The average degree

is purely a function the number of nodes N and the number of links L,

$$E[D] = \sum_{k=1}^{k=d_{\max}} k \cdot \Pr[D = k] = \frac{2L}{N}$$

where $d_{\max}$ is the maximum degree in a graph.

The *link density* p is equal to the number of links L in the graph divided by the maximal possible number of links $\binom{N}{2}$ that may exist in a graph with the same size

$$p = \frac{2L}{N(N-1)} = \frac{E[D]}{N-1}$$

The *clustering coefficient* of a node $c_G(v)$ characterizes the density of connections in the environment of a node v and is defined as the ratio of the number of links y connecting the $d_v > 1$ neighbors of v over the total possible $\frac{d_v(d_v-1)}{2}$, thus $c_G(v) = \frac{2y}{d_v(d_v-1)}$. The clustering coefficient $C(G)$ of a graph is the average clustering coefficient of nodes whose degree is larger than 1, given as

$$C(G) = \frac{1}{N - |\mathcal{N}^{(1)}|} \sum_{v \in \mathcal{N} - \mathcal{N}^{(1)}} c_G(v)$$

where $\mathcal{N}$ is the set of all nodes and $\mathcal{N}^{(1)}$ is the set of degree 1 nodes.

The *hopcount* H_N of a shortest path is the number of links contained in that path. The hopcount distribution $\Pr[H_N = k]$ is the histogram of the hopcount between all possible node pairs in the graph. The average $E[H_N]$ and the variance $\mathrm{Var}[H_N]$, sometimes, can be used to characterize the hopcount distribution. The largest hopcount $h_{\max}$ between any pair of nodes is also referred to as the *diameter* of a graph.

A good measure for "link/node importance" is the *betweenness* $B_l(B_n)$ of a link (node), which is defined as the number of shortest paths between all possible pairs of nodes in the network that traverse the link (node). The betweenness $B_l(B_n)$ which incorporates global information is a simplified quantity to assess the maximum possible traffic. Assuming that a unit packet is transmitted between each node pair, the betweenness B_l is the total amount of packets passing through a link.

The *node connectivity* $\kappa(G)$ and the *link connectivity* $\lambda(G)$ are the minimal number of nodes and links that have to be removed in order to disconnect a network. They seem natural quantifiers for robustness, but difficult to compute for large networks.

3.2   Spectral measures

Let G be a graph and let $\mathcal{N}$ denote the set of nodes and $\mathcal{L}$ the set of links, with $N = |\mathcal{N}|$ nodes and $L = |\mathcal{L}|$ links, respectively. The graph G can be represented by the adjacency matrix A, a $N \times N$ matrix, consisting of elements a_{ij} that are either

one or zero depending on whether there is a link between node i and j. The Laplacian matrix of G with N nodes is a $N \times N$ matrix $Q = \Delta - A$, where $\Delta = \mathrm{diag}(d_i)$ and d_i is the degree of node $i \in \mathcal{N}$. The set of N eigenvalues of the Laplacian matrix $\mu_N = 0 \leq \mu_{N-1} \leq ... \leq \mu_1$ is called the Laplacian spectrum of G. The second smallest eigenvalue μ_{N-1}, also called after Fiedler's seminal paper [42], the algebraic connectivity, can be denoted as $\mu_{N-1} = a(G)$ for simplicity. We denote the set of eigenvalues of the adjacency matrix, $\lambda_N \leq \lambda_{N-1} \leq \cdots \leq \lambda_1$, where λ_1 is called the spectral radius. The smaller the spectral radius is, the more robust a network is against the spread of viruses [71]. The eigenvalues of the adjacency matrix are real while the eigenvalues of the Laplacian matrix are real and non-negative [75]. The information of the adjacency matrix A can be equally represented by the set of (either Laplacian or adjacency) eigenvalues and the corresponding eigenvectors.

3.3 Example: the hopcount

The world is becoming more dependent on computer and communication networks, ranging from the traditional telephony network to the Internet. An issue of significant theoretical and practical interest for both consumers and service providers is the path stability (or route stability), where a path refers to the routing path determined by the specific routing protocol. The path stability can be quantitatively measured by the probability that a routing path is still available when the network is prone to link failures. In fact, the path stability quantifies the effect of link failures on the current routing path and characterizes the end-to-end performance of a communication network. For example, links break frequently in ad-hoc networks due to the mobility or the lack of energy of nodes. Once a routing path is unavailable because of the failures of one or more links on the path, the corresponding route rediscovery broadcasts would lead to significant delays and energy consumptions in a dynamic network with a large diameter. Here, we are going to illustrate how a robustness aspect of a network, the path stability, can be characterized by the topological metric hopcount.

We concentrate on the following problem: two arbitrary nodes are communicating via a given routing path in a network G of size N. At some time later, each link has an independent probability p_R to fail. Other link failure patterns can be envisaged, but most of them will complicate the analysis. Moreover, in practice, it is difficult to obtain accurate information of possible other correlated link failures. We define two status of a link: the working status (colored by blue) and the failed status (colored by red). The status of a link is independent of other links. A link can be either working or failed. The link capacity is assumed to be infinite. Hence, the path is stable if none of its links fail. What is then the path stability $\Pr[Y = 0]$, i.e. the probability that no red links (link failures) appear in the routing path, where Y is the number of red links (failed links) appearing in the path. More general, what is the probability distribution

$\Pr[Y = k]$ of the number of red links (failed links) encountered by a path.

We start with a general analysis with the only assumption of independent link failures. Applying the law of total probability, the probability that k red links appear in a path is

$$\Pr[Y = k] = \sum_{j=1}^{N-1} \Pr[Y = k | H_N = j] \Pr[H_N = j]$$

where H_N is the hopcount or number of links of the path. In case of a random distribution of red links and given the hopcount H_N of the shortest path, the probability distribution of the number of red/blue links that appear in the shortest path has the binomial distribution,

$$\Pr[Y = k | H_N = j] = \binom{j}{k} p_R^k (1 - p_R)^{j-k}$$

Hence,

$$\Pr[Y = k] = \sum_{j=1}^{N-1} \binom{j}{k} p_R^k (1 - p_R)^{j-k} \Pr[H_N = j]$$

$$= \frac{p_R^k}{(1 - p_R)^k} \sum_{j=1}^{N-1} \binom{j}{k} \Pr[H_N = j](1 - p_R)^j \qquad (3.1)$$

which depends purely on the hopcount distribution. Based on the definition of a probability generating function

$$\varphi_H(z) = \sum_{j=1}^{N-1} \Pr[H_N = j] \, z^j$$

we find that

$$\frac{d^k}{dz^k} \varphi_H(z) = \sum_{j=1}^{N-1} j(j-1) \cdots (j-k+1) \Pr[H_N = j] \, z^{j-k}$$

$$= \frac{k!}{z^k} \sum_{j=1}^{N-1} \binom{j}{k} \Pr[H_N = j] \, z^j$$

Using this expression in that of $\Pr[Y = k]$ yields

$$\Pr[Y = k] = \frac{p_R^k}{k!} \frac{d^k}{dz^k} \varphi_H(z) \Big|_{z=1-p_R} \qquad (3.2)$$

The path stability follows

$$\Pr\left[Y=0\right] = \varphi_H(1-p_R) \tag{3.3}$$

According to (3.1), (3.2) and (3.3), the path stability $\Pr\left[Y=0\right]$ as well as the distribution $\Pr[Y=k]$ of link failures in a path depend purely on the hopcount distribution $\Pr[H_N = j]$ or equivalently, on the probability generating function of the hopcount $\varphi_H(z)$ in the graph. This exemplifies the relation between a robustness aspect of a network, namely, the path stability, and the topological metric hopcount.

3.4 Example: the algebraic connectivity $a(G)$

The algebraic connectivity can characterize network robustness regarding to the following two dynamic processes: synchronization of dynamic processes at the nodes of a network and random walks on graphs. Random walks on graphs model, for example, the dispersion phenomena or exploring graph properties [31]. A network has a more robust synchronized state if the algebraic connectivity of the network is large [98][97]. Random walks move and disseminate efficiently in topologies with large algebraic connectivity.

Besides, the algebraic connectivity is also widely studied in various areas of mathematics [28], mainly discrete mathematics and combinatorial optimization, with interpretation in several physical and chemical problems. In this section, we present those mathematical results, which may reveal the topological implications of the algebraic connectivity or imply the robustness characterized by the algebraic connectivity.

3.4.1 Advantages

The algebraic connectivity characterizes the robustness with respect to the connectivity of a network, as follows from two properties:

- Fiedler [42] showed that the algebraic connectivity is only equal to zero if G is disconnected and that the multiplicity of zero as an eigenvalue of the Laplacian matrix Q is equal to the number of disconnected components of G.

- Over many years, network properties as connectivity and "the ease to tear a network apart" have been studied in graph theory. A commonly agreed metric to reflect these properties is the algebraic connectivity. The higher the algebraic connectivity is, the more difficult it is to break up the network in separate components.

Other metrics, like the minimum degree $d_{\min}(G)$, node $\kappa(G)$ and link $\lambda(G)$ connectivity, may characterize the network robustness against node or link removal, with respect to the most vulnerable part of a network. The algebraic connectivity is upper bounded by these metrics and is illustrated to be a better robustness measure, because

- Fielder [42] proved that in an incomplete graph, $a\left(G\right) \leq \kappa(G) \leq \lambda(G) \leq d_{\min}(G)$, where the node $\kappa(G)$ and the link $\lambda(G)$ connectivity are the minimal number of nodes and links that have to be removed in order to disconnect a network. The algebraic connectivity is a more useful robustness measure with respect to the connectivity than the node and link connectivity. For example, all trees have the minimum node and link connectivity 1, while the star has a larger algebraic connectivity than a path. A star is more robust than a path in the sense that only the removal of the highest degree node in a star will disconnect the network, while the removal of any node (except the two degree 1 node) always disconnects the path.

- The algebraic connectivity is non-decreasing by adding links. Let $G + e$ denotes a graph obtained from G by adding a link e between two nodes that are not connected. The following interlacing property is well known (see, for example [38])

Theorem 1 *Let G be a general graph of N nodes. Let $G + e$ be the graph obtained by adding the edge e in G. Then the eigenvalues of G interlace those of $G + e$, that is,*

$$\mu_N(G) \leq \mu_N(G + e) \leq \mu_{N-1}(G) \leq \mu_{N-1}(G + e)$$
$$\leq \mu_{N-2}(G)... \leq \mu_1 \leq \mu_1(G + e)$$

Since

$$\sum\nolimits_{i=1}^{N} \mu_i(G) = 2L$$
$$\sum\nolimits_{i=1}^{N} \left(\mu_i(G + e) - \mu_i(G)\right) = 2$$

at least one inequality in Theorem 1 must be strict and

$$0 \leq \mu_{N-1}(G + e) - \mu_{N-1}(G) \leq 2 \tag{3.4}$$

This is a fundamental feature expected for any good robustness metric of transportation networks, because the traffic capacity is generally non-decreasing after adding links.

- Finally, we demonstrate in Section 4.3 that compared to other Laplacian eigenvalues, the algebraic connectivity better quantifies the network robustness with respect to traffic engineering. A network with large algebraic connectivity tends to possess a dense core and sparse border structure. Hence, the traffic is more uniformly distributed, when traffic is injected between each node pair.

3.4.2 Disadvantages

According to Theorem 1, if the third smallest Laplacian eigenvalue is equal to the second smallest (the algebraic connectivity), $\mu_{N-2}(G) = \mu_{N-1}(G)$, then $\mu_{N-1}(G) = \mu_{N-1}(G+e)$ the algebraic connectivity remains the same wherever a link is added. This is observed in graphs such as the d-lattice, the ring, or the star. Hence, the algebraic connectivity is sometimes not sensitive to link addition. Furthermore, many real-world networks possess degree 1 nodes, such as stubs or end-users. For these networks $a(G) \leq 1$, and thus small, since $a(G) \leq d_{\min}(G)$. Hence, the algebraic connectivity seems a less suitable robustness quantifier. However, a properly designed robust network does not contain these obvious weaknesses. The extreme example is adding one link and one node to a complete graph (that is maximally "robust" in the sense that $a(G) \leq a(K_N) = N$). That resulting graph suddenly features a low algebraic connectivity. But, this is intuitively quite obvious: no network designer that aims to create a robust network would allow for these weak links, especially in the design of the core of a communications network.

In summary, the algebraic connectivity characterizes the network robustness regarding to the synchronization of dynamic processes at the nodes of a network and random walks on graphs. It quantifies the network robustness in terms of connectivity, the connection of the weakest part of a network and traffic engineering. Although the algebraic connectivity is not a sensitive measure to compare the robustness of current real-world networks, it is a good robustness metric to optimize, at least when designing the core of a network.

Significant progress has been made in understanding the quantification of network robustness, in terms of the relationship between the structural properties of networks and the performance of the services. The robustness quantification with respect to specific performance aspects of a service is also discussed in Part III of the thesis.

Part II

Robustness Optimization

Chapter 4

Graphs with given diameter maximizing $a(G)$

Dependent on the specific performance aspect that a service requires, a metric R can be selected to characterize the network robustness. The next questions are "what is the optimal network topology G that maximizes the robustness measure $R(G)$?" and "Given the network infrastructure, how can we improve its robustness via structure modification like adding links?" Such robustness optimization problems, that will be investigated in Chapter 4 and Chapter 5, are not only of great interest for practical network design and development, but also of theoretical importance in graph theory, discrete mathematics and combinatorial optimization. We confine to the spectral metric, the algebraic connectivity, in view of its wealthy robustness implications presented in Section 3.4.

The diameter D of a graph is the maximum distance in terms of the number of hops or links over all pairs of nodes in G. The diameter is one of the graph metrics that is not only of theoretical interest but that also has many practical applications. In communication networks, the diameter plays a key role in network design when the network performance, such as the delay or signal degradation, is proportional to the number of links that a packet traverses [21]. In order to guarantee certain quality of service, the diameter should be, somehow, limited. The complete graph or clique K_N has the maximal algebraic connectivity $a(K_N) = N$. However, real-world networks are always far sparser and their diameters are mostly larger. In order to construct a certain relative large diameter, links have to be removed, but this reduces the algebraic connectivity. It is essential to understand the relationship between the maximal algebraic connectivity $a_{\max}(N, D)$ and the diameter D, at constant N.

In this chapter, we propose a class of graphs $G_D^*(n_1, n_2, ..., n_D, n_{D+1})$, containing a chain of $D + 1$ cliques $K_{n_1}, K_{n_2}, ..., K_{n_{D+1}}$, where neighboring cliques are fully-interconnected. The class of graphs has diameter D and size $N = \sum_{1 \le i \le D+1} n_i$. This structure was employed by Van Dam [91] to determine the graphs with the maximal

spectral radius λ_1 among those on N nodes and diameter D. Here, we prove that this class of graphs can achieve the maximal number of links, the minimum average hopcount, and more interestingly, the maximal of any Laplacian eigenvalue among all graphs with N nodes and diameter D. We determine analytically the graph with the largest algebraic connectivity $a_{\max}(N, D)$ among graphs with N nodes and diameter $D < 4$. For other diameters, numerically searching for the maximum of any Laplacian eigenvalue is feasible, because (a) the searching within the class $G_D^*(n_1, n_2, ..., n_{D+1})$ is much smaller than within all graphs with N nodes and diameter D; (b) we reduce the calculation of the Laplacian spectrum from a $N \times N$ to a $(D+1) \times (D+1)$ Jacobian matrix. These maximal Laplacian eigenvalues as well as the corresponding optimal graphs, obtained either theoretically or by numerical searching, are applied to (1) investigate the topological features of graphs that maximize different Laplacian eigenvalues; (2) study the correlation between the maximum algebraic connectivity $a_{\max}(N, D)$ and N as well as D and (3) evaluate two upper bounds of the algebraic connectivity that are proposed in the literature.

4.1 The class of graphs G_D^*

4.1.1 Definition

Definition 2 *The class of graphs $G_D^*(n_1, n_2, ..., n_{D+1})$ is composed of $D + 1$ cliques $K_{n_1}, K_{n_2}, ..., K_{n_D}$ and $K_{n_{D+1}}$, where the variable $n_i \geq 1$ with $1 \leq i \leq D + 1$ is the size or the number of nodes in the $i - th$ clique. Each clique K_{n_i} is fully connected with its neighboring cliques $K_{n_{i-1}}$ and $K_{n_{i+1}}$ for $2 \leq i \leq D$. Two graphs G_1 and G_2 are fully connected if each node in G_1 is connected to all the nodes in G_2.*

Two examples, $G_{D=4}^*(n_1 = 3, n_2 = 1, n_3 = 2, n_4 = 1, n_5 = 2)$ and $G_{D=4}^*(n_1 = 1, n_2 = 3, n_3 = 2, n_4 = 2, n_5 = 1)$, are shown in Figure 4.1. Obviously, the class of graphs $G_D^*(n_1, n_2, ..., n_{D+1})$ has diameter D, which equals the distance between nodes in clique K_{n_1} and nodes in $K_{n_{D+1}}$. The size of each clique must be larger than or equal to one, i.e. $n_i \geq 1$ for $1 \leq i \leq D + 1$, else, the diameter of the graph is smaller than D. The degree of any node in K_{n_i} is $n_i - 1 + n_{i+1} + n_{i-1}$ for $2 \leq i \leq D$. The degree is $n_1 - 1 + n_2$ for any node in K_{n_1} and is $n_{D+1} - 1 + n_D$ for nodes in clique $K_{n_{D+1}}$.

4.1.2 Properties

Each node in the class of graphs $G_D^*(n_1, n_2, ..., n_{D+1})$ is fully connected within the clique and with neighboring cliques. We now define a node shifting action performed on a graph of the class $G_D^*(n_1, n_2, ..., n_{D+1})$. The resultant graph also belongs to this class and differs from the initial graph in that one node is shifted to a neighboring clique.
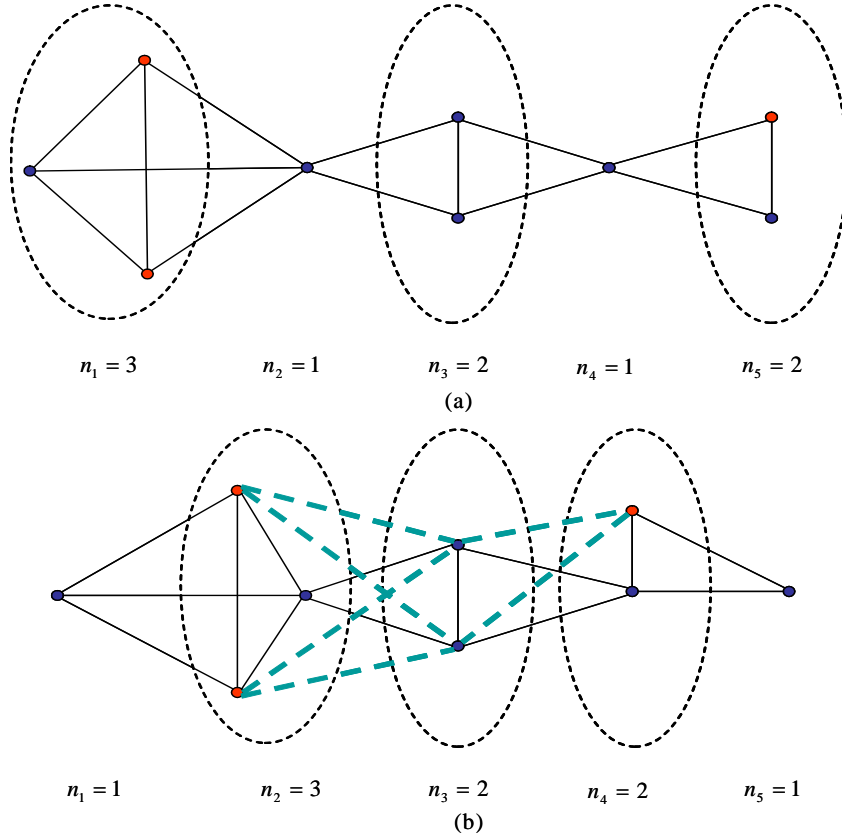
Figure 4.1: The graph (a) $G_{D=4}^*(n_1 = 3, n_2 = 1, n_3 = 2, n_4 = 1, n_5 = 2)$ and (b) $G_{D=4}^*(n_1 = 1, n_2 = 3, n_3 = 2, n_4 = 2, n_5 = 1)$.

Definition 3 *Node shifting within the class $G_D^*(n_1, ..., n_{i-1}, n_i, n_{i+1}, ..., n_{D+1})$: Any node in clique K_{n_i} for $2 \leq i \leq D$ can be shifted to its neighboring clique $K_{n_{i+1}}$ (or $K_{n_{i-1}}$) by removing links between this node and all the nodes in clique $K_{n_{i-1}}$ (or $K_{n_{i+1}}$) and by adding links between this node and all the nodes in clique $K_{n_{i+2}}$ ($K_{n_{i-2}}$), if $n_i > 1$. The resultant graph after one of such node shifting actions is $G_D^*(n_1, ..., n_{i-1}, n_i - 1, n_{i+1} + 1, ..., n_{D+1})$ or $G_D^*(n_1, n_2, ..., n_{i-1} + 1, n_i - 1, n_{i+1}, ..., n_{D+1})$. A node in clique K_{n_1} (or $K_{n_{D+1}}$) can only be shifted to clique K_{n_2} (or K_{n_D}) by adding links between it and all nodes in clique K_{n_3} (or $K_{n_{D-1}}$), which results in $G_D^*(n_1 - 1, n_2 + 1, ..., n_{D-1}, n_D, n_{D+1})$ (or $G_D^*(n_1, n_2, ..., n_{D-1}, n_D + 1, n_{D+1} - 1)$).*

Figure 4.2 illustrates an example of node shifting. From (a) $G_{D=4}^*(n_1 = 3, n_2 = 1, n_3 = 2, n_4 = 1, n_5 = 2)$ to (b) $G_{D=4}^*(n_1 = 3, n_2 = 1, n_3 = 1, n_4 = 2, n_5 = 2)$, a node (red) in K_{n_3} is shifted to K_{n_4} by removing the link (marked with cross) between that node and nodes in clique K_{n_2} and by adding links (the blue dotted line) between the
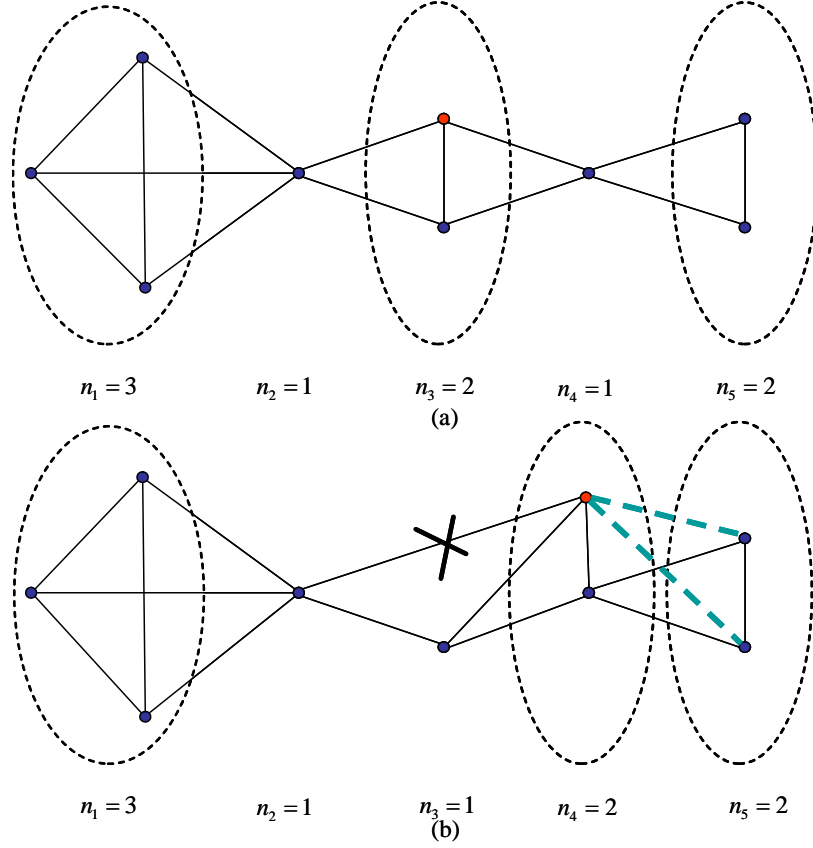
Figure 4.2: The graph (a) $G^*_{D=4}(n_1 = 3, n_2 = 1, n_3 = 2, n_4 = 1, n_5 = 2)$ and (b) $G^*_{D=4}(n_1 = 3, n_2 = 1, n_3 = 1, n_4 = 2, n_5 = 2)$. The line with cross mark is to be removed and the dotted blue links are added.

node and all nodes in K_{n_5}. In fact, any two graphs in the class $G^*_D(n_1, n_2, ..., n_D, n_{D+1})$ with the same number N of nodes can be transformed from one to the other by a set of node shifting actions. For example, Figure 4.1(b) can be obtained from Figure 4.1(a) by shifting two nodes from K_{n_1} to K_{n_2} and one node from K_{n_5} to K_{n_4}. When a node in clique K_{n_i}, where $2 \leq i \leq D$ and $n_i > 1$, is shifted to clique $K_{n_{i+1}}$, n_{i-1} links are removed and n_{i+2} links are added. Hence, if we shift $m < n_i$ nodes from clique K_{n_i} to clique $K_{n_{i+1}}$, $n_{i-1} \cdot m$ links are removed and $n_{i+2} \cdot m$ links are added. This node shifting operation will be frequently used to prove several interesting properties of the class $G^*_D(n_1, n_2, ..., n_D, n_{D+1})$.

Based on the sizes of the first and last clique, the class of graphs $G^*_D(n_1, n_2, ..., n_D, n_{D+1})$ can be divided into two sets: 1) $n_1 = n_{D+1} = 1$, e.g. Figure 4.1(b), and 2) at least one of n_1, n_{D+1} is larger than 1, e.g. Figure 4.1(a). The set 1 is generally denser than the

set 2, in the sense that

Lemma 4 *A graph $G_D^*(n_1, n_2, ..., n_D, n_{D+1})$, where at least one of n_1 and n_{D+1} is larger than one, is a subgraph of $G_D^*(1, n_1 - 1 + n_2, ..., n_{D-1}, n_D + n_{D+1} - 1, 1)$.*

Proof. According to the definition of node shifting, links are only added and not removed, when a node is shifted from K_{n_1} to K_{n_2} or from $K_{n_{D+1}}$ to K_{n_D}. $G_D^*(1, n_1 - 1 + n_2, ..., n_{D-1}, n_D + n_{D+1} - 1, 1)$ can be obtained from $G_D^*(n_1, n_2, ..., n_{D-1}, n_D, n_{D+1})$ by shifting $n_1 - 1$ nodes from K_{n_1} to K_{n_2} and by shifting $n_{D+1} - 1$ nodes from $K_{n_{D+1}}$ to K_{n_D} by purely adding links. Hence, $G_D^*(n_1, n_2, ..., n_{D-1}, n_D, n_{D+1})$ is a subgraph of $G_D^*(1, n_1 - 1 + n_2, ..., n_{D-1}, n_D + n_{D+1} - 1, 1)$, when either n_1 or n_{D+1} is larger than one. ∎

Figure 4.1 gives an example of Lemma 4, i.e. $G_{D=4}^*(n_1 = 3, n_2 = 1, n_3 = 2, n_4 = 1, n_5 = 2)$ is a subgraph of $G_{D=4}^*(n_1 = 1, n_2 = 3, n_3 = 2, n_4 = 2, n_5 = 1)$. Both graphs contain the same set of nodes, while the latter consists of more links, the blue dotted ones.

The motivation to study the set of graphs $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$ lies in the following properties.

Theorem 5 *Any graph $G(N, D)$ with N nodes and diameter D is a subgraph of at least one graph in the class $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$.*

Proof. There is at least one node pair in $G(N, D)$ that is D hops away from each other, because the diameter of $G(N, D)$ is D. We select a node s from one such node pair and denote it as cluster $C_1 = s$. We define the set of clusters C_i $(2 \le i \le D + 1)$ as the set of $|C_i|$ nodes that is i hops away from s or cluster C_1. There can be more than one node that is D hops away from s, when $|C_{D+1}| \ge 1$. First, $G(N, D)$ is a subgraph of the graph $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1})$ when $n_i = |C_i|$ for $1 \le i \le D + 1$, because of two reasons: (a) Within each cluster C_i of $G(N, D)$, for $1 \le i \le D + 1$, these $|C_i|$ nodes are at best fully connected as in the corresponding clique K_{n_i} with size $n_i = |C_i|$ in $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1})$. (b) in $G(N, D)$, nodes in cluster C_i $(2 \le i \le D)$ can not be connected to nodes in other clusters except for C_{i-1} and C_{i+1}, or else, the distance between $C_1 = s$ and nodes in C_{D+1} is smaller than D. Similarly, each clique K_{n_i} is only but fully connected to its neighboring cliques $K_{n_{i-1}}$ and $K_{n_{i+1}}$ in $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1})$.

Based on Lemma 4, $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1})$ is a subgraph of $G_D^*(1, n_2, ..., n_{D-1}, n_D + n_{D+1} - 1, 1)$. Hence, any graph $G(N, D)$ with N nodes and diameter D is a subgraph of at least one graph in the class $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$. ∎

Since $\sum_{i=1}^{D+1} n_i = N$ and $n_1 = n_{D+1} = 1$ always hold, the graph $G_D^*(1, n_2, ..., n_D, 1)$ contains $D - 2$ variables: the sizes of the cliques and $n_i > 0$ for $1 \le i \le D + 1$.

Fiedler [42] showed that, if G_1 is a subgraph of G with the same size, then $a(G_1) \le a(G)$. Hence, by virtue of Theorem 5, we have:

Corollary 6 *The maximum algebraic connectivity of the graphs in the class $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$ is also the maximum among all the graphs with the same size N and diameter D, i.e. $a_{\max}(G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)) = a_{\max}(N, D)$.*

However, given size N and diameter D, the graph that has the maximum algebraic connectivity $a_{\max}(N, D)$ may not be unique. For example, the graph in $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$ maximizing the algebraic connectivity $a_{\max}(N, D)$ may possess the same algebraic connectivity after a set of links is deleted. In other words, different graphs may have the same algebraic connectivity $a_{\max}(N, D)$.

Theorem 7 *The maximum of any eigenvalue $\mu_i(G_1)$, $i \in [1, N]$ achieved in the class $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$ is also the maximum among all the graphs with N nodes and diameter D.*

Proof. Based on the well-known interlacing property introduced in Theorem 1, if G_1 is a subgraph of G with the same size N, $\mu_i(G_1) \leq \mu_i(G)$, for $i \in [1, N]$. Together with Theorem 5, the proof can be completed. ∎

Theorem 8 *The maximum number of links in a graph with given size N and diameter D is $L_{\max}(N, D) = \binom{N-D+2}{2} + D - 3$, which can only be obtained by either $G_D^*(1, ..., 1, n_j = N - D, 1, ..., 1)$ with $j \in [2, D]$, where only one clique has size larger than one, or by $G_D^*(1, ..., 1, n_j > 1, n_{j+1} > 1, 1, ..., 1)$ with $j \in [2, D-1]$ where only two cliques have size larger than one and they are next to each other.*

Proof. First, according to Theorem 5, the maximum number of links $L_{\max}(N, D)$ can only be achieved within the class $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$. Second, any other graph $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$, where more than one clique has size larger than one, can be transformed into $G_D^*(1, ..., 1, n_j = N - D, 1, ..., 1)$ by a set of node shifting operations. (a) When progressing from clique K_{n_2} to clique $K_{n_{D-1}}$, we label the first encountered clique that has size larger than one as K_{n_r} such that $n_i = 1$ for $i < r$. (b) We shift all but one (i.e. $n_r - 1$) nodes in clique K_{n_r} to clique $K_{n_{r+1}}$ by deleting $(n_r - 1) \cdot n_{r-1} = (n_r - 1)$ links and by adding $(n_r - 1) \cdot n_{r+2}$ links. The process (a and b) is repeated until there is only one clique having size larger than 1. Since $n_i \geq 1$ for $i \in [1, D+1]$ according to the definition of the class $G_D^*(n_1, n_2, ..., n_D, n_{D+1})$, $(n_r - 1) \cdot n_{r+2} \geq n_r - 1$. The inequality holds when $n_{r+2} > 1$, which happens at least one time during the recursive node shifting except for $G_D^*(1, ..., 1, n_j > 1, n_{j+1} > 1, 1, ..., 1), j \in [1, D-2]$ where only two cliques have size larger than one and they are next to each other. Hence, $G_D^*(1, ..., 1, n_j = N - D, 1, ..., 1), j \in [2, D]$ and $G_D^*(1, ..., 1, n_j > 1, n_{j+1} > 1, 1, ..., 1), j \in [2, D-1]$ possess the maximum number of links among graphs of size N and diameter D. The maximum number of links is $L_{\max}(N, D) = \binom{N-D}{2} + 2(N - D) + D - 2 = \binom{N-D+2}{2} + D - 3$. ∎

Theorem 9 *The minimum average hopcount in graphs with size N and diameter D can be only obtained by $G_D^*(1, ..., 1, n_{\frac{D}{2}+1} = N - D, 1, ..., 1)$ when D is even, or when D is odd, by $G_D^*(1, ..., 1, n_{\lfloor \frac{D}{2} \rfloor+1} \geq 1, n_{\lceil \frac{D}{2} \rceil+1} \geq 1, 1, ..., 1)$, where only the two cliques in the middle have size larger than one. The minimum average hopcount is*

$$
\min_{\mathcal{G} \in G(N,D)} E[H(\mathcal{G})] =
\begin{cases}
\frac{N-D-1}{N(N-1)} \left(\frac{D^2}{2} + N \right) + \frac{\sum_{i=1}^{D} i(D-i+1)}{\binom{N}{2}}, & \text{when } D \text{ is even} \\[2ex]
\frac{N-D-1}{N(N-1)} \left(2 \left\lfloor \frac{D}{2} \right\rfloor^2 + N + D \right) + \frac{\sum_{i=1}^{D} i(D-i+1)}{\binom{N}{2}}, & \text{when } D \text{ is odd}
\end{cases}
$$

Proof. First, according to Theorem 5 and the fact that adding links can always reduce the average hopcount, the minimum average hopcount in graphs with given size N and diameter D can be only be achieved within the class $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$. Second, within the set $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$, any graph can be transformed into $G_D^*(1, ..., 1, n_{\frac{D}{2}+1} = N - D, 1, ..., 1)$ for even D, or into $G_D^*(1, ..., 1, n_{\lfloor \frac{D}{2} \rfloor+1} \geq 1, n_{\lceil \frac{D}{2} \rceil+1} \geq 1, 1, ..., 1)$ for odd D via the following node shifting, where the average hopcount can always be reduced. We consider first the case that D is odd. We repeat the node shifting process (a) and (b) in the proof of Theorem 8 for $r \leq \lfloor \frac{D}{2} \rfloor$, until $n_i = 1$ for $i < \lfloor \frac{D}{2} \rfloor + 1$ and all the remaining nodes are shifted into clique $\lfloor \frac{D}{2} \rfloor + 1$. When a node is shifted from K_{n_r} to $K_{n_{r+1}}$, its distance to any node in clique $i < r$ is increased by one, while its distance to any node in clique $i > r + 1$ is reduced by one. Hence, via such a node shifting operation, the sum of the hopcounts between all nodes pairs is reduced by $\sum_{j=r+2}^{D+1} n_j - \sum_{j=1}^{r-1} n_j \geq \sum_{j=\lceil \frac{D}{2} \rceil+1}^{D+1} n_j - \sum_{j=1}^{\lfloor \frac{D}{2} \rfloor-1} 1 > 0$, because $r \leq \lfloor \frac{D}{2} \rfloor$ and $n_j \geq 1$ for $j \in [1, D+1]$. Similarly, from clique K_{n_D} to clique $K_{\lceil \frac{D}{2} \rceil+2}$, we denote the first encountered clique that has size larger than one as K_{n_r}. The $n_r - 1$ nodes in clique K_{n_r} are shifted to clique $K_{n_{r-1}}$. This shifting process is recursively carried out until $n_i = 1$ for $i > \lceil \frac{D}{2} \rceil + 1$ and all other nodes are shifted to the clique $K_{\lceil \frac{D}{2} \rceil+1}$. Shifting one node from clique K_{n_r} to clique $K_{n_{r-1}}$, where $\lceil \frac{D}{2} \rceil + 1 < r \leq D$, reduces the sum of the hopcounts between all nodes pairs by $\sum_{j=1}^{r-2} n_j - \sum_{j=r+1}^{D+1} n_j \geq \sum_{j=1}^{\lceil \frac{D}{2} \rceil} n_j - \sum_{j=\lceil \frac{D}{2} \rceil+2}^{D+1} 1 > 0$. The average hopcount can always be reduced as long as a node is shifted. Therefore, $G_D^*(1, ..., 1, n_{\lfloor \frac{D}{2} \rfloor+1} \geq 1, n_{\lceil \frac{D}{2} \rceil+1} \geq 1, 1, ..., 1)$ has the minimum average hopcount. The size of clique $\lfloor \frac{D}{2} \rfloor + 1$ and clique $\lceil \frac{D}{2} \rceil + 1$ have no effect on the average hopcount due

to the symmetry of G_D^*. Taking $n_{\lfloor \frac{D}{2} \rfloor + 1} = N - D$ and $n_{\lceil \frac{D}{2} \rceil + 1} = 1$, we have

$$
\min_{\mathcal{G} \in G(N,D)} E[H(\mathcal{G})] = \frac{(N - D - 1)\left(\sum_{i=1}^{\lfloor \frac{D}{2} \rfloor} i + \sum_{i=1}^{\lceil \frac{D}{2} \rceil} i\right) + \binom{N-D}{2} + \sum_{i=1}^{D} i(D - i + 1)}{\binom{N}{2}}
$$

$$
= \frac{N - D - 1}{N(N-1)}\left(2\left\lfloor \frac{D}{2} \right\rfloor^2 + N + D\right) + \frac{\sum_{i=1}^{D} i(D - i + 1)}{\binom{N}{2}}
$$

When D is even, the clique $K_{\lfloor \frac{D}{2} \rfloor + 1} = K_{\lceil \frac{D}{2} \rceil + 1}$ are the same. Similarly, any other graph $G_D^*(1, n_2, ..., n_D, 1)$ can be transformed into $G_D^*(1, ..., 1, n_{\frac{D}{2}+1} = N - D, 1, ..., 1)$ by nodes shifting, which can only decrease the average hopcount. When D is even, we have

$$
\min_{\mathcal{G} \in G(N,D)} E[H(\mathcal{G})] = \frac{2(N - D - 1)\sum_{i=1}^{\frac{D}{2}} i + \binom{N-D}{2} + \sum_{i=1}^{D} i(D - i + 1)}{\binom{N}{2}}
$$

$$
= \frac{N - D - 1}{N(N-1)}\left(\frac{D^2}{2} + N\right) + \frac{\sum_{i=1}^{D} i(D - i + 1)}{\binom{N}{2}}
$$

∎

In summary, the class $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$ can achieve the maximum of any Laplacian eigenvalue μ_i, $1 \le i \le N - 1$, the maximum link density, the minimum average hopcount among all graphs with given size N and diameter D. The graphs that possess the maximum link density and the minimum average hopcount are rigorously determined in Theorem 8 and 9. In the sequel, we focus on the Laplacian spectrum of the class $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$.

4.2   Laplacian Eigenvalues of G_D^*

Theorem 10 *The characteristic polynomial of the Laplacian $Q_{G_D^*} = \Delta_{G_D^*} - A_{G_D^*}$ of $G_D^*(n_1, n_2, ..., n_{D-1}, n_D, n_{D+1})$ equals*

$$
\det\left(Q_{G_D^*} - \mu I\right) = p_D(\mu) \prod_{j=1}^{D+1} (d_j + 1 - \mu)^{n_j - 1} \tag{4.1}
$$

where d_j denotes the degree of a node in clique j. The polynomial $p_D(\mu) = \prod_{j=1}^{D+1} \theta_j$ is of degree $D + 1$ in μ and the function $\theta_j = \theta_j(D; \mu)$ obeys the recursion

$$
\theta_j = (d_j + 1 - \mu) - \left(\frac{n_{j-1}}{\theta_{j-1}} + 1\right)n_j \tag{4.2}
$$

with initial condition $\theta_0 = 1$ and with the convention that $n_0 = n_{D+2} = 0$.

Proof. See Appendix A.1. ∎

Theorem 10 shows that the Laplacian $Q_{G_D^*}$ has eigenvalues at $n_{j-1}+n_j+n_{j+1} = d_j+1$ with multiplicity $n_j - 1$ for $1 \le j \le D + 1$, with the convention that $n_0 = n_{D+2} = 0$. The less trivial zeros are solutions of the polynomial $p_D(\mu) = \prod_{j=1}^{D+1}\theta_j$, where θ_j is recursively defined via (4.2). Since all the explicit eigenvalues $\mu_j = d_j + 1$ are larger than zero and since $\mu = 0$ is an eigenvalue of any Laplacian, the polynomial $\prod_{j=1}^{D+1}\theta_j$ must have a zero at $\mu = 0$. Thus, the polynomial of interest is

$$p_D(\mu) = \prod_{j=1}^{D+1}\theta_j(D;\mu) = \sum_{k=0}^{D+1} c_k(D)\mu^k = \prod_{k=1}^{D+1}(z_k - \mu) \tag{4.3}$$

where the dependence of θ_j on the diameter D and on μ is explicitly written and where the product with the zeros $z_{D+1} \le z_D \le \cdots \le z_1$ follows from the definition of the eigenvalue equation (see [67, p. 435-436]). Moreover, each $z_j \in [0, N]$ because each Laplacian eigenvalue of any graph is contained in the interval $[0, N]$.

Corollary 11 *The three Laplacian eigenvalues of G_D^*, the two smallest Laplacian eigenvalues $\mu_N = 0$ and the algebraic connectivity $a = \mu_{N-1}$ and the largest one μ_1, are equal to the zero $z_{D+1} = 0$, z_D and z_1 of the polynomial $p_D(\mu)$, respectively.*

Proof. Since all the explicit Laplacian eigenvalues $d_j + 1$ of G_D^* in (4.1) are larger than zero and since $\mu = 0$ is an eigenvalue of any Laplacian, the polynomial $p_D(\mu)$ must have a zero at $\mu = 0$. Grone and Merris [48] succeeded to improve Fiedler's lower bound and proved that, for any graph, $\mu_{N-1} \le d_{\min}$, where $d_{\min}$ is the minimum degree in the graph. All trivial eigenvalues are larger than the minimum degree since $d_j + 1 > d_j \ge d_{\min}$, which implies that the algebraic connectivity is $a = \mu_{N-1} = z_D$, the smallest positive zero of $p_D(\mu)$. The largest Laplacian eigenvalue obeys $\mu_1 \ge d_{\max} + 1$. Brouwer and Haemers [15] further show that the equality holds if and only if there is a node connecting to all the other nodes in the graph, i.e. $D = 2$. Hence, when the diameter $D > 2$, the largest eigenvalue is always a nontrivial eigenvalue, i.e. $\mu_1 = z_1$. When $D = 2$, the zeros of

$$p_D(\mu) = \mu\left(\mu^2 - (N + n_2)\mu + Nn_2\right) = \mu(\mu - N)(\mu - n_2)$$

are $z_3 = 0, z_2 = n_2$ and $z_1 = N$. Since the largest eigenvalue $\mu_1 \in [0, N]$, $\mu_1 = z_1$. ∎

Furthermore, $p_D(\mu)$ is shown in Appendix B to belong to a set of orthogonal polynomials. All the non-trivial eigenvalues of $Q_{G_D^*}$ are also eigenvalues of the (much simpler and smaller) Jacobian matrix $-\widetilde{M}$, where

$$\widetilde{M} = \begin{bmatrix} -n_2 & \sqrt{n_1 n_2} & & & \\ \sqrt{n_1 n_2} & -(n_1 + n_3) & \sqrt{n_2 n_3} & & \\ & \ddots & \ddots & \ddots & \\ & & \sqrt{n_{D-1}n_D} & -(n_{D-1} + n_{D+1}) & \sqrt{n_D n_{D+1}} \\ & & & \sqrt{n_D n_{D+1}} & -n_D \end{bmatrix}$$

Therefore, exhaustively numerical searching for the maximum of any Laplacian eigenvalue is feasible because of two reasons: (a) the searching space within $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$ is much smaller than the searching within all graphs with N nodes and diameter D. (b) the calculation of the Laplacian spectrum is reduced from a $N \times N$ matrix to a $(D+1) \times (D+1)$ tri-diagonal matrix.

