TOPOLOGY OF MOLECULAR NETWORKS

TOPOLOGY OF MOLECULAR NETWORKS

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K. C. A. M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op maandag 10 maart 2014 om 15:00 uur

door

Wynand WINTERBACH

Meester van Wetenskap in Toegepaste Wiskunde, Universiteit Stellenbosch, Zuid-Afrika
geboren te King Williams Town, Zuid-Afrika.

Dit proefschrift is goedgekeurd door de promotor:

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

Copromotor: Dr. D. de Ridder

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. P. F. A. Van Mieghem,	Technische Universiteit Delft, promotor
Dr. ir. D. de Ridder,	Technische Universiteit Delft, copromotor
Prof. dr. ir. G. W. Klau,	Vrije Universiteit Amsterdam
Prof. dr. ir. J. J. Heijnen,	Technische Universiteit Delft
Prof. dr. B. Mons,	Leids Universitair Medisch Centrum
Prof. dr. H. J. Jensen,	The Imperial College of Science, Technology and Medicine, Verenigd Koninkrijk
Dr. P. Dini,	London School of Economics and Political Science, Verenigd Koninkrijk
Prof. dr. L. Wessels,	Technische Universiteit Delft, reservelid

Prof. dr. ir. M. J. T. Reinders en Dr. ir. H. Wang hebben als begeleiders in belangrijke mate aan de totstandkoming van het proefschrift bijgedragen.

Keywords: Graph, network, biology, molecular, disease, robustness

Printed by: Proefschriftmaken.nl

Front & Back: W. Winterbach

ISBN 978-94-6186-289-1

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

To my mother and father.

Topology of Molecular Networks

February 25, 2014

CONTENTS

1

BACKGROUND

Scientists have long attempted to understand biological systems in terms of simple building blocks and their interactions. The molecular biology research program was started in the 1930's under the assumption that biological function emerged from interactions between a few fundamental biological molecules. The research program quickly led to the discovery of until-then-unknown fundamental molecules and interaction pathways. But molecular biology was a reductionist science and was ill suited to study complex, indirect molecular interactions – such systems require consideration of all pathways between all molecules. High-throughput techniques invented at the end of the 20th century enabled measurement of many simultaneous interactions and revealed that many phenotypic traits arise from indirect interactions.

To understand these indirect interactions, new analysis techniques focused on interactions, instead of only molecules, were needed. Graph theory, the branch of mathematics that deals with the abstract analysis of networks, proved to be a good research tool as interaction datasets were already generally represented as networks. Eventually, researchers started considering molecular interaction networks as objects of study in their own right. Given that the functions of systems can often be inferred from their structure, it was natural to ask whether structures in molecular networks could be used to predict biological properties of the underlying systems. Thus was born the field of network biology.

This thesis is motivated by the observation that biological system function can often be predicted from the system structure. Networks are abstract representations of system interactions; networks related to biological systems may therefore contain structural signatures that could aid in the prediction of biological function. An important consequence would be that, since biological systems are known to be robust, certain structural signatures in biological networks may be associated with robustness, a finding that would aid in the design of more robust human-made networks.

The content of this thesis centers around molecular interaction networks. In the remainder of this introduction, we briefly discuss the types of molecular interactions that are considered. In order to analyze the molecular interaction networks we use tools from

graph theory for which we also give a brief overview.

1.1. MOLECULAR INTERACTION NETWORKS

Here, we assume a basic familiarity with the field of molecular biology and refer the interested reader to [1] for a thorough introduction to the field.

The fundamental molecules of interest in network biology are DNA, RNA, proteins and any small molecules interacting with these molecules. A simplified overview of common interactions between these molecules is shown in Figure 2.1a. Whilst this depiction includes only a few molecules, a typical cell contains between thousands and millions of distinct molecules. As molecular biology does not cope well with the simultaneous consideration of so many molecules, researchers in the field tend to focus on small subsystems of interacting molecules such as the MAPK-ERK pathway shown in Figure 2.1b.

In contrast to molecular biology, the aim of network biology is the analysis of networks of hundreds or thousands of molecular interactions and therefore it considers much larger systems than that of Figure 2.1b. Neither Figure 2.1a nor Figure 2.1b is suitable as network model for use in network biology. Figure 2.1b models a sequence of events and contains non-pairwise relations. Figure 2.1a could, in principle, be analyzed using graph theory but its many node types and even greater number of interaction types limit the applicability of such analyses. Network biology generally focuses on networks containing one or two kinds of molecule and one or two kinds of interaction. The network in Figure 2.1c is a simplification of Figure 2.1b that models only protein binding relations; due to its homogeneity, larger versions containing thousands of protein binding relations are well suited for graph theoretic analyses.

Due to the variety of molecules and molecular interactions in the cell, many kinds of molecular interaction network are studied. A number of commonly studied molecular interaction network types are enumerated in the list below:

Association networks represent any kind of relation between molecules (e.g. binding, co-expression and structural similarities). Examples of association networks are **gene co-expression networks** and **protein similarity networks**; in fact the entire network in Figure 2.1a could be seen as one large, if very imprecise, association network.

Functional networks model functional relations between pairs of molecules (usually genes or proteins). A link implies that both are involved in the same function, process or phenotype. For example, **Genetic interaction networks** represent interactions where a pair of genetic mutations leads to an epistatic effect, i.e., worse or better than expected based on the single mutation.

Protein-protein interaction (PPI) networks are undirected networks that model protein binding (in Figure 2.1a, protein interactions are shown in the strip labeled "Protein" as dashed lines without arrow heads). PPI networks are derived from high-throughput experiments using techniques such as yeast two-hybrid screening, mass spectrometry and tandem affinity purification [2]. **Signaling networks** are related to to PPI networks but represent signal transduction between proteins (and other molecules) instead of binding. Since signal transduction is directional

(that is, proceeds from a signal source and ends at a final signal sink), signaling networks are directed.

Transcription-regulatory (TR) networks are bipartite networks with one set of nodes representing genes and the other representing transcription factors (TFs). TFs are products of genes (modeled by gene-TF links; in Figure 2.1a these links are indirect and are a combination of Genome-RNA links – black, solid lines representing transcription – and RNA-Protein links – dashed lines representing translation) whilst genes are regulated by TFs (modeled by TG-gene links; in Figure 2.1a, the two black links stretching from proteins to the genome). Data for such networks is derived through the process of chromatin immunoprecipitation (ChIP) [3]. **Gene regulatory (GR) networks** are related to TR networks but contain only genes. Often, their links represent indirect regulatory relationships.

Metabolic networks are bipartite networks that model the relationships between the chemical reactions that occur in cells and the substrates involved in the reactions (in Figure 2.1a, substrates are represented as diamonds, reactions as enzymes/proteins and the chemical relationships as solid gray lines). Simplified, non-bipartite metabolic networks containing only metabolites or only reactions are also often studied.

1.2. GRAPH THEORY

In graph theory, networks are studied as abstract representations of relationships between objects [4]. The objects are known as **nodes**, the full set of which is represented as $\mathcal{N}$. Relationships are represented as **links** connecting pairs of nodes; the set of links is denoted $\mathcal{L}$. When nodes u and v are linked (i.e. $\{u, v\} \in \mathcal{L}$), u is said to be a **neighbor** of v and vice-versa. The number of neighbors of a node u is called its **degree**. In Figure 2.1c, EGF is a neighbor of both EGFR and GRB2 and therefore has degree 2.

Networks in the above description do not model non-symmetric relationships and are therefore also known as **undirected networks**. When non-symmetric relationships, such as the flow directions of chemical reactions in the metabolic layer of Figure 2.1a are to be modeled, **directed networks** are used.

Graph Theory allows one to discover structural similarities between superficially very different networks. Examples of structural properties that one might compare include link distribution, the number of tightly connected communities and the average shortest path between arbitrary nodes. Typically, structural properties are represented by means of **metrics**, simpler scalar or vector measures of the properties in question.

The most fundamental network metrics are number of nodes, denoted N, and the number of links, denoted L. Network density is the ratio of the the link count L to the maximum number of links possible in the network ($N(N-1)/2$). A number of network metrics that are commonly used in network biology are 1) the distribution of node degrees, 2) the clustering coefficient, a measure of how densely connected the neighbors of nodes are and 3) the average length of the shortest distances between all pairs of nodes.

The research in this thesis starts with an analysis of the **degree assortativity** of a network, a metric that measures the tendency of nodes with similar degrees to be connected.

1.3. OUTLINE OF THESIS

We commence by reviewing the state of the art in network biology research in § 2. In § 2, we also show that the way in which researchers apply network biology has changed significantly since the early days of the field.

Our own research starts with an investigation into degree assortativity as a measure of network robustness. However, assortativity is a relatively new metric whose properties are not yet fully understood. In § 3, we study what it means for a network to have high or low assortativity. The question of whether networks can reconfigure themselves gradually to attain high or low assortativity values is considered in § 4.

Having established that assortativity can be well optimized through a series of small changes in network structure, we were interested in whether it is a factor in network evolution since evolution itself is a series of small changes. In § 5, we develop a framework for measuring network robustness and proceed to test the effect of changes in assortativity on instances of random network and a metabolic network. The aim of this chapter is to investigate whether molecular networks are structurally more robust than expected by chance.

In § 6, we explore the extent to which structural properties of metabolic networks are correlated with their biological function. The assumption underlying this work is that molecular networks are robust and that metrics associated with robustness could be used to assess robustness of other networks.

In § 6, structural properties of entire metabolic networks are correlated with metabolic function. Here, the aim is to discover whether structural properties describing entire metabolic networks are predicitive of biological function of the whole system. The focus on the whole makes this a *global* approach.

In contrast, the penultimate chapter of this thesis, § 7, takes a *local* approach. Here, we ask whether the topological characteristics of small regions of protein interaction networks correlate with their having certain biological traits.

The thesis concludes with § 8 where we summarize the most important findings of the thesis. We also outline what we consider to be interesting open problems in the field of network biology.

REFERENCES

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell* (Taylor & Francis, 2007).

[2] J. De Las Rivas and C. Fontanillo, *Protein–protein interactions essentials: Key concepts to building and analyzing interactome networks*, PLoS Computional Biology **6**, e1000807 (2010).

[3] O. Aparicio, J. V. Geisberg, E. Sekinger, A. Yang, Z. Moqtaderi, and K. Struhl, *Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo*, in *Current Protocols in Molecular Biology* (John Wiley & Sons, Inc., 2005).

[4] D. B. West, *Introduction to graph theory*, Vol. 2 (Prentice hall Englewood Cliffs, 2001).

2

TOPOLOGY OF MOLECULAR INTERACTION NETWORKS

Wynand WINTERBACH, Piet VAN MIEGHEM, Marcel REINDERS, Huijuan WANG, Dick DE RIDDER

2.1. ABSTRACT

Molecular interactions are often represented as network models which have become the common language of many areas of biology. Graphs serve as convenient mathematical representations of network models and have themselves become objects of study. Their topology has been intensively researched over the last decade after evidence was found that they share underlying design principles with many other types of networks.

Initial studies suggested that molecular interaction network topology is related to biological function and evolution. However, further whole-network analyses did not lead to a unified view on what this relation may look like, with conclusions highly dependent on the type of molecular interactions considered and the metrics used to study them. It is unclear whether global network topology drives function, as suggested by some researchers, or whether it is simply a byproduct of evolution or even an artefact of representing complex molecular interaction networks as graphs.

Nevertheless, network biology has progressed significantly over the last years. We review the literature, focusing on two major developments. First, realizing that molecular interaction networks can be naturally decomposed into subsystems (such as modules and pathways), topology is increasingly studied locally rather than globally. Second, there is a move from a descriptive approach to a predictive one: rather than correlating

This chapter was published in BMC Systems Biology **7**, 90 (2013) [1].

biological network topology to generic properties such as robustness, it is used to predict specific functions or phenotypes.

Taken together, this change in focus from globally descriptive to locally predictive points to new avenues of research. In particular, multi-scale approaches are developments promising to drive the study of molecular interaction networks further.

2.2. INTRODUCTION

Over the last half century, our understanding of life at the molecular level has advanced tremendously. This is made possible by continuously improving technology for measuring the presence or concentrations of molecules at a genome-wide level, such as the microarray (transcriptomics), mass spectrometry (proteomics, metabolomics) and next-generation sequencing (genomics). Perhaps more importantly from a systems biology perspective, similar technology and protocols have been developed to measure interactions among molecules, leading to so-called *interactomics* [2]. Protein-protein interactions are measured using yeast-two-hybrid technology and tandem affinity purification amongst others [3], and stored in a variety of databases [4]; interactions between DNA and proteins, such as histones and transcription factors, are found using yeast-one-hybrid and chromatin immunoprecipitation [5] and deposited in databases such as JASPAR [6] and FactorBook [7]; enzyme-metabolite interactions are measured using enzymatic assays and can be found in for example, BRENDA [8], KEGG [9] and MetaCyc [10]. Besides physical interactions, many indirect interactions have been reported, such as genetic interactions [11], general epistatic interactions [12] and predicted functional interactions [13].

This molecular interaction data is the cornerstone of many computational approaches aiming to analyze, model, interpret and predict biological phenomena, many at a genome-wide scale [14]. Interactions are often thought of as constituting networks, a view already proposed quite early [15] which recently came to full fruition [16]. Networks are now used as vehicles for modeling, storing, reporting, transmitting and interpreting molecular interactions [17]. Often they are represented as graphs, although this is not straightforward for many molecular interactions. For example, metabolic networks, representing physical interactions between enzymes and metabolites as well as conversions between metabolites, are ideally represented by hypergraphs [18] but are often reduced to simple graphs [19] for further analysis.

Although graphs are convenient representations of molecular interaction networks, it was quickly realized that they could be treated similarly to large systems of interacting particles: small sets of interactions might be difficult to understand, but statistical properties relating to all interactions could contain valuable information [20]. This led to **network biology** [21]: a combination of systems biology, graph theory and computational and statistical analyses in which the topology of the graphs representing molecular interaction networks themselves became the subject of study. In subsequent work, statistically maintained properties, such as scale-freeness, were found in molecular networks of different types. In similar analyses, graphs were mined for statistically over-represented network motifs [22], small subgraphs, suggesting that certain interaction patterns are common to many networks [23].

Despite their apparent universality, it proved difficult to derive biological conclu-

sions from the patterns discovered in these initial global statistical analyses of molecular interaction networks. They may therefore be labeled as *descriptive*, pointing at generic underlying properties rather than leading to verifiable hypotheses. In time, molecular interactions networks were studied more locally, leading to more tangible biological insights. For example, clustering was used to discover significant biological modules and their interconnection patterns, which shed some light on evolutionary constraints of organisms [24]. Ranking of nodes by topological features (such as degree) was shown to relate to biological importance of a gene or protein and may for example be used to prioritize targets for development of pharmaceuticals [25]. We label such approaches *suggestive*. Finally, by studying networks even more locally, typically neighborhoods surrounding a few nodes, it has become possible to derive *predictive* results from molecular interaction networks. A typical approach is to compute a topological fingerprint of the neighborhood around a node; nodes are found to be functionally similar when their fingerprints are similar [26].

Over the past decade, network biology has thus transformed from being an initially descriptive approach to a predictive tool that is routinely applied to discover biologically relevant facts. In this survey, we chart this progression, showing that it corresponds well to a focus change from global to local. Many reviews of developments in network biology have appeared over the last years; here we list those most closely related to ours. Pržulj [27] reviews the use of protein interaction networks in network biology, touching on some of the techniques discussed throughout this review and calling for more integration of biological knowledge with network theory. A review of network theory from the perspective of data mining may be found in Pavlopoulos *et al.* [28]. This review covers a variety of network metrics with an especially strong focus on clustering and node centrality. Likewise, Cho *et al.* [14] review several data-mining approaches applicable to molecular networks. A related topic is that of random molecular networks, which serve as benchmarks against which data mining results are measured. Such networks are generally produced through processes mimicking evolution, several of which are reviewed by Foster *et al.* [29] and Sun & Kim [30]. Finally, many recent reviews focus on the use of network biology in diagnosing disease [31–33], in particular network-based disease markers.

Our review adds to the existing literature by taking a high-level view of network biology as moving from descriptive to predictive, and by maintaining a clear focus on research exploiting the topology of molecular interaction graphs. The remainder of the paper is organized as follows: in Section 2.3, a brief overview of relevant biological and mathematical theory is presented. Sections 2.4-2.6 then give a chronological overview of research on the graph topology of molecular interaction networks, moving from descriptive to suggestive and predictive. We end with a conclusion and outlook in Section 2.7.

2.3. NETWORK BIOLOGY

For the purposes of this review, we define network biology to be the study of the topology of graph representations of molecular interaction networks, both to describe such networks and as a tool to make biological predictions. We briefly review graph theory and discuss graph representations of molecular interaction networks.

2.3.1. GRAPH THEORY

Graph theory is the study of **graphs**: structures representing relationships between pairs of objects. The set $\mathcal{N}$ of objects in a graph G are called **nodes**; the relationships between the objects are captured by a set $\mathcal{L}$ of node pairs called **links**. When nodes u and v are linked (i.e. $\{u, v\} \in \mathcal{L}$), u is said to be a **neighbor** of v and vice-versa. In **directed graphs**, used for modeling non-symmetric relationships such as activation or repression, each link is directed and has a source node (origin) and a target node (destination). The number of neighbors of a node u is called its **degree**. Figure 2.2 shows examples of directed graphs. **Weighted graphs** model non-binary relations by associating scalars or **weights** with links. An example is the affinity with which proteins bind to one another. Box 2.3.1 lists some metrics often used to study graphs. Many more metrics in the context of network biology are covered in [28].

An **induced subgraph** G' of G is a subset of the nodes of G, along with all links whose endpoint nodes are both in G'. In a **bipartite graph**, the nodes can be split into two sets such that no two vertices in the same set are adjacent. A complete bipartite graph in which all nodes from the first set are connected to all nodes in the second is said to be **complete**.

Degree Distribution The statistical distribution followed by the degrees of the nodes in a network. Many real-world networks have degree distributions that depart sharply from those of classical random network models (Box 2.4.1).

Path Metrics In an unweighted graph G, the shortest path between nodes u and v is the minimum number of links one must traverse to move from u to v. If G is weighted, the shortest path is that with the minimal sum of link weights. The **average shortest path** or **characteristic path length** is the average length of all shortest paths (between all node pairs) in a network

Centrality Metrics A centrality metric gives a ranking of nodes according to their "importance". The simplest measure is **degree centrality** – the degree of a node specifies its importance. **Closeness centrality** is the reciprocal of the sum of the shortest paths to all other nodes (i.e.a node whose closeness centrality is high is close to many nodes). **Betweenness centrality** is the fraction of shortest paths passing through a node. **Eigenvector centrality** and **Pagerank** are measures of how frequently one arrives at a node when performing a random walk on a network.

Box 2.3.1: Graph metrics reduce structural properties of network to (vectors of) real numbers, facilitating the comparison of different networks.

2.3.2. MOLECULAR INTERACTION NETWORKS

Molecular biology is the study of all cellular processes involving DNA, RNA, proteins and metabolites. A simplified overview of common interactions between these molecules is shown in Figure 2.1 (a). Although simplified, models such as Figure 2.1 (a) are still complex. Researchers generally study models with fewer molecules and interactions, such as the signaling pathway model in Figure 2.1 (b).

Association networks Association networks model *any* kind of relation between molecules (e.g. binding, co-expression and structural similarities). Examples of association networks are **gene co-expression networks** and **protein similarity networks**.

Functional networks Functional networks model functional relations between pairs of molecules (usually genes or proteins). A link implies that both are involved in the same function, process or phenotype. **Genetic interaction networks** represent interactions where a double mutation leads to an epistatic effect, i.e., worse or better than expected based on the single mutation.

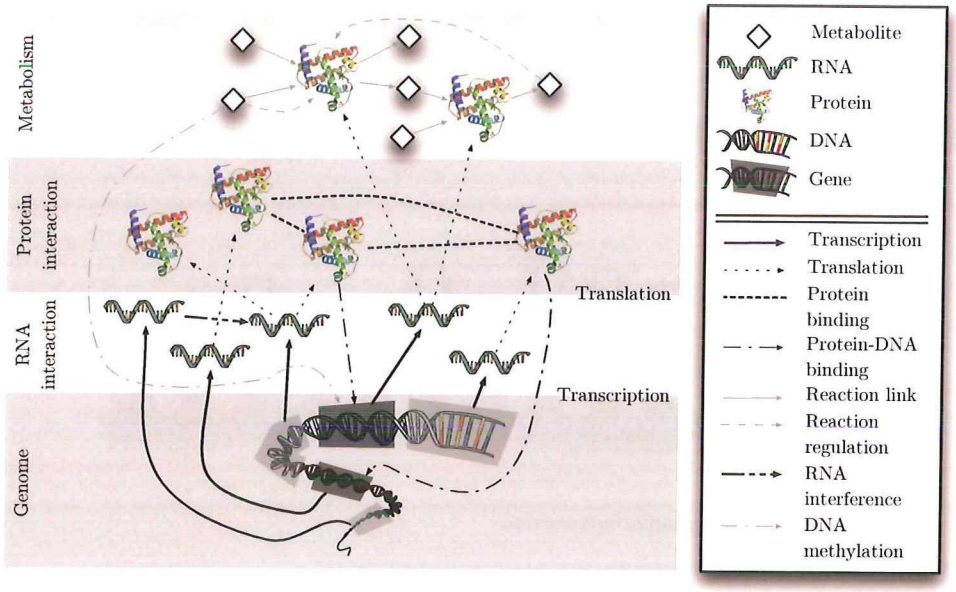
Protein-protein Interaction Networks (PPI Networks) Protein-protein interaction networks are undirected networks that model protein binding. PPI networks are derived from high-throughput experiments using techniques such as yeast two-hybrid screening, mass spectrometry and tandem affinity purification [3]. **Signaling networks** are related to protein interaction networks, but their links are directed according to the flow of molecular signals.

Transcription-regulatory Networks (TR Networks) Transcription-regulatory networks are bipartite networks with one set of nodes representing genes and the other representing transcription factors (TFs). TFs are products of genes (modeled by gene-TF links) whilst genes are regulated by TFs (modeled by TF-gene links). Data for such networks is derived through the process of chromatin immunoprecipitation (ChIP) [34]. **Gene regulatory (GR) networks** are related to TR networks but contain only genes. Their links represent indirect regulatory relationships.

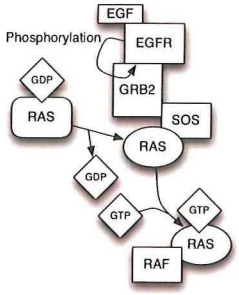
Metabolic Networks Metabolic Networks are bipartite networks that model the relationships between the chemical reactions that occur in cells and the substrates involved in the reactions (the solid gray lines in Figure 2.1 (a)). Reduced, non-bipartite metabolic networks containing only metabolites or only reactions are also often studied.

Box 2.3.2: Commonly studied molecular interaction networks.

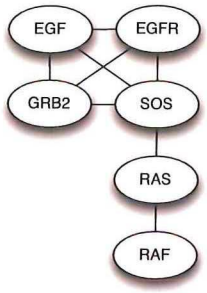
Both Figures 2.1 (a) and (b) focus on interactions and can therefore be represented as networks. But neither is a graph, since Figure 2.1 (b) contains non-pairwise relationships and Figure 2.1 (a) contains multiple types of relationships while both contain multiple types of nodes. Complex interaction models that distinguish between node and link types are useful when the focus of study is on a small molecular subsystem but a hindrance when the aim is the discovery of *interaction patterns* across large sets of interactions. When pattern discovery is the aim, networks are reduced to graphs by including only links and nodes modeling one or two concepts and by converting non-pairwise links to pairwise links. The graph in Figure 2.1 (c) is one possible simplification of the pathway in Figure 2.1 (b). While **network** and **graph** are thus two distinct concepts, we will henceforth use the term **network** to refer to both concepts. Box 2.3.2 lists several such networks commonly studied.



(a) Simple overview of molecular interactions in the cell.

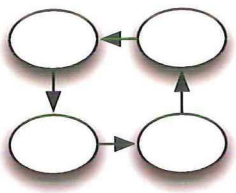


(b) Part of the MAPK/ERK pathway modeled as a network.

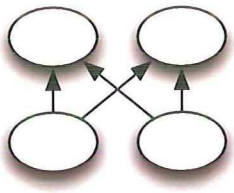


(c) Homogenous protein interaction graph representation of part of the MAPK/ERK pathway.

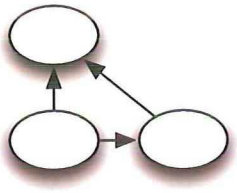
Figure 2.1: From biological models to networks.



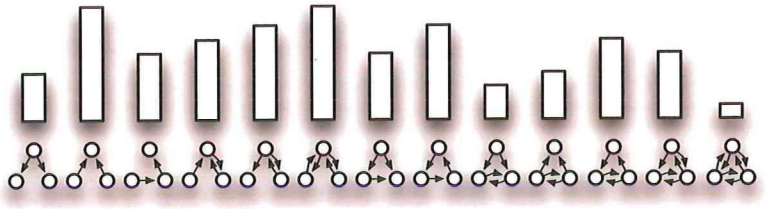
(a) A four-node feed-back motif.



(b) A four-node bi-fan motif.



(c) A three-node feed-forward motif.



(d) Three-node motif signature for a network.

Figure 2.2: Some motifs thought to be overrepresented in molecular interaction networks. Arrowheads indicate link directionality.

2.4. DESCRIPTIVE ANALYSIS

During the 1990's, researchers in various scientific fields started studying macro-scale systems in which individual entities locally interact in simple ways, leading to complex behavior emerging at a global scale. Examples include telecommunications networks [20, 43], social relationship structures [36] and biological interactions from the molecular to the ecological scale [22].

The structure of the above networks departed significantly from the random network models – the Erdős-Renyí model [35] and the Watts-Strogatz model [36] – commonly used in that day to model large networks (see Box 2.4.1). Real-world networks had short average path lengths and degree distributions approximating power laws[20]. The slopes of the degree distributions, when plotted on log-log axes, tended to fall within a narrow range, regardless of the numbers of nodes in these networks. This independence of scale or **scale-freeness** was thought be indicative of networks formed through gradual growth processes based on **preferential attachment**: every time a node is added to a network, it is linked to existing nodes with probabilities proportional to the degrees of those nodes [20, 21].

In biology, initial studies on molecular interaction networks matched the topologies observed in other real-world networks. Gene co-expression networks [44], protein-

Erdős-Renyí (ER) [35] The oldest class of random networks. To construct a graph instance, links are added between each pair of nodes with probability p (a parameter).

Watts-Strogatz (WS) [36] A kind of generalization of ER networks in which links of a regular lattice are rewired. Characterized by high clustering coefficients and short average path lengths.

Barabási-Albert (BA) [20] A class of random networks constructed one node at a time, with new nodes preferentially attaching to existing high-degree nodes. These networks are scale-free (i.e.hub-like) and more closely resemble molecular interaction network networks than ER or WS networks.

Duplication-divergence These networks, inspired by gene duplication and subsequent divergence (in sequence, interaction and function) [37] are generated by duplicating nodes and randomly removing/adding links. Architecturally, duplication-divergence networks are similar to Barabási-Albert networks [38, 39]

Fixed node degrees Random networks characterized by their specific node degree sequences that are generated either by randomly rewiring the links of an existing network [40] or through the configuration model [41, 42].

Box 2.4.1: In graph theory, topological characteristics of a network are often compared to those of instances of random network models. Listed are a few widely used random network models in which nodes represent a single concept; these are generally unsuitable for generating networks in which nodes correspond to multiple concepts (e.g.metabolites and reactions in metabolic networks) since additional structural constraints apply to their connectivity.

protein interaction networks [45], metabolic networks [46] and transcription regulation networks [21] all contain aspects of scale-free networks. Nevertheless, although various random network models reproduce some salient properties of molecular networks, each has been criticized for not being consistent with other important aspects of molecular networks [47–50].

Molecular networks are often also highly clustered, implying **modular** design (see Box 2.4.2) and supporting the idea that biological systems are modular at all levels [51]. An early study on the *S.cerevisiae* PPI network showed proteins with similar functional annotations to be highly connected, strongly suggesting modularity [26]. Similarly, in the yeast TR network, highly co-expressed genes were found to be clustered [52]. Evidence for hierarchical modularity was found in a PPI network [53] and in the metabolic networks of several organisms [54]. In general, molecular interaction networks were increasingly thought to consist of modules, linked through connector or linker nodes [55]. In other words, molecular networks are networks of networks that can tolerate disruptions to individual modules but whose functions are sensitive to disruptions module of connectors.

Although early attempts at understanding molecular interaction networks took a top-down approach, characterizing networks using global metrics such as their degree distributions, it was soon suggested that global behavior of the cell could be the result

Modules are induced subgraphs whose link density is high in comparison to the rest of the graph. This definition is deliberately vague, as what constitutes a module depends on the context and the algorithm used to discover modules.

Motifs are small subgraphs, usually of 3 or 4 nodes, whose over- or underrepresentation may indicate that their structures are important or detrimental to the system [22]. Usually, all distinct motifs in a network are counted, yielding a motif signature for the network that may then be compared to signatures obtained by sampling from an appropriate random network null model (see Box 2.4.1) to determine over- or underrepresentation. A signature for all motifs on 3 nodes is shown in Figure 2.2 (d). Motif signatures can be used to characterize networks.

Graphlets are similar to motifs but always fully connected. As with motifs, graphlets are used to construct signatures that capture the local characteristics of a network [56].

Box 2.4.2: Modules, motifs and graphlets. These concepts are used to decompose networks into smaller units that are easier to study.

of local features [57], a bottom-up view. One view was that behavior of molecular interaction networks emerges from the interactions of many small subgraphs or **motifs** (see Box 2.4.2), in the same way that the behavior of a computer results from the interactions of simple logic circuits [22]. Statistical overrepresentation of a motif is thought to be evidence that the motif offers a functional advantage to its host organism. Such motifs – feed-back loops, feed-forward loops and bi-fan motifs (see Figure 2.2) – all have analogues in the electronic world [22]. This fitted well with the increasing popularity of systems biology [58] that advocated an engineering-inspired approach to study biology. Simple motifs may act as sign-sensitive delay mechanisms or as input response-accelerators, depending on their mix of activators and repressors [23]. More complex motifs may even act as logic circuits, switches and memory states, making them interesting building blocks for synthetic biology [59].

Motifs can also be used to characterize networks more globally. Global motif signatures were found to be unique for different types of networks [22] but conserved between organisms [60], providing further evidence that motifs embody underlying design principles in different types of molecular interaction networks, that are preserved across evolution [23].

The global, module and motif views led to the idea that molecular networks are organized at multiple levels of complexity [61]. At the local level, motifs act as small control circuits or building blocks. Motifs aggregate into modules that, through the interactions of their motifs, implement more complex biological processes. At the global level, modules are connected to each other – and may thus exchange information or molecules – through a small number of linker nodes. The fact that certain topological features, such as scale-free degree distributions, are common among molecular networks suggests that the designs of these networks are shaped at all levels by evolutionary mechanisms.

The case for an architecture based on a hierarchy of motifs, modules and global

properties was strong and it appeared to be universal, so that its presence came to be assumed. At the local level, overrepresented motifs were used to filter spurious links from noisy high-throughput networks by rejecting links that did not form part of motif structures [62]. At the global level, the assumption of power-law degree distributions led researchers to propose the evolutionary processes of duplication and divergence as leading to preferential attachment in the formation of molecular networks [37].

2.4.1. LIMITS TO THE DESCRIPTIVE APPROACH

Details of the multi-layered view were increasingly disputed as data quality improved and as researchers revisited interpretations of older findings. At the global level, the most contested trait was that of scale-freeness, a property found to arise under many circumstances, challenging its significance [63]. Careful examination of molecular interaction data showed that some non-scale-free distributions fit degree distributions of molecular networks as well as scale-free distributions [64, 65]. More contentious was the suggestion that some global features are modeling artifacts. The hub-like architecture of protein interaction networks was questioned, since no protein can realistically bind to the number of proteins suggested by hub nodes; hub nodes are more likely to represent groups of proteins that only appear to be individuals owing to experimental limitations [47]. Likewise, metabolic networks do not display short average path lengths when metabolite paths are traced; shortest path algorithms on metabolic networks do not take into account the requirement that all metabolites be present for a reaction to occur and their direct application to these networks is meaningless [18].

At the module level, it was found that modules are less clearly delineated than previously assumed. There appeared to be many connections between modules, making it difficult to distinguish linker nodes [66]. Without linker nodes, assignment of nodes to modules is more difficult, leading to "fuzzy" modules. Motifs were also criticized. The bi-fan motif, found to be overrepresented in molecular networks [22] and assumed to be functionally important, was shown to have no characteristic behavior when considered as a dynamic system [67]. If motifs lack characteristic behavior, aggregates of motifs, such as motif clusters, cannot be assumed to implement specialized biological functions. Motif signatures (Box 2.4.2 and Figure 2.2 (d)) of networks were argued to be by-products of simple evolutionary mechanisms (such as gene duplication and divergence) [68]. Evolution may thus not be driven by motifs; rather, motifs may be the inevitable result of the self-organizing effects of evolution.

Although there is less universal structure in molecular networks than once thought, the original multi-layered model is still useful, albeit with some modifications. There is much evidence that molecular networks are not scale-free, but they are generally heavy-tailed [65], meaning that they have a few hubs and many low-degree nodes. Motifs may not be simple biological circuits [22], but they established the idea that local structure is important; one way in which this was later exploited was to compute node signatures for use in function prediction in molecular networks [56] and alignment of molecular networks [69]. Perhaps the most important contribution of the layered view was the idea that molecular networks are organized at multiple levels; the molecular organization of the cell cannot be understood at one scale only.

2.4.2. TOPOLOGICAL FEATURES AS TARGET OR BY-PRODUCT OF EVOLUTION

The global approach was not meant to be purely descriptive: its original goal was the discovery of universal architectural features. Universality suggests that organisms are selected *because* they posses such features and would provide clues about the topological requirements that are essential to life.

One property thought to emerge from natural selection is *robustness*, the ability to maintain function under perturbations [70]. Network biologists have sought to explain robustness in terms of topological characteristics. In PPI networks, the number of interaction partners of nodes initially appeared to correlate with their essentiality [57]: robustness may come from the fact that PPI networks have few hubs and many low-degree nodes. In metabolic networks, almost the opposite is true, with networks being susceptible to disruption of low-degree linker nodes that connect metabolic modules [71]. However, in both cases the systems are resilient to most perturbations but susceptible to targeted attacks, a property known as *highly optimized tolerance* [72].

After-the-fact attempts to match topology to properties such as robustness were eventually called into question. *In silico* evolution experiments with simple gene-regulatory networks showed that many such structural features emerge from network dynamics rather than selective pressure [73]. Other such network evolution experiments suggested that the drivers could be simple processes such as reuse, genetic drift and mutation [68, 74, 75]. Even higher-level organization such as modularity is thought to arise from such simple processes [24]. A study comparing a metabolic network to a network of atmospheric chemical reactions found large topological similarities and concluded that many large-scale topological features have no functional nor evolutionary significance, the so-called **neutral theory of chemical reaction networks**[76]. In bacteria, horizontal gene transfer is thought to play an important role in module formation, as cells adopt clusters of foreign genetic material wholesale in reaction to environmental variability [77]. Nevertheless, the extent of this influence was recently questioned, stressing possible interplay between variability and gene transfer [78, 79].

Not all network features emerge through network dynamics. Selection pressure does seem necessary for the fine-tuning of topological features and may in some cases be responsible for the difference between a robust and fragile network [80]. In simulations of metabolic network evolution, hubs emerge when networks are selected for their ability to grow [81]. In models of GR network evolution, sparsity (i.e.low link counts) emerges when selectional stability (which models energy minimization of the mutation process) is enforced [82]. Even modularity may rely on selection pressure, albeit in a more subtle form. When networks are evolved and selected for their ability to prosper in varying conditions, modularity is found to emerge and, crucially, to be maintained [83]. A similar result was obtained by subjecting randomly generated metabolic networks (i.e., not generated by a procedure mimicking evolution) to a range of environments and assessing the amount of biomass they produced [84].

2.5. SUGGESTIVE ANALYSIS

Since the early days of network biology, data mining was used to discover unexpected (ir)regularities in molecular interaction networks. Some findings were already discussed in Section 2.4 (the use of clustering to discover functional annotation, the existence of hub proteins). While data mining techniques shed light on aspects of biological function, they do not necessarily lead to directly testable hypotheses. In this sense, we call the methods in this section "suggestive". We describe four strategies for extracting network regularities: significant feature detection, clustering, central and hub node discovery and network homology.

Significant Feature Detection The idea behind this strategy is that unlikely patterns in molecular networks are indicative of underlying "design" processes (such as evolution). The likelihood of a feature is determined by considering its distribution in network instances generated using a random network model (see Box 2.4.1). In early work, PPI networks were rewired (link pairs were shuffled) to generate random networks [40]. The connections between high-degree nodes in the original protein interaction network were found to be statistically unlikely in rewired networks, leading to the hypothesis that interactions between high-degree proteins are suppressed in evolution, perhaps to control cross-talk in the cell. Modules and motifs [22] can also be considered as significant features. Some of the clustering algorithms mentioned earlier in this section explicitly assess cluster significance as a function of its likelihood [85].

Such significant features can sometimes be biologically interpreted. Statistical analysis of miRNA targets in a human signaling network found that miRNAs tend to target proteins that are part of positive feedback motifs [86]. Similarly, cancer genes tend to be part of positive feedback motifs whilst genes that are highly methylated tend to be part of negative feedback motifs [87]. In both of these cases, the motifs are interpreted as amplification or dampening circuits, analogous to electronic circuits. An interesting recent view is that individual motifs are not necessarily significant but that large clusters of positive or negative feedback motifs act as stochastic amplifiers or dampers, respectively [88].

The advantage of significant feature detection lies in its simplicity: existing techniques are used to analyze and compare the input network and networks derived from a random model. But this is also its main drawback: choosing an incorrect random network model can make features appear significant when they are not.

Clusters Modules in complex systems tend to be highly internally connected whilst sharing only a few connections with the outside world. Graph clustering is an approach to discover such modules by decomposing a network into a number of subnetworks or **clusters** that are internally highly connected. The "big data" era has inspired development of clustering algorithms that efficiently deal with large datasets.

In network biology, general clustering algorithms have been used to discover functional modules in gene co-expression networks [89] and genomic cooccurence networks [90]. Since proteins in complexes highly interact with one another, graph clustering has also been used to discover protein complexes in PPI networks [55]. Here we mention a few of such general clustering algorithms; the interested reader is referred to [91] for a more

thorough overview. Most modern clustering algorithms are based on physical models, data mining techniques or spatial partitioning. Physics-inspired approaches include spin models [92, 93], random walk models [94, 95] and synchronization models [96]. Data mining approaches treat cluster discovery as a problem of significant feature discovery. A few clustering algorithms discussed below are (at least partially) based on this idea. Spatial partitioning approaches associate distance metrics on pairs of nodes that are then clustered using approaches such as k-means clustering. A number of such distance metrics are discussed later in the context of "neighborhood homology" later in this review.