4.3 The maximum of any Laplacian eigenvalue

Theorem 7 shows that the maximum of any eigenvalue among all graphs with size N and diameter D can be achieved within the class $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$. What is the topological implication when different eigenvalues are optimized? Table 4.1 presents the different topologies with $D = 6$ that optimize the i-th largest Laplacian eigenvalue μ_i and the spacing $\mu_i - \mu_{i+1}$, for $i \in [N-1, N-7]$.

Table 4.1: Graphs with $D = 6$ that optimize the i-th largest Laplacian eigenvalue μ_i or the spacing $\mu_i - \mu_{i+1}$.

N = 50						N = 100					
value to optimize	n_2	n_3	n_4	n_5	n_6	value to optimize	n_2	n_3	n_4	n_5	n_6
$\mu_{N-1} = \mu_{N-1} - \mu_N$	6	11	14	11	6	$\mu_{N-1} = \mu_{N-1} - \mu_N$	13	22	28	22	13
μ_{N-2}	16	15	1	1	15	μ_{N-2}	32	32	1	1	32
$\mu_{N-2} - \mu_{N-1}$	16	15	1	1	15	$\mu_{N-2} - \mu_{N-1}$	32	32	1	1	32
μ_{N-3}	1	22	1	1	23	μ_{N-3}	1	47	1	1	48
$\mu_{N-3} - \mu_{N-2}$	1	22	1	1	23	$\mu_{N-3} - \mu_{N-2}$	1	47	1	1	48
μ_{N-4}	1	22	1	23	1	μ_{N-4}	1	48	1	47	1
$\mu_{N-4} - \mu_{N-3}$	1	22	1	23	1	$\mu_{N-4} - \mu_{N-3}$	1	48	1	47	1
μ_{N-5}	1	1	44	1	1	μ_{N-5}	1	1	94	1	1
$\mu_{N-5} - \mu_{N-4}$	1	1	44	1	1	$\mu_{N-5} - \mu_{N-4}$	1	1	94	1	1
μ_{N-6}	1	1	1	30	15	μ_{N-6}	1	1	34	61	1
$\mu_{N-6} - \mu_{N-5}$	1	1	43	1	2	$\mu_{N-6} - \mu_{N-5}$	2	1	93	1	1
μ_{N-7}	1	1	1	35	10	μ_{N-7}	1	1	1	38	57
$\mu_{N-7} - \mu_{N-6}$	2	1	42	1	2	$\mu_{N-7} - \mu_{N-6}$	2	1	92	1	2

The graph in the class $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$ that optimizes μ_i for $i \leq N - 5$ possesses the maximal number of links, i.e. only one or two adjacent cliques have size larger than one, according to Theorem 8. In fact, μ_i for $i = N - 6, N - 7$ can be optimized by any graph in the class $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$ that has the maximal number of links, not only the graph listed in the table. Theorem 10 shows

that the Laplacian $Q_{G_D^*}$ has eigenvalues at $n_{j-1} + n_j + n_{j+1} = d_j + 1$ with multiplicity $n_j - 1$ for $1 \leq j \leq D + 1$. Graphs that maximize the number of links have the maximal trivial eigenvalue $N - D + 2$ with the maximal multiplicity $N - D + 1$. Hence, a large set of eigenvalues, but not the largest one[1] μ_1, can be optimized by graphs possessing the maximal number of links. Graph that optimizes the eigenvalue μ_i, at the same time, maximizes the corresponding spacing $\mu_i - \mu_{i+1}$, for $i \geq N - 5$. However, when $i < N - 5$, the graph that optimizes the eigenvalue μ_i, has spacing $\mu_i - \mu_{i+1} = 0$, which is far from the maximal spacing.

The graph that maximizes the algebraic connectivity μ_{N-1}, has larger sizes for cliques in the middle. It is dense in the core and sparse at borders. Such structure is robust for information transportation in the sense that traffic is more uniformly distributed, when traffic is injected between each node pair. Contrary, graphs that maximize other eigenvalues or spacing, have cliques with small size ($n_j = 1$) around the middle, which have to carry much more traffic and become the bottleneck for transportation. Graphs with many cliques of size one are vulnerable, because removal of such clique - which is in fact a node - disconnects the rest of the graph. Hence, the comparison of topologies in table 4.1 provides us with an extra motivation to quantify network robustness by the algebraic connectivity. Since $G_D^*(n_1, n_2, ..., n_D, n_{D+1})$ has

$$L = \sum_{i=2}^{D} \binom{n_i}{2} + \sum_{i=1}^{D} n_i n_{i+1}$$

links, the number of links in the graph that maximizes the algebraic connectivity is far smaller than the maximum for $D > 3$ according to Theorem 8. Therefore, graphs that maximize the algebraic connectivity are robust for transportation, while, at the same time, efficient in the number of links.

4.4 The maximum algebraic connectivity $a_{\mathrm{max}}(N, D)$

4.4.1 Exact computation of $a_{\mathrm{max}}(N, D)$ for diameter $D = 2, 3$

Before we start the $D = 2, 3$ cases, it should be mentioned that the graph $G(N, D = N - 1)$ with N nodes and diameter $N - 1$ is unique: a line graph. The algebraic connectivity of a line graph [28] is well-known: $a_{\mathrm{max}}(N, D = N - 1) = 2\left(1 - \cos\frac{\pi}{N}\right)$.

The complete Laplacian spectrum of $G_{D=2}^*(n_1, n_2, n_3)$ follows from Theorem 10 and the polynomial $p_D(\mu)$, as the zeros at $\mu_1 = n_1 + n_2$ with multiplicity $n_1 - 1$, $\mu_2 = n_1 + n_2 + n_3 = N$ with multiplicity $n_2 - 1$, $\mu_3 = n_2 + n_3$ with multiplicity $n_3 - 1$ and the simple zeros of

$$p_D(\mu) = \mu\left(\mu^2 - (N + n_2)\mu + Nn_2\right) = \mu(\mu - N)(\mu - n_2)$$

---

[1]The largest eigenvalue μ_1 is always a nontrivial one according to Theorem 10. Hence, a lower bound for the maximal possible μ_1 follows $\mu_{1\,\mathrm{max}} \geq N - D + 2$.

which are $z_3 = 0, z_2 = n_2$ and $z_1 = N$. Clearly, since $n_1 + n_2 + n_3 = N$, the largest possible algebraic connectivity $a_{\max}(N, D = 2) = n_2$ is $N - 2$.

This result is more directly found from Corollary 6. The maximum algebraic connectivity in graphs with N nodes and diameter D can be achieved in the class $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$, which is unique, i.e. $G_{D=2}^*(n_1 = 1, n_2 = N-2, n_3 = 1)$ for $D = 2$. The graph $G_{D=2}^*(n_1 = 1, n_2 = N - 2, n_3 = 1)$ is a clique of size N without one link $K_N - \{(i, j)\}$, which is the complementary graph of a path (or a clique) $\{(i, j)\}$ of two nodes. The sum of the algebraic connectivity of two complementary graphs is N [28]. Hence, $G_{D=2}^*(1, N - 2, 1) = K_N - \{(i, j)\}$ has the maximum algebraic connectivity $N - 2$ among graphs with N nodes and diameter $D = 2$, i.e. $a_{\max}(N, D = 2) = N - 2$.

Theorem 12 *For graphs with N nodes and diameter $D = 3$, the graph $G_{D=3}^* \left(1, \left\lfloor \frac{N-2}{2} \right\rfloor, N - 2 - \left\lfloor \frac{N-2}{2} \right\rfloor, 1\right)$ has the maximum algebraic connectivity with $\left\lfloor \frac{N-2}{2} \right\rfloor - 1 \leq a_{\max}(N, D = 3) \leq \left\lfloor \frac{N-2}{2} \right\rfloor$.*

Proof. Theorem 10 shows that the characteristic polynomial of the corresponding Laplacian matrix of $G_{D=3}^*(n_1 = 1, n_2, n_3, n_4 = 1)$ satisfies (4.1). The algebraic connectivity of $G_{D=3}^*(n_1 = 1, n_2, n_3, n_4 = 1)$ is the smallest zero z_3 of the polynomial,

$$q_3 \left(\mu\right) = p_3 \left(\mu\right)/\mu = \mu^3 - (2N - n_1 - n_4)\mu^2 + (n_2^2 + n_3^2 + n_1 n_2 + n_1 n_3 + n_1 n_4 + 3n_2 n_3 + n_2 n_4 + n_3 n_4)\mu - N n_2 n_3$$

that, here with $n_1 = n_4 = 1$, $n_2 = m$ and $n_3 = N - 2 - m$ reduces to

$$q_3 \left(\mu\right) = \mu^3 - 2(N-1)\mu^2 + ((N-2)(N+1) + (m-1)(N-m-3))\mu - Nm(N-m-2) \quad (4.4)$$

Second, we only need to consider the case $m \leq \left\lfloor \frac{N-2}{2} \right\rfloor$ because the $m > \left\lfloor \frac{N-2}{2} \right\rfloor$ can be reduced to the case $m \leq \left\lfloor \frac{N-2}{2} \right\rfloor$ by swapping the clique K_{n_1} and K_{n_4}. We will now show that, for $m \leq \left\lfloor \frac{N-2}{2} \right\rfloor$, the smallest zero z_3 of (4.4) satisfies $m - 1 < z_3 < m$.

All zeros of the orthogonal polynomial $p_D \left(\mu\right)$ are simple and non-negative. The sign of $q_3 \left(\mu\right)$ for $\mu = m$, $\mu = N - 1$ and for $\mu = N$ follows from

$$q_3(m) = -m < 0 \tag{4.5}$$

$$q_3(N - 1) = m(N - 2 - m) > 0 \tag{4.6}$$

$$q_3(N) = -N < 0 \tag{4.7}$$

Likewise, we find that

$$q_3(m - 1) = -2m + 2m^2 - 3Nm + N^2 \tag{4.8}$$

which obtains, as a function of m, a minimum at $m_0 = \frac{2+3N}{4}$. Thus, for $0 \leq m \leq m_0$, the function $q_3(m-1)$ is decreasing in m. Because $m \leq \lfloor \frac{N-2}{2} \rfloor$, it follows $m \leq \frac{N-2}{2} = m^*$ and that $q_3(m^* - 1) = 4$. Finally, because $m \leq m^* < m_0$ it follows that

$$q_3(m-1) > 0 \tag{4.9}$$

From (4.5), (4.9), (4.6) and (4.7), it follows that $q_3(\mu)$ has simple zeros $z_3 < z_2 < z_1$ satisfying $m - 1 < z_3 < m$, $m < z_2 < N - 1$ and $N - 1 < z_1 < N$. Hence, among the class $G_D^*(1, n_2, n_3, 1)$, the largest algebraic connectivity can be obtained by $G_{D=3}^* \left(1, \lfloor \frac{N-2}{2} \rfloor, N - 2 - \lfloor \frac{N-2}{2} \rfloor, 1\right)$, where m is maximized.

Finally, according to Corollary 6, the algebraic connectivity of $G_{D=3}^* \left(1, \lfloor \frac{N-2}{2} \rfloor, N - 2 - \lfloor \frac{N-2}{2} \rfloor, 1\right)$ is also the maximum $a_{\max}(N, D = 3)$ of all the graphs with N nodes and diameter $D = 3$, and $\lfloor \frac{N-2}{2} \rfloor - 1 \leq a_{\max}(N, D = 3) \leq \lfloor \frac{N-2}{2} \rfloor$. ∎

4.4.2 $a_{\max}(N, D)$ in relation to N and D

The maximum algebraic connectivity $a_{\max}(N, D)$ used in this section is obtained via exhaustive searching in $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$ for $D \geq 4$. We study first the $a_{\max}(N, D)$ in relation to N. As shown in Figure 4.3, the maximal algebraic connectivity
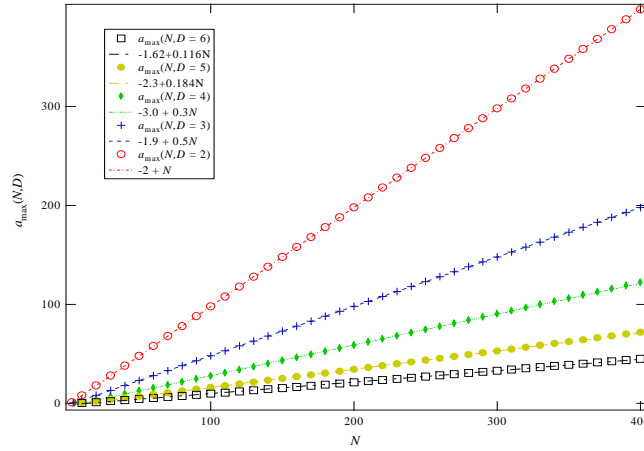
Figure 4.3: The $a_{\max}(N, D)$ (marker) for $2 \leq D \leq 6$ and the corresponding linear fitting (dotted line).

seems linear in N for constant D, i.e. $a_{\max}(N, D) \cong b + \epsilon \cdot N$. The slope ϵ decreases fast from $\epsilon = 1$ to $\epsilon = 0.12$ when the diameter increases from $D = 2$ to $D = 6$. When $D = 2$, $a_{\max}(N, D = 2) = N - 2$, which is determined in Section 4.4.1. When $D = 3$, we have that $\epsilon = 0.5$, which follows from Theorem 12, i.e. $\lfloor \frac{N-2}{2} \rfloor - 1 \leq a_{\max}(N, D = 3) \leq \lfloor \frac{N-2}{2} \rfloor$. Moreover,

Theorem 13 *For a constant diameter D and large N, all non-trivial eigenvalues of the Laplacian matrix of any graph in the class $G_D^*(n_1, n_2, ..., n_D, n_{D+1})$ scale linearly with N, the number of nodes.*

Proof. Each non-trivial eigenvalue μ of the Laplacian satisfies the eigenvalue equation $-\widetilde{M}y = \mu y$, where y is the corresponding normalized eigenvector such that the Euclidian norm $y^T y = \|y\|_2^2 = 1$. We now define the rational number $\alpha_j = \frac{n_j}{N}$, for each $1 \leq j \leq D + 1$. It follows $N = \sum_{j=1}^{D+1} n_j$ and $n_j \geq 1$ that $0 < \alpha_j < 1$ and $\sum_{j=1}^{D+1} \alpha_j = 1$. The Jacobian matrix then becomes $\widetilde{M} = N.\widetilde{R}$, where

$$
\widetilde{R} = \begin{bmatrix}
-\alpha_2 & \sqrt{\alpha_1 \alpha_2} & & & \\
\sqrt{\alpha_1 \alpha_2} & -(\alpha_1 + \alpha_3) & \sqrt{\alpha_2 \alpha_3} & & \\
& \ddots & \ddots & \ddots & \\
& & \sqrt{\alpha_{D-1} \alpha_D} & -(\alpha_{D-1} + \alpha_{D+1}) & \sqrt{\alpha_D \alpha_{D+1}} \\
& & & \sqrt{\alpha_D \alpha_{D+1}} & -\alpha_D
\end{bmatrix}
$$

For small N, the dependence[2] of α_j on N will influence μ. For large N, on the other hand, since the norm of $\widetilde{R}$ is bounded for a constant D independent of N, and the eigenvector y is normalized, we observe from the eigenvalue equation that $\mu = -N y^T \widetilde{R} y = N(c + o(1))$, where c is only dependent on D. This means that, for large N, the eigenvalue μ scales linearly with N. ∎

Combining Theorem 13 and 6 and Corollary 11 implies that, for large N, the highest possible achievable algebraic connectivity in networks $G(N, D)$ is a linear function of N, provided the diameter D is independent from N.

We start the investigation of the relation between $a_{\max}(N, D)$ and diameter D by proving that the maximum of any eigenvalue $\mu_{i\max}(G(N, D))$, $i \in [1, N]$ is non-increasing as the diameter D increases. The proof is based on the following clique merging operation.

Definition 14 *Clique merging: In any graph with diameter D of the class $G_D^*(n_1, n_2, ..., n_i, n_{i+1}, ..., n_{D+1})$, any two adjacent cliques K_{n_i} and $K_{n_{i+1}}$ can be merged into one clique, resulting into a graph with diameter $D-1$, i.e. $G_{D-1}^*(n_1, n_2, ..., n_i + n_{i+1}, ..., n_{D+1})$. The merging of clique K_{n_i} and $K_{n_{i+1}}$ is obtained by adding $n_i n_{i+2}$ links such that clique K_{n_i} is fully meshed with clique $K_{n_{i+2}}$ (if $i + 2 \leq D + 1$) and by adding $n_{i-1} n_{i+1}$ links such that the clique $K_{n_{i+1}}$ is fully meshed with clique $K_{n_{i-1}}$ (if $1 \leq i - 1$).*

Figure 4.4 presents an example of clique merging. Clique K_{n_3} and K_{n_4} in Figure 4.4(a) $G_{D=4}^*(n_1 = 3, n_2 = 1, n_3 = 2, n_4 = 1, n_5 = 2)$ are merged into one clique, which results in Figure 4.4(b) $G_{D=3}^*(n_1 = 3, n_2 = 1, n_3 + n_4 = 3, n_5 = 2)$. The clique merging consists of purely adding links (the blue dotted line).

---

[2]In particular, for $n_1 = n_{D+1} = 1$, the dependence on N is obvious because $\alpha_1 = \alpha_{D+1} = \frac{1}{N}$.
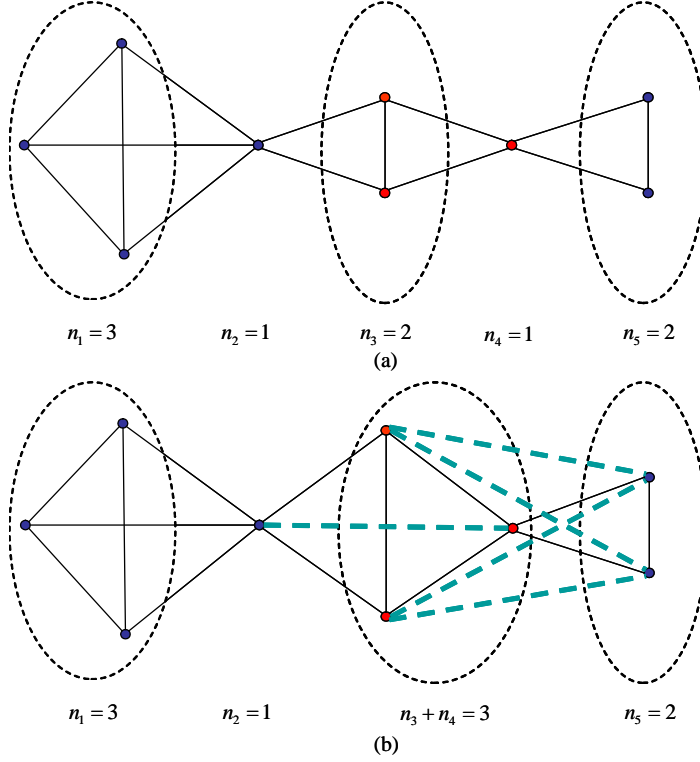
Figure 4.4: (b) $G^*_{D=4}(n_1 = 3, n_2 = 1, n_3 + n_4 = 3, n_5 = 2)$ is obtained by merging clique K_{n_3} and K_{n_4} in (a) $G^*_{D=3}(n_1 = 3, n_2 = 1, n_3 = 2, n_4 = 1, n_5 = 2)$ via adding the blue dotted links.

Theorem 15 *Given the network size N, the maximum of any eigenvalue $\mu_{i\,\max}(G(N, D))$, $i \in [1, N]$ is non-increasing as the diameter D increases, i.e. $\mu_{i\,\max}(G(N, D + 1)) \leq \mu_{i\,\max}(G(N, D))$.*

Proof. Assume that the graph $G^*_{D+1}(n'_1 = 1, n'_2, ..., n'_i, ..., n'_{D+1}, n'_{D+2} = 1)$ possesses the maximum eigenvalue $\mu_{i\,\max}(G(N, D + 1))$, $i \in [1, N]$ among all graphs with size N and diameter $D + 1$. Any two adjacent cliques can be merged by only adding links, which results in $G^*_D(n'_1 = 1, n'_2, ..., n'_i + n'_{i+1}, ..., n'_{D+1}, n'_{D+2} = 1)$. Hence, the graph $G^*_{D+1}(n'_1 = 1, n'_2, ..., n'_i, ..., n'_{D+1}, n'_{D+2} = 1)$ is a subgraph of $G^*_D(n'_1 = 1, n'_2, ..., n'_i + n'_{i+1}, ..., n'_{D+1}, n'_{D+2} = 1)$. According to the interlacing property in the proof of Theorem 7, we have $\mu_{i\,\max}(G(N, D + 1)) \leq \mu_i(G^*_D(n'_1 = 1, n'_2, ..., n'_i + n'_{i+1}, ..., n'_{D+1}, n'_{D+2} = 1))$. Furthermore, $G^*_D(n'_1 = 1, n'_2, ..., n'_i + n'_{i+1}, ..., n'_{D+1}, n'_{D+2} = 1)$ does not necessarily possess the maximum eigenvalue $\mu_{i\,\max}(G(N, D))$, i.e. $\mu_i(G^*_D(n'_1 = 1, n'_2, ..., n'_i + n'_{i+1}, ..., n'_{D+1}, n'_{D+2} = 1)) \leq \mu_{i\,\max}(G(N, D))$. Thus, $\mu_{i\,\max}(G(N, D+1)) \leq \mu_{i\,\max}(G(N, D))$. ∎
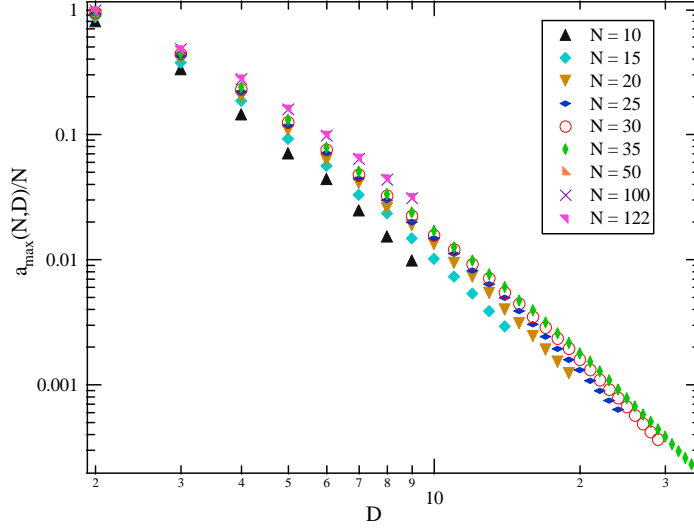
Figure 4.5: The scaled maximal algebraic connectivity $a_{\max}(G(N, D))/N$ (marker) as a function of the diameter D in log-log scale.

In view of the linear relation between $a_{\max}(N, D)$ and N, we present in Figure 4.5 the scaled maximal algebraic connectivity $a_{\max}(G(N, D))/N$ in relation with the diameter D, when $10 \leq N \leq 122$. The maximal algebraic connectivity $a_{\max}(G(N, D))$ is presented for all possible diameters, i.e. $1 \leq D \leq N - 1$ when $10 \leq N \leq 35$ and for $D < 10$ when N is large. The decrease of $a_{\max}(G(N, D))/N$ as a function of D is always slower than an exponential $b \exp(-\epsilon N)$ and close to (but faster than) a power law $bN^{-\epsilon}$. For large N, the scaled algebraic connectivity $a_{\max}(G(N, D))/N$ is expected to follow a universal function of diameter D.

The corresponding clique sizes of $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$ that maximizes the algebraic connectivity are partially given in Appendix C and completely documented in [96]. A symmetric clique size $(n_1, n_2, ..., n_{D+1})$ or a symmetric structure seems to be necessary to maximize the algebraic connectivity $a_{\max}(N, D)$. The graphs that achieve the maximum algebraic connectivity in $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$ have relative large sizes for cliques close to the middle.

4.4.3 Two proposed upper bounds for $a(N, D)$

Here, we discuss two upper bounds that are proposed in the literature [60][76]. Based on the upper bound

$$D \leq \left\lfloor \frac{\cosh^{-1}(N-1)}{\cosh^{-1}\left(\frac{\mu_1 + a}{\mu_1 - a}\right)} \right\rfloor + 1$$

given by Chung [21], where μ_1 is the largest eigenvalue of the Laplacian Q and a is the algebraic connectivity, Lin and Zhan [60] obtain an upper bound on $\frac{a}{\mu_1}$

$$\frac{a}{\mu_1} \leq \frac{\cosh\left(\frac{\cosh^{-1}(N-1)}{D-1}\right) - 1}{\cosh\left(\frac{\cosh^{-1}(N-1)}{D-1}\right) + 1}$$

Combining a simple upper bound on μ_1

$$\mu_1 \leq N \tag{4.10}$$

Lin and Zhan [60] arrive at an upper bound of the algebraic connectivity in relation to D and N

$$a(G(N, D)) \leq a_{up}(N, D) = N \frac{\cosh\left(\frac{\cosh^{-1}(N-1)}{D-1}\right) - 1}{\cosh\left(\frac{\cosh^{-1}(N-1)}{D-1}\right) + 1} \tag{4.11}$$

For $D = 2$, $a_{\max}(N, D = 2) = N - 2$, which is equal to the upper bound (4.11).

Figure 4.6 illustrates that $a_{up}(N, D)$ loosely bounds the largest possible algebraic connectivity $a_{\max}(N, D)$. The upper bound $a_{up}(N, D)$ increases approximately linearly with N for $3 \leq D \leq 6$ and the corresponding slope is much higher than that of $a_{\max}(N, D)$. When $D = 3$, each node in clique K_{n_2} and K_{n_3} of $G^*_{D=3}(1, n_2, n_3, 1)$ possesses the maximum degree $d_{\max} = N - 2$. When $D = 4$, the maximum degree of $G^*_{D=4}(1, n_2, n_3, n_4, 1)$ is $d_{\max} = N - 3$, which corresponds to a node in clique K_{n_3}. Since, $\mu_1 \geq d_{\max} + 1$ [48][64], the class $G^*_D(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$ has $\mu_1 \approx N$, for $D = 3, 4$. Therefore, the relative loose bound of (4.11) is not introduced by the $\mu_1 \approx N$ approximation of (4.10), when $D = 3, 4$.

Alon and Milman [76] present another upper bound of the algebraic connectivity in relation to diameter D and the maximum degree $d_{\max}$

$$a(G) \leq \frac{2d_{\max}}{D^2} \left(\log_2 N^2\right)^2 \tag{4.12}$$

Hence,

$$G^*_{D=3}(1, n_2, n_3, 1) \leq \frac{2(N-2)}{9} \left(\log_2 N^2\right)^2$$
$$G^*_{D=4}(1, n_2, n_3, n_4, 1) \leq \frac{2(N-3)}{16} \left(\log_2 N^2\right)^2$$
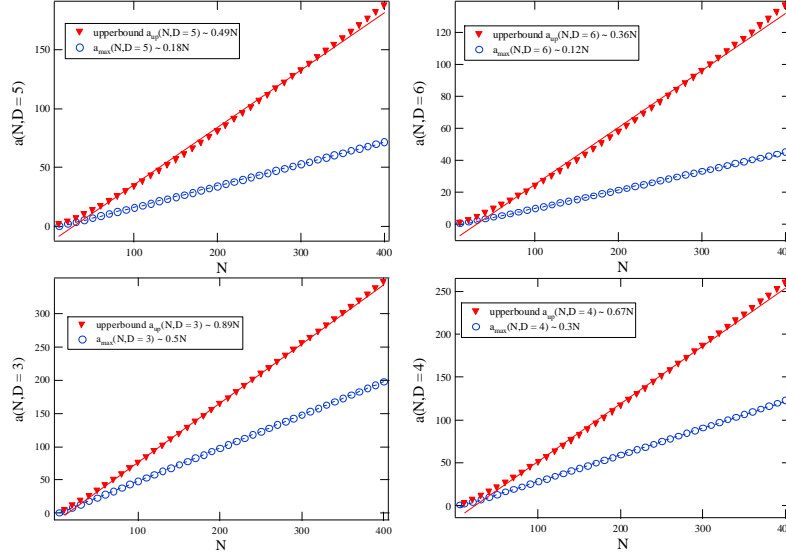
Figure 4.6: Comparison of $a_{\max}(N, D)$ and the upper bound of $a(N, D)$ when $3 \le D \le 6$.

which bounds the $a_{\max}(N, D)$ even loser, especially for large N.

However, we should mention that the two upper bounds (4.11) and (4.12) may be tight in other cases. In view of the relative loose upper bounds, at least for smaller diameter $D \le 6$, the largest possible algebraic connectivity $a_{\max}(N, D)$ or its approximations derived from data fitting is of great interest. We refer to [96], where $a_{\max}(N, D)$ as well as its the corresponding topology are presented for a wide range of diameter D and size N.

4.5   Conclusion

We propose a class of graphs $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$, within which the largest number of links, the minimum average hopcount, and more interestingly, the maximum of any Laplacian eigenvalue among all graphs with N nodes and diameter D can be achieved. The largest possible algebraic connectivity $a_{\max}(N, D)$ is rigorously determined for diameter $D = 2, 3$ and $D = N - 1$. For other diameters, the maximum of any Laplacian eigenvalue can be searched within $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$, which is feasible due to the reduction in the Laplacian eigenvalue computation from a $N \times N$ to a $(D + 1) \times (D + 1)$ matrix.

Combining both the theoretical and numerical results, we have (1) illustrated the different topological features of graphs that maximize different Laplacian eigenvalues, which provides an evidence that graphs with large algebraic connectivity are robust with

respect to traffic engineering; (2) presented the relation between the maximum algebraic connectivity $a_{\max}(N, D)$ and the size N as well as the diameter D; (3) compared two upper bounds of the algebraic connectivity proposed in literature with the largest possible $a_{\max}(N, D)$ for small diameter. This is a first step to explore the application of these maximal possible Laplacian eigenvalues via the class $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$. Rich mathematical results related to the characteristic polynomial of both the Laplacian and adjacency matrix are documented in [73], which is, however, still far from being able to analytically determine the graph optimizing a given eigenvalue. More numerical results about the $a_{\max}(N, D)$ as well as the corresponding graph are being collected and updated in [96].
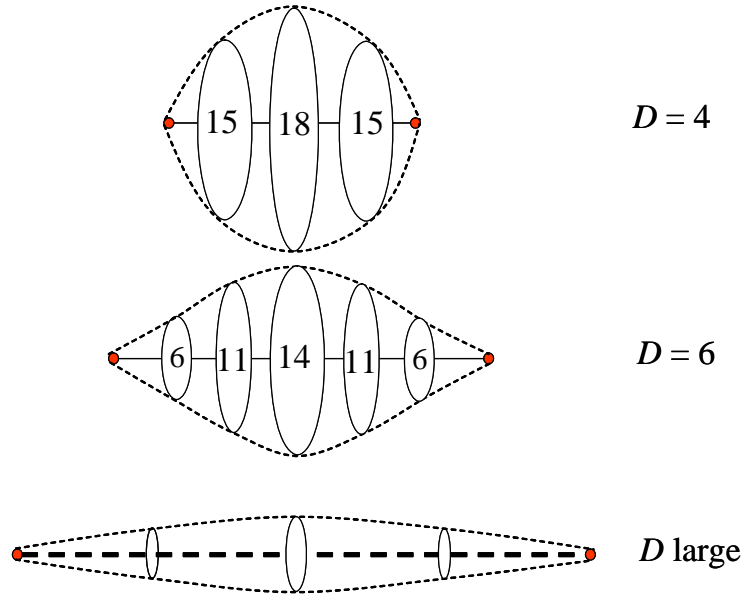
Figure 4.7: The optimal graphs G_D^* with $N = 50$ and diameter D.

By choosing the algebraic connectivity as the metric for network robustness, the resulting optimal graphs G_D^* with given number of node N and diameter D are mostly symmetric and the cliques in the middle have a larger size than cliques close to the borders. If D is sufficiently long, the optimal structure is homogeneous and only deviations occur at the ends, as shown in Figure 4.7. These features are also observed in those long molecules in nature [6] represented by remarkable homogeneous strings which suggest extremal properties. We may regard evolution as an optimization process over years. Then, the homogeneous string structure in proteins, DNA structures and our optimal graphs seem the results of optimizations in a same direction. Thus, the importance of algebraic connectivity may be far beyond the current understanding.

Chapter 5

Optimize $a(G)$ via link addition

To improve the performance of an existing large real-world network, instead of substituting the infrastructure for the optimal one that maximizes the robustness metric $R(G)$ as discussed in Chapter 4, a minor modification on the current network, i.e. adding a small number of links, is usually required due to economic concerns. An important question is how can we refine the network to improve its robustness. In this chapter, we confine our problem as the following: "where should we add a link to a network G such that the algebraic connectivity $a(G)$ can be increased the most?" The investigation on adding one link to improve the algebraic connectivity will also provide insights on how to dynamically add a set of links one by one so that the algebraic connectivity is maximally increased.

The number of possibilities of adding a link to a network $G(N, L)$ with N nodes and L links is $\binom{N}{2} - L$. For large realistic (hence sparse) networks, it is infeasible to compare all these possibilities and find the optimal one. Hence, we propose two strategies of adding a link to optimize the algebraic connectivity. The node pair (i, j) where a link can be added, can be characterized by various topological metrics such as the node degree, or the distance between the node pair. We propose two strategies based on these topological metrics. Strategies are compared together with random link addition in various classes of networks.

5.1 Strategies of adding a link to optimize $a(G)$

In this section, we investigate the increase of the algebraic connectivity due to the addition of a link between $\{i, j\}$ in relation to the topological characteristics of $\{i, j\}$. We propose link addition strategies based on structural metrics as well as on spectral metrics of $\{i, j\}$, respectively.

5.1.1 Structural metrics based strategy

Since $\mu_{N-1}(G) \leq \mu_{N-1}(G+e) \leq \mu_{N-2}(G)$ according to Theorem 1, we define the normalized increase of the algebraic connectivity as $0 \leq \mu^* = \frac{\mu_{N-1}(G+e)-\mu_{N-1}(G)}{\mu_{N-2}(G)-\mu_{N-1}(G)} \leq 1$. A link $e = (i,j)$ can be added if i and j are not yet connected in G. We start with a specific network to explore all possible realizations of adding a link to a graph in order to gain some insights in the relationship between μ^* and the topological characteristics of $\{i,j\}$ such as the degree, the clustering coefficient and the node betweenness of node i and j, and the hopcount between node i and j.

The correlation between μ^* introduced by the addition of link (i,j) and topological metrics of the node pair $\{i,j\}$ is supposed to be topology dependent. We consider the following topologies: the Erdös-Rényi random graph $G_p(N)$ and the BA power law graph.

Erdös-Rényi random graph $G_p(N)$

Given one Erdös-Rényi random graph $G = G_{0.6}(100)$, in each realization, a link $e = (i,j)$ is added to G provided it does not exist in G, and we calculate μ^* and topological metrics of $\{i,j\}$. All possible realizations of adding a link to G are considered. Figure 5.1(a)

(a)   (b)   (c)
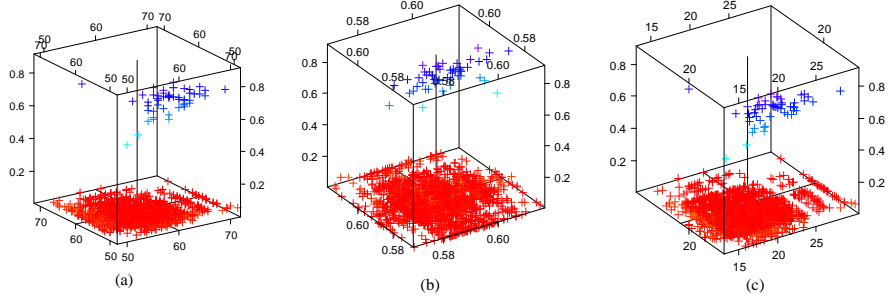
Figure 5.1: The increase of the algebraic connectivity μ^* (the z axis) due to the addition of link $\{i,j\}$ in relation with (a) the degree of node i and j; (b) the clustering coefficient of node i and j (c) the betweenness of node i and j separately in x and y axis in an Erdös-Rényi random graph $G_{0.6}(100)$.

shows that μ^* is large if and only if one node of the pair $\{i,j\}$ has small degree. A similar correlation between μ^* and the node betweenness is observed in Figure 5.1(c). Furthermore, Figure 5.2 illustrates that the node betweenness is positively correlated with the degree of the node. Hence, the degree will be considered instead of the node betweenness whose computational complexity is much higher. The correlation between μ^* and the clustering coefficient is relatively weak as illustrated in Figure 5.1(b). There
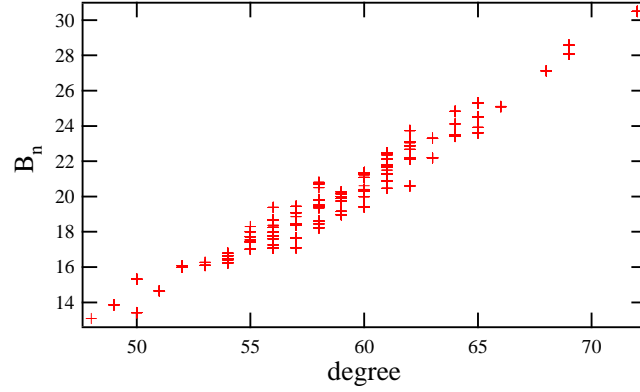
Figure 5.2: The degree and the betweenness of a node in an Erdös-Rényi random graph $G_{0.6}(100)$.

exists hardly correlation between μ^* and the hopcount between i and j, which is mostly equal to 2 in such a dense graph.

BA power law graph

We perform the same experiment in a BA power law graph with $N = 100$ and $m = 3$. A small m is selected such that the power law degree distribution can be observed for $N = 100$.
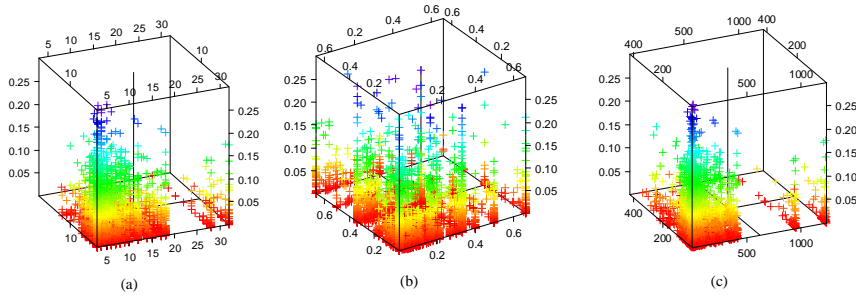
Figure 5.3: The increase of the algebraic connectivity μ^* (the z axis) due to the addition of link $\{i, j\}$ in relation with (a) the degree of node i and j; (b) the clustering coefficient of node i and j (c) the betweenness of node i and j seperately in x and y axis in a BA power law graph with $N = 100$ and $m = 3$.

As shown in Figure 5.3, different from an Erdös-Rényi random graph, an relative high increase in μ^* implies low degree of both node i and j. However, the inverse

does not hold. Adding a link to two low degree nodes only modestly increases μ^*. In fact, many nodes in a BA power law graph possess a low degree. Again, the positive correlation between betweenness and degree of a node observed in Figure 5.4, persuades us to consider the node degree which is simpler to compute than the node betweenness. Similarly, no clear correlation between μ^* and the clustering coefficient as well as the hopcount between i and j appear in Figure 5.3 (b) and 5.5.
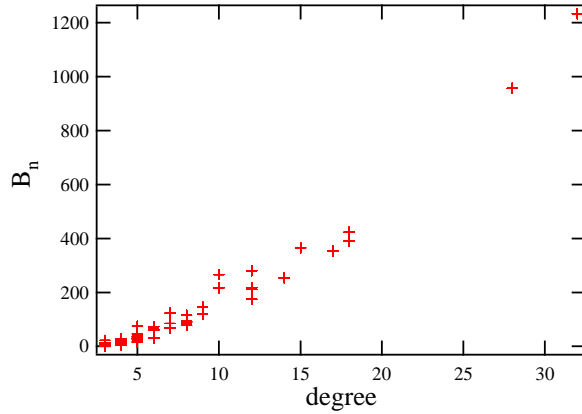
Figure 5.4: The degree and the node betweenness of a node in a BA model with $N = 100$ and $m = 3$.

In view of the correlation between μ^* and the topological metrics regarding to node i and j where a link is added in both the Erdös-Rényi random graph and the BA power law graph, we propose a structural metric based strategy that adds a link between a minimum degree node and a random other node, which are not originally connected. This strategy is also motivated by the upper bound $a(G) \leq d_{\min}(G)$, where the algebraic connectivity is shown to be limited by the lowest degree nodes. Finally, the metric degree is the simplest to calculate and it can be obtained only from the node and its neighbors, i.e. local information.

5.1.2   Fiedler vector based strategy

Maas [61] showed that after inserting an edge between node i and j, the upper and lower bounds of the algebraic connectivity $\mu_{N-1}(G + e)$

$$\min \left\{ \mu_{N-1}(G) + \frac{\varepsilon \eta^2}{\varepsilon + (2 - \eta^2)}, \mu_{N-2}(G) - \varepsilon \right\} \leq$$
$$\mu_{N-1}(G + e) \leq \min \left\{ \eta^2 + \mu_{N-1}(G), \mu_{N-2}(G) \right\}$$
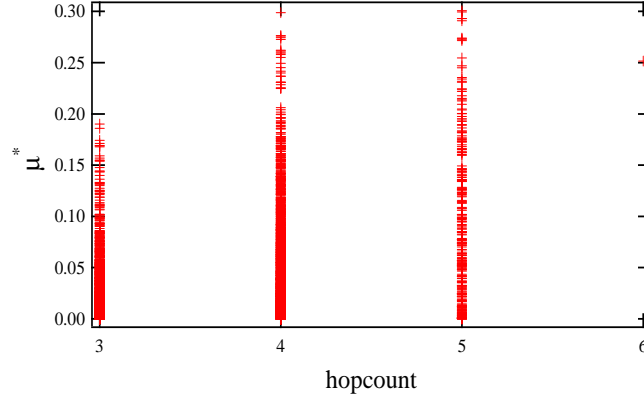
Figure 5.5: The increase of the algebraic connectivity μ^* due to the addition of link (i, j) in relation with the hopcount between i and j in a BA model with $N = 100$ and $m = 3$.

are related to $\eta = |u_i - u_j|$, the absolute difference between the $i-th$ and $j-th$ elements of the Fiedler vector u of G. The Fiedler vector u is the eigenvector corresponding to the second smallest eigenvalue (i.e., the algebraic connectivity) of the Laplacian matrix of a graph. In the lower bound, the first term increases with increasing ε whereas the second one decreases. The highest lower bound can be achieved by a choice of ε that makes both terms equal:

$$\varepsilon = \frac{\beta - 2}{2} + \left(\frac{(\beta - 2)^2}{4} + \beta \left(2 - \eta^2 \right) \right)^{\frac{1}{2}} \geq 0$$

where $\beta = \mu_{N-2} - \mu_{N-1} \geq 0$. The higher η is, the lower is ε, and the higher the highest lower bound is. Higher η also contributes possibly to a higher upper bound. Hence, $\mu_{N-1}(G + e)$ tends to be large if $\eta = |u_i - u_j|$ is large.

This is further illustrated by Figure 5.6, where all realizations of adding a link are performed in both an Erdös-Rényi random graph $G_{0.6}(100)$ and a BA power law graph with $N = 100$ and $m = 3$. The μ^* is positively correlated with $\eta = |u_i - u_j|$.

Therefore, we propose the strategy of adding a link to the node pair $\{i, j\}$ with the highest $\eta = |u_i - u_j|$, such that the algebraic connectivity can be increased the most. As shown in Figure 5.6, if we apply the strategy of adding a link to the node pair with the highest $|u_i - u_j|$, i.e. $|u_i - u_j|_{\max} = 0.986$ and $|u_i - u_j|_{\max} = 0.468$ in these two graphs, the algebraic connectivity is increased by $\mu^* = 0.859$ and $\mu^* = 0.293$, which is close to the corresponding maximal possible increase $\mu^*_{\max} = 0.917$ and $\mu^*_{\max} = 0.301$, respectively.
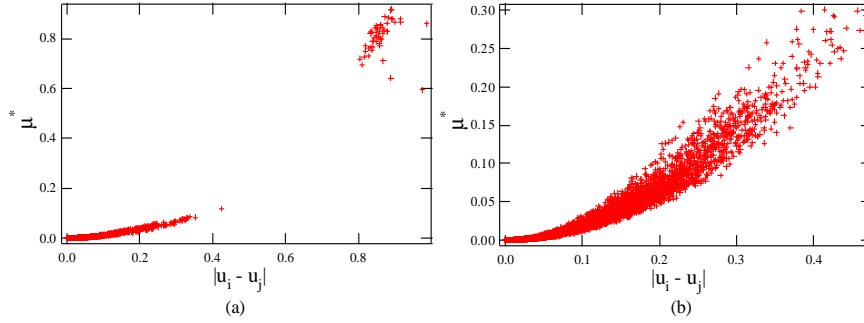
Figure 5.6: The increase of the algebraic connectivity μ^* due to the addition of link $\{i, j\}$ in relation with $|u_i - u_j|$ in (a) an Erdös-Rényi random graph $G_{0.6}(100)$ (b) a BA power law graph with $N = 100$ and $m = 3$.

5.2   Strategy evaluation

The strategies proposed are based on the correlation study in specific graphs. Since correlations between metrics are topology dependent, a strategy, that works in one topology, may not work in another network. Hence, in this section, we examine the validity of these two strategies in the class of the Erdös-Rényi random graphs, BA power law graphs and k-ary tree, where each node is connected to k children. From the Erdös-Rényi random graph, over the BA model to the k-ary tree, both the randomness of node interconnections and link density decrease.

We carried out 10^4 iterations for each simulation. In each iteration, an Erdös-Rényi random graph $G_p(N)$ or a BA power law graph is generated and the second and the third smallest Laplacian eigenvalues are calculated. First, we add a link between a random non-connected node pair without any strategy and obtain the normalized increase of the algebraic connectivity $\mu^*_{(0)}$. Second, we perform strategy 1: add a link between the node with the minimum degree and a random non-connected node. The corresponding increase of the algebraic connectivity is $\mu^*_{(1)}$. Third, we perform strategy 2: add a link to the node pair $\{i, j\}$ with the highest $\eta = |u_i - u_j|$ and obtain $\mu^*_{(2)}$.

5.2.1   Erdös-Rényi random graph $G_p(N)$

Simulations are carried out for Erdös-Rényi random graph with $p = 0.6, N = 50, 100,$ 200, 400, 800. Besides, with fixed $N = 200$, we perform simulations for $p = 0.2, 0.4, 0.6$ and 0.8. We evaluate the gain of the strategy by comparing with the link addition without any strategy, i.e. $\Pr[\mu^*_{(s)} - \mu^*_{(0)} \geq x]$, $s = 1$ and 2 for strategy 1 and 2, respectively. Any link addition, including the optimal one follows $\mu^* \leq 1$. Hence, $\mu^*_{(s)} - \mu^*_{(0)} < 1$. Figure 5.7 and 5.8 show that both strategies perform better than

random link addition, with probability larger than 0.9, i.e. $\Pr[\mu^*_{(s)} - \mu^*_{(0)} \geq 0] > 0.9$. Both strategies work better for smaller networks as shown in Figure 5.7. Strategy 2 performs better than strategy 1 in Erdös-Rényi random graphs with various network size N and link density p. The effect of the link density is not obvious in Erdös-Rényi random graphs which are generally dense.
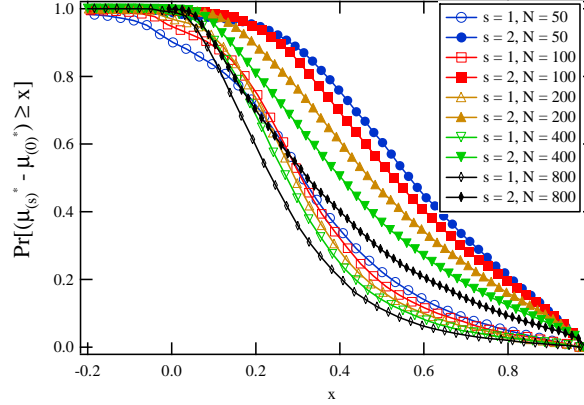
Figure 5.7: Gain $\Pr[\mu^*_{(s)} - \mu^*_{(0)} \geq x]$ of strategy 1 ($s = 1$) and strategy 2 ($s = 2$) in the Erdös-Rényi random graph $G_p(N)$, where $p = 0.6$.

As shown in Table 5.1, the effect of strategy 1 and 2 is evident in $E[\mu^*_{(1)}]$ and $E[\mu^*_{(2)}]$ compared to $E[\mu^*_{(0)}]$ when no strategy is applied. Strategy 2 performs better, but it requires the knowledge of the whole network topology. Only local information, the degree, is needed in strategy 1.

Table 5.1: Strategies comparison in Erdös-Rényi random graphs.

$p = 0.6$	$N = 50$	$N = 100$	$N = 200$	$N = 400$	$N = 800$
$E[\mu^*_{(2)}]$	0.644	0.594	0.532	0.469	0.401
$E[\mu^*_{(1)}]$	0.407	0.387	0.364	0.340	0.303
$E[\mu^*_{(0)}]$	0.049	0.0214	0.00951	0.0042	0.00169
$N = 200$	$p = 0.2$	$p = 0.4$	$p = 0.6$	$p = 0.8$	
$E[\mu^*_{(2)}]$	0.588	0.544	0.532	0.579	
$E[\mu^*_{(1)}]$	0.481	0.396	0.364	0.391	
$E[\mu^*_{(0)}]$	0.012	0.0092	0.0095	0.011	
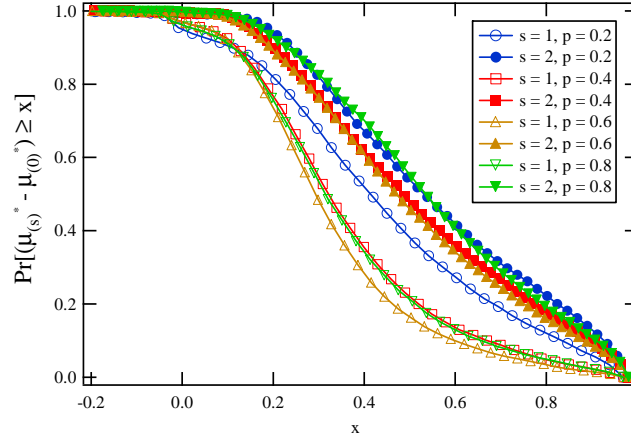
Figure 5.8: Gain $\Pr[\mu^*_{(s)} - \mu^*_{(0)} \geq x]$ of strategy 1 ($s = 1$) and strategy 2 ($s = 2$) in the Erdös-Rényi random graph $G_p(N)$, where $N = 200$.

5.2.2 BA power law graph

The same simulations are carried out in the BA model with $N = 100, 200, 400$ and $m = 3, 4, 5$. As illustrated in Figure 5.9, strategy 1 behaves the same as link addition without strategy, while, strategy 2 performs better with probability round 0.9 than random link addition, i.e. $\Pr[\mu^*_{(s)} \geq \mu^*_{(0)}] \approx 0.9$. Furthermore, m (or equivalently the link density) has no effect on the performance of both strategies. Figure 5.10 shows that strategy 2 works slightly better on BA power law graphs with smaller size N.
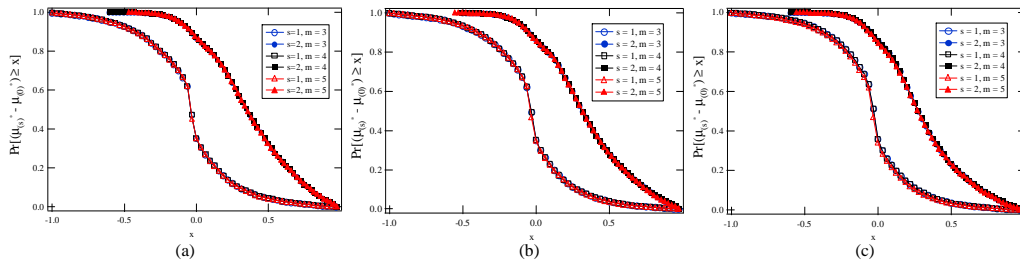
Figure 5.9: Gain $\Pr[\mu^*_{(s)} - \mu^*_{(0)} \geq x]$ of strategy 1 ($s = 1$) and strategy 2 ($s = 2$) in the BA model with (a) $N = 100$ (b) $N = 200$ and (c) $N = 400$.
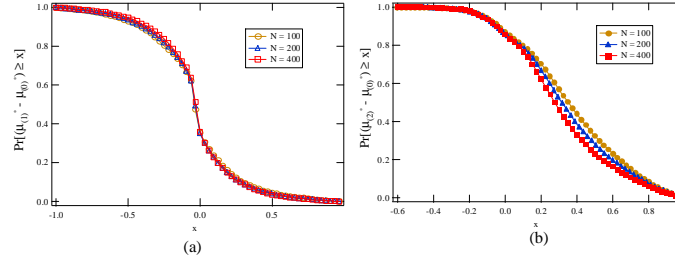
Figure 5.10: (a) Gain $\Pr[\mu_{(1)}^* - \mu_{(0)}^* \geq x]$ of strategy 1 and (b) gain $\Pr[\mu_{(2)}^* - \mu_{(0)}^* \geq x]$ of strategy 2 in the BA model with $m = 3$.

Table 5.2: Strategies comparison in BA power law graphs.

$m = 3$	$N = 100$	$N = 200$	$N = 400$
$E[\mu_{(2)}^*]$	0.558	0.521	0.477
$E[\mu_{(1)}^*]$	0.153	0.146	0.142
$E[\mu_{(0)}^*]$	0.170	0.161	0.148

5.2.3 K-ary tree

We investigate the k-ary tree [67] of depth[1] d where each node has exactly k children. In a k-ary tree the total number of nodes is

$$N(d) = 1 + k + k^2 + \cdots + k^d = \begin{cases} \frac{k^{d+1}-1}{k-1}, & k \neq 1 \\ 1 + d, & k = 1 \end{cases}$$

Since the interconnections of nodes in a k-ary tree is determined, we examine one specific k-ary tree with $k = 2, d = 7$ and try all the possibilities of adding a link. On average, the algebraic connectivity can be increased by $E[\mu_{(0)}^*] = 0.137$. On average, strategy 1 improves the algebraic connectivity by $E[\mu_{(1)}^*] = 0.129 < E[\mu_{(0)}^*]$, which is not better than random link addition. Given the level of the node pair, where a link is added, μ^* is found to be higher if these two nodes share fewer common parents. The optimal link addition is between two nodes both in level 2. Hence, none of them has degree 1. As shown in Figure 5.11, the positive correlation between μ^* and $|u_i - u_j|$ still remains, but weaker than in the Erdös-Rényi random graph and the BA power law graph. The increase of algebraic connectivity achieved by strategy 2 is $\mu_{(2)}^* = 0.24$, round half of the maximal possible increase $\mu_{\max}^* = 0.51$.

---

[1]The depth d is the number of hops (or links) from the root to a node at the leaves.
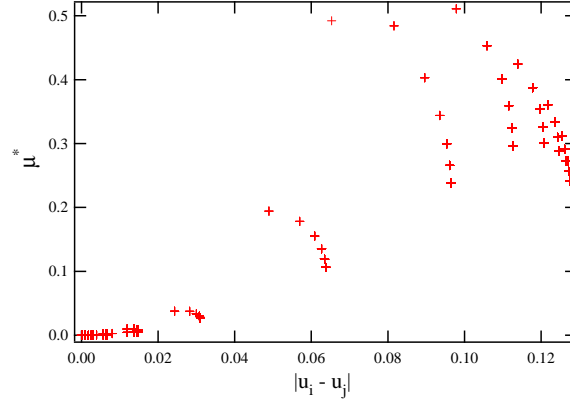
Figure 5.11: The increase of the algebraic connectivity μ^* due to the addition of link $\{i, j\}$ in relation with $|u_i - u_j|$ in a k-ary tree with $k = 2, d = 7$.

5.2.4   Comparison with optimal link addition

Our strategies are evaluated via $\Pr[\mu^*_{(s)} - \mu^*_{(0)} \geq x]$, $s = 1$ and 2 for strategy 1 and 2 respectively, by comparing them with the random link addition. One may be curious to know how far our strategies are from the optimal link addition. The maximal possible increase or optimal increase of the algebraic connectivity $\mu^*_{\max}$ can be obtained by trying all possible ways of adding a link to a network. Due to the computational complexity, we compare our strategies with the optimal link addition in 10^4 Erdös-Rényi random graphs $G_{0.6}(100)$ and in 10^4 BA power law graphs with $N = 100, 200$ and $m = 5$. As

Table 5.3: Strategies comparison with optimal link addition

	$G_{0.6}(100)$	BA model $N = 100, m = 5$	BA model $N = 200, m = 5$
$E[\mu^*_{\max}]$	0.62	0.87	0.88
$E[\mu^*_{(2)}]$	0.59	0.56	0.52
$E[\mu^*_{(1)}]$	0.39	0.15	0.14
$E[\mu^*_{(0)}]$	0.021	0.17	0.16

shown in Table 5.3, strategy 2 performs generally better than strategy 1, and is close to the optimal link addition, especially in Erdös-Rényi random graphs.

5.3 Conclusion

The difficulty of applying strategies to optimize the algebraic connectivity (or in general a topological metric R) by adding a link $e = (i, j)$ based on topological characteristics of node pair $\{i, j\}$ is due to the fact that the correlation between metrics (e.g. between μ^* and degrees of the node pair) is topology dependent. In other words, a strategy, that works for a given type of graphs, may not work in other classes of networks.

Strategy 1 seems to perform better in the class of dense graphs such as the Erdös-Rényi random graphs, although the effect of the specific link density of Erdös-Rényi random graph is not obvious. However, strategy 1 looses its effect in sparse networks like the BA power law graph and the k-ary tree, where many nodes possess the minimal degree. Another extreme example is $G = K_N \backslash \{i, j\}$, the complete graph except one link. Each node has the maximal degree $N - 1$ except that node i and j have the minimal degree $N - 2$. Adding a link between i and j leads to the maximal increase of the algebraic connectivity $\mu_{N-1}(G + e) = \mu_{N-1}(G) + 2$ and $\mu^* = 1$ according to Barik and Pati [82].

Strategy 2 seems to be applied better in graphs with more randomness in interconnections of nodes. From the Erdös-Rényi random graph, over the BA model to the k-ary tree, the randomness of node interconnections decreases. Strategy 2 performs better in Erdös-Rényi random graphs and BA power law graphs than in the k-ary tree, which is regular in node interconnections. In other words, the positive correlation between μ^* and $|u_i - u_j|$ is weaker in the k-ary tree. The worst case is the lattice. A lattice is even more regular than a k-ary tree in the sense that more nodes have the same degree except nodes in the border. The second and third smallest eigenvalue of a lattice are the same. The algebraic connectivity remains the same wherever a link is added. Hence, the correlation between μ^* and $|u_i - u_j|$ disappears completely.

Strategy 2 performs generally better than strategy 1 in Erdös-Rényi random graph, BA model and k-ary tree. However, strategy 2 requires the Fiedler vector, or the whole network topology, while only the node degree is needed in strategy 1, which can be applied even when the network is partially known. The computational complexity[2] is reduced from $O(N^5)$ of the optimal link addition searching to $O(N^3)$ by strategy 2 and to $O(N)$ by strategy 1. Finally, strategy 2 is shown to be close to the optimal link addition, especially in Erdös-Rényi random graphs.

---

[2]The computational complexity of eigenvalues is $O(N^3)$.

Part III

Interplay between network and service

Chapter 6

The observable part of a Network

In communications networks, traffic is usually carried along the shortest paths such that resources of a network are most efficiently used. In this chapter, we study the structure of the overlay $G_{\cup spt}$ formed by the union of all shortest path trees SPT in a graph $G(N, L)$ with N nodes and L links, where a SPT is the union of the shortest paths from one node to all the other nodes. The relation between the overlay $G_{\cup spt}$ and the underlying graph or substrate $G(N, L)$ is shown in Figure 6.1. The overlay $G_{\cup spt}$,
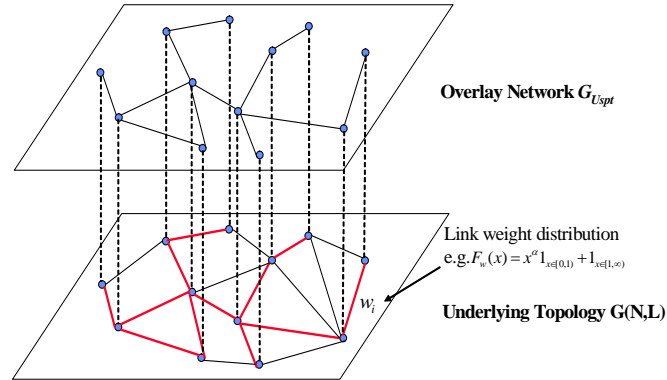
Figure 6.1: The relation between the overlay network and the underlying topology.

which is also the union of the shortest paths between all possible pairs of nodes, can be regarded as the "transport overlay network" on top of the network or substrate $G(N, L)$. In the Internet, for example, traffic is carried along the overlay $G_{\cup spt}$, composed of a fraction of the links in the underlying network, which is just the maximal part of the Internet that we can actually observe by traceroute measurements [86]. We discuss three potential applications that motivate a study of $G_{\cup spt}$: performance or robustness characterization, traffic engineering and Internet topology interference.

The importance of overlay networks is believed to grow in the future. One example of an overlay network is peer-to-peer networks [7] with n distributed systems sharing resources such as content, CPU cycles and storage, where n is smaller than the number of nodes N in the underlying network. The peer-to-peer overlay network can be regarded as a union of paths connecting these n nodes. Another type of overlay network is a virtual private network (VPN), a private network that uses a public network (usually the Internet or the telephony network) to connect remote sites or users together. The physical networks traversed by both the peer-to-peer and the VPN overlay networks are a subgraph of $G_{\cup spt}$. The overlay $G_{\cup spt}$, not the substrate, determines the network's performance, because any link removed in $G_{\cup spt}$ will definitely impact at least those flows of traffic that pass over that link.

The topology of the overlay $G_{\cup spt}$ is determined by the underlying graph $G(N, L)$ as well as the link weight structure. Link weight tuning is regarded as part of the routing in the service. In this chapter, we investigate how the topology of the overlay network $G_{\cup spt}$ changes by tuning the link weight structure. Current best-effort routing simply computes appropriate paths based on a single, relatively static measure (e.g. the delay, the monetary cost, etc.). QoS routing takes into account multiple measures including both the applications' requirements and the availability of network resources. Hence, we investigate two approaches of link weight tuning as given in Section 2.2: (a) the link weight of each link is an i.i.d. polynomial random variable

$$F_w(x) = x^\alpha 1_{x\in[0,1]} + 1_{x\in(1,\infty)}, \quad \alpha > 0$$

which can be tuned via the extreme value index α and (b) each link is specified by a 2-dimensional link weight vector $\overrightarrow{w}(u \to v) = [w_1(u \to v), w_2(u \to v)]$, where the component w_1 and w_2 are correlated uniformly distributed random variables $\in [0,1]$ with correlation coefficient ρ. We tune the correlation coefficient $\rho \in [-1, 1]$ to vary the link weight structure. The $\alpha \to \infty$ regime is entirely determined by the topology of the graph because the link weight structure does not differentiate between links. The overlay network is equal to the underlying graph. In the $\alpha \to 0$ regime, all flows are transported the minimum spanning tree (MST), a tree spanning over all the nodes and whose total weight is minimum. Any failure in a node or link disconnects the MST into two parts and may result in obstruction of transport in the network. The $\alpha \to 0$ regime may constitute a weak regime although it is highly efficient: only $N-1$ links are used which means that a minimum of links need to be controlled and/or secured. From a traffic engineering point of view, choosing α around 1 will lead to the use of more paths and, hence, a more balanced overall network load than in the $\alpha \to 0$ regime. The precise traffic distribution among these links in the overlay will be explored in Chapter 8.

The final motivation applies to interfering the Internet topology. Recently, Lakhina *et al.* [59] have pointed to the effect of biases when trying to construct the Internet

topology from a few source trees that span a huge number of destinations, basically because the links close to the source have a substantially higher probability to be detected than links close to the destination. The potential dramatic effect of biases was recently rigorously analyzed in a mathematical analysis first by Clauset and Moore [22], and later extended by Achlioptas *et al.* [2]. They showed that, irrespective of the degree distribution of the substrate, the inferred topology deduced from only a few source trees to many destinations is likely to possess a power law degree distribution. For example, a random network with a Poisson degree distribution and a regular graph (with constant degrees) all lead to an observed power law degree distribution. These analyses place doubt on the believed power law degree structure of the Internet. Biases can be circumvented if sources and destinations are regarded as equally important. By constructing the union of all paths between $m < N$ testboxes, a sampled overlay network $G_{\cup_m spt}$ that is a subgraph of $G_{\cup spt}$ is obtained. An interesting issue, how large needs m to be such that $G_{\cup_m spt}$ sufficiently resembles properties of $G_{\cup spt}$, will be addressed in Chapter 7. In this chapter, we shed some light on how different $G_{\cup spt}$ can be compared to the underlying network and also show measurement results in Section 6.6 of a partial overlay $G_{\cup_m spt}$.

As we tune the index α of the polynomial link weights, we investigate the properties of the overlay networks $G_{\cup spt(\alpha)}$ on top of the complete graphs K_N, lattices and power law graphs. Structural properties of the overlay networks are examined by (a) the number of links, (b) the degree distribution, and (c) the spectrum. The effect of one and two dimensional link weight tuning on the overlay networks $G_{\cup spt}$ are compared in the dense substrates, i.e. the Erdös-Rényi random graphs.

6.1 The union of the shortest path trees $G_{\cup spt}$: Theory

Any set of links $l\,(i \to j)$, $l\,(j \to k)$ and $l\,(i \to k)$ between three nodes i, j and k in the union $G_{\cup spt}$ of the shortest path trees obeys the triangle inequality,

$$w\,(i \to k) \leq w\,(i \to j) + w\,(j \to k)$$

otherwise the link $l\,(i \to k)$ is not the shortest path from i to k and, consequently, does not belong to $G_{\cup spt}$. Hence, if $l\,(i \to j)$, $l\,(j \to k)$, $l\,(i \to k) \in \mathcal{L}_{G_{U spt}}$ where $\mathcal{L}_{G_{U spt}}$ is the set of links of the graph $G_{\cup spt}$, then

$$\Pr\left[w\,(i \to j) + w\,(j \to k) \geq w\,(i \to k)\right] = 1$$

A tree is a graph without cycles. Similarly as for a path, the weight of a tree T is $w(T) = \sum_{(i \to j) \in T} w(i \to j)$. A spanning tree T_G is a tree that contains or spans all N nodes of the graph G. A minimum spanning tree (MST) T_G^* is a minimum weight

spanning tree in G such that $w(T_G^*) \leq w(T_G)$ for all T_G in G. Since we assume networks are equipped with undirected i.i.d. link weights, the probability to have more than one shortest path or more than one MST is negligibly small. Algorithms to compute a shortest path (such as Dijkstra's) and a MST (such as Prim's and Krushkal's) are nicely explained in [25] and applied to data communication networks in [66]. If the MST does not exist, the graph is disconnected.

Theorem 16 *A minimum spanning tree belongs to* $G_{\cup spt}$.

Proof. The proof is by contradiction. Suppose that a MST does not belong to $G_{\cup spt}$. This means that there is at least one link $l(i \to j) \in MST$ which does not belong to the union of shortest path trees, $l(i \to j) \notin G_{\cup spt}$. Hence, the link $l(i \to j)$ is not the shortest path $P_{i \to j}^*$ from node i to node j implying that

$$w\left(P_{i \to j}^*\right) < w(i \to j)$$

In that case, we can lower the weight of the MST which equals

$$w(MST) = \sum_{(k \to l) \in MST} w(k \to l)$$

$$= w(i \to j) + \sum_{k \to l \in MST; k \to l \neq i \to j} w(k \to l)$$

by changing[1] $w(i \to j)$ for $w\left(P_{i \to j}^*\right)$. However, this is impossible since w_{MST} is, by definition, the tree that minimizes the above sum. This proves the theorem. $\blacksquare$

Any link $l(i \to j)$ with link weight $w(i \to j)$ in the $G_{\cup spt}$ must be the shortest path $P_{i \to j}^*$ between i and j because a link in the $G_{\cup spt}$ must belong to a shortest path and a subsection of a shortest path is also a shortest path. Conversely, if a link $l(i \to j)$ is the shortest path $P_{i \to j}^*$ between i and j, it must belong to the $G_{\cup spt}$, because the $G_{\cup spt}$ is the union of shortest paths between all possible source and destination nodes. Therefore, the event that a link $l(i \to j)$ is observed in (i.e. belongs to) the $G_{\cup spt}$ is equivalent to the event $\{P_{i \to j}^* = l(i \to j)\}$ that the link $l(i \to j)$ is the shortest path $P_{i \to j}^*$ between i and j. Hence, $\Pr[P_{i \to j}^* = l(i \to j)]$ is also the probability that a link can be observed.

Theorem 17 *In any graph with positive i.i.d. link weights w specified by the probability density function $f_w(x)$, the probability that a link $l(i \to j)$ between node i and j is the shortest path $P_{i \to j}^*$ between i and j is*

$$\Pr[P_{i \to j}^* = l(i \to j)] = \int_0^\infty \frac{f_w(x) \Pr[w(P_{i \to j}^*) > x]}{\Pr[w(i \to j) > x] + \frac{1 - p_{ij}}{p_{ij}}} dx \tag{6.1}$$

where $p_{ij} = \Pr[l(i \to j) \text{ exists}]$.

---

[1] Actually, since $P_{i \to j}^*$ must consist of at least two links, only those that are necessary to obtain a tree are needed in the MST such that we can further lower w_{MST}.

Proof. See Appendix A.3. ∎

Since $\Pr[P^*_{i\to j} = l\,(i \to j)] \leq 1$ and $\int_0^\infty f_w(x)dx = 1$, we see in (6.1) that

$$\frac{\Pr[w(P^*_{i\to j}) > x]}{\Pr[w(i \to j) > x] + \frac{1-p_{ij}}{p_{ij}}} \leq 1$$

A slightly tighter bound follows from the probability P_c in (A.8) and the left hand side of (A.11), that is bounded by p_{ij}, such that

$$\Pr[w(P^*_{i\to j}) > x] \leq p_{ij}\Pr[w(i \to j) > x] + 1 - p_{ij} \tag{6.2}$$

In words, the probability that the weight of the shortest path exceeds x is always less than or equal to the probability that an arbitrary link weight exceeds x (because of the assumption of i.i.d. link weights) multiplied by the probability of the existence of the link plus the probability $1 - p_{ij}$. The bound (6.2) is sharpest in case $p_{ij} = 1$, thus, in case the direct link $i \to j$ exists surely.

Corollary 18 *In any graph with N nodes and with positive i.i.d. link weights, we can write*

$$\Pr[P^*_{i\to j} = l\,(i \to j)] = \Pr[H_N = 1]$$
$$= -\int_0^\infty f_{w(P^*_{i\to j})}(x)\log(1 - p_{ij}F_w(x))\,dx \tag{6.3}$$

where H_N denotes the hopcount of a shortest path and $F_w(x) = \Pr[w(i \to j) \leq x]$ is the link weight distribution.

Proof. See Appendix A.4. ∎

Corollary 18 establishes a relation between the probability that the hopcount of the shortest path equals 1 in terms of the distribution of the link weights and of the weight of the shortest path.

When multiplying all the link weights by a factor $\frac{1}{b}$ where $b > 0$, relation (6.1) remains unchanged. For, since $f_{\frac{w}{b}}(x) = bf_w(bx)$, we have

$$\int_0^\infty f_{\frac{w}{b}}(x)\frac{\Pr\left[\frac{w(P^*_{i\to j})}{b} > x\right]}{\Pr\left[\frac{w(i\to j)}{b} > x\right] + A}dx$$
$$= \int_0^\infty f_w(bx)\frac{\Pr\left[w(P^*_{i\to j}) > bx\right]}{\Pr[w(i \to j) > bx] + A}d(bx)$$

and substitution of $u = bx$ leads to (6.1). This fact is, of course, natural because the shortest path does not change in structure and in the number of hops when all links are scaled or expressed in a different unit.

6.1.1   Example

If link weights are exponentially distributed,

$$\Pr[w(i \to j) > x] = \exp(-\lambda x)$$

where $E[w] = \frac{1}{\lambda}$ and $\Pr[l(i \to j) \text{ exists}] = p_{ij}$ is the link density, then (6.3) gives with $W_N = w(P^*_{i \to j})$

$$\Pr[H_N = 1] = -\int_0^\infty f_{W_N}(x) \log\left(1 - p_{ij} + p_{ij}e^{-\lambda x}\right) dx \tag{6.4}$$

Applied to the complete graph K_N where $p_{ij} = 1$ leads to

$$\Pr[H_N = 1] = \lambda \int_0^\infty x f_{W_N}(x)\, dx = \lambda E[W_N]$$

Invoking [67, Chapter 15]

$$E[W_N] = \frac{1}{\lambda(N-1)} \sum_{n=1}^{N-1} \frac{1}{n}$$

we find that

$$\Pr[P^*_{i \to j} = l(i \to j)] = \frac{E[W_N]}{E[w]} = \frac{1}{N-1} \sum_{n=1}^{N-1} \frac{1}{n} \tag{6.5}$$

 Alternatively, the probability density function of hopcount of the shortest path in K_N with exponential link weights is [67, Chapter 15]

$$\Pr[H_N = k] = \frac{N}{N-1} \frac{(-1)^{N-(k+1)} S_N^{(k+1)}}{N!}$$

where $S_N^{(k)}$ is the Stirling number of the first kind [1]. The second Stirling number of the first kind can be explicitly written as

$$S_N^{(2)} = (-1)^N (N-1)! \sum_{n=1}^{N-1} \frac{1}{n}$$

we obtain

$$\Pr[P^*_{i \to j} = l(i \to j)] = \Pr[H_N = 1] = \frac{1}{N-1} \sum_{n=1}^{N-1} \frac{1}{n}$$

which is, indeed, the same as (6.5).

6.1.2   The number of observable links in a network

The number of observable links, or the number of links in $G_{\cup spt}$, denoted by L_o, in any network G is, by definition,

$$L_o = \sum_{(i \to j) \in \mathcal{L}} 1_{\left\{ l(i \to j) = P^*_{i \to j} | l(i \to j) \text{ exists} \right\}} \leq L \tag{6.6}$$

where L is the number of links in the substrate network G. If all links in G have equal probability to exist $(p_{ij} = p)$, by taking the expectation in (6.6), the average number of observable links is

$$E[L_o] = \sum_{(i \to j) \in \mathcal{L}} \frac{\Pr[P^*_{i \to j} = l\,(i \to j)]}{p}$$
$$= \binom{N}{2} \Pr[P^*_{i \to j} = l(i \to j)] = \binom{N}{2} \Pr[H_N = 1] \tag{6.7}$$

Hence, the probability of link observability, $\Pr[P^*_{i \to j} = l(i \to j)]$, is equal to the average number of links in $G_{\cup spt}$ divided by the total number of node pairs $\binom{N}{2}$ in the substrate network G.

Clearly, since $G_{\cup spt}$ is connecting all nodes and the number of links in a tree – which is the minimum number of links to connect all nodes – is $N - 1$, we have

$$N - 1 \leq L_o \leq L \tag{6.8}$$

The total number of links in a square lattice with N nodes is $L = 2(N - \sqrt{N})$. Applying the bounds (6.8) for L_o to a square lattice,

$$N - 1 \leq L_o \leq 2N - 2\sqrt{N} \tag{6.9}$$

which shows that $L_o = cN$ is linear to first order in N. An estimate of c is given in Section 6.3.1.

For exponential link weights, the average number of observable links in the complete graph K_N, or equivalently in large Erdös-Rényi random graphs, equals

$$E[L_o] = \frac{N}{2} \sum_{n=1}^{N-1} \frac{1}{n} \simeq \frac{N}{2} (\gamma + \ln N) \tag{6.10}$$

which follows from (6.5), (6.7) and where $\gamma = 0.57721...$ is the Euler constant. For a polynomial link weight distribution (2.1), we have approximately [68] to highest order in N and for α around 1 only, that

$$\Pr[H_N = k] \simeq \frac{1}{N} \frac{(\ln \alpha N)^{\alpha k}}{\Gamma(\alpha k + 1)}$$

and, hence, for α around 1 and large N,

$$E\left[L_o\left(\alpha\right)\right] \simeq \frac{N}{2}\frac{(\ln \alpha N)^\alpha}{\Gamma(\alpha + 1)}$$

If $\alpha \to \infty$, all link weights are the same and equal to 1, $L_o = L$, which shows that equality in (6.8) can occur. For that extreme case, (6.7) tells us that any link is a shortest path between its end nodes.

6.1.3   The degree distribution and beyond

Theorem 19 *The degree distribution of a node in the overlay $G_{\cup spt}$ is equal to its degree distribution when that node is the root of a SPT.*

Proof. We construct $G_{\cup spt}$ in two steps. First, the SPT rooted at a particular node, say node 1, is calculated. Second, the SPT rooted at each other node is computed. Since $G_{\cup spt}$ is the union of the SPTs rooted at all the nodes, the union of shortest paths obtained in these two steps is just $G_{\cup spt}$. All the links in $G_{\cup spt}$ connected to node 1 are found in the first step. For, assume that the link $1 \to j$ is only found in the second step. Since $1 \to j \in G_{\cup spt}$, it must be the shortest path between 1 and j. Thus, it must belong to the SPT rooted at 1 in the first step. Hence, the degree of a node in $G_{\cup spt}$ is equal to its degree when it is the root of a SPT.   ■

A direct consequence of this proof is

Corollary 20 *A link $l\left(i \to j\right) \in G_{\cup spt}$ if and only if $l\left(i \to j\right)$ is a first hop link in the SPT rooted at node $i \in G_{\cup spt}$.*

Theorem 21 *The degree distribution in the overlay $G_{\cup spt}$ on top of the complete graph K_N equipped with exponentially distributed link weights is*

$$\Pr[D_{G_{\cup spt}} = k] = \frac{(-1)^{N-1-k}S_{N-1}^{(k)}}{(N-1)!} \tag{6.11}$$

Proof. In [67, Section 16.6.3], it is shown that the degree distribution of the root in the SPT in the complete graph with exponential link weights is given by the right hand side of (6.11). Application of Theorem 19 proves this Theorem 21.   ■

Recall that a regular link weight distribution is linear around zero and the uniform and exponential distributions belong to the regular link weight distribution according to Section 2.2.1.

Corollary 22 *For large N, the degree distribution in the overlay $G_{\cup spt}$ on top of the Erdös-Rényi random graph $G_p(N)$ with link density p above the disconnectivity threshold p_c and equipped with i.i.d. regular link weights is*

$$\Pr[D_{G_{\cup spt}} = k] \sim \frac{(-1)^{N-1-k}S_{N-1}^{(k)}}{(N-1)!} \tag{6.12}$$

Proof. In [67, Chapter 16], it is shown that the SPT in the complete graph K_N with exponential link weights is precisely a uniform recursive tree[2] URT for any N. In [92], a URT is shown to be asymptotically the SPT in the Erdös-Rényi random graph $G_p(N)$ (see e.g. [13]) with *any* link density p above the disconnectivity threshold $p_c \sim \frac{\log N}{N}$ and with exponential links weights. The exponential distribution is a regular distribution and the shortest path is mainly determined by small link weights in the substrate graph. For sufficiently large and dense, connected graphs, and in view of the i.i.d. assumption, there are enough small link weights very near zero. Hence, under these assumptions, any regular link weight distribution will lead asymptotically to the same SPT. In conclusion, with regular link weights and large N, we have, structurally, that $SPT_{K_N} \simeq SPT_{G_p(N)} \simeq URT$. Therefore, the degree distribution of $G_{\cup spt}$ in K_N or $G_p(N)$ with regular link weights, is asymptotically equal to the degree distribution of the root in the URT [67], which is equal to (6.11). ■

Conjecture 23 *For large N, the overlay $G_{\cup spt}$ on top of a connected Erdös-Rényi random graph $G_p(N)$ with link density $p \in (p_c, 1]$ and equipped with i.i.d. regular link weights is a **connected** Erdös-Rényi random graph $G_{p_c}(N)$ where p_c the disconnectivity threshold.*

    Arguments: First, relation (6.5) states that each link in the underlying complete graph K_N has a probability to appear in the overlay $G_{\cup spt}$ equal to $\frac{1}{N-1} \sum_{n=1}^{N-1} \frac{1}{n} \sim p_c$, for large N.

    Second, for large N and $p = \frac{\log N}{N}$, the binomial degree distribution of the Erdös-Rényi random graph $G_p(N)$ tends to a Poisson distribution with mean $\log N$. Hence, for large N,

$$\Pr\left[D_{G_{\frac{\log N}{N}}} = k\right] = \binom{N-1}{k} p^k (1-p)^{N-1-k} \Bigg|_{p=\frac{\log N}{N}}$$

$$\sim \frac{(\log N)^k}{Nk!}$$

In addition, in [67, Section 16.3.1], it is shown that also (6.11) tends to a same Poisson distribution,

$$\Pr[D_{G_{\cup spt}} = k] = \frac{(-1)^{N-1-k} S_{N-1}^{(k)}}{(N-1)!} \sim \frac{(\log N)^k}{Nk!}$$

In summary, for large N, the URT is also asymptotically the shortest path tree in a connected Erdös-Rényi random graph $G_p(N)$ with link density $p \in (p_c, 1]$. Thus, since

---

[2]A URT of size N is a random tree rooted at some source node and where at each stage a new node is attached uniformly to one of the existing nodes until the total number of nodes is equal to N.

$G_{\cup spt}$ is surely connected[3], for large N, each link in the substrate $G_p(N)$ has a same probability of appearing in $G_{\cup spt}$ equal to $p \sim p_c$ and the degree distribution of $G_{\cup spt}$ equals that of $G_{p_c}(N)$.

In contrast to the Erdös-Rényi random graph $G_p(N)$ where all links *are* independent for any N, the links in $G_{\cup spt}$ on top of the complete graph K_N are not independent because each of them is a shortest path link, a fact that correlates all these links. It still remains to prove that links in the overlay $G_{\cup spt}$ are asymptotically independent. Lemma's 25, 26, and 27 on the uncorrelation of links in $G_{\cup spt}$ are presented in Appendix A.5 as partial arguments. If the asymptotic independence of links can be proved (which would turn the conjecture into a theorem), then, for large N, the three properties (a link density p_c, a Poissonean degree distribution and asymptotic independence of the links) together with the connectedness of $G_{\cup spt}$ will demonstrate that $G_{\cup spt}$ is a connected Erdös-Rényi random graph with $p = p_c$.   □

Simulations in Section 6.4 further illustrate this Conjecture. We expect that Conjecture 23 may hold for a broader class than Erdös-Rényi random graphs: namely all substrate topologies that are *homogeneous* (i.e. the SPT rooted at any node has a same structure or any node perceives, views the network in a same way) and *dense* (i.e. "enough" link). Conjecture 23 explains why the role of the simple Erdös-Rényi random graph $G_p(N)$ is more important in overlay networks, such as e.g. peer-to-peer networks (see also Figures 6.14, 6.15), than in substrate topologies, where only a few complex networks belong to the class of Erdös-Rényi random graphs. Finally, the asymptotic results in this section motivate why a confinement to the complete graph (in later sections) is much less restrictive than it appears at first glance.

6.2   Simulation scenarios

Both the underlying topology and the link weight structure, as introduced in Chapter 2, are key determinants of the overlay networks $G_{\cup spt}$. Three classes of topologies are considered: the complete graphs K_N, lattices and power law graphs. Power law graphs are generated according to both the Havel-Hakimi algorithm and the Barabási-Albert model. The Havel-Hakimi graphs are more attractive because (a) different from the BA power law graphs where $\tau_{BA} = 3$, the Havel-Hakimi graphs can be built according to power law degree distributions with various exponent τ. (b) they show already power law behavior for small N.

For each simulation, 10^4 iterations are carried out. Within each iteration, the specified underlying topology is generated randomly and the polynomial link weights with parameter α or the 2-dimensional uniform link weights with correlation $\rho \in [-1, 1]$ are assigned independently to each link in the graph. The $G_{\cup spt}$ is found by calculating the

---

[3]By definition, $\Pr[G_{p_c}$ is connected$] = \frac{1}{2}$ when $N \to \infty$. Hence, roughly half of the Erdös-Rényi random graphs $G_p(N)$ are connected if $p \sim p_c$ for large N.

shortest paths between all pairs of nodes. For one dimensional link weight, the shortest path can be calculated by Dijkstra's algorithm [30] or the high precision Dijkstra algorithm [70] when α is small, i.e. link weights are close to 0 but they differ significantly with each other. For 2-dimensional link weight, the shortest path or optimal path can be found via algorithm SAMCRA [69], as introduced in Section 2.3. Properties of the overlay networks are examined by (a) the number of links, (b) the degree distribution, and (c) the spectrum.

6.3 Properties of $G_{\cup spt}$ with $\alpha = 1$

For $\alpha = 1$ in (2.1), we obtain the uniform distribution on $[0, 1]$. In both the simulation and the analysis, the average number of links $E[L_o]$ and the degree distribution of $G_{\cup spt}$ is examined. Exact results exist for the complete graph K_N.

6.3.1 The average number of links
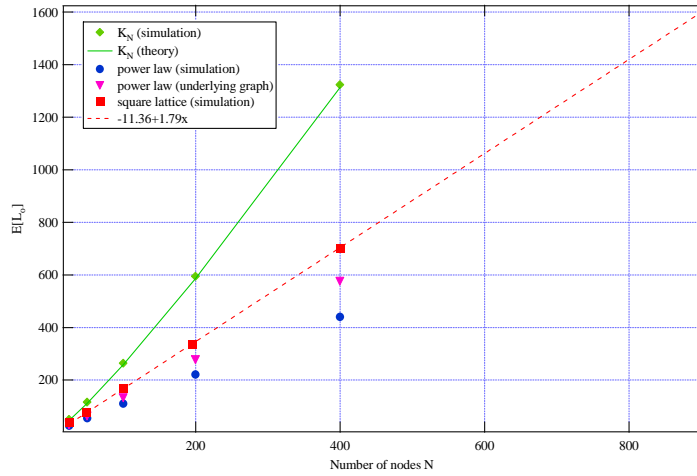
Figure 6.2: Average number of links in the $G_{\cup spt}$ on top of complete graphs, square lattices and Havel-Hakimi power law graphs.

Figure 6.2 shows that the simulation and the theory (6.5) of the average number of links $E[L_o]$ in $G_{\cup spt}$ of the complete graph K_N nicely match.

Simulations in Figure 6.2 also show that, in square lattices, the average number of the observed links via $G_{\cup spt}$ is linear with the number of nodes N. Hence, with (6.7)

$$\Pr[P^*_{i \to j} = l\,(i \to j)] \cdot L = cN \tag{6.13}$$
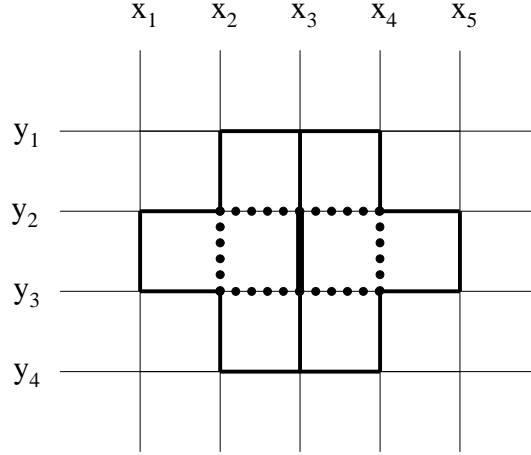
Figure 6.3: Link Observability in a square lattice.

where fitting the simulations yields $c = 1.79$, while (6.9) implies that $1 \leq c \leq 2$.