Whilst general algorithms can be applied to molecular networks, clustering algorithms that exploit the specific structure of molecular networks may achieve better results. MCODE is a heuristic algorithm developed to detect complexes in protein interaction networks [97]. Other examples include Restricted Neighborhood Search Clustering [98] and CODENSE, an algorithm for finding dense subgraphs [99]. A number of algorithms based on local neighborhood statistics were proposed as well, for example to find subgraphs of PPI networks that are active according to high-throughput measurements (ActiveModules [100] and MATISSE [101]). More generally, a likelihood score for the density of a subgraph can be used in (greedy) optimization algorithms to mine dense subgraphs, such as in CEZANNE, which finds functional modules in gene co-expression networks [101].

Besides fully connected clusters, clusters that resemble bi-cliques (complete bi-partite subgraphs, see Section 2.3.1) have been shown to be common and biologically relevant in protein interaction networks [102]. Furthermore, clusters in bipartite networks such as TR and metabolic networks are also manifested as bi-clique-like networks. Algorithms have been proposed to mine such (bi-)clique clusters [103, 104]. Specialized algorithms for bipartite networks have also been developed, such as SAMBA, that integrates additional biological data to discover modules [105].

A still-difficult problem is the discovery of overlapping clusters. Many molecules are components of multiple modules (e.g. proteins are part of multiple protein complexes, metabolites are inputs to multiple metabolic reactions) whilst most existing clustering algorithms place each molecule in exactly one cluster. A relatively simple approach is to group molecules in topics and to apply node-based clustering on the topics; a node that belongs to topics in different clusters would be a member of (at least) two clusters. Recent research uses the more restricted case of edge clustering (which is equivalent to topic clustering on topics of two nodes each) with good success [106–108].

Clustering is a useful technique to gain understanding of the modular construction of a molecular network, but caution is required. Recovered clusters may not reflect actual biological modules; inaccurate clustering can arise from badly chosen clustering criteria (in particular from criteria unrelated to biological constraints) [109]. Algorithms that produce overlapping clusters may assign nodes to too many or too few clusters and rigorous techniques for handling such problems are still lacking.

Central Nodes and Hubs Early findings in network biology suggested that some nodes are more important or *central* [110] (see Box 2.3.1) in molecular interaction networks. This manifestation of highly optimized tolerance entails that the survival of an organism

depends more on the presence of a few central nodes than on most other, less central nodes. First, it was found that disrupting the highly connected, "hub-like" p53 gene in the human signaling leads to cancer [111]. It was subsequently shown that the number of interaction partners of a protein (i.e.,*degree centrality*) in the *S.cerevisiae* protein interaction network is correlated with its lethality [57]. Research on protein interaction networks [112], co-expression networks [113] and synthetic genetic interaction networks [114] showed similar correlations. Furthermore, the number of interaction partners was shown to be negatively correlated with the rate of evolution in protein interaction networks [115], metabolic networks [116] and transcription-regulatory networks [117], further supporting the idea that central nodes are important.

Closeness centrality was used to find central metabolites in metabolic networks [118]. Betweenness centrality was used to identify bottleneck nodes – nodes of low degree whose removal is fatal to the organism [119]. Both of these metrics fit the interpretation of central nodes as being chemical flow routers. In signaling networks, disruption of central nodes has been linked to cancer, suggesting that they act as information coordinators/routers [120, 121]. However, not all centrality measures can be easily related to routing, examples of which include subgraph centrality [122], coreness centrality [123], bipartivity (the fraction of closed loops including the node that are of even length) [124] and node hierarchy [125].

In spite of the initial positive findings, further experiments on *S. cerevisiae* showed little correlation between protein degree and essentiality [126], a finding strengthened by computer simulations of gene expression [127]. This cast doubt on the use of centrality measures alone to predict node functionality. Some researchers have sought to refine the notion of centrality by considering interaction patterns of central nodes: those that interact with many interaction partners simultaneously are called "party" hubs whilst those that interact with a few of their partners at a time are called "date" hubs [128]. Party hubs are thought to be global coordinators that connect components within network modules whilst date hubs may be local coordinators that connect network modules [128]. However, this distinction has been challenged with the availability of new data that does not show such clear distinctions between central nodes [129].

Even if node centrality is not as well correlated with node function as hoped, research in this field has shown that hubs do tend to be more essential than non-hubs. Furthermore, subversion of central nodes has been implicated in the formation of cancer [120, 130], suggesting possibly useful drug targets.

It has been suggested that a simple explanation for the essentiality of high degree nodes is that they are more likely to interact with essential complexes and their removal breaks such complexes [126]. The implication is that local topology is a deciding factor in essentiality. Indeed, versions of existing centrality measures modified to take more local information into account are better at predicting which nodes are essential [131]. However, it is important not to conflate node essentiality, a concept tied to survivability, with the influence that a node exerts on a network. The latter concept is discussed in the next section in the guise of "controllability".

Global Homology The principle of homology states that biological systems related by evolution are structurally similar. Its converse – structural similarities imply common

heritage – is often used to predict the function of unknown proteins and genes. In networks, topological similarity can likewise be used to infer functional similarity. Using this approach, metabolic networks of 43 organisms were found to display hierarchical modularity [54]; these modules were found to center around core metabolites [132]. In the same vein, the connectivity of a protein in a PPI network was shown to be proportional to its age. In a study on three species, common proteins are likely to be older than those present in only a single species [133].

The approaches above focus on high-level similarities between networks without attempting to match individual nodes in the networks. By performing such alignments, clustering and significant feature detection applied across species can lead to more insight. In an early example, the glycolytic pathways of 17 organisms were aligned [134] and revealed many interesting differences between species in this essential part of metabolism. Alignment of the *E.coli* metabolic network to those of other organisms identified enzymes whose genes were candidates for horizontal gene transfer [39]. The average degree of these candidates is higher than that of other enzymes, implying that they are central to metabolism. Thus, ancestors to *E.coli* replaced their central enzymes with better functioning enzymes from other species.

Data Mining in Biological Networks Suggests Biological Findings Data mining techniques have been successfully applied in network biology to suggest biological functions for genes and proteins. The common theme is that instead of considering global properties of biological networks, they focus on subnetworks, from individual nodes to neighborhoods and features shared between networks. This increased focus allows the derivation of more tangible biological results. However, when analyses are based on comparisons to random network models (Box 2.4.1), such as in significant feature detection, the problem of telling these apart from evolutionary by-products remains.

2.6. PREDICTIVE ANALYSIS

The data mining approaches discussed in Section 2.5 reveal the large-scale organization of molecular networks in some detail but do not, in general, yield testable biological hypotheses. Approaches that do give such results tend to be based on network generalizations of existing principles in molecular biology: guilt-by-association, homology and differential analysis.

Guilt-by-association The principle of guilt-by-association is based on the observation that if most of the interaction partners of a molecule are associated with some property (such as a specific biological process or molecular function [135]), the molecule itself is also likely to be associated with that property [136]. Guilt-by-association has been used to assign functions to proteins with unknown roles based on the functions shared by the majority of their direct neighbors (i.e.interaction partners) in protein interaction networks [26]. The properties shared by the majority of a node's neighbors do not necessarily yield the best annotations [137] and more sophisticated approaches, such as Markov random fields trained on node neighborhoods [138], have been developed as alternatives.

By only taking direct interactions into account, the above applications of guilt-by-association ignore the impact of potentially informative indirect interactions. So-called n-hop features have been used to predict disease associations of proteins in PPI networks [139]. Another technique for incorporating indirect neighbors is graph diffusion, an idea derived from the study of diffusion in physical systems. Here, properties of nodes are diffused across links in a network; properties that diffuse in high quantities to nodes with unknown roles are used to annotate these nodes [140]. In both n-hop methods and graph diffusion, interaction strength between nodes depends on the path structure between the nodes.

Path structure need not be the only determinant of interaction strength. Nodes that are members of the same biological module may have similar functions [26]. Thus, a node whose role is unknown can be annotated with the functions appearing most frequently in the module(s) to which it belongs. Whilst we do not know what the biological modules are, we can compute approximate modules through clustering. Such an approach has been used to annotate unknown proteins in *S.cerevisiae* protein interaction networks [103]. Guilt-by-association is a simple and effective technique that extends naturally to networks. However, it is only effective when the roles of the majority of molecules in a network are known, limiting the technique to well-studied organisms.

Neighborhood Homology Since the use of homology is pervasive in biology, we expect the principle to extend to networks. Indeed, in Section 2.5 it was already discussed how networks found in different organisms have similar structural properties. Predictive approaches use topological and possibly biological similarity to match similar nodes across different networks. Once nodes are aligned, the function of a protein or gene whose role is unknown can be predicted, if the function of its matched node in the other network is known.

The first network alignment algorithms operated at a local level, attempting to match only small parts of entire networks to one another [69, 141]. Global alignment is more difficult, because networks to be aligned generally differ in size. Moreover, homology is not a one-to-one relation: many nodes may align to many nodes. There are two main approaches for performing global alignment:

1. Cluster the nodes in each network and compute topological matching scores on the clusters [142, 143] ("matching clusters").

2. Select groups of nodes in different networks that are pairwise similar in local neighborhoods and possibly biological labels [144, 145] ("clustering matches").

The first type of algorithm has the disadvantage that the clustering step precedes matching and thus ignores potentially useful information. Many algorithms of the second type associate feature vectors of topological (and possibly biological) attributes with nodes that are then used to compute node similarity. Various metrics have been used [146]. The Jaccard coefficient, a measure of overlap between sets of binary attributes, has been widely used, an example of which was the prediction of protein function in human PPI networks [147]. The h-confidence metric [148] is a data-mining tool for discovering associations and has been used in protein function prediction. Specialized metrics, such

as the graphlet distance (tailored to graphlet signatures[56]) have been used to discover genes implicated in cancer [149].

Variations of clustering algorithms, looking for dense subgraphs within one network, have been proposed to mine subgraphs similar in two networks. For example, the PathBlast algorithm combined a statistical score for protein similarity and probability of a reported protein interaction to mine pathways or complexes occurring in PPI networks of different species [141]. Similar approaches were applied to assign functions to proteins [150] and to align metabolic pathways [151].

Differential Analysis Diagnosis of many diseases (such as cancer) is based on the fact they influence the regulation programs of cells. Traditionally, this involved finding changed expression of marker genes, or specific gene mutations, i.e.focusing on the nodes in the network. Network biology allows additional focus on node relations, making it possible to diagnose molecular diseases that cannot be well characterized by the traditional techniques [152]. This so-called differential analysis, finding changes in network structure [32], is currently complicated by the fact that construction of high-quality molecular networks requires considerable time and resources. One common way around this is to use an existing high-quality network, typically a PPI or TR network, as a scaffold onto which noisy high-throughput patient data (typically gene expression or methylation data) is overlaid. If multiple measurements are available for each patient, gene coexpression/comethylation values can be computed and overlaid as link weights on PPI links.

Expression changes of genes/proteins linked to central nodes in molecular networks have been proven to be reliable markers of disease. Differential expression around topologically central nodes in protein interaction networks has been used to diagnose cancer [153, 154]. Disease central nodes (i.e., nodes implicated in disease) have been similarly used in the diagnosis of breast cancer and leukemia [155]. More recently, coexpression changes around biologically central nodes, such as signaling hubs, have shown to be even more reliable disease markers [156, 157].

More elaborate differential approaches consider changes in expression patterns of subnetworks, instead of only central nodes. Automatic extraction of such subnetworks based on topology and measurements such as gene expression has revealed subnetworks associated with cancer (in which differential gene/protein expression could be used for diagnosis of the disease) [87, 158] as well as subnetworks that are implicated in heart failure [159]. An alternative to automatic extraction is to use biological modules based on theoretical knowledge; such an approach has been used in cancer prognosis [160].

Differential diagnosis, despite its relative newness has quickly grown to a large field. Our discussion is necessarily limited by the scope of this review; the interested reader is referred to recent reviews that consider the discipline in much more depth [32, 33, 161].

Relating Topology to Biological Properties Leads to Predictive Power The data mining techniques discussed in Section 2.5 are mostly based on topological information. In contrast, the predictive approaches discussed above depend on additional biological information. This approach to network biology clearly yields more testable hypotheses

than the suggestive and descriptive approaches.

Since we do not, in general, have good models of biological function at large scales, predictive approaches are most often applied to small groups of nodes or subnetworks. There are exceptions with metabolic networks being the most prominent. Flux balance analysis (FBA) [162, 163] is a framework for computing steady-state reaction rates in such networks based on reaction stoichiometry, assuming the cell attempts to achieve some objective such as maximum growth. FBA is often used in a predictive way, but has also been applied in a "suggestive" setting, e.g.to study robustness of metabolic networks [71]. FBA allows one to take additional physical constraints into account, such as thermodynamic interactions [164] or responses to signaling [165]; for an extensive overview see [166].

The biggest problem with incorporating additional biological knowledge into existing models is that, for any given biological attribute, we seldom have complete data. Two recent ideas, "controllability" and "observability", potentially allow to use partial (local) knowledge to predict global state. Controllability refers to "driver" nodes that have a large influence on the state of a system [167]; observability is almost complementary, focusing on a small set of appropriately chosen observation nodes whose properties allow reconstruction of the global system state [168]. These techniques promise to allow associating local information with driver/observation nodes and to predict global properties from limited available data.

2.7. CONCLUSION AND OUTLOOK

In this review, we have summarized common research themes in the field of network biology. We find a slow movement from global to local analysis, arguing that this trend emerged from a need to draw more concrete biological knowledge from networks.

The survey findings seem to suggest that one must either choose between untestable abstract hypotheses about large-scale topological patterns or small-scale results that neglect large-scale topology. But the successes of local techniques lie not in their focus on the local but because *they tightly couple topological observations to biological knowledge*. From this starting point, we see two broad research directions for improving the explanatory power of large-scale topology patterns. The first approach is theoretical and is aimed at making descriptive and suggestive techniques more predictive, whilst the second approach is practical and extends the predictive techniques to work at larger topological scales.

The theoretical research direction entails the improvement of network evolution models in order that they reproduce as much of the topological aspects of real molecular networks as possible. Better models of network evolution can better reveal the topological features that are by-products of evolution, permitting researchers to concentrate on explaining topological results that cannot be explained by the models. An additional benefit is that these models could themselves lead to biological insight.

In the practical direction, we propose the application of predictive techniques to various "resolutions" of molecular networks, that is, multi-resolution analysis. Lower resolution versions of a network are typically obtained by grouping subnetworks into meta-nodes (by analogy, the entire street network of a city is represented by a single city node in national road maps). How nodes are grouped depends on the topological

properties that must be maintained in low-resolution network versions. Node clustering techniques from Section 2.5 can be used to produce low-resolution networks by grouping node clusters into meta-nodes. Another promising technique that aims to maintain random-walk properties is **spectral coarse graining** [169].

The two research directions outlined above are by no means the only possible paths for developing network biology. Rather, they show this young field still has much potential for development; we expect that future researchers will bring us unexpected biological insights with the help of network biology.

REFERENCES

[1] W. Winterbach, P. Van Mieghem, M. Reinders, H. Wang, and D. de Ridder, *Topology of molecular interaction networks,* BMC Systems Biology **7**, 90 (2013).

[2] M. Cusick, N. Klitgard, M. Vidal, and D. Hill, *Interactome: Gateway into systems biology,* Human Molecular Genetics **14**, R171 (2005).

[3] J. De Las Rivas and C. Fontanillo, *Protein–protein interactions essentials: Key concepts to building and analyzing interactome networks,* PLoS Computional Biology **6**, e1000807 (2010).

[4] S. Orchard, S. Kerrien, S. Abbani, B. Aranda, J. Bhate, S. Bidwell, A. Bridge, L. Briganti, F. Brinkman, G. Cesareni, A. Chatr-aryamontri, E. Chautard, C. Chen, M. Dumousseau, J. Goll, R. Hancock, L. Hannick, I. Jurisica, J. Khadake, D. Lynn, U. Mahadevan, L. Perfetto, A. Raghunath, S. Ricard-Blum, B. Roechert, L. Salwinski, V. Stumpflen, M. Tyers, P. Uetz, I. Xenarios, and H. Hermjakob, *Protein interaction data curation: The International Molecular Exchange (IMEx) consortium,* Nature Methods **9**, 345 (2012).

[5] B. Dey, S. Thukral, S. Krishnan, M. Chakrobarty, S. Gupta, C. Manghani, and V. Rani, *DNA-protein interactions: Methods for detection and analysis,* Molecular and Cellular Biochemistry **365**, 279 (2012).

[6] E. Portales-Casamar, S. Thongjuea, A. Kwon, D. Arenillas, X. Zhao, E. Valen, D. Yusuf, B. Lenhard, W. Wasserman, and A. Sandelin, *JASPAR 2010: The greatly expanded open-access database of transcription factor binding profiles,* Nucleic Acids Research **38**, D105 (2010).

[7] J. Wang, J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield, M. C. Greven, B. G. Pierce, X. Dong, A. Kundaje, Y. Cheng, O. J. Rando, E. Birney, R. M. Myers, W. S. Noble, M. Snyder, and Z. Weng, *Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors,* Genome Research **22**, 1798 (2012).

[8] M. Scheer, A. Grote, A. Chang, I. Schomburg, C. Munaretto, M. Rother, C. Söhngen, M. Stelzer, J. Thiele, and D. Schomburg, *BRENDA, the enzyme information system in 2011,* Nucleic Acids Research **39**, D670 (2011).

[9] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, *KEGG for integration and interpretation of large-scale molecular datasets*, Nucleic Acids Research **40**, D109 (2012).

[10] R. Caspi, T. Altman, K. Dreher, C. A. Fulcher, P. Subhraveti, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, *et al.*, *The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases*, Nucleic Acids Research **40**, D742 (2012).

[11] M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, H. Ding, J. L. Koh, K. Toufighi, S. Mostafavi, J. Prinz, R. P. St. Onge, B. VanderSluis, T. Makhnevych, F. J. Vizeacoumar, S. Alizadeh, S. Bahr, R. L. Brost, Y. Chen, M. Cokol, R. Deshpande, Z. Li, Z.-Y. Lin, W. Liang, M. Marback, J. Paw, B.-J. San Luis, E. Shuteriqi, A. H. Y. Tong, N. van Dyk, I. M. Wallace, J. A. Whitney, M. T. Weirauch, G. Zhong, H. Zhu, W. A. Houry, M. Brudno, S. Ragibizadeh, B. Papp, C. Pál, F. P. Roth, G. Giaever, C. Nislow, O. G. Troyanskaya, H. Bussey, G. D. Bader, A.-C. Gingras, Q. D. Morris, P. M. Kim, C. A. Kaiser, C. L. Myers, B. J. Andrews, and C. Boone, *The genetic landscape of a cell*, Science **327**, 425 (2010), http://www.sciencemag.org/content/327/5964/425.full.pdf.

[12] E. S. Snitkin and D. Segrè, *Epistatic interaction maps relative to multiple metabolic phenotypes*, PLoS Genetics **7**, e1001294 (2011).

[13] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. Jensen, and C. von Mering, *The STRING database in 2011: Functional interaction networks of proteins, globally integrated and scored*, Nucleic Acids Research **39**, D561 (2011).

[14] D.-Y. Cho, Y.-A. Kim, and T. Przytycka, *Chapter 5: Network biology approach to complex diseases*, PLoS Computional Biology **8**, e1002820 (2012).

[15] N. Rashevsky, *Topology and life: In search of general mathematical principles in biology and sociology*, Bulletin of Mathematical Biology **16**, 317 (1954).

[16] T. Ideker and D. Lauffenburger, *Building with a scaffold: Emerging strategies for high- to low-level cellular modeling*, Trends in Biotechnology **21**, 255 (2003).

[17] R. Albert, *Network inference, analysis, and modeling in systems biology*, The Plant Cell **19**, 3327 (2007).

[18] M. Arita, *The metabolic world of* escherichia coli *is not small*, Proceedings of the National Academy of Sciences of the USA **101**, 1543 (2004).

[19] T. Aittokallio and B. Schwikowski, *Graph-based methods for analysing networks in cell biology*, Briefings in Bioinformatics **7**, 243 (2006), http://bib.oxfordjournals.org/content/7/3/243.full.pdf+html.

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

[21] A.-L. Barabási and Z. N. Oltvai, *Network biology: Understanding the cell's functional organization, Nature Reviews Genetics*, Nature Reviews Genetics **5**, 101 (2004).

[22] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, *Network motifs: Simple building blocks of complex networks*, Science **298**, 824 (2002).

[23] U. Alon, *Network motifs: Theory and experimental approaches*, Nature Reviews Genetics **8**, 450 (2007).

[24] K. Takemoto, *Metabolic network modularity arising from simple growth processes*, Physical Review E **86**, 036107 (2012).

[25] J. Chen, B. Aronow, and A. Jegga, *Disease candidate gene identification and prioritization using protein interaction networks*, BMC Bioinformatics **10**, 73 (2009).

[26] B. Schwikowski, P. Uetz, and S. Fields, *A network of protein-protein interactions in yeast*, Nature Biotechnology **18**, 1257 (2000).

[27] N. Pržulj, *Protein-protein interactions: Making sense of networks via graph-theoretic modeling*, Bioessays **33**, 115 (2011).

[28] G. Pavlopoulos, M. Secrier, C. Moschopoulos, T. Soldatos, S. Kossida, J. Aerts, R. Schneider, and P. Bagos, *Using graph theory to analyze biological networks*, BioData Mining **4**, 10 (2011).

[29] D. V. Foster, S. A. Kauffman, and J. E. S. Socolar, *Network growth models and genetic regulatory networks*, Physical Review E **73**, 031912 (2006).

[30] M. Sun and P. Kim, *Evolution of biological interaction networks: From models to real data*, Genome Biology **12**, 235 (2011).

[31] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, *Network medicine: A network-based approach to human disease*, Nature Reviews Genetics **12**, 56 (2011).

[32] T. Ideker and N. J. Krogan, *Differential network biology*, Molecular Systems Biology **8** (2012).

[33] M. W. Gonzalez and M. G. Kann, *Chapter 4: Protein interactions and disease*, PLoS Computational Biology **8**, e1002819 (2012).

[34] O. Aparicio, J. V. Geisberg, E. Sekinger, A. Yang, Z. Moqtaderi, and K. Struhl, *Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo*, in *Current Protocols in Molecular Biology* (John Wiley & Sons, Inc., 2005).

[35] B. Bollobás, *Random graphs*, 2nd ed., Cambridge Studies in Advanced Mathematics (Cambridge University Press, 2001).

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

[37] A. Rzhetsky and S. M. Gomez, *Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome,* Bioinformatics **17**, 988 (2001).

[38] S. Teichmann and M. Babu, *Gene regulatory network growth by duplication,* Nature Genetics **36**, 492 (2004).

[39] S. Light, P. Kraulis, and A. Elofsson, *Preferential attachment in the evolution of metabolic networks,* BMC Genomics **6**, 159 (2005).

[40] S. Maslov and K. Sneppen, *Specificity and stability in topology of protein networks,* Science **296**, 910 (2002).

[41] M. Newman, *Random graphs as models of networks,* in *Handbook of graphs and networks: From the genome to the internet,* edited by S. Bornholdt and H. G. Schuster (Wiley-VCH, Berlin, 2003) pp. 35–68.

[42] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras, *Generation of uncorrelated random scale-free networks,* Physical Review E **71**, 027103 (2005).

[43] R. Albert, H. Jeong, and A.-L. Barabási, *Internet: Diameter of the world-wide web,* Nature **401**, 130 (1999).

[44] S. L. Carter, C. M. Brechbühler, M. Griffin, and A. T. Bond, *Gene co-expression network topology provides a framework for molecular characterization of cellular state,* Bioinformatics **20**, 2242 (2004).

[45] J.-C. Rain, L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schachter, Y. Chemama, A. Labigne, and P. Legrain, *The protein–protein interaction map of* helicobacter pylori, Nature **409**, 211 (2001).

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

[47] C.-J. Tsai, B. Ma, and R. Nussinov, *Protein-protein interaction networks: How can a hub protein bind so many different partners?* Trends in Biochemical Sciences **34**, 594 (2009).

[48] G. Lima-Mendez and J. van Helden, *The powerful law of the power law and other myths in network biology. Mol. BioSyst.*, Molecular Biosystems **5**, 1482 (2009).

[49] A. Samal and O. C. Martin, *Randomizing genome-scale metabolic networks,* PLoS ONE **6**, e22295 (2011).

[50] G. Basler, O. Ebenhöh, J. Selbig, and Z. Nikoloski, *Mass balanced randomization of metabolic networks,* Bioinformatics (2011), 10.1093/bioinformatics/btr145.

[51] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, *From molecular to modular cell biology,* Nature **402**, C47 (1999).

[52] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai, *Revealing modular organization in the yeast transcriptional network,* Nature Genetics **31**, 370 (2002).

[53] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley, K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg, *A protein interaction map of* drosophila melanogaster, Science **302**, 1727 (2003).

[54] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, *Hierarchical organization of modularity in metabolic networks,* Science **297**, 1551 (2002).

[55] A. W. Rives and T. Galitski, *Modular organization of cellular networks,* Proceedings of the National Academy of Sciences of the USA **100**, 1128 (2003).

[56] T. Milenković and N. Pržulj, *Uncovering biological network function via graphlet degree signatures.* Cancer Informatics **6**, 257 (2008).

[57] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, *Lethality and centrality in protein networks.* Nature **411**, 41 (2001).

[58] H. Kitano, *Computational systems biology,* Nature **420**, 206 (2002).

[59] R. Entus, B. Aufderheide, and H. Sauro, *Design and implementation of three incoherent feed-forward motif based biological concentration sensors,* Systems and Synthetic Biology **1**, 119 (2007), 10.1007/s11693-007-9008-6.

[60] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, *Superfamilies of evolved and designed networks,* Science **303**, 1538 (2004).

[61] M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann, *Structure and evolution of transcriptional regulatory networks.* Current Opinion in Structural Biology **14**, 283 (2004).

[62] B. Zhang and S. Horvath, *A general framework for weighted gene co-expression network analysis.* Statistical Applications in Genetics and Molecular Biology **4** (2005), 10.2202/1544-6115.1128.

[63] E. Fox Keller, *Revisiting "scale-free" networks,* BioEssays **27**, 1060 (2005).

[64] R. Khanin and E. Wit, *How scale-free are biological networks.* Journal of Computational Biology **13**, 810 (2006).

[65] A. L. G. de Lomana, Q. K. Beg, G. de Fabritiis, and J. Villà-Freixa, *Statistical analysis of global connectivity and activity distributions in cellular networks*, Journal of Computational Biology (2010).

[66] N. N. Batada, T. Reguly, A. Breitkreutz, L. Boucher, B.-J. Breitkreutz, L. D. Hurst, and M. Tyers, *Still stratus not altocumulus: Further evidence against the date/party hub distinction*, PLoS Biology **5**, e154+ (2007).

[67] P. J. Ingram, M. P. Stumpf, and J. Stark, *Network motifs: Structure does not determine function*, BMC Genomics **7**, 108+ (2006).

[68] P. D. Kuo, W. Banzhaf, and A. Leier, *Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence*, Biosystems **85**, 177 (2006).

[69] J. Berg and M. Lässig, *Local graph alignment and motif search in biological networks*. Proceedings of the National Academy of Sciences of the USA **101**, 14689 (2004).

[70] C. H. Waddington, *Canalization of development and the inheritance of acquired characters*, Nature **150**, 563 (1942).

[71] A. G. Smart, L. A. N. Amaral, and J. M. Ottino, *Cascading failure and robustness in metabolic networks*, Proceedings of the National Academy of Sciences of the USA **105**, 13223 (2008).

[72] T. Hase, H. Tanaka, Y. Suzuki, S. Nakagawa, and H. Kitano, *Structure of protein interaction networks and their implications on drug design*, PLoS Computational Biology **5**, e1000550 (2009).

[73] A. Bergman and M. L. Siegal, *Evolutionary capacitance as a general feature of complex gene networks*, Nature **424**, 549 (2003).

[74] S. Maslov, S. Krishna, T. Y. Pang, and K. Sneppen, *Toolbox model of evolution of prokaryotic metabolic networks and their regulation*, Proceedings of the National Academy of Sciences of the USA **106**, 9743 (2009).

[75] T. Y. Pang and S. Maslov, *A toolbox model of evolution of metabolic pathways on networks of arbitrary topology*, PLoS Computational Biology **7**, e1001137 (2011).

[76] S. H. Lee, S. Bernhardsson, P. Holme, B. J. Kim, and P. Minnhagen, *Neutral theory of chemical reaction networks*, New Journal of Physics **14**, 033032 (2012).

[77] A. Kreimer, E. Borenstein, U. Gophna, and E. Ruppin, *The evolution of modularity in bacterial metabolic networks*, Proceedings of the National Academy of Sciences of the USA **105**, 6976 (2008).

[78] W. Zhou and L. Nakhleh, *Convergent evolution of modularity in metabolic networks through different community structures*, BMC Evolutionary Biology **12**, 181 (2012).

[79] K. Takemoto, *Does habitat variability really promote metabolic network modularity?* PLoS ONE **8**, e61348 (2013).

[80] S. Ciliberti, O. C. Martin, and A. Wagner, *Robustness can evolve gradually in complex regulatory gene networks with varying topology,* PLoS Computional Biology **3**, e15+ (2007).

[81] T. Pfeiffer, O. S. Soyer, and S. Bonhoeffer, *The evolution of connectivity in metabolic networks,* PLoS Biology **3**, e228 (2005).

[82] R. D. Leclerc, *Survival of the sparsest: Robust gene networks are parsimonious,* Molecular Systems Biology **4** (2008), 10.1038/msb.2008.52.

[83] N. Kashtan and U. Alon, *Spontaneous evolution of modularity and network motifs,* Proceedings of the National Academy of Sciences of the USA **102**, 13773 (2005).

[84] A. Samal, A. Wagner, and O. Martin, *Environmental versatility promotes modularity in genome-scale metabolic networks,* BMC Systems Biology **5**, 135 (2011).

[85] R. Sharan, I. Ulitsky, and R. Shamir, *Network-based prediction of protein function.* Molecular Systems Biology **3** (2007), 10.1038/msb4100129.

[86] Q. Cui, Z. Yu, E. O. Purisima, and E. Wang, *Principles of microrna regulation of a human cellular signaling network,* Molecular Systems Biology **2** (2006).

[87] Q. Cui, Y. Ma, M. Jaramillo, H. Bari, A. Awan, S. Yang, S. Zhang, L. Liu, M. Lu, M. O'Connor-McCourt, E. O. Purisima, and E. Wang, *A map of human cancer signaling,* Molecular Systems Biology **3** (2007).

[88] M. Kittisopikul and G. M. Süel, *Biological role of noise encoded in a genetic network motif,* Proceedings of the National Academy of Sciences of the USA **107**, 13300 (2010).

[89] A. Ben-Dor, R. Shamir, and Z. Yakhini, *Clustering gene expression patterns,* Journal of Computational Biology **6**, 281 (1999).

[90] B. Snel, P. Bork, and M. A. Huynen, *The identification of functional modules from the genomic association of genes,* Proceedings of the National Academy of Sciences of the USA **99**, 5890 (2002).

[91] S. Fortunato, *Community detection in graphs,* Physics Reports **486**, 75 (2010).

[92] M. Blatt, S. Wiseman, and E. Domany, *Superparamagnetic clustering of data,* Physical Review Letters **76**, 3251 (1996).

[93] S.-W. Son, H. Jeong, and J. D. Noh, *Random field ising model and community structure in complex networks,* The European Physical Journal B - Condensed Matter and Complex Systems **50**, 431 (2006).

[94] S. van Dongen, *Graph clustering by flow simulation,* Ph.D. thesis, University of Utrecht (2000).

[95] W. E, T. Li, and E. Vanden-Eijnden, *Optimal partition and effective dynamics of complex networks*, Proceedings of the National Academy of Sciences of the USA **105**, 7907 (2008).

[96] A. Pluchino, V. Latora, and A. Rapisarda, *Changing opinions in a changing world: A new perspective in sociophysics*, International Journal of Modern Physics C **16**, 515 (2005).

[97] G. Bader and C. Hogue, *An automated method for finding molecular complexes in large protein interaction networks*, BMC Bioinformatics **4**, 2+ (2003).

[98] A. D. King, N. Pržulj, and I. Jurisica, *Protein complex prediction via cost-based clustering*, Bioinformatics **20**, 3013 (2004).

[99] H. Hu, X. Yan, Y. Huang, J. Han, and X. J. Zhou, *Mining coherent dense subgraphs across massive biological networks for functional discovery*, Bioinformatics **21**, i213 (2005).

[100] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, *Discovering regulatory and signalling circuits in molecular interaction networks*, Bioinformatics **18**, S233 (2002).

[101] I. Ulitsky and R. Shamir, *Identifying functional modules using expression profiles and confidence-scored protein interactions*, Bioinformatics **25**, 1158 (2009).

[102] A. Thomas, R. Cannings, N. Monk, and C. Cannings, *On the structure of protein–protein interaction networks*, Biochemical Society Transactions **31**, 1491 (2003).

[103] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, and R. Chen, *Topological structure analysis of the protein–protein interaction network in budding yeast*, Nucleic Acids Research **31**, 2443 (2003).

[104] H. Liu, J. Liu, and L. Wang, *Searching maximum quasi-bicliques from protein-protein interaction network*, Journal of Biomedical Science and Engineering **1**, 200 (2008).

[105] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir, *Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data*, Proceedings of the National Academy of Sciences of the USA **101**, 2981 (2004).

[106] R. W. Solava, R. P. Michaels, and T. Milenković, *Graphlet-based edge clustering reveals pathogen-interacting proteins*, Bioinformatics **28**, i480 (2012).

[107] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, *Link communities reveal multiscale complexity in networks*, Nature **466**, 761 (2010).

[108] E. Becker, B. Robisson, C. E. Chapple, A. Guénoche, and C. Brun, *Multifunctional proteins revealed by overlapping clustering in protein interaction network*, Bioinformatics **28**, 84 (2012).

[109] S. Fortunato and M. Barthélemy, *Resolution limit in community detection*, Proceedings of the National Academy of Sciences of the USA **104**, 36 (2007).

[110] L. C. Freeman, *Centrality in social networks conceptual clarification*, Social Networks **1**, 215 (1978-1979).

[111] B. Vogelstein, D. Lane, and A. J. Levine, *Surfing the p53 network*, Nature **408**, 307 (2000).

[112] H. Yu, D. Greenbaum, H. X. Lu, X. Zhu, and M. Gerstein, *Genomic analysis of essentiality within protein networks*, Trends in Genetics **20**, 227 (2004).

[113] S. Bergmann, J. Ihmels, and N. Barkai, *Similarities and differences in genome-wide expression data of six organisms*, PLoS Biology **2**, e9 (2003).

[114] A. H. Tong, G. Lesage, G. D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G. F. Berriz, R. L. Brost, M. Chang, Y. Chen, X. Cheng, G. Chua, H. Friesen, D. S. Goldberg, J. Haynes, C. Humphries, G. He, S. Hussein, L. Ke, N. Krogan, Z. Li, J. N. Levinson, H. Lu, P. Ménard, C. Munyana, A. B. Parsons, O. Ryan, R. Tonikian, T. Roberts, A.-M. Sdicu, J. Shapiro, B. Sheikh, B. Suter, S. L. Wong, L. V. Zhang, H. Zhu, C. G. Burd, S. Munro, C. Sander, J. Rine, J. Greenblatt, M. Peter, A. Bretscher, G. Bell, F. P. Roth, G. W. Brown, B. Andrews, H. Bussey, and C. Boone, *Global mapping of the yeast genetic interaction network*, Science **303**, 808 (2004).

[115] H. B. Fraser, A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman, *Evolutionary rate in the protein interaction network*, Science **296**, 750 (2002).

[116] D. Vitkup, P. Kharchenko, and A. Wagner, *Influence of metabolic network structure and function on enzyme evolution*, Genome Biology **7**, R39 (2006).

[117] Y. Wang, E. A. Franzosa, X.-S. Zhang, and Y. Xia, *Protein evolution in yeast transcription factor subnetworks*, Nucleic Acids Research **38**, 5959 (2010).

[118] H.-W. Ma and A.-P. Zeng, *The connectivity structure, giant strong component and centrality of metabolic networks*, Bioinformatics **19**, 1423 (2003).

[119] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein, *The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics*. PLoS Computational Biology **3**, e59+ (2007).

[120] L. Li, C. Tibiche, C. Fu, T. Kaneko, M. F. Moran, M. R. Schiller, S. S.-C. Li, and E. Wang, *The human phosphotyrosine signaling network: Evolution and hotspots of hijacking in cancer*, Genome Research **22**, 1222 (2012).

[121] E. Wang, *Understanding genomic alterations in cancer genomes using an integrative network approach*, Cancer letters (2013).

[122] E. Estrada and J. A. Rodríguez Velázquez, *Subgraph centrality in complex networks*, Physical Review E **71**, 056103+ (2005).

[123] S. Wuchty and E. Almaas, *Peeling the yeast protein network,* Proteomics **5**, 444 (2005).

[124] E. Estrada, *Protein bipartivity and essentiality in the yeast protein-protein interaction network.* Journal of Proteome Research **5**, 2177 (2006).

[125] N. Bhardwaj, P. M. Kim, and M. B. Gerstein, *Rewiring of transcriptional regulatory networks: Hierarchy, rather than connectivity, better reflects the importance of regulators,* Science Signaling **3**, ra79 (2010).

[126] X. He and J. Zhang, *Why do hubs tend to be essential in protein networks?* PLoS Genetics **2**, e88 (2006).

[127] M. Siegal, D. Promislow, and A. Bergman, *Functional and evolutionary inference in gene networks: Does topology matter?* Genetica **129**, 83 (2007).

[128] J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal, *Evidence for dynamically organized modularity in the yeast protein–protein interaction network,* Nature **430**, 88 (2004).

[129] S. Agarwal, C. M. Deane, M. A. Porter, and N. S. Jones, *Revisiting date and party hubs: Novel approaches to role assignment in protein interaction networks,* PLoS Computional Biology **6**, e1000817+ (2010).

[130] D. Breitkreutz, L. Hlatky, E. Rietman, and J. A. Tuszynski, *Molecular signaling network complexity is correlated with cancer patient survivability,* Proceedings of the National Academy of Sciences of the USA **109**, 9209 (2012).

[131] K. Park and D. Kim, *Localized network centrality and essentiality in the yeast–protein interaction network,* Proteomics **9**, 5143 (2009).

[132] P. Holme, M. Huss, and H. Jeong, *Subnetwork hierarchies of biochemical pathways,* Bioinformatics **19**, 532 (2003).

[133] E. Eisenberg, *Preferential attachment in the protein network evolution,* Physical Review Letters **91** (2003), 10.1103/PhysRevLett.91.138701.