We will further determine the constant c in (6.13). As shown in Figure 6.3, the link $l((x_3, y_2), (x_3, y_3))$ can be observed via $G_{\cup spt}$, if it is the shortest path between node (x_3, y_2) and node (x_3, y_3). The link l can be dominated by a shorter path such as $P_{h=3}^1 = P_{(x_3,y_2) \to (x_2,y_2) \to (x_2,y_3) \to (x_3,y_3)}$ or $P_{h=3}^2 = P_{(x_3,y_2) \to (x_4,y_2) \to (x_4,y_3) \to (x_3,y_3)}$ which are the only three hops paths between (x_3, y_2) and (x_3, y_3). Furthermore, each link in these two paths can again be dominated by a three hops path, which is shown in bold line in Figure 6.3. Recursively, these links can also be dominated further. Each level of domination results in a longer hopcount of the shortest path between two adjacent nodes. If we define $w(P_{h=3})$ as the weight of a three hops path, then $w(P_{h=3})$ is the sum of three independent uniform random variables and its *pdf* [67] is

$$f_{w(P_{h=3})}(x) = \sum_{j=0}^{3} \binom{3}{j} \cdot (-1)^j \cdot \frac{(x-j)^2}{2} 1_{(x-j) \geq 0}$$

and

$$\Pr[w(P_{h=3}) \leq x] = \int_0^x f_{w(P_{h=3})}(y) dy = \frac{x^3}{6}$$

Since the two 3 hops paths are independent, the probability that one of these two three hops paths is smaller than x is

$$\Pr[\min_{1 \leq k \leq 2} w(P_{h=3}^k) \leq x] = 1 - (1 - \Pr[w(P_{h=3}) \leq x])^2$$

$$= \frac{x^3}{3} - \frac{x^6}{36}$$

For any link, the probability that there exists a three hops path shorter than this direct link $w(l) \in [0, 1]$ is

$$\Pr[\min_{1 \leq k \leq 2} w\left(P_{h=3}^k\right) \leq w(l)]$$

$$= \int_0^1 \Pr[\min_{1 \leq k \leq 2} w\left(P_{h=3}^k\right) \leq x]dx = 0.08$$

If there exists a three hops path shorter than the direct link, the direct link is definitely not observed. However, the hopcount of the shortest path between these two adjacent nodes can be longer than 3, since links in the three hops path can be dominated on their turn. When both the three hops paths are longer than the direct link, the direct link is not necessary the shortest, because paths with hopcount larger than 3 can be even shorter, which is, however, very unlikely to happen for uniform i.i.d. link weights. Hence, the probability that a link can be observed can be approximated by the upper bound:

$$\lim_{N \to \infty} \Pr[P_{i \to j}^* = l(i \to j)] \leq 1 - \Pr[\min_{1 \leq k \leq 2} w\left(P_{h=3}^k\right) \leq w(l)] = 0.92 \qquad (6.14)$$

where $N \to \infty$ means that each node has four neighbors and links at the border are not taken into account. Combining (6.14) and (6.13) results in

$$\lim_{N \to \infty} \Pr[P_{i \to j}^* = l(i \to j)] = \lim_{N \to \infty} \frac{c \cdot N}{L}$$

$$= \lim_{N \to \infty} \frac{c}{2} \cdot \frac{\sqrt{N}}{\sqrt{N} - 1} = \frac{c}{2}$$

and, in $c = 1.84$ which is close to the simulation result $c = 1.79$. In addition, (6.14) also shows that $L_o \simeq L$. In other words, the observable square lattice is very near to the substrate, in contrast to K_N as observed from (6.10).

In Figure 6.2, the underlying Havel-Hakimi power law graph with $\tau = 2.4$ is shown to be sparse and $E[L_o]$ of the corresponding $G_{\cup spt}$ approaches $N - 1$. It is natural that the $G_{\cup spt}$ is close to a tree, because the sparse underlying power law graph is already tree-like.

6.3.2 The degree distribution

The simulation result of the degree distribution of $G_{\cup spt}$ in K_N is shown in Figure 6.4 together with the result calculated by Theorem 21. As N increases, the degree distribution of $G_{\cup spt}$ tends to that of the URT root. These simulations support that $SPT_{K_N} \simeq URT$ with any regular link weights as explained in Section 6.1.3.

The degree distribution of $G_{\cup spt}$ in a square lattice is shown in Figure 6.5. The good approximation (6.14) encourages us to further simplify the analysis for the degree
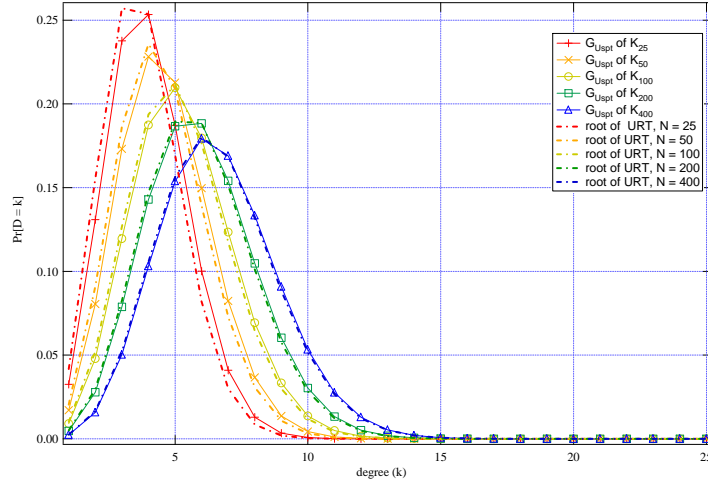
Figure 6.4: Degree distribution of $G_{\cup spt}$ in K_N and degree distribution of the root of the corresponding URT.

distribution in a finite square lattice. There are three kinds of nodes in a lattice: $n_2 = 4$ nodes with degree $D = 2$; $n_3 = 4 * (\sqrt{N} - 2)$ nodes with degree $D = 3$; $n_4 = N - 4 * (\sqrt{N} - 1)$ nodes with degree $D = 4$. Two kinds of links exist: (a) links at the border, that have a probability $p_2 \approx 1 - \int_0^1 \Pr[P_{h=3} \leq x]dx = 0.96$ to be observed via $G_{\cup spt}$ and (b) central links, that have a probability $p_1 \approx 0.92$ to be the shortest path between its end nodes. We assume that the observation of one link via $G_{\cup spt}$ will not influence the probability of other links being observed, since the hopcount of the shortest path in a square lattice is usually large and a link with small link weight is unlikely to attract the other shortest paths to pass through it. Then the following can be obtained:

$$
\begin{aligned}
&\Pr[D = 1] = 1 - \Pr[d = 2] - \Pr[d = 3] - \Pr[d = 4] \\
&\Pr[D = 2] = \tfrac{1}{N}\binom{4}{2} \cdot n_4 \cdot p_1^2(1 - p_1)^2 + \\
&\tfrac{1}{N}\left(n_3 \cdot (p_2^2(1 - p_1) + 2p_1p_2(1 - p_2)) + n_2 \cdot p_2^2\right) \\
&\Pr[D = 3] = \tfrac{1}{N}\left(n_3 \cdot p_1 p_2^2 + \binom{4}{1} \cdot n_4 \cdot p_1^3(1 - p_1)\right) \\
&\Pr[D = 4] = \tfrac{1}{N} \cdot n_4 \cdot p_1^4
\end{aligned}
\tag{6.15}
$$

We explain $\Pr[D = 3]$. A node with degree 3 in the $G_{\cup spt}$ is either a node with degree 3 in the square lattice with all three links being observed, or a node with degree 4 in the square lattice with three of its four links being observed. With the set (6.15), we calculate the degree distribution of $G_{\cup spt}$ in a square lattice with $N = 400$ nodes and compare those values with the simulation result in Table 6.1, which, again, shows a good agreement.
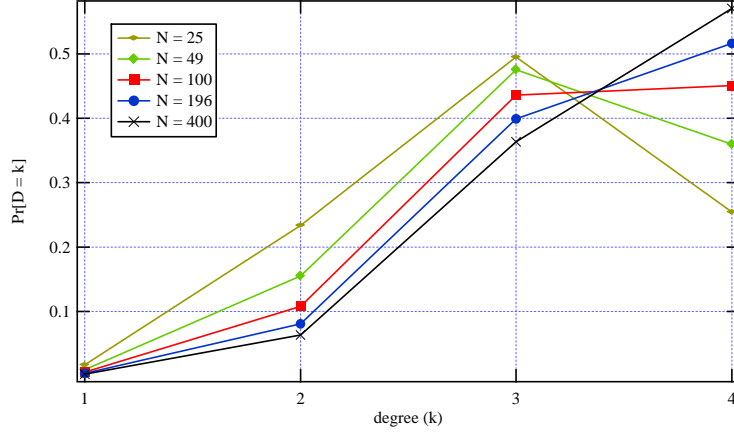
Figure 6.5: Degree distribution of G_{Uspt} in square lattices.

Table 6.1: Degree distribution of $G_{\cup spt}$ in a square lattice with $N = 400$.

	$\Pr[D=1]$	$\Pr[D=2]$	$\Pr[D=3]$	$\Pr[D=4]$
Simulation	0.003	0.063	0.36	0.57
theory	0.004	0.062	0.354	0.580

The degree distribution of $G_{\cup spt}$ in Havel-Hakimi power law graph substrates with $\tau = 2.4$ is shown in Figure 6.6. The degree distribution of a power law graph is $\Pr[D=i] = ci^{-\tau}$ (by definition) which is shown in bold line in the figure for $N = 100$. Nodes with degree 1 in the underlying graph must remain the same in the $G_{\cup spt}$ in order for $G_{\cup spt}$ to be connected. A node with higher degree in the underlying graph may have only one link in $G_{\cup spt}$, thus, degree 1 in $G_{\cup spt}$. Hence, as shown in Figure 6.6, compared to the degree distribution of the underlying topology, $\Pr[D=1]$ in the $G_{\cup spt}$ increases while $\Pr[D=i]$ for $i > 1$ decreases. However, such difference is not substantial. The overlay network exhibits a degree distribution visually similar to the underlying graph, because as shown in Section 6.3.1, the underlying topology is sparse and already tree-like. A similar conclusion is reached for the Barabási-Albert preferential attachment model as illustrated in Figure 6.7, where we have chosen a larger number of $m = 4$ links that is attached at each time which is equivalently to a large link density. This demonstrates that the overlay on top of a power law graph is very close to the substrate in terms of degree distribution, even for $m = 4$. These observations are consistent with results in [59], where even a subgraph of $G_{\cup spt}$ has similar degree distribution as that of the underlying power law graph with $w = 1$ or $\alpha \to \infty$ [3].
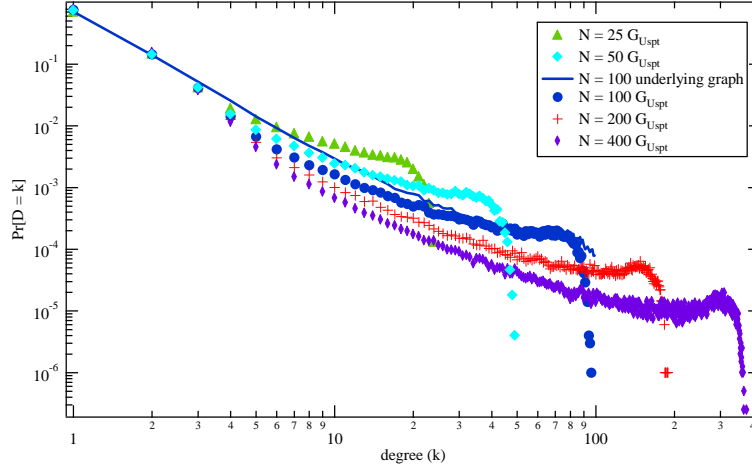
Figure 6.6: Degree distribution of $G_{\cup spt}$ in Havel-Hakimi power law graphs with $\tau = 2.4$.

6.4 Properties of $G_{\cup spt}$ with varying α

When $\alpha \to 0$, all links will be close to 0, but, relatively, they differ significantly with each other. If $\alpha \to \infty$, it follows from (2.1) that $w = 1$ almost surely for all links. Hence, the overlay $G_{\cup spt}$ is the same as the underlying topology, since the link weight structure does not differentiate between links. Hence, the $\alpha \to \infty$ regime is not further considered. Van Mieghem and Magdalena [70] have found that, by tuning the extreme value index α of the polynomial link weight distribution, a phase transition occurs around a critical extreme value index α_c. The critical extreme value index α_c is defined as $F_T(\alpha_c) = \frac{1}{2}$ where $F_T(\alpha) = \Pr\left[G_{\cup spt(\alpha)} = MST\right]$. When $\alpha > \alpha_c$, the overlay $G_{\cup spt(\alpha)}$ contains more than $N - 1$ links whereas for $\alpha < \alpha_c$, all transport traverses a critical backbone consisting of $N - 1$ links, which is the MST, as follows from Theorem 16. Here, we extend the analysis of [70] in two ways: (a) we include, besides the complete graphs K_N and square lattices, also cubic lattices and Havel-Hakimi power law graphs; (b) by a spectral analysis, we further obtain insights in the structure of the overlay $G_{\cup spt(\alpha)}$.

6.4.1 Phase transition in the $G_{\cup spt(\alpha)}$ structure

Instead of the number of links in $G_{\cup spt(\alpha)}$, when α is small, we study the probability that the overlay $G_{\cup spt(\alpha)}$ is a tree. As shown in Figure 6.8, normalized by α_c, the same phase transition curve is observed for all these three types of topologies. As α increases, the transport is more likely to traverse over more links and the overlay $G_{\cup spt(\alpha)}$ will less likely become a tree. These additional simulations over those reported in [70] strengthen
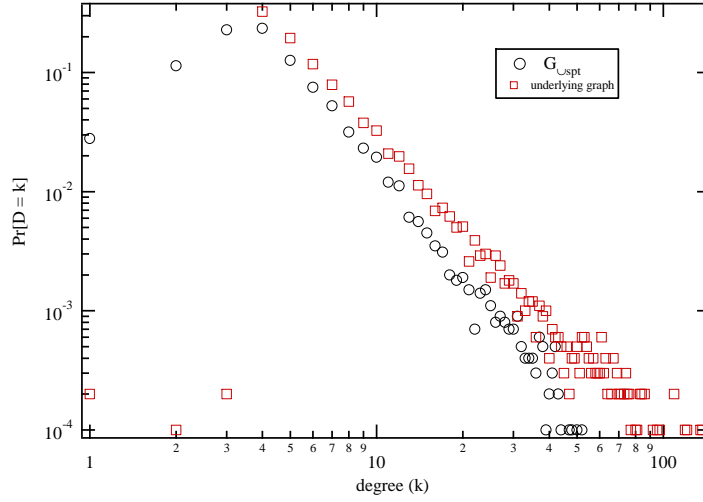
Figure 6.7: Degree distribution in Barabási-Albert power law graphs where $N = 800$ and the initial number to start the preferential growth is $m_0 = m = 4$.

the belief that the curve $F_T(\alpha) \approx 2^{-\left(\frac{\alpha}{\alpha_c}\right)^2}$ is universal for all graphs that are not trees.

For each substrate topology, the critical extreme value α_c is shown in Figure 6.9 as a function of N on a log-log scale. Each curve is fitted with a line and indicates that $\alpha_c \simeq bN^{-\beta}$ where b and β depend on the underlying substrate. In spite of the fact that this phase transition is constructed, the exponent β seems to lie in the interval $[\frac{1}{2}, \frac{2}{3}]$, which agrees surprisingly well with critical exponents observed in nature (see e.g. [11]). As explained in [70], it is computationally difficult to determine α_c for large N. This limits the extent to which a seemingly power law can be observed. In [20], on the other hand, the characteristics – basically hopcount and weight – of the shortest path between an arbitrary source and destination are different above and below a cross-over point $\alpha_{co} = cN^{-\gamma}$, where $\gamma = \frac{1}{3}$ for Erdös-Rényi random graphs. The scaling laws for α_c and α_{co} are, however, different[4], which may point, at first glance, to an anomaly. The inconsistency may be due to either the finite size of networks in our numerical simulations or the approximations in the analysis of [20]. The link density does not influence properties of the shortest paths as discussed in Section 6.1. Thus, we carry out simulations to determine the critical extreme value α_c on Erdös-Rényi random graphs till $N = 800$ nodes. The critical point scales as $\alpha_c = bN^{-\beta}$, where $\beta \simeq 0.61$ is

---

[4]The α_{co} corresponds to the cross-over from the weak disorder regime to the strong disorder limit. In the weak disorder regime, all links contribute to the total weight of a shortest path. In the strong disorder limit, a single link weight dominates the sum of the weights along the paths. All shortest paths belong the the MST in the stronge disorder limit. Thus, the scaling laws for α_c and α_{co} should be the same.
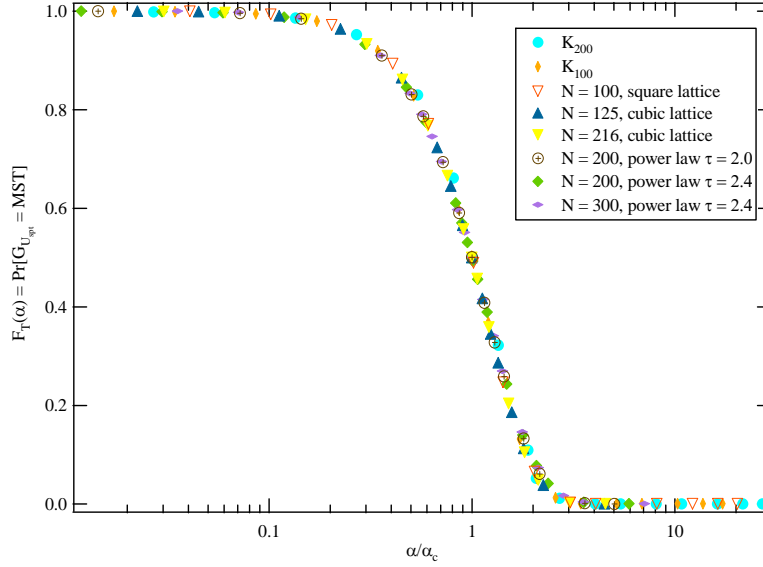
Figure 6.8: The probability distribution $F_T(\alpha)$ as a function of the normalized α/α_c.

still far above $\gamma = \frac{1}{3}$ and close to $\beta \simeq 0.63$ for complete graphs as shown in Figure 6.9.

The higher the α_c-curve in Figure 6.9, the faster $G_{\cup spt(\alpha)}$ tends to a tree if α decreases. The power law graphs seem to possess the highest α_c because they are already tree-like. Although when $\alpha = 1$, the overlay $G_{\cup spt(\alpha)}$ of the complete graph K_N contains more links than that of the square lattice (compare (6.10) with (6.9)), Figure 6.9 illustrates that, since $\alpha_c(K_N) > \alpha_c$ (square lattice), the overlay $G_{\cup spt(\alpha)}$ of K_N will tend faster to a tree if α decreases. We can only give a speculative explanation. When α is very small around the α_c, link weights are small, but they differ significantly from each other. In fact, if $\alpha \to 0$, the ratio $\frac{\sqrt{\mathrm{Var}[w]}}{\mathbb{E}[w]} \sim \frac{1}{\sqrt{\alpha}}$ diverges which means that, in this limit, the link weights possess extremely strong fluctuations. Since the number of possible trees in K_N is much larger than that in the square lattice, it is more probable to find a spanning tree only composed of these extremely small links and that tree is the MST.

6.4.2   The spectrum of the adjacency matrix of $G_{\cup spt(\alpha)}$

The spectrum, the eigenvalues of the adjacency matrix, of $G_{\cup spt(\alpha)}$ in the complete graph K_{50} are displayed in Figure 6.10. For K_{50}, the critical extreme value index is $\alpha_c = 0.09$. Figure 6.8 shows that the onset of the phase transition is somewhere between $3\alpha_c$ and $4\alpha_c$. Figure 6.10 (a) shows that, for $\alpha \in [0.01, 0.3]$, the corresponding $G_{\cup spt(\alpha)}$ have almost the same spectrum, except that, the peaks at ± 1.4 diminish as α increases. The nodes with small degrees are most likely responsible [32] for the delta
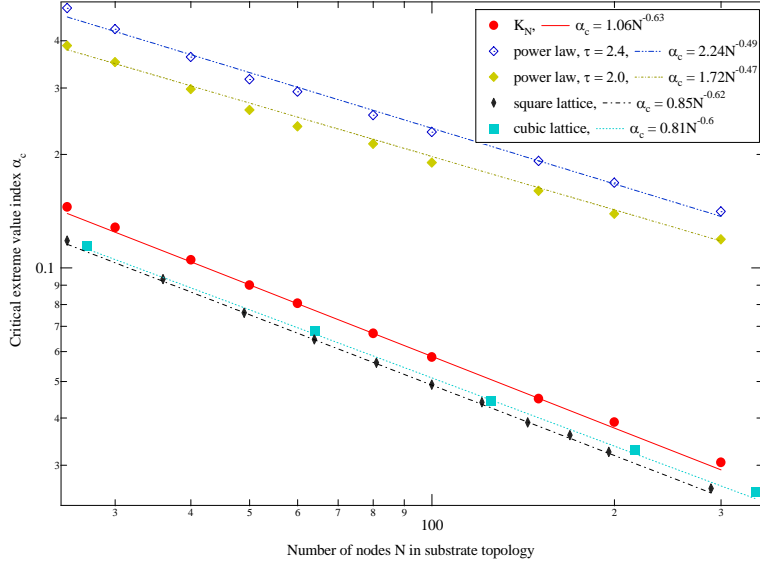
Figure 6.9: The critical extreme value α_c as a function of N.

peak at[5] $\lambda = 0$. For example, the local configurations with two and more dead-end nodes produce eigenvalues $\lambda = 0$, where the dead-end node is a node with degree 1. The corresponding eigenvectors have non-zero components only at the dead-end nodes [49][95]. The spectrum of a tree is symmetric [26], because any tree is a bi-partite graph and any bi-partite graph is symmetric around $\lambda = 0$. In Figure 6.10 (b) when $\alpha = 0.4$, these two peaks at ± 1.4 are smoothed out and the spectrum is not symmetric. It indicates that when α is smaller than the onset value of the phase transition (case (a) in Figure 6.10), the $G_{\cup spt(\alpha)}$ seem to possess similar topological, tree-like structure. Since very few trees can be uniquely specified by their spectrum [90], the spectrum is not well suited to reveal the specifics of the MST. The spectrum of the $G_{\cup spt(\alpha=1)}$ in K_{50} with regular link weights illustrated in Figure 6.10 (c) is close to the spectrum of a random graph according to the Wigner's Semicircle Law [101][67, Appendix B]. This correspondence is an additional illustration of Conjecture 23. When α is large, link weights are ineffective in that $\lim_{\alpha \to \infty} G_{\cup spt(\alpha)} = G_{\text{substrate}}$. The spectrum of $G_{\cup spt(\alpha)}$ in Figure 6.10 (d) is, indeed, close to the spectrum of substrate K_N, that has $N - 1$ eigenvalues at -1 and 1 eigenvalue at $N - 1$. While peaks in the spectrum reflect structure and regularity in the graph, a bulk almost symmetrical around zero, which

---

[5]In general, each time when two rows in the adjacency matrix A are the same, the rank of A decreases with 1, which is equivalent to an increase in the multiplicity of the eigenvalue $\lambda = 0$, since $\det A = \prod_{j=1}^{N} \lambda_j$.
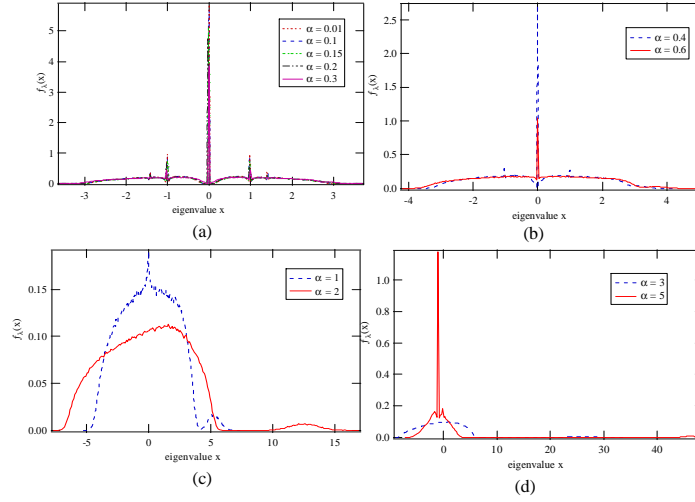
Figure 6.10: The spectrum of the $G_{\cup spt(\alpha)}$ in the complete graph K_{50}: $(a)0.01 \le \alpha \le 0.3$; $(b)\alpha = 0.4, 0.6$; $(c)\alpha = 1, 2$; $(d)\alpha = 3, 5$.

ultimately tends to a semicircle, points to uncorrelated randomness. The latter is a characteristic property of an Erdös-Rényi random graph. Figure 6.10 thus shows, as a function of α, transitions of $G_{\cup spt(\alpha)}$ between two graph types, a tree and the complete graph, with apparent maximum randomness for regular link weights ($\alpha = 1$).

The spectrum of the $G_{\cup spt(\alpha)}$ in a square lattice with 49 nodes are displayed in Figure 6.11. When α is small, as shown in Figure 6.11 (a) and (b), the transition of the spectrum of $G_{\cup spt(\alpha)}$ is similar to that in the complete graph. As studied in Section 6.3.1, on average, 92% of the links in the underlying square lattice can be observed via the overlay $G_{\cup spt}$ when $\alpha = 1$. The spectrum of a square lattice with N nodes [26] comprises the eigenvalues

$$\lambda_{ij} = 2\cos\frac{2\pi}{\sqrt{N}}i + 2\cos\frac{2\pi}{\sqrt{N}}j \qquad i, j \in \{1, \ldots, \sqrt{N}\}$$

which correspond to the peaks in the spectrum of $G_{\cup spt(\alpha \ge 2)}$. Compared to the complete graph, in a square lattice, the overlay $G_{\cup spt}$ approaches the underlying topology at a smaller α-values, $\alpha \ge 2$ for the square lattice while $\alpha \ge 5$ for K_N.

The spectrum of the $G_{\cup spt(\alpha)}$ in a Havel-Hakimi power law graph is almost the same as that of the underlying graph for any α, because the underlying graph is already close to a tree. For the spectrum of a Barabási-Albert power law graph, we refer to [39]. As mentioned before, since the spectrum for most trees is not unique [90], a spectral analysis of tree-like graphs is not the best way to deduce specific properties.
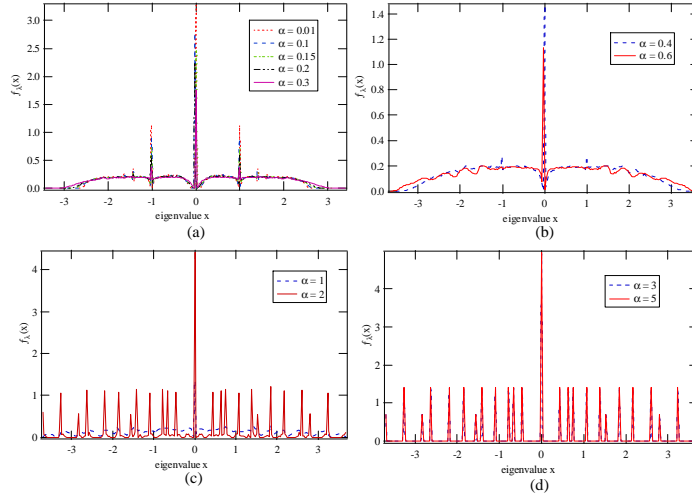
Figure 6.11: The spectrum of the $G_{\cup spt(\alpha)}$ in a square lattice with 49 nodes: $(a)0.01 \leq \alpha \leq 0.3; (b)\alpha = 0.4, 0.6; (c)\alpha = 1, 2; (d)\alpha = 3, 5$.

6.5 Two dimensional link weight tuning

We assign two uniformly distributed link weights with correlation ρ to each link in the underlying graph and the correlation coefficient ρ can be varied within $[-1, 1]$. When $\rho = 1$, the two link weights of each link are the same, which reduces to the one dimensional uniformly distributed link weight analyzed in Section 6.3. When $\rho = -1$, $\overrightarrow{w}(u \rightarrow v) = [w_1(u \rightarrow v), w_2(u \rightarrow v)] = [w_1(u \rightarrow v), 1 - w_1(u \rightarrow v)]$. We recall the definition [69] of the path length function $l(\mathcal{P}) = \max_{1 \leq i \leq 2} \left[\frac{w_i(\mathcal{P})}{L_i} \right]$. We assume the same constraint for these two link weight measures $L_1 = L_2$, which are large such that the shortest path always satisfies the constraints. Hence, the path length function to find the optimal path can be reduced to $l(\mathcal{P}) = \max \binom{w_1(\mathcal{P})}{w_2(\mathcal{P})}$, where $w_i(\mathcal{P}) = \sum_{(u \rightarrow v) \in \mathcal{P}} w_i(u \rightarrow v)$.

The path length of a $h \geq 2$ hop path is

$$l_h(\mathcal{P}) = \max \binom{w_1(\mathcal{P})}{w_2(\mathcal{P})} = \max \binom{w_1(\mathcal{P})}{h - w_1(\mathcal{P})} \geq \frac{h}{2} \geq 1$$

while the path length of a one hop path

$$l_1(\mathcal{P} = u \rightarrow v) = \max \binom{w_1(u \rightarrow v)}{1 - w_1(u \rightarrow v)} < 1 \leq l_h(\mathcal{P})$$

Hence, when $\rho = -1$, the link between the source and destination, if exists, is always the shortest path. All links in the underlying graph will appear in the overlay $G_{\cup spt}$. In

other words, the overlay $G_{\cup spt(\rho=-1)}$ is the same as the substrate, which corresponds to $G_{\cup spt(\alpha \to \infty)}$.
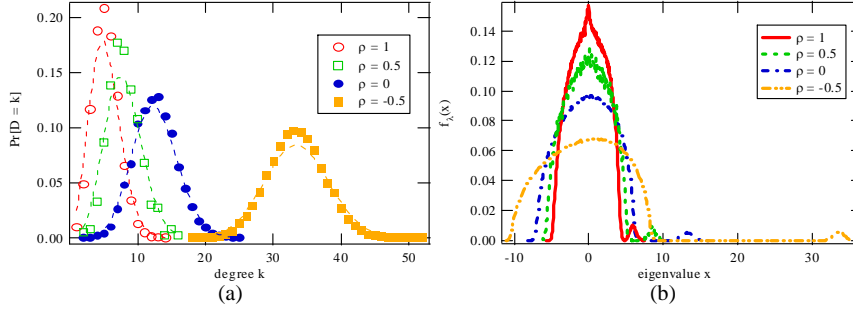
Figure 6.12: Degree distribution and spectrum of the overlay $G_{\cup spt}$ in K_{100} with 2-dimensional correlated uniformly distributed link weights.

When the underlying topology is the complete graph, the degree distribution of the overlay $G_{\cup spt}$ displayed in Figure 6.12(a) is close to the binomial distribution, the degree distribution of an Erdös-Rényi random graph. The dotted line is the degree distribution of the Erdös-Rényi random graph $\Pr[D = k] = \binom{N-1}{k} p^k (1-p)^{N-1-k}$ where p is set as the link density[6] of the corresponding overlay $G_{\cup spt(\rho)}$. As shown in Figure 6.12(b), the spectrum of the overlay $G_{\cup spt(\rho)}$ with different correlation coefficient ρ is close to the spectrum of an Erdös-Rényi random graph according to the Wigner's Semicircle Law. Both the degree distribution and the spectrum indicate that the overlay $G_{\cup spt(\rho)}$ on top of K_{100} with 2-dimension correlated uniform link weights is close to an Erdös-Rényi random graph. The same behavior has been observed when the substrate is not the complete graph but the Erdös-Rényi random graph: the overlay $G_{\cup spt(\rho)}$ is also always close to an Erdös-Rényi random graph.

The link density of the overlay $G_{\cup spt}$, the number of links L_o in the overlay $G_{\cup spt}$ divided by the maximum number of links in a graph $\frac{N(N-1)}{2}$, is plotted in Figure 6.13. When $\rho = -1$, the overlay $G_{\cup spt(\rho=-1)}$ is equal to the substrate $G_p(N)$. The link density of the overlay is then determined by the link density of the underlying topology. According to conjecture 23, when $\rho = 1$, the overlay $G_{\cup spt}$ is a **connected** Erdös-Rényi random graph $G_{p_c}(N)$ with link density p_c, which is independent of the link density of the substrate. The link density of the overlay $G_{\cup spt(\rho)}$ decreases exponentially from the link density p of the substrate to $p_c \sim \frac{\log N}{N}$ as a function of the correlation coefficient ρ.

---

[6]It is the average link density $E[\frac{2L_o}{N(N-1)}]$ of the 10^4 generated overlay $G_{\cup spt}$ in each simulation, where L_o is the number of links in the overlay $G_{\cup spt}$.
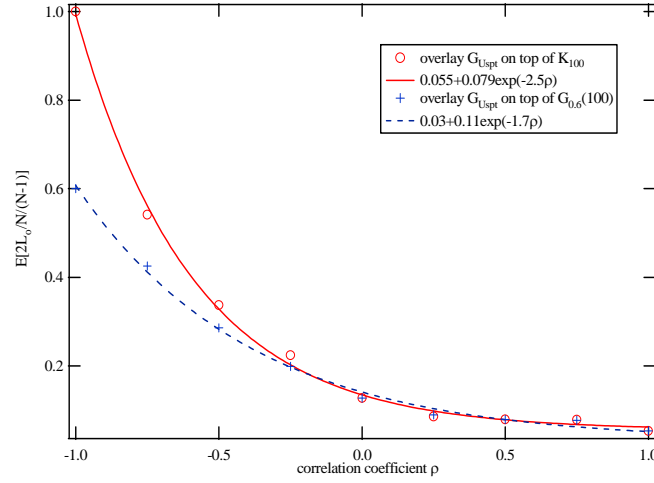
Figure 6.13: Link density of the overlay $G_{\cup spt}$ on K_{100} and $G_{0.6}(100)$ with 2-dimensional correlated uniformly distributed link weights.

6.6 Conclusion

The union of all shortest path trees $G_{\cup spt}$ constitutes the observable part of a network provided traffic flows follow shortest paths. We focus on the one dimensional link weight tuning and we have studied two properties of the $G_{\cup spt}$, the average number of links $E[L_o]$ and the degree distribution, and simulated the spectrum of the adjacency matrix of $G_{\cup spt(\alpha)}$ as a function of the extreme value index of the link weight structure. Different underlying topologies and the link weight structure were treated independently. The minimum spanning tree belongs to the $G_{\cup spt}$. Any link $l\,(i \rightarrow j)$ with link weight $w(i \rightarrow j)$ in the $G_{\cup spt}$ must be the shortest path between i and j and if a link is the shortest path between its end nodes, it must belong to the $G_{\cup spt}$. Most of the theory is based on these two points. Apart from the Theorems and Corollaries presented, the Conjecture 23 and the seemingly universality of $F_T\,(\alpha) \approx 2^{-\left(\frac{\alpha}{\alpha_c}\right)^2}$ in the phase transition appearing in the structure of $G_{\cup spt}$ are considered important new findings. For example, Conjecture 23 which has assumed an i.i.d. link weight structure, claims the appearance of the random graph $G_p(N)$ in many application such as, for example, peer-to-peer networks [7] and ad-hoc networks [50]. The universality of $F_T\,(\alpha)$ in the phase transition points to the possibility to control the network structure or to steer or balance transport by tuning the link weight structure.

The overlay $G_{\cup spt}$ is, actually, the *maximally* measurable part of the substrate topology. For example, the RIPE traceroute measurement configuration[7] only measures the

---

[7]RIPE Test Traffic Measurements, http://www.ripe.net/ripencc/mem-services/ttm/.
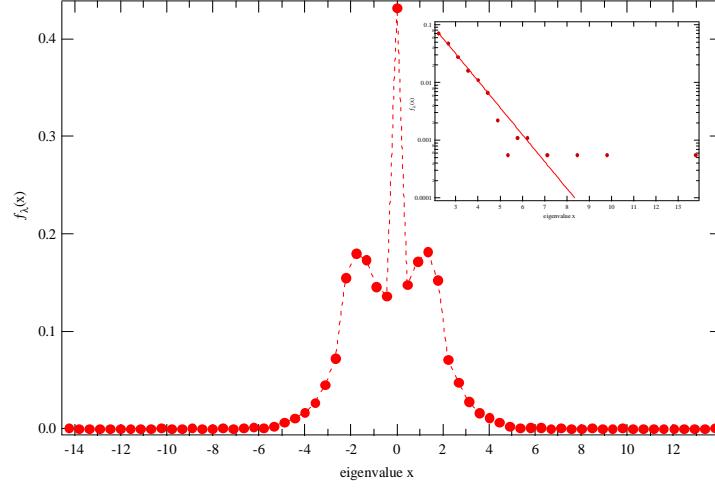
Figure 6.14: The spectrum of G_{RIPE} measured by RIPE. The insert graph is the tail part, fitted with $f_\lambda(x) \simeq 0.8e^{-x}$. The theory of [32] is not applicable to exponentially decaying tails in $f_\lambda(x)$.

union $G_{\cup_m spt}$ of shortest paths between each pair of a small group of $m << N$ nodes, while the number of nodes in the underlying graph N is much larger. Considerable attention has been devoted to the properties of graphs derived from Internet measurements. But how accurate does the measured subgraph reflect the underlying graph [59]?

The spectrum of the topology G_{RIPE} measured by RIPE [53] is shown in Figure 6.14, and is very akin to that measured on Planet lab[8] in Figure 6.15. More details are given in Table 6.2.

Table 6.2: Measurement description.

	RIPE	PlanetLab
date	9-18-2005	11-10-2004
m	70	79
#spts in union	67	76
N	4058	4214
L	6151	6994
$\Pr[D = k]$	$\sim e^{-0.39k}$	$\sim e^{-0.44k}$

Each of the m testboxes acts as a source and sends traffic to other testboxes. After

[8]http://www.planet-lab.org

removing error measurements in the trace-routes, the overlay G_{RIPE} and $G_{\text{PlanetLab}}$ are constructed as the union of (only) #spts trees. Both G_{RIPE} and $G_{\text{PlanetLab}}$ are subgraphs of $G_{\cup spt}$ on top of the underlying Internet topology.
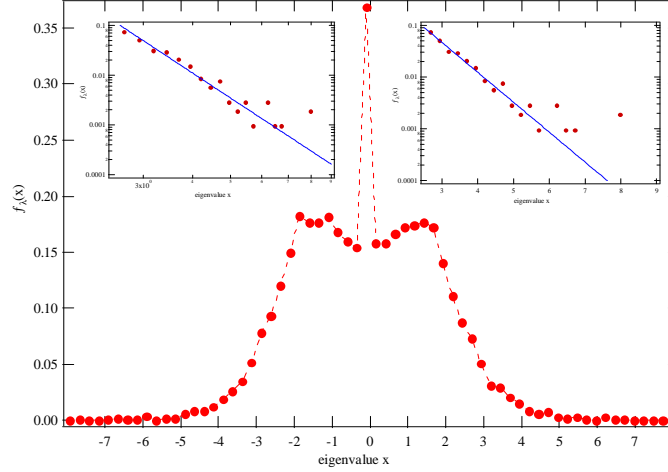
Figure 6.15: The spectrum of the overlay measured on PlanetLab. The deep tail region can be fitted both by an exponential $f_\lambda(x) = 2.5e^{-1.3x}$ and a power law $f_\lambda(x) = 1.5\lambda^{-5.2}$. Applying the conversion $f_\lambda(x) \simeq 2xf_D(x^2)$ from [32] to the power law results in a degree pdf $f_D(x) \sim x^{-3.1}$. The power exponent $\tau = 3.1$ exceeds the commonly accepted $2.2 \leq \tau_{\text{Internet}} \leq 2.5$.

The spectra of G_{RIPE} and $G_{\text{PlanetLab}}$ seem to give support for the Conjecture 23, since the partial overlay $G_{\cup_m spt}$ $(m << N)$ seems approximately close to $G_{\cup spt}$ (see Figure 6.10 (c)) and reveals more features of the overlay than that of the underlying topology, which is overwhelmingly shown in the literature to belong to the class of power law or scale free graphs. The spectrum of scale-free networks exhibit a power law tail in the region of large eigenvalues [39][32]. The spectra of G_{RIPE} and $G_{\text{PlanetLab}}$, however, may possess an exponential tail, most likely because the value of m is still not large enough to cover $G_{\cup spt}$ sufficiently and to observe the power law behavior [54]. The relationship between a partial overlay $G_{\cup_m spt}$, the complete overlay $G_{\cup_N spt}$ and the underlying topology or substrate will be studied in Chapter 7.

Conjecture 23 implies that the observed network will have a low maximum degree and a Poisson degree distribution, contrary to the view, promoted by some papers [59][22][2], claiming that the bias may lead to an observed power law degree distribution irrespective of the degree distribution of the substrate. Most published work on sampling bias focuses on unweighted graphs and the bias originates purely from the sampling methods (such as the union of paths from a small set of sources to a relatively larger set of destinations). Here, we study the "bias" introduced by the link weight

structure of the substrate. The overlay constructed as the union of shortest paths between all node pairs is exactly the same as the substrate if the substrate is unweighted ($\alpha \to \infty$ case). In weighted graphs, links with high weight rarely appear in the observed network.

Finally, the differences between one and two dimensional link weight tuning in dense substrate, i.e. the Erdös-Rényi random graph $G_p(N)$, is summarized in table 6.3.

Table 6.3: The structure of the overlay $G_{\cup spt}$ on top of $G_p(N)$ via one and two dimensional link weight tuning.

one dimensional $\alpha \in [0, \infty)$	two dimensional $\rho \in [-1, 1]$
$G_{\cup spt(\alpha < \alpha_c)} = MST$	
$G_{\cup spt(\alpha = 1)} \simeq G_{p_c}(N)$	$G_{\cup spt(\rho = 1)} \simeq G_{p_c}(N)$
$G_{\cup spt(1 < \alpha < \infty)} \neq ER$	$G_{\cup spt(-1 < \rho < 1)} \simeq ER$
$G_{\cup spt(\alpha \to \infty)} = G_p(N)$	$G_{\cup spt(\rho = -1)} = G_p(N)$

With one dimensional link weight tuning, the overlay $G_{\cup spt(\alpha)}$ varies from the sparsest structure, the MST, to the densest, the substrate $G_p(N)$. Via two dimensional link weight tuning, the overlay $G_{\cup spt(\rho)}$ is always close to the Erdös-Rényi random graph. The link density of $G_{\cup spt(\rho)}$ varies within $[\frac{\log N}{N}, p]$, which is smaller than the link density range of $G_{\cup spt(\alpha)}$: $[\frac{2}{N}, p]$.

Chapter 7

Sampling networks by $G_{\cup_m spt}$

Topologies of complex networks ranging from biological networks, artificial networks to social networks have been accumulated by active investigation in recent years. However, many surveyed networks to date are, in fact, subnets of the actual network, which we call the *"underlying network"*. For example, only a subset of the molecular entities in a cell have been sampled in protein interaction, gene regulation and metabolic networks. The topology of the Internet is inferred by aggregating paths or traceroutes [86], which reveals only a part of the whole Internet. Thus, these identified networks are sampled networks of the underlying networks according to different mapping or sampling methods.

In this work, we study the bias phenomenon of a sampling method that originated from the Internet. The topology of the Internet has typically been measured by the union of sampling traceroutes, which are approximately shortest paths. Mainly two sampling methods exist: (a) The topology is built from the union of traceroutes from a small set of sources to a larger set of destinations as in the CAIDA skitter project[1]. The sampled map can be modeled as the union of the spanning trees rooted at the sources. (b) The traceroute measurements are carried out between each pair of a set $\mathcal{M}$ of m testboxes or testbeds. The sampled network, denoted as $G_{\cup_m spt}$, is the union of m shortest paths trees $SPTs$, where each SPT is the union of shortest paths from the root $\in \mathcal{M}$ to the other $m-1$ testboxes $\in \mathcal{M}$. Equivalently, $G_{\cup_m spt}$ is the union of shortest paths between each node pair in the set $\mathcal{M}$ of m testboxes. The RIPE NCC and the PlanetLab measurement architectures, as discussed in Section 6.6, are examples of this type. The methodology in (a) has been argued and even proved to introduce such intrinsic biases that statistical properties of the sampled topology may sharply differ from that of the underlying graph (see e.g. [59][2][22]). While most related works on Internet exploration have been devoted to the sampling method (a), we investigate the other sampling method (b). Although the number of destinations may be limited

---

[1]http://www.caida.org

to the number m of measurement boxes, the spurious effects in (a), where nodes and links closer to the sources are more likely to be sampled than those surrounding the destinations, can be reduced.

With statistical and graph theory methodologies, we investigate this sampling method (m shortest path trees) on a wide class of networks: the weighted Erdös-Rényi random graphs, which represent dense and homogeneous networks, and the unweighted real-world complex networks which are generally sparse and inhomogeneous graphs. Various underlying networks are investigated, because network sampling is a generic problem residing in various disciplines and the actual underlying network topology is mostly uncertain. Here, we focus on the sampling bias (the incompleteness of the network mapping) introduced purely by the sampling method. Technical limitations in the topology measurements may also introduce significant sampling bias. For example, the network measured by traceroute represents the interconnections of IP addresses. The bias in mapping the router level Internet topology depends highly on the alias resolution technique, which maps IP addresses to the corresponding routers [83]. Such specific technical concerns, which vary in the measuring of different complex networks, are not explored in this chapter.

The sampled network $G_{\cup_m spt}$ depends on the set $\mathcal{M}$ of m boxes as well as the underlying network. In this work, we focus on the effect of the testboxes, in particular, 1) the subgraph $G_{\mathcal{M}}$ of the underlying network, consisting of the set $\mathcal{M}$ and the direct links between nodes of set $\mathcal{M}$, and 2) the relative size m/N of set $\mathcal{M}$, where N is the size of the underlying network. With a given set of testboxes, the sampling bias varies for different networks. The kind of networks with small sampling bias will also be briefly mentioned.

The main contributions of this study can be summarized as follows.

1. Introduction of a general framework for network sampling on both weighted and unweighted complex networks.

2. Establishment of the correlation between the interconnections of set $\mathcal{M}$, i.e. the subgraph $G_{\mathcal{M}}$, and the sampled network $G_{\cup_m spt}$.

3. Illustration of the detection/measuring effort (the relative size m/N of set $\mathcal{M}$) to obtain an increasingly accurate view of a given network.

4. Characterization of networks bearing small sampling bias when m/N is small and the corresponding proposal of testbox placement for good network topology measurements.
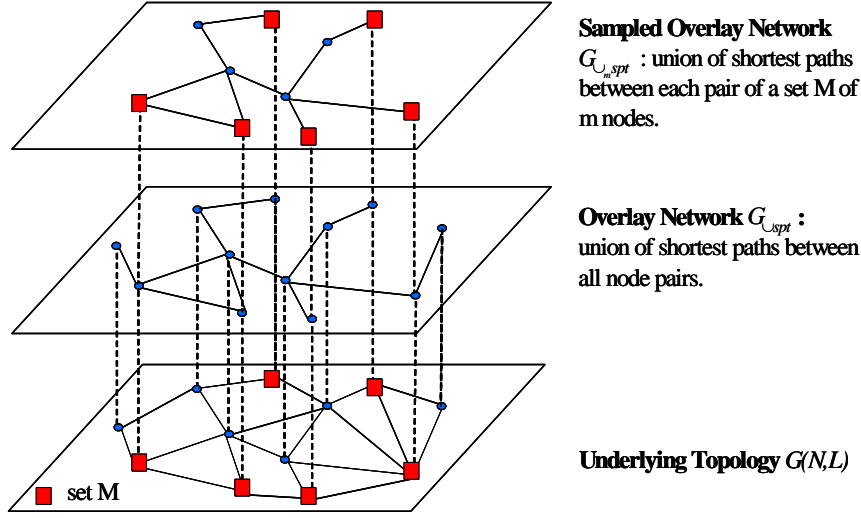
Figure 7.1: The relation between the sampled overlay network, the overlay network and the underlying graph.

7.1 Modeling the sampling process of large networks

Assuming that traceroutes used in RIPE NCC and the PlanetLab are shortest paths, the sampled topology is then the union $G_{\cup_m spt}$ of shortest paths between each pair of a small group of $m \ll N$ nodes, while the number of nodes in the underlying graph N is much larger. When $m = N$, the graph $G_{\cup_m spt}$ becomes $G_{\cup spt}$, the union of all shortest paths between any node pair. $G_{\cup spt}$ is thus the maximal measurable or observable part of a network by traceroute measurements, which has been extensively investigated in Chapter 6. An example to represent the relation between the sampled overlay network $G_{\cup_m spt}$, the overlay network $G_{\cup spt}$ and the underlying graph (or substrate) is shown in Figure 7.1. The performance of the network for a given service can be characterized by the structural properties of $G_{\cup spt}$. Hence, the sampling bias refers to the difference between the sampled overlay $G_{\cup_m spt}$ and the overlay network $G_{\cup spt}$. We show in Section 7.3 that the sampling bias can be quantitatively characterized by $\frac{E[L_{mspt}]}{E[L_o]}$, where L_{mspt} and L_o are the number of links in $G_{\cup_m spt}$ and $G_{\cup spt}$. When the underlying graph is an unweighted network, the overlay network is equal to the underlying graph $G_{\cup spt} = G(N, L)$, because each link (i, j) in $G(N, L)$ is the shortest path between node i and j.

We consider two classes of substrates: the Erdös-Rényi random graphs $G_p(N)$ equipped with i.i.d. uniform link weight within $[0, 1]$ as motivated in Chapter 2 and the unweighted real-world complex networks that are enumerated in Section 2.1.5. Furthermore, other network models, such as power law graphs are usually sparse. The sampling via $G_{\cup_m spt}$ of a sparse network is the same no matter this network is weighted

Figure 7.2: Sketch of the sampled overlay $G_{\cup_m spt}$ and the overlay $G_{\cup spt}(N)$ on top of the underlying graph $G(N, L)$ and the overlay $G_{\cup spt}(m)$ on the subgraph $G_{\mathcal{M}}$.

or not, because paths between any node pair are likely unique. Hence, in the class of the weighted networks, we consider the Erdös-Rényi random graph $G_p(N)$, which is dense. And some of the unweighted real-world networks possess a power law degree distribution.

7.2   Effect of $G_{\mathcal{M}}$ on the sampled overlay $G_{\cup_m spt}$

Recall that a network is mapped as $G_{\cup_m spt}$, the union of shortest paths between each pair of a set $\mathcal{M}$ of m testboxes. The overlay network $G_{\cup spt}$ is the union of the shortest paths between all node pairs. We examine first the effect of $G_{\mathcal{M}}$ on the sampled overlay $G_{\cup_m spt}$ when the underlying network or substrate is a weighted Erdös-Rényi random graph. As shown in Figure 7.2, the subgraph $G_{\mathcal{M}}$ of a underlying network $G(N, L)$ is the set $\mathcal{M}$ and the direct links between nodes of set $\mathcal{M}$. The maximal observable part of the subgraph $G_{\mathcal{M}}$ is the overlay network $G_{\cup spt}$ upon $G_{\mathcal{M}}$. It is now denoted as $G_{\cup spt}(m)$ to include the number of nodes in the overlay network and $G_{\cup spt}(m) \subset G_{\mathcal{M}}$. The sampled overlay $G_{\cup_m spt}$ and the overlay $G_{\cup spt}(N)$ are constructed based on the shortest paths computed in the underlying network $G(N, L)$ while the overlay $G_{\cup spt}(m)$ on the subgraph $G_{\mathcal{M}}$ is based on the shortest path computation in the subgraph $G_{\mathcal{M}}$. Similar to the overlay $G_{\cup spt}(m)$, the sampled network $G_{\cup_m spt}$ is also the union of shortest path between each node pair of the set $\mathcal{M}$, however, upon the underlying network $G(N, L)$ instead of upon the subgraph $G_{\mathcal{M}}$. We now examine the similarity or difference between $G_{\cup spt}(m)$ and $G_{\cup_m spt}$.
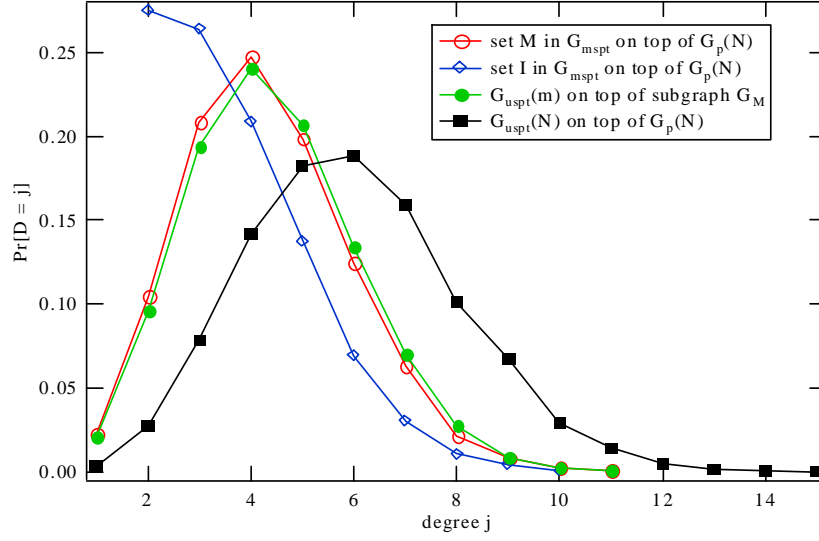
Figure 7.3: Degree distribution of (a) the sampled overlay $G_{\cup_m spt}$ upon $G_{0.6}(200)$ (b) overlay $G_{\cup spt}(N)$ upon $G_{0.6}(200)$ and (c) overlay $G_{\cup spt}(m)$ upon the subgraph $G_{\mathcal{M}}$, where $m = 40$.

Each simulation on Erdös-Rényi random graphs consists of 10^4 iterations. Within each iteration, a set $\mathcal{M}$ of $m = 40$ nodes is uniformly chosen out of the generated substrate $G_{0.6}(200)$ and an *i.i.d.* uniform link weight is assigned to each link. We compute three netwo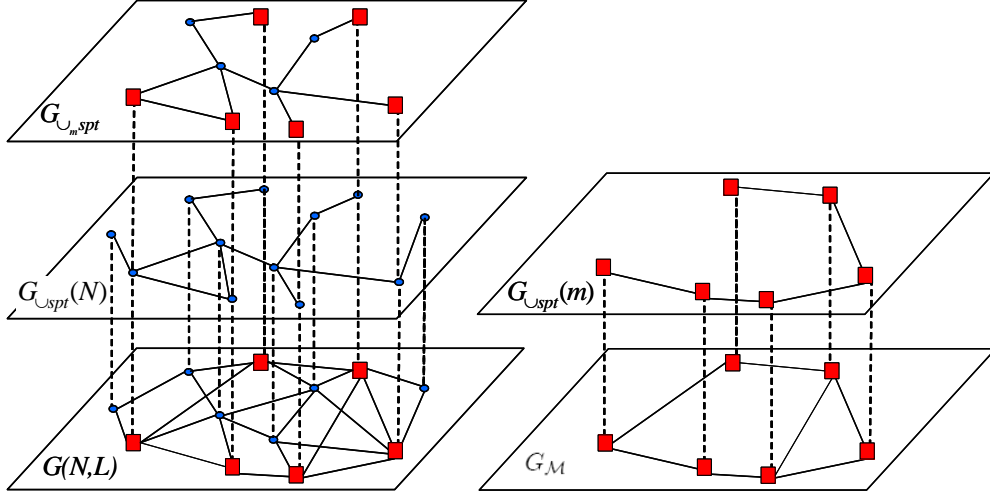rks (a) the sampled overlay $G_{\cup_m spt}$ and (b) the overlay $G_{\cup spt}(N)$ on top of the underlying graph $G(N, L)$ and (c) the overlay $G_{\cup spt}(m)$ on the subgraph $G_{\mathcal{M}}$. The degree distributions of these three networks are displayed in Figure 7.3. We denote $D_{\mathcal{M}}$ as the degree of set $\mathcal{M}$ in the sampled overlay $G_{\cup_m spt}$. The degree distribution of $D_{\mathcal{M}}$ is much closer to the degree distribution of the overlay $G_{\cup spt}(m)$ on top of $G_{\mathcal{M}}$ than that of the overlay $G_{\cup spt}(N)$. Beside the set $\mathcal{M}$, the other nodes in the sampled overlay $G_{\cup_m spt}$ belong to set $\mathcal{I}$. The degree distribution $D_{\mathcal{I}}$ of set $\mathcal{I}$ performs even worse to represent the overlay $G_{\cup spt}(N)$ as compared to set $\mathcal{M}$.

We further investigate the resemblance in degree distribution between $D_{\mathcal{M}}$ and the overlay $G_{\cup spt}(m)$ on the subgraph $G_{\mathcal{M}}$ over more Erdös-Rényi random graphs: $G_{0.2}(400)$ and $G_{0.2}(800)$ with different size m of the set $\mathcal{M}$. Figure 7.4 illustrates that the set $\mathcal{M}$ in the sampled overlay $G_{\cup_m spt}$ and the overlay $G_{\cup spt}(m)$ upon $G_{\mathcal{M}}$ possess almost the same degree distribution. The degree distribution of the overlay $G_{\cup spt}(m = 10, 20, 50)$ upon $G_{\mathcal{M}}$ is calculated based on Corollary 22, using

$$\Pr[D_{G_{\cup spt}}(m) = k] = \frac{(-1)^{m-1-k} S_{m-1}^{(k)}}{(m-1)!}$$
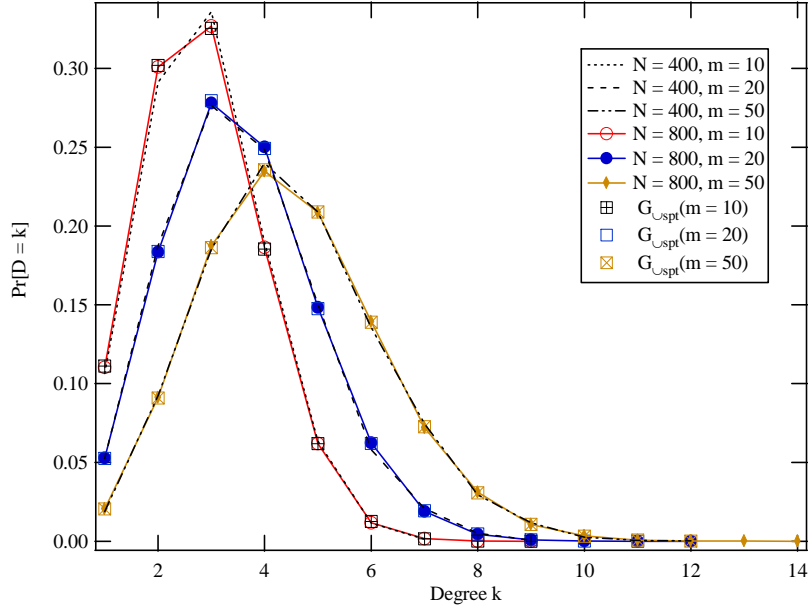
Figure 7.4: Degree distribution $D_{\mathcal{M}} \left(N_{mspt} \right)$ of set $\mathcal{M}$.

It seems that $\Pr \left[D_{\mathcal{M}} = k \right] = \Pr[D_{G_{\cup spt}}(m) = k]$. The degree distribution of the set $\mathcal{M}$ in the sampled overlay $G_{\cup_m spt}$ is independent of the size N of the underlying network: the set $\mathcal{M}$ follows a same degree distribution in $G_{\cup_m spt}(N = 400)$, $G_{\cup_m spt}(N = 800)$ and $G_{\cup_m spt}(N = m) = G_{\cup spt}(m)$. Hence, we claim the following conjecture:

Conjecture 24 *Consider the sampled overlay graph $G_{\cup_m spt}$ on top of an Erdös-Rényi random graph $G_p(N)$ with link density p above the disconnectivity threshold p_c and equipped with i.i.d. regular link weights. The degree distribution of $D_{\mathcal{M}}$ of set $\mathcal{M}$ in the sampled overlay graph $G_{\cup_m spt}$ is independent of the size N of the network and*

$$\Pr \left[D_{\mathcal{M}} = k \right] = \Pr[D_{G_{\cup spt}}(m) = k] = \frac{(-1)^{m-1-k} S_{m-1}^{(k)}}{(m-1)!}$$

As presented in Appendix A.6, two extreme cases $\Pr \left[D_{\mathcal{M}} = 1 \right]$ and $\Pr \left[D_{\mathcal{M}} = m - 1 \right]$ can be proved. The Conjecture 24 states that the degree distribution of the set $\mathcal{M}$ is independent of the size of the underlying topology, but only of the number m of measurement nodes in $\mathcal{M}$. This "intermediate node invariant" degree property could be used, in principle, to reduce or infer $G(N, L)$ and the link weight structure. In other words, if the so measured $G_{\cup spt}(m)$ statistically has the same degree distribution as the set $\mathcal{M}$ of $G_{\cup_m spt}$, the network is possibly homogeneous and equipped with i.i.d. regular link weights.

On top of a dense homogeneous network equipped with i.i.d. regular link weights, the set $\mathcal{M}$ of the sampled overlay network well reflects the local overlay $G_{\cup spt}(m)$ on top of a subgraph $G_{\mathcal{M}}$ in the degree distribution, although $m \ll N$. It seems that the testboxes, i.e. the subgraph $G_{\mathcal{M}}$ (or, equivalently, $G_{\cup spt}(m)$ upon the subgraph $G_{\mathcal{M}}$) do effect the sampled overlay $G_{\cup_m spt}$ in the degree distribution of set $\mathcal{M}$. The Erdös-Rényi random graph is homogenous and so is the subgraph $G_{\mathcal{M}}$. Hence, the resemblance in degree distribution between $D_{\mathcal{M}}$ and the overlay $G_{\cup spt}(m)$ may originate from the fact that both $G_{\cup spt}(m)$ and $G_{\cup_m spt}$ take into account the union of $m(m-1)/2$ shortest paths.

In a real-world unweighted network, the overlay network $G_{\cup spt}(N)$ is equal to the substrate $G(N, L)$ and the overlay network $G_{\cup spt}(m)$ on top of subgraph $G_{\mathcal{M}}$ is $G_{\mathcal{M}}$ itself. For unweighted networks, we have

$$G_{\mathcal{M}} = G_{\cup spt}(m) \subset G_{\cup_m spt} \subset G_{\cup spt}(N) = G(N, L)$$

where $G_{\cup spt}(m) \subset G_{\cup_m spt}$ is due to the fact that any link (i, j) in an unweighted graph is the shortest path between its end nodes i and j. The structure of $G_{\cup_m spt}$ varies between $G_{\mathcal{M}}$ and the substrate $G(N, L)$. Hence, the subgraph $G_{\mathcal{M}}$ is correlated with the sampled network $G_{\cup_m spt}$, in the sense that $G_{\mathcal{M}} = G_{\cup spt}(m) \subset G_{\cup_m spt}$. As a larger proportion of the substrate is observed, the sampled overlay $G_{\cup_m spt}$ resembles the underlying network $G_{\cup spt}(N) = G(N, L)$ more.

7.3 Effect of the relative size m/N of the testboxes

In this section, we first explain why $E[L_{mspt}]/E[L_o]$ quantifies the sampling bias well. Then, we investigate the effect of the relative size m/N of the testboxes on the sampling bias. Given the ratio m/N, the sampling bias differs for various networks depending on their topologies. We will briefly discuss which type of network tends to possess small sampling bias.

7.3.1 Characterizing the sampling bias by $E[L_{mspt}]/E[L_o]$

The sampling bias refers to the difference between the sampled overlay $G_{\cup_m spt}$ and the overlay network $G_{\cup spt}$. The relation $G_{\cup_m spt} \subset G_{\cup spt}(N)$ holds for both weighted Erdös-Rényi random graphs and unweighted networks. Hence, the ratio of the average number of links in the $G_{\cup_m spt}$ and $G_{\cup spt}$, $E[L_{mspt}]/E[L_o]$ can reasonably well characterize[2] the sampling bias of a network, where $E[L_o] = L$ in case the network is unweighted.

First, Figure 7.5 shows that the probability distribution of the normalized number of links $L^*_{mspt} = \frac{L_{mspt} - E[L_{mspt}]}{\sigma[L_{mspt}]}$ and the normalized number of nodes $N^*_{mspt} = \frac{N_{mspt} - E[N_{mspt}]}{\sigma[N_{mspt}]}$

---

[2]$E[L_{mspt}]/E[L_o]$ is a statistical property which takes into account different realizations of the set $\mathcal{M}$ selection as well as the link weight assignment.
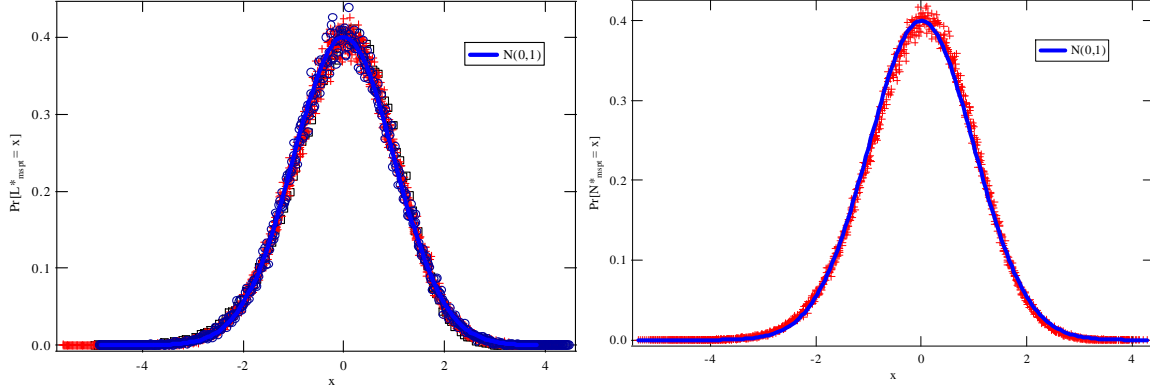
Figure 7.5: Probability distribution of the normalized number of links(left) L^*_{mspt} and nodes(right) N^*_{mspt} in $G_{\cup_m spt}$ on top of $G_{0.2}(800)$ and $m = 10, 20, ..., 60$.

in $G_{\cup_m spt}$ are both close to the Gaussian distribution $N(0, 1)$. Moreover, their average and standard deviation, which determine the distribution, follow $\sigma(L_{mspt}) \ll E[L_{mspt}]$ and $\sigma(N_{mspt}) \ll E[N_{mspt}]$ as illustrated in Figure 7.6 and 7.7. Hence, the random variables L_{mspt} and N_{mspt} are close to their mean $E[L_{mspt}]$ and $E[N_{mspt}]$, which are thus the appropriate quantities to be studied.

Furthermore, we investigate the sampling bias via $E[L_{mspt}]/E[L_o]$ instead of the number of nodes $E[N_{mspt}]/N$. The relation between $E[N_{mspt}]$ and $E[L_{mspt}]$ follows from the basic law of the degree:

$$\sum_{j=1}^{m} d_{j\in\mathcal{M}} + \sum_{j=m+1}^{N_{mspt}} d_{j\in\mathcal{I}} = 2L_{mspt}$$

Taking the expectation yields

$$m \cdot E[D_{\mathcal{M}}] + E[\sum_{j=m+1}^{N_{mspt}} d_{j\in\mathcal{I}}] = 2E[L_{mspt}]$$

Assume that N_{mspt} and $d_{j\in\mathcal{I}}$ are only weakly dependent such that we may apply Wald's identity [67, Chapter 1],

$$2E[L_{mspt}] \simeq m \cdot E[D_{\mathcal{M}}] + (E[N_{mspt}] - m) \cdot E[D_{\mathcal{I}}(N_{mspt})]$$

or

$$E[L_{mspt}] \simeq \frac{1}{2}E[D_{\mathcal{I}}(N_{mspt})] \cdot E[N_{mspt}] + \frac{m}{2}(E[D_{\mathcal{M}}] - E[D_{\mathcal{I}}(N_{mspt})]) \qquad (7.1)$$
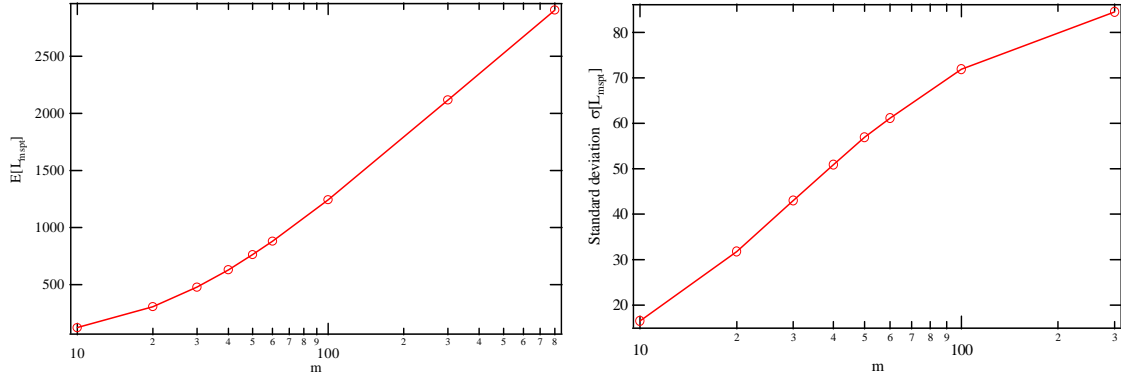
Figure 7.6: Average and standard deviation of the number of links in $G_{\cup_m spt}$ on top of $G_{0.2}(800)$.

Under the assumption of weakly dependence between N_{mspt} and $d_{j \in \mathcal{I}}$, a linear relation exists between $E[L_{mspt}]$ and $E[N_{mspt}]$ with slope equal to $E[D_{\mathcal{I}}(N_{mspt})]/2$, where $E[D_{\mathcal{I}}(N_{mspt})]$ is a function of m. For example, we consider the substrate $G_{0.2}(800)$ equipped with i.i.d. uniformly distributed link weights. The left and right sides of (7.1) are shown to be almost the same in the table below, which justifies the weak dependency assumption.

m	10	20	30	40	50	60	100	300
left side of (7.1)	124.4	308.3	479.6	630.6	762.6	881.4	1242.6	2111.7
right side of (7.1)	124.6	308.6	479.6	630.6	763.4	881.1	1244	2117.2

7.3.2 Sampling of the weighted Erdös-Rényi random graph

The average number of links in the SPT rooted at a source to m uniformly chosen nodes in the complete graph K_N, or approximately in $G_p(N)$, with uniform link weights is given in [67, Chapter 17],

$$g_N(m) = \frac{mN}{N-m} \sum_{k=m+1}^{N} \frac{1}{k} \simeq \frac{mN}{N-m} \log \frac{N}{m} \tag{7.2}$$

Hence, the number of links in each of the m $SPTs$ of $G_{\cup_m spt}$ is, on average, equal to $g_N(m-1)$. The maximum number of links that can be detected in case $m = N$ via $G_{\cup spt}$ is $E[L_o]$ given by (6.10). Since L_{mspt} is not decreasing in m, we have that

$$g_N(m-1) \leq E[L_{mspt}] \leq E[L_o]$$

and

$$E[L_{mspt}] \leq m \cdot g_N(m-1)$$
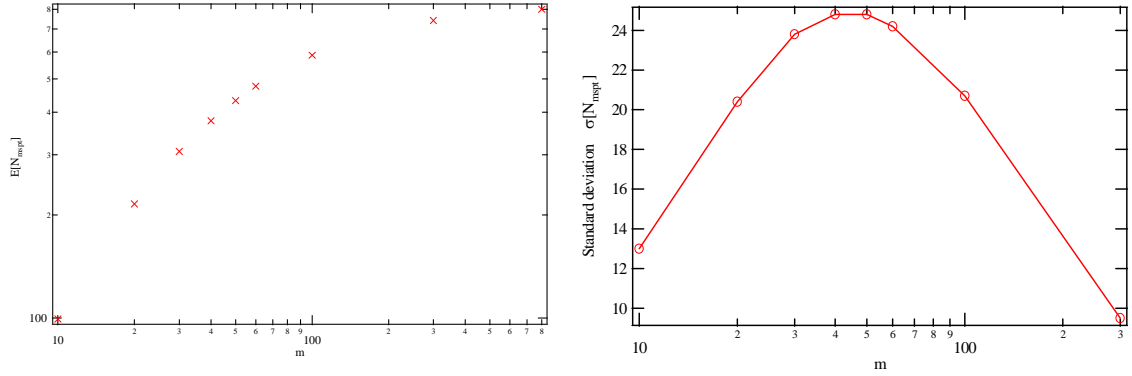
Figure 7.7: The average and standard deviation of the number of nodes N_{mspt} in $G_{\cup_m spt}$ on top of $G_{0.2}(800)$.

Hence, for large N,

$$\frac{(m-1)N}{N-m+1} \log \frac{N}{m-1} \leq E[L_{mspt}] \leq \frac{N}{2}(\gamma + \ln N)$$

where $\gamma = 0.57721...$ is the Euler constant. The ratio $E[L_{mspt}]/E[L_o]$ quantifies the sampling bias, while the ratio $E[L_{mspt}]/(m \cdot g_N(m-1))$ reflects the extent of overlap between these m $SPTs$. As shown in Figure 7.8, for the substrate $G_{0.2}(800)$ and
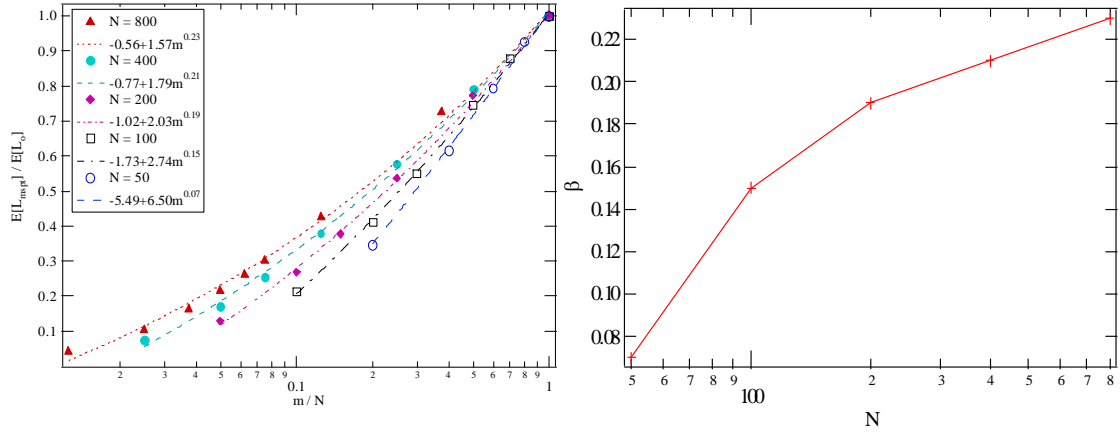
Figure 7.8: The ratio $E[L_{mspt}]/E[L_o]$ and the power exponent β in the corresponding fitting $E[L_{mspt}]/E[L_o] = O(m^\beta)$, where the substrate is $G_{0.2}(N)$.

$m = 60$, 30% of links in $G_{\cup spt}$ have already been observed. For $m = 120$, about 40% links

are discovered. Indeed, for any network, the larger m is, the smaller the sampling bias is, because $\lim_{m \to N} G_{\cup_m spt} = G_{\cup spt}(N)$. For $N = 800$, the ratio $E[L_{mspt}]/E[L_o] = O(m^\beta)$ with $\beta \approx 0.23$, which implies that "the discovering rate of new links" decreases with m. In other words, to obtain an increasingly accurate view of the network, a higher detection/measuring effort is needed, in fact, much higher than proportional. Since $E[L_{mspt}]/E[L_o] = O(m^\beta)$, we found via simulation that the exponent β increases with N. When $A = \frac{m}{N} \to 0$, the shortest paths between nodes of set $\mathcal{M}$ seldom overlap,

$$E[L_{mspt}] \simeq \binom{m}{2} E[H_N] = \frac{A^2 N^2}{2} E[H_N]$$

Using (6.7) and [67, Section 16.3], we have

$$\frac{E[L_{mspt}]}{E[L_o]} \simeq \frac{\frac{A^2 N^2}{2} E[H_N]}{\frac{N(N-1)}{2} p_o} \simeq \frac{\frac{A^2 N^2}{2} (\ln N + \gamma)}{\frac{N}{2} (\ln N + \gamma)} = A^2 N$$

where p_o is the link density of the overlay $G_{\cup spt}$. Hence, for a small m/N, large networks tend to have a small sampling bias or large $E[L_{mspt}]/E[L_o]$. Moreover, a sparse overlay network characterized by a small p_o tends to have a small sampling bias, as observed in real-world complex network sampling in Section 7.3.3.

7.3.3   Sampling of the real-world complex networks

On top of each real-world network, we increase the size of the set $\mathcal{M}$ from $\frac{m}{N} = 5\%$ to $\frac{m}{N} = 35\%$ with a step of 5%. Given $\frac{m}{N}$, each simulation consists of 40 realizations[3] of the random selection of set $\mathcal{M}$. The average proportion of links $E[L_{mspt}]/L$ discovered in the corresponding sampled overlay $G_{\cup_m spt}$ is plotted as a function of $\frac{m}{N}$ in Figure 7.9. Similar to the weighted Erdös-Rényi random graph, to obtain an increasingly accurate view of the network, a higher than linear detection/measuring effort m/N is needed.

With a given proportion m/N of uniformly distributed testboxes in a network, the sampling bias $E[L_{mspt}]/L$ depends purely on the topology of the network. We compare the real-world complex networks (the first 8 networks) mentioned in Section 2.1.5 to see which kind of network tends to possess small sampling bias. We examined the following topological metrics which are considered relevant in the networking literature [62]:

- number of nodes N and links L.

- Average degree $E[D] = 2L/N$ and link density $p = \frac{L}{\binom{N}{2}}$.

---

[3]20 or 10 iterations are carried out for large networks with $N > 3000$.
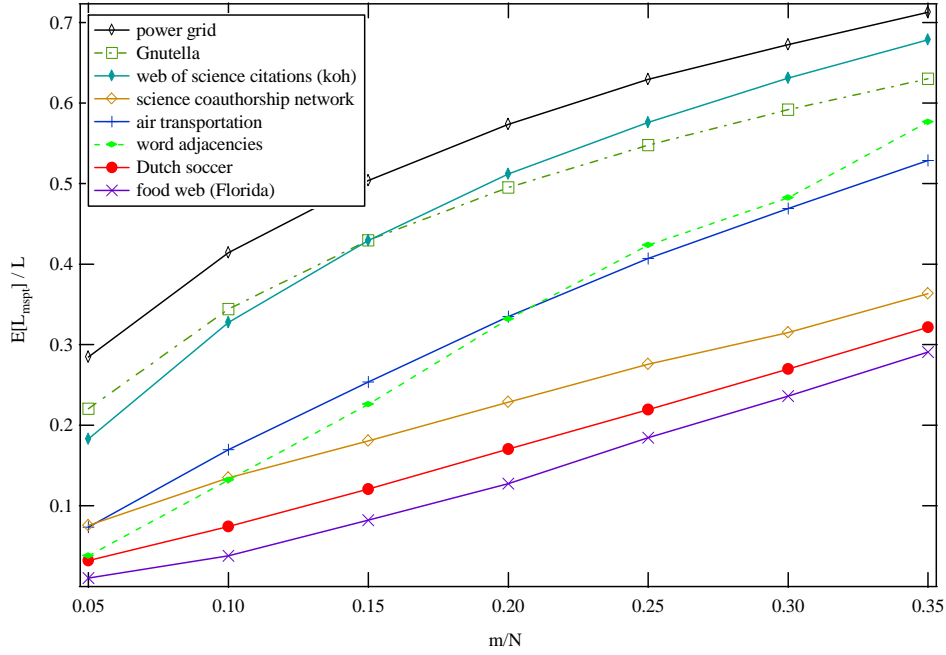
Figure 7.9: The average proportion of links $E[L_{mspt}]/L$ discovered via $G_{\cup_m spt}$ as a funtion of the relative size m/N of set $\mathcal{M}$.

- The average hopcount $E[H_N]$ of the shortest path between each node pair and the largest hopcount $h_{\max}$ or the diameter of a graph. Actually, we assign independently to each link a unit link weight plus a small uniform random variable within $[-\frac{1}{N}, \frac{1}{N}]$, such that a unique shortest path is found between each node pair.

- The clustering coefficient $C(G)$ of a graph.

Table 7.1 presents the topological metrics of the real complex networks, in the decreasing order of their corresponding $E[L_{mspt}]/L$ at $m/N = 5\%$ shown in Figure 7.9. Recall that a larger proportion $E[L_{mspt}]/L$ of the substrates observed via $G_{\cup_m spt}$ implies a lower sampling bias. Figure 7.9 and Table 1 show that a network tends to have a small sampling bias if its link density p is low and the average hopcount $E[H_N]$ is large, especially for small m/N. Indeed, when $A = \frac{m}{N} \to 0$, the shortest paths between set $\mathcal{M}$ seldom overlap and

$$\frac{E[L_{mspt}]}{L} \simeq \frac{\frac{A^2 N^2}{2} E[H_N]}{\frac{N(N-1)}{2} p} \simeq \frac{A^2 E[H_N]}{p} \tag{7.3}$$

For any m, the proportion of observed links $\frac{E[L_{mspt}]}{L}$ can be upper bounded by (7.3). When m is larger, the shortest paths between set $\mathcal{M}$ overlap more, and $\frac{E[L_{mspt}]}{L}$ is far

Table 7.1: Topological metrics of the real complex networks

Table 1	N	L	C	$E[H_N]$	$h_{\max}$	$E[D]$	p
Power grid	4941	6594	0.11	18.99	46	2.67	0.00054
Gnutella Crawl2	1568	1906	0.04	6.10	21	2.43	0.0016
Web of Science Citations(koh)	3704	12673	0.30	3.67	12	6.84	0.0018
Science coauthorship network	379	914	0.80	6.03	17	4.82	0.0128
Air Transportation	2179	31326	0.59	3.02	8	28.75	0.0132
Word adjacencies	112	425	0.19	2.51	5	7.59	0.068
Dutch soccer	685	10310	0.75	4.45	11	30.10	0.044
Food web(Florida)	128	2075	0.33	1.76	3	32.42	0.26

smaller than its upper bound (7.3). For larger m/N, the sampling bias of these networks may have a different order. No clear correlation between the sampling bias and other metrics have been found.

In summary, in both the weighted Erdös-Rényi random graph and unweighted real-world networks, to obtain an increasingly accurate view of the network, a higher than linear detection/measuring effort m/N is needed. When m/N is small, the sampling bias depends purely on the average hopcount $E[H_N]$ and the link density p (or p_o) of an unweighted network (or of the overlay $G_{\cup spt}$ upon a weighted network). Indeed, a larger average hopcount $E[H_N]$ and a small p or p_o imply a small sampling bias. For small m/N, the sampling bias of the weighted Erdös-Rényi random graph is positively correlated with N.

7.4 Conclusions

In this chapter, we study a network sampling method originated from the Internet, namely $G_{\cup_m spt}$ the union of m shortest path trees, or equivalently, the union of shortest paths between each pair of a set M of m testboxes. The analysis covers a wide class of networks, ranging from real-world unweighted complex networks to the weighted Erdös-Rényi random graphs.

The interconnections of set $\mathcal{M}$, i.e. the subgraph $G_{\mathcal{M}}$, are correlated with the sampled network $G_{\cup_m spt}$ as follows: When the underlying network is a real-world unweighted network $G(N, L)$, $G_{\mathcal{M}}$ is a subgraph of the sampled overlay $G_{\cup_m spt}$. Surprisingly, when the underlying network is an Erdös-Rényi random graph equipped with i.i.d. regular link weights, the set $\mathcal{M}$ in the sampled overlay graph $G_{\cup_m spt}$ follows the same degree distribution as the overlay $G_{\cup spt}(m)$ upon $G_{\mathcal{M}}$. The degree distribution of $D_{\mathcal{M}}$ of the

set $\mathcal{M}$ in the sampled overlay graph $G_{\cup_m spt}$ is independent of the size N of the network.

To obtain an increasingly accurate view of a given network, a higher detection/measuring effort (the size m of set $\mathcal{M}$) is needed, in fact, higher than proportional.

When m/N is small, as in RIPE NCC and the PlanetLab measurement where the number m of testboxes (hundreds) is much smaller the number of routers in the Internet (hundreds of thousands), the sampling bias tends to be small if the average hopcount $E[H_N]$ is large and the link density p, or link density p_o of the overlay network $G_{\cup spt}$, is small. Hence, a large number of testboxes placed far from each other is preferable for good network topology measurements. Furthermore, the sampled overlay network consists of a large number, m, of shortest paths that either start or end at each testbox. Links connected to the testboxes are more likely to be sampled than the other links. Hence, placing testboxes at hubs (nodes with a high degree in the underlying network) may contribute to a small sampling bias. In the sampled overlay $G_{\cup_m spt}$, the set of m textboxes tend to possess a higher average degree than the other (intermediate) nodes, if the underlying network is dense[4], as observed in Figure 7.3.

---

[4]When the underlying network is sparse, the uniformly distributed testboxes tend to possess a small degree in the underlying network, which limites the number of links incident to the testboxes to be sampled. On the other hand, those few high degree nodes in the underlying network are likely to appear in the sampled overlay as the intermediate nodes. Hence, in the sampled overlay network, the average degree of the intermediate nodes may be higher than that of the testboxes.

Chapter 8

Betweenness centrality in a weighted network

In large complex networks, not all links have equal importance. For example, if two clusters are connected by one link, the removal of this link will disable all the traffic flowing between these two clusters. In contrast, the removal of a link connecting to a dead-end whose degree is one, will have no effect on the other parts of the network. The importance of links is of primary interest for network resilience to attacks [5][23] and immunization against epidemics [81]. A good measure for "link/node importance" is the betweenness $B_l(B_n)$ of a link (node), which is defined as the number of shortest paths between all possible pairs of nodes in the network that traverse the link (node). The betweenness $B_l(B_n)$ which incorporates global information is a simplified quantity to assess the maximum possible traffic. Assuming that a unit packet is transmitted between each node pair, the betweenness B_l is the total amount of packets passing through a link.

Recall that the overlay network $G_{\cup spt}$ is the union of the shortest paths between all possible node pairs. The importance and structure features of the overlay network have been extensively studied in Chapter 6. Since all the traffic traverses only the overlay $G_{\cup spt}$ and all the nodes in the substrate also appear in the overlay $G_{\cup spt}$, the betweenness of a node in the substrate is equal to the betweenness of that node in the overlay $G_{\cup spt}$. A link in the substrate has betweenness 0 if it does not belong to the overlay $G_{\cup spt}$. Otherwise, its link betweenness is the same as that in the overlay $G_{\cup spt}$.

In this chapter, the traffic through the network is examined by the betweenness B_l of links in the overlay $G_{\cup spt}$. The study of betweenness usually deals with scale-free trees [14, 40, 41, 88] or scale-free networks [57, 45] whose degree distribution follow a power law, i.e. $\Pr[D = k] \sim k^{-\tau}$. However, the overlay $G_{\cup spt}$ that we are going to examine possesses different degree distribution e.g. uniform, exponential or power law distribution. The structure of the overlay network $G_{\cup spt}$ can be controlled e.g. by tuning the extreme value index α of the i.i.d. polynomial link weights defined in

(2.1). In the strong disorder limit ($\alpha \to 0$, or $\alpha < \alpha_c$ for large networks), the overlay $G_{\cup spt}(\alpha \to 0)$ becomes the MST. The betweenness of the MST for various network models and real-world complex networks follow surprisingly a power law. This power law betweenness distribution for MST holds more generally than in Erdös-Rényi random graph and scale-free networks as found in [46]. Besides, the relationship between the structural characteristics and its betweenness distribution is investigated. We study the correlation between the link weights and the corresponding link betweenness when the system is in weak disorder, instead of the correlation between the node betweenness and the node degree as in [9, 46].

8.1   Simulation scenarios

Here, we explore the betweenness distribution of the overlay network $G_{\cup spt}$, when the i.i.d. polynomial link weights are tuned via the index α. When $\alpha \to 0$ (or in the $\alpha < \alpha_c$ regime for large networks), all flows are transported over the MST. Hence, $G_{\cup spt(\alpha \to 0)} =$ MST is also called an overlay tree. When $\alpha > \alpha_c$, transport in the network traverses many links. The $\alpha \to 0$ (or $\alpha < \alpha_c$ for large networks) regime corresponds to the strong disorder limit, where the total weight of a path is characterized by the maximum link weight along the path. The shortest path in this case is the path with the minimum value of the maximum link weight. When all links contributes to the total weight of the shortest path, the system is weak disordered, e.g. $\alpha > \alpha_c$. In fact, other distributions that could lead to strong disorder [20] would arrive at similar betweenness behavior, because the MST is probabilistically the same for various i.i.d. link weights distributions [72].

For the underlying topology, or the substrate, we consider the following complex network models: the Erdös-Rényi random graph $G_p(N)$, the square and the cubic lattice and the Barabási-Albert (BA) power law model. We carried out 10^4 iterations for each simulation. Within each iteration, we randomly generate an underlying topology. Polynomial link weights with parameter α are assigned independently to each link. The overlay $G_{\cup spt}$ as well as its betweenness is found by calculating the shortest paths between all node pairs with Dijkstra's algorithm for weak disorder regime. For the strong disorder limit $\alpha \to 0$, $G_{\cup spt(\alpha \to 0)} =$ MST is found by Kruskal's algorithm [25] on the corresponding network with uniform link weights, because with i.i.d. link weights, the structure of the MST is probabilistically the same for various link weights distributions [72]. The betweenness distribution of the overlay tree $G_{\cup spt(\alpha \to 0)}$ is determined totally by the structure of the overlay tree. Furthermore, we also perform the same statistical analysis of the overlay trees $G_{\cup spt(\alpha \to 0)}$ on top of real-world networks that are enumerated in Section 2.1.5. On top of each, usually large network, 100 realizations of i.i.d. uniform link weights assignments are carried out.