[134] T. Dandekar, S. Schuster, B. Snel, M. Huynen, and P. Bork, *Pathway alignment: Application to the comparative analysis of glycolytic enzymes.* The Biochemical journal **343 Pt 1** (1999).

[135] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, M. Harris, D. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. Matese, J. Richardson, M. Ringwald, G. Rubin, and G. Sherlock, *Gene ontology: Tool for the unification of biology,* Nature Genetics **25**, 25 (2000).

[136] S. Oliver, *Proteomics: Guilt-by-association goes global,* Nature **403**, 601 (2000).

[137] J. Gillis and P. Pavlidis, *"guilt by association" is the exception rather than the rule in gene networks,* PLoS Computational Biology **8**, e1002444 (2012).

[138] Y. A. I. Kourmpetis, A. D. J. van Dijk, M. C. A. M. Bink, R. C. H. J. van Ham, and C. J. F. ter Braak, *Bayesian markov random field analysis for protein function prediction based on network data*, PLoS ONE **5**, e9293 (2010).

[139] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal, *Towards a proteome-scale map of the human protein-protein interaction network*, Nature **437**, 1173 (2005).

[140] Y. Qi, Y. Suhail, Y.-y. Lin, J. D. Boeke, and J. S. Bader, *Finding friends and enemies in an enemies-only network: A graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions*, Genome Research **18**, 1991 (2008).

[141] B. P. Kelley, B. Yuan, F. Lewitter, R. Sharan, B. R. Stockwell, and T. Ideker, *Pathblast: A tool for alignment of protein interaction networks*, Nucleic Acids Research **32**, W83 (2004).

[142] H. T. T. Phan and M. J. E. Sternberg, *Pinalog: A novel approach to align protein interaction networks—implications for complex detection and function prediction*, Bioinformatics **28**, 1239 (2012).

[143] Y.-K. Shih and S. Parthasarathy, *Scalable global alignment for multiple biological networks*, BMC Bioinformatics **13**, S11 (2012).

[144] O. Kuchaiev, T. Milenković, V. Memišević, W. Hayes, and N. Pržulj, *Topological network alignment uncovers biological function and phylogeny*, Journal of The Royal Society Interface **7**, 1341 (2010).

[145] R. Patro and C. Kingsford, *Global network alignment using multiscale spectral signatures*, Bioinformatics **28**, 3105 (2012).

[146] G. Pandey, S. Manocha, G. Atluri, and V. Kumar, *Enhancing the functional content of protein interaction networks*, arXiv preprint arXiv:1210.6912 (2012).

[147] M. E. Sardiu, Y. Cai, J. Jin, S. K. Swanson, R. C. Conaway, J. W. Conaway, L. Florens, and M. P. Washburn, *Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics*, Proceedings of the National Academy of Sciences of the USA **105**, 1454 (2008).

[148] G. Pandey, M. Steinbach, R. Gupta, T. Garg, and V. Kumar, *Association analysis-based transformations for protein interaction networks: A function prediction case study*, in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07 (ACM, New York, NY, USA, 2007) pp. 540–549.

[149] T. Milenković, V. Memišević, A. K. Ganesan, and N. Pržulj, *Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data*, Journal of The Royal Society Interface **7**, 423 (2010).

[150] S. Bandyopadhyay, R. Sharan, and T. Ideker, *Systematic identification of functional orthologs based on protein network comparison*, Genome Research **16**, 428 (2006).

[151] Y. Li, D. de Ridder, M. de Groot, and M. Reinders, *Metabolic pathway alignment between species using a comprehensive and flexible similarity measure*, BMC Systems Biology **2**, 111 (2008).

[152] T. Ideker and R. Sharan, *Protein networks in disease*, Genome Research **18**, 644 (2008).

[153] S. Wachi, K. Yoneda, and R. Wu, *Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues*, Bioinformatics **21**, 4205 (2005).

[154] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. L. Wrana, *Dynamic modularity in protein interaction networks predicts breast cancer outcome*, Nature Biotechnology **27**, 199 (2009).

[155] D. Tuck, H. Kluger, and Y. Kluger, *Characterizing disease states from topological properties of transcriptional regulatory networks*, BMC Bioinformatics **7**, 236 (2006).

[156] C. Yao, H. Li, C. Zhou, L. Zhang, J. Zou, and Z. Guo, *Multi-level reproducibility of signature hubs in human interactome for breast cancer metastasis*, BMC Systems Biology **4**, 151 (2010).

[157] J. Li, A. E. G. Lenferink, Y. Deng, C. Collins, Q. Cui, E. O. Purisima, M. D. O'Connor-McCourt, and E. Wang, *Identification of high-quality cancer prognostic markers and metastasis network modules*, Nature Communications **1**, 34 (2010).

[158] H.-Y. Y. Chuang, E. Lee, Y.-T. T. Liu, D. Lee, and T. Ideker, *Network-based classification of breast cancer metastasis*. Molecular Systems Biology **3** (2007), 10.1038/msb4100180.

[159] C.-C. Lin, J.-T. Hsiang, C.-Y. Wu, Y.-J. Oyang, H.-F. Juan, and H.-C. Huang, *Dynamic functional modules in co-expressed protein interaction networks of dilated cardiomyopathy*, BMC Systems Biology **4**, 138 (2010).

[160] S. Efroni, C. F. Schaefer, and K. H. Buetow, *Identification of key processes underlying cancer phenotypes using biologic pathway analysis*, PLoS ONE **2**, e425 (2007).

[161] D.-Y. Cho, Y.-A. Kim, and T. M. Przytycka, *Chapter 5: Network biology approach to complex diseases*, PLoS Computational Biology **8**, e1002820 (2012).

[162] R. U. Ibarra, J. S. Edwards, and B. O. Palsson, Escherichia coli *k-12 undergoes adaptive evolution to achieve* in silico *predicted optimal growth*, Nature **420**, 186 (2002).

[163] N. E. Lewis, K. K. Hixson, T. M. Conrad, J. A. Lerman, P. Charusanti, A. D. Polpitiya, J. N. Adkins, G. Schramm, S. O. Purvine, D. Lopez-Ferrer, K. K. Weitz, R. Eils, R. Konig, R. D. Smith, and B. O. Palsson, *Omic data from evolved* e. coli *are consistent with computed optimal growth from genome-scale models*, Molecular Systems Biology **6** (2010).

[164] A. Hoppe, S. Hoffmann, and H.-G. Holzhutter, *Including metabolite concentrations into flux balance analysis: Thermodynamic realizability as a constraint on flux distributions in metabolic networks*, BMC Systems Biology **1**, 23 (2007).

[165] M. W. Covert, N. Xiao, T. J. Chen, and J. R. Karr, *Integrating metabolic, transcriptional regulatory and signal transduction models in* escherichia coli, Bioinformatics **24**, 2044 (2008).

[166] N. E. Lewis, H. Nagarajan, and B. O. Palsson, *Constraining the metabolic genotype–phenotype relationship using a phylogeny of* in silico *methods*, Nature Reviews Microbiology **10**, 291 (2012).

[167] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabasi, *Controllability of complex networks*, Nature **473**, 167 (2011).

[168] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási, *Observability of complex systems*, Proceedings of the National Academy of Sciences of the USA **110**, 2460 (2013).

[169] D. Gfeller and P. De Los Rios, *Spectral coarse graining of complex networks*, Physical Review Letters **99**, 038701 (2007).

3

ASSORTATIVITY OF COMPLEMENTARY GRAPHS

Huijuan WANG, Wynand WINTERBACH, Piet VAN MIEGHEM,

3.1. ABSTRACT

Newman's measure for (dis)assortativity, the linear degree correlation ρ_D, is widely studied although analytic insight into the assortativity of an arbitrary network remains far from well understood. In this paper, we derive the general relation (3.3), (3.4) and Theorem 1 between the assortativity $\rho_D(G)$ of a graph G and the assortativity $\rho_D(G^c)$ of its complement G^c . Both $\rho_D(G)$ and $\rho_D(G^c)$ are linearly related by the degree distribution in G. When the graph $G(N,p)$ possesses a binomial degree distribution as in the Erdős-Rényi random graphs $G_p(N)$, its complementary graph $G_p^c(N) = G_{1-p}(N)$ follows a binomial degree distribution as in the Erdős-Rényi random graphs $G_{1-p}(N)$. We prove that the maximum and minimum assortativity of a class of graphs with a binomial distribution are asymptotically antisymmetric: $\rho_{\max}(N,p) = -\rho_{\min}(N,p)$ for $N \to \infty$. The general relation (3.4) nicely leads to (a) the relation (3.12) and (3.18) between the assortativity range $\rho_{\max}(G) - \rho_{\min}(G)$ of a graph with a given degree distribution and the range $\rho_{\max}(G^c) - \rho_{\min}(G^c)$ of its complementary graph and (b) new bounds (3.8) and (3.17) of the assortativity. These results together with our numerical experiments in over 30 real-world complex networks illustrate that the assortativity range $\rho_{\max} - \rho_{\min}$ is generally large in sparse networks, which underlines the importance of assortativity as a network characterizer.

This chapter was published in The European Physical Journal B **83**, 2 (2011) [1].

3.2. INTRODUCTION

"Mixing" in complex networks [2, 3] refers to the tendency of network nodes to connect preferentially to other nodes with either similar or opposite properties. Networks whose nodes preferentially connect to nodes with (dis)similar properties, are called (dis)assortative. When the property of interest is the degree of a node, the linear degree correlation coefficient ρ_D measures the assortativity in node degree of a network, which is computed in [4] as

$$\rho_D = 1 - \frac{\sum_{i\sim j}(d_i - d_j)^2}{\sum_{i=1}^{N} d_i^3 - \frac{1}{2L}\left(\sum_{i=1}^{N} d_i^2\right)^2} \tag{3.1}$$

where d_j is the degree of node j and $i \sim j$ denotes that node i and j are linked. Although (3.1) is well suited to computation, it is difficult to interpret. Using derivations from [4], (3.1) can be written as

$$\rho_D = \frac{E[D_i D_j] - \mu_{D_i}^2}{E[D_i^2] - \mu_{D_i}^2} \tag{3.2}$$

where D_i and D_j are the degrees of connected nodes. This expression is exactly the correlation coefficient of the degrees of connected nodes. Networks in which nodes of similar degrees tend to be connected have positive correlation coefficients and are said to be are assortative, whereas networks in which nodes of different degrees tend to be connected have negative correlation coefficient and are said to be disassortative.

Network assortativity was widely studied after it was realized that the degree distribution alone provides an insufficient characterization of complex networks. Networks with the same degree distribution may still differ significantly in various topological features. Consequently, many investigations have focused on (a) exploring the relation between assortativity and other topological properties as well as spectra of networks [5][6][4] and (b) understanding the effect of assortativity on dynamic network processes such as the epidemic spreading [7] and percolation phenomena [8]. Relations between degree correlation and other topological or dynamic features are mostly studied experimentally [5] or in a specific network model [8][7]. Recently, we have verified spectral bounds for the assortativity [4] and we have studied how the modularity changes under degree-preserving rewiring [9], which alters the assortativity of the graph.

Analytic insight in degree correlations in an arbitrary network is still lacking. In this work, we analytically explore the relation between the assortativity $\rho_D(G)$ of graph G and $\rho_D(G^c)$ of its complement G^c. Let G be a graph or a network and let $\mathcal{N}$ denote the set of $N = |\mathcal{N}|$ nodes and $\mathcal{L}$ the set of $L = |\mathcal{L}|$ links. An undirected graph G can be represented by an $N \times N$ symmetric adjacency matrix A, 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. The complement G^c of G is a graph containing all the nodes in G and all the links that are *not* in G. Thus, the adjacency matrix of G^c is $A(G^c) = J - I - A(G)$, where J is the all-one matrix and I is the identity matrix.

Furthermore, the general relation (3.4) between $\rho_D(G)$ and $\rho_D(G^c)$ that we derived is further applied to the complementary classes of graphs with a binomial degree distribution. The binomial degree distribution is a characteristic of an Erdős-Rényi random

graph $G_p(N)$, which has N nodes and any two nodes are connected independently with a probability p. Such a random construction leads to a zero assortativity as proved in [4]. However, the class of graphs $G(N,p)$ with the same binomial degree distribution $\Pr[D_G = k] = \binom{N-1}{k} p^k (1-p)^{N-1-k}$ as Erdős-Rényi random graphs $G_p(N)$ and obtained, for instance, by degree-preserving rewiring features an assortativity that may vary within a wide range between $\min \rho_D$ and $\max \rho_D$. The complementary class $G(N, 1-p)$ possesses also a binomial degree distribution $\Pr[D_{G^c} = k] = \binom{N-1}{k} (1-p)^k p^{N-1-k}$ characterized by N and $1-p$. We derive the relation between the assortativity of a graph with a binomial degree distribution and that of its complementary graph. This relation enabled us to prove, interestingly, that the maximum and minimum achievable assortativity of a class of graphs with a binomial degree distribution is symmetric around 0, $\max \rho_D(N,p) = -\min \rho_D(N,p)$, which is also numerically illustrated.

The general relation (3.4) between $\rho_D(G)$ and $\rho_D(G^c)$ also allows us to derive new bounds of the assortativity and to relate the assortativity ranges $\max \rho_D(G) - \min \rho_D(G)$ and $\max \rho_D(G^c) - \min \rho_D(G^c)$ of two complementary classes of graphs, each with a given degree vector or a degree distribution.

The importance of investigating the assortativity and assortativity range relation of complementary graphs lies in the following aspects. A) Computational complexity of assortative(disassortative) degree-preserving rewiring, which increases (decreases) the assortativity of network whilst the degree of each node remains the same, is higher in a dense network than that in a sparse network [10][4]. Most real-world networks are sparse. However, hierarchical networks at a higher aggregation level tend to be denser. Moreover, most studied brain networks and biological networks are originally weighted networks. These networks are usually transformed into an unweighted network by different link weight thresholds so that classical networking theories can be applied. For each weighted network, unweighted networks usually have to be derived at different link densities without losing the information of the weighted network. Thus, they can be dense with link density ranging over $0.5 < p \ll 1$ and they may even follow a binomial degree distribution [11]. Hence, the assortativity relation between complementary graphs allows the assortative (disassortative) degree-preserving rewiring in a dense network to be derived from the disassortative (assortative) rewiring in its complement with less computational complexity. B) The maximum $\max \rho_D$ and minimum $\min \rho_D$ assortativity reveals to what extent a degree vector d may characterize a graph. A small range $\max \rho_D - \min \rho_D$ emphasizes the determining role of the degree vector d, whereas the opposite underlines the importance of the assortativity. Also, experiments suggest that most complex networks (see the Table 3.1 in Appendix 3.11) can be degree-preservingly rewired in two opposite ways so that $\rho_D < 0$ and, alternatively, so that $\rho_D > 0$. Given this experimental observation, we can say that $\max \rho_D > 0$ and $\min \rho_D < 0$ for the degree vector d of a complex network. Consequently, a small $\max \rho_D - \min \rho_D$ means that the degree vector is "hard" to correlate, because ρ_D needs to be close to zero. Apart from degree vectors $d = r.u$ of regular graphs of degree r, where u is the all-one vector and $d_j = r$ for each component/node j, it would be interesting to find examples of degree vectors of *complex networks* for which $\min \rho_D > 0$. Such degree vectors would generate and characterize a class of strict assortative graphs, where $\min \rho_D > 0$. A non-trivial example of a strict disassortative class of (almost) regular graphs is analyzed in Appendix 3.10, while

the Table 3.1 in Appendix 3.11 shows a couple of real-world complex networks that generate a strict disassortative class. The difference $\max\rho_D - \min\rho_D$ may be regarded as a metric of a given degree vector d that reflects the adaptivity in (dis)assortativity under degree-preserving rewiring. Moreover, the quantity

$$r_G = \frac{\max\rho_D - \rho_D}{\max\rho_D - \min\rho_D}$$

determines the relative maximum assortativity deficiency of a graph, which measures the remaining degree-preserving rewiring left to achieve the maximum assortative state. If degree-preserving rewiring can be considered as an evolutionary process of a network, then r_G quantifies the life-time or the evolutionary state of the network. For example, the functional human brain network of a newly born baby is approximately randomized, with $\rho_D \approx 0$. The learning process rewires the brain and changes ρ_D. Suppose that learning during growth increases ρ_D in that it structures the functional brain, then $1 - r_G$ measures the effect of learning. The maximum possible trained functional brain possesses an assortativity of $\max\rho_D$, which corresponds to learning efficiency $1 - r_G$ equal to 1.

3.3. ASSORTATIVITY OF COMPLEMENTARY GRAPHS

3.3.1. RELATED BY DEGREE SEQUENCE

A node i with degree d_i in graph G has degree $N-1-d_i$ in the corresponding complementary graph G^c. All connected node pairs in G^c are non-connected node pairs $i \nsim j$ in G. Therefore, the assortativity of the complementary graph can be written from (3.1) as

$$\rho_D(G^c) = 1 - \frac{\sum_{i\nsim j}(d_i - d_j)^2}{\sum_{i=1}^{N}(N-1-d_i)^3 - \frac{1}{N(N-1)-2L}\left(\sum_{i=1}^{N}(N-1-d_i)^2\right)^2}$$

where d_i refers to the degree of node i in the original graph G. The variance $\mathrm{Var}[D] = \sigma^2[D]$ of the degree D of an arbitrary node[1] can be written as a function of the degree differences between all node pairs

$$\sigma^2[D] = \sum_{j=2}^{N}\sum_{k=1}^{j-1}\left(\frac{d_j - d_k}{N}\right)^2$$

which is derived in [12]. Furthermore, since

$$N^2\sigma^2[D] = N^2\left(E[D^2] - E^2[D]\right) = N\sum_{i=1}^{N} d_i^2 - 4L^2$$

(where E denotes expectation and L the number of links in the network) we have

$$\sum_{i\sim j}(d_i - d_j)^2 + \sum_{i\nsim j}(d_i - d_j)^2 = N\sum_{i=1}^{N} d_i^2 - 4L^2$$

[1] We use capital letters for random variables and small letters for specific realizations.

Hence,

$$\rho_D(G^c) = 1 - \rho_D(G)\left(\frac{\sum_{i=1}^{N} d_i^3 - \frac{1}{2L}\left(\sum_{i=1}^{N} d_i^2\right)^2}{\sum_{i=1}^{N}(N-1-d_i)^3 - \frac{1}{N(N-1)-2L}\left(\sum_{i=1}^{N}(N-1-d_i)^2\right)^2}\right) + \frac{\left(\sum_{i=1}^{N} d_i^3 - \frac{1}{2L}\left(\sum_{i=1}^{N} d_i^2\right)^2\right) - \left(N\sum_{i=1}^{N} d_i^2 - 4L^2\right)}{\sum_{i=1}^{N}(N-1-d_i)^3 - \frac{1}{N(N-1)-2L}\left(\sum_{i=1}^{N}(N-1-d_i)^2\right)^2} \tag{3.3}$$

where (3.1) has been introduced.

3.3.2. RELATED BY DEGREE DISTRIBUTION

We can rephrase expression (3.3) in terms of random variables. According to [4],

$$\sum_{i=1}^{N} d_i^3 - \frac{1}{2L}\left(\sum_{i=1}^{N} d_i^2\right)^2 = 2L\sigma^2\left[D_{l^+}(G)\right]$$

$$\sum_{i=1}^{N}(N-1-d_i)^3 - \frac{1}{N(N-1)-2L}\left(\sum_{i=1}^{N}(N-1-d_i)^2\right)^2 = (N(N-1)-2L)\,\sigma^2\left[D_{l^+}(G^c)\right]$$

$$N\sum_{i=1}^{N} d_i^2 - 4L^2 = N^2\sigma^2[D]$$

where $\sigma^2[D_{l^+}(G)]$ and $\sigma^2[D_{l^+}(G^c)]$ are the variances of the degrees at one side of an arbitrary link in G and in G^c, respectively. Thus, (3.3) becomes

$$\rho_D(G^c) = -\rho_D(G)\frac{2L\sigma^2[D_{l^+}(G)]}{(N(N-1)-2L)\,\sigma^2[D_{l^+}(G^c)]} + 1 - \frac{N^2\sigma^2[D(G)] - 2L\sigma^2[D_{l^+}(G)]}{(N(N-1)-2L)\,\sigma^2[D_{l^+}(G^c)]} \tag{3.4}$$

which holds for any graph. Observe that, except for $\rho_D(G)$, all factors and terms in (3.3) and (3.4) are constant for a given degree vector. This means that the assortativity $\rho_D(G^c)$ of the complement G^c of a graph linearly varies with the assortativity $\rho_D(G)$ of the graph G, and vice versa.

Theorem 1 *The assortativity relation between complementary graphs (3.4) can be further expressed as a function of the degree distribution* $\Pr[D = k]$ *in the original graph G where*

$$\frac{2L\sigma^2[D_{l^+}(G)]}{(N(N-1)-2L)\,\sigma^2[D_{l^+}(G^c)]} = \frac{E[D^3] - \frac{E^2[D^2]}{E[D]}}{\frac{(N-1)^2E[D^2]-(N-1)^2E^2[D]+(N-1)E[D^2]E[D]-E^2[D^2]}{(N-1-E[D])} - E[D^3]} \tag{3.5}$$

$$\frac{N^2\sigma^2[D(G)] - 2L\sigma^2[D_{l^+}(G)]}{(N(N-1)-2L)\,\sigma^2[D_{l^+}(G^c)]} = \frac{NE[D^2] - NE^2[D] - E[D^3] + \frac{E^2[D^2]}{E[D]}}{\frac{(N-1)^2E[D^2]-(N-1)^2E^2[D]+(N-1)E[D^2]E[D]-E^2[D^2]}{(N-1-E[D])} - E[D^3]} \tag{3.6}$$

Proof. See Appendix 3.7. ■

Relations (3.3), (3.4) and Theorem 1 are equivalent and explicitly reflect how the assortativity $\rho_D(G^c)$ and $\rho_D(G)$ of complementary graphs are linearly related.

3.3.3. BOUNDS FOR THE ASSORTATIVITY

Given a degree distribution or degree sequence, the assortativity $\rho_D(G)$ of a graph may range within

$$-1 \le \min\rho_D \le \rho_D(G) \le \max\rho_D \le 1$$

and, likewise, the assortativity of its complementary graph $\rho_D(G^c)$ may vary within

$$-1 \le \min\rho_D^c \le \rho_D(G^c) \le \max\rho_D^c \le 1$$

where $\max\rho_D$ and $\min\rho_D$ ($\max\rho_D^c$ and $\min\rho_D^c$) are the maximum and minimum achievable assortativity of the (complementary) class of graphs with a given degree vector d.

When $\rho_D(G) = -1$, (3.4) shows, that

$$4L\sigma^2[D_{l^+}(G)] \le N^2\sigma^2[D(G)] \le 2N(N-1)\sigma^2\left[D_{l^+}(G^c)\right]$$

and, when $\rho_D(G) = 1$, that

$$N^2\sigma^2[D(G)] \le 2\left(N(N-1)-2L\right)\sigma^2\left[D_{l^+}(G^c)\right]$$

Thus, if $\min\rho_D = -1$ and $\max\rho_D = 1$,

$$4L\sigma^2[D_{l^+}(G)] \le N^2\sigma^2[D(G)] \le 2\left(N(N-1)-2L\right)\sigma^2\left[D_{l^+}(G^c)\right]$$

Alternatively, after inverting (3.4),

$$\begin{aligned}\rho_D(G) \;=\; & -\frac{(N(N-1)-2L)\,\sigma^2[D_{l^+}(G^c)]}{2L\sigma^2[D_{l^+}(G)]}\rho_D(G^c)+1 \\ & -\frac{N^2\sigma^2[D(G)]-(N(N-1)-2L)\,\sigma^2[D_{l^+}(G^c)]}{2L\sigma^2[D_{l^+}(G)]}\end{aligned} \tag{3.7}$$

we find the bounds for the assortativity and disassortativity of any graph G,

$$\mathfrak{r}_{\min} \le \rho_D(G) \le \mathfrak{r}_{\min} + \frac{(N(N-1)-2L)\,\sigma^2[D_{l^+}(G^c)]}{L\sigma^2[D_{l^+}(G)]} \tag{3.8}$$

where

$$\mathfrak{r}_{\min} = 1 - \frac{N^2\sigma^2[D(G)]}{2L\sigma^2[D_{l^+}(G)]} = 1 - \frac{N\sigma^2[D(G)]}{p(N-1)\sigma^2[D_{l^+}(G)]} \approx 1 - \frac{1}{p}\cdot\frac{\sigma^2[D(G)]}{\sigma^2[D_{l^+}(G)]} \tag{3.9}$$

Thus, we conclude that

$$\begin{aligned}&\min\rho_D \ge \max(-1, \mathfrak{r}_{\min}) \\ &\max\rho_D \le \min\left(1, \mathfrak{r}_{\min} + \frac{2(1-p)}{p}\cdot\frac{\sigma^2[D_{l^+}(G^c)]}{\sigma^2[D_{l^+}(G)]}\right)\end{aligned} \tag{3.10}$$

where $p = L/\binom{N}{2}$ is the link density. The assortativity range $0 \leq \max\rho_D - \min\rho_D \leq 2$ of the class of graphs G and the assortativity range $0 \leq \max\rho_D^c - \min\rho_D^c \leq 2$ of its complementary class can be related by (3.4) as

$$\left(\max\rho_D^c - \min\rho_D^c\right) = \frac{2L\sigma^2\left[D_{l^+}(G)\right]}{\left(N(N-1)-2L\right)\sigma^2\left[D_{l^+}(G^c)\right]}\left(\max\rho_D - \min\rho_D\right) \tag{3.11}$$

or, inverted

$$\left(\max\rho_D - \min\rho_D\right) = \left(\frac{1}{p} - 1\right)\frac{\sigma^2\left[D_{l^+}(G^c)\right]}{\sigma^2\left[D_{l^+}(G)\right]}\left(\max\rho_D^c - \min\rho_D^c\right) \tag{3.12}$$

where both $\sigma^2\left[D_{l^+}(G^c)\right]$ and $\sigma^2\left[D_{l^+}(G)\right]$ have been expressed as a function of the degree distribution of the original graph in Appendix 3.7. The assortativity range $\max\rho_D - \min\rho_D$ is small if (a) the variance $\sigma^2\left[D_{l^+}(G^c)\right]$ is small, (b) $\sigma^2\left[D_{l^+}(G)\right]$ is large and/or the link density p is high (close to 1).

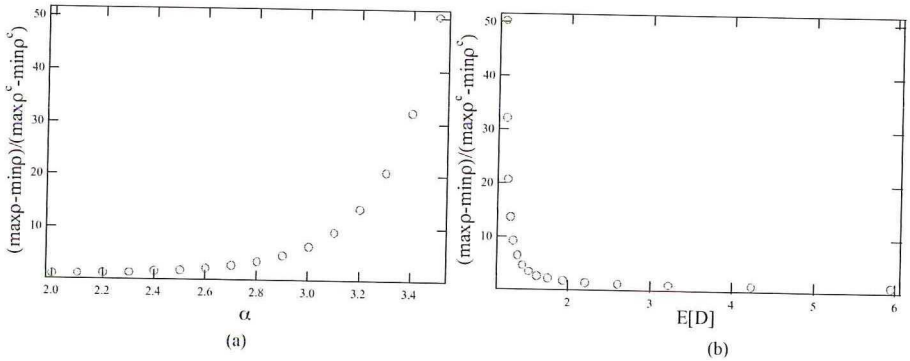


Figure 3.1: The ratio $\Delta = \frac{\max\rho_D - \min\rho_D}{\max\rho_D^c - \min\rho_D^c}$ in graphs with $N = 10000$ nodes and with a power-law degree distribution versus (a) the exponent α of the degree distribution and versus (b) the average degree $E[D]$.

The ratio $\frac{\sigma^2[D_{l^+}(G^c)]}{\sigma^2[D_{l^+}(G)]}$ has been extensively analyzed in Appendix 3.8, in general as well as in graphs with a binomial or a power-law degree distribution. When a graph has a binomial degree distribution $\Pr[D_G = k] = \binom{N-1}{k}p^k(1-p)^{N-1-k}$, $\frac{\sigma^2[D_{l^+}(G^c)]}{\sigma^2[D_{l^+}(G)]} = 1$ as derived both in Section 3.4.1 (rigorously) and in Appendix 3.8 (asymptotically). When a graph has a power-law degree distribution $\Pr[D = k] = ck^{-\alpha}$, where $c = 1/\sum_{k=1}^{N-1} k^{-\alpha}$ and $1 \leq \alpha \leq 3$, $\frac{\sigma^2[D_{l^+}(G^c)]}{\sigma^2[D_{l^+}(G)]} \to 0$ if the graph is large and sparse as proved in Appendix 3.8. We further quantitatively investigate the assortativity range ratio

$$\Delta = \frac{\max\rho_D - \min\rho_D}{\max\rho_D^c - \min\rho_D^c} = \left(\frac{1}{p} - 1\right)\frac{\sigma^2\left[D_{l^+}(G^c)\right]}{\sigma^2\left[D_{l^+}(G)\right]}$$

in graphs with a power-law or binomial degree distribution. In binomial graphs, $\Delta = \frac{1}{p} - 1$. In graphs with $N = 10000$ nodes and with a power-law degree distribution, the ratio Δ, expressed as a function of the degree distribution, can be numerically computed. We

consider power-law graphs with an exponent $2 \leq \alpha \leq 3.5$, since most real-world graphs have $2 \leq \alpha \leq 3$. As shown in Fig. 3.1(a), the ratio of the assortativity range Δ increases as the power exponent α, or the heterogeneity increases. The assortativity of the compliment may still vary within a certain range upbounded by $2/\Delta$ when $2 \leq \alpha \leq 3$, whereas $2/\Delta$ goes fast to zero when $\alpha > 3$. The link density is smaller for a larger exponent α. Hence, the ratio Δ decreases as the average degree/link density increases, as depicted in Fig. 3.1(b).

In general, a sparse network, favors a large assortativity range. This effect of a (small) link density is more evident in graphs with a binomial degree distribution than that in power-law graphs, since $\frac{\sigma^2[D_{l^+}(G^c)]}{\sigma^2[D_{l^+}(G)]}$ is far smaller in power-law graphs. As shown in Fig. 3.3 and 3.4, a power-law graph, indeed, has a smaller assortativity range compare to the binomial graph with the same link density.

When p is large, a non-trivial bound can be derived from (3.12)

$$\max \rho_D - \min \rho_D \leq 2\left(\frac{1}{p} - 1\right)\frac{\sigma^2[D_{l^+}(G^c)]}{\sigma^2[D_{l^+}(G)]} \tag{3.13}$$

Most real-world networks are sparse. However, hierarchical network at a higher aggregation level or the unweighted networks transformed from the original weighted e.g. brain and biological networks, likely have a link density $0.5 < p \ll 1$, as discussed in Section 3.2. The assortativity of such a dense network can be derived from its complement with less computational complexity by the assortativity relation (3.3), (3.4) or Theorem 1. A non-trivial bound of the assortativity range tends to be achieved via the assortativity range relation (3.12). When $p \to 1$, the range of variability in the degrees of a graph with a number of links $L \sim O(N^2)$ is narrow and the assortativity is close to zero as illustrated in Fig. 3.6.

3.4. GRAPHS WITH A BINOMIAL DEGREE DISTRIBUTION

Consider the class of graphs $G(N,p)$ with a binomial degree distribution $\Pr[D_G = k] = \binom{N-1}{k}p^k(1-p)^{N-1-k}$ characterized by N and p as in the Erdős-Rényi (ER) random graphs $G_p(N)$. Its complementary class of graphs $G(N,1-p)$ also possess a binomial degree distribution $\Pr[D_{G^c} = k] = \binom{N-1}{k}(1-p)^k p^{N-1-k}$ with parameter N and $1-p$ as followed by the ER random graphs $G_{1-p}(N)$. The assortativity of connected ER random graphs is zero [4]. However, the assortativity of graphs like $G(N,p)$ conditioned only by a degree distribution can vary with in a large range. Besides its theoretical beauty, the binomial distribution has been observed in e.g. peer-to-peer networks [13] and the unweighted functional brain networks [11].

3.4.1. ASSORTATIVITY OF COMPLEMENTARY GRAPHS

We first explore the relation between the assortativity $\rho_D(G(N,p))$ and $\rho_D(G^c(N,p)) = \rho_D(G(N,1-p))$ of two complementary graphs each having a binomial degree distribution characterized by (N,p) and $(N,1-p)$ respectively, based on Theorem 1.

For a binomial degree distribution $\Pr[D_G = k] = \binom{N-1}{k} p^k (1-p)^{N-1-k}$, it follows that

$$E[D^3] = (N-1)p(1-6p+3Np+6p^2-5Np^2+N^2p^2)$$
$$E[D^2] = (N-1)p(1-2p+Np)$$
$$E[D] = (N-1)p$$

Substituted into Theorem 1 and further into (3.4), we find

$$\frac{\sigma^2[D_{l^+}(G^c)]}{\sigma^2[D_{l^+}(G)]} = 1 \tag{3.14}$$

$$\rho_D(G^c(N,p)) = \rho_D(G(N,1-p)) = -\frac{p}{1-p}\rho_D(G(N,p)) - \frac{2}{(N-2)(1-p)} \tag{3.15}$$

If a graph with a binomial degree distribution is assortative $\rho_D(G(N,p)) > 0$, its complementary graph is definitely disassortative $\rho_D\left(G^c(N,p)\right) < 0$, because $\frac{2}{(N-2)(1-p)} > 0$. The reverse does not hold when N is small. However, the bound

$$\rho_D(G(N,1-p)) \leq -\frac{p}{1-p}\rho_D(G(N,p))$$

is attained asymptotically for $N \to \infty$,

$$\lim_{N\to\infty} \rho_D(G(N,1-p)) = -\frac{p}{1-p} \lim_{N\to\infty} \rho_D(G(N,p)) \tag{3.16}$$

Moreover, from (3.16), we obtain the bounds

$$\max\left(-1, 1-\frac{1}{p}\right) \leq \lim_{N\to\infty} \rho_D(G(N,p)) \leq \min\left(1, \frac{1}{p}-1\right) \tag{3.17}$$

demonstrating that $\lim_{N\to\infty, p\to 1} \rho_D(G(N,p)) = 0$. In other words, the linear degree correlation coefficient of the complete graph is zero. Only for $p > \frac{1}{2}$, these bounds (3.17) are non-trivial. When p is small, a large assortativity range can be expected.

3.4.2. MAXIMUM AND MINIMUM ASSORTATIVITY

Given a class of graphs with a binomial degree distribution $\Pr[D_G = k] = \binom{N-1}{k} p^k (1-p)^{N-1-k}$, the maximal and minimal achievable assortativity is denoted by $\max\rho(N,p)$ and $\min\rho(N,p)$. The complementary class of graphs achieve the maximal and minimal assortativity $\max\rho(N,1-p)$ and $\min\rho(N,1-p)$. Relation (3.15) shows that $\max\rho(N,1-p) = -\frac{p}{1-p}\min\rho(N,p)$ and $\min\rho(N,1-p) = -\frac{p}{1-p}\max\rho(N,p)$. Thus,

$$\max\rho(N,1-p) - \min\rho(N,1-p) = \frac{p}{1-p}\left(\max\rho(N,p) - \min\rho(N,p)\right) \tag{3.18}$$

which is a special case of (3.11) for graphs with a binomial degree distribution. When p is small, the assortativity range is far larger than that in the complementary class of graphs.

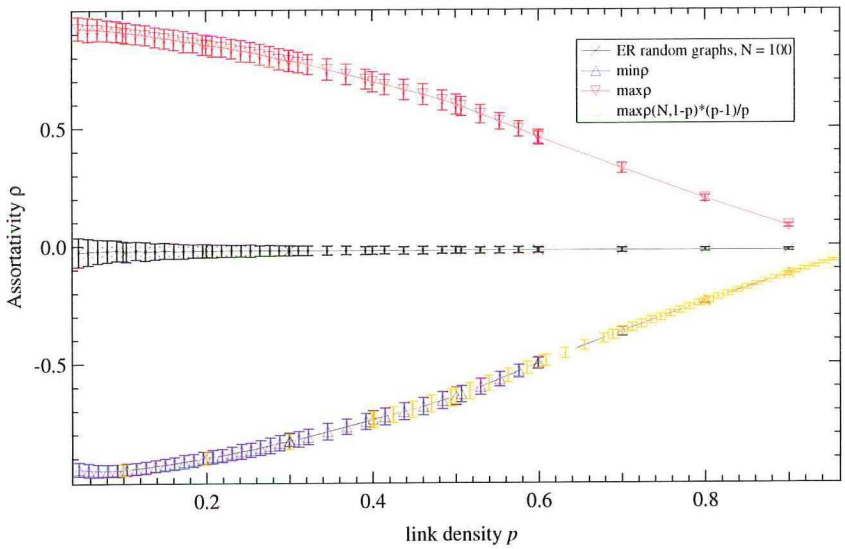


Figure 3.2: The average maximum $\max\rho(N=100,p)$ and minimum $\min\rho(N=100,p)$ assortativity of graphs with a binomial degree distribution versus the link density p. Verification of (3.16): $\frac{p-1}{p}\max\rho(N,p) = \min\rho(N,p)$.

The complementary classes of graphs $G(N,p)$ and $G(N,1-p)$ both follow a binomial degree distribution. They differ only in link density p. A small link density p contributes to a wide range of assortativity as illustrated in Fig. 3.2.

Most real-world networks are mostly sparse. Thus, their assortativity ranges expected to be larger than that of their corresponding complementary graphs according to (3.12) and (3.18). Furthermore, we will prove the following theorem:

Theorem 2 *For binomially distributed nodal degrees, the maximum $\rho_{\max}(N,p)$ and minimum assortativity $\rho_{\min}(N,p)$ tend to be symmetric around the $\rho_D = 0$ axis for large N. Specifically, it holds that*

$$\lim_{N\to\infty} \max\rho(N,p) + \min\rho(N,p) = 0$$

when the link density $p \in (0,1)$.

Proof. See Appendix 3.9. ■

Numerical computations in Fig. 3.2, indeed, illustrate that, approximately for finite N,

$$\max\rho(N,p) \simeq -\min\rho(N,p)$$

for any link density p. The values of $\max\rho(N,p)$ and $\min\rho(N,p)$ in Fig. 3.2 are computed with the exact algorithm explained in [4].

3.5. REAL-WORLD COMPLEX NETWORKS

This section illustrates how the assortativity of a graph and of its complement changes under degree-preserving rewiring, during which the degree of each node in the graph does not change. Fig. 3.3 shows that, for an ER random graph with $N = 500$ nodes, $L = 1984$ links and link density $p = 0.016$, the assortativity of the complement decreases much slower than that the assortativity of the original graph increases under degree-preserving rewiring. Relation (3.4), indeed, confirms that the assortativity of the complement must decrease, when $\rho_D(G)$ increases. The slower observed speed is due to the factor $\frac{p}{1-p}$ in (3.15) which is small for a small p. In general, assortativity of the complement changes much slower than that the assortativity of the original graph changes under degree-preserving rewiring, if the factor $\frac{2L\sigma^2[D_{l+}(G)]}{(N(N-1)-2L)\sigma^2[D_{l+}(G^c)]}$ in relation (3.4), which is a constant under degree-preserving rewiring, is small.