8.2 Link weight versus link betweenness

Does a low link weight implies a high link betweenness B_l? When polynomial link weights are independently assigned to links in the substrate, we randomly choose a link in each overlay network $G_{\cup spt}$. The betweenness of this link and the corresponding link weight are plotted in Figure 8.1. According to [74], $\alpha_c = 0.2$ for Erdös-Rényi random
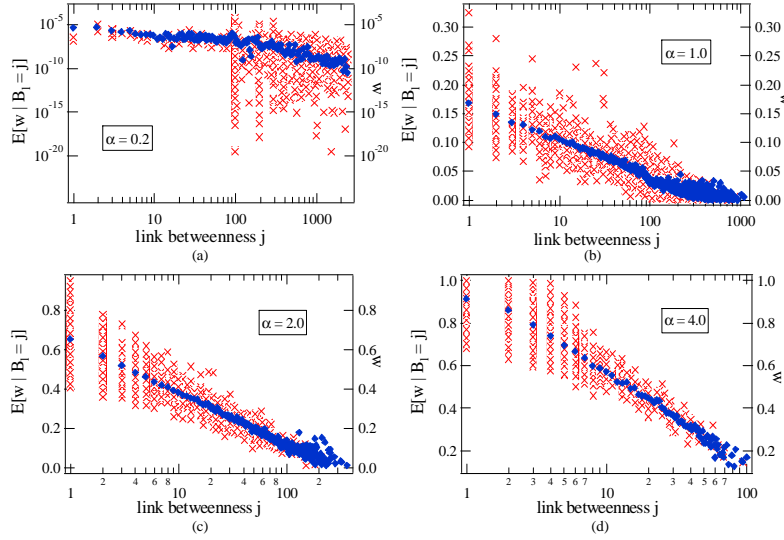
Figure 8.1: The link weight w (cross) versus its link betweenness j and $E[w|B_l = j]$ (square) the average link weight of links with given betweenness j in the overlay $G_{\cup spt}$ on top of Erdös-Rényi random graph $G_{0.4}(100)$.

graph with $N = 100$ nodes[1]. When the system is weakly disordered, i.e. $\alpha > \alpha_c$ (Figure 8.1 b-d), a link with lower link weight is more likely to have higher betweenness. However, when $\alpha = 0.2$ (Figure 8.1(a)), where link weights possess relatively strong fluctuations, the correlation between link weight and betweenness disappears. Hence, a negative correlation exists between the link weight and its betweenness for the weak disorder regime. The correlation becomes stronger as α increases, as illustrated in table 8.1 where the linear correlation coefficient is equal to the covariance between the two random variables divided by the product of their standard deviations. The increasing strength of the correlation for larger α is also reflected by Figure 8.1, where as α increases, the plot of link weights become narrower.

The correlation between the weight and the betweenness of the link is shown to be dependent on the underlying graph as well as on the extreme value index α of

---

[1]α_c depends on the network topology as well as the size N of the network

Table 8.1: The correlation coefficient between weight and betweenness of a link.

α	0.2	1.0	2.0	4.0	8.0	16.0
$G_{\cup spt}$ on $G_{0.4}(100)$	-0.06	-0.61	-0.70	-0.78	-0.84	-0.84
$G_{\cup spt}$ on square lattice $N = 100$	-0.22	-0.53	-0.54	-0.53	-0.53	-0.53
$G_{\cup spt}$ on cubic lattice $N = 125$	-0.18	-0.60	-0.66	-0.67	-0.68	-0.68
$G_{\cup spt}$ on BA $N = 100, m = 3$	-0.12	-0.53	-0.66	-0.60	-0.50	-0.49

the polynomial link weight distribution. For homogeneous network such as the Erdös-Rényi random graph and lattice, the correlation coefficient increases monotonically as α increases. However, in a the nonhomogeneous topology like the BA power law substrate, the correlation coefficient decreases after a maximum has been reached. In a homogeneous network, when α is large, a link with lower link weight tends to attract more traffic. While in a non-homogeneous topology, the relative importance of a link or its connectivity in substrate is also a determinant factor for its betweenness. In short, both the nonhomogeneity of the underlying topology and the link weight disorder (e.g. a smaller α) contribute to the nonhomogeneity of the overlay $G_{\cup spt}$, which reduces the correlation between link weight and betweenness.

8.3   Link betweenness distribution of the overlay $G_{\cup spt}$

The link betweenness represents the total traffic passing through a link if a unit packet is transmitted between each node pair. Hence, the link betweenness distribution reflects how the traffic is distributed over the network.

8.3.1   Overlay $G_{\cup spt}$ on top of complex network models

As shown in Figure 8.2 (a), the traffic on the overlay $G_{\cup spt}$ on top of $G_{0.4}(100)$ varies less for large α, because the betweenness is distributed within a small range. When α is small, as shown in Figure 8.2 (b), the betweenness is ranging approximately over $[1, 2500]$ for $N = 100$ and peaks appear on the betweenness at $n(N - n)$, where $1 \leq n \leq N - 1$.

A link is called critical if its removal will disconnect the overlay $G_{\cup spt}$ into two clusters with n and $N - n$ nodes. The betweenness of such critical link is $n(N - n)$, because all the traffic with source and destination separated in these two clusters will traverse this link. However, if a link has betweenness $n(N - n)$, the removal of this link does not necessarily disconnect the overlay graph.
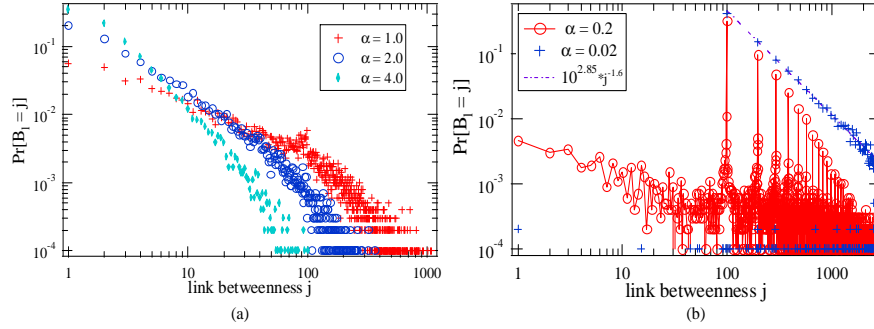
Figure 8.2: *Pdf* of link betweenness B_l in the overlay $G_{\cup spt}$ on top of $G_{0.4}(100)$. The *pdf* for $\alpha = 0.02$ is linear fitted by the dashed line.

As we decrease the extreme value index α, the overlay $G_{\cup spt}$ contains less links and it becomes tree-like or even an exact tree. Any link in a tree is critical. We consider for example, the Erdös-Rényi random graph $G_{0.4}(100)$. When $\alpha = 0.2$, the average number of links in the overlay is 107.2. Within such a sparse overlay topology, a link is very likely to be critical, which contributes to the peaks in Figure 8.2(b). A sparse overlay $G_{\cup spt}$ is composed of the minimum spanning tree and few shortcuts, that direct a small part of the traffic. The largest link betweenness 2500 comes from the critical link which could separate the overlay network into two clusters each with $\frac{N}{2} = 50$ nodes. A link has higher betweenness if it is critical and the maximal link betweenness is achieved when $n = \left\lfloor \frac{N}{2} \right\rfloor$. Hence, the betweenness of any link in a graph with N nodes obeys

$$B_l \leq \left\lfloor \frac{N}{2} \right\rfloor \left(N - \left\lfloor \frac{N}{2} \right\rfloor \right) \tag{8.1}$$

When the overlay becomes a tree, the magnitude of peaks at $n(N-n)$ depends on the structure of the tree. For example, if the overlay network is a star with N nodes, the link betweenness is always $N-1$. And if $G_{\cup spt}$ is a line graph, the betweenness of a link is $n(N-n)$ with n uniformly distributed over $[1, N-1]$. We find that the betweenness distribution of the overlay tree $G_{\cup spt(\alpha \to 0)}$ on top of the Erdös-Rényi random graph $G_p(N)$ follows a power law

$$\Pr[B_l = j] = c_0 j^{-\epsilon}, \quad N - 1 \leq j \leq \left\lfloor \frac{N}{2} \right\rfloor \left(N - \left\lfloor \frac{N}{2} \right\rfloor \right) \tag{8.2}$$

with exponent $c = 1.6$. Further, we observe that the overlay tree $G_{\cup spt(\alpha \to 0)}$ on top of other complex network models such as the lattice, cubic lattice or a BA model also seems to possess a power law betweenness distribution as illustrated in Figure 8.3. The lower bound $N-1$ of the betweenness in a tree is attained at a link connected to a degree
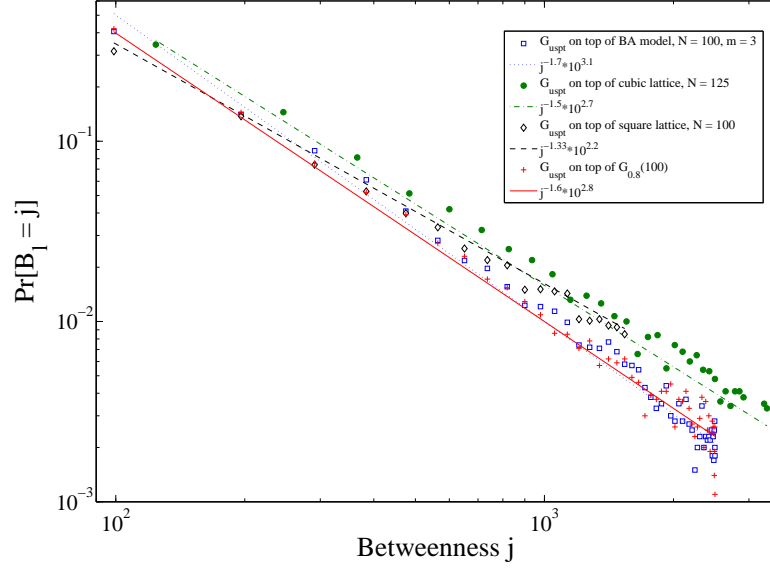
Figure 8.3: Link betweenness distribution (markers) of overlay tree $G_{\cup spt(\alpha \to 0)}$ on top of complex network models and the corresponding linear fitting (dashed lines)

1 node while the upper bound obeys (8.1). The exponent c we found for Erdös-Rényi ($c = 1.6$) lattice ($c = 1.33$) and BA model ($c = 1.7$) with $N \sim 100$ are the same as observed in [103] with $N \sim 8100$. The scaling exponent c seems insensitive to the size N of the network. Additional simulations for Erdös-Rényi random graph suggest that the exponent c is independent of the size N of the underlying graph as well as the link density p, if p is larger than the disconnectivity threshold $p_c \sim \ln N / N$. For example, the power exponent $c = 1.6$ remains the same for the substrate $G_{0.4}(100), G_{0.4}(50)$, $G_{0.8}(100)$ and the Erdös-Rényi random graph in [46] with $N = 10^4$ nodes and $L = 2N$ links.

8.3.2 Overlay tree $G_{\cup spt(\alpha \to 0)}$ on top of real networks

As found in section 8.3.1 and Figure 8.3, an overlay tree $G_{\cup spt(\alpha \to 0)}$ follow a power law betweenness distribution when the substrate is an Erdös-Rényi random graph, a square or cubic lattice or a BA power law graph. It would be especially interesting to examine whether the power law link betweenness distribution still holds for the overlay trees $G_{\cup spt(\alpha \to 0)}$ on top of real-world networks. Hence, we perform a statistical analysis of the real-world networks that are listed in Section 2.1.5.

As shown in Figure 8.4, 8.5 and 8.6, the betweenness distribution of these overlay trees on top of real networks follows, surprisingly, for almost all a power law, while
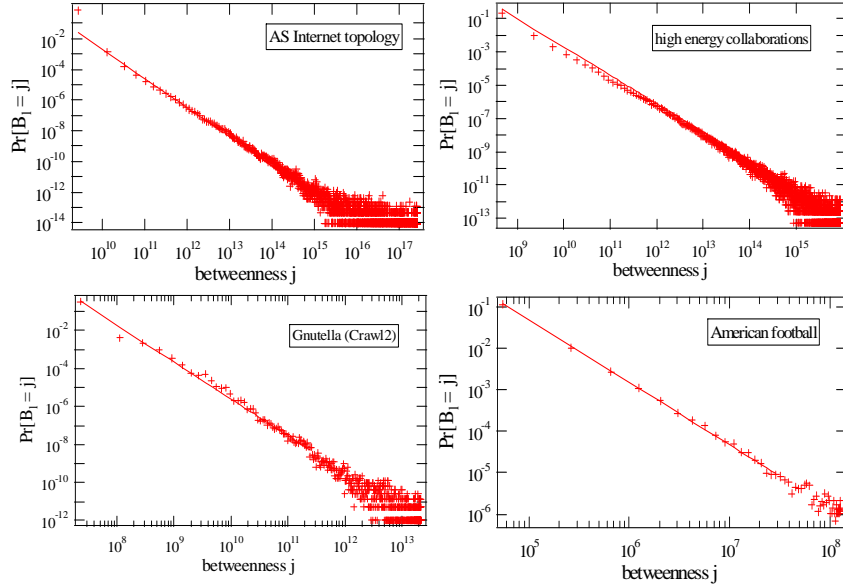
Figure 8.4: Betweenness distribution (+) of $G_{\cup spt(\alpha \to 0)}$ on top of real network topologies. The line is the linear curve fitting.

their corresponding degree distribution of the tree (see Figure 8.7 to 8.9) may differ significantly.

The power law betweenness distribution of the overlay tree $G_{\cup spt(\alpha \to 0)}$ or MST implies that a set of links in the MST possess a much higher betweenness. In [103], it is found that the infinite incipient percolation cluster (IIC), a subgraph of the MST has a significantly higher average betweenness than the entire MST, and the betweenness distribution of the IIC also satisfies a power law. But why does the betweenness distribution of a MST follow a power law? Is that due to the network topology, a particular link weight distribution function or the fact that link weights are independently and identically distributed? The betweenness of the overlay tree follows a power law distribution no matter the substrate is a traditional complex network model or a real network, provided the substrate is denser than a tree. When the substrate is close to a tree, the overlay tree is almost the same as the substrate and the corresponding betweenness distribution does not necessarily follow a power law. Hence, the power law betweenness distribution does not hold for any tree structure but seems to hold for the overlay tree $G_{\cup spt(\alpha \to 0)}$ on top of a substrate which is not too sparse. With i.i.d. link weights, the structure of the overlay tree or MST is probabilistically the same for various link weight distributions, because the ranking of the link weights suffices to construct the MST. Therefore, the i.i.d. link weights compared to the network topology and link weight distribution, contribute more to the power law betweenness distribution of the MST
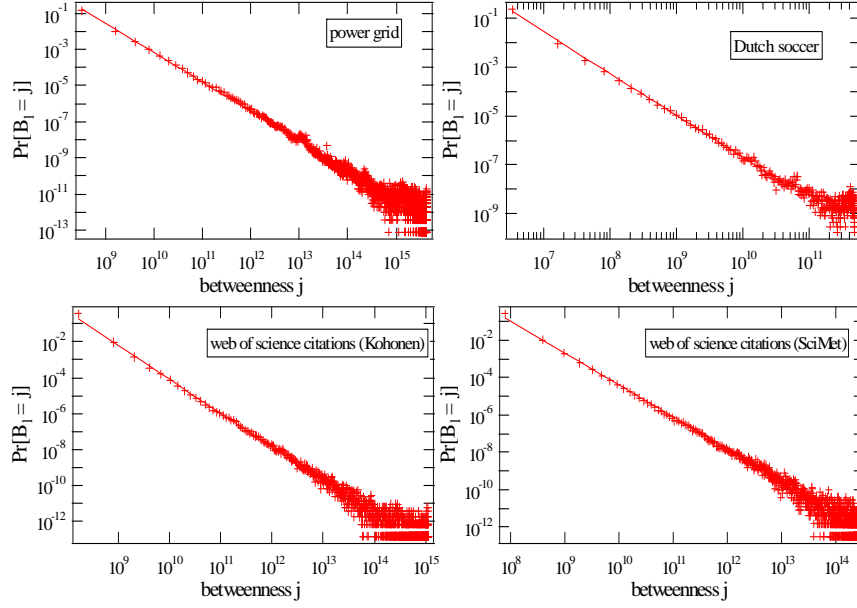
Figure 8.5: Betweenness distribution $(+)$ of $G_{\cup spt(\alpha \to 0)}$ on top of real network topologies. The line is the linear curve fitting.

for various networks. In fact, with i.i.d. link weights, the equivalent Kruskal growth process of the MST starts from N individual nodes and in each step an arbitrary link in the substrate is added while links generating loops are forbidden. However, the power exponent c of the betweenness distribution of a MST is determined by the network topology, due to the exclusion of links that generating loops in the growth process of the MST. The relationship between the topological characteristics of a network and the exponent c of the betweenness distribution of the corresponding MST is studied in Section 8.4.2.

8.4 Betweenness distribution of trees

Since the path between each node pair is unique in a tree and is independent of link weights, the betweenness of a tree depends purely on its tree structure. In the strong disorder limit $(\alpha \to 0)$, the betweenness distribution depends on the structure of $G_{\cup spt(\alpha \to 0)}$ or MST. In this way, we are able to compare the tree structure of overlay $G_{\cup spt(\alpha \to 0)}$ to other classes of trees via the link betweenness distribution. Although trees are special graphs, real-world networks like the Autonomous Systems in the Internet [16] can be modeled by trees or tree-like graphs with a negligible number of shortcuts.

In this section we compare the following trees: (a) the three tree models: the k-
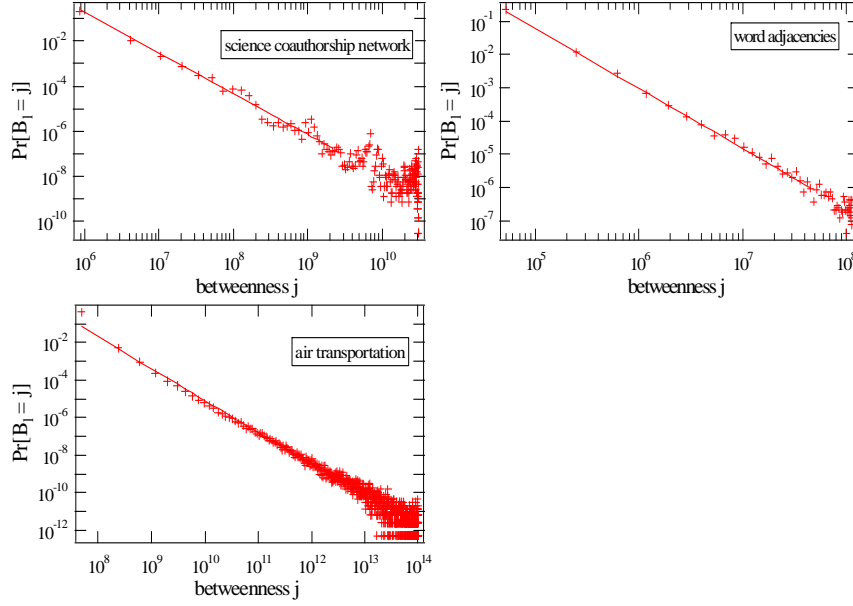
Figure 8.6: Betweenness distribution $(+)$ of $G_{\cup spt(\alpha \to 0)}$ on top of real network topologies. The line is the linear curve fitting.

ary tree, the scale-free trees and the uniform recursive tree URT; (b) the overlay tree $G_{\cup spt(\alpha \to 0)}$ on top of complex network models: the Erdös-Rényi random graph, the square or cubic lattice, and the BA power law model; (c) the overlay tree $G_{\cup spt(\alpha \to 0)}$ on top of real complex networks. The class (b) and (c) have been shown to possess power law betweenness distribution. Hence, it is interesting to first examine whether the class (a) has such power law betweenness distribution.

A link l in any tree connects two clusters with size $1 \leq |C_l| \leq \frac{N}{2}$ and $N - |C_l|$. The betweenness of a link l is $B_l = |C_l| (N - |C_l|)$, because traffic traverses the link l if and only if the source and destination lie in the two clusters separated by l. If $|C_l| = o(N)$, which holds for all but a few large clusters, then we have $B_l \sim |C_l| \cdot N$ for large N.

8.4.1 Betweenness distribution of tree models

k-ary tree

In a k-ary tree the total number of nodes is

$$N(d) = 1 + k + k^2 + \cdots + k^d = \begin{cases} \frac{k^{d+1}-1}{k-1}, & k \neq 1 \\ 1 + d, & k = 1 \end{cases}$$
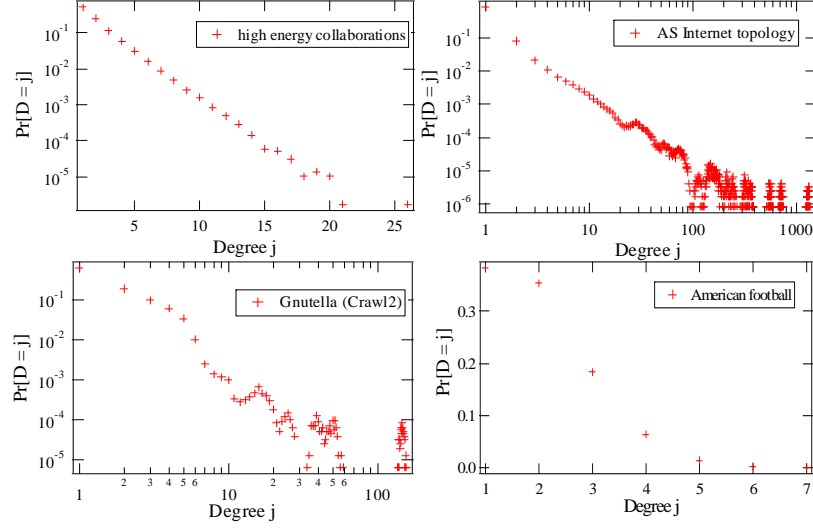
Figure 8.7: Degree distribution of $G_{\cup spt(\alpha \to 0)}$ on top of real network topologies.

A link is called the $j - th$ level link if it connects two nodes which is j and $j - 1$ hops away from the root. The removal of a $j - th$ level link disconnect the graph into two clusters: one is a k-ary tree of depth $d - j$ with $N(d - j)$ nodes and the other cluster has $N(d) - N(d - j)$ nodes. Since there are k^j $j - th$ $(1 \leq j \leq d)$ level links

$$\Pr[|C_l| = N(d - j)] = \frac{k^j}{N(d) - 1}$$

Hence,

$$\Pr[|C_l| = n] = \frac{k^{d+1}}{(N(d) - 1)(kn - n + 1)}, \quad n = N(d - j) \text{ and } 1 \leq j \leq d$$

The approximate betweenness distribution

$$\Pr[B_l \sim n \cdot N] = \frac{k^{d+1}}{(N(d) - 1)(kn - n + 1)}, \quad n = N(d - j) \text{ and } 1 \leq j \leq d$$

follows an inverse power law with exponent $c = 1$. Two exceptions are: the line graph where $k = 1, \Pr[B_l = n(N - n)] = \frac{1}{N-1}, 1 \leq n \leq N - 1$ and the star where $k = N - 1, \Pr[B_l = N - 1] = 1$.

A rigorous analysis based on

$$\Pr[B_l = N(d - j)(N(d) - N(d - j))] = \frac{k^j}{N(d) - 1}, \qquad 1 \leq j \leq d \qquad (8.3)$$

is given in Appendix A.8.
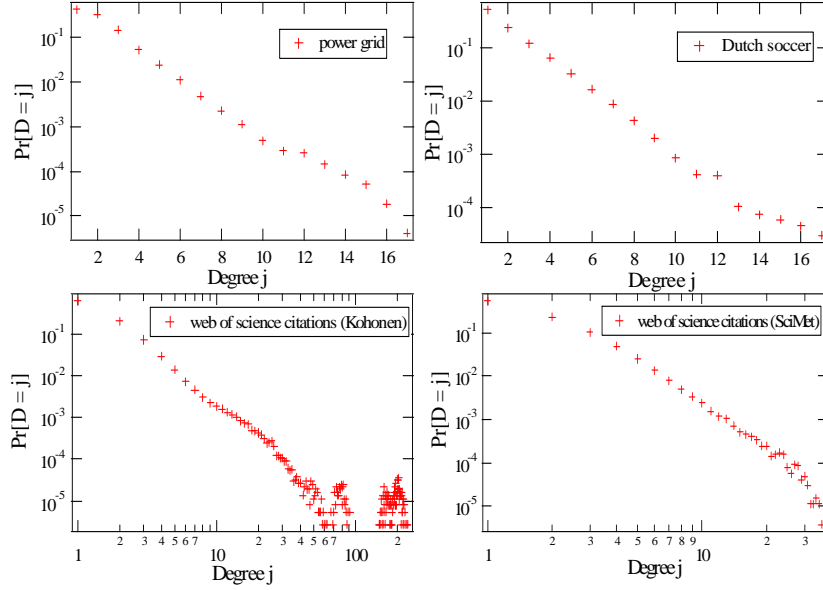
Figure 8.8: Degree distribution of $G_{\cup spt(\alpha \to 0)}$ on top of real network topologies.

Scale-free trees

A scale-free tree contains initially only one node, the root. Then, at each step a new node is attached to one of the existing nodes. The probability that a new node connects to a certain existing node is proportional to the attractiveness of the old node, defined as

$$A(v) = a + q$$

where $a > 0$ denotes the initial attractiveness and q is the in-degree of node v, the number of links connected to the node. The corresponding in-degree distribution [33] is

$$\Pr[D_{in} = q] = (q + a)^{-(2+a)}$$

Early in 2002, the power law betweenness distribution with $c = 2$ for the scale-free trees is solved analytically by Goh et al. [47]. Here, we relate the betweenness distribution to the subtree size distribution, which is derived by Fekete and Vattay [40]. In our notation, the probability distribution of the size of a subtree rooted at a random node in a scale-free tree with N nodes is:

$$\Pr\left[\left|\mathcal{T}^{(N)}\right| = k\right] = \frac{N - \beta}{N - 1} \cdot \frac{1 - \beta}{(k - \beta)(k + 1 - \beta)} \tag{8.4}$$

$$\approx (1 - \beta)\frac{1}{k^2} \tag{8.5}$$
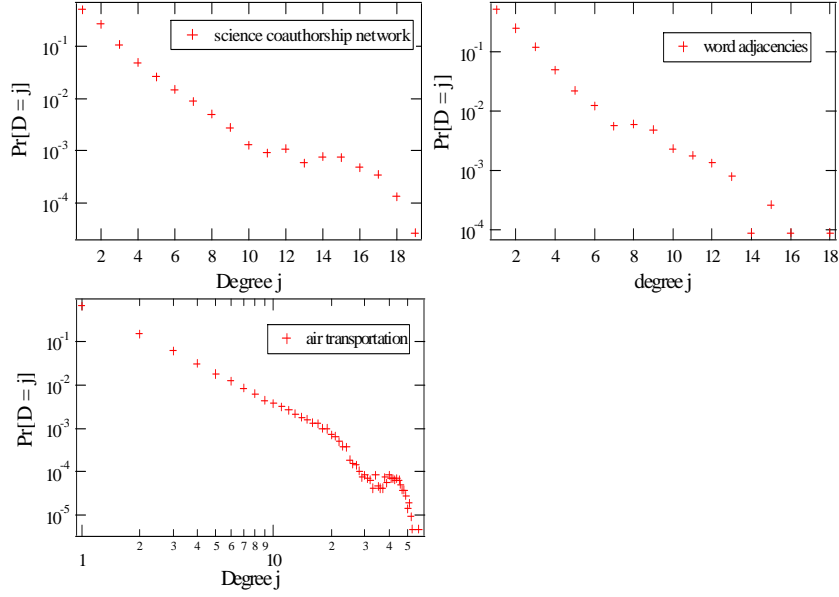
Figure 8.9: Degree distribution of $G_{\cup spt(\alpha \to 0)}$ on top of real network topologies.

where $\beta = \frac{1}{1+a} \in [0,1]$. When $\beta = \frac{1}{2}$, the scale-free tree is exactly the BA tree, with $m = 1$ in the BA model. When $\beta = 0$, the tree becomes a uniform recursive tree URT. Hence, the probability that a link has load approximately kN will be

$$\Pr[B_l = kN] \approx (1 - \beta)\frac{1}{k^2}$$

The inverse square power law betweenness distribution with $c = 2$ holds for the class of scale-free trees where the scaling property of the degree can be finely tuned by the initial attractiveness a. Further as shown in [41], if $N \ll B_l \ll \left(\frac{N}{2}\right)^2$, its complementary distribution can be approximated by power law $\Pr[B_l \geq x] = (1-\alpha)N\frac{1}{B_l}$ which leads to our $c = 2$ scaling for the probability distribution of B_l. The link and node betweenness distribution is considered to be same in a tree [56]. Szabó et al. [88] found the scaling exponent $c = 2$ for node betweenness in a BA tree with a "mean-field" approximation. The rigorous proof of the heuristic result of [88] has been provided by Bollobás and Ridordan in [14].

An URT ($\beta = 0$) possesses in fact exponential degree distribution. A rigorous derivation of link betweenness distribution for URT is given in Appendix A.7.

8.4.2 Comparison of betweenness distribution of overlay trees and tree models

All the three classes of trees have been shown to follow approximately a power law betweenness distributions. The power law betweenness distribution has been proved for class (a) tree models in section 8.4.1, while for class (b) overlay tree on top of network models and (c) overlay tree on top of real networks. It seems to arise from the random sampling of the overlay tree (caused by the i.i.d. link weights) as explained in section 8.3.2.

The slope of the betweenness distribution in log-log scale, or equivalently, the power exponent c of the corresponding power law distribution (8.2), characterizes the variance of the traffic carried along links in the network. High values of c can be interpreted as a high concentration of traffic on the most important links. The betweenness distribution of a tree depends purely on the structure of the tree. Hence, we further examine the relationship between the scaling exponent c and the corresponding tree structure which can be partially characterized by the average hopcount $E[H]$ of the shortest path and the standard deviation $sdev[D]$ of the degree, because the average degree in any tree is $E[D] = 2(N-1)/N = 2 - \frac{2}{N}$.

Table 8.2: Topological characteristics of tree models and overlay tree on top of network models.

	N	c	$E[H]$	$sdev[D]$
BA tree	100	2.3	4.7	2.38
URT	100	2.1	6.6	1.36
$G_{\cup spt(\alpha<\alpha_c)}$ on BA model ($m=3$)	100	1.7	9.6	1.04
$G_{\cup spt(\alpha<\alpha_c)}$ on $G_{0.8}(100)$	100	1.6	9.8	1.04
$G_{\cup spt(\alpha<\alpha_c)}$ on cubic lattice	125	1.5	12.8	0.92
$G_{\cup spt(\alpha<\alpha_c)}$ on square lattice	100	1.3	13.4	0.81
$G_{\cup spt(\alpha<\alpha_c)}$ on square lattice	144	1.3	16.8	0.82
k-ary tree, $1 < k < N-1$	100	1	$E[H(k)]$	$\sqrt{k-1}$

We compare class (a) and (b) in Table 8.2 and class (c) in Table 8.3. With a similar number of nodes in Table 8.2:

- The scaling exponent c seems to be negatively correlated with the $E[H]$ except for the k-ary tree.

- The scaling exponent c seems to be positively correlated with the $sdev[D]$ standard deviation of the degree except for the k-ary tree. The higher the variance of the degree is, the more traffic among links varies.

- The scaling exponent c seems to be insensitive to the size N of the tree. A same slope c is obtained for different substrate size, e.g. the k-ary tree, and the $G_{\cup spt(\alpha \to 0)}$ on top of network models as mentioned in Section 8.3.1. However, the $E[H]$ behaves as a function of N and the $sdev[D]$ can slightly depend on N with the fixed average $E[D] \approx 2$. Hence, the correlation between c and $E[H]$ as well as $sdev[D]$ may become weaker or even disappear when networks with different sizes are considered, which will be further examined for real-world networks in Table 8.3.

- The URT and the class of scale-free trees (e.g. the BA tree) discussed in Section 8.4.1, have $c \to 2.0$ for large N and $N \ll B_l \ll \left(\frac{N}{2}\right)^2$. Compared to URT, the degree of the BA tree varies more and has higher scaling exponent c (see Table 8.2), when the complete range $B_l \in \left(N - 1, \left\lfloor\frac{N}{2}\right\rfloor \left(N - \left\lfloor\frac{N}{2}\right\rfloor\right)\right]$ is taken into account.

The scaling exponent c of betweenness distribution varies from $c = 1$ for the k-ary tree to $c = 2$ for scale-free trees.

Table 8.3: Topological characteristics of overlay tree on top of real-world networks. The overlay tree of networks that are marked with a star possesses a power law degree distribution.

Table2	N	L	c	$E[H]$	$sdev[D]$
Internet As topology*	12254	25319	1.9	12.2	16
Web of Science Citations(koh)*	3704	12673	1.9	14.6	6.0
Gnutella Crawl2*	1568	1906	1.9	11.6	4.3
Science coauthorship network	379	914	1.8	14.1	1.6
Word adjacencies	112	850	1.8	7.7	1.6
Air Transportation*	2179	31326	1.7	17.9	2.8
Web of Science Citations(scimet)*	2678	10385	1.7	22.18	1.9
High Energy Collaborations	5835	13815	1.7	31.1	1.5
Dutch soccer	685	10310	1.7	22.7	1.4
Power grid	4941	6594	1.6	58.9	1.2
American football	115	613	1.5	11.8	1.0

For overlay trees on top of real networks $G(N, L)$ with N nodes and L links in Table 8.3:

- The exponent c ranges from 1.5 to 1.9, while the network size varies from $N = 112$ to $N = 12254$. The scaling exponent c does not seem to be dependent on the size N of the topology.

- The negative correlation between hopcount $E[H]$ and c disappear because $E[H]$ is positively correlated with N.

- The positive correlation between $sdev[D]$ and c still holds for most of the considered networks.

- The overlay trees possess different degree distributions as plotted in Figure 8.7 to 8.9. The overlay tree of networks that are marked with a star in Table 8.3 possesses a power law degree distribution. Hence, the betweenness distribution of scale-free networks does not necessarily follow the same power law exponent c, while a similar exponent c can be obtained in networks with different degree distributions.
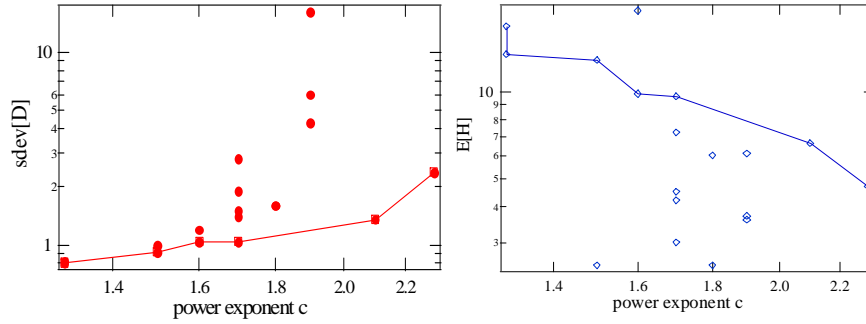
Figure 8.10: Relationship between the power exponent c of the betweenness and the standard deviation $sdev[D]$ and the average hopcount $E[H]$ of the tree.

The relationship between the $sdev[D]$ as well as the $E[H]$ and the scaling exponent c of betweenness distribution is given in Figure 8.10. Points lying on the line are for networks listed in Table 8.2 with similar topology size N. The approximately positive correlation between $sdev[D]$ and c can be observed for all the three classes. Since the average degree $E[D] = 2(N-1)/N \approx 2$ in a tree is almost constant, a higher degree variance implies more nodes with higher degree or/and more nodes with degree 1. The betweenness of links connected to a degree 1 node is always the minimum $N-1$ while the traffic passing through a high degree node is splited by links connected to this node. Both contribute to a higher scaling exponent c.

8.5   Conclusion

In this chapter, we examine the traffic in a weighted network via the link betweenness distribution of the corresponding transport overlay network $G_{\cup spt}(\alpha)$, the union of all shortest paths. In the strong disorder regime, all transport flows over the overlay tree $G_{\cup spt(\alpha \to 0)}$ =MST. Important new findings are the power law betweenness distribution specified in (8.2) of trees: tree models such as scale-free trees and k-ary trees; overlay trees on top of traditional network models; overlay trees on top of real-world complex networks. The scaling exponent $1 < c \leq 2$ for large networks is shown to be positively correlated with the $sdev[D]$ of the corresponding tree and is insensitive to the network size N. Equipped with i.i.d. link weights, the overlay tree is, in fact, a random minimum spanning tree (RMST). We conjecture that the scaling exponent c may be used to characterize these tree structures and probably the underlying topology. First, recall that any link in a tree connects two clusters with size $1 \leq |C_l| \leq \frac{N}{2}$ and $N - |C_l|$ and $B_l \sim |C_l| \cdot N$ in Section 8.4. The power law betweenness distribution implies approximately a same power law scaling for $\Pr[|C_l| = n] \sim n^{-c}$ the probability distribution of cluster size. Second, for Internet As topology, our power law scaling of betweenness with $c = 1.9$ is the same as $\Pr[S = n] \sim n^{-1.9 \pm 0.1}$ the probability of finding n points downhill [16], a signature of the intrinsic fractal properties of webs. And recently, Kitsak et al. [57] have brought fractal properties of networks into the betweenness analysis.

In the weak disorder regime, traffic flows over more links than that of the MST. The negative correlation between link weight and betweenness also depends on α, the strength of link weight disorder and the structure of the underlying topology. Both a stronger disorder in link weights and the nonhomogeneity of the substrate reduce the correlation.

Chapter 9

Conclusions

The fundamental question "how robust a network is for a given service?" pervades many fields in science: biology, physics, sociology etc., because the performance of various services or dynamic processes always depends on the network topology. We define robustness as follows, *a network is more robust if the service on the network performs better*. The performance of the service is assessed when the network is either in a conventional state or under perturbations or challenges such as failures and malicious attacks. In view of the interdisciplinary nature of this network robustness problem, we investigate it according to our general framework: (1) robustness quantification in Chapter 3 (2) optimization in Chapter 4, 5 and (3) the interplay between network and service in Chapter 6, 7 and 8.

In this final chapter of the thesis, we provide for each direction of this framework an overview of the obtained results. The descriptions in this chapter are at a higher level of abstraction than the conclusion in each chapter. A wide range of research questions may fit in this framework. Thus, we strive to depict the broad picture of network robustness, by the line of the research performed in this thesis, together with the essential ongoing work and potential extensions.

9.1 Robustness quantification

The quantification of network robustness is, actually, to associate or attribute the performance of a service to a topological metric $R(G)$. Due to the substantial diversity of services, any topological metric $R(G)$ may possibly represent certain network robustness for a service. For example, the network robustness with respect to the path stability subject to link failures is shown to be directly related to the metric hopcount. In this thesis, we focus on a spectral measure, the algebraic connectivity. Literature shows that the algebraic connectivity characterizes the network robustness regarding to (a) the synchronization of dynamic processes at nodes of a network (b) random walks on

graphs and (c) the connectivity of a network. Moreover, we illustrate that the algebraic connectivity may quantify network robustness in terms of traffic engineering. A large algebraic connectivity implies that the traffic tends to be homogeneously distributed due to the corresponding dense core and sparse border structure. The algebraic connectivity may be not a sensitive robustness measure because it is upper bounded by the minimum degree of a network. However, it seems a suitable robustness metric from the design point of view, because no network designer allow for these weak nodes with very low degree, especially when designing the core of a communications network.

Significant research effort has been made to understand the relationship between the structural properties of networks captured by certain topological measures and the nature of the dynamics or services taking place on these networks. The most challenging subject is probably the robustness quantification for human brain network, i.e. the relation between the functioning of human brain and the topological features of the brain network, the interconnections of billions of neurons.

9.2   Robustness optimization

If robustness is quantified by the algebraic connectivity, the complete graph is the most robust network. However, when the diameter of a graph is constrained to guarantee certain end-to-end quality of service, it is non-trivial to find out the graph optimizing the algebraic connectivity. Only upper bounds of the algebraic connectivity have been obtained in mathematics over the past fifteen years. We propose a class of graphs with given diameter D and N nodes, within which the largest number of links, the minimum average hopcount, and more important, the maximum of any Laplacian eigenvalue (including the algebraic connectivity) among all graphs on N nodes and diameter D can be achieved. The corresponding optimal graphs that achieve the maximum/minimum values of these measures are not necessarily the same. They can be either analytically determined or searched numerically out of the class of graphs, especially due to our reduction in the eigenvalue computation. The importance of these results is three fold (a) the maximal achievable algebraic connectivity is sharper than any upper bound proposed in literature; (b) the relation between the maximum/minimum value of a measure and the size N (or diameter D) provides insights on how to normalize this measure such that networks with different sizes (or diameters) can be compared; (c) features of the network optimizing the algebraic connectivity contribute directly to the design of robust networks.

Furthermore, we investigate how to refine a network to optimize its robustness. In particular, how to add a link to optimize the algebraic connectivity. We propose a Fiedler vector based strategy and a degree based strategy. The former performs generally better than the latter in various classes of graphs and is close to the optimal link addition in Erdös-Rényi random graphs. The latter is, however, simple to compute.

In general, it is difficult to design strategies to optimize any topological measure R based on topological features, because the correlations of topological metrics are dependent on topology. Thus, the performance of a strategy varies for different types of networks.

9.3 Interplay between the network and service

Network and service are the key determinants of the overall performance. Many services provided by communications and transportation networks are based on the shortest path routing. Link weight tuning, a mechanism to control traffic, is also considered as part of the service. We investigate the interplay between the network structure and the link weight structure in a shortest path based service in determining the following performance aspects: the structural features of the transport overlay network, the sampling bias of measuring a network topology and the traffic distribution, motivated respectively by potential applications in overlay networks, Internet topology interference and traffic engineering.

The transport overlay network $G_{\cup spt}$, the union of shortest paths between all node pairs, determines the network's performance. Overlay networks such as peer-to-peer networks or virtual private networks can be regarded as a subgraph of $G_{\cup spt}$. We show that the overlay $G_{\cup spt}$ in an Erdös-Rényi random graph $G_p(N)$ equipped with regular i.i.d. link weights or with 2-dimensional correlated uniform link weights, is a connected Erdös-Rényi random graph, which emphasizes the important role of random graphs in overlay networks. By tuning the power exponent α of polynomial link weights in different graphs, a universal phase transition in the overlay structure occurs around a critical extreme value index α_c. If $\alpha > \alpha_c$, transport in the network traverses many links whereas for $\alpha < \alpha_c$, all transport flows over a critical backbone: the Minimum Spanning Tree. The existence of a controllable phase transition in networks may allow network operators to steer and balance flows in their network.

Many network topology measurements capture or sample only a partial view of the actual network structure, which we call the underlying network. In RIPE NCC and the PlanetLab measurement architectures, the Internet is mapped as $G_{\cup_m spt}$, the union of shortest paths between each pair of a set $\mathcal{M}$ of m testboxes. The sampling bias refers to the difference between the sampled overlay $G_{\cup_m spt}$ and the overlay network $G_{\cup spt}$, the maximally observable part of a network. It depends on both the selection of the testboxes, which can be regarded as the service, and the underlying networks. We examine firstly the effect of the service, i.e. the set of testboxes, on $G_{\cup_m spt}$: (a) the interconnections of set $\mathcal{M}$ are demonstrated to be correlated with the sampled network $G_{\cup_m spt}$; (b) in order to obtain an increasingly accurate view of a given network, a higher than linear detection/measuring effort (the relative size m/N of set $\mathcal{M}$) is needed. Finally, when the relative size m/N of set $\mathcal{M}$ is small, the sampling bias tends to be small if the network possesses large average hopcount $E[H_N]$ and small link density.

Hence, a large number of testboxes placed far from each other is preferable for good network topology measurements. Moreover, placing testboxes at hubs (nodes with a high degree in the underlying network) may contribute to a small sampling bias, because links connected to the testboxes are more likely to be sampled than the other links.