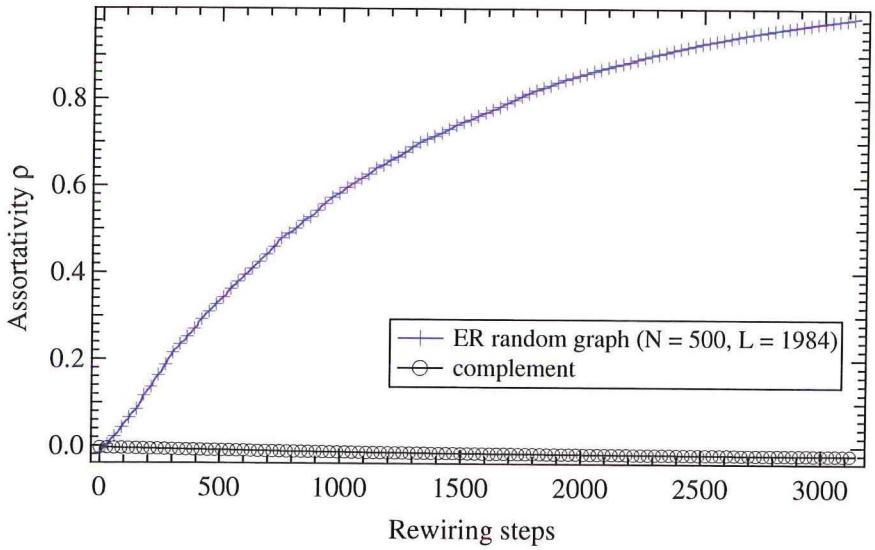


Figure 3.3: The assortativity of the Erdős-Rényi random graph with $N = 500$ nodes and $L = 1984$ links and its complement versus the number of rewiring steps in an assortative degree-preserving rewiring procedure.

The relation (3.18) and Fig. 3.2 demonstrate that a small link density (as in Fig. 3.3) corresponds to a large assortativity range $\max\rho - \min\rho$ and that the corresponding link density $1-p$ in the complement leads to a small $\rho_{\max} - \rho_{\min}$. This also explains in Fig. 3.3 why the assortativity of the graph increases much faster than the corresponding decrease in the complement during the degree-preserved rewiring process. Fig. 3.4 shows the same tendency in a Barabási-Albert graph [14] of the same size (N and L).

Fig. 3.5 illustrates for over thirty real-world complex networks how the assortativity ρ_D lies within the maximum possible range $\rho_{\max} - \rho_{\min}$. As shown in the corresponding table 3.1, the link density $p = L/\binom{N}{2} = \frac{E[D]}{N-1}$ in these complex networks is small, ranging from $4 \cdot 10^{-4} \le p \le 0.37$, such that the bound (3.13) for the assortativity range $\max\rho - \min\rho$ is here not confined by p. We observe that there are 6 strict disassortative

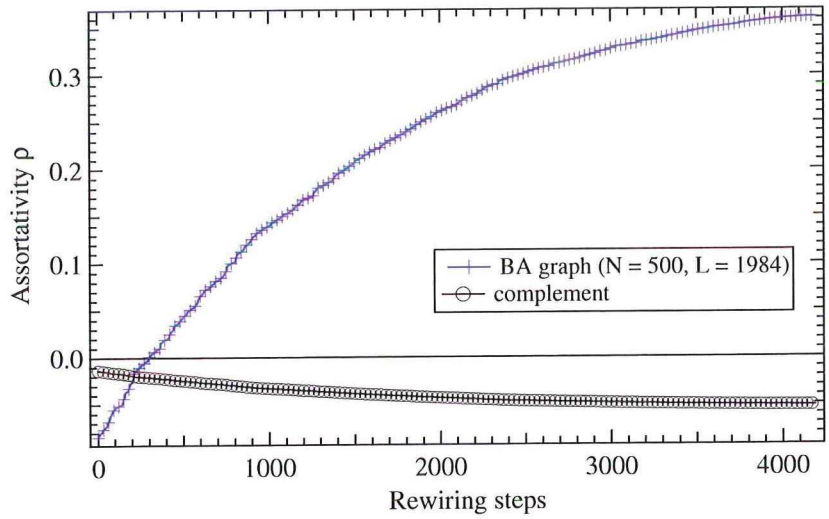


Figure 3.4: The assortativity of the Barabási-Albert random graph with $N = 500$ nodes and $L = 1984$ links and its complement versus the number of rewiring steps in an assortative degree-preserving rewiring procedure.

networks, where $\rho_{\max} < 0$. The assortativity range in those networks is small compared to the majority of complex networks. Moreover, they seem to possess a few very large degree nodes and many small degree nodes. So far, we have not found a strict assortative network, where $\min \rho > 0$. This observation supports the explanation in [4] why most real-world networks favor disassortativity due to a stronger connectivity and higher diversity than in assortative graphs. It would be interesting to know whether strict assortative, connected complex networks actually do exist. Assortativity range of the complements of these real-world networks, as shown in Fig. 3.6, are mostly small and around zero. This is due to the effect of a large link density p on the assortativity range relation between complementary graphs (3.12). However, the degree distribution plays an important role in determining the assortativity range, which explains possible large assortativity range even in dense networks (e.g. network 11-13).

3.6. CONCLUSION

The general relations (3.3), (3.4) and Theorem 1 between the assortativity $\rho_D(G)$ and $\rho_D(G^c)$ of two complementary graphs are considered important new findings. Based on these relations, we further derive bounds for the assortativity (3.8) and the relation (3.11) between assortativity range of two complementary graphs with a given degree distribution. The influence of link density and degree distribution on the assortativity and on the assortativity range of two complementary graphs is explicitly revealed.

Properties of complementary graphs are widely studied in Erdős-Rényi (ER) random graphs, because the complementary graph of an ER random graphs $G_p(N)$ is again an Erdős-Rényi random graph $G_{1-p}(N)$. Actually, the assortativity of an ER random graph is proved in [4] to be zero due to the random construction. However, constrained only by

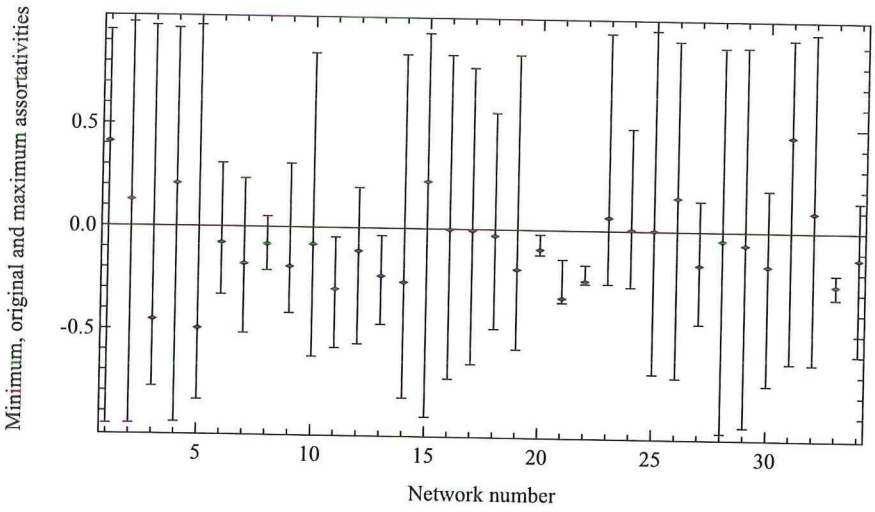


Figure 3.5: The minimum ($\min\rho$), original (ρ_D) and maximum ($\max\rho$) assortativity for various complex networks, described in Section 3.11. The values are computed by a heuristic, greedy degree-preserving rewiring algorithm.

a binomial degree distribution as in the ER random graphs $G_p(N)$, the assortativity of a graph $G(N,p)$ may vary within a wide range. The complementary graph $G(N,1-p)$ also possesses a binomial degree distribution, but characterized by N and link density $1-p$. The relation between $\rho_D(G(N,p))$ and $\rho_D(G(N,1-p))$ in this case can be simplified into (3.16). As a consequence, the maximum and minimum assortativity of a class of graphs with a binomial distribution are proved to be symmetric, $\max\rho(N,p) = -\min\rho(N,p)$ and the range $\max\rho(N,p) - \min\rho(N,p)$ is shown in (3.18) to be smaller for a large p.

A degree distribution is normally considered as a first order metric to characterize a network, while the assortativity as a second order descriptor. A narrow assortativity range $\max\rho - \min\rho$ of graphs with a given degree distribution implies that the degree distribution alone specifies the other properties well and is thus representative. Our results, (3.12) and (3.18), illustrate that a high link density confines the possible assortativity range more than a low link density. This, again, strengthens the importance of assortativity as a network characterizer, since most real-world networks are sparse. Finally, in over 30 real-world complex networks, the assortativity range $\max\rho - \min\rho$ is generally found to be large, except for a few strict disassortative graphs ($\max\rho < 0$). As we did not encounter strict assortative graphs ($\min\rho > 0$), it may be worthwhile to ponder whether they exist. Assortativity range relation 3.12 allows us to derive a non-trivial bound in one of the two complementary graphs, mostly the dense one. Exploring a better assortativity bound for sparse networks is deemed as an interesting future work. The ratio $\frac{\sigma^2[D_{l^+}(G^c)]}{\sigma^2[D_{l^+}(G)]}$ in the assortativity range relation has been explicitly expressed as functions of degree moments. Further quantitative studies on this ratio in network models as well as in real-world networks will provides more insights.

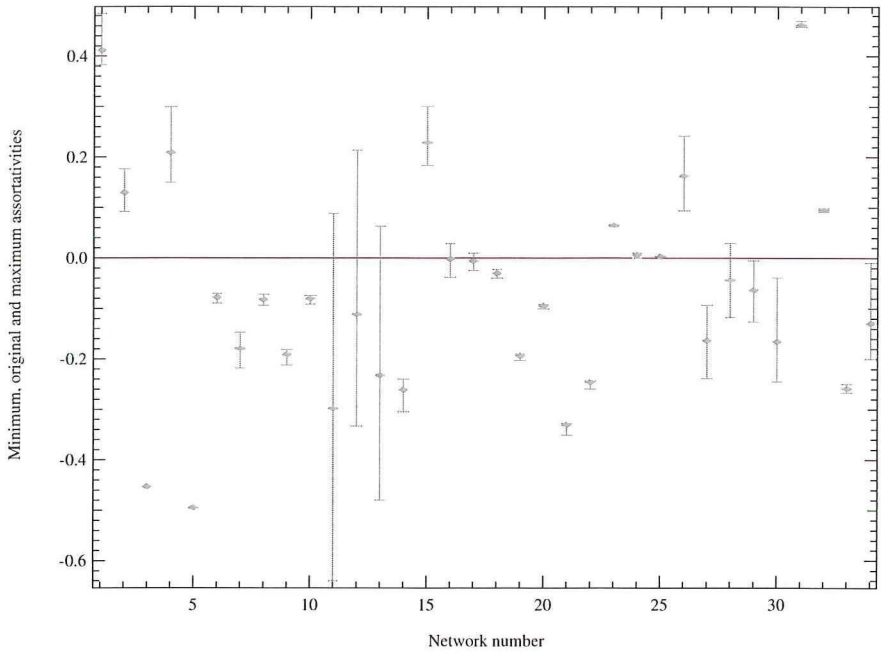


Figure 3.6: The minimum ($\min\rho$), original (ρ_D) and maximum ($\max\rho$) assortativity for the complements of various complex networks, described in Section 3.11. The values are derived from those of Fig. 3.5 by (3.3). They can be equivalently computed by the heuristic, greedy degree-preserving rewiring algorithm.

3.7. PROOF OF THEOREM 1

Consider an arbitrary link l in G with right endnode l^+. The probability that this link l is connected to a node $j = l^+$ with degree k equals

$$\Pr[D_{l^+}(G) = k] = \sum_{j=1}^{N} \Pr\left[\text{node } j \text{ is } l^+ | D_j = k\right] \Pr\left[D_j = k\right]$$

Each link l consists of two half links connected to node l^- and node l^+. With the basic law of the degree is $\sum_{j=1}^{N} D_j = 2L$, we have

$$\Pr\left[\text{node } j \text{ is } l^+ | D_j = k\right] = \frac{k}{2L}$$

Since each nodal degree D_j is distributed as the degree D of an arbitrary node in G, $\Pr\left[D_j = k\right] = \Pr[D = k]$ and we end up with

$$\Pr[D_{l^+}(G) = k] = \frac{Nk\Pr[D = k]}{2L} = \frac{k\Pr[D = k]}{E[D]}$$

$$\Pr[D_{l^+}(G^c) = k] = \frac{k\Pr[D = N-1-k]}{N-1-E[D]}$$

These expressions allow us to derive $\sigma^2[D_{l^+}(G)]$ and $\sigma^2[D_{l^+}(G^c)]$ in (3.4) as a function of the degree distribution $\Pr[D=k]$:

$$E[D_{l^+}(G)] = \sum_{k=0}^{N-1} \frac{Nk^2 \Pr[D=k]}{2L} = \frac{E[D^2]}{E[D]}$$
$$E[D^2_{l^+}(G)] = \sum_{k=0}^{N-1} \frac{Nk^3 \Pr[D=k]}{2L} = \frac{E[D^3]}{E[D]}$$
$$\sigma^2[D_{l^+}(G)] = \frac{E[D^3]E[D] - E^2[D^2]}{E^2[D]}$$

Similarly,

$$E[D_{l^+}(G^c)] = \sum_{k=0}^{N-1} \frac{k^2 \Pr[D=N-1-k]}{N-1-E[D]} = \frac{(N-1)^2 + E[D^2] - 2(N-1)E[D]}{N-1-E[D]}$$
$$E[D^2_{l^+}(G^c)] = \frac{(N-1)^3 + 3(N-1)E[D^2] - 3(N-1)^2E[D] - E[D^3]}{N-1-E[D]}$$
$$\sigma^2[D_{l^+}(G^c)] = \frac{1}{(N-1-E[D])^2}\Big((N-1)^2E[D^2] - (N-1)E[D^3] + (N-1)E[D^2]E[D] - (N-1)^2E^2[D] + E[D^3]E[D] - E^2[D^2]\Big)$$

They, together, lead to Theorem 1.

3.8. THE RATIO $\frac{\sigma^2[D_{l^+}(G^c)]}{\sigma^2[D_{l^+}(G)]}$

The ratio $\frac{\sigma^2[D_{l^+}(G^c)]}{\sigma^2[D_{l^+}(G)]}$ can be written as a function of the moments of the degree is the original graph G

$$\frac{\sigma^2[D_{l^+}(G^c)]}{\sigma^2[D_{l^+}(G)]} = \frac{E^2[D]}{(N-1-E[D])^2} + (N-1)\frac{(N-1)Var[D] - \{E[D^3] - E[D^2]E[D]\}}{(N-1-E[D])^2} \cdot \frac{E^2[D]}{E[D^3]E[D] - E^2[D^2]}$$

We express the variances $\sigma^2[D_{l^+}(G^c)]$ and $\sigma^2[D_{l^+}(G)]$ in terms of the centered moments $\mu_k = E\left[(D-E[D])^k\right]$ for $k \geq 2$. In particular, denoting the average degree by $\mu = E[D]$, we have that

$$E[D^2] = \mu_2 + \mu^2 = \mu^2 + \mathrm{Var}[D]$$
$$E[D^3] = E[(D-\mu+\mu)^3] = E[(D-\mu)^3 + 3(D-\mu)^2\mu + 3(D-\mu)\mu^2 + \mu^3]$$
$$= \mu_3 + \mu^3 + 3\mu_2\mu = \mu^3 + 3\mu\mathrm{Var}[D] + \mu_3$$

where the skewness $\frac{\mu_3}{\mu_2^{3/2}}$ measures the lack of symmetry of the degree distribution around the mean. Then,

$$\begin{aligned}\frac{\sigma^2[D_{l^+}(G^c)]}{\sigma^2[D_{l^+}(G)]} &= \frac{\mu^2}{(N-1-\mu)^2} + (N-1)\frac{(N-1)\mu_2 - \{\mu_3+\mu^3+2\mu_2\mu\}}{(N-1-\mu)^2}\cdot\frac{\mu^2}{\mu_3\mu+\mu_2\mu^2-\mu_2^2} \\ &= \frac{\mu^2}{(N-1-\mu)^2} + (N-1)\frac{(N-1-2\mu)\mu_2-\mu_3-\mu^3}{(N-1-\mu)^2}\cdot\frac{1}{\frac{\mu_3}{\mu}+\mu_2\left(1-\frac{\mu_2}{\mu^2}\right)} \\ &= \frac{\mu^2}{(N-1-\mu)^2} + \frac{\mu_2-\frac{\mu_3+\mu^3}{N-1-2\mu}}{\frac{(N-1-\mu)^2}{(N-1)(N-1-2\mu)}\left(\left(1-\frac{\mu_2}{\mu^2}\right)\mu_2+\frac{\mu_3}{\mu}\right)}\end{aligned}$$

We consider large and sparse graphs such that

$$\frac{\sigma^2[D_{l^+}(G^c)]}{\sigma^2[D_{l^+}(G)]} = \frac{1-\frac{\mu_3+\mu^3}{N\mu_2}}{\left(1-\frac{\mu_2}{\mu^2}\right)+\frac{\mu_3}{\mu\mu_2}} \tag{3.19}$$

When the degree distribution is symmetrical around the mean such that $\mu_3 = 0$,

$$\frac{\sigma^2[D_{l^+}(G^c)]}{\sigma^2[D_{l^+}(G)]} = \frac{\mu^2}{\mu^2-\mu_2}$$

Moreover, if the symmetrical degree distribution follows a binomial distribution where $\mu = Np$ and $\mu_2 = \mu(1-p)$,

$$\frac{\sigma^2[D_{l^+}(G^c)]}{\sigma^2[D_{l^+}(G)]} = 1$$

which is the same as (3.14), rigorously derived in Section 3.4.1.

For a power-law distribution $\Pr[D=k] = ck^{-\alpha}$ and $c = \frac{1}{\sum_{k=1}^{N-1}k^{-\alpha}} \simeq \frac{1}{\zeta(\alpha)}$, we have that $E[D] = \mu = c\sum_{k=1}^{N-1}k^{-(\alpha-1)} \simeq \frac{\zeta(\alpha-1)}{\zeta(\alpha)}$ and $E[D^m] = c\sum_{k=1}^{N-1}k^{-(\alpha-m)} \simeq \frac{\zeta(\alpha-m)}{\zeta(\alpha)}$, where the approximation sign is only valid provided $\alpha - m > 1$. Then,

$$\begin{aligned}\mu_3 &= E\left[(D-\mu)^3\right] = E[D^3] - 3\mu E[D^2] - \mu^3 \\ &= c\sum_{k=1}^{N-1}k^{-(\alpha-3)} - 3c^2\sum_{k=1}^{N-1}k^{-(\alpha-1)}\sum_{k=1}^{N-1}k^{-(\alpha-2)} - c^3\left(\sum_{k=1}^{N-1}k^{-(\alpha-1)}\right)^3\end{aligned}$$

For large, but finite N, we approximate as

$$\sum_{k=1}^{N-1}k^{-\alpha} \simeq \int_1^N \frac{dx}{x^\alpha} = \frac{N^{1-\alpha}-1}{1-\alpha} = \frac{1-N^{1-\alpha}}{\alpha-1}$$

and

$$\frac{\mu_3}{\mu_2} = \frac{\sum_{k=1}^{N-1} k^{-(\alpha-3)} - 3c\sum_{k=1}^{N-1} k^{-(\alpha-1)}\sum_{k=1}^{N-1} k^{-(\alpha-2)} - c^2\left(\sum_{k=1}^{N-1} k^{-(\alpha-1)}\right)^3}{\sum_{k=1}^{N-1} k^{-(\alpha-2)} - c\left(\sum_{k=1}^{N-1} k^{-(\alpha-1)}\right)^2}$$
$$\simeq \frac{\frac{1-N^{4-\alpha}}{\alpha-4} - 3\frac{\alpha-1}{1-N^{1-\alpha}}\frac{1-N^{2-\alpha}}{\alpha-2}\frac{1-N^{3-\alpha}}{\alpha-3} - \left(\frac{\alpha-1}{1-N^{1-\alpha}}\right)^2\left(\frac{1-N^{2-\alpha}}{\alpha-2}\right)^3}{\frac{1-N^{3-\alpha}}{\alpha-3} - \frac{\alpha-1}{1-N^{1-\alpha}}\left(\frac{1-N^{2-\alpha}}{\alpha-2}\right)^2}$$

For $1 < \alpha < 2$ and large, but finite N, we have

$$\frac{\mu_3}{\mu_2} \simeq \frac{\frac{N^{4-\alpha}}{4-\alpha} - 3\frac{(\alpha-1)}{(2-\alpha)(3-\alpha)}N^{5-2\alpha} - \frac{(\alpha-1)^2}{(2-\alpha)^3}N^{6-3\alpha}}{\frac{N^{3-\alpha}}{3-\alpha} - \frac{(\alpha-1)}{(2-\alpha)^2}N^{4-2\alpha}}$$
$$= N\frac{\frac{1}{4-\alpha} - 3\frac{(\alpha-1)}{(2-\alpha)(3-\alpha)}N^{1-\alpha} - \frac{(\alpha-1)^2}{(2-\alpha)^3}N^{2(1-\alpha)}}{\frac{1}{3-\alpha} - \frac{(\alpha-1)}{(2-\alpha)^2}N^{1-\alpha}} \simeq \frac{3-\alpha}{4-\alpha}N > 0$$

Similarly,

$$\frac{\mu_3 + \mu^3}{N\mu_2} \simeq \frac{3-\alpha}{4-\alpha}$$
$$\frac{\mu_3}{\mu\mu_2} \simeq \frac{(3-\alpha)(2-\alpha)}{(4-\alpha)(\alpha-1)}N^{\alpha-1}$$

and

$$\frac{\mu_2}{\mu^2} \simeq \frac{\frac{1-\alpha}{\alpha-3}N^{\alpha-1} - \left(\frac{\alpha-1}{\alpha-2}\right)^2}{\left(\frac{\alpha-1}{\alpha-2}\right)^2}$$

Together with (3.19), we have

$$\frac{\sigma^2[D_{l^+}(G^c)]}{\sigma^2[D_{l^+}(G)]} \simeq \frac{1 - \frac{3-\alpha}{4-\alpha}}{\left(1 - \frac{\frac{1-\alpha}{\alpha-3}N^{\alpha-1} - \left(\frac{\alpha-1}{\alpha-2}\right)^2}{\left(\frac{\alpha-1}{\alpha-2}\right)^2}\right) + \frac{(3-\alpha)(2-\alpha)}{(4-\alpha)(\alpha-1)}N^{\alpha-1}} = O\left(N^{1-\alpha}\right) \to 0$$

When $2 < \alpha < 3$, we prove in a similar way that

$$\frac{\sigma^2[D_{l^+}(G^c)]}{\sigma^2[D_{l^+}(G)]} = O(N^{-1}) \to 0$$

3.9. PROOF OF THEOREM 2

First, we note from (3.15) that

$$\max\rho(G^c(N,p)) = -\frac{p}{1-p}\min\rho(G(N,p)) - \frac{2}{(N-2)(1-p)}$$

Let $R_N(p) = \max\rho(N,p) + \min\rho(N,p)$. From (3.15), it follows that, $R_N(p) = -\frac{p}{1-p}R_N(1-p) - \frac{4}{(N-2)(1-p)}$. By setting $p = \frac{1}{2}$, one obtains $R_N(\frac{1}{2}) = -R_N(\frac{1}{2}) - \frac{8}{N-2}$, showing that $R_N(\frac{1}{2}) = -\frac{4}{N-2}$.

The link density $p = \frac{L}{\binom{N}{2}} \in \mathbb{Q}$ is a rational number, which tends to a real number when $N \to \infty$. Assume that $\max\rho(N,p)$ is differentiable with respect to p, then so are $\min\rho(N,p)$ and $R_N(p)$. Thus,

$$\frac{d^n R_N(p)}{dp^n} = \frac{d^n}{dp^n}\left(-\frac{p}{1-p}R_N(1-p) - \frac{4}{(N-2)(1-p)}\right)$$

By applying Leibniz' rule, we have for $n \geq 1$

$$\frac{d^n}{dp^n}\left(-\frac{p}{1-p}R_N(1-p)\right) = -\sum_{j=0}^{n}\binom{n}{j}\frac{d^{n-j}}{dp^{n-j}}\left(\frac{p}{1-p}\right)\frac{d^j}{dp^j}R_N(1-p)$$

For $m > 0$, we use $\frac{d^m}{dp^m}\left(\frac{p}{1-p}\right) = \frac{d^m}{dp^m}\left(1 - \frac{1}{1-p}\right) = -\frac{d^m}{dp^m}\left(\frac{1}{1-p}\right) = (-1)^m m!(1-p)^{-m-1}$ such that

$$\frac{d^n}{dp^n}\left(-\frac{p}{1-p}R_N(1-p)\right) = -\left(\frac{p}{1-p}\right)\frac{d^n R_N(1-p)}{dp^n} + \frac{(-1)^n n!}{(1-p)^{n+1}}\sum_{j=0}^{n-1}\frac{1}{j!}\frac{d^j R_N(1-p)}{dp^j}(1-p)^j$$

Hence,

$$\frac{d^n R_N(p)}{dp^n} = -\left(\frac{p}{1-p}\right)\frac{d^n R_N(1-p)}{dp^n} + \frac{(-1)^n n!}{(1-p)^{n+1}}\sum_{j=0}^{n-1}\frac{1}{j!}\frac{d^j R_N(1-p)}{dp^j}(1-p)^j + \frac{(-1)^n 4(n!)}{(N-2)(1-p)^{n+1}} \tag{3.20}$$

Setting $n = 1$ renders

$$\frac{dR_N(p)}{dp} = -\frac{p}{1-p}\frac{dR_N(1-p)}{dp} - \frac{1}{(1-p)^2}R_N(1-p) - \frac{4}{(N-2)(1-p)^2}$$

Evaluation at $p = \frac{1}{2}$ (with $R_N(\frac{1}{2}) = -\frac{4}{N-2}$) yields $\left.\frac{dR_N(p)}{dp}\right|_{p=\frac{1}{2}} = -\left.\frac{dR_N(1-p)}{dp}\right|_{p=\frac{1}{2}}$, which shows that $\left.\frac{dR_N(p)}{dp}\right|_{p=\frac{1}{2}} = 0$. Since $\left.\frac{dR_N(p)}{dp}\right|_{p=\frac{1}{2}} = 0$, it also follows from (3.20) that $\left.\frac{d^2R_N(p)}{dp^2}\right|_{p=\frac{1}{2}} = 0$ and in fact, by iteration, that $\left.\frac{d^n R_N(p)}{dp^n}\right|_{p=\frac{1}{2}} = 0$. The Taylor expansion of $R_N(p)$ around $p = \frac{1}{2}$,

$$R_N(p) = \sum_{n=0}^{\infty}\frac{1}{n!}\left.\frac{d^n R_N(p)}{dp^n}\right|_{p=\frac{1}{2}}\left(p-\frac{1}{2}\right)^n = R_N\left(\frac{1}{2}\right) = -\frac{4}{N-2}$$

demonstrates that $R_N(p) = \max\rho(N,p) + \min\rho(N,p) = -\frac{4}{N-2}$. Hence, the maximum and minimum assortativity are symmetric around $\rho_D = 0$ when $N \to \infty$, in which case the assumption of differentiability with respect to p also holds. This proves Theorem 2.

3.10. EXAMPLE OF A STRICT DISASSORTATIVE GRAPH CLASS

Consider the connected graphs in which $N-2$ nodes have degree r and the two remaining nodes, 1 and 2, have degree d_1 and d_2. Thus, the basic law of the degree tells us that

$$2L = (N-2)r + d_1 + d_2$$

There are only two configurations possible that lead to a different sum $S = \sum_{i\sim j}(d_i - d_j)^2$ in (3.1): (a) when node 1 and node 2 are not mutually connected and (b) when they are. In the first case,

$$S_1 = d_1(r-d_1)^2 + d_2(r-d_2)^2$$

and in the second case,

$$\begin{aligned} S_2 &= (d_1-d_2)^2 + (d_1-1)(r-d_1)^2 + (d_2-1)(r-d_2)^2 \\ &= S_1 + (d_1-d_2)^2 - (r-d_1)^2 - (r-d_2)^2 \end{aligned}$$

Now,

$$(d_1-d_2)^2 - (r-d_1)^2 - (r-d_2)^2 = -2(r-d_1)(r-d_2)$$

such that

$$S_2 = S_1 - 2(r-d_1)(r-d_2) \tag{3.21}$$

The basic law of the degree $2L = Nr + (d_1 - r) + (d_2 - r)$ allows us to eliminate d_2,

$$S_2 = S_1 + 2(r-d_1)(2L-Nr) + 2(r-d_1)^2 \tag{3.22}$$

If $r > d_1$, then it follows from (3.22) that $S_2 > S_1$ and, further from (3.21), that then $r < d_2$. If $r = d_1$, then $S_2 = S_1$. If $r < d_1$ and $r < d_2$ or $r > d_1$ and $r > d_2$, then (3.21) shows that $S_2 < S_1$.

After choosing the S_1 configuration, we rewrite (3.1) as

$$\rho_D = \frac{V - S_1}{V}$$

The denominator V in (3.1) is, with

$$\sum_{i=1}^{N} d_i^2 = (N-2)r^2 + d_1^2 + d_2^2$$

$$\sum_{i=1}^{N} d_i^3 = (N-2)r^3 + d_1^3 + d_2^3$$

equal to

$$V = \sum_{i=1}^{N} d_i^3 - \frac{1}{2L}\left(\sum_{i=1}^{N} d_i^2\right)^2 = (N-2)r^3 + d_1^3 + d_2^3 - \frac{\left((N-2)r^2 + d_1^2 + d_2^2\right)^2}{2L}$$

Hence,

$$S_1 - V = d_1(r-d_1)^2 + d_2(r-d_2)^2 - (N-2)r^3 - d_1^3 - d_2^3 + \frac{\left((N-2)r^2 + d_1^2 + d_2^2\right)^2}{2L}$$

from which

$$2L(S_1 - V) = \left((N-2)r^2 + d_1^2 + d_2^2 - rL\right)^2 + rL\{2d_1(r-d_1) + 2d_2(r-d_2) - rL\}$$

Using $2L = (N-2)r + d_1 + d_2$ yields, after some tedious manipulations,

$$2L(S_1 - V) = \left\{d_1^2 + d_2^2 - r(d_1 + d_2)\right\}^2$$

In conclusion, $S_1 - V \geq 0$ and only zero if $d_1 = d_2 = r$. Hence, since $V \geq 0$ (as shown in [4, 12] and since $\rho_D = (V - S_1)/V$, we conclude that $\rho_{D1} < 0$. If $S_2 > S_1$, then $S_2 - V > 0$ such that we find a strict disassortative class. The analysis above shows that this happens if $d_1 < r < d_2$.

3.11. TABLE OF ASSORTATIVITIES FOR COMPLEX NETWORKS

# Name	N	L	$E[D]$	ρ_0	$\rho_{\min}$	$\rho_{\max}$	$\Delta\rho$
Proteins							
1 1AOR	97	212	4.37	0.412	-0.959	0.955	1.91
2 1a4j	95	213	4.48	0.129	-0.959	0.992	1.95
3 1atn	5015	5128	2.05	-0.453	-0.778	0.977	1.75
4 1eaw	53	123	4.64	0.209	-0.952	0.965	1.92
5 3cro	1856	1966	2.12	-0.495	-0.842	0.979	1.82
Software call graphs							
6 AbiWord	1093	1765	3.23	-0.0777	-0.33	0.309	0.639
7 Digital Material	187	269	2.88	-0.179	-0.516	0.235	0.751
8 MySql	1500	4202	5.60	-0.0825	-0.21	0.0521	0.262
9 VTK	786	1370	3.49	-0.191	-0.418	0.309	0.727
10 XMMS	1097	1894	3.45	-0.0809	-0.627	0.848	1.48
Food webs							
11 Everglades	69	880	25.5	-0.298	-0.584	-0.0462	0.538
12 Florida	128	2075	32.4	-0.112	-0.565	0.196	0.761
13 St Marks	54	350	13.0	-0.232	-0.467	-0.0361	0.431
Telecommunications networks							
14 ARPANET80	71	86	2.42	-0.261	-0.824	0.845	1.67
15 Surfnet	65	111	3.42	0.229	-0.916	0.950	1.87
Electronic circuits							
16 s208	122	189	3.10	-0.00201	-0.729	0.845	1.57
17 s420	252	399	3.17	-0.00591	-0.657	0.783	1.44
18 s838	512	819	3.20	-0.03	-0.483	0.567	1.05
Peer-to-peer networks							
19 Gnutella 1	737	803	2.18	-0.193	-0.582	0.848	1.43
20 Gnutella 2	1568	1906	2.43	-0.0946	-0.122	-0.0211	0.101
21 Gnutella 3	435	459	2.11	-0.33	-0.351	-0.141	0.210
22 Gnutella 4	653	738	2.26	-0.246	-0.259	-0.168	0.0913
Power grids							
23 Western European power grid level 2 AL	3690	4206	2.28	0.0649	-0.259	0.958	1.22
24 Western European power grid level 3 AL	756	786	2.08	0.00648	-0.273	0.497	0.770
25 Western US power grid	4941	6594	2.67	0.00346	-0.695	0.975	1.67
Miscellaneous networks							
26 American football contest network	115	613	10.7	0.162	-0.713	0.924	1.64
27 C. elegans neural network	297	2148	14.5	-0.163	-0.449	0.149	0.598
28 Dolphin social network	62	159	5.13	-0.0436	-0.979	0.895	1.87
29 Dutch football player co-appearance network	685	10310	30.1	-0.0634	-0.95	0.897	1.85
30 Les Miserable co-appearance network	77	254	6.60	-0.165	-0.746	0.202	0.949
31 Network science collaboration network	1461	2742	3.75	0.462	-0.638	0.935	1.57
32 Western European railway network level 2 AL	697	785	2.25	0.0954	-0.642	0.963	1.61
33 Word adjacency network – Japanese texts	2704	7998	5.92	-0.259	-0.321	-0.204	0.117
34 Word adjacency network – David Copperfield	112	425	7.59	-0.129	-0.598	0.147	0.745

Table 3.1: Various real-world networks whose maximum and minimum assortativities were computed heuristically by greedy degree-preserving rewiring. Although the heuristic algorithm cannot guarantee to find the optimal assortativity results, it achieves results that are close to that of the exact algorithm.

REFERENCES

[1] H. Wang, W. Winterbach, and P. Mieghem, *Assortativity of complementary graphs*, The European Physical Journal B **83**, 203 (2011).

[2] M. E. J. Newman, *Assortative mixing in networks*, Phys. Rev. Lett. **89**, 208701 (2002).

[3] M. E. J. Newman, *Mixing patterns in networks*, Phys. Rev. E **67**, 026126 (2003).

[4] P. Van Mieghem, H. Wang, X. Ge, S. Tang, and F. A. Kuipers, *Influence of assortativity and degree-preserving rewiring on the spectra of networks*, The European Physical Journal B **76**, 643 (2010).

[5] Z. Jing, T. Lin, Y. Hong, L. Jian-Hua, C. Zhi-Wei, and L. Yi-Xue, *The effects of degree correlations on network topologies and robustness*, Chinese Physics **16**, 3571 (2007).

[6] L. Li, D. Alderson, J. C. Doyle, and W. Willinger, *Towards a theory of scale-free graphs: Definition, properties, and implications*, Internet Mathematics **2**, 431 (2005), http://www.tandfonline.com/doi/pdf/10.1080/15427951.2005.10129111 .

[7] M. Boguñá, R. Pastor-Satorras, and A. Vespignani, *Absence of epidemic threshold in scale-free networks with degree correlations*, Phys. Rev. Lett. **90**, 028701 (2003).

[8] J. D. Noh, *Percolation transition in networks with degree-degree correlation*, Phys. Rev. E **76**, 026116 (2007).

[9] P. Van Mieghem, X. Ge, P. Schumm, S. Trajanovski, and H. Wang, *Spectral graph analysis of modularity and assortativity*, Phys. Rev. E **82**, 056113 (2010).

[10] S. Zhou and R. J. Mondragón, *Structural constraints in complex networks*, New Journal of Physics **9**, 173 (2007).

[11] H. Wang, L. Douw, J. M. Hernández, J. C. Reijneveld, C. J. Stam, and P. Van Mieghem, *Effect of tumor resection on the characteristics of functional brain networks*, Phys. Rev. E **82**, 021924 (2010).

[12] P. Van Mieghem, *Graph spectra for complex networks* (Cambridge University Press, 2010).

[13] M. Castro, M. Costa, and A. Rowstron, *Should we build gnutella on a structured overlay?* SIGCOMM Comput. Commun. Rev. **34**, 131 (2004).

[14] A.-L. Barabási and R. Albert, *Emergence of scaling in random networks*, Science **286**, 509 (1999), http://www.sciencemag.org/content/286/5439/509.full.pdf .

4

DO GREEDY ASSORTATIVITY OPTIMIZATION ALGORITHMS PRODUCE GOOD RESULTS?

Wynand WINTERBACH, Dick DE RIDDER, Huijuan WANG, Marcel REINDERS, Piet VAN MIEGHEM,

4.1. ABSTRACT

We consider algorithms for generating networks that are extremal with respect to degree assortativity. Networks with maximized and minimized assortativities have been studied by other authors. In these cases, networks are rewired whilst maintaining their degree vectors. Although rewiring can be used to create networks with high or low assortativities, it is not known how close the results are to the true maximum or minimum assortativities achievable by networks with the same degree vectors.

We introduce the first algorithm for computing a network with maximal or minimal assortativity on a given vector of N valid node degrees. We compare the assortativity metrics of networks obtained by this algorithm to assortativity metrics of networks obtained by a greedy assortativity-maximization algorithm. The algorithms are applied to Erdős-Rényi networks, Barabási-Albert and a sample of real-world networks. For the Erdős-Rényi and Barabási-Albert networks considered, we find that the mean difference of the assortativity metrics produced by the two methods decreases faster than $O(N^{-1.1})$. We also find that the number of rewirings considered by the greedy approach must scale with the number of links in order to ensure a good approximation.

This chapter was published in The European Physical Journal B **85**, 5 (2012) [1].

4.2. INTRODUCTION

Networks play an ever-larger role in the analysis of various systems. Examples are biological systems, social networks and computer networks. Comparison of such networks is difficult since they vary in size (both in node and link counts) and link configurations. Topological metrics provide one way of comparing different networks by encoding their properties as scalars or vectors: two networks with similar metrics could be considered equivalent, depending on the context.

Degree distributions of networks are an often-used metric for characterizing networks. Such first-order descriptions are not always enough to describe the topology of networks. Thus, it may be necessary to consider second-order measures in addition to degree distributions. One such measure is Newman's degree assortativity [2] (a special case of assortative mixing [3]), a relatively new metric that measures the extent to which nodes with similar degrees are connected by links. The limits of this metric are not yet as well studied as those of other metrics. Extremal graph theory provides a framework for studying these limits. A typical approach in extremal studies is the generation of networks that are extremal with respect to the metric being studied. As an example, in Wang *et al.* [4], the maximum and minimum assortativities achievable by networks with binomial degree distributions are shown to vary greatly with the densities of the networks. This is a non-obvious result, illustrating that assortativity measures have to be considered relative to a given network structure. We consider two methods for obtaining networks with maximal degree assortativity subject to fixed degree vectors: a greedy algorithm based on link rewiring and an exact algorithm based on weighted b-matching.

Watts and Strogatz [5] introduced link rewiring as a technique for generating random networks. During rewiring, a link is chosen at random and one of its end-points is replaced by a random node in the same network provided that no self-loops or duplicate links are introduced (that is, the network must remain simple). Due to the way that rewiring works, the node and link counts are invariant. Evans [6] and Lindquist *et al.* [7] exploited this property and studied rewiring as a mechanism for optimizing metrics subject to fixed node and link counts.

Degree-preserving rewiring is a restriction of link rewiring where a pair of links is chosen at random and a random end-point from the first link is exchanged for a random end-point from the second link. Maslov and Sneppen [8] introduced degree-preserving rewiring as a technique for generating null models. Their aim was to determine the likelihood of features observed in protein-protein interaction networks (relative to the null models). By requiring that degrees are preserved, the rewiring procedure is able to generate random networks that can be characterized by their degree sequences. The utility of this is evident from the fact that two of the most well-known classes of random networks are characterized by their degree distributions: Erdős-Rényi networks and Barabási-Albert networks.

Degree-preserving rewiring forms the basis of a simple technique for optimizing the degree-assortativity of a network (with a constant degree vector): a number of such rewiring steps are applied such that each rewiring increases/decreases the assortativity. This is essentially the approach taken by our greedy algorithm. Menche *et al.* [9] implemented a heuristic degree-preserving rewiring algorithm based on simulated annealing that used to produce networks with maximized and minimized assortativities, focusing

on the class of scale free networks. However, as they did not have an exact algorithm, they could not compare the results of their heuristic algorithm to exact results.

In this paper, we consider the open question of how good a simple greedy assortativity maximization approach is. To this end, we present a novel exact algorithm for computing the maximum degree-preserved assortativity of a network. Using ensembles of Erdős-Rényi and Barabási-Albert networks as well as a number of real-world networks, we compare results from the greedy algorithm to those of the exact algorithm. We show that while a greedy rewiring process does not, in general, attain optimum assortativity, it achieves very good approximations.

4.3. ASSORTATIVITY MAXIMIZATION ALGORITHMS

4.3.1. EXACT ALGORITHM

Van Mieghem *et al.* [10] have shown that the assortativity $\rho(G)$ of a network $G(\mathcal{N},\mathcal{L})$ with $N = |\mathcal{N}|$ nodes and $L = |\mathcal{L}|$ links can be expressed as

$$\rho(G) = 1 - \frac{\sum_{i\sim j}(d_i - d_j)^2}{\sum_{i=1}^{N} d_i^3 - \frac{1}{2L}\left(\sum_{i=1}^{N} d_i^2\right)^2} \tag{4.1}$$

$$= 1 - \frac{\sum_{i=1}^{N} d_i^3 - 2\sum_{i\sim j} d_i d_j}{\sum_{i=1}^{N} d_i^3 - \frac{1}{2L}\left(\sum_{i=1}^{N} d_i^2\right)^2}. \tag{4.2}$$

where d_i is the degree of the i-th node and $i \sim j$ means that node i and node j are joined by a link. Under degree-preserving rewiring, $\sum_{i\sim j} d_i d_j$ is the *only* variable part of the expression, attaining a maximum when the assortativity of G is maximized. Now consider the weighted complete network K_G whose nodes have the same labels $n_1, n_2, \ldots, n_N$ as G and in which the link $\{n_i, n_j\} \in \mathcal{L}(K_G)$ has weight $w(i,j) = d_i d_j$. Thus, G is an unweighted subnetwork of K_G. Let G_w be equal to G except that it has the same link weights as K_G (thus, G_w is simply a weighted subnetwork of K_G). The sum of the link weights in G_w is exactly $\sum_{i\sim j} d_i d_j = \sum_{i\sim j} w(i,j)$. Thus, $\sum_{i\sim j} d_i d_j$ can be maximized by finding the maximum weight subnetwork in K_G whose degree vector matches that of G.

The degree-constrained weighted degree subnetwork problem is equivalent to the weighted perfect b-matching problem [11] which can be efficiently computed: for example, the algorithm of Miller and Pekny [12] has a worst-case time complexity of $\max\{O(NL\log d_{\max}), O(N^2 L)\}$ where $d_{\max_i} = \max d_i$. Since the algorithm is always applied to the network K_G, $L = O(N^2)$ rendering the running time $O(N^4)$.

We were unable to find any usable implementations of Miller and Pekny's algorithm. The algorithm is difficult to implement correctly. Consequently, we took a simpler route, due to Shiloach [13], wherein b-matching problems are converted to 1-matching problems. We then applied Kolmogorov's [14] very fast $O(N^3)$ Blossom V matcher. In spite of the speed of Blossom V, the initial $O(N^2)$ transformation resulted in a running time of $O(N^6)$, limiting the sizes of the instances that we could investigate. See Supplemental Material at [URL will be inserted by publisher] for a description of Shiloach's transformation.

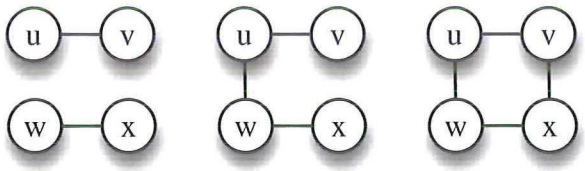


Figure 4.1: The only link configurations that permit link rewirings.

4.3.2. GREEDY ALGORITHM

Like the exact assortativity maximization algorithm, the greedy algorithm modifies the topology of a given network in order to maximize the term $\sum_{i\sim j} d_i d_j$ in (4.2). As opposed to the exact algorithm which computes an entirely new link configuration, the greedy algorithm increases the term $\sum_{i\sim j} d_i d_j$ by rewiring pairs of links in a sequence of steps.

In an optimistic rewiring strategy, a pair of links $\{u, v\}$, $\{w, x\} \in \mathscr{L}(G)$ is selected such that u, v, w and x are distinct. If $\{u, x\} \notin \mathscr{L}(G)$ and $\{w, v\} \notin \mathscr{L}(G)$, $\{u, v\}$ and $\{w, x\}$ can be rewired to (that is, replaced by) $\{u, x\}$, $\{w, v\}$. The four-node configurations in Figure 4.1 can all be rewired in this fashion. Let d_u, d_v, d_w and d_x be the degrees of u, v, w, x in G. If $-d_u d_v - d_w d_x + d_u d_x + d_w d_v > 0$, the rewiring increases the term $\sum_{i\sim j} d_i d_j$ and therefore the change is made. Otherwise, the rewiring is rejected. There are eleven non-isomorphic four-node configurations of which only three – those in Figure 4.1 – permit pair-wise link rewiring. Inspection reveals that the symmetry of the first and last of these configurations allow for two possible rewirings, whereas the middle configuration allows only for one rewiring.

The greedy algorithm searches the input network for the configurations in Figure 4.1 whose links can be rewired to increase the assortativity. In each iteration of the algorithm, a random assortativity-increasing configuration is selected to ensure that different invocations of the greedy algorithm can sample different parts of the rewiring space. A simple way to facilitate this selection is to maintain a set R of rewirable link pairs from which selections can be made (R is in fact a network with links from the input network as its nodes; the links in R correspond to rewirable link pairs in the input network). After a pair of links $\{u, v\}$, $\{w, x\}$ is rewired, all rewirable configurations containing at least two nodes in $\{u, v, w, x\}$ have to be re-evaluated for rewirability. Those that are no longer rewirable are removed from R whilst those that become rewirable are added to R. The nodes of a rewirable link pair in R induce one of the rewirable configurations in Figure 4.1. The reason for focusing on rewirable link pairs as opposed to rewirable configurations, is that the first and last of the rewirable configurations in Figure 4.1 may be rewired in two ways and it is easier to consider each of the two rewirings as a separate element in the set R.

Explicitly maintaining R is expensive, at least initially (before any rewirings) when it may be that $|R| = O(N^4)$. However, when $|R|$ is large, keeping track of R is unnecessary as there is a good chance of finding rewirable link pairs when randomly sampling links from the network. Since not every random sampling will yield a rewirable link pair, sampling is repeated up to a pre-specified number of times s; if a valid rewiring is found, it

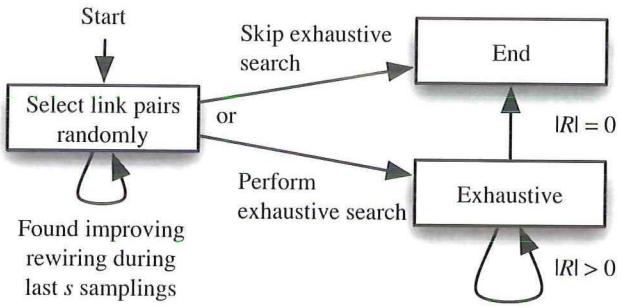


Figure 4.2: State diagram for the greedy assortativity maximization algorithm.

is applied and the algorithm starts with a new iteration. As the greedy algorithm progresses, the number of rewirable link pairs $|R|$ decreases, rendering it less and less likely for a randomly sampled pair of links to be rewirable. Eventually, s random samplings will fail to discover rewirable link pairs.

At this point, R can be constructed explicitly, since $|R|$ should be small enough. From this point onwards, all link pairs are sampled from R and the algorithm proceeds until $|R| = 0$. The algorithm naturally decomposes into two states. In the first state, links are sampled at random from the input network; in the second, the set R is constructed and links are subsequently sampled from R. We refer to the first state as the *random selection* state and the second as the *exhaustive* state (since it continues until no more assortativity increasing configurations exist). Note that although $|R|$ may be small, constructing R requires $O(L^2)$ time, as all link pairs have to be enumerated.

The execution time on a large network is considerable and therefore such an exhaustive state is impractical for real-world assortativity-maximization algorithms. Our motivation for including it was to study whether algorithms without exhaustive states might miss good, difficult to find solutions. The exhaustive step is optional in our greedy algorithm, allowing exhaustive and non-exhaustive results to be compared.

Combining all of this leads to the state diagram in Figure 4.2. When the exhaustive state is skipped, the greedy algorithm is a simple optimization algorithm whose results are unlikely to best those of more sophisticated algorithms, such as the algorithm of Menche *et al.* [9]. When the exhaustive state is engaged, our algorithm has the opportunity to find rewirings that will be missed by algorithms based on random link pair selection.

4.4. APPROACH SETUP

4.4.1. DATA SETS

We investigate ensembles of Erdős-Rényi and Barabási-Albert networks, as well as a number of real-world networks. Erdős-Rényi networks [15] are a 2-parameter family of random networks denoted $G_p(N)$. The parameter N is the number of nodes in the network whilst the parameter p is the probability that a pair of nodes are connected by a link. We considered networks of size $N \in \{25, 50, 80, 100, 150, 200\}$ and $p \in [0.05, 0.95]$. We

also considered networks of size $N \in \{250, 300, 350, 400, 450, 500\}$ for $p = 0.05$; we were forced to limit p due to the excessive computation time required for larger p.

Barabási-Albert networks [16] are a 2-parameter family of random scale-free networks. As before, the parameter N denotes the number of nodes in the network. The parameter m represents the degree of nodes added in the growth process (Barabási-Albert networks are grown one node at a time). For these networks, we considered instances with $N \in [25, 1000]$ (including most values of N for which the Erdős-Rényi experiments were computed) and $m \in \{2, 3, 4\}$.

Random network ensembles were constructed for each pair of parameters: $\{N, p\}$ for Erdős-Rényi networks and $\{N, m\}$ for Barabási-Albert networks. With the exception of a few cases, at least 10^4 ensemble instances were generated for each parameter pair. Only 10^3 Erdős-Rényi networks with $N = 200$ and $p > 0.1$ were generated due to the long running times required on these networks.

The real-world networks that we considered come from a number of different domains and include protein-protein interaction networks, software call graphs, food webs, telecommunications networks and electronic circuits.

4.4.2. ALGORITHM SETUP

The greedy algorithm was executed in both its exhaustive and non-exhaustive modes. In the non-exhaustive mode, we considered various upper bounds to the number of random samplings: $s \in \{100, 1000, 10000, 100000\}$. In the exhaustive mode, $s = 100000$ random samplings were allowed before the algorithm switched to the exhaustive state.

4.4.3. MEASURED DATA

We considered the means and standard deviations of the differences between the assortativities as computed by the exact and greedy algorithms for each network instance (in a given ensemble of networks). A simple approach is to consider $\mathrm{E}[\rho - \rho']$ and $\mathrm{Var}[\rho - \rho']$. Here, ρ is a random variable representing the maximum assortativity of an ensemble of networks as computed by the exact algorithm. Similarly, ρ' is a random variable representing the maximum assortativity of the ensemble as computed by the greedy algorithm. Wang *et al.* [4] show that the range of degree assortativities achievable by networks with binomial degree distributions (which include Erdős-Rényi networks) vary greatly with their density and can often be much smaller than the possible assortativity range of $[-1, 1]$. In particular, as the density increases, the range shrinks. This variation in ranges skews the results, as the absolute differences may appear to be small whilst they are in fact large relative to the attainable assortativity range. To account for this, we normalize the mean and variance by dividing by $\mathrm{E}[\rho - \rho_0]$ and $\mathrm{Var}[\rho - \rho_0]$ respectively. Here, ρ_0 is a random variable representing the (original) assortativities of networks in the ensemble under investigation.

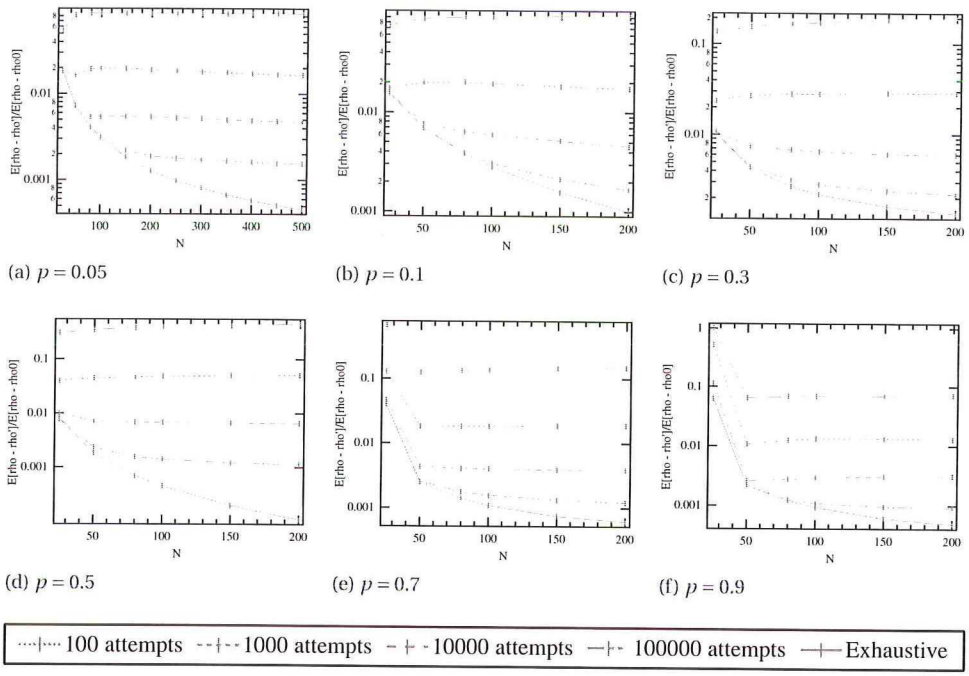


Figure 4.3: Means of relative differences in solutions obtained by the exact and greedy algorithms for various values of N and p. These plots apply to Erdős-Rényi networks. Each plot corresponds to a fixed p.

4.5. RESULTS

4.5.1. ERDŐS-RÉNYI NETWORKS

RESULTS AS FUNCTIONS OF N

First, we consider how the performance of the greedy algorithm changes as node counts increase. The normalized mean differences between the exact and greedy algorithms are shown in Figure 4.3 as functions of N for a few representative values of p. Likewise, the normalized variance differences for the same values of p are shown in Figure 4.4.

These plots paint a favorable picture for the greedy approach, as it performs well even when the number of random samplings s is small. The downward slopes corresponding to some of the non-exhaustive results seem to suggest that they improve as N increases. However, the Barabási-Albert (Section 4.5.2) and real-world (Section 4.5.3) results show opposing trends. It may be that the Erdős-Rényi networks we tested are either too small or that the structure of Erdős-Rényi networks particularly favors the random link selection scheme employed by the non-exhaustive phase of the greedy algorithm.

The variance plots in Figure 4.4 show more marked increases than the means plots for the non-exhaustive greedy approach. The reason for this is simple: as the networks become larger, the non-exhaustive greedy algorithm becomes less likely to find a good sequence of rewirings (as there are many more such sequences). However, the plots also

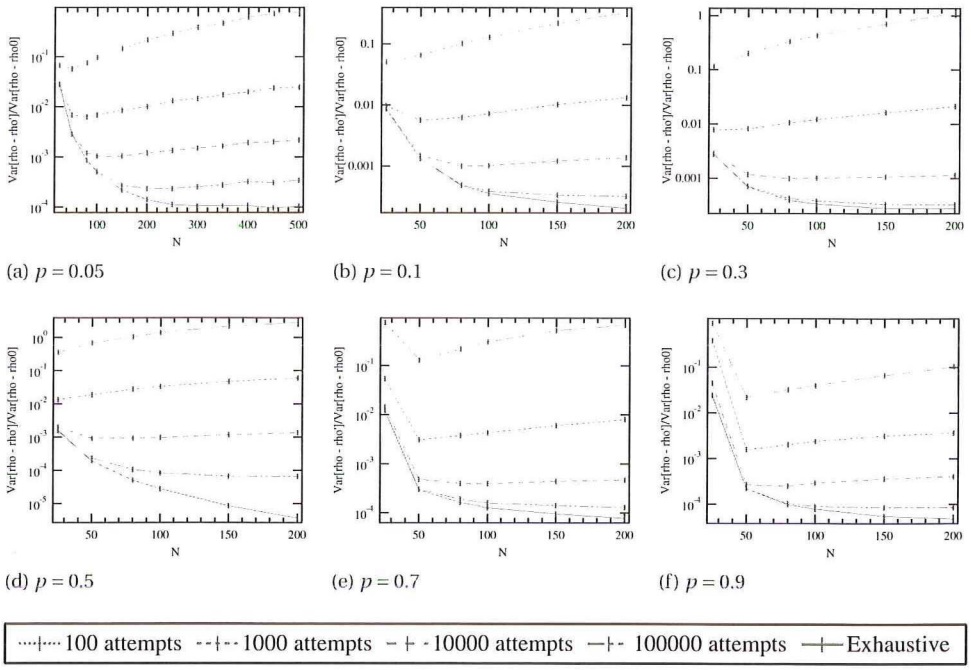


Figure 4.4: Variances of relative differences in solutions obtained by the exact and greedy algorithms for various values of N and p. These plots apply to Erdős-Rényi networks. Each plot corresponds to a fixed p.

suggest a remedy – the algorithm should simply consider more rewirings.

We fitted the power function $y+ax^{-\alpha}$ to each of the exhaustive results and found that all of the functions decrease faster than $O(N^{-\alpha})$, $\alpha > 1.1$. This cements the observation that one is assured of good results if the greedy algorithm makes enough rewirings and that these results get better for larger networks.

RESULTS AS FUNCTIONS OF p

In Section 4.5.1, we considered the performance of the greedy algorithm in terms of node counts. Here, we consider the performance relative to network density. The normalized differences between the exact and greedy algorithms are shown in Figure 4.5. Starting with $N \geq 50$, there are peaks and dips around $p = 0.5$. When the number of random selection trials s is small, the greedy results display peaks, whilst when s is large the results display dips.

A partial explanation for why this happens lies in the number of rewirable configurations available in networks with $p = 0.5$ and in the probability of finding a rewirable link pair during random link selection. The number of rewirable configurations in an Erdős-Rényi network is approximately:

$$3(1-p)^2p^4 + 4(1-p)^3p^3 + 3(1-p)^4p^2 \tag{4.3}$$

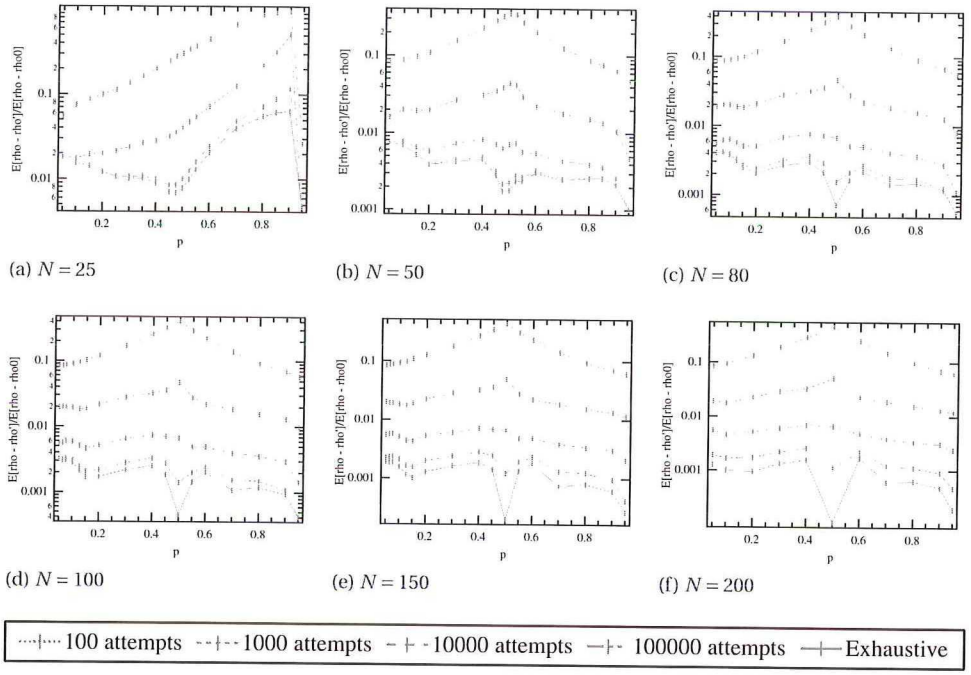


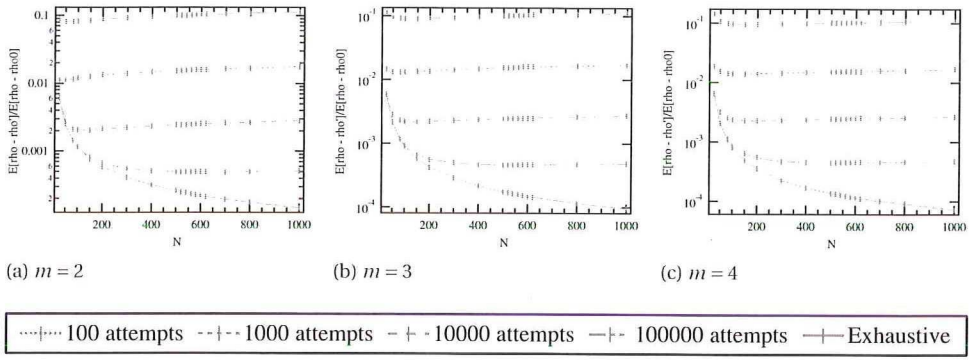
Figure 4.5: Means of relative differences in solutions obtained by the exact and greedy algorithms for various values of N and p. These plots apply to Erdős-Rényi networks. Each plot corresponds to a fixed N.

(the coefficients count the number of isomorphic networks for each of the three configurations). The expression attains a maximum at $p = 0.5$ in the range $p \in [0, 1]$. Thus, an algorithm that is able to find all possible rewirings has ample opportunity for maximizing the assortativity and is less penalized for bad rewiring choices early in the rewiring process. As rewiring proceeds, the number of rewirable configurations decreases (non-linearly) and the probability of finding such rewirable configurations decreases to the point where the non-exhaustive greedy algorithm will fail to find them. Thus, while there may be many rewirable configurations, they are greatly outnumbered by the total number of link pairs.

Some caveats apply to expression (4.3). First, it is a mean-field approximation of the number of rewirable configurations (see Figure 4.1). Second, the expression is not valid for networks that have been rewired (since these networks are no longer Erdős-Rényi network). However, numerical simulations show that when $p = 0.5$, the number of rewirable configurations is indeed maximized (data not shown).

4.5.2. BARABÁSI-ALBERT NETWORKS

To ensure that the results observed for Erdős-Rényi networks are not merely accidental, we also considered Barabási-Albert networks. The sparsity of Barabási-Albert networks



(a) $m = 2$ (b) $m = 3$ (c) $m = 4$

Figure 4.6: Means of relative differences in solutions obtained by the exact and greedy algorithms for various values of N and m in Barabási-Albert networks. Each plot corresponds to a fixed m and each is a function of N.

allowed us to investigate networks with up to 1000 nodes. The means of the differences between the exact and greedy algorithms for Barabási-Albert networks are shown in Figure 4.6 as functions of N (for each m). Most of what applies to the Erdős-Rényi results also applies to the Barabási-Albert results: the greedy algorithm approximates the exact algorithm well and the exhaustive greedy results tend towards the exact greedy results as N increases. Here, we also fitted the power function $y + ax^{-\alpha}$ to the exhaustive results, finding the results decrease faster than $O(N^{-\alpha})$, $\alpha > -1.2$. Possibly due to the use of larger N or possibly due to the different structure of Barabási-Albert networks, these results depart from the Erdős-Rényi results in one key area: there are subtle but constant increases in the non-exhaustive results as N increases. The implication is that the number of samplings s performed by non-exhaustive assortativity-maximization algorithms must be a function $s(N, L)$ of the number of nodes and links in network.

4.5.3. REAL-WORLD NETWORKS

Finally, we applied our algorithms to some real-world networks (see Supplemental Material at [URL will be inserted by publisher] for details). These networks are from diverse areas, making them a good testbed for confirming the trends observed for Erdős-Rényi and Barabási-Albert networks. The real-world network results are shown in Figure 4.7. The networks were sorted in terms of their link counts. These counts span two orders of magnitude, starting at 45 links at the left and ending with 5128 links on the right. The real-world network results confirm our earlier observations (albeit in terms of link counts). On the one hand, the exhaustive greedy algorithm fares progressively better as link counts increase. On the other hand, non-exhaustive runs of the greedy algorithm with fixed random sampling bounds s fare worse as N increases (although this is not so clear when $s = 100000$; this is likely because the link counts are not sufficient to show the same trends as for smaller s). Thus, for increasing link counts, the penalty incurred by the greedy algorithm requires increases in s.

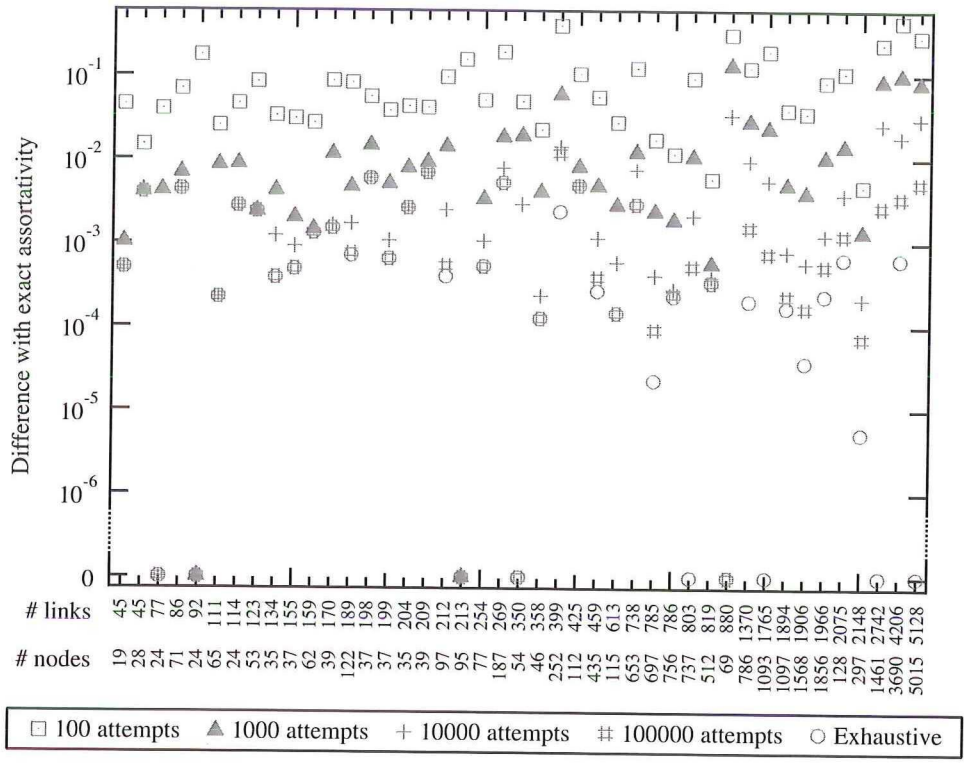


Figure 4.7: Means of relative differences in solutions obtained by the exact and greedy algorithms for a number of real-world networks.

4.6. CONCLUSION

In this paper, we performed the first comparative study between greedy and exact algorithms for maximizing the assortativity of networks under the constraint that their degree vectors remain unchanged. We have focussed only on the maximization of assortativity but our results hold equally for the minimization of assortativity. A few sign changes in our algorithms is all that is required to convert them to minimization algorithms. We applied the algorithms to Erdős-Rényi, Barabási-Albert and real-world networks of varying sizes and link configurations. The overall theme is clear: the greedy assortativity-maximization algorithm approximates the exact algorithm well. We have shown that for all the considered Erdős-Rényi and Barabási-Albert networks, the average difference between the results decreases faster than $O(N^{-1.1})$. The results support heuristic approaches such as those of Menche *et al.* [9], provided that the number of steps s is increased as the network size N increases. Our work raises some interesting questions:

- How many steps s does the greedy algorithm require to obtains results within a

given tolerance of the exact algorithm?

- How bad can the results of a single greedy algorithm run be?
- How much better are sophisticated heuristic algorithms than our simple greedy algorithm?

Any approach to these questions would benefit from a faster exact assortativity-maximization implementation, such as the algorithm of Miller and Pekny [12]. Armed with such an implementation, one could investigate (hopefully much) larger networks.

REFERENCES

[1] W. Winterbach, D. Ridder, H. Wang, M. Reinders, and P. Mieghem, *Do greedy assortativity optimization algorithms produce good results?* The European Physical Journal B **85**, 1 (2012).

[2] M. E. J. Newman, *Mixing patterns in networks,* Physical Review E **67**, 026126 (2003).

[3] M. E. J. Newman, *Mixing patterns in networks,* Physical Review E **67**, 026126 (2003).

[4] H. Wang, W. Winterbach, and P. Van Mieghem, *Assortativity of complementary graphs (to appear),* The European Physical Journal B (2011).

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

[6] T. S. Evans, *Exact solutions for network rewiring models,* The European Physical Journal B – Condensed Matter and Complex Systems **56**, 65 (2007).

[7] J. Lindquist, J. Ma, P. van den Driessche, and F. H. Willeboordse, *Network evolution by different rewiring schemes,* Physica D: Nonlinear Phenomena **238**, 370 (2009).

[8] S. Maslov and K. Sneppen, *Specificity and stability in topology of protein networks,* Science **296**, 910 (2002).

[9] J. Menche, A. Valleriani, and R. Lipowsky, *Asymptotic properties of degree-correlated scale-free networks,* Physical Review E **81**, 046103+ (2010).

[10] P. Van Mieghem, H. Wang, X. Ge, S. Tang, and F. A. Kuipers, *Influence of assortativity and degree-preserving rewiring on the spectra of networks,* The European Physical Journal B - Condensed Matter and Complex Systems , 1 (2010).

[11] C. H. Papadimitriou and K. Steiglitz, *Combinatorial optimization: algorithms and complexity* (Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1982).

[12] D. L. Miller and J. F. Pekny, *A Staged Primal-Dual Algorithm for Perfect b-Matching with Edge Capacities,* INFORMS Journal on Computing **7**, 298 (1995).

[13] Y. Shiloach, *Another look at the degree constrained subgraph problem,* Information Processing Letters **12**, 89 (1981).

[14] V. Kolmogorov, *Blossom v: A new implementation of a minimum cost perfect matching algorithm*, Mathematical Programming Computation **1**, 43 (2009).

[15] B. Bollobás, *Random Graphs* (Cambridge University Press, 2001).

[16] A. L. Barabasi and R. Albert, *Emergence of scaling in random networks*, Science **286**, 509 (1999).

5

ROBUSTNESS ENVELOPES OF NETWORKS

Stojan TRAJANOVSKI, Javier MARTÍN-HERNÁNDEZ, Wynand WINTERBACH, Piet VAN MIEGHEM

5.1. ABSTRACT

We study the robustness of networks under node removal, considering random node failure, as well as targeted node attacks based on network centrality measures. Whilst both of these have been studied in the literature, existing approaches tend to study random failure in terms of average-case behavior, giving no idea of how badly network performance can degrade purely by chance. Instead of considering average network performance under random failure, we compute approximate network performance probability density functions as functions of the fraction of nodes removed. We find that targeted attacks based on centrality measures give a good indication of the worst-case behavior of a network. We show that many centrality measures produce similar targeted attacks and that a combination of degree centrality and eigenvector centrality may be enough to evaluate worst-case behavior of networks. Finally, we study the robustness envelope and targeted attack responses of networks that are rewired to have high and low degree assortativities, discovering that moderate assortativity increases confer more robustness against targeted attacks whilst moderate decreases confer more robustness against random uniform attacks.

This chapter was published without its appendix in the Journal of Complex Networks **1**, 44 (2013) [1].

5.2. INTRODUCTION

In a world where critical infrastructure is composed of and controlled by complex networks, techniques for determining network robustness are essential for the design of reliable infrastructure. After an architecture-dependent number of failures, a network can no longer perform its core function. For example, a telecommunications network whose hubs are removed may be partitioned into many disconnected parts, effectively rendering communication impossible. Appropriate performance metrics can quantify the robustness of a network to such failures.

Network failure is caused by unintentional failures and intentional attacks. Unintentional failures include human error, manufacturing defects and worn-out mechanical parts. These kinds of failures appear randomly and are characterized as *random attacks* [2, 3]. Intentional attacks, on the other hand, are not random and are aimed at maximizing damage. In the literature, they are known as *targeted attacks* [4–6].

In this paper, we study the robustness of network topologies under various challenges. We apply our methodology to random network models and real networks. Our contributions can be summarized as follows: (1) instead of only considering a network average performance, we perform a more comprehensive and granular statistical analysis which shows how all the realizations of random and worst-/best- case targeted removals affect the network performance, but also how do the realizations differ from one another; (2) by studying centrality rankings similarities, we show that some are redundant and degree centrality and eigenvector centrality may be enough to evaluate worst-case behavior of networks; (3) by changing a network by assortativity optimization degree-preserving rewiring, we find that moderate assortativity increases confer more robustness against targeted attacks whilst moderate decreases confer more robustness against random uniform attacks.

The paper is organized as follows. In Section 5.3, we review existing robustness frameworks. Our robustness envelope metrics are presented in Section 5.4. In Section 5.5 metric envelopes of random networks as well as real-world networks are studied. In Section 5.6, we consider the extent to which different targeted attack strategies overlap. Section 5.7 explores changes to the envelope of a network under degree-preserving rewiring. The paper concludes with Section 5.8.

5.3. RELATED WORK

Network robustness has been studied by a number of researchers but the lack of a common vocabulary has made cooperation difficult. Several terms related to robustness have been proposed over the last fifty years, including *reliability, resilience, safety, maintainability, dependability* and *degree-distribution entropy* [7–10]. Meyer [11] studied robustness in the context of his performability framework [12], whilst Cholda *et al.* [13] surveyed various robustness frameworks. In previous research [14–16], maintenance of *connectivity* under failure has typically been used to characterize network robustness. Connectivity has been studied from a probabilistic point of view in the context of graph percolation [17, 18] and reliability polynomials [19]. Most probabilistic studies assume that link failures are independent and that failures occur with the same, fixed probability.

Since the behaviors of topological metrics depend on the characteristics of the net-

works to which they are applied, robustness profiles based on these metrics also depend on these characteristics. Therefore, researchers have studied robustness in the context of various network types. Callaway *et al.* [20] and Holme *et al.* [4] have studied the robustness of random networks and power-law graphs. In particular, Cohen *et al.* have examined the robustness of the Internet and other power-law networks under random [2] and targeted [6] failures. Recently, the robustness of time-evolving networks or temporal graphs [21, 22] has been researched in [3, 23]. A method based on the cumulative change of the giant component under targeted attacks has been proposed by Schneider *et al.* [24]. Çetinkaya *et al.* [25] developed a framework for analyzing packet loss relative to node and link failure. They consider packet loss under global targeted and random failure, as well as attacks contained within geographic regions. Our approach is similar to their approach, although we consider not only average network performance under random attacks but the density function given the probability that a metric will assume a given value after a given fraction of node removals.

5.4. ENVELOPE COMPUTATION AND COMPARISON

In this section, we propose a framework to quantify network robustness. We assume that a network can be expressed as a graph G, defined by a set $\mathcal{N}$ of N *nodes* interconnected by a set $\mathcal{L}$ of L *links*. With this formalism, various aspects of the network can be described by means of graph *metrics* which are typically real-valued functions of the network.

5.4.1. ROBUSTNESS AND THE R-VALUE

We define robustness as the maintenance of function under node or link removal. In this context, function is measured by one or more graph metrics. As in [9], we express robustness as a real-valued function R of graph metrics, normalized to the range $[0,1]$. A value of $R = 0$ means that the network is completely non-functional, whereas $R = 1$ means that the network is fully functional.

Here, we consider two different R-values, computed using the 1) *size of the giant component* and 2) *efficiency*. The choice of these metrics is arbitrary and it depends on the network function. The method presented translates easily to other sets of metrics.