The traffic distribution in the network is examined by the link betweenness distribution in the overlay network $G_{\cup spt}$. An important new finding is the power law betweenness distribution $\Pr[B_l = j] \sim j^{-c}$ followed by: tree models such as scale-free trees and k-ary trees; overlay trees $G_{\cup spt(\alpha < \alpha_c)}$ on top of traditional network models; overlay trees on top of real-world complex networks, as long as the underlying network is not as sparse as a tree. The exponent c seems to be positively correlated with the degree variance of the overlay tree and to be insensitive of the size N of a network. In the weak disorder regime, where $\alpha > \alpha_c$, we show that a link with smaller link weight tends to carry more traffic. This negative correlation between link weight and betweenness depends on α and the structure of the underlying topology.

Our investigation on the interplay between network and service, again, reveals the relation between the performance of a service and network structural features, thus, contributes to robustness quantifications. For example, the sampling bias is highly related to the average hopcount $E[H_N]$ and the link density p of a network. The exponent c of the power law betweenness distribution, or equivalently, the variance of the traffic, can be reasonably characterized by the degree variance of the overlay tree.

Finally, we would like to highlight the most important contributions of this thesis:

- The algebraic connectivity seems a suitable robustness metric from the design point of view.

- We propose a class of graphs with given diameter D and N nodes, within which, many robust features, i.e. the largest number of links, the minimum average hopcount, and the maximum of any Laplacian eigenvalue (including the algebraic connectivity) among all graphs on N nodes and diameter D, can be achieved.

- We propose a Fiedler vector based strategy and a degree based strategy to optimize the algebraic connectivity by link addition and the former performs generally better than the latter. We point out that the performance of a strategy varies for different types of networks, because the correlations of topological metrics are topology dependent.

- By tuning the power exponent α of polynomial link weights, a universal phase transition in the structure of $G_{\cup spt}$, i.e. $F_T(\alpha) = \Pr[G_{\cup spt}$ is a tree$]$, is observed in various graphs. This may allow network operators to steer and balance flows in their network.

- A power law betweenness distribution is followed by: tree models such as scale-free trees and k-ary trees; overlay trees $G_{\cup spt(\alpha < \alpha_c)}$ on top of traditional network

models; overlay trees on top of real-world complex networks, as long as the network is not as sparse as a tree.

The importance of algebraic connectivity may be far beyond the current understanding. Our clique chain structures G_D^* with N nodes and diameter D can optimize the algebraic connectivity. The optimal graphs are mostly symmetric and the cliques in the middle have a larger size than cliques close to the borders. If D is sufficiently long, the optimal structure is homogeneous and only deviations occur at the ends. These features are also observed in those long molecules in nature represented by remarkable homogeneous strings which suggest extremal properties. We may regard evolution as an optimization process over years. Then, the homogeneous string structure in proteins, DNA structures and our optimal graphs seem the results of optimizations in a same direction.

A challenge posed for robustness characterization is that the correlations between topological metrics are topology dependent. In order to precisely quantify network robustness, e.g. by multiple metrics coupled with a weighing vector, the correlations among the set of topological measures are necessarily independent of network structures. Furthermore, the comparison of networks with different sizes or number of links still remains as an open question, because the scaling of a robustness measure even in terms of the network size may depend on the properties of networks. This stresses the importance to capture universal behaviors of robustness measures and to discover generic features of complex networks.

Appendix A

Proofs

A.1 Proof of Theorem 10

The adjacency matrix of $G_D^*(n_1, n_2, ..., n_{D+1})$ is

$$
A_{G_D^*} = \begin{bmatrix}
\widetilde{J}_{n_1 \times n_1} & J_{n_1 \times n_2} & & & & & & \\
J_{n_2 \times n_1} & \widetilde{J}_{n_2 \times n_2} & J_{n_2 \times n_3} & & & & & \\
& & \ddots & & & & & \\
& & & J_{n_i \times n_{i-1}} & \widetilde{J}_{n_i \times n_i} & J_{n_i \times n_{i+1}} & & \\
& & & & & \ddots & & \\
& & & & & J_{n_D \times n_{D-1}} & \widetilde{J}_{n_D \times n_D} & J_{n_D \times n_{D+1}} \\
& & & & & & J_{n_{D+1} \times n_{D+1}} & \widetilde{J}_{n_{D+1} \times n_{D+1}}
\end{bmatrix}
$$

where $\widetilde{J} = J - I$. The eigenvalues of corresponding Laplacian $Q_{G_D^*} = \Delta_{G_D^*} - A_{G_D^*}$ are the solutions of $\det\left(Q_{G_D^*} - \mu I\right) = 0$, where

$$
\det\left(Q_{G_D^*} - \mu I\right) =
$$
$$
\begin{bmatrix}
\delta_1 I - J & -J_{n_1 \times n_2} & & & & & & \\
-J_{n_2 \times n_1} & \delta_2 I - J & -J_{n_2 \times n_3} & & & & & \\
& & \ddots & & & & & \\
& & & -J_{n_i \times n_{i-1}} & \delta_i I - J & -J_{n_i \times n_{i+1}} & & \\
& & & & & \ddots & & \\
& & & & & -J_{n_D \times n_{D-1}} & \delta_D I - J & -J_{n_D \times n_{D+1}} \\
& & & & & & -J_{n_{D+1} \times n_{D+1}} & \delta_{D+1} I - J
\end{bmatrix}
$$

where we have defined

$$
\begin{aligned}
\delta_1 &= n_1 + n_2 - \mu \\
\delta_i &= n_{i-1} + n_i + n_{i+1} - \mu \text{ for } i \in [2, D] \\
\delta_{D+1} &= n_D + n_{D+1} - \mu
\end{aligned}
$$

The dimensions of the block diagonal matrix are $n_i \times n_i$ (and omitted to make the matrix fit on the page). Clearly, the degree d_i of a node in clique i equals δ_i when $\mu = 1$. The submatrix of $Q_{G_D^*} - \mu I$ consisting of the last $D + 2 - j$ block rows and block columns is denoted by T_j; thus, T_j is the right bottom $\sum_{i=j}^{D+1} n_i \times \sum_{i=j}^{D+1} n_i$ sub-matrix of $Q_{G_D^*} - \mu I$, where $1 \leq j \leq D + 1$ and $T_1 = Q_{G_D^*} - \mu I$. Applying (A.6) yields

$$
\det T_1 = \det\left(\delta_1 I - J_{n_1 \times n_1}\right) \det\left(T_2 - \begin{bmatrix} J_{n_2 \times n_1} \\ 0_{\left(\sum_{i=3}^{i=D+1} n_i\right) \times n_1} \end{bmatrix} \left(\delta_1 I - J_{n_1 \times n_1}\right)^{-1} \begin{bmatrix} J_{n_2 \times n_1} \\ 0_{\left(\sum_{i=3}^{i=D+1} n_i\right) \times n_1} \end{bmatrix}^T\right)
$$

$$
= \det\left(\delta_1 I - J_{n_1 \times n_1}\right) \det\left(T_2 - \begin{bmatrix} J_{n_2 \times n_1}\left(\delta_1 I - J_{n_1 \times n_1}\right)^{-1} J_{n_1 \times n_2} & 0_{n_1 \times \left(\sum_{i=3}^{i=D+1} n_i\right)} \\ 0_{\left(\sum_{i=3}^{i=D+1} n_i\right) \times n_2} & 0_{\left(\sum_{i=3}^{i=D+1} n_i\right) \times \left(\sum_{i=3}^{i=D+1} n_i\right)} \end{bmatrix}\right)
$$

Using (A.2) and (A.4) results in

$$
J_{n_2 \times n_1} \left(\delta_1 I - J_{n_1 \times n_1}\right)^{-1} J_{n_2 \times n_1} = \frac{n_1}{\delta_1 - n_1} J_{n_2 \times n_2}
$$

while application of (A.5) yields

$$
\det\left(\delta_1 I - J_{n_1 \times n_1}\right) = \delta_1^{n_1 - 1}\left(\delta_1 - n_1\right)
$$

Thus, with the definition

$$
\theta_1 = \delta_1 - n_1
$$

we obtain

$$
\det T_1 = \delta_1^{n_1 - 1} \theta_1 \det \widetilde{T}_2
$$

where

$$
\widetilde{T}_2 = T_2 - \begin{bmatrix} \frac{n_1}{\theta_1} J_{n_2 \times n_2} & 0_{n_2 \times \left(\sum_{i=3}^{i=D+1} n_i\right)} \\ 0_{\left(\sum_{i=3}^{i=D+1} n_i\right) \times n_2} & 0_{\left(\sum_{i=3}^{i=D+1} n_i\right) \times \left(\sum_{i=3}^{i=D+1} n_i\right)} \end{bmatrix}
$$

We observe that the matrix at the right hand side has the same structure as that of T_2, except that the first block row and block column is now $\delta_2 I - \left(1 + \frac{n_1}{\theta_1}\right) J_{n_2 \times n_2} = $

$\delta_2 I - \frac{\delta_1}{\delta_1 - n_1} J_{n_2 \times n_2}$. We apply the same operations on

$$\det \widetilde{T}_2 = \det \left(\delta_2 I - \frac{\delta_1}{\delta_1 - n_1} J_{n_2 \times n_2} \right) \times$$

$$\det \left(T_3 - \left[\begin{array}{c} J_{n_3 \times n_2} \\ 0_{\left(\sum_{i=4}^{i=D+1} n_i \right) \times n_2} \end{array} \right] \left(\delta_2 I - \frac{\delta_1}{\delta_1 - n_1} J_{n_2 \times n_2} \right)^{-1} \left[\begin{array}{c} J_{n_3 \times n_2} \\ 0_{\left(\sum_{i=4}^{i=D+1} n_i \right) \times n_2} \end{array} \right]^T \right)$$

$$= \delta_2^{n_2 - 1} \left(\delta_2 - \frac{\delta_1 n_2}{\delta_1 - n_1} \right) \det \left(T_3 - \left[\begin{array}{cc} \frac{n_2}{\delta_2 - \frac{\delta_1 n_2}{\delta_1 - n_1}} J_{n_3 \times n_3} & 0_{n_3 \times \left(\sum_{i=4}^{i=D+1} n_i \right)} \\ 0_{\left(\sum_{i=4}^{i=D+1} n_i \right) \times n_3} & 0_{\left(\sum_{i=4}^{i=D+1} n_i \right) \times \left(\sum_{i=4}^{i=D+1} n_i \right)} \end{array} \right] \right)$$

$$= \delta_2^{n_2 - 1} \theta_2 \det \left(T_3 - \left[\begin{array}{cc} \frac{n_2}{\theta_2} J_{n_3 \times n_3} & 0_{n_3 \times \left(\sum_{i=4}^{i=D+1} n_i \right)} \\ 0_{\left(\sum_{i=4}^{i=D+1} n_i \right) \times n_3} & 0_{\left(\sum_{i=4}^{i=D+1} n_i \right) \times \left(\sum_{i=4}^{i=D+1} n_i \right)} \end{array} \right] \right)$$

where $\theta_2 = \delta_2 - \frac{\delta_1 n_2}{\delta_1 - n_1} = \delta_2 - \left(\frac{n_1}{\theta_1} + 1 \right) n_2$. Since the matrix

$$\widetilde{T}_3 = T_3 - \left[\begin{array}{cc} \frac{n_2}{\theta_2} J_{n_3 \times n_3} & 0_{n_3 \times \left(\sum_{i=4}^{i=D+1} n_i \right)} \\ 0_{\left(\sum_{i=4}^{i=D+1} n_i \right) \times n_3} & 0_{\left(\sum_{i=4}^{i=D+1} n_i \right) \times \left(\sum_{i=4}^{i=D+1} n_i \right)} \end{array} \right]$$

again possesses a similar structure, we claim that

$$\widetilde{T}_j = T_j - \left[\begin{array}{cc} \frac{n_{j-1}}{\theta_{j-1}} J_{n_j \times n_j} & 0_{n_j \times \left(\sum_{i=j+1}^{i=D+1} n_i \right)} \\ 0_{\left(\sum_{i=j+1}^{i=D+1} n_i \right) \times n_j} & 0_{\left(\sum_{i=j+1}^{i=D+1} n_i \right) \times \left(\sum_{i=j+1}^{i=D+1} n_i \right)} \end{array} \right]$$

obeys the recursion

$$\det \widetilde{T}_j = \delta_j^{n_j - 1} \theta_j \det \widetilde{T}_{j+1} \tag{A.1}$$

where θ_j is defined by the recursion (A.1) with the convention that $n_0 = 0$ and $\theta_0 = 1$, because $\theta_1 = \delta_1 - n_1$.

We have shown that (A.1) holds for $j = 1$ and $j = 2$. Assuming that (A.1) holds

for j (induction argument), we compute $\det \widetilde{T}_{j+1}$ similarly,

$$\det \widetilde{T}_{j+1} = \det \left(\delta_{j+1} I - \left(\frac{n_j}{\theta_j} + 1 \right) J_{n_{j+1} \times n_{j+1}} \right) \times$$

$$\det \left(T_{j+2} - \left[\begin{array}{c} J_{n_{j+2} \times n_{j+1}} \\ 0_{\left(\sum_{i=j+3}^{i=D+1} n_i \right) \times n_{j+1}} \end{array} \right] \left(\delta_{j+1} I - \left(\frac{n_j}{\theta_j} + 1 \right) J_{n_{j+1} \times n_{j+1}} \right)^{-1} \times \left[\begin{array}{c} J_{n_{j+2} \times n_{j+1}} \\ 0_{\left(\sum_{i=j+3}^{i=D+1} n_i \right) \times n_{j+1}} \end{array} \right]^T \right)$$

$$= \delta_{j+1}^{n_j} \left(\delta_{j+1} - \left(\frac{n_j}{\theta_j} + 1 \right) n_{j+1} \right)$$

$$\times \det \left(T_{j+2} - \left[\begin{array}{cc} \frac{n_{j+1}}{\delta_{j+1} - \left(\frac{n_j}{\theta_j} + 1 \right) n_{j+1}} J_{n_{j+2} \times n_{j+2}} & 0_{n_{j+2} \times \left(\sum_{i=j+3}^{i=D+1} n_i \right)} \\ 0_{\left(\sum_{i=j+3}^{i=D+1} n_i \right) \times n_{j+2}} & 0_{\left(\sum_{i=j+3}^{i=D+1} n_i \right) \times \left(\sum_{i=j+3}^{i=D+1} n_i \right)} \end{array} \right] \right)$$

$$= \delta_{j+1}^{n_j} \theta_{j+1} \det \left(T_{j+2} - \left[\begin{array}{cc} \frac{n_{j+1}}{\theta_{j+1}} J_{n_{j+2} \times n_{j+2}} & 0_{n_{j+2} \times \left(\sum_{i=j+3}^{i=D+1} n_i \right)} \\ 0_{\left(\sum_{i=j+3}^{i=D+1} n_i \right) \times n_{j+2}} & 0_{\left(\sum_{i=j+3}^{i=D+1} n_i \right) \times \left(\sum_{i=j+3}^{i=D+1} n_i \right)} \end{array} \right] \right)$$

where $\theta_{j+1} = \delta_{j+1} - \left(\frac{n_j}{\theta_j} + 1 \right) n_{j+1}$. Hence, (A.1) holds also for $j+1$, and by induction for any $1 \leq j \leq D + 1$. But,

$$\widetilde{T}_{D+1} = T_{D+1} - \frac{n_D}{\theta_D} J_{n_{D+1} \times n_{D+1}}$$

$$= \delta_{D+1} I - \left(1 + \frac{n_D}{\theta_D} \right) J_{n_{D+1} \times n_{D+1}}$$

such that, with (A.5),

$$\det \widetilde{T}_{D+1} = \delta_{D+1}^{n_{D+1}-1} \left(\delta_{D+1} - \left(1 + \frac{n_D}{\theta_D} \right) n_{D+1} \right) = \delta_{D+1}^{n_{D+1}-1} \theta_{D+1}$$

Iterating (A.1) back finally yields

$$\det \left(Q_{G_D^*} - \mu I \right) = \prod_{j=1}^{D+1} \delta_j^{n_j-1} \prod_{j=1}^{D+1} \theta_j$$

which is (4.1).

A.2 Results from linear algebra

If

$$X_{m \times m} = (J - (\lambda + 1) I)_{m \times m} \tag{A.2}$$

then the inverse matrix of X is

$$X^{-1} = -\frac{1}{(\lambda+1)(\lambda+1-m)}\left(J + (\lambda+1-m)I\right)_{m \times m} \tag{A.3}$$

We now compute

$$Y = J_{(N-m) \times m} X^{-1}_{m \times m} J_{m \times (N-m)}$$
$$= -\frac{1}{(\lambda+1)(\lambda+1-m)} J_{(N-m) \times m}\left(J_{m \times m} + (\lambda+1-m)I_{m \times m}\right)J_{m \times (N-m)}$$

Using $J_{k \times n} J_{n \times l} = n J_{k \times l}$ gives

$$Y = -\frac{1}{(\lambda+1)(\lambda+1-m)}\left(m J_{(N-m) \times m} + (\lambda+1-m)J_{(N-m) \times m}\right)J_{m \times (N-m)}$$
$$= -\frac{1}{(\lambda+1)(\lambda+1-m)}\left(m^2 J_{(N-m) \times (N-m)} + m(\lambda+1-m)J_{(N-m) \times (N-m)}\right)$$

hence

$$Y = -\frac{m}{(\lambda+1-m)} J_{(N-m) \times (N-m)} \tag{A.4}$$

Finally, it is shown in [67, p. 481] that,

$$\det(J - xI)_{n \times n} = (-1)^n x^{n-1}(x-n) \tag{A.5}$$

and we will need [65]

$$\det \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \det A \det\left(D - CA^{-1}B\right) \tag{A.6}$$

where $D - CA^{-1}B$ is called the Schur complement of A.

A.3 Proof of Theorem 17

By the law of total probability, we can write

$$\Pr[P^*_{i \to j} = l\,(i \to j)] = \int_0^\infty f_{w(i \to j)}(x) \times$$
$$\Pr[P^*_{i \to j} = l(i \to j)|w(i \to j) = x]dx \tag{A.7}$$

where $f_{w(i \to j)}(x)$ is the pdf of the weight of a link in the graph.

Let us assume that $P^*_{i \to j,h>1}$ is the shortest path among paths with more than 1 hop and let us denote the path weight by $w(P^*_{i \to j,h>1})$. Provided that the direct link between

i and j exists, then the shortest path equals the direct link if its weight is smaller than any path with more than one hop and vice versa,

$$P^*_{i \to j} = P^*_{i \to j, h>1} \cdot 1_{\{w(i \to j) \geq w(P^*_{i \to j, h>1})\}}$$
$$+ l(i \to j) \cdot 1_{\{w(i \to j) < w(P^*_{i \to j, h>1})\}}$$

from which the conditional probability

$$\lim_{\Delta x \to 0} \Pr[P^*_{i \to j} = l(i \to j)|\{x \leq w(i \to j) \leq x + \Delta x\}$$
$$\cap \{l(i \to j) \text{ exists}\}] = \Pr[w(P^*_{i \to j, h>1}) > x]$$

is immediate. Since the link weights are i.i.d. and also independent of a specific link, the conditional probability P_c is

$$P_c = \Pr[P^*_{i \to j} = l(i \to j)|\{x \leq w(i \to j) \leq x + \Delta x\}$$
$$\cap \{l(i \to j) \text{ exists}\}]$$
$$= \frac{\Pr[P^*_{i \to j} = l(i \to j)|\{x \leq w(i \to j) \leq x + \Delta x\}]}{\Pr[l(i \to j) \text{ exists}]} \tag{A.8}$$

where the last step follows from the fact that the event $\{P^*_{i \to j} = i \to j\}$ is contained in the event $\{l(i \to j) \text{ exists}\}$. Hence,

$$\lim_{\Delta x \to 0} \Pr\left[P^*_{i \to j} = l(i \to j)|\{x \leq w(i \to j) \leq x + \Delta x\}\right]$$
$$= \Pr\left[l(i \to j) \text{ exists}\right] \Pr[w(P^*_{i \to j, h>1}) > x] \tag{A.9}$$

The sample space Ω consists of four mutually exclusive events:

$$\{w(P^*_{i \to j, h>1}) \leq x, w(i \to j) \leq x \cap l(i \to j) \text{ exists}\}$$
$$\{w(P^*_{i \to j, h>1}) > x, w(i \to j) \leq x \cap l(i \to j) \text{ exists}\}$$
$$\{\{w(i \to j) > x \cap l(i \to j) \text{ exists}\} \cup$$
$$\{l(i \to j) \text{ does not exist}\}, w(P^*_{i \to j, h>1}) \leq x\}$$
$$\{\{w(i \to j) > x \cap l(i \to j) \text{ exists}\} \cup$$
$$\{l(i \to j) \text{ does not exist}\}, w(P^*_{i \to j, h>1}) > x\}$$

Therefore, the related probability measure is

$$\Pr[w(P^*_{i \to j}) \leq x]$$
$$= \Pr[w(P^*_{i \to j, h>1}) \leq x, \{\{w(i \to j) > x \cap l(i \to j) \text{ exists}\}$$
$$\cup \{l(i \to j) \text{ does not exist}\}]$$
$$+ \Pr[w(P^*_{i \to j, h>1}) \leq x, w(i \to j) \leq x \cap l(i \to j) \text{ exists}]$$
$$+ \Pr[w(P^*_{i \to j, h>1}) > x, w(i \to j) \leq x \cap l(i \to j) \text{ exists}]$$
$$= \Pr[w(P^*_{i \to j, h>1}) \leq x, \{\{w(i \to j) > x \cap l(i \to j) \text{ exists}\}$$
$$\cup \{l(i \to j) \text{ does not exist}\}]$$
$$+ \Pr[w(i \to j) \leq x] \Pr[l(i \to j) \text{ exists}] \tag{A.10}$$

where in the last step, we have again used the law of total probability. Further, the event

$$
\begin{aligned}
&\{\{w(i \to j) > x \cap l(i \to j) \text{ exists}\} \\
&\quad \cup \{l(i \to j) \text{ does not exist}\}, w(P^*_{i \to j, h>1}) \le x\} \\
&= \{w(P^*_{i \to j, h>1}) \le x\} \cap \{\{w(i \to j) > x \cap l(i \to j) \text{ exists}\} \\
&\quad \cup \{l(i \to j) \text{ does not exist}\}\}
\end{aligned}
$$

and the first two events are independent because link weights are independently and identically distributed and $l(i \to j)$ is different from $P^*_{i \to j, h>1}$. Hence, (A.10) reduces to

$$
\begin{aligned}
&\Pr[w(P^*_{i \to j}) \le x] \\
&= (\Pr[w(i \to j) > x] \cdot \Pr[l(i \to j) \text{ exists}] \\
&\quad + \Pr[l(i \to j) \text{ does not exist}]) \cdot \Pr[w(P^*_{i \to j, h>1}) \le x] \\
&\quad + \Pr[w(i \to j) \le x] \Pr[l(i \to j) \text{ exists}]
\end{aligned}
$$

from which

$$
\begin{aligned}
&\Pr[w(P^*_{i \to j, h>1}) \le x] \\
&= \frac{\Pr[w(P^*_{i \to j}) \le x] - \Pr[w(i \to j) \le x] \cdot p_{ij}}{\Pr[w(i \to j) > x] \cdot p_{ij} + 1 - p_{ij}}
\end{aligned}
$$

where $p_{ij} = \Pr[l(i \to j) \text{ exists}]$. Finally, the conditional probability (A.9) becomes

$$
\begin{aligned}
&\lim_{\Delta x \to 0} \Pr\left[P^*_{i \to j} = l(i \to j) | \{x \le w(i \to j) \le x + \Delta x\}\right] \\
&= \frac{p_{ij} \cdot \Pr[w(P^*_{i \to j}) > x]}{\Pr[w(i \to j) > x] \cdot p_{ij} + 1 - p_{ij}}
\end{aligned}
\tag{A.11}
$$

and substituted into (A.7) leads to (6.1). $\qquad\square$

A.4 Proof of Corollary 18

The event $\{P^*_{i \to j} = l(i \to j)\}$ is equivalent to the event $\{H_N = 1\}$ because the direct link corresponds to a one hop shortest path. The second equality is demonstrated as follows. Since $\frac{d}{dx} \Pr[w(i \to j) > x] = -f_{w(i \to j)}(x)$, we can write (6.1) as

$$
\begin{aligned}
&\Pr[P^*_{i \to j} = l(i \to j)] \\
&= -\int_0^\infty \frac{\Pr[w(P^*_{i \to j}) > x]}{\Pr[w(i \to j) > x] + A} d(\Pr[w(i \to j) > x] + A)
\end{aligned}
$$

where $A = \frac{1-p_{ij}}{p_{ij}}$. Partial integration yields,

$$\Pr[P^*_{i\to j} = l(i \to j)]$$
$$= -\Pr[w(P^*_{i\to j}) > x]\log(\Pr[w(i \to j) > x] + A)\big|^{x=\infty}_{x=0}$$
$$-\int_0^\infty f_{w(P^*_{i\to j})}(x)\log(\Pr[w(i \to j) > x] + A)dx$$

The limit $x \to 0$ gives $\log(1 + A) = -\log p_{ij}$ and it remains to show that the limit $x \to \infty$ vanishes. Since for any x and any probability distribution w holds that

$$-\log(\Pr[w > x] + A) = \int_0^x \frac{f_w(u)\,du}{\Pr[w > u] + A}$$
$$\leq \frac{1}{\Pr[w > x] + A}\int_0^x f_w(u)\,du$$
$$= \frac{\Pr[w \leq x]}{\Pr[w > x] + A}$$

we observe that

$$-\log(\Pr[w(i \to j) > x] + A) \leq \frac{1 + A}{\Pr[w(i \to j) > x] + A} - 1$$

such that

$$\Pr[w(P^*_{i\to j}) > x]\log\Pr[w(i \to j) > x] \cdot (-1)$$
$$\leq \frac{(1 + A)\Pr[w(P^*_{i\to j}) > x]}{\Pr[w(i \to j) > x] + A} - \Pr[w(P^*_{i\to j}) > x]$$

Using (A.11), we have

$$-\Pr[w(P^*_{i\to j}) > x]\log\Pr[w(i \to j) > x]$$
$$\leq \lim_{\Delta x \to 0}\Pr\left[P^*_{i\to j} = l(i \to j)|\{x \leq w(i \to j) \leq x + \Delta x\}\right] \cdot$$
$$(1 + A) - \Pr[w(P^*_{i\to j}) > x]$$

For $x \to \infty$, the both probabilities at right hand side tend to zero. Hence,

$$\Pr[P^*_{i\to j} = l(i \to j)]$$
$$= -\int_0^\infty f_{w(P^*_{i\to j})}(x)\log\left(-\Pr[w(i \to j) \leq x] + \frac{1}{p_{ij}}\right)dx$$
$$-\log p_{ij}$$

After writing

$$\log\left(-\Pr[w(i \to j) \leq x] + \frac{1}{p_{ij}}\right)$$
$$= -\log p_{ij} + \log\left(1 - p_{ij}F_w(x)\right)$$

we arrive at (6.3). □

A.5 Asymptotic uncorrelation of links in $G_{\cup spt}$

Lemma 25 *All first hop links in a same URT are independent.*

Proof: see [67, p. 371]. □

This independence in the URT is only true for the first hop nodes, and not for higher hop nodes since the latter depend on the specific structure of the URT.

Lemma 26 *Two links in a same URT, of which only one is a first hop link, are asymptotically (for large N) uncorrelated.*

Proof. Drmota and Hwang [35] have, for large N, computed the asymptotic correlation coefficient $\rho\left(X_N^{(k)}, X_N^{(j)}\right)$ of the number of nodes $X_N^{(k)}$ at hopcount k (called the k-th level set of the URT) from the root in a URT, based on the exact probability generating function $E\left[x^{X_N^{(k)}} y^{X_N^{(j)}}\right]$ derived by van der Hofstad *et al.* in [93]. For large N and small hopcounts $k = o\left(\log N\right)$ and $k = o\left(j\right)$ where j can range over all levels, $\rho\left(X_N^{(k)}, X_N^{(j)}\right)$ tends to zero, which implies that the level set k and j are asymptotically uncorrelated[1]. Since there is a one-to-one correspondence between nodes and links in a tree because each node (apart from the root) has precisely one ancestor, the correlation between nodes transfers to a correlation between links. Hence, *all* higher hop links in the URT are asymptotically independent from the first hop links. This proves the Lemma. ■

The proof actually demonstrates more than necessary: instead of uncorrelation between two links, it shows uncorrelation between all higher hop links.

Lemma 27 *Two different arbitrary links in the overlay $G_{\cup spt}$ on top of the complete graph K_N are asymptotically (for large N) pairwise uncorrelated.*

---

[1]The correlation coefficient $\left|\rho\left(X_N^{(k)}, X_N^{(j)}\right)\right| \to 1$ if $k = O\left(\log N\right)$ and $j = O\left(\log N\right)$, implying that levels around the average hopcount $E\left[H_N\right] \sim \log N$ (containing most of the nodes) are strongly correlated, and this is mainly a consequence of the growth rule of the URT [?, Sec. 16.2.2.] and of the "conservation of nodes over the levels", $\sum_{k=0}^{N-1} X_N^{(k)} = N$. Each URT of $G_{\cup spt}$ thus contains highly correlated level sets, but the individual links (not paths) in $G_{\cup spt}$ seem far less correlated as suggested by Conjecture 23.

Proof. We denote two arbitrary links of $G_{\cup spt}$ by $l_1 = l\,(n_1 \to x)$ and $l_2 = l\,(n_2 \to y)$ where node $x \neq y$. We distinguish between two cases, $n_1 = n_2$ and $n_1 \neq n_2$.

If $n_1 = n_2$, then both l_1 and l_2 are first hop links in the same URT and, by Lemma 25, independent.

If $n_1 \neq n_2$, the links l_1 and l_2 do not share a common node and there are two cases: (1) l_2 does not belong to the URT rooted at n_1 (and vice versa), in which case l_1 and l_2 are either independent or at most asymptotically uncorrelated (see proof Lemma 26) because both links l_1 and l_2 may appear in a URT rooted at another node n_3. (2) $l_2 \in URT_{n_1}$, the URT rooted at node n_1. Lemma 26 then shows that l_2 and l_1 are asymptotically independent. ∎

Recall that pairwise uncorrelation is weaker than pairwise independence, which in turn does not necessarily imply independence.

A.6 Proof of extreme cases of conjecture 24

To simplify the proof, instead of $D_{\mathcal{M}}$, we use $D_N(m)$ to denote the degree of set $\mathcal{M}$ in the overlay $G_{\cup_m spt}$, where N is the number of nodes in the underlying graph and m is the number of testboxes.

A.6.1 Proof of the Corollary for $k = 1$

Firstly, we prove the conjecture for $\Pr[D_N(m) = 1]$. Van der Hofstad et al. [29] have shown that $p_n\,(i) = \frac{n-i}{ni}$ is the probability that the paths from the root to i uniformly chosen nodes *that may include the root* in a URT of size n share a common link. If one of the i nodes equals the root, there is no link in common because there is no path from the root to itself. Denote by $A_{\text{No Root}}$ the event that the paths from the root to m uniformly chosen nodes *that do not include the root* in a URT of size n share a common link and by A_{Root} the event that the paths from the root to m uniformly chosen nodes *that may include the root* in a URT of size n share a common link. The probability that the root is one of the m nodes is $\Pr[\text{root}] = \frac{m}{n}$. Then

$$\Pr\,[A_{\text{No Root}}] = \Pr\,[A_{\text{Root}}|\text{No root}] = \frac{\Pr\,[A_{\text{Root}} \cap \{\text{No root}\}]}{\Pr\,[\text{No root}]}$$

If one of the m nodes is the root, there is no link in common. That event is not included in A_{Root}, which means that

$$\Pr\,[A_{\text{Root}} \cap \{\text{No root}\}] = \Pr\,[A_{\text{Root}}] = p_n\,(m)$$

and that

$$\Pr\,[A_{\text{No Root}}] = \frac{p_n\,(m)}{1 - \frac{m}{n}} = \frac{\frac{n-m}{n \cdot m}}{1 - \frac{m}{n}} = \frac{1}{m} = p_n^*\,(m)$$

Finally, we arrive at $p_n^*(m)$, the probability that the paths from the root to m uniformly chosen nodes *that do not include the root* in a URT of size n share a common link. If these paths share a link, then the number of links connected to the root and traversed by these paths must be one. Therefore, the probability $\Pr[D = 1]$ of the set $\mathcal{M}$ in a underlying graph with N nodes is

$$\Pr[D_N(m) = 1] = p_N^*(m-1) = \frac{1}{m-1}$$

In the *URT* with m nodes, according to (6.11) the probability $\Pr[D_{G_{\cup spt}} = 1] = \frac{1}{m-1}$, which explain the match of the first node in Figure 7.4. $\square$

A.6.2 Proof of the Corollary for $k = m - 1$

The extreme case $\Pr[D_N(m) = m - 1]$ is proved by using the URTs separation theorem [67, Theorem 16.2.1] and considering Figure 18.3 in [67]. A URT of size N can be separated in a URT T_1 of size k and a URT T_2 of size $N - k$ that incorporates the root (see Figure 18.3 in [67, Theorem 16.2.1]). The maximum degree of the root is achieved in two cases: (a) there is precisely 1 node of $\mathcal{M}$ in T_1 and $m - 2$ in T_2 or (b) there is none in T_1 and all $m - 1$ are in T_2. If there is more than 1 node of $\mathcal{M}$ in T_1, the degree of the root $D_N(m)$ is smaller than $m - 1$, because we need to have $m - 1$ separate clusters attached to the root that each contain precisely one node of $\mathcal{M}$. Thus,

$$\Pr[D_N(m) = m - 1] = \sum_{k=1}^{N-1} \Pr[D_{N-k}(m-1) = m - 2] \frac{\binom{k}{1}\binom{N-k-1}{m-2}}{\binom{N-1}{m-1}} \Pr[T_1 = k] +$$
$$+ \sum_{k=1}^{N-1} \Pr[D_{N-k}(m) = m - 1] \frac{\binom{k}{0}\binom{N-k-1}{m-1}}{\binom{N-1}{m-1}} \Pr[T_1 = k]$$

because the number of ways to distribute $m-1$ nodes over $N-1$ places that are different from the root such that there is 1 of the m in T_1 and the other $m-2$ in T_2 is $\binom{k}{1}\binom{N-k-1}{m-2}$ and there are $\binom{N-1}{m-1}$ ways to distribute $m - 1$ nodes over $N - 1$ places. Further, the URTs separation theorem implies that $\Pr[T_1 = k] = \frac{1}{N-1}$. This gives the recursion,

$$\Pr[D_N(m) = m - 1] = \frac{1}{(N-1)\binom{N-1}{m-1}} \sum_{k=1}^{N-1} \left\{ k \Pr[D_{N-k}(m-1) = m - 2] \binom{N-k-1}{m-2} \right.$$
$$\left. + \Pr[D_{N-k}(m) = m - 1] \binom{N-k-1}{m-1} \right\}$$
$$= \frac{1}{(N-1)\binom{N-1}{m-1}} \sum_{q=m-1}^{N-1} \left\{ (N-q)\Pr[D_q(m-1) = m - 2] \binom{q-1}{m-2} \right.$$
$$\left. + \Pr[D_q(m) = m - 1] \binom{q-1}{m-1} \right\}$$

where, in the last line, we have incorporated that $\Pr\left[D_q\left(m-1\right)=m-2\right]=0$ if $q<m-1$. From (6.11), the initial condition is $\Pr\left[D_m\left(m\right)=m-1\right]=\frac{1}{(m-1)!}$.

Further,

$$
(N-1)\binom{N-1}{m-1}\Pr\left[D_N\left(m\right)=m-1\right]=(N-1)\sum_{q=m-1}^{N-1}\Pr\left[D_q\left(m-1\right)=m-2\right]\binom{q-1}{m-2}
$$

$$
+\sum_{q=m-1}^{N-1}\left\{\Pr\left[D_q\left(m\right)=m-1\right]\binom{q-1}{m-1}\right.
$$

$$
\left.-(q-1)\Pr\left[D_q\left(m-1\right)=m-2\right]\binom{q-1}{m-2}\right\}
$$

After substitution of $N\to N+1$ in the above and subtracting the above yields, for the left-hand side,

$$
L=N\binom{N}{m-1}\Pr\left[D_{N+1}\left(m\right)=m-1\right]-(N-1)\binom{N-1}{m-1}\Pr\left[D_N\left(m\right)=m-1\right]
$$

and the right-hand side

$$
R=Q+\binom{N-1}{m-1}\Pr\left[D_N\left(m\right)=m-1\right]
$$

$$
-(N-1)\binom{N-1}{m-2}\Pr\left[D_N\left(m-1\right)=m-2\right]
$$

with

$$
Q=N\sum_{q=m-1}^{N}\Pr\left[D_q\left(m-1\right)=m-2\right]\binom{q-1}{m-2}-(N-1)\sum_{q=m-1}^{N-1}\Pr\left[D_q\left(m-1\right)=m-2\right]\binom{q-1}{m-2}
$$

$$
=N\left[\sum_{q=m-1}^{N}\Pr\left[D_q\left(m-1\right)=m-2\right]\binom{q-1}{m-2}-\sum_{q=m-1}^{N-1}\Pr\left[D_q\left(m-1\right)=m-2\right]\binom{q-1}{m-2}\right]
$$

$$
+\sum_{q=m-1}^{N-1}\Pr\left[D_q\left(m-1\right)=m-2\right]\binom{q-1}{m-2}
$$

$$
=N\Pr\left[D_N\left(m-1\right)=m-2\right]\binom{N-1}{m-2}+\sum_{q=m-1}^{N-1}\Pr\left[D_q\left(m-1\right)=m-2\right]\binom{q-1}{m-2}
$$

Simplified,

$$
\begin{aligned}
L\&R &= N \binom{N}{m-1} \Pr\left[D_{N+1}\left(m\right) = m-1\right] - N \binom{N-1}{m-1} \Pr\left[D_N\left(m\right) = m-1\right] \\
&= \binom{N-1}{m-2} \Pr\left[D_N\left(m-1\right) = m-2\right] \\
&\quad + \sum_{q=m-1}^{N-1} \Pr\left[D_q\left(m-1\right) = m-2\right] \binom{q-1}{m-2}
\end{aligned}
$$

Repeating the same procedure to remove the last remaining sum gives, for the left hand side,

$$
\begin{aligned}
L &= (N+1)\binom{N+1}{m-1}\Pr\left[D_{N+2}\left(m\right)=m-1\right] - (N+1)\binom{N}{m-1}\Pr\left[D_{N+1}\left(m\right)=m-1\right] \\
&\quad - N\binom{N}{m-1}\Pr\left[D_{N+1}\left(m\right)=m-1\right] + N\binom{N-1}{m-1}\Pr\left[D_N\left(m\right)=m-1\right] \\
&= (N+1)\binom{N+1}{m-1}\Pr\left[D_{N+2}\left(m\right)=m-1\right] - (2N+1)\binom{N}{m-1}\Pr\left[D_{N+1}\left(m\right)=m-1\right] \\
&\quad + N\binom{N-1}{m-1}\Pr\left[D_N\left(m\right)=m-1\right]
\end{aligned}
$$

The right hand side becomes,

$$
\begin{aligned}
R &= \binom{N}{m-2}\Pr\left[D_{N+1}\left(m-1\right)=m-2\right] - \binom{N-1}{m-2}\Pr\left[D_N\left(m-1\right)=m-2\right] \\
&\quad + \sum_{q=m-1}^{N}\Pr\left[D_q\left(m-1\right)=m-2\right]\binom{q-1}{m-2} - \sum_{q=m-1}^{N-1}\Pr\left[D_q\left(m-1\right)=m-2\right]\binom{q-1}{m-2} \\
&= \binom{N}{m-2}\Pr\left[D_{N+1}\left(m-1\right)=m-2\right] - \binom{N-1}{m-2}\Pr\left[D_N\left(m-1\right)=m-2\right] \\
&\quad + \binom{N-1}{m-2}\Pr\left[D_N\left(m-1\right)=m-2\right] \\
&= \binom{N}{m-2}\Pr\left[D_{N+1}\left(m-1\right)=m-2\right]
\end{aligned}
$$

Combining both sides gives,

$$\binom{N}{m-2} \Pr\left[D_{N+1}\left(m-1\right) = m-2\right] = (N+1)\binom{N+1}{m-1}\Pr\left[D_{N+2}\left(m\right) = m-1\right]$$

$$-\left(2N+1\right)\binom{N}{m-1}\Pr\left[D_{N+1}\left(m\right) = m-1\right]$$

$$+N\binom{N-1}{m-1}\Pr\left[D_{N}\left(m\right) = m-1\right]$$

By defining

$$r\left[N,m\right] = \binom{N-1}{m-1}\Pr\left[D_{N}\left(m\right) = m-1\right]$$

we arrive at the recursion,

$$r\left[N+1,m-1\right] = (N+1)\,r\left[N+2,m\right] - (2N+1)\,r\left[N+1,m\right] + N r\left[N,m\right] \quad \text{(A.12)}$$

with initial condition

$$r\left[m,m\right] = \frac{1}{(m-1)!}$$

What we claim is that $\Pr\left[D_{N}\left(m\right) = m-1\right] = \Pr\left[D_{m}\left(m\right) = m-1\right]$ for all N, which means that

$$r\left[N,m\right] = \binom{N-1}{m-1}\Pr\left[D_{m}\left(m\right) = m-1\right] = \binom{N-1}{m-1}r\left[m,m\right] = \binom{N-1}{m-1}r\left[m,m\right]$$

Introduced in (A.12) gives

$$\binom{N}{m-2}r\left[m-1,m-1\right] = (N+1)\binom{N+1}{m-1}r\left[m,m\right] - (2N+1)\binom{N}{m-1}r\left[m,m\right]$$

$$+N\binom{N-1}{m-1}r\left[m,m\right]$$

or

$$\binom{N}{m-2}(m-1) = (N+1)\binom{N+1}{m-1} - (2N+1)\binom{N}{m-1} + N\binom{N-1}{m-1}$$

The relation is, indeed, an identity. $\square$

A.7   Link betweenness distribution of URT

A URT [67] of size N is a random tree rooted at some node A. At each stage a node is attached uniformly to one of the existing nodes until the total number of nodes is

equal to N. When the $j - th$ node is attached, the corresponding $j - th$ attached link is also added except that no link is added when we start from the root or the $1st$ node. In a tree, the traffic traverses the link if and only if the source and destination lies in different clusters separated by this link. In a URT, we define $\left|\mathcal{T}_j^{(N)}\right|$ as the size of the subtree rooted at the $j - th$ attached node. The removal of the $j - th$ ($2 \leq j \leq N$) attached link will separate the graph into two clusters with size $\left|\mathcal{T}_j^{(N)}\right|$ and $N - \left|\mathcal{T}_j^{(N)}\right|$. Correspondingly, the betweenness of the $j - th$ ($2 \leq j \leq N$) attached link is $\left|\mathcal{T}_j^{(N)}\right| \cdot \left(N - \left|\mathcal{T}_j^{(N)}\right|\right)$. The probability distribution of the size of the subtree [29] equals:

$$\Pr\left[\left|\mathcal{T}_j^{(N)}\right| = k\right] = \frac{(j-1)(N-j)!(N-k-1)!}{(N-1)!(N-j-k+1)!} = \frac{\binom{N-k-1}{j-2}}{\binom{N-1}{j-1}} \tag{A.13}$$

Using the law of total probability [67], we have for the URT that

$$\Pr\left[B_l = k(N-k)\right] = \sum_{j=2}^{N} \Pr\left[B_l = k(N-k)\,|l=j\right]\Pr\left[l=j\right], \quad 1 \leq k \leq \left\lfloor \frac{N}{2} \right\rfloor$$

A random link l is the j-th attached link or attaches the j-th node to the URT with probability $\Pr\left[l=j\right] = \frac{1}{N-1}$.