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

2) **Efficiency.** The efficiency [26] of a given network G is the mean of the reciprocals of all the hopcounts in a network

$$E[1/H] = \frac{\sum_{1 \le i < j \le N} 1/h_{i,j}}{\binom{N}{2}}.$$

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

5.4.2. NETWORK PERTURBATIONS OR CHALLENGES

A perturbation or challenge P is defined as a set of elementary changes [9]. Elementary changes include: (1) addition of a node, (2) removal of a node, (3) addition of a link, (4) removal of a link and, (5) in weighted networks, a change in the weight of a link (or node). We consider only node removals, but our analysis can be extended to all five perturbation types. A *realization* is a vector $[P_1, P_2, \ldots, P_N]$ of perturbations, where P_i is a subset of i nodes. In addition, a realization is called *successive* iff $P_1 \subset P_2 \subset \ldots \subset P_N$. Since every perturbation has an associated R-value, any realization can also be expressed as a sequence of R-values denoted $\{R[k]\}_{0 \leq k \leq 1}$, where k is the fraction of removed nodes.

5.4.3. RANDOM ATTACKS AND TARGETED ATTACKS

Network perturbations are classified either as random (un-intentional) failures [2] or as targeted attacks [4, 5].

RANDOM ATTACKS

Assuming that the nature of the attacks is unknown and attacks occur independently, $R[k]$ is a random variable. We employ *probability density function* (*PDF*), which is the probability of a random variable to fall within a particular region. The PDF of this $R[k]$ is computed using all subsets of $\lfloor kN \rfloor$ (i.e., the integer part of kN) nodes of the set $\mathscr{P}_r$ of all possible perturbations. The envelope for a graph G is constructed using all $R[k]$ for $k \in \{\frac{1}{N}, \frac{2}{N}, \ldots, 1\}$, where boundaries are given by the extreme R-values

$$R_{min}^{(\mathscr{P}_r)}[k] = [min(R[\tfrac{1}{N}]), min(R[\tfrac{2}{N}]), \ldots, min(R[1])]$$

and

$$R_{max}^{(\mathscr{P}_r)}[k] = [max(R[\tfrac{1}{N}]), max(R[\tfrac{2}{N}]), \ldots, max(R[1])].$$

Such boundaries can be seen in FIG. 5.1. Although extreme R-values give the best- and worst-case metrics for a network after a given number of perturbations, we are just as often interested in the expected R-value resulting from k perturbations

$$R_{avg}^{(\mathscr{P}_r)}[k] = [E[R[\tfrac{1}{N}]], E[R[\tfrac{2}{N}]], \ldots, E[R[1]])].$$

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

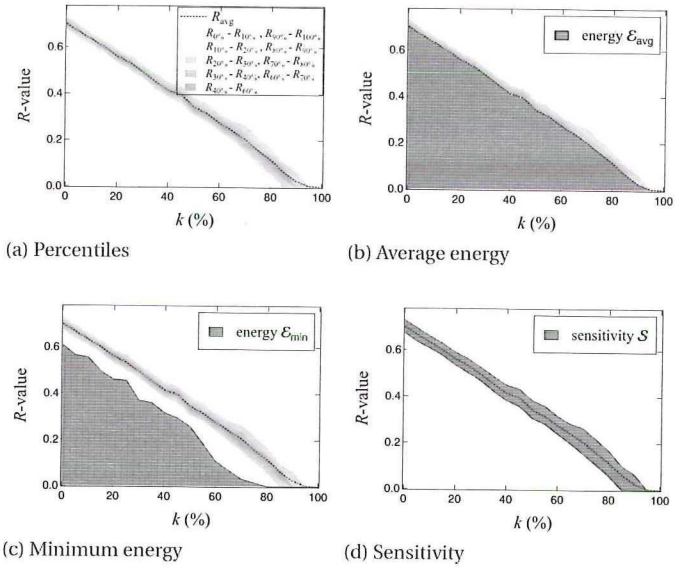
$$R_{m\%}^{(\mathscr{P}_r)}[k] = [R_{m\%}[\tfrac{1}{N}], R_{m\%}[\tfrac{2}{N}], \ldots, R_{m\%}[1])]$$

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

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

We refer to $R_{m\%}[k]$ as an m-percentile. By definition $R_{0\%}[k] = R_{min}[k]$, and $R_{100\%}[k] = R_{max}[k]$. The dark-gray areas in FIG. 5.1a are bounded by low-percentile lines whereas the lighter-gray areas correspond to higher-percentile lines.

In the case where $\lfloor kN \rfloor$ nodes in the network are attacked, $\binom{N}{\lfloor kN \rfloor}$ R-values need to be computed. It has been shown that the problem of finding a set of nodes minimizing $R[k]$ is NP-complete [27]. For this reason, we perform random sampling to approximate the PDF of $R[k]$ and targeted attacks to approximate the maxima and minima of the PDFs.



(a) Percentiles (b) Average energy

(c) Minimum energy (d) Sensitivity

Figure 5.1: Depictions of the robustness envelopes defined in Section 5.4. The x-axis represents the fraction of attacked nodes.

TARGETED ATTACKS

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

Centrality measures may provide a set of such rankings. We consider five different measures: (a) *node degree*; (b) *betweenness* [28]; (c) *closeness* [29] and (d) *eigenvector centrality* [30]. In Section 5.6 we study the extent to which these rankings overlap.

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

5.4.4. COMPARISON OF NETWORKS VIA ENVELOPES

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

sensitivity $\mathcal{S}$.

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

$$\mathcal{E}^{(\mathcal{P})} = \frac{1}{K}\sum_{k=1}^{K} R^{(\mathcal{P})}[k] \tag{5.1}$$

where $K = |\mathcal{P}|$. Energy expresses how robust, on average, a graph is against a given type of attack. For instance, if $\mathcal{E}^{(\mathcal{P})}_{G_1} > \mathcal{E}^{(\mathcal{P})}_{G_2}$, G_1 has higher energy than G_2 with respect to the perturbation $\mathcal{P}$. Other examples of energy include those computed from the maximal realization $\mathcal{E}^{(\mathcal{P})}_{max}$, minimal realization $\mathcal{E}^{(\mathcal{P})}_{min}$, expected realization $\mathcal{E}^{(\mathcal{P})}_{avg}$, and m-percentile realization $\mathcal{E}^{(\mathcal{P})}_{m\%}$, as illustrated in FIGS. 5.1b-5.1c.

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

$$\mathcal{S}^{(\mathcal{P})} = \mathcal{E}^{(\mathcal{P})}_{80\%} - \mathcal{E}^{(\mathcal{P})}_{20\%}. \tag{5.2}$$

The sensitivity $\mathcal{S}$ indicates how likely the R-value is to shift upon random removals, as illustrated in FIG. 5.1d. The smaller the sensitivity, the narrower the uncertainty of the R-value, thus the better the robustness. The sensitivity together with the percentiles of R-values express the variability of different random attacks in a given network.

5.5. ROBUSTNESS OF RANDOM AND REAL NETWORKS

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

We consider four network models with different structural properties: Erdős-Rényi networks, Watts-Strogatz networks, Barabási-Albert networks, and lattices. Erdős-Rényi networks [31, 32] are a 2-parameter family of random networks denoted $G_p(N)$. The parameter N is the number of nodes in the network whilst the parameter p is the probability that two nodes are connected by a link. Watts-Strogatz $W(N, q, p)$ networks [33] are a family of networks with small-world properties, whose main features are small average shortest paths and high clustering coefficients. Initially, a Watts-Strogatz instance is a regular ring lattice in which each node is connected to q neighbors. The topology is then randomized by replacing, with a probability p, an incident node of each link with a random node, provided that no self-loops or multiple links between nodes are introduced. Barabási-Albert networks [34] are a family of *scale-free* networks whose architectures emerge from preferential attachment. Initially a Barabási-Albert network instance has m_0 nodes. The remaining $N - m_0$ nodes are added one at a time, each one connected by m links to already-placed nodes with probabilities proportional to the degrees of those nodes. We also consider rectangular lattice networks. A lattice $L_{N\times M}$ has NM nodes; the central $(N-2)(M-2)$ nodes have degree 4; the $2(N+M-2)$ non-corner nodes have degree 3 and the 4 corner nodes have degree 2.

The instances of the network models considered in this paper all have $N = 100$ nodes, except for the lattices. We consider (sparse) networks with $L \approx 500$ links, as well as (relatively dense) networks with $L \approx 3200$ links. The parameter choices of our network models

are therefore chosen to generate networks with (approximately) these link counts. The rewiring probability for the Watts-Strogatz instances is chosen to be $p = 0.1$, leading to networks with high clustering coefficients and low average hop-counts (this is called the *small-world* regime). The lattice network does not accept any input parameters, hence we displayed two arbitrarily chosen lattices: a square-like with 20 by 20 nodes, and a stretched lattice with 100 by 10 nodes.

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

Network	N	L	Description
USp	4941	6594	Western US power grid network [33]
CA	5242	14484	Co-authorship network [35]
EUr	8730	11350	Western Europe railway network
EUp	9168	10417	Western Europe power grid network

In addition to instances of random network models, we consider four real-world networks. First, are the high-voltage power grids of the Western United States [33] and of Western Europe [36]. In the remainder of the paper, we refer to these two networks as USp and EUp respectively. Nodes represent power stations, transformers and generators and links represent high-voltage connections between nodes. Second, we study a social collaboration network from ArXiv that covers papers joining authors in the field of Relativity and Quantum Cosmology [35] in the period January 1993 to April 2003. We refer to this network as CA. Here, two nodes are joined if the two authors appear as co-authors in at least one paper. Finally, we consider the Western European Railway network, referred to as EUr. The nodes in the network represent railway stations and links represent railway tracks between stations. The size of each real network is given in Table 5.1.

5.5.1. THEORETICAL PRELIMINARIES

Let us denote by $G(N;k)$ a network with N nodes which has had a fraction k of its nodes attacked. Before any attacks, the network is thus denoted by $G(N;0)$. We are interested in calculating the change of the network metric $R = R^{(\mathscr{P})}_{G(N;k)}$ as a function of the percentage of attacked nodes k. Denote by $\mathscr{T}$ the set of nodes that have been attacked and denote by $\mathscr{N}\setminus\mathscr{T}$ the nodes that have not been attacked. Here, $\mathscr{N}$ is the set of all nodes in the network. The number of attacked nodes in $G(N,k)$ is $m = |\mathscr{T}| = \lfloor kN \rfloor$ and therefore the number of nodes that have not been attacked is $N - m = |\mathscr{N}\setminus\mathscr{T}| = N - \lfloor kN \rfloor$.

A metric, such as efficiency, whose value is the average over all node pairs is dealt with in a similar fashion. Denote by R_{ij} the contribution of a pair of nodes i and j ($i \neq j$) to the R-value. If either node i or j has been removed (that is, $i \in \mathscr{T}$ or $j \in \mathscr{T}$), $R_{ij} = 0$. Thus,

$$R = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} R_{ij} = \frac{1}{N(N-1)} \sum_{i,j \in \mathscr{N}\setminus\mathscr{T}, i \neq j} R_{ij}. \tag{5.3}$$

5.5.2. ANALYTICAL RESULTS FOR ERDŐS-RÉNYI NETWORKS

Here, we provide analytical results for the robustness of Erdős-Rényi random networks relative to the efficiency and size of the giant component. In the case of random re-

moval, where $k\%$ of the nodes are discarded, the resulting network has N' nodes of degree 0. The remaining nodes form an Erdős-Rényi random network $G_p(N-m)$ with the same link density p because the link between two nodes from $\mathcal{N}\setminus\mathcal{T}$ appears with a fixed probability p. Targeted attacks afford no such easy analysis, making them much less analytically tractable.

Efficiency. The average efficiency is the reciprocal of the mean hopcount, which is approximately $h_{ij} \approx \frac{\ln(N)}{\ln(Np)}$ for an arbitrary pair of nodes i and j in a connected Erdős-Rényi network [37, 38]. Consequently, the efficiency e_{ij} for the pair i,j is $e_{ij} = \frac{1}{h_{ij}} \approx \frac{\ln(Np)}{\ln(N)}$. Consider the independent, random removal of $k\%$ of the nodes. The resulting network is an Erdős-Rényi network $G_p(N-\lfloor kN\rfloor)$ with $N' = \lfloor kN\rfloor$ isolated nodes. Thus, the efficiency e_{ij} of an arbitrary pair of nodes i and j is approximately

$$e_{ij} = \begin{cases} \frac{\ln((N-\lfloor kN\rfloor)p)}{\ln(N-\lfloor kN\rfloor)}, \text{ for } i,j \in \mathcal{N}\setminus\mathcal{N}' \\ 0, \text{ otherwise.} \end{cases} \tag{5.4}$$

Substituting (5.4) into (5.3), yields

$$\begin{aligned} E[1/H] &= \frac{\sum_{i,j\in\mathcal{N}\setminus\mathcal{N}', i\neq j} e_{ij}}{N(N-1)} = \frac{\sum_{i,j\in\mathcal{N}\setminus\mathcal{N}', i\neq j} \frac{\ln((N-\lfloor kN\rfloor)p)}{\ln(N-\lfloor kN\rfloor)}}{N(N-1)} \approx \frac{\frac{\ln((1-k)Np)}{\ln((1-k)N)} \sum_{i,j\in\mathcal{N}\setminus\mathcal{N}', i\neq j} 1}{N(N-1)} \\ &= \frac{\frac{\ln((1-k)Np)}{\ln((1-k)N)} N'(N'-1)}{N(N-1)} \approx \frac{\frac{\ln((1-k)Np)}{\ln((1-k)N)} (N-kN)^2}{N^2} = (1-k)^2 \frac{\ln\left((1-k)Np\right)}{\ln\left((1-k)N\right)}. \end{aligned} \tag{5.5}$$

The shape of (5.5) is validated by FIGS. 5.3a and 5.3b.

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

$$S \leq 1-k \tag{5.6}$$

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

$$S = 1-k, \text{ if } p > \frac{\ln(N-\lfloor kN\rfloor)}{N-\lfloor kN\rfloor}. \tag{5.7}$$

The function $\frac{\ln(N-\lfloor kN\rfloor)}{N-\lfloor kN\rfloor}$ increases with the percentage of attacked nodes k. Thus, for fixed values of p and N and large enough values of k, $p \leq \frac{\ln(N-\lfloor kN\rfloor)}{N-\lfloor kN\rfloor}$. As this is the connectivity threshold for Erdős-Rényi networks, we find that $S < 1-k$. The "dips" in the lines $R = 1-k$ for large k in FIGS. 5.2a and 5.2b are manifestations of disconnected giant components. As can be seen in FIG. 5.2a, when p is small, disconnection happens for smaller values of k. The size of the giant component is approximately [37]

$$S = 1 - e^{-p(N-\lfloor kN\rfloor)S}$$

which explains the "dip" in the linear line $R = (1-k)$. In the analysis for the size of the giant component, we consider $R = S$, however a slightly similar approach is comparing the absolute values by taking $R = S/S[0]$, where $S[0]$ is the size of the giant component in the original network. Clearly, both approaches are identical if the original network does not have disconnected parts.

5.5.3. ROBUSTNESS OF RANDOM NETWORK MODEL INSTANCES

In this section, we interpret simulation results of the random network model instances. The properties of the network models considered in the analysis are stated at the beginning of this section (Section 5.5). The simulations have been repeated 1000 times to obtain the energy, the sensitivity and R values.

SIZE OF THE GIANT COMPONENT.

Energy analysis: The maximum energies for all strategies and networks exceed 0.460 (0.5 is the maximum energy attainable for the giant component, as the slope of R-value cannot exceed $(1-k)$. The R-values for the giant component are shown in FIG. 5.2 and Table 1 in the online supplement of the paper. For almost all networks, there are sequences of node removals that render large giant components. In addition, lattice networks show interesting behavior: there seems to be a phase transition around 50% as seen in FIG. 5.2g. After randomly removing more than 50% of the nodes, all the topologies lose energy at an increased rate, due to the loss of connectivity. This result is in accordance with percolation theory [39], where the critical probability of bond percolation equals $0.5N$.

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

Targeted versus random attacks: Amongst the random network models, the ratio $\mathscr{E}_{min}/\mathscr{E}_{avg}$ attains the highest value for Barabási-Albert networks (an unfavorable condition), followed by Erdős-Rényi networks and finally Watts-Strogatz networks. As with efficiency, the lattice network has the highest ratio $\mathscr{E}_{min}/\mathscr{E}_{avg}$ for all targeted strategies, peaking at 1.42 for node-degree targeted attacks. Again, this means that, for grid networks, the targeted strategies perform worse (on average) than a random strategy. The ratio $(\mathscr{E}_{max}-\mathscr{E}_{min})/\mathscr{S}$ is the highest for Barabási-Albert networks, followed by Erdős-Rényi networks, Watts-Strogatz networks and lattices. Targeted attacks have the largest impact on Barabási-Albert networks, whilst Erdős-Rényi networks are the least affected. Table 5.2 shows that the most destructive perturbations are those based on degree and betweenness centrality.

EFFICIENCY

As can be seen from FIG. 5.3, amongst the sparse networks, the lattice has the lowest average efficiency energy, followed by G_{WS} (with $q = 10$). Both of these networks are fairly regular (G_{WS} has a low rewiring probability in our paper) leading us to conclude that regularity does not confer robustness in terms of efficiency. G_{BA} networks are the most robust to random attacks as well as being the most sensitive, making them the most vulnerable to targeted attacks. Again, G_{ER} networks win in terms of energy and sensitivity, making them robust both to random and targeted attacks.

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

	G_{ER}	G_{WS}	G_{BA}	Lattice
Betweenness	$-R^{(top)}$	$-R^{(top)}$	$-R^{(top)}$	
Closeness				$-R^{(top)}$
Degree			$-R^{(top)}$	
Eigenvector				

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

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

	G_{ER}	G_{WS}	G_{BA}	Lattice
Betweenness	$-R^{(top)}$	$-R^{(top)}$	$-R^{(top)}$	
Closeness	$+R^{(bot)}$	$-R^{(top)}, +R^{(bot)}$	$+R^{(bot)}$	$+R^{(bot)}$
Degree	$-R^{(top)}$		$-R^{(top)}$	
Eigenvector	$+R^{(bot)}$	$+R^{(bot)}$	$+R^{(bot)}$	$-R^{(top)}$

5.5.4. ROBUSTNESS OF REAL NETWORKS

In this section, we compare the robustness profiles of real-world networks to the robustness profiles of the network models presented in the previous section. Many numerical details regarding the *energy* and the *sensitivity* are given in Table 3 of the online supplement of the paper.

THE SIZE OF THE GIANT COMPONENT

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

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

As before in Section 5.5.2, the most effective attack strategies are the node degree and node betweenness attacks. The least effective attack strategy is the node closeness

attack (e.g. FIG. 5.4c), which leaves the size of the giant component nearly untouched for all real networks.

EFFICIENCY

The network with the highest absolute efficiency value is the co-authorship network. As before, this is due to the high link density and the presence of many cliques. Remarkably, all four real-world networks show rapid decreases in efficiency after only $\approx 10\%$ of their nodes are removed. This behavior is similar to that observed for the Barabási-Albert model: in this case, the removal of $\approx 20\%$ of the nodes causes a large drop in efficiency. But more importantly, the dramatic drop in the R-value occurs for both random and (most) targeted strategies. FIG. 5.4f illustrates this effect, also seen in the $\mathscr{E}_{min}/\mathscr{E}_{avg}$ ratios in the online supplement of the paper (Table 3 there). In conclusion, sparse real-world networks are easily disconnected, regardless of the type of attack. As with the results in Section 5.5.2, the attack with the lowest *min R-value* is the node betweenness attack.

5.6. SIMILARITY OF NODE-CENTRALITY MEASURES

Centrality measures express the relative importance of nodes within a graph. Different centrality measures rank nodes differently. To quantify the similarity of centrality rankings, we define a centrality similarity metric.

For two node rankings $A = [a_{(1)}, a_{(2)}, \ldots, a_{(N)}]$ and $B = [b_{(1)}, b_{(2)}, \ldots, b_{(N)}]$, $M_{A,B}(k)$ is the percentage of nodes in $\{a_{(1)}, a_{(2)}, \ldots, a_{(\lfloor kN \rfloor)}\}$ that also appear in $\{b_{(1)}, b_{(2)}, \ldots, b_{(\lfloor kN \rfloor)}\}$.

The measure $M_{A,B}(k)$ is different from the scalar correlation of topological metrics [40]. When we compare all the nodes ($k = 100\%$), we have a full overlap and $M_{A,B}(100\%) = 1$. In other words, $M_{A,B}(k)$ gives the percentage of overlapping nodes from the top $k\%$ of nodes in the rankings A and B. For instance, it reveals whether the nodes with the highest betweenness values are also those with the highest degrees.

The results of $M_{A,B}(k)$ for real-world networks are given in FIG. 5.5. From the figure, we observe that

$M_{\text{closeness, eigenvector}}(k)$ generally has the highest value and that it is closely followed by $M_{\text{degree, betweenness}}(k)$. On the other hand, $M_{\text{betweenness, eigenvector}}(k)$ shows that there is little overlap between the node rankings derived from the betweenness and eigenvector centrality measures. In both the US and the European power grid networks (FIGS. 5.5c and 5.5d), $M_{\text{degree, betweenness}}(k)$ attains large values. On the other hand, in the citation and railway networks (FIGS. 5.5a and 5.5b), $M_{\text{closeness, eigenvector}}(k)$ attains large values.

The measure $M_{A,B}(k)$ is small when the rankings A and B differ in the nodes that are deemed central. In such cases, both centrality measures should be used as attack strategies, since each strategy could have a different effect in a network.

5.7. ROBUSTNESS OPTIMIZATION BY DEGREE-PRESERVING REWIRING

We demonstrate the use of our robustness framework by studying changes in the metric envelope of a network as it is rewired (through degree-preserving transformations) in order to increase or decrease its *degree assortativity* [41, 42].

5.7.1. DEGREE ASSORTATIVITY

Degree assortativity measures the tendency of links to connect nodes with similar degrees. Formally, it is defined [41] as

$$\rho_D = 1 - \frac{\sum_{i\sim j}(d_i - d_j)^2}{\sum_{i=1}^{i=N} d_i^3 - \frac{1}{2L}(\sum_{i=1}^{N} d_i^2)^2}$$

where $i \sim j$ denotes a link between nodes n_i and n_j, d_i the degree of node n_i and $D = [d_1, d_2, \ldots, d_N]$ the degree-sequence of the network. The degree assortativity has been shown [43] to be an important indicator for the epidemic spread such that assortative networks spread are more prone to the propagation of epidemics. Moreover, the close relation between the degree assortativity and the modularity, which is an indicator for network clusterness, has been studied in [44].

5.7.2. DEGREE-PRESERVING REWIRING

Degree-preserving rewiring [42] allows for the modification of the link architecture of a network without changing its degree sequence. In a rewiring step, a pair of links $\{u, v\}$, $\{w, x\}$ in a network G is selected such that u, v, w and x are distinct nodes. If $\{u, x\} \notin \mathscr{L}(G)$ and $\{w, v\} \notin \mathscr{L}(G)$, $\{u, v\}$ and $\{w, x\}$ can be rewired to (that is, replaced by) $\{u, x\}$, $\{w, v\}$.

5.7.3. REWIRING ALGORITHM FOR ASSORTATIVITY OPTIMIZATION

We used the greedy degree-preserving rewiring algorithm of [45] to optimize degree assortativity. In each iteration, the algorithm samples up to s pairs of links. If a sampled pair of links is rewirable and if the rewiring leads to a desired change in the degree assortativity (see Lemma 1 in [42]) of the network, the change is made. If, after s sampling attempts, no such pair of links is found, the algorithm terminates.

5.7.4. EXPERIMENT SETUP

Using our simple algorithm, we maximized and minimized the degree assortativity of an Erdős-Rényi graph as well as a Barabási-Albert graph. The number of rewirings needed to achieve high or low degree assortativity can number in the hundreds or even thousands. Therefore, it is impractical to study the robustness profiles of the networks associated with each rewiring step. For each network, we study five snapshots: (1) a rewired network whose assortativity is fully maximized; (2) a rewired network whose assortativity is halfway between the fully maximized value and that of the original network; (3) the original network; (4) a rewired network whose assortativity is halfway between the fully minimized assortativity value and that of the original network; and (5) a rewired network whose assortativity is fully minimized. Snapshots of the G_{ER}, along with corresponding *energy* and *sensitivity* changes for the giant component and efficiency are shown in FIG. 5.6. The analogues for G_{BA} are shown in FIG. 5.7.

5.7.5. INTERPRETATION

As assortativity is maximized, the $\mathscr{E}_{avg}$ of both the giant component and efficiency decrease (the black lines in FIGS. 5.6 and 5.7). In the intermediate assortativity-maximized

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

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

Our observations suggest that networks whose assortativities are moderately maximized (through degree-preserving transformations) are more tolerant to targeted attacks whilst having worse average-case robustness. On the other hand, networks whose assortativities are moderately minimized are more tolerant to random attacks (and less tolerant to targeted attacks). These observations match those of Friedel and Zimmer [46], who researched the role of assortativity in protein interaction networks.

5.8. CONCLUSIONS

Within the topological robustness framework [9, 10], we have extended and detailed the concept of robustness envelopes. We studied the robustness envelopes of sparse and dense instances of well-known random classes of networks, as well as four real-world networks. Our envelope approach shows that although networks may have similar average-case performance under attack, they may differ significantly in their sensitivities to certain attack sequences. We also contrasted robustness envelopes of the studied networks to their responses when subjected to targeted attacks. The targeted attacks are all based on node centrality measures.

We found that targeted attack strategies often lead to performance degradation beyond the limits of the robustness envelopes that we computed, leading us to conclude that centrality-based targeted attacks are sufficient for studying the worst-case behavior of real-world networks. In this regard, our analysis suggests that real-world networks are susceptible to rapid degradation under targeted attacks. The overlap between centrality rankings reveals that attack strategies based on different centrality measures may have very similar results. We argue that degree centrality and eigenvector centrality strike a good balance between differences in attack sequences and in computational power required.

Finally, we investigated envelopes and targeted attack patterns of networks whose structures were modified, through degree-preserving rewiring, to optimize their assortativity. We found that by slightly increasing degree assortativity, our networks became more resilient against targeted attacks, if somewhat less resilient against random attacks. The converse was true when decreasing degree assortativity.

An interesting question for future research is whether it is possible to design an efficient method for increasing the worst-case robustness of a network (through rewiring)

without adversely affecting its mean robustness.

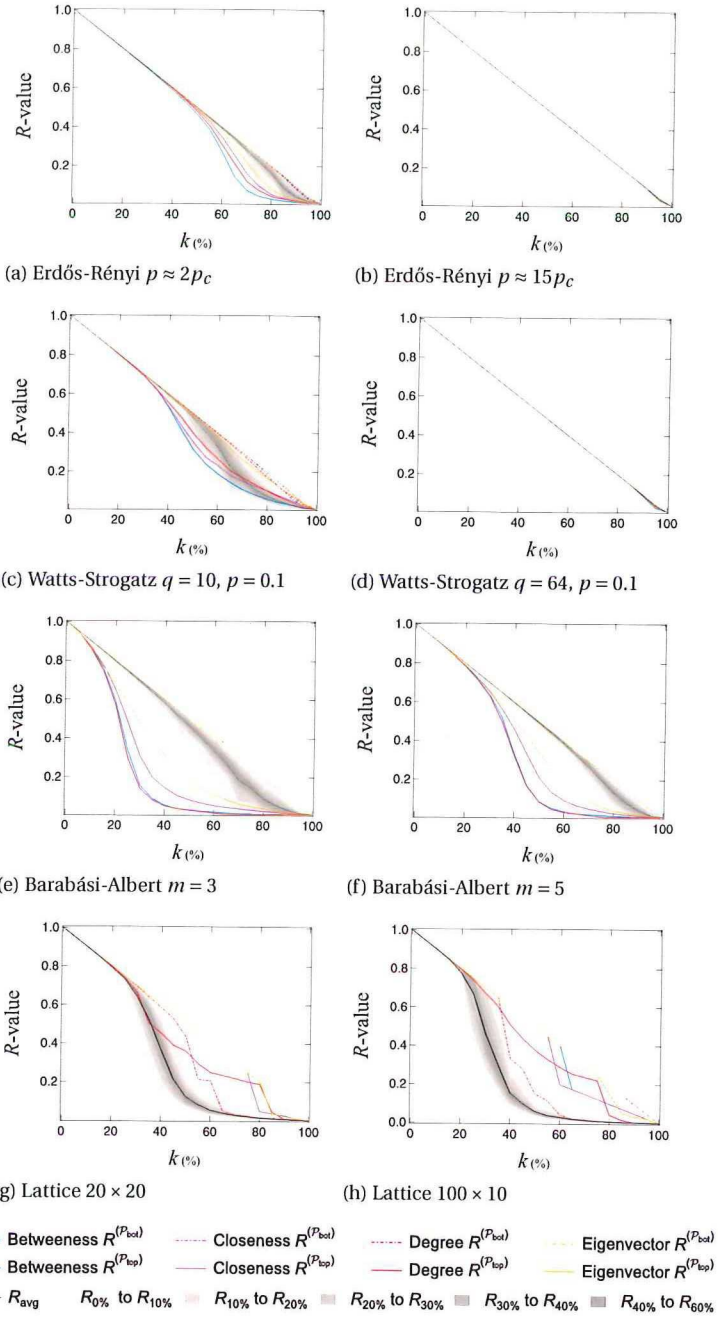
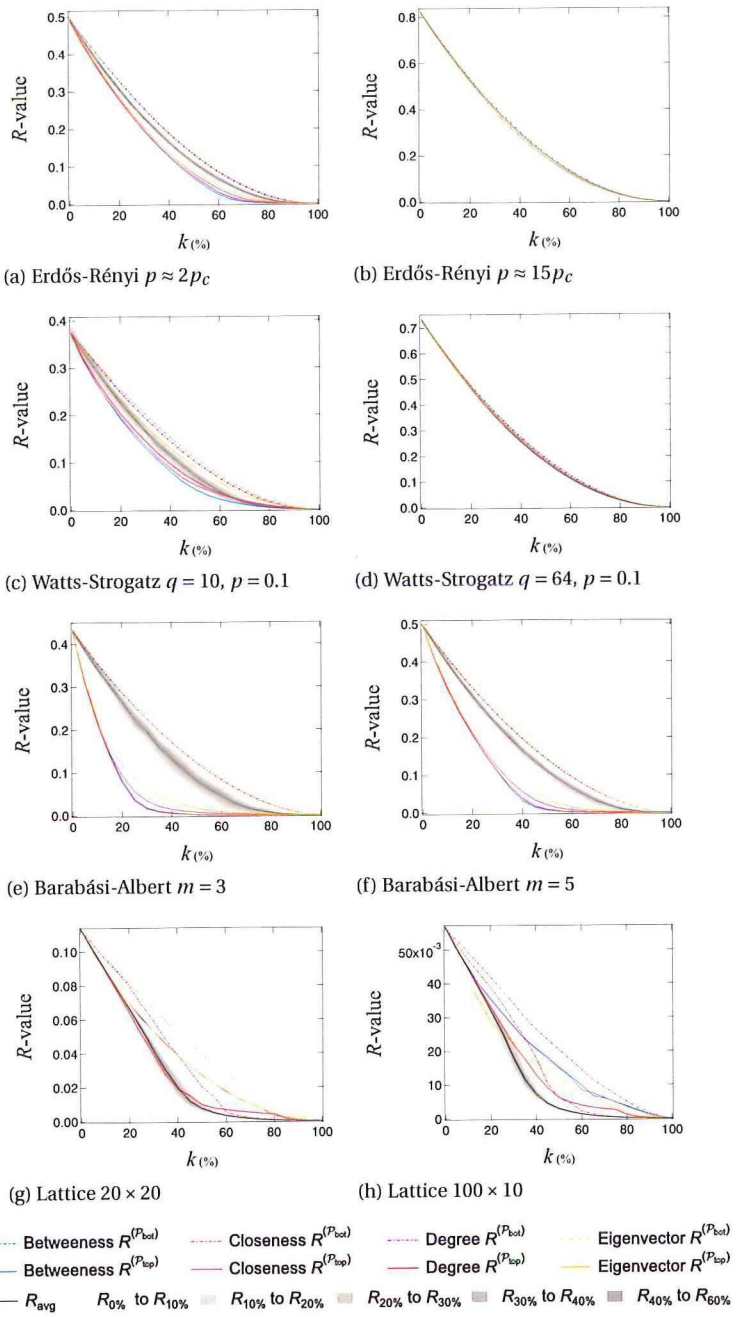


Figure 5.2: The R-values for the giant component size. The network model considered and its property (the link density p for Erdős-Rényi, the number of neighbors q per node and the rewiring probability p in Watts-Strogatz and m the number of links of a newly added node in Barabási-Albert model) is given in sub-captions (a) - (h). The x-axis is the percentage of removed nodes either at random ore according to a centrality measure as it is shown in the legend.



(a) Erdős-Rényi $p \approx 2p_c$

(b) Erdős-Rényi $p \approx 15p_c$

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

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

(e) Barabási-Albert $m = 3$

(f) Barabási-Albert $m = 5$

(g) Lattice 20×20

(h) Lattice 100×10

Figure 5.3: The R-values for the efficiency. The network model considered and its property (the link density p for Erdős-Rényi, the number of neighbors q per node and the rewiring probability p in Watts-Strogatz and m the number of links of a newly added node in Barabási-Albert model) is given in sub-captions (a) - (h). The x-axis is the percentage of removed nodes either at random ore according to a centrality measure as it is shown in the legend.

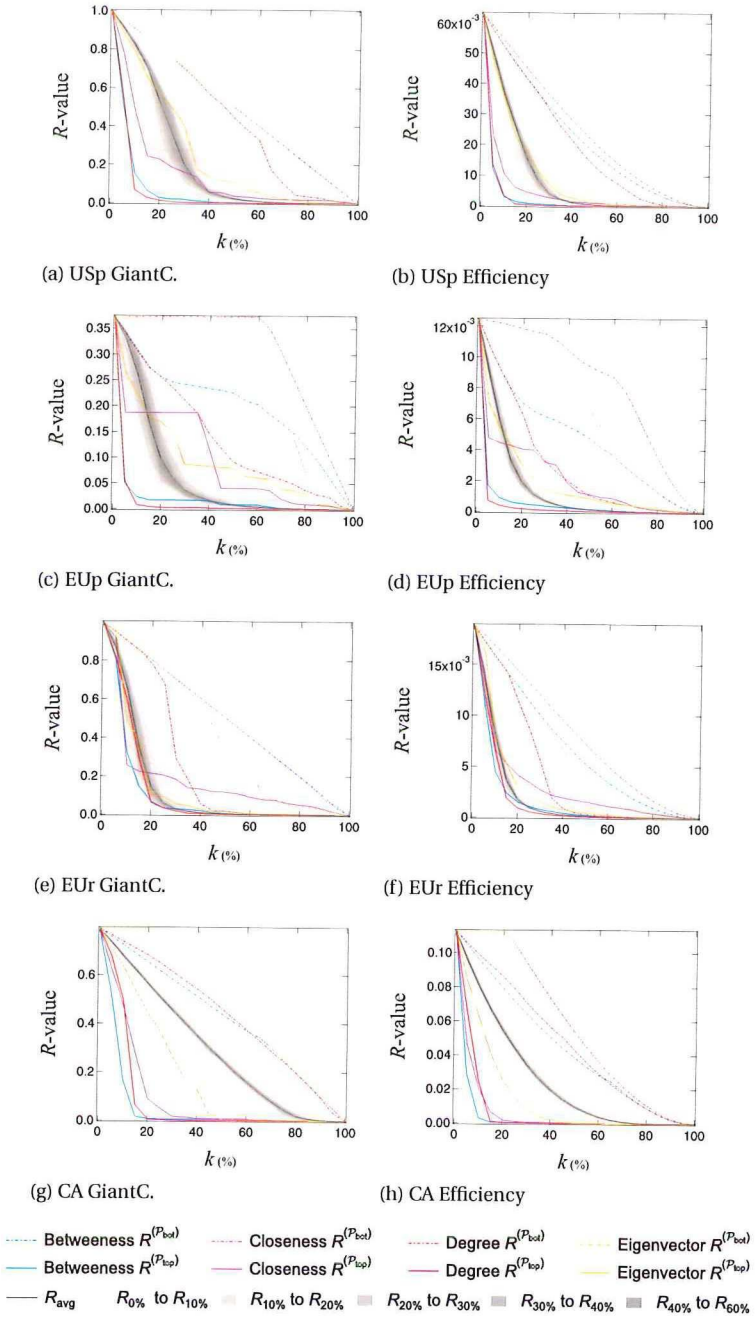
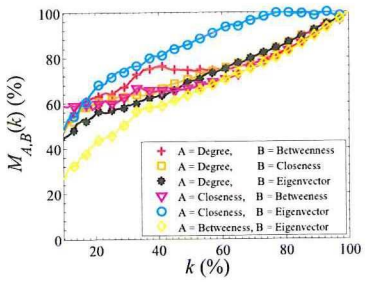
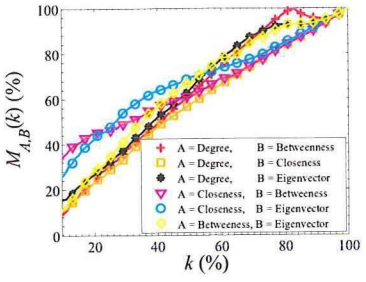


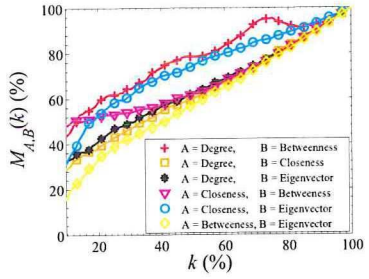
Figure 5.4: R-values for the real-world networks. The network considered and the metric reflecting R-value are given in sub-captions (a) - (h). The x-axis is the percentage of removed nodes either at random ore according to a centrality measure as it is shown in the legend.



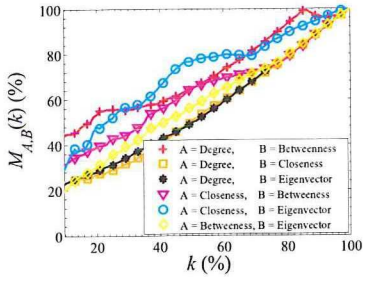
(a) Collaboration network



(b) European railway network



(c) US power grid



(d) West. European power grid

Figure 5.5: Similarities of centrality rankings $M_{A,B}(k)$ for real networks. Each plot shows the overlap of nodes (y-axis in %) from the first k% nodes (x-axis) ranked according to centrality ranking A and the first k% nodes ranked according to centrality ranking B for a given network.

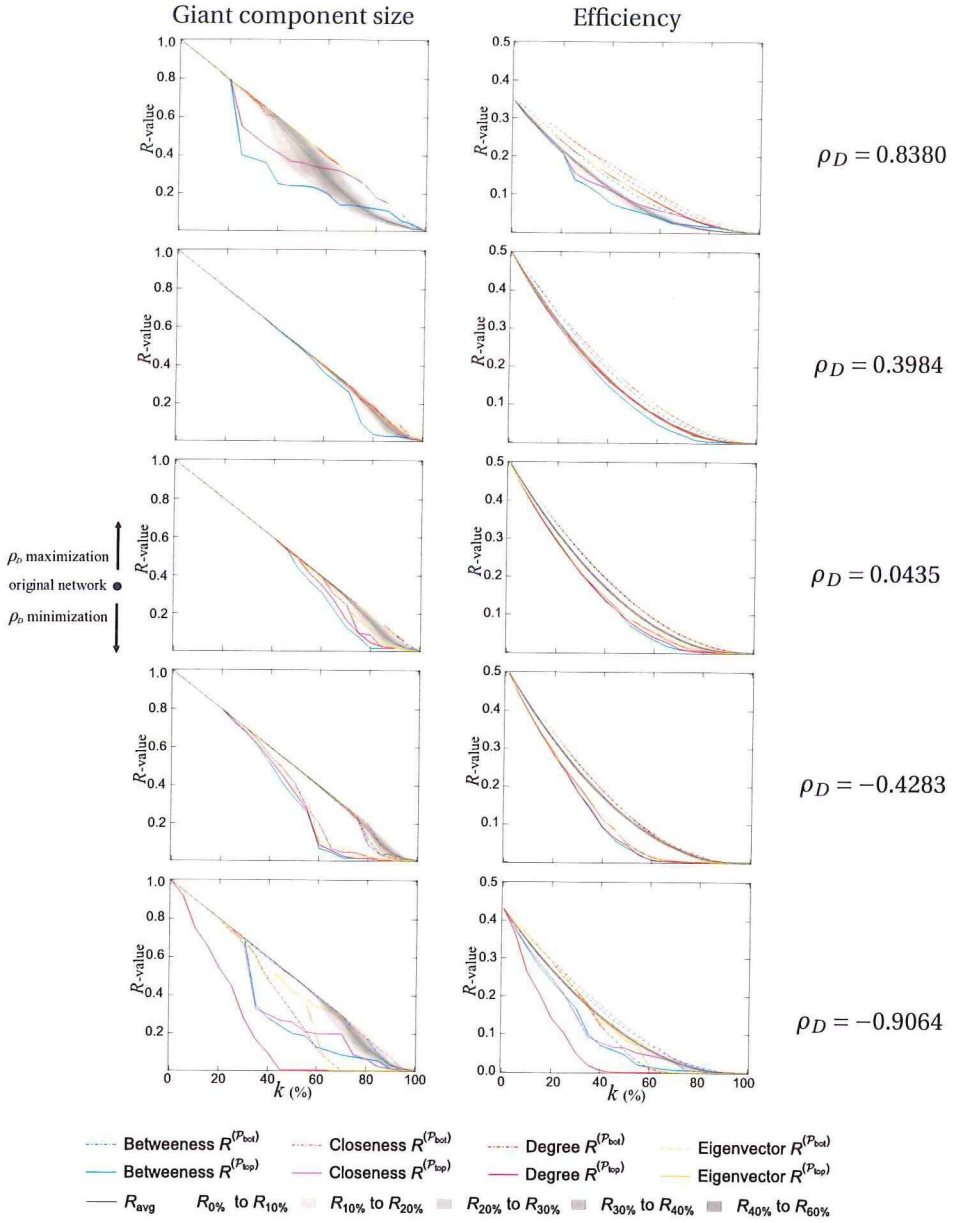


Figure 5.6: The influence of degree-preserving assortativity-optimization on the robustness of an Erdős-Rényi network. Robustness is measured relative to the giant component size (left) and the efficiency (right). In the first (top) row, a rewired network whose assortativity is fully maximized; in the second row, a rewired network whose assortativity is halfway between the fully maximized value and that of the original network; in the third (middle) row, the original network; in the fourth row, a rewired network whose assortativity is halfway between the fully minimized assortativity value and that of the original network; and in the fifth (bottom) row, a rewired network whose assortativity is fully minimized. The legend is the same as the ones in FIGS. 5.2, 5.3 and 5.4.

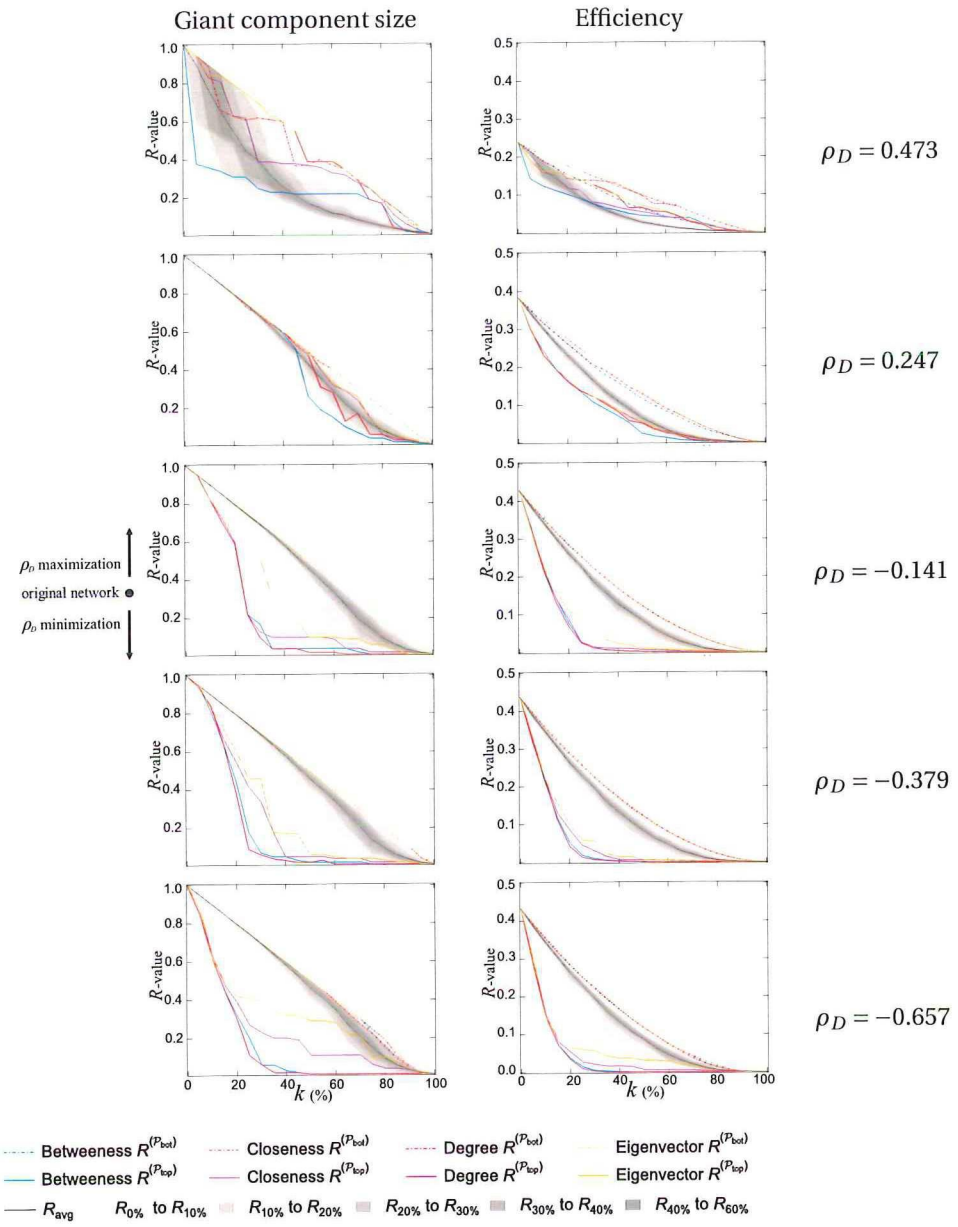


Figure 5.7: The influence of degree-preserving assortativity-optimization on the robustness of a Barabási-Albert graph. Robustness is measured relative to the giant component size (left hand) and the efficiency (right). In the first (top) row, a rewired network whose assortativity is fully maximized; in the second row, a rewired network whose assortativity is halfway between the fully maximized value and that of the original network; in the third (middle) row, the original network; in the fourth row, a rewired network whose assortativity is halfway between the fully minimized assortativity value and that of the original network; and in the fifth (bottom) row, a rewired network whose assortativity is fully minimized. The legend is the same as the ones in FIGS. 5.2, 5.3 and 5.4.

5.9. APPENDIX: ROBUSTNESS ENVELOPES OF BIOLOGICAL NETWORKS

In addition to the published material in this chapter, we performed additional experiments to bridge the purely topological approach to robustness discussed in this chapter and the biologically oriented approach discussed in the next chapter. To this end, robustness envelopes based on efficiency and giant component size were constructed for a number of metabolic networks considered in Chapter 6. The depth of the technical matter in this section is limited as all relevant concepts are thoroughly dealt with in § 6.3.

5.9.1. DATASET AND NETWORK CONSTRUCTION

The dataset used to construct networks is the genome-wide yeast metabolic model, described in § 6.3.4. A minimal metabolic model represents metabolites and reactions acting on metabolites. Ideally, a network representation models both metabolites and reactions. Since an arbitrary metabolite may be involved in many reactions and an arbitrary reaction may act on many metabolites, neither can be represented as a link; instead, both metabolites and reactions are modeled as nodes. A network containing both is called a *metabolite-reaction network* and is denoted G_B. Since networks containing multiple types of nodes are difficult to analyze, simplified networks containing only metabolites (so-called *metabolite networks*, denoted G_M) or only reactions (so-called *reaction networks*, denoted G_R) are also studied. In a metabolite network, metabolites that are connected by a reaction are linked whilst in a reaction network, reactions that are connected by a metabolite are linked. This simplification comes at the cost of detail. In particular, reaction networks are very dense (since a few highly connected metabolites lead to a high number of connections between reactions), rendering them less useful in the analysis of robustness of metabolic systems (since dense networks are robust). In this appendix, only metabolite-reaction and metabolite networks are considered.

Many metabolic reactions require the input of energy-carrying metabolites. A small set of commonly-occuring metabolites, known as *currency metabolites*, serve this role. A direct network rendering of a metabolic system includes such currency metabolites but since their main role is in energy provision rather than chemical transformation, we did not include them in our metabolic networks.

Metabolic systems are directional (that is, they have inputs and outputs), implying that their network representations ought also to be directed. However, the analyses earlier in this chapter used only undirected networks. To bridge the gap, both undirected and directed versions of metabolic networks were analyzed. This leads to a total of four networks that were analyzed; the networks, along with their basic topological properties are listed in Table 5.4. The discrepancy between the number of nodes in the directed and undirected versions of the metabolite-reaction network results from the fact that bidirectional reactions are represented by two nodes in the directed version whilst each reaction is represented by exactly one node in the undirected version. This aspect of modeling is more thorougly covered in § 6.3.2..

Network	N	L
Metabolite-reaction (directed)	2914	4347
Metabolite-reaction (undirected)	2368	3154
Metabolite (directed)	1102	2494
Metabolite (undirected)	1102	1789

Table 5.4: Metabolic networks derived from the iND750 dataset [47] and analyzed in this appendix.

5.9.2. Robustness-envelope analysis

As with all the other networks in this chapter, robustness envelopes based on efficiency and giant component size were computed on the four metabolic networks from Table 5.4. There are however two additional aspects to consider:

Metric calculation on directed networks Efficiency generalizes naturally to directed networks, as the metric on which it is based, hopcount, generalizes to directed networks. Giant component size can be defined in various ways in directed networks. Here, we are interested in the number of nodes reachable from any given starting node, as a crude proxy of the number of metabolites that a metabolic system can produce from a given input.

Nodes to remove from the metabolite-reaction network The metabolite-reaction networks are different from the other networks considered in this chapter in that they contain two types of nodes, metabolites and reactions. In other words, metabolite-reaction networks are bipartite and no two metabolites are connected, nor are any two reactions. Therefore, if all metabolites are removed, the network is completely disconnected; likewise for reactions. As in the next chapter, we consider removal of only metabolites and removal of only reactions but never removal of both types at the same time.

5.9.3. Results and Discussion

Robustness envelopes for the directed metabolic networks are shown in Figure 5.8 whilst envelopes for the undirected metabolite-reaction network are shown in Figure 5.9. In all cases, the giant component envelopes were normalized by dividing with the number of nodes in the network whilst the efficiency envelopes reflect the real efficiency values (since this metric falls in the range $[0,1]$, we opted to forgo normalization to facilitate comparison between the networks).

Figure 5.10 contains robustness envelopes of undirected metabolite network along with robustness envelopes of four rewired versions of the network (§ 5.7.4), analogous to Figure 5.6 and Figure 5.7.

The original metabolic networks (Figure 5.8, Figure 5.9 and the third row in Figure 5.10) are similar to (most of the) real-world networks (Figure 5.4) and Barabási-Albert networks (Figure 5.2 and Figure 5.3) in that they are robust against random failure but fragile against targeted attacks. This is demonstrated by fact that the solid colored lines representing targeted attacks fall below the black line and shaded gray regions repre-

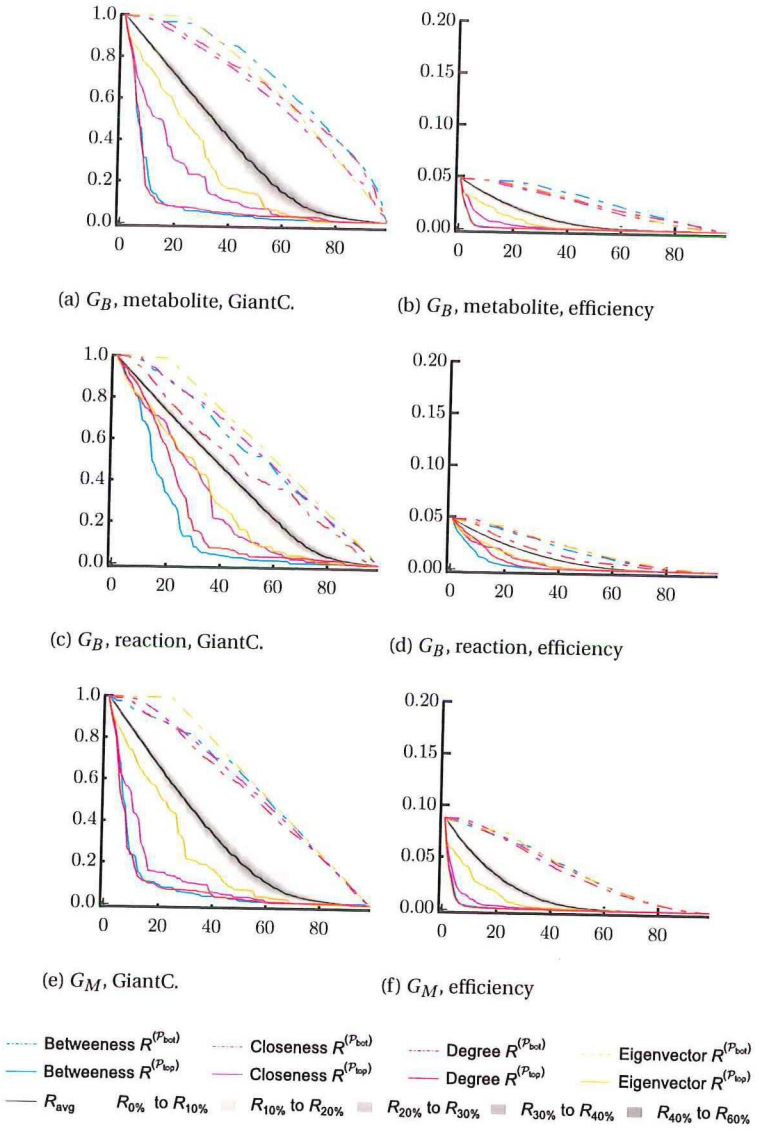
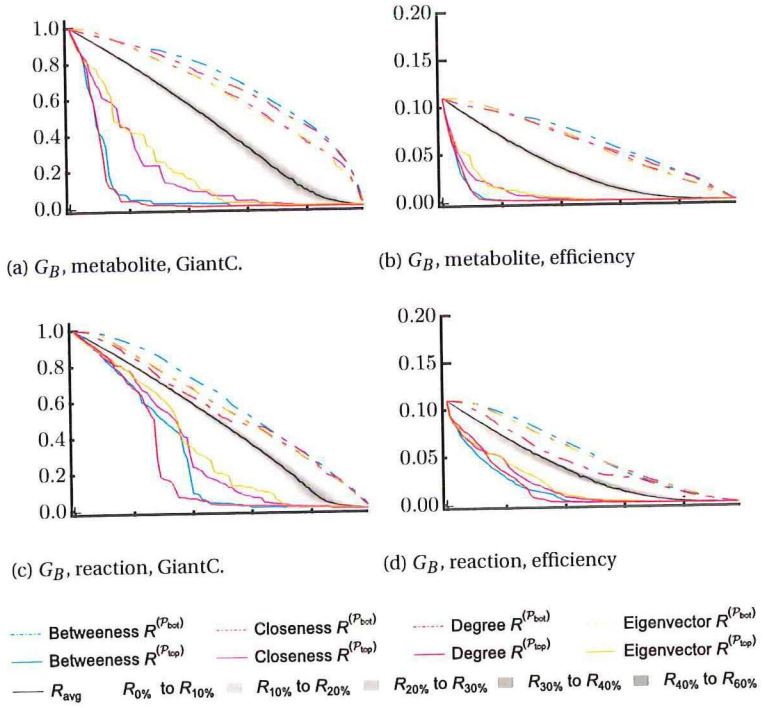


Figure 5.8: R-values for the directed metabolic networks. The network considered and the metric reflecting R-value are given in sub-captions (a)–(f). The x-axis is the percentage of removed nodes either at random ore according to a centrality measure as it is shown in the legend.



(a) G_B, metabolite, GiantC.

(b) G_B, metabolite, efficiency

(c) G_B, reaction, GiantC.

(d) G_B, reaction, efficiency

Figure 5.9: R-values for the undirected version of the metabolite-reaction network, G_B. The network considered and the metric reflecting R-value are given in sub-captions (a)–(d). The x-axis is the percentage of removed nodes either at random ore according to a centrality measure as it is shown in the legend.

senting the effects of random removal. Robustness envelopes for efficiency of metabolite networks (Figure 5.8, Figure 5.9 and Figure 5.10) have larger starting values than all real-world networks except the co-authorship network, CA, (Figure 5.4) whilst they have lower starting values than Barabási-Albert networks (Figure 5.7). The differences in network efficiency are explained by the fact that metabolic networks are more dense than the real-world networks (except for the co-authorship network) whilst they are less dense than the Barabási-Albert networks. On the whole, the robustness envelopes of the metabolic networks do not indicate that they are obviously more robust networks than the networks considered earlier in the chapter.

An interesting difference between the assortativity-optimized metabolite networks of Figure 5.10 have larger starting values than all real-world networks except the co-authorship network, CA, (Figure 5.4) whilst they have lower starting values than Barabási-Albert networks (Figure 5.7). The differences in network efficiency are explained by the fact that metabolic networks are more dense than the real-world networks (except for the co-authorship network) whilst they are less dense than the Barabási-Albert networks. On the whole, the robustness envelopes of the metabolic networks do not indicate that they are obviously more robust networks than the networks considered earlier in the chapter.

An interesting difference between the assortativity-optimized metabolite networks of Figure 5.10 and the assortativity-optimized networks of Figure 5.6 and Figure 5.7, is that, at least in terms of the two envelope metrics, the original network (that is, the network corresponding to the third row) is inferior to the rewired network with slightly decreased assortativity (fourth row). This suggests that metabolic systems are not necessarily optimized for shortest path lengths (something that was also shown in simulations by Arita [48]) nor for connectivity, suggesting that one needs to consider robustness relative to metrics that matter for biological function.

Although this robustness envelope study is limited in scope, there are no clear results suggesting that metabolite networks are topologically special. Of course, molecular networks evolved in contexts where their architectures contributed to the *biological function* of their host organisms and not in contexts where simple topological properties were selected for. This does not mean that topology does not play an important role, rather, it means that biological function should be tied to topology in order to discover the structures that influence biological function. This is the approach considered in the next chapter.

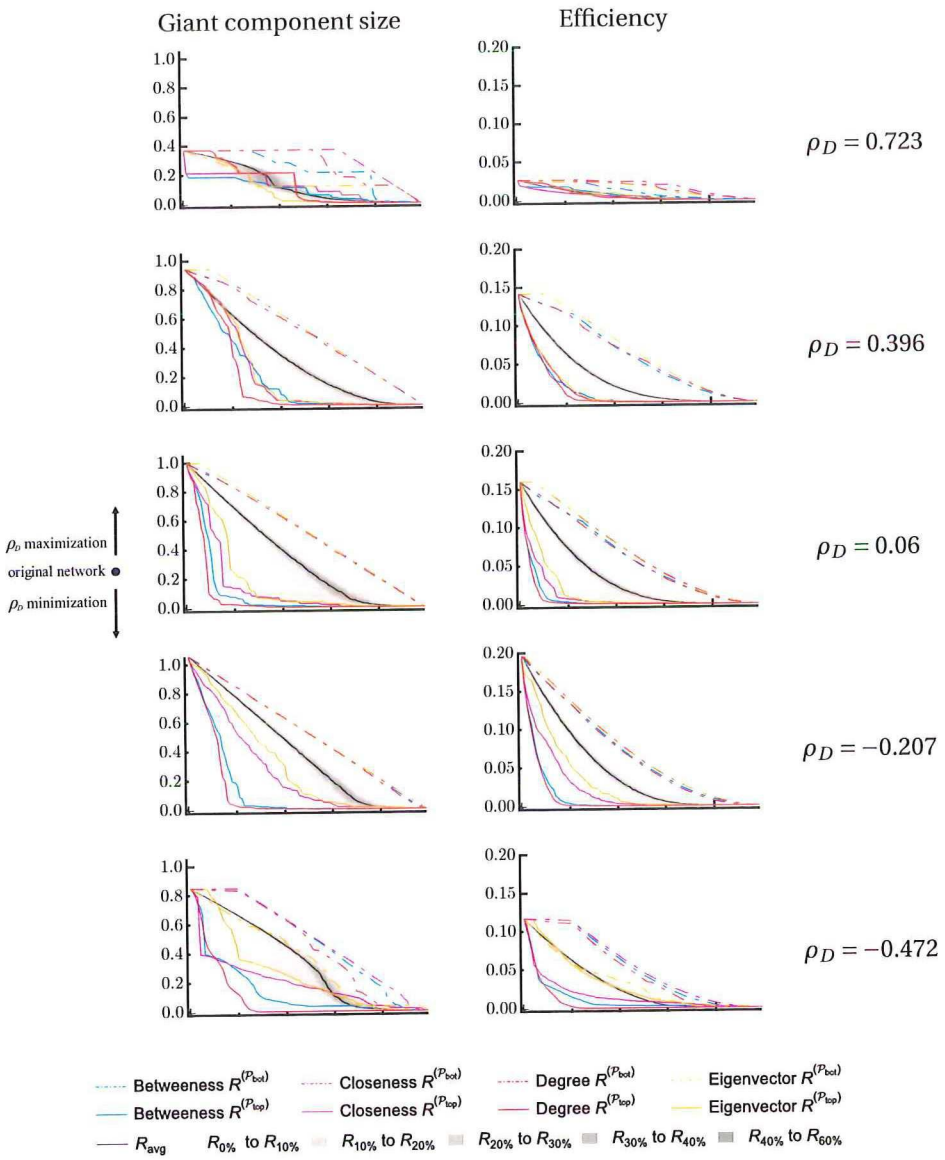


Figure 5.10: The influence of degree-preserving assortativity-optimization on the robustness of a metabolite network. Robustness is measured relative to the giant component size (left) and the efficiency (right). In the first (top) row is a rewired copy whose assortativity is fully maximized; in the second row, a rewired network whose assortativity is halfway between the fully maximized value and that of the original network; in the third (middle) row, the original network; in the fourth row, a rewired network whose assortativity is halfway between the fully minimized assortativity value and that of the original network; and in the fifth (bottom) row, a rewired network whose assortativity is fully minimized. The legend is the same as the ones in FIGS. 5.2, 5.3 and 5.4.

REFERENCES

[1] S. Trajanovski, J. Martín-Hernández, W. Winterbach, and P. Van Mieghem, *Robustness envelopes of networks*, Journal of Complex Networks **1**, 44 (2013), http://comnet.oxfordjournals.org/content/1/1/44.full.pdf+html.

[2] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, *Resilience of the internet to random breakdowns*, Phys. Rev. Lett. **85**, 4626 (2000).

[3] S. Scellato, I. Leontiadis, C. Mascolo, P. Basu, and M. Zafer, *Understanding robustness of mobile networks through temporal network measures*, in *Proceedings of INFOCOM* (IEEE, 2011).

[4] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, *Attack vulnerability of complex networks*, Phys. Rev. E **65** (2002), 10.1103/PhysRevE.65.056109, cond-mat/0202410.

[5] X. Huang, J. Gao, S. V. Buldyrev, S. Havlin, and H. E. Stanley, *Robustness of interdependent networks under targeted attack*, Phys. Rev. E **83**, 065101 (2011).

[6] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, *Breakdown of the internet under intentional attack*, Phys. Rev. Lett. **86**, 3682 (2001).

[7] A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr, *Basic concepts and taxonomy of dependable and secure computing*, IEEE Transactions on Dependable and Secure Computing **1**, 11 (2004).

[8] M. Menth, M. Duelli, R. Martin, and J. Milbrandt, *Resilience analysis of packet-switched communication networks*, IEEE/ACM Transactions on Networking **17**, 1950 (2009).

[9] P. Van Mieghem, C. Doerr, H. Wang, J. Martín-Hernández, D. Hutchison, M. Karaliopoulos, and R. E. Kooij, *A Framework for Computing Topological Network Robustness*, Tech. Rep. 20101218 (Networks Architectures and Services, Delft University of Technology, 2010).

[10] C. Doerr and J. Martín-Hernández, *A computational approach to multi-level analysis of network resilience*, in *Third International Conference on Dependability (DEPEND)* (2010).

[11] J. F. Meyer, *Performability: a retrospective and some pointers to the future*, Performance Evaluation **14**, 139 (1992).

[12] J. F. Meyer, *On evaluating the performability of degradable computing systems*, IEEE Transactions on Computers **C-29**, 720 (1980).

[13] P. Cholda, A. Mykkeltveit, B. Helvik, O. Wittner, and A. Jajszczyk, *A survey of resilience differentiation frameworks in communication networks*, Communications Surveys Tutorials, IEEE **9**, 32 (2007).

[14] A. Satyanarayana and A. Prabhakar, *New topological formula and rapid algorithm for reliability analysis of complex networks,* IEEE Transactions on Reliability **R-27**, 82 (1978).

[15] R. Wilkov, *Analysis and design of reliable computer networks,* IEEE Transactions on Communications **20**, 660 (1972).

[16] S. Rai and K. K. Aggarwal, *An efficient method for reliability evaluation of a general network,* IEEE Transactions on Reliability **R-27**, 206 (1978).

[17] H. L. Frisch and J. M. Hammersley, *Percolation processes and related topics,* SIAM Journal on Applied Mathematics **11**, 894 (1963).

[18] M. F. Sykes and J. W. Essam, *Exact Critical Percolation Probabilities for Site and Bond Problems in Two Dimensions,* Journal of Mathematical Physics **5**, 1117 (1964).

[19] L. B. Page and J. E. Perry, *Reliability polynomials and link importance in networks,* IEEE Transactions on Reliability **43**, 51 (1994).

[20] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts, *Network robustness and fragility: Percolation on random graphs,* Physical Review Letters **85**, 5468 (2000).

[21] V. Kostakos, *Temporal graphs,* Physica A **388**, 1007 (2009).

[22] J. Tang, M. Musolesi, C. Mascolo, and V. Latora, *Temporal Distance Metrics for Social Network Analysis,* in *Proceedings of WOSN '09* (Barcelona, Spain, 2009).

[23] S. Trajanovski, S. Scellato, and I. Leontiadis, *Error and attack vulnerability of temporal networks,* Phys. Rev. E **85**, 066105 (2012).

[24] C. M. Schneider, A. A. Moreira, J. S. Andrade, S. Havlin, and H. J. Herrmann, *Mitigation of malicious attacks on networks,* Proceedings of the National Academy of Sciences **108**, 3838 (2011).

[25] E. Çetinkaya, D. Broyles, A. Dandekar, S. Srinivasan, and J. Sterbenz, *A comprehensive framework to simulate network attacks and challenges,* in *International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)* (2010) pp. 538 –544.

[26] V. Latora and M. Marchiori, *Efficient Behavior of Small-World Networks,* Physical Review Letters **87** (2001).

[27] T. N. Dinh, Y. Xuan, M. T. Thai, P. M. Pardalos, and T. Znati, *On new approaches of assessing network vulnerability: Hardness and approximation,* IEEE/ACM Transactions on Networking **PP**, 1 (2011).

[28] L. C. Freeman, *Centrality in social networks conceptual clarification,* Social Networks **1**, 215 (1978-1979).

[29] J. Scott, *Social network analysis: a handbook* (SAGE Publications, 2000).

[30] P. Van Mieghem, *Graph Spectra for Complex Networks* (Cambridge University Press, 2011).

[31] P. Erdős and A. Rényi, *On the evolution of random graphs,* Publications of the Mathematical Institute of the Hungarian Academy of Sciences **5**, 17 (1960).

[32] E. N. Gilbert, *Random graphs,* Annals of Mathematical Statistics **30**, 1141 (1959).

[33] D. J. Watts and S. H. Strogatz, *Collective dynamics of small world networks,* Nature , 440 (1998).

[34] R. Albert and A.-L. Barabasi, *Statistical Mechanics of Complex Networks,* Review of Modern Physics **74**, 47 (2002).

[35] J. Leskovec, J. Kleinberg, and C. Faloutsos, *Graph evolution: Densification and shrinking diameters,* ACM Trans. Knowl. Discov. Data **1** (2007), http://doi.acm.org/10.1145/1217299.1217301.

[36] A. Jamakovic and S. Uhlig, *On the relationships between topological metrics in real-world networks,* Networks and Heterogeneous Media **3**, 345 (2008).

[37] P. Van Mieghem, *Performance Analysis of Communications Networks and Systems* (Cambridge University Press, UK, 2006).

[38] F. Chung and L. Lu, *The average distances in random graphs with given expected degrees,* Internet Mathematics **1**, 15879 (2002).

[39] R. Smythe and J. Wierman, *First-passage percolation on the square lattice,* Lecture Notes in Mathematics **671** (1978).

[40] C. Li, H. Wang, W. de Haan, C. J. Stam, and P. Van Mieghem, *The correlation of metrics in complex networks with applications in functional brain networks,* Journal of Statistical Mechanics: Theory and Experiment **2011**, P11018 (2011).

[41] M. E. J. Newman, *Mixing patterns in networks,* Phys. Rev. E **67**, 026126 (2003).

[42] P. Van Mieghem, H. Wang, X. Ge, S. Tang, and F. A. Kuipers, *Influence of assortativity and degree-preserving rewiring on the spectra of networks,* The European Physical Journal B **76**, 643 (2010).

[43] G. D'Agostino, A. Scala, V. Zlatić, and G. Caldarelli, *Robustness and assortativity for diffusion-like processes in scale-free networks,* EPL (Europhysics Letters) **97**, 68006 (2012).

[44] P. Van Mieghem, X. Ge, P. Schumm, S. Trajanovski, and H. Wang, *Spectral graph analysis of modularity and assortativity,* Phys. Rev. E **82**, 056113 (2010).

[45] W. Winterbach, D. de Ridder, H. Wang, M. Reinders, and P. Van Mieghem, *Do greedy assortativity optimization algorithms produce good results?* The European Physical Journal B **85**, 1 (2012).

[46] C. Friedel and R. Zimmer, *Influence of degree correlations on network structure and stability in protein-protein interaction networks*, BMC Bioinformatics **8**, 297+ (2007).

[47] palsson, *Organisms - systems biology research group*, (2009), `http://systemsbiology.ucsd.edu`.

[48] M. Arita, *The metabolic world of* escherichia coli *is not small*, Proceedings of the National Academy of Sciences of the USA **101**, 1543 (2004).

6

METABOLIC NETWORK DESTRUCTION: RELATING TOPOLOGY TO ROBUSTNESS

Wynand WINTERBACH, Huijuan WANG, Marcel REINDERS, Piet VAN MIEGHEM, Dick DE RIDDER

6.1. ABSTRACT

Biological networks exhibit intriguing topological properties such as small-worldness. In this paper, we investigate whether the topology of a particular type of biological network, a metabolic network, is related to its robustness. We do so by perturbing a metabolic system *in silico*, one reaction at a time and studying the correlations between growth, as predicted by flux balance analysis, and a number of topological metrics, as computed from three network representations of the metabolic system.

We find that a small number of metrics correlate with growth and that only one of the network representations stands out in terms of correlated metrics. The most correlated metrics point to the importance of hub nodes in this network, so-called "currency metabolites". Since they are responsible for interconnecting distant functional modules in the network, they are important points in the network for predicting if reaction removal affects growth. A second set of correlations in contrast is related to "loner" nodes

This chapter was published in Nano Communications Networks **2**, 2-3 (2011) [1].

that uniquely connect important pathways and thus correspond to essential steps in metabolism.

Source code and data are available upon request.

6.2. INTRODUCTION

In the last decade, advances in high-throughput biological measurement systems have made it possible to extract large-scale networks from biological systems. Jeong *et al.* [2] were among the first to study the topologies of metabolic networks, networks of interconversions of small compounds. The metabolic networks of the 43 organisms that they studied gave evidence of a scale-free structure. Characteristic properties of these so-called "small-world" networks are their power-law distributed node degrees and their small average shortest path lengths.

Subsequently, researchers studied the topologies of a number of other types of biological networks [3–5]. Much of this work confirmed the Jeong *et al.* results: scale-free behavior was everywhere. Even the Internet and some power grids are thought to display scale-free behavior [6]. These latter networks have expanded in a seemingly organic fashion through a process of *preferential attachment* – new nodes are more likely to attach to existing high-degree nodes than to low-degree nodes. This expansion process forms the basis of Barabási and Albert's [6] random network model. They show that it leads to the characteristic power-law node-degree distribution and small-world properties. Although Kim *et al.* [7] and Lima-Mendez *et al.* [8] argue that biological networks do not develop through simple processes of preferential attachment, the presence of similar topological elements, such as hub nodes, begs the question whether these topological properties confer some benefit or whether certain topologies are inherently suited for particular functionality.

In an effort to understand the relationship between the function of a network and its topological properties, Milo *et al.* [9] introduced the concept of *motifs*. A motif is a small sub-network (3–5 nodes) whose over-representation may be indicative of its role in maintaining function at a local level. They found that certain motifs occur more often in biological networks than expected by chance and that they may correspond to certain desired behavior such as response acceleration, signal delay and stability. Prill *et al.* [10] took this idea further and claimed that certain motifs were inherently more prone to display stable behavior than others. By abstracting away from the underlying functionality, they demonstrated that such relations held to some extent over a variety of biological networks. However, Ingram *et al.* [11] considered gene networks and compared the results of a differential equation model of gene expression to specific motif counts in the gene network but found no correlation. Lima-Mendez *et al.* [8] argue that global topological properties cannot explain the function of networks. While they claim that the significance of motif frequencies may have been overestimated (since the frequencies only capture global properties), they do consider a localized approach to be more promising as the key to understanding biological networks lies in understanding local details.

In our work, we take a global approach and investigate to what extent network topology can be related to more systems-level network properties shared by the various network types studied by Barabási *et al.* An interesting property in this respect is that of *robustness*. Stelling *et al.* [12] and Kitano [13] define robustness as the ability of a system

to maintain its function in the face of perturbations or uncertainty. Biological systems are known to be robust [14] to many forms of perturbation while being highly sensitive to other forms, so-called "highly optimized tolerance" [12]. The question is whether there is something in the topology of these networks that confers robustness to the overall system.

In this paper, we study the relationship between the robustness of a micro-organism (baker's yeast, *Saccharomyces cerevisiae*) and the topologies of network representations of its metabolic system. Microbial metabolic systems provide a good test bed, since an often assumed functional objective – growth – is easily expressed in terms of fluxes through these systems. Furthermore, good quality metabolic datasets are readily available and resulting flux models can be studied computationally with high efficacy.

To study the link between network topology and robustness, we propose an *in silico* metabolic system perturbation experiment. We define robustness as the ability of the yeast cell to maintain growth under reaction removals. First, we show how its metabolic system can be represented by three different networks. Then, through a number of trials, reactions are removed from the metabolic system until growth ceases. This provides a number of snapshots of partially "destructed" metabolic systems. For each snapshot, growth and a number of network-wide topological metrics can be computed. By calculating correlations between growth and these metrics, we find that most of the topological metrics are not related to function. The strongest correlations point to the importance of both "hub" nodes (so-called "currency metabolites") and "loner" nodes.

6.3. METHOD

6.3.1. COMPUTING FUNCTION

In this work, we define robustness as the maintenance of cell growth under perturbations to the organism's metabolic system when reactions are removed from the metabolic network. A metabolic system with r reactions and m metabolites is modeled by a set of m differential equations:

$$\frac{\mathrm{d}X_i}{\mathrm{d}t} = s_{syn}v_{syn} - s_{deg}v_{deg} - s_{use}v_{use} + s_{trans}v_{trans} \tag{6.1}$$

that specify how the concentration X_i of a metabolite i changes in time. v_{syn} is the rate of metabolite synthesis, v_{deg} is the degradation rate, v_{use} is the rate of consumption (by other reactions) and v_{trans} is the rate of transport across the cell boundary (into the cell). v_{syn}, v_{deg} and v_{use} are generally non-linear functions whose behavior is governed by the kinetic parameters of the enzymes catalyzing the reactions in which they take part and by concentrations of other metabolites. Because the kinetic parameters are not generally known and must be estimated, it is difficult to solve the differential equations directly. s_{syn}, s_{deg}, s_{use} and s_{trans} are *stoichiometric coefficients*[1] (reaction rates are measured in μmol gDW^{-1} h, i.e., micromoles per gram of dry weight per hour).