For $k \in \left[1, \left\lfloor \frac{N}{2} \right\rfloor\right]$ and $k \neq \frac{N}{2}$, the conditional probability

$$\Pr\left[B_l = k(N-k)\,|l=j\right] = \Pr\left[\left|\mathcal{T}_j^{(N)}\right| = k\right] + \Pr\left[\left|\mathcal{T}_j^{(N)}\right| = N-k\right]$$

because only if the size of the subtree rooted at node j is of size $\left|\mathcal{T}_j^{(N)}\right| = k$ or of size $\left|\mathcal{T}_j^{(N)}\right| = N-k$, the betweenness of the link $l = j$ equals $k(N-k)$. Combining both yields

$$\Pr[B_l = k(N-k)] = \frac{1}{N-1}\sum_{j=2}^{N}\Pr\left[\left|\mathcal{T}_j^{(N)}\right| = k\right] + \Pr\left[\left|\mathcal{T}_j^{(N)}\right| = N-k\right]$$

Substituting (A.13) gives

$$\Pr[B_l = k(N-k)] = \frac{(N-k-1)!}{(N-1)(N-1)!}\sum_{j=2}^{N}\frac{(j-1)(N-j)!}{(N-j-k+1)!} + \frac{(k-1)!}{(N-1)(N-1)!}\sum_{j=2}^{N}\frac{(j-1)(N-j)!}{(k+1-j)!}$$

We use the identity

$$\sum_{j=n}^{m} j\binom{a-j}{b-j} = n\binom{a+1-n}{b-n} + \binom{a+1-n}{b-1-n} - m\binom{a-m}{b-1-m} - \binom{a+1-m}{b-1-m} \tag{A.14}$$

and obtain

$$\sum_{j=2}^{N} \frac{(j-1)(N-j)!}{(N-j-k+1)!} = (k-1)! \sum_{j=1}^{N-1} j \binom{N-1-j}{N-k-j}$$

$$= (k-1)! \left(\binom{N-1}{N-k-1} + \binom{N-1}{N-k-2} \right) = (k-1)! \binom{N}{N-k-1}$$

$$= \frac{(k-1)!N!}{(k+1)!(N-k-1)!}$$

Similarly,

$$\sum_{j=2}^{N} \frac{(j-1)(N-j)!}{(k+1-j)!} = (N-1-k)! \sum_{j=1}^{N-1} j \binom{N-1-j}{k-j}$$

$$= (N-1-k)! \left(\binom{N-1}{k-1} + \binom{N-1}{k-2} \right) = (N-1-k)! \binom{N}{k-1}$$

$$= \frac{(N-1-k)!N!}{(k-1)!(N-k+1)!}$$

Hence,

$$\Pr\left[B_l = k(N-k)\right] = \frac{(N-k-1)!}{(N-1)(N-1)!} \frac{(k-1)!N!}{(k+1)!(N-k-1)!}$$

$$+ \frac{(k-1)!}{(N-1)(N-1)!} \frac{(N-1-k)!N!}{(k-1)!(N-k+1)!}$$

$$= \frac{N}{(N-1)} \left(\frac{1}{(k+1)k} + \frac{1}{(N-k+1)(N-k)} \right)$$

$$= \frac{N}{(N-1)\,k(N-k)} \left(\frac{N-k}{k+1} + \frac{k}{N-k+1} \right)$$

While for $k = \left\lfloor \frac{N}{2} \right\rfloor = \frac{N}{2}$, the probability has to be halved,

$$\Pr\left[B_l = k(N-k)\right] = \frac{N}{(N-1)\,k(k+1)}$$

Hence,

$$\Pr\left[B_l = k(N-k)\right] = \begin{cases} \frac{N}{(N-1)k(N-k)} \left(\frac{N-k}{k+1} + \frac{k}{N-k+1} \right), & k \in \left[1, \left\lfloor \frac{N}{2} \right\rfloor\right] \text{ and } k \neq \frac{N}{2} \\ \frac{N}{(N-1)k(k+1)}, & \text{if } k = \left\lfloor \frac{N}{2} \right\rfloor = \frac{N}{2} \end{cases}$$

$$(A.15)$$

A.8 Link betweenness distribution of a *k*-ary tree

If the link betweenness distribution (8.3) of a k-ary tree follows a power law of the form $y = c_0 x^c$, then for any two points (x_1, y_1) and (x_2, y_2) on this curve, we have $\frac{y_1}{y_2} = \left(\frac{x_1}{x_2}\right)^c$. Two nodes are selected: $\left(N(d) - 1, \frac{k^d}{N(d)-1}\right)$ corresponding to $j = d$ in (8.3) and a random node $\left(N(d-j)(N(d) - N(d-j)), \frac{k^j}{N(d)-1}\right)$.

$$\frac{N(d-j)(N(d) - N(d-j))}{N(d) - 1} = \frac{\left(k^{d-j+1} - 1\right)\left(k^{d+1} - k^{d-j+1}\right)}{(k^{d+1} - 1 - k + 1)(k - 1)} = \frac{\left(k^{d-j+1} - 1\right)\left(k^{d+1} - k^{d-j+1}\right)}{k\left(k^d - 1\right)(k - 1)}$$

$$= \frac{k^d\left(k^{d-j+1} - 1\right)\left(1 - k^{-j}\right)}{\left(k^d - 1\right)(k - 1)}$$

For large networks with large k and d,

$$\frac{N(d-j)(N(d) - N(d-j))}{N(d) - 1} \simeq \frac{\left(k^{d-j+1} - 1\right)\left(1 - k^{-j}\right)}{(k - 1)} = \frac{k^{d-j}\left(k - k^{-(j-1)} - k^{-(j-d)} + k^{-d}\right)}{(k - 1)}$$

$$\simeq k^{d-j} = \left(\frac{\frac{k^j}{N(d)-1}}{\frac{k^d}{N(d)-1}}\right)^{-1}$$

Hence, the link betweenness distribution of a k-ary tree is not a precise power distribution, but it is close to a power law with exponent $c = -1$, especially for larger k and d. The first and last point of link betweenness corresponds to $j = 1$ and $j = N$. Since

$$\frac{N(d-1)(N(d) - N(d-1))}{N(d) - 1} = \left.\frac{k^d\left(k^{d-j+1} - 1\right)\left(1 - k^{-j}\right)}{\left(k^d - 1\right)(k - 1)}\right|_{j=1} = k^{d-1} = \left(\frac{\left.\frac{k^j}{N(d)-1}\right|_{j=1}}{\frac{k^d}{N(d)-1}}\right)^{-1}$$

The first and the last points always lie on a power law curve with exponent $c = -1$. Hence, an exceptional case is for $d = 2$, which is an exact power law although d is small.

Appendix B

Orthogonal polynomials

In the sequel, we will show that $p_D(\mu)$ belongs to a set of orthogonal polynomials. We refer to Szegö's classical book [89] for the beautiful theory of orthogonal polynomials.

B.1 The recursive nature of (4.2)

Lemma 28 *For all $j \geq 0$, the functions $\theta_j(D; x)$ are rational functions*

$$\theta_j(D; x) = \frac{t_j(D; x)}{t_{j-1}(D; x)} \tag{B.1}$$

where $t_j(x)$ is a polynomial of degree j in $x = -\mu$ and $t_0(D; x) = 1$.

Proof: It holds for $j = 1$ as verified from (4.2) because $\theta_0(D; x) = 1$. Let us assume that (4.2) holds for $j - 1$ (induction argument). Substitution of (B.1) into the right hand side of (4.2),

$$\theta_j(D; x) = \begin{cases} \frac{(x + n_{j-1} + n_{j+1}) t_{j-1}(D;x) - n_{j-1} n_j t_{j-2}(D;x)}{t_{j-1}(D;x)} & 1 \leq j \leq D \\ \frac{(x + n_D) t_D(D;x) - n_D n_{D+1} t_{D-1}(D;x)}{t_D(D;x)} & j = D + 1 \end{cases}$$

indeed shows that the left hand side is of the form (B.1) for j. This demonstrates the induction argument and proves the lemma. $\square$

Introducing (B.1) into the definition (4.3) yields

$$p_D(-x) = \frac{\prod_{j=1}^{D+1} t_j(D; x)}{\prod_{j=1}^{D+1} t_{j-1}(D; x)} = t_{D+1}(D; x)$$

143

We rewrite (B.1) as $t_j\left(D;x\right) = \theta_j\left(D;x\right)t_{j-1}\left(D;x\right)$ and with (4.2), we obtain the set of polynomials

$$
\begin{cases}
\quad t_{D+1}\left(D;x\right) = \left(x + n_D\right)t_D\left(D;x\right) - n_D n_{D+1} t_{D-1}\left(D;x\right) \\
t_j\left(D;x\right) = \left(x + n_{j-1} + n_{j+1}\right)t_{j-1}\left(D;x\right) - n_{j-1}n_j t_{j-2}\left(D;x\right) \quad \text{for } 1 \le j \le D \\
\qquad\qquad t_1\left(D;x\right) = \left(x + n_2\right)t_0\left(D;x\right)
\end{cases}
$$

$$\text{(B.2)}$$

where $t_0\left(D;x\right) = 1$. By iterating the equation upwards, we find that

$$
t_j\left(D;0\right) = \begin{cases} \displaystyle\prod_{m=2}^{j+1} n_m & 1 \le j \le D \\ \qquad 0 & j = D+1 \end{cases}
$$

$$\text{(B.3)}$$

Thus, $t_{D+1}\left(D;0\right) = 0$ (and thus $\theta_{D+1}\left(D;0\right) = 0$) implies that $p_D\left(\mu\right)$ must have a zero at $\mu = 0$, which is, indeed, a general property of any Laplacian as mentioned in Corollary 11. From (B.1), it then follows that

$$
\theta_j\left(D;0\right) = n_{j+1} > 0
$$

For a fixed D, the sequence $\{t_j\left(D;x\right)\}_{0 \le j \le D+1}$ is an orthogonal set of polynomials because it obeys Favard's three-term recurrence relation (see e.g. [43]). The zeros of any set of orthogonal polynomials are all simple, real and lying in the orthogonality interval $[a,b]$, which is here for the Laplacian equal to $[0,N]$. Moreover, the zeros of $t_j\left(D;x\right)$ and $t_{j-1}\left(D;x\right)$ are interlaced. In other words, in between two zeros of $t_{j-1}\left(D;x\right)$, there is precisely one zero of $t_j\left(D;x\right)$ and between two zeros of $t_j\left(D;x\right)$ there is at least one zero of $t_k\left(D;x\right)$ with $k > j$. This property shows that the set $\{t_j\left(D;x\right)\}_{0 \le j \le D+1}$ is finite and cannot be extended beyond $D+1$, because the smallest zero of the highest degree polynomial $t_{D+1}\left(D;x\right)$ coincides with the lower boundary of the orthogonality interval.

B.2 Jacobi Matrix of the set $\{t_j\left(D,x\right)\}_{1 \le j \le D+1}$

As known in the theory of orthogonal polynomials [43], it is instructive to rewrite the j-equation in (B.2) as

$$
xt_{j-1}\left(D;x\right) = n_{j-1}n_j t_{j-2}\left(D;x\right) - \left(n_{j-1} + n_{j+1}\right)t_{j-1}\left(D;x\right) + t_j\left(D;x\right)
$$

and in matrix form by defining the vector

$$
\tau\left(D;x\right) = \begin{bmatrix} t_0\left(D;x\right) & t_1\left(D;x\right) & \cdots & t_{D-1}\left(D;x\right) & t_D\left(D;x\right) \end{bmatrix}^T
$$

we arrive at

$$
x \begin{bmatrix} t_0 \\ t_1 \\ \vdots \\ t_{D-1} \\ t_D \end{bmatrix} = \begin{bmatrix} -n_2 & 1 & & & & \\ n_1 n_2 & -(n_1 + n_3) & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & n_{D-1} n_D & -(n_{D-1} + n_{D+1}) & 1 \\ & & & n_D n_{D+1} & -n_D \end{bmatrix} \begin{bmatrix} t_0 \\ t_1 \\ \vdots \\ t_{D-1} \\ t_D \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ t_{D+1} \end{bmatrix}
$$

where $t_j = t_j(D; x)$. Thus, the three-term recursion set of polynomials (B.2) is written in matrix form as

$$
x\tau(D; x) = M\tau(D; x) + t_{D+1}(D; x) e_{D+1} \tag{B.4}
$$

where the basic vector $e_{D+1} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T$ and the $(D+1) \times (D+1)$ Jacobi matrix is

$$
M = \begin{bmatrix} -n_2 & 1 & & & \\ n_1 n_2 & -(n_1 + n_3) & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & n_{D-1} n_D & -(n_{D-1} + n_{D+1}) & 1 \\ & & & n_D n_{D+1} & -n_D \end{bmatrix} \tag{B.5}
$$

When $x = z_k$ is a zero of $t_{D+1}(D; x) = p_D(-x)$, then (B.4) reduces to the eigenvalue equation

$$
M\tau(D; z_k) = z_k \tau(D; z_k)
$$

such that z_k is an eigenvalue of M belonging to the eigenvector $\tau(D; z_k)$. This eigenvector is never equal to the zero vector because the first component $t_0(x; D) = 1$. The special case where $z_{D+1} = 0$ leads again to (B.3) and all components of $\tau(D; 0)$ are positive.

There must be a similarity transform to make the matrix M symmetric (since all eigenvalues are real). The simplest similarity transform is $H = \operatorname{diag}(h_1, h_2, \ldots, h_{D+1})$ such that

$$
\widetilde{M} = HMH^{-1} = \begin{bmatrix} -n_2 & \frac{h_1}{h_2} & & & \\ \frac{h_2}{h_1} n_1 n_2 & -(n_1 + n_3) & \frac{h_2}{h_3} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{h_D}{h_{D-1}} n_{D-1} n_D & -(n_{D-1} + n_{D+1}) & \frac{h_D}{h_{D+1}} \\ & & & \frac{h_{D+1}}{h_D} n_D n_{D+1} & -n_D \end{bmatrix}
$$

Thus, in order to have $\widetilde{M} = \widetilde{M}^T$, we need to require that $\left(\widetilde{M}\right)_{i,i-1} = \left(\widetilde{M}\right)_{i-1,i}$ for all $2 \leq i \leq D+1$, implying that

$$
\frac{h_i}{h_{i-1}} n_{i-1} n_i = \frac{h_{i-1}}{h_i}
$$

whence,

$$\frac{h_{i-1}}{h_i} = \sqrt{n_{i-1}n_i}$$

and $h_i = \frac{1}{\sqrt{n_{i-1}n_i}} h_{i-1}$ for $2 \leq i \leq D+1$ and $h_1 = 1$. Thus,

$$H = \mathrm{diag}\left(1, \frac{1}{\sqrt{n_1 n_2}}, \ldots, \frac{1}{\sqrt{n_1 n_j}\prod_{k=2}^{j-1} n_k}, \ldots, \frac{1}{\sqrt{n_1 n_{D+1}}\prod_{k=2}^{D} n_k}\right)$$

and the eigenvector belonging to zero equals

$$\widetilde{\tau}(D;0) = H\tau(D;0) = \begin{bmatrix} 1 & \sqrt{\frac{n_2}{n_1}} & \cdots & \sqrt{\frac{n_{D-1}}{n_1}} & \sqrt{\frac{n_D}{n_1}} \end{bmatrix}^T$$

After the similarity transform H, the result is

$$\widetilde{M} = HMH^{-1} = \begin{bmatrix} -n_2 & \sqrt{n_1 n_2} & & & \\ \sqrt{n_1 n_2} & -(n_1 + n_3) & \sqrt{n_2 n_3} & & \\ & \ddots & \ddots & \ddots & \\ & & \sqrt{n_{D-1} n_D} & -(n_{D-1} + n_{D+1}) & \sqrt{n_D n_{D+1}} \\ & & & \sqrt{n_D n_{D+1}} & -n_D \end{bmatrix}$$

In summary, all non-trivial eigenvalues of $Q_{G_D^*}$ are also eigenvalues of the (much simpler and smaller) matrix $-M$ or $-\widetilde{M}$.

Appendix C

The graph maximizing the algebraic connectivity

We proved in Chapter 4 that the maximum algebraic connectivity of the class $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$ is also the maximum $a_{\max}$ over all graphs $G(N, D)$ with N nodes and diameter D. The tables below present the clique sizes of $G_D^*(n_1 = 1, n_2, ..., n_D, n_{D+1} = 1)$, which achieves the maximal algebraic connectivity $a_{\max}$ among all graphs with size N and diameter D.

	$a_{\max}(G(N = 26, D))$	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	n_{10}
$D = 2$	24	1	24	1							
$D = 3$	11.1345	1	12	12	1						
$D = 4$	5.6834	1	7	10	7	1					
$D = 5$	3.1264	1	5	7	7	5	1				
$D = 6$	1.8566	1	3	6	6	6	3	1			
$D = 7$	1.1555	1	2	5	5	5	5	2	1		
$D = 8$	0.781781	1	2	3	5	4	5	3	2	1	
$D = 9$	0.517162	1	1	3	4	4	4	4	3	1	1

	$a_{\max}(G(N = 50, D))$	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	n_{10}
$D = 2$	48	1	48	1							
$D = 3$	23.074278	1	24	24	1						
$D = 4$	12.641101	1	15	18	15	1					
$D = 5$	7.080889	1	9	15	15	9	1				
$D = 6$	4.290025	1	6	11	14	11	6	1			
$D = 7$	2.764758	1	5	8	11	11	8	5	1		
$D = 8$	1.859022	1	3	7	9	10	9	7	3	1	
$D = 9$	1.320825	1	3	5	7	9	9	7	5	3	1

	$a_{\max}(G(N=100,D))$	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	n_{10}
$D=2$	98	1	98	1							
$D=3$	48.0385	1	49	49	1						
$D=4$	27.6754	1	31	36	31	1					
$D=5$	15.8799	1	19	30	30	19	1				
$D=6$	9.7886	1	13	22	28	22	13	1			
$D=7$	6.3833	1	9	17	23	23	17	9	1		
$D=8$	4.358863	1	7	13	19	20	19	13	7	1	
$D=9$	3.098801	1	5	10	16	18	18	16	10	5	1

	$a_{\max}(G(N=122,D))$	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	n_{10}
$D=2$	120	1	120	1							
$D=3$	59.031762	1	60	60	1						
$D=4$	34.442561	1	39	42	39	1					
$D=5$	19.858188	1	24	36	36	24	1				
$D=6$	12.266200	1	16	27	34	27	16	1			
$D=7$	8.021537	1	11	20	29	28	21	11	1		
$D=8$	5.499296	1	8	16	23	26	23	16	8	1	
$D=9$	3.910465	1	6	13	18	23	22	19	13	6	1

Appendix D

Symbols

α	extreme value index of the polynomial distribution
γ	Euler constant
λ_i	i-th largest eigenvalue of the adjacency matrix
μ_i	i-th largest eigenvalue of the Laplacian matrix
μ^*	normalized increase of algebraic connectivity
ρ	correlation coefficient of 2-dimensional link weight
τ	exponent of a power law degree distribution
$a(G) = \mu_{N-1}$	algebraic connectivity of a graph G
A	adjacency matrix
B_l	betweenness of a link
$c_G(v)$	clustering coefficient of a node v
$C(G)$	clustering coefficient of a graph
d	depth of a k-ary tree
D_G	degree of a random node in a graph G
D	diameter of a graph
$E[X]$	statistical expectation of random variable X
$F_w(x)$	probability distribution function $\Pr[w \leq x]$ of link weight
G	graph
$G_p(N)$	Erdös-Rényi random graph with link density p and N nodes
H_N	hopcount of a shortest path
K_N	complete graph with N nodes
$l(i \rightarrow j)$	link between node i and j
$\mathcal{L}$	set of L links
L	number of links
$\mathcal{M}$	set of m testboxes
$\mathcal{N}$	set of N nodes
N	number of nodes

p_c — disconnectivity threshold

$P_{i \to j}$ — a path between node pair i and j

$P_{i \to j}^*$ — the shortest path between node pair i and j

Q — Laplacian matrix

$R(G)$ — a robustness measure of network G

u_i — the i-th element of the Fiedler vector

$w(i \to j)$ — weight of link $l(i \to j)$

$w(P_{i \to j}^*)$ — weight of the shortest path $P_{i \to j}^*$

$w(G)$ — the total link weight of graph G

Y — the number of link failures in a shortest path

Bibliography

[1] M. Abramowitz and J. A. Stegun. *Handbook of Mathematical Functions.* Dover Publications Inc., New York, 1999.

[2] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, 2005.

[3] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 171–180, 2000.

[4] R. Albert and A. Barabasi. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[5] R. Albert, H. Jeong, and A.-L. Barabàsi. Error and attack tolerance of complex networks. *Nature*, 406:378–482, 2000.

[6] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of The Cell.* Garland Science, New York, fifth edition, 2008.

[7] S. Androutsellis-Theotokis and D. Spinellis. A survey of peer-to-peer content distribution technologies. *ACM Computing Surveys*, 36(4):335–371, 2004.

[8] A.-L. Barabasi. *Linked, The new science of networks.* Perseus Publishing, Cambridge, Massachusetts, April 2002.

[9] M. Barthélemy. Betweenness centrality in large complex networks. *Eur. Phys. J. B*, 38:163–168, 2004.

[10] A. Beygelzimer, G. Grinstein, R. Linsker, and I. Rish. Improving network robustness. *Physica A*, 357(3-4):593–612, Nov. 2005.

[11] J. J. Binney, N.J. Dowrick, A.J. Fisher, and M.E.J. Newman. *The Theory of Critical Phenomena: An Introduction to the Renormalization Groups.* Clarendon Press, Oxford, 1992.

[12] S. Boccaletti, V. Latora, Y. Morenod, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424:175–308, 2006.

[13] B. Bollobás. *Random Graphs*. Cambridge University Press, Cambridge, UK, second edition, 2001.

[14] B. Bollobás and O. Riordan. Shortest paths and load scaling in scale-free trees. *Phys. Rev. E*, 69(036114), 2004.

[15] A. E. Brouwera and W. H. Haemers. A lower bound for the laplacian eigenvalues of a graph - proof of a conjecture by guo. *Linear Algebra and its Applications*, 429(8-9):2131–2135, 2008.

[16] G. Caldarelli, R. Marchetti, and L. Pietronero. The fractal properties of internet. *Europhys. Lett.*, 52:386–391, 2000.

[17] M. Castro, M. Costa, and A. Rowstron. Should we build gnutella on a structured overlay? *ACM SIGCOMM Computer Communications Review*, 34(1):131–136, 2004.

[18] G. Chartrand and L. Lesniak. *Graphs and Digraphs*. Chapman and Hall/CRC, 1996.

[19] G. Chartrand and O. R. Oellermann. *Applied and Algorithmic Graph Theory*. Mcgraw-Hill College, 1992.

[20] Y. Chen, E. Lopez, S. Havlin, and H. E. Stanley. Universal behavior of optimal paths in weighted networks with general disorder. *Phys. Rev. Lett.*, 96(068702), 2006.

[21] F. R. K. Chung, V. Faber, and T. A. Manteuffel. An upper bound in the diameter of a graph from eigenvalues associated with its laplacian. *SIAM J. Discrete Math.*, 7:443–457, 1994.

[22] A. Clauset and C. Moore. Accuracy and scaling phenomena in internet mapping. *Phys. Rev. Lett.*, 94(018701), 2005.

[23] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin. Breakdown of the internet under intentional attack. *Phys. Rev. Lett.*, 86(3682), 2001.

[24] V. Colizza, R. Pastor-Satorras, and A. Vespignani. Reaction–diffusion processes and metapopulation models in heterogeneous networks. *Nature Physics*, 3:276 – 282, 2007.

[25] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *An Introduction to Algorithms.* MIT Press, Boston, 1991.

[26] D. M. Cvetkovic, M. Doob, and H. Sachs. *Spectra of Graphs.* VEB Deutscher Verlag der Wissenschaften, Berlin, 1980.

[27] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167 – 242, Jan. 2007.

[28] N.M.M. de Abreu. Old and new results on algebraic connectivity of graphs. *Linear Algebra and its Applications*, 423(1):53–73, 2007.

[29] R. Van der Hofstad, G. Hooghiemstra, and P. Van Mieghem. Size and weight of shortest path trees with exponential link weights. *Combinatorics, Probability and Computing*, 15:903–926, 2006.

[30] E. W. Dijkstra. A note on two problems in connection with graphs. *Num. Math.*, 1:269–271, 1959.

[31] L. Donetti, F. Neri, and M. A. Muñoz. Optimal network topologies: Expanders, cages, ramanujan graphs, entangled networks and all that. *J. Stat. Mech.*, (P08007), 2006.

[32] S. N. Dorogovtsev, A. V. Goltsev, J. F. F. Mendes, and A. N. Samukhi. Spectra of complex networks. *Phys. Rev. E*, 68(046109), 2003.

[33] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Structure of growing networks with preferential linking. *Phys. Rev. Lett.*, 85(4633), 2000.

[34] S.N. Dorogovtsev and J.F.F. Mendes. *Evolution of Networks, From Biological Nets to the Internet and WWW.* Oxford University Press, Oxford, 2003.

[35] M. Drmota and H-K. Hwang. Profiles of random trees: Correlation and width of random recursive trees and binary search trees. *Advances in Applied Probability*, 37:321–341, 2005.

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

[37] P. Erdös and A. Rényi. The evolution of random graphs. *Magyar Tud. Akad. Mat. Kutato Int. Kozl.*, 5:17–61, 1960.

[38] Y. Fan. On spectral integral variations of graphs. *Linear and Multilinear Algebra*, 50:133–142, 2002.

[39] I. J. Farkas, I. Deréyi, A.-L. Barabási, and T. Vicsek. Spectra of Şreal-worldŤ graphs: Beyond the semicircle law. *Phys. Rev. E*, 64(026704), 2001.

[40] A. Fekete and G. Vattay. Distribution of edge load in scale-free trees. *Phys. Rev. E*, 73(046102), 2006.

[41] A. Fekete, G. Vattay, and L. Kocarev. Traffic dynamics in scale-free networks. *Complexus*, 3:97–107, 2006.

[42] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Math. J.*, 23:298–305, 1973.

[43] W. Gautschi. *Orthogonal Polynomials: Computation and Approximation.* Oxford University Press, Oxford, UK, 2004.

[44] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99:7821–7826, 2002.

[45] K.-I. Goh, B. Kahng, and D. Kim. Universal behavior of load distribution in scale-free networks. *Phys. Rev. Lett.*, 87(278701), 2001.

[46] K.-I. Goh, J. D. Noh, B. Kahng, and D. Kim. Load distribution in weighted complex networks. *Phys. Rev. E*, 72(017102), 2005.

[47] K.-I. Goh, E. Oh, H. Jeong, B. Kahng, and D. Kim. Classification of scale-free networks. *Proc. Natl. Acad. Sci. U.S.A.*, 99(12583), 2002.

[48] R. Grone and R. Merris. The laplacian spectrum of a graph ii. *SIAM J. Discrete Math.*, 7:221–229, 1994.

[49] L. He, X. Liu, and G. Strang. Trees with cantor eigenvalue distribution. *Stud. Appl. Math.*, 110(123), 2003.

[50] R. Hekmat and P. Van Mieghem. Connectivity in wireless ad hoc networks with a log-normal radio model. *Mobile Networks and Applications*, 11(3):351–360, 2005.

[51] C. Hugenii. *Horoloquium Oscilatorium.* Apud F. Muguet, Parisiis, 1673.

[52] A. Jamakovic, R.E. Kooij, P. Van Mieghem, and E.R. van Dam. Robustness of networks against viruses: the role of the spectral radius. *Proceedings of the 13th Annual Symposium of the IEEE/CVT Benelux*, 2006.

[53] M. Janic, F. Kuipers, X. Zhou, and P. Van Mieghem. Implications for qos provisioning based on traceroute measurements. *Proceedings of 3nd International Workshop on Quality of Future Internet Services*, pages 3–14, 2002.

[54] M. Janic and P. Van Mieghem. On properties of multicast routing trees. *International Journal of Communication Systems*, 19(1):95–114, 2006.

[55] J. O. Kephart and S. R. White. Directed-graph epidemiological models of computer viruses. *Proc. of the 1991 Computer Society Symposium on Research in Security and Privacy*, pages 343–359, 1991.

[56] D.-H. Kim, J. D. Noh, and H. Jeong. Scale-free trees: The skeletons of complex networks. *Phys. Rev. E*, 70(046126), 2004.

[57] M. Kitsak, S. Havlin, G. Paul, M. Riccaboni, F. Pammolli, and H. E. Stanley. Betweenness centrality of fractal and nonfractal scale-free model networks and tests on real networks. *Phys. Rev. E*, 75(056115), 2007.

[58] F.A. Kuipers and P. Van Mieghem. Conditions that impact the complexity of qos routing. *IEEE/ACM Transaction on Networking*, 13(4):717–730, Aug. 2005.

[59] A. Lakhina, J. Byers, M. Crovella, and P. Xie. Sampling biases in ip topology measurements. *Proc. of IEEE INFOCOM*, 2003.

[60] W. Lin and X. Zhan. Polarized networks, diameter, and synchronizability of networks. *arXiv.org:cond-mat/0604295*, 2006.

[61] C. Maas. Transportation in graphs and the admittance spectrum. *Discrete Applied Mathematics*, 16(1):31–49, Jan. 1987.

[62] P. Mahadevan, D. Krioukov, M. Fomenkov, B. Huffaker, X. Dimitropoulos, K. Claffy, and A. Vahdat. The internet as-level topology: Three data sources and one definitive metric. *ACM SIGCOMM Computer Communications Review*, 36(1):17–26, 2006.

[63] R. M. May. *Stability and Complexity in Model Ecosystems*. Princeton University Press, Princeton, 1973.

[64] R. Merris. Laplacian matrices of graphs: a survey. *Linear Algebra and its Applications*, 197-198:143–176, 1994.

[65] C. D. Meyer. *Matrix Analysis and Applied Linear Algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2000.

[66] P. Van Mieghem. *Data Communications Networking*. Techne Press, Amsterdam, 2006.

[67] P. Van Mieghem. *Performance Analysis of Communications Systems and Networks*. Cambridge University Press, Cambridge, UK, 2006.

[68] P. Van Mieghem, G. Hooghiemstra, and R. W. van der Hofstad. A scaling law for the hopcount. *Delft University of Technology*, (report2000125), 2000.

[69] P. Van Mieghem and F.A. Kuipers. Concepts of exact quality of service algorithms. *IEEE/ACM Transaction on Networking*, 12(5):851–864, 2004.

[70] P. Van Mieghem and S. M. Magdalena. A phase transition in the link weight structure of networks. *Phys. Rev. E*, 72(056138), 2005.

[71] P. Van Mieghem, J. S. Omic, and R. E. Kooij. Virus spread in networks. *IEEE/ACM Transaction on Networking*, 17(1):1–14, February 2009.

[72] P. Van Mieghem and S. van Langen. Influence of the link weight structure on the shortest path. *Phys. Rev. E*, 71(056113), 2005.

[73] P. Van Mieghem and H. Wang. Spectra of a new class of graphs with extremal properties. *submitted*, 2008.

[74] P. Van Mieghem and H. Wang. The observable part of a network. *IEEE/ACM Transaction on Networking*, accepted, 2009.

[75] B. Mohar, Y. Alavi, G. Chartrand, O.R. Oellermann, and A.J. Chwenk. The laplacian spectrum of graphs. *Graph Theory, Combinatorics and Applications*, 2:871–898, 1991.

[76] V.D. Milman N. Alon. isoperimetric inequalities for graphs and superconcentrators. *J. Combin. Theory. Ser. B*, 38:73–88, 1985.

[77] M. E. J. Newman. Scientific collaboration networks: I. network construction and fundamental results. *Phys. Rev. E*, 64(016131), 2001.

[78] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74(036104), 2006.

[79] T. Ohira and R. Sawatari. Phase transition in a computer network traffic model. *Phys. Rev. E*, 58:193–195, 1998.

[80] R. Oliveira, B. Zhang, and L. Zhang. Observing the evolution of internet as topology. *ACM SIGCOMM*, 2007.

[81] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86(3200), 2001.

[82] S. Pati S. Barik. On algebraic connectivity and spectral integral variations of graphs. *Linear Algebra and its Applications*, 397:209–222, 2005.

[83] R. Sherwood, A. Bender, and N. Spring. Discarte: A disjunctive internet cartographer. *ACM SIGCOMM'08*, 2008.

[84] R. T. Smythe and John C. Wierman. *First-Passage Percolation on the Square Lattice. In: Lecture Notes in Mathematics*, volume Vol. 671. Springer, Berlin, Heidelberg, New York, 1978.

[85] C. J. Stam and J. C. Reijneveld. Graph theoretical analysis of complex networks in the brain. *Nonlinear Biomedical Physics*, 1(3), July 2007.

[86] W. Richard Stevens. *TCP/IP Illustrated, volume 1, The Protocols.* Addsion Wesley, Reading , Massachusetts, 1994.

[87] S. H. Strogatz. Exploring complex networks. *Nature*, 410:268–276, March 2001.

[88] G. Szabó, M. Alava, and J. Kertész. Shortest paths and load scaling in scale-free trees. *Phys. Rev. E*, 66(026101), 2002.

[89] G. Szegö. *Orthogonal Polynomials*, volume 23. American Mathematical Society, Providence, Rhode Island, fourth edition, 1978.

[90] E. D. van Dam and W. H. Haemers. Which graphs are determined by their spectrum? *Linear Algebra and its Applications*, 373:241–272, 2003.

[91] E.R. van Dam. Graphs with given diameter maximizing the spectral radius. *Linear Algebra and its Applications*, 426:454–457, Oct. 2007.

[92] R. van der Hofstad, G. Hooghiemstra, and P. Van Mieghem. First passage percolation on the random graph. *Probability in the Engineering and Informational Sciences (PEIS)*, 15:225–237, 2001.

[93] R. van der Hofstad, G. Hooghiemstra, and P. Van Mieghem. On the covariance of the level sizes in random recursive trees. *Advances in Applied Probability*, 20:519–539, 2002.

[94] P. VanMieghem, H. Wang, C. Doerr, J. M. Hernandes, D. Hutchison, and M. Karaliopoulos. A framework for network robustness. *EU EP7 ResumeNet*, 2009.

[95] D. Vukadinovi, P. Huang, and T. Erlebach. On the spectrum and structure of internet topology graphs. *Lect. Notes Comput. Sci.*, 83(2346), 2002.

[96] H. Wang and P. Van Mieghem. Graphs with given diameter maximizing the algebraic connectivity -numerical results. *Delft University of Technology*, report20081118, 2008.

[97] X. Wang and G. Chen. Synchronization in scale-free dynamical networks: robustness and fragility. *IEEE Trans. Circ. Syst. I.*, 49:54–62, 2002.

[98] X. Wang and G. Chen. Synchronization in small-world dynamical networks. *Int. J. Bifurcation Chaos Appl. Sci. Eng.*, 12(1):187 – 192, 2002.

[99] D. J. Watts. *Small Worlds, The Dynamics of Networks between Order and Randomness.* Princeton University Press, Princeton, New Jersey, 1999.

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

[101] E. Wigner. On the distribution of the roots of certain symmetric matrices. *Ann. Math.*, 67(325-327), 1958.

[102] E. O. Wilson. *Consilience: The Unity of Knowledge.* Knopf, New York, 1998.

[103] Z. Wu, L. A. Braunstein, S. Havlin, and H. E. Stanley. Transport in weighted networks: Partition into superhighways and roads. *Phys. Rev. Lett.*, 96(148702), 2006.

Samenvatting (Summary in Dutch)

Onze maatschappij hangt nu meer dan ooit afhankelijk van grote netwerken, zoals transportnetwerken, het Internet en hoogspanningsnetwerken. Omdat netwerken altijd de werking van de dienst beïnvloeden, worden ingenieurs geconfronteerd met fundamentele kwesties als "hoe kunnen we de robuustheid van een netwerk voor een bepaalde dienst evalueren?", "hoe ontwerpt men een robuust netwerk?". Robuustheid is een belangrijke kwestie voor veel complexe netwerken, waarin diverse dynamische processen plaatsvinden. In dit werk defnieren we robuustheid als volgt: een netwerk is meer robuust indien de dienst over het netwerk beter presteert, waar de prestatie wordt gemeten (a) in de conventionele staat, of (b) tijdens perturbaties, bijv. tijdens uitval, virusverspreiding etc. In dit proefschrift onderzoeken wij een bepaalde stroming van netwerkrobuustheid binnen een generiek raamwerk: robuustheidkwantificering, optimalisatie en de interactie tussen dienst en netwerk.

Significante vooruitgang is geboekt in het begrijpen van de relatie tussen de structurele netwerkeigenschappen en de prestatie van dynamica of diensten die plaatsvinden in deze netwerken. We nemen aan dat netwerkrobuustheid gekwantificeerd kan worden door een topologische maat van het netwerk. Een kort overzicht van topologische maten wordt gepresenteerd. Iedere maat kan de netwerkrobuustheid representeren met betrekking tot een bepaald aspect van de prestatie van de dienst. Wij richten ons op de maat die bekend staat als algebraische connectiviteit. Bewijs verzamelt uit de literatuur toont aan dat de algebraische connectiviteit de netwerkrobuustheid karakteriseert met betrekking tot het synchroniseren van dynamische processe in knooppunten, random walks in grafen en de connectiviteit van een netwerk. Bovendien tonen we aan dat bij een gegeven diameter, grafen met een grote algebraische connectiviteit geneigd zijn om compact te zijn in de kern en opener aan de rand. Dergelijke structuren distribueren het verkeer op homegene wijze en zijn derhalve robuust in termen van traffic engineering.

Hoe ontwerpen we een robuust netwerk met betrekking tot de maatstaf algebraische connectiviteit? Ten eerste, de complete graaf heeft de maximale algebraische connectiviteit, terwijl diens hoge linkdichtheid het gebruik onpraktisch maakt vanwege de hoge kosten om een link te construeren. Normaal gesproken zijn er beperkingen opgesteld voor andere netwerkeigenschappen om realistische eisen toe te voegen. Bijvoorbeeld, een beperking voor de diameter kan garanderen dat bepaalde kwaliteitsnormen wor-

den behaald met betrekking tot bijvoorbeeld vertragingen. We stellen een klasse voor van gegroepeerde ketenstructuren welke de algebraische connectiviteit en vele andere robuuste eigenschappen optimaliseren voor alle grafen met diameter D en grootte N. De optimale graaf binnen deze klasse kan bepaald worden via analyse of simulatie. Ten tweede, de volledige vervanging van een bepaalde infrastuctuur is kostbaar. Derhalve ontwerpen we strategieën voor het optimaliseren van de robuustheid met gebruik van kleine topologische aanpassingen. Deze strategieën worden geevalueerd in diverse klassen van grafen.

De kwantificering van robuustheid, of evenzo, de associatie van de prestatie van een dienst met een topologische maat, kan impliciet zijn. In dit geval verkennen we de interactie tussen topologie en dienst bij het bepalen van de algemene prestatie. Vele diensten in communicatie- en transportnetwerken zijn gebaseerd op kortste pad routering. Het gewicht van een link, zoals vertraging of bandbreedte, is voornamelijk een maat die geoptimaliseerd is via kortste pad routering. Dus, het fijnstellen van het gewicht van de link, een mechanisme om het verkeer te controleren, wordt mede gezien als een onderdeel van de dienst. De interactie tussen dienst (kortste pad routering en gewicht aanpassing) en topologie wordt onderzocht voor de volgende prestatieaspecten: (a) de structuur van het transport overlay netwerk, welke de vereniging is van de korste paden tussen alle paren van knooppunten en (b) de verkeersverdeling in het overlay netwerk. Belangrijke, nieuwe bevindingen zijn (i) de universele faseovergang in overlay structuren als we de structuur van linkgewichten aanpassen voor verschillende klassen van netwerken en (ii) de power-law verdeling van het verkeer in overlay netwerken wanneer de gewichten van de links sterk variëren in verschillende klassen van netwerken. Bovendien beschouwen we de dienst die de netwerktopologie meet als de unie van de kortste paden van een set van testboxen (knooppunten). De gemeten topologie is een subgraaf van het overlay netwerk, welke wederom een subgraaf is van het werkelijke netwerk. De prestatie wordt onderzocht in termen van vertekeningen tijdens het meten van de netwerktopologie. Ons werk voegt substantieel toe aan de kennis van de invloed van de dienst (selectie van de testbox) en de werkelijke netwerkstructuur op de prestatie met betrekking tot de vertekening van de metingen. Ons onderzoek naar de interactie tussen dienst en netwerk openbaart wederom de associatie tussen de prestatie van een dienst en een bepaald topologisch kenmerk, en dus, draagt bij aan de kwantificering van netwerkrobuustheid.

De multidisciplinaire aard van dit onderzoek ligt niet alleen in de aanwezigheid van robuustheidskwesties in vele complexe netwerken, maar ook in het feit dat vorderingen in andere disciplines, zoals grafentheorie, combinatoriek, lineaire algebra en statistische fysica op grote schaal worden toegepast in dit proefschrift om optimalisatieproblemen en prestatie van grote netwerken te kunnen bestuderen.

Acknowledgements

This thesis presents the most interesting work I have performed during my studies towards a Ph.D. degree. It brings back all the vivid memories over these four years, one of the happiest period in my life. Here, I would like to thank many people, without whom this thesis would not be possible.

In the first place, I would like to thank my supervisor and promotor Prof. Piet Van Mieghem. I feel lucky to be his Ph.D. as well as M.Sc. student. I am thankful to him for his guidance, patience, for sharing ideas and knowledge in countless inspiring and open discussions, for his comments written on piles of manuscripts. His enthusiasm makes our research much more enjoyable.

I would like to express my gratitude to the members of my Ph.D defense committee, for their time and effort spent on this thesis and for the nice conversations we had. It is my great honor to have them involved in this thesis. In particular I would like to thank Prof. Kees Stam for our pleasant cooperation, which brings me to the promising area of brain networks. I am indebted to Prof. Caterina Scoglio for her willingness to travel to Delft from the US. My special thanks go to Dr. Rob Kooij for his inspiring and unlimited ideas, for his encouraging attitude towards work and life.

Working on the 19th floor of the EWI building in TUDelft is very enjoyable. I would like to thank all the NAS and WMC people for their help, for the happiness they bring to me. I am grateful to the cooperations with Fernando, Almerima, Javier, Siyu, Wynand and Dajie. My gratitude goes to Fernando and Tom for translating part of this thesis into Dutch, and to Javier for his nice figures. My thanks go to Laura, Marjon and Wendy for their help in all kinds of documents and procedures. I would like to thank my dear Jasmina and Jing for our joy.

Finally, I would like to thank my parent and parent-in-law for their unlimited love and support; my husband for his care and humor. My sincere gratitude goes to my good friends, the Guos and the Sus, for the time we have enjoyed together.

Curriculum Vitae

Huijuan Wang was born in Harbin, China, on April 28, 1981. She graduated with her Bachelor of Science B.Sc. degree (cum laude) in Electrical Engineering at the Harbin Engineering University in July 2003. She graduated and received her M.Sc. degree (cum laude) in Electrical Engineering at the Delft University of Technology, the Netherlands, in the year 2005. After finishing her study she joined the Network Architecture and Services (NAS) Group to work towards a Ph.D. degree. Under the supervision of the NAS group chair Piet Van Mieghem, she was working on the NWO project ROBUNET. Her core research area is the performance analysis of large complex networks like the Internet, transportation networks and brain functional networks. Specifically, her focus is on the evaluation of how robust or good the network architecture is for the quality of a service, on the design of a robust network structure and on the interplay between the network structure and the service in determining the overall performance.

Publications

Journal

- P. Van Mieghem and H. Wang, "The Observable Part of a Network", IEEE/ACM Transaction on Networking, vol. 17, No. 1, pp. 93-105, 2009.

- H. Wang, J. Martin Hernandez and P. Van Mieghem, "Betweenness Centrality in Weighted Networks", Physical Review E, vol. 77 (046105), April, 2008.

- H. Wang and P. Van Mieghem, "Graphs with given diameter maximizing the algebraic connectivity", submitted to COMBINATORICA, 2008.

- P. Van Mieghem and H. Wang, "Spectra of a new class of graphs with extremal properties", submitted to SIAM Journal on Discrete Mathematics, 2008.

- H. Wang and P. Van Mieghem, "Sampling networks by the union of m shortest path trees", submitted to Computer Networks, 2008.

Conference

- H. Wang and P. Van Mieghem, 2008, "Algebraic Connectivity Optimization via Link Addition", Bionetics 2008, Japan, November, 2008.

- S. Tang, H. Wang and P. Van Mieghem, "The Effect of Peer Selection with Hopcount or Delay Constraint on Peer-to-Peer Networking", IFIP Networking, Singapore, May, 2008.

- H. Wang and P. Van Mieghem, "Constructing the Overlay Network by Tuning Link Weights", ChinaCom2007, China, August, 2007.

- H. Wang and P. Van Mieghem, "Shifting the Link Weights in Networks", ChinaCom2007, China, August, 2007.

- J. Martin Hernandez, T. Kleiberg, H. Wang and P. Van Mieghem, "A Qualitative Comparison of Power Law Generators", International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS 2007), USA, July, 2007.

- A. Jamakovic, H. Wang and P. Van Mieghem, "Topological characteristics of the Dutch road infrastructure", Seminar Infrastructure Reliability, Netherlands, 2006.

- F. A. Kuipers, H. Wang and P. Van Mieghem, "The Stability of Paths in a Dynamic Network", CoNEXT 2005, France, October, 2005.