We assume that $s_{trans}v_{trans}$ is a constant value b_i, allowing (6.1) to be written in vector form as $\mathrm{d}\mathbf{X}/\mathrm{d}t = \mathbf{S}\cdot\mathbf{v}+\mathbf{b}$, with $\mathbf{S}$ the $m \times r$ *stoichiometric matrix*, $\mathbf{v}$ an $r \times 1$ vector of

[1] These are derived from the chemical mass balance coefficients: e.g. $2H_2 + O_2 \rightarrow 2H_2O$ corresponds to the stoichiometric coefficient vector $[-2\ \ -1\ \ 2]$.

$$\mathbf{S} = \begin{array}{c} \\ m_1 \\ m_2 \\ m_3 \\ m_4 \\ m_5 \\ m_6 \\ m_7 \\ m_8 \end{array} \begin{array}{c} \begin{array}{ccc} R_a & R_b & R_c \end{array} \\ \begin{pmatrix} -1 & 0 & 0 \\ -1 & 0 & 0 \\ 1 & -3 & -1 \\ 1 & 0 & -1 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \end{pmatrix} \end{array}$$

(a) The stoichiometric matrix from (6.4).

$$\mathbf{S}' = \begin{array}{c} \\ m_1 \\ m_2 \\ m_3 \\ m_4 \\ m_5 \\ m_6 \\ m_7 \\ m_8 \end{array} \begin{array}{c} \begin{array}{cccc} R_a^+ & R_a^- & R_b & R_c \end{array} \\ \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 1 & -1 & -3 & -1 \\ 1 & -1 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

(b) The stoichiometric matrix from (6.6).

Figure 6.1: Stoichiometric matrices of the toy problem in Section 6.3.1.

reaction rates (fluxes) and **b** the vector of boundary transport reaction rates. We will use a small example to make the form of **S** clear (and later to show how networks are derived from **S**). Consider the metabolic system:

$$\begin{aligned} m_1 + m_2 &\xleftrightarrow{R_a} m_3 + m_4 \\ 3m_3 + m_5 &\xrightarrow{R_b} m_6 + 2m_7 \\ m_3 + m_4 &\xrightarrow{R_c} m_7 + m_8 \end{aligned} \tag{6.2}$$

The corresponding **S** matrix is shown in Figure 6.1a. Since each column is labeled by a reaction R_i, we refer to the corresponding flux value in **v** as v_i. At steady-state $\mathrm{d}\mathbf{X}/\mathrm{d}t = \mathbf{0}$, rendering the linear system:

$$\mathbf{S} \cdot \mathbf{v} + \mathbf{b} = \mathbf{0}. \tag{6.3}$$

Since **S** and **b** are constant, **v** can be determined without any knowledge of enzyme kinetics (in flux balance analysis, the unknowns are reaction rates rather than metabolite concentrations). Due to the small size of the example, **S** is overdetermined (i.e., there are fewer reactions than metabolites; opposite of much of systems biology, in flux balance analysis reaction rates are unknown rather than metabolite concentations). In real biological networks however, stoichiometric matrices are under-determined. Such systems generally have infinitely many solutions but biologists are only interested in biologically significant ones. A common (biological) assumption is that microbial cells attempt to maximize the rate of their biomass production or in other words, growth. Growth can be expressed as a linear combination $\mathbf{c}^T \cdot \mathbf{v}$ of certain key reaction rates in the metabolic system. The reaction rates can then be computed by a linear program:

$$\begin{aligned} &\text{Maximize} && \mu = \mathbf{c}^T \cdot \mathbf{v} \\ &\text{subject to} && \mathbf{S} \cdot \mathbf{v} + \mathbf{b} = \mathbf{0} \end{aligned} \tag{6.4}$$

Positive components of **v** correspond to forward-acting reactions, whilst negative components correspond to reactions running in reverse. In (6.4), the components of

v may assume negative and positive values meaning that any reaction can, in principle, occur in either direction. Due to thermodynamics, some reactions are very unlikely to occur in reverse (in the example, only reaction R_a is reversible). These constraints are modeled by restricting rates of non-reversible reactions to be non-negative. Thus for each non-reversible reaction R, the constraint $v_R \geq 0$ is added, rendering the linear system:

$$\begin{aligned} &\text{Maximize} && \mu = \mathbf{c}^T \cdot \mathbf{v} \\ &\text{subject to} && \mathbf{S} \cdot \mathbf{v} + \mathbf{b} = \mathbf{0} \\ & && v_{R_i} \geq 0 \text{ for each non-reversible reaction } R_i \end{aligned} \tag{6.5}$$

In addition, biological constraints limit the rates of some reactions. These inequalities are simply added to the list of constraints of the linear program. This steady-state framework for computing metabolic fluxes by optimizing some criterion is known as *flux balance analysis*. Orth *et al.* [15] give a good overview of the framework.

TESTING ROBUSTNESS

We test robustness by iteratively removing reactions and recalculating (6.5) until growth μ drops below a low threshold value (1×10^{-9} μmol gDW^{-1} h). This produces a sequence $T = \{s_1, s_1, s_2, \ldots, s_n\}$ which is referred to as the *trial* T. A *step* is a reaction label index: step s_i corresponds to the removal of reaction R_{s_i}. Removal of a reaction is modeled by removing its corresponding column from **S**. The steps in a trial are associated with a sequence of linear programs $P_0, P_1, P_2, \ldots, P_n$, where P_0 is the unmodified linear program (from which no reaction has been removed) and P_i is the linear program resulting from the removal of the reactions $R_{s_1}, R_{s_2}, \ldots, R_{s_i}$ for $i \geq 1$.

Pseudo-code for the algorithm is shown in Algorithm 1. This algorithm computes the results for one trial. The input is a description of the metabolic system σ and a network metric (that takes a network as input and produces an output of type $\mathcal{O}$). The i-th iteration of the loop corresponds to step s_i.

The function "random-reaction" in Algorithm 1 chooses a random enzyme-catalyzed reaction with uniform probability. Reactions that are not mediated by enzymes but occur due to chemical processes such as diffusion are never removed.

6.3.2. TOPOLOGY

To be able to calculate topological properties of the metabolic system, the stoichiometric matrix **S** should be represented as a network. However, **S** cannot be directly represented as a network since a reaction may interact with more than two metabolites and a metabolite may interact with more than two reactions. A natural representation of such a system is a *hyper-network* in which a link may connect more than two nodes. The stoichiometric matrix represents a hyper-network where the columns are links and the rows are nodes. The links are directed: negative values in a column represent source nodes and positive values represent target nodes. Let u be a node, and let L be a set of links that have u as their source nodes, then the target nodes of L are the out-neighbors of u. The in-neighbors are defined analogously, with u as the target node.

Algorithm 1 destruction-trial(σ : metabolic system, metric : network $\rightarrow \mathcal{O}$)

```
X ← empty-list() {List of growth values}
M ← empty-list() {List of metric values}
P ← to-linear-program(σ) {Compute P_o}
μ ← growth-rate(P)
while μ > 1 × 10^-9 do {One step s_i in the current trial}
   R ← random-reaction(σ) {Pick R_{s_i}}
   σ ← remove-reaction(σ, R)
   P ← to-linear-program(σ) {Compute P_i}
   μ ← growth-rate(P)
   g ← network(σ)
   m ← metric(g)
   X ← append-to-list(X, μ)
   M ← append-to-list(M, m)
end while
return X, M
```

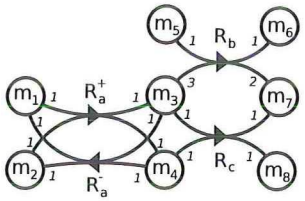
Note that the stoichiometric matrix derived from the linear programming formulation does not capture the reversibility of reactions (such as R_a in the example) because a reaction R_i is considered to act in reverse when its rate v_i in the linear program solution is negative. We therefore reformulate the linear program such that $\mathbf{v} \geq \mathbf{0}$ (i.e., all fluxes are positive). A reversible reaction R_i is converted to a pair of reactions R_i^+ and R_i^-; then if $\mathbf{c}_i$ is the column vector in $\mathbf{S}$ corresponding to R_i, $\mathbf{c}_i$ is replaced by two column vectors $\mathbf{c}_i^+$ and $\mathbf{c}_i^-$ (corresponding to R_i^+ and R_i^- respectively) such that $\mathbf{c}_i^+ = \mathbf{c}_i$ (the forward reaction) and $\mathbf{c}_i^- = -\mathbf{c}_i$ (the reverse reaction); for example, column R_a in Figure 6.1a is replaced by the columns R_a^+ and R_a^- in Figure 6.1a. Converting $\mathbf{S}$ leads to the stoichiometric matrix $\mathbf{S}'$ in Figure 6.1b. The hyper-network is shown in Figure 6.2a.

The linear program (6.5) is modified with the new stoichiometric matrix $\mathbf{S}'$ and non-negative flux constraints, giving:

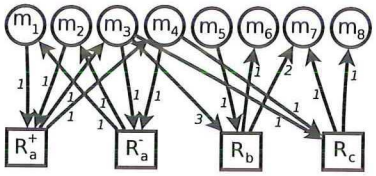
$$\begin{aligned} &\text{Maximize} && \mu = \mathbf{c}^T \cdot \mathbf{v} \\ &\text{subject to} && \mathbf{S}' \cdot \mathbf{v} + \mathbf{b} = \mathbf{0} \\ & && \mathbf{v} \geq \mathbf{0} \end{aligned} \tag{6.6}$$

Network theory provides many tools for studying the topological properties of normal networks, whilst there are very few metrics that can be computed on hyper-networks. Thus we considered three possible network representations of the hyper-networks specified by the stoichiometric matrix $\mathbf{S}'$. First, a hyper-network $H(\mathcal{M}, \mathcal{R})$ can be modeled as a bipartite network $G_B(\mathcal{M} \cup \mathcal{R}, \mathcal{L})$. The nodes in the set $\mathcal{M}$ represent the metabolites in H, whilst the nodes in the set $\mathcal{R}$ represent reaction links in H. Conversion of the hyper-network H in Figure 6.2a produces the biparitite network G_B in Figure 6.2b. We refer to this network as the *metabolite-reaction network*[2] as it contains both metabolite nodes

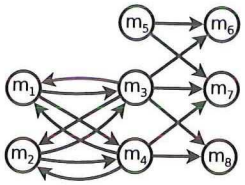
[2] This representation is the Petri-net representation [16, 17] of the metabolic system.



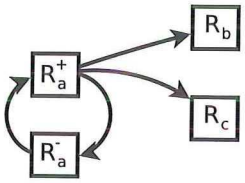
(a) The hyper-network H specified by $\mathbf{S}'$.



(b) G_B: the bipartite representation of H.



(c) G_M: the one-mode reduction of the metabolite nodes in G_B.



(d) G_R: the one-mode reduction of the reaction nodes in G_B.

Figure 6.2: The hyper-network and networks derivable from $\mathbf{S}'$ in (6.6).

$\mathscr{M}$ and reaction nodes $\mathscr{R}$.

Although standard network theory techniques can be applied to G_B, its bipartite nature makes some metrics difficult or impossible to compute. For example, the clustering coefficient for any node in a bipartite network is 0. For this reason, we also considered one-mode reductions of G_B. An $\mathscr{M}$-node ($\mathscr{R}$-node) *one-mode* reduction $G'(\mathscr{N}, \mathscr{L}')$ of $G_B(\mathscr{M} \cup \mathscr{R}, \mathscr{L})$ is a network that contains only nodes from the set $\mathscr{M}$ (the set $\mathscr{R}$) such that for each directed link $l = (n_1, n_2) \in \mathscr{L}'$ there is a node $n_3 \in \mathscr{R}$ ($n_3 \in \mathscr{M}$) such that $(n_1, n_3) \in \mathscr{L}$ and $(n_3, n_2) \in \mathscr{L}$ (note that there may be many nodes n_3 that satisfy this condition). We call the $\mathscr{M}$-node one-mode reduction simply the *metabolite network* G_M (shown in Figure 6.2c) and likewise the $\mathscr{R}$-node one-mode reduction simply the *reaction network* G_R (illustrated in Figure 6.2d).

Note that it is possible to represent the link weights of the hyper-network H in its bipartite representation G_B: such a mapping can be seen in Figure 6.2b. However, there is no obvious way to map these weights to G_M or G_R. For this paper, we opted to consider only unweighted networks. Furthermore, note that when a reaction is removed from the metabolic system, the corresponding networks G_B, G_M and G_R may become disconnected. For a given network, all metrics are applied to the largest component whilst the small components are ignored.

TOPOLOGICAL METRICS

For every step of each trial, a number of topological metrics were computed for each of the three network representations (where possible). Since $G_B = G_B(\mathscr{M} \cup \mathscr{R}, \mathscr{L})$ contains two types of nodes, the metrics are applied separately to its reaction nodes $\mathscr{R}$ and metabolite nodes $\mathscr{M}$, giving two sets of results.

The metrics employed are listed in Table 6.1. These metrics divide into two groups: those that associate a value $c(G)$ with a network G and those that associate values $\{c(n_1), c(n_2),$

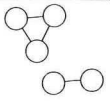
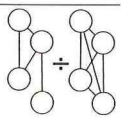
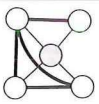
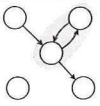
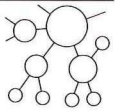
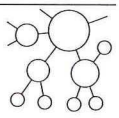
Newman's assortativity coefficient [18]	scalar
How likely is it for nodes with similar degrees to be connected to each other. Calculated for the out-degrees, in-degrees and undirected degrees of nodes, it is computed as $r = \sum_{ij} d_i d_j (b_{ij} - a_i a_j)/\sigma_a^2$, where a_i is the distribution of the degree of node n_i without the link $\{i, j\}$, b_{ij} is the joint probability distribution of the degrees between n_i and n_j without the link $\{i, j\}$ and σ_a^2 is the variance of degrees.	
Transitivity	scalar
The number of triangles in the network divided by the maximum possible number of triangles in the network. This is computed on undirected versions of the networks.	
Clustering coefficient	node
For a node n, the number of links spanning n's neighbors divided by the maximum possible number of links that can span n's neighbors. The mean clustering coefficient is equal to the transitivity.	
Reciprocity	scalar
The ratio of reciprocal pairs to all possible reciprocal pairs. A pair of nodes n_1 and n_2 is reciprocal if there are bi-directed links (n_1, n_2) and (n_2, n_1).	
Betweenness centrality	node
In a network G, the betweenness centrality $C_B(n)$ of a node n is computed as the fraction of shortest paths in G that include n: $C_B(n) = \sum_{s \neq n \neq t \in \mathcal{N}} \sigma_{s,t}(n)/\sigma_{s,t}$, where $\mathcal{N}$ is the node-set of G, $\sigma_{s,t}(n)$ is the number of shortest paths that include n and $\sigma_{s,t}$ is the total number of shortest paths between the nodes s and t.	
Eigenvector centrality & Largest eigenvalue of the adjacency matrix	node
For a graph G, the largest eigenvalue is that of the corresponding adjacency matrix $\mathbf{A}$. The eigenvector centrality of a node n_i is defined as the i-th component of the eigenvector corresponding to that eigenvalue.	

Table 6.1: A list of the various network metrics that were calculated on the networks G_B, G_M and G_R. Metrics that are calculated for a network as a whole are marked "scalar" whilst those that are calculated for every node are marked "node".

$\ldots, c(n_N)\}$ with the nodes $\mathcal{N} = \{n_1, n_2, \ldots, n_N\}$ of G. In order to compare this latter group of metrics to growth values, the node values (for a given metric c) have to be reduced to a single value $c^*(G) = f(c(n_1), c(n_2), \ldots, c(n_N))$ (where f is function of N arguments that produces a single real value $c^*(G) \in \mathbb{R}$). A simple choice is to let f compute the minimum, mean or maximum values of $\{c(n_1), c(n_2), \ldots, c(n_N)\}$ (thereby yielding three metrics). This is the approach that we took. Some metrics associate vectors of values with each node; thus, if the metric $\mathbf{c}$ associates a vector with a node, the result will be a set of vectors $\{\mathbf{c}(n_1), \mathbf{c}(n_2), \ldots, \mathbf{c}(n_N)\}$). The hop-count is such a metric, since it associates a vector of hop-count values $\mathbf{c}(n)$ with a node n containing the hop-counts to all other nodes in the network. We took the approach of first reducing the vectors to scalars – thus we converted $\{\mathbf{c}(n_1), \mathbf{c}(n_2), \ldots, \mathbf{c}(n_N)\}$ to $\{c'(n_1), c'(n_2), \ldots, c'(n_N)\}$ where c' is a function that

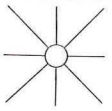
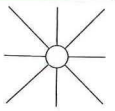
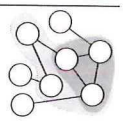
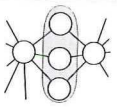
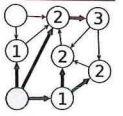
In- and out-degrees	node
The out-degree d_i^{out} of a node is the number of links leaving a node, i.e., $d_i^{\text{out}} = \sum_{(n_i,n_j)\in L(G)} 1$. Likewise, the in-degree d_i^{in} is the number of links entering a node, i.e., $d_i^{\text{in}} = \sum_{(n_j,n_i)\in L(G)} 1$.	
Average in- and out-degrees of incoming and outgoing neighbors	node
For a network G, the average out-degree of out-going neighbors of a node n_i is $\sum_{(n_i,n_j)\in L(G)} d_j^{\text{out}}/d_i$ whilst the average in-degree of in-coming neighbors is $\sum_{(n_j,n_i)\in L(G)} d_j^{\text{in}}/d_i$ where d_i^{out} is the out-degree of n_i and d_i^{in} is the in-degree of n_i.	
Coreness	node
A k-core is a subset of nodes in which each node has a degree of at least k. A node has a coreness value of c if it is in a c-core but not in a $c+1$-core.	
Dice similarity	node
If the neighbors of two nodes are the sets X and Y, the Dice similarity of the nodes is $2\lvert X \cap Y\rvert/(\lvert X\rvert+\lvert Y\rvert)$, i.e., a measure of how similar their neighbor sets are. Since this metric is defined for pairs of nodes, a vector of metrics is associated with each node. We compute the Dice similarity for all outgoing neighbors, all incoming neighbors and also for the combination of these.	
Reciprocal node hop-count	node
The hop-count between a pair of nodes is equal to the number of links on a shortest path between them. For each node there is a vector of hop-counts to all other nodes, reduced to a single value by taking the mean. Because the networks are directed, there are nodes which are unreachable from other nodes and are thus at an infinite distance. We therefore used reciprocal hop-count values, converting infinite distances to zero distances.	

Table 6.1: (continued) A list of the various network metrics that were applied to the networks G_B, G_M and G_R. Metrics that are calculated for a network as a whole are marked "scalar" whilst those that are calculated for every node are marked "node".

reduces vectors to real values. As above, we performed the reductions by computing the minima, maxima and means of the vectors. Once this initial reduction is performed, we can proceed as before (by reducing the sets of node values to single values). Note that this double reduction scheme can lead to confusing metric names. To take the example of the hop-count again, we could proceed by first computing the means of the hop-count vectors associated with each node and then we could compute the minimum over these mean values. In this case, we would refer to the minimum of the mean hop-count, or in the naming convention used in the results section, "mean hop-count $\triangledown$". Likewise, we refer to the mean of the mean hop-count as "mean hop-count $\square$" and the maximum of the mean hop-count as "mean hop-count $\triangle$".

In our experiments, many reactions have zero reaction rates (as predicted by the flux balance linear program) in all trials. These reactions contribute links and nodes to the network representations whilst their removal cannot influence growth. We excluded these reactions when constructing $G_B = G_B(\mathcal{M} \cup \mathcal{R}, \mathcal{L})$ by letting $\mathcal{R}$ be the set of all reactions that have non-zero reaction rates in at least one step of one trial and $\mathcal{M}$ the metabolites that interact with the reactions in $\mathcal{R}$. Note that this is only a global pre-

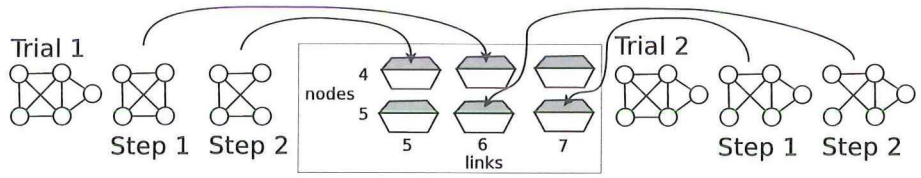


Figure 6.3: An example of binning for two steps of two trials.

processing step; in each individual trial, reactions are randomly chosen without regard to whether they are active at that time or not.

6.3.3. RELATING GROWTH AND TOPOLOGY

For each trial (i.e., sequence of reaction removals) we compute a sequence of growth values (computed from the linear program discussed in Section 6.3.1) and three sequences of networks, one for each representation. For each network, a set of topological metrics is calculated. This allows us to relate growth to topology.

An obvious first choice for calculating the relationship is, for each individual trial, to compute correlation coefficients ρ between the growth sequence and each of the sequences of topological metrics. However, apparent correlations found by this method may simply be side-effects of the network size decreasing as we remove reactions. We can reduce the impact of this incidental correlation by binning the steps from all of the trials: trial-step pairs whose corresponding networks have similar numbers of nodes and links are placed into the same bin. This process is illustrated in Figure 6.3: here one sees network sequences from two trials placed into bins (the bin width here is 1 for both nodes and links). In our experiments, we used a bin width of 2 nodes × 4 links – i.e., in a bin, node counts can differ by 1 and link counts by 3.

Since a bin contains numerous steps, it is possible to correlate growth with any of the topological metrics. We used the Pearson correlation coefficient to compute, for a given topological metric, a correlation value ρ_i for every bin i. An example of bin correlations is shown in Table 6.2 (here binning is only performed using link counts, with a bin width of 4 links). The per-bin results for each metric were then averaged, weighted by the number n_i of items in each bin: $\bar{\rho} = \left(\sum_i n_i \rho_i\right) / \left(\sum_i n_i\right) = 0.27$ in our example. For each topological metric, this yields one value $\bar{\rho}$ indicating the strength of its binned correlation with growth.

For all of the metrics that we studied, there were one or more bins for which correlations could not be computed, since the growth and/or metric values in the bin were constant. In this case, the Pearson correlation coefficient is not defined. These bins were excluded from the calculation of $\bar{\rho}$. We also required correlations to be:

- *reliable*, i.e., calculated on a sufficient number of data-points, by demanding that at least 90% of all steps fall in bins on which correlations are defined; and
- *consistent*, by requiring that at least 90% of all steps fall in bins whose correlations have the same sign.

Metrics that did not pass this test were not considered.

Bin number	1	2	3	4	5
# links	2685-2688	2689-2692	2693-2696	2697-2700	2701-2704
Correlation ρ_i	0.342	0.286	0.322	0.236	0.172
# items in bin n_i	889	935	907	959	936

Table 6.2: Bins showing Pearson correlations ρ between growth and an unspecified network metric.

6.3.4. EXPERIMENTAL SETUP

We used the genome-scale metabolic data set which is available from the UCSD Systems Biology Research Group website [19]. The website provides a minimal aerobic growth environment which was used for our experiments. In this experiment,

- the rate of the ATP maintenance reaction (ATPM) is $1\,\mu\text{mol}\,\text{gDW}^{-1}\,\text{h}$ whilst the acetyl-CoA hydrolase (ACOAH) and the glutamate synthase for NADH (GLUSx) reactions are disabled;
- the reaction rates of reactions that transport O_2, NH_4^+, SO_4^{2-}, P_i, H_2O, K, Na and CO_2 are unconstrained.

6.4. RESULTS AND DISCUSSION

6.4.1. METRICS CORRELATE WITH NETWORK SIZE

We initially performed one thousand *in silico* reaction removal trials and for each trial computed the Pearson correlation ρ between the growth values of the trial and the metrics in Table 6.1 as computed on G_B, G_M and G_R (where applicable). The average metric correlations over 200 random trials for G_M are shown in Figure 6.4 (here, we have only aggregated node-wise metrics using the mean, as described in Section 6.3.2). Many metrics stand out as strongly correlated.

We found that most of these correlations are due to the reduction of the number of nodes and/or links in G_B, G_M and G_R associated with each step in a destruction trial of a metabolic system. This growth-size relationship confounds the search for metrics that correlate with growth, since any apparent correlation ρ may be due solely to the correlation between the metric and the number of nodes/links in the network.

Removal of this effect by metric normalization is non-trivial, since the relationship between a given metric and the number of nodes/links in a network is, in general, non-linear. Furthermore, any technique that reduces this effect, must use topological information; but then this information itself is affected by the changing topology. We therefore devised a "binning" procedure to calculate alternative correlation measures $\bar{\rho}$ in which this effect is reduced (as described in Section 6.3.3). In the remainder, all results reported employ this binned correlation measure.

6.4.2. TOPOLOGY IS WEAKLY CORRELATED WITH FUNCTION

Next, we calculated correlations $\bar{\rho}$ (using the binning procedure) between growth and each metric. The results for $G_B = G_B(\mathscr{M} \cup \mathscr{R}, \mathscr{L})$ are shown in Figure 6.5 (recall that there are two sets of results for G_B: one for the metabolite nodes $\mathscr{M}$ and one for the reaction nodes $\mathscr{R}$) whilst the correlations for G_M are shown in Figure 6.6. There are no

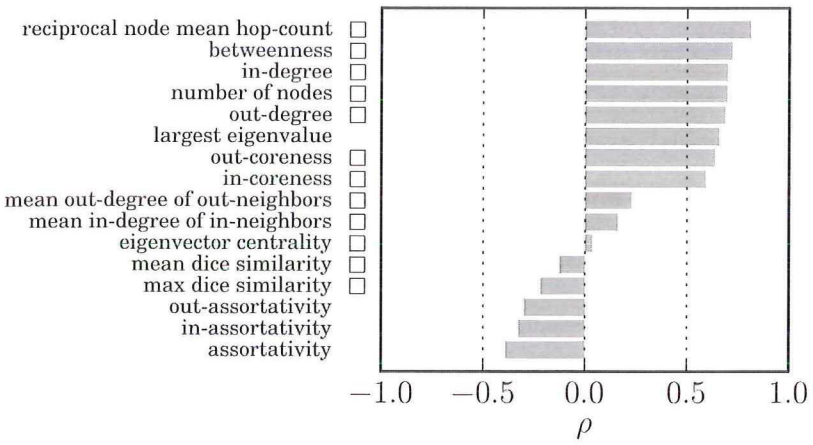


Figure 6.4: Network metrics are correlated with network size. This gives the appearance of strong correlations between growth and metrics. These metrics were all calculated for G_M. The symbol □ indicates that node values were reduced to single values by computing their means as discussed in Section 6.3.2.

correlations for G_R that satisfy the reliability and consistency requirements described in Section 6.3.3. First we discuss these results from a purely topological perspective and then we interpret the biological aspects.

The results show that most metrics do not correlate well with growth. An obvious first explanation for this lack of correlation is that it is possible to remove a reaction without affecting growth (since the reaction may be part of a bypass that is not used when the cell is functioning normally). However, at a deeper level, the low correlations may be explained by the indirect relationship between the flux balance analysis framework (which measures function) and the network (on which topological metrics are measured). In flux balance analysis, growth is the objective function of a linear program in terms of metabolic fluxes, whilst the topologies of the metabolic networks are only functions of the stoichiometric matrix. While the objective function may be changed (perhaps to study a scenario other than growth maximization) the topology remains unchanged. Thus, correlations between the objective function and topological metrics depend to some extent on the objective function.

6.4.3. THE METABOLITE-REACTION NETWORK G_B IS THE BEST REPRESENTATION

Here we investigate some of the $\bar{\rho}$ correlations observed in Figure 6.5 and Figure 6.6. We generally limit our discussion to metrics for which $|\bar{\rho}| \geq 0.2$.

Metabolite-reaction network G_B As discussed in Section 6.3.2, correlations for the metabolite nodes $\mathscr{M}$ and the reaction nodes $\mathscr{R}$ were computed separately. First, the results for the metabolite nodes are considered, followed by the reaction node results.

Metabolite nodes $\mathscr{M}$: there are a number of relatively strong correlations for nodes in

$\mathscr{M}$, mostly falling into two groups:

- For both the metabolite nodes, so-called "hub" nodes provide shortcuts through which shortest paths are routed. Removal of a reaction node that interacts with a hub node may therefore remove a shortcut through which some shortest paths are routed. Thus, the mean reciprocal hop-count is decreased (and the mean hop-count is increased). In the remainder of the paper, all correlations associated with hub nodes are colored light gray.
- So-called "loner" nodes are nodes with low in-coming and/or out-going degrees. Some of these nodes are on important pathways and can cause growth to decrease when they are no longer produced (i.e., when their incoming links are removed) or consumed (i.e., when their outgoing links are removed) by any reactions. As a result, they are often implicated in correlations using the minimum function (those indicated by $\triangledown$). Correlations associated with loner nodes are colored dark gray.

Reaction nodes $\mathscr{R}$: only a few reliable, consistent correlations were found for the reaction nodes $\mathscr{R}$ in G_B. Of these, the mean reciprocal hop-count is the only reaction node metric that stands out, owing its presence to the metabolite hubs which provide shortcuts between a large number of reaction nodes.

Metabolite network G_M The correlation results for G_M are shown in Figure 6.6. G_M has more high-degree nodes than G_B and these are at least partially responsible for the strongest correlations. As with its progenitor G_B, the hub nodes in G_M provide shortcuts and thus provide the basis for the strong mean reciprocal node mean hop-count correlations.

The out-degree of out-neighbors correlations are due either to hub nodes themselves or nodes attached to the hub nodes (in particular hydrogen). The Dice similarity correlations are also the result of hub nodes – for example, the maximum mean Dice similarity is the result of a certain node (Asparagine) which is connected to a number of hub nodes; therefore it shares neighbors with many other nodes.

There are no apparent loner-node related correlations amongst the top correlations ($|\bar{\rho}| \geq 0.2$). However, the three correlations immediately following the top correlations (the minimum in-degree, hop-count and out-degree) are due to loner nodes.

Reaction network G_R The reaction network G_R yielded apparently no reliable, consistent correlations. As G_R is much denser than either G_B or G_M, each reaction removal forces a node to be removed from G_R. This leads to larger changes in G_R relative to the other networks; a property that may in part explain the difficulty of finding a connection between topology and growth in this representation.

Metabolite relationships hold the key to understanding the topology of metabolic systems The most interesting results are associated with the metabolite nodes. As mentioned in Section 6.3.1, there are more reactions than metabolites in metabolic systems. A reaction ties together a small number of metabolites while there are metabolites that are involved in many reactions. In other words, metabolites bind the network at a high

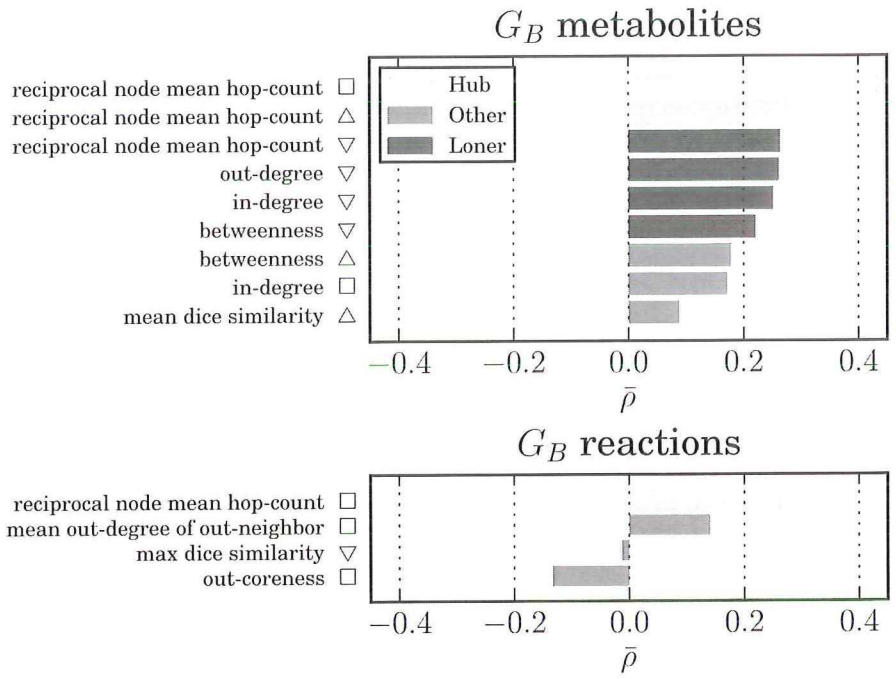


Figure 6.5: $\bar{\rho}$ measures between growth and topological metrics for G_B. The symbols ▽, □ and △ indicate that node values were reduced to single values by computing their minima, means and maxima respectively, as discussed in Section 6.3.2. As the legend shows, the light gray bars correspond to hub nodes, the dark gray bars correspond to loner nodes and the medium gray bars correspond to metrics that were either not interpreted or that do not fit the hub/loner distinction.

level and are responsible for global connectivity. This leads us to conclude that the metabolite-reaction network G_B and the metabolite network G_M are the most useful representations for our purposes. The reaction network G_R is less interesting, as no reliable, consistent correlations were found. Reactions are, of course, essential to the metabolic system, but metabolites tell the most interesting story.

Because G_B is the most accurate representation of the metabolic system and because of its strong correlations, we consider the metabolite nodes of G_B to be the most promising entities for studying metabolism.

6.4.4. THE STRONGEST CORRELATIONS POINT TO CURRENCY METABOLITES

Many of the hub metabolite nodes implicated in the previous section correspond to so-called *currency metabolites*. We know from biology that currency metabolites play a crucial role in metabolism: they are energy carriers or co-factors that are used by many reactions. Holme *et al.* [20] found the currency metabolites of *S. cerevisiae* to be H, H_2O, ATP, ADP, AMP, NAD, NADH, NADP, NADPH, CoA, CO_2, O_2, P_i, PP_i and NH_4^+ (for this set they used the undirected version of G_M with information taken from the BiGG database).

To validate the role of these metabolites, we repeated our experiments with currency

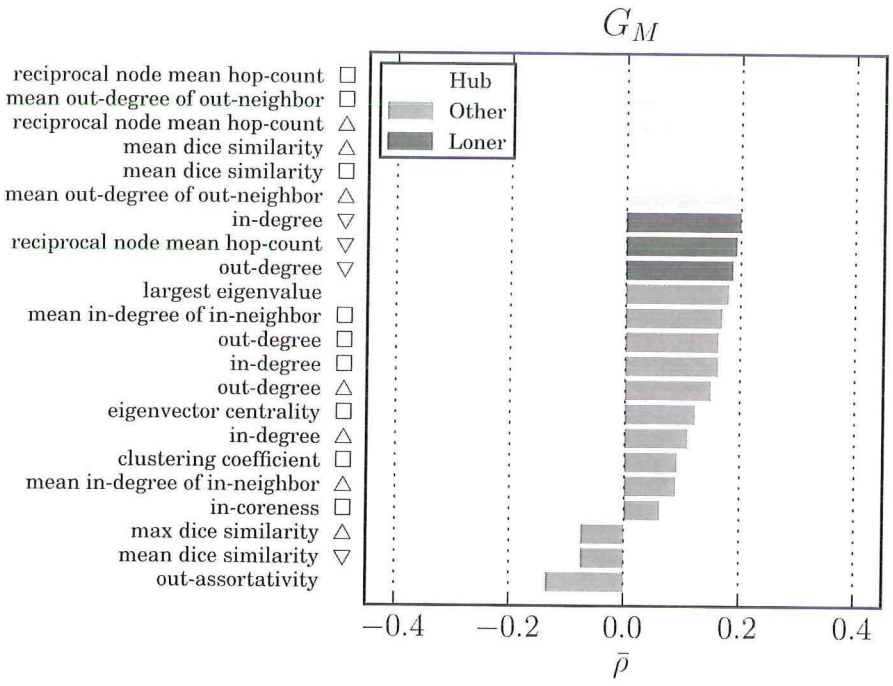


Figure 6.6: $\bar{\rho}$ measures between growth and topological metrics for G_M. The labels and colors are explained in Figure 6.5.

metabolites removed from G_B, G_M and G_R. Note that the metabolites were not removed from the flux balance linear program, as this would lead to incorrect chemical equations and it would change the computed growth. The five most significant $\bar{\rho}$ correlations for each of G_B, G_M and G_R are shown in Figure 6.7 (note that the reaction nodes in G_B were omitted, as all $\bar{\rho}$ correlations for these nodes fell below 0.2). Correlations that are neither the direct result of hub nodes nor loner nodes are shown as medium gray bars in the figure.

There are a number of interesting differences in the correlations brought about by currency metabolite removal:

- Most of the strong correlations due to hub nodes have been strongly reduced. The exception is the mean reciprocal mean hop-count correlation in G_M which remains approximately the same, in contrast with the correlation of the same metric in G_B. This hints at second-order network structure (as opposed to first-order hub structure) that is important in routing shortest paths.

- Removal of hub nodes removes shortcuts that route many shortest paths. The shortest paths are therefore more "spread out" through the metabolite network. This leads to a relative increase in node betweenness values and a concomitant increasing influence of arbitrary nodes on the average betweenness. Although this effect is most pronounced for G_M, it is also present for the metabolite nodes of G_B.

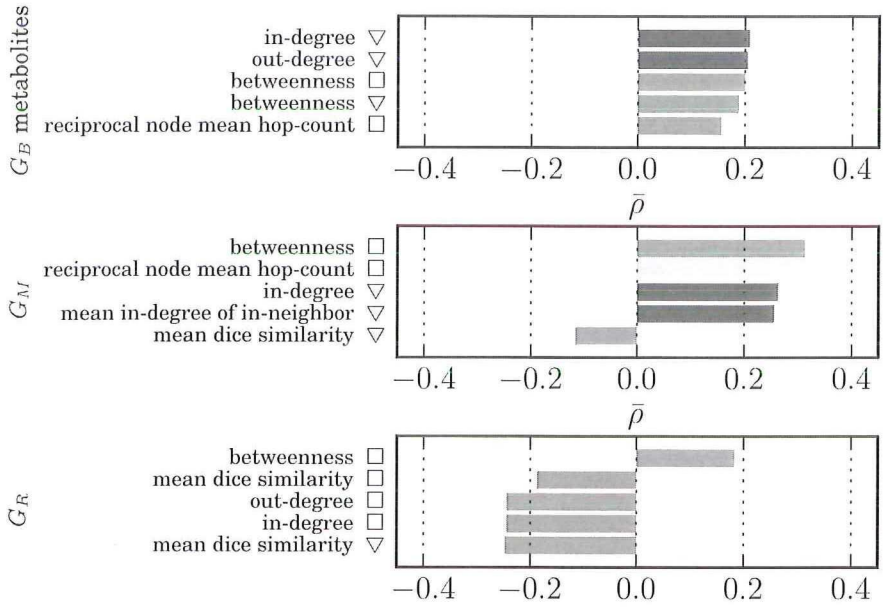


Figure 6.7: The most significant $\bar{\rho}$ correlations between growth and topological metrics for networks lacking currency metabolites. The labels and colors are explained in Figure 6.5.

- As G_R is less dense due to currency metabolite removal, a number of reliable, consistently correlated metrics could now be found. The majority of reaction nodes have degrees below the mean degree, so that a reaction removal is likely to increase the mean in- and out-degrees. Likewise the minimum and mean Dice similarities are likely to be increased, since the low-degree nodes have low Dice similarities. The correlations are not obviously due to hub nodes or loner nodes.

6.5. CONCLUSIONS

The goal of this study was to determine whether topology and robustness of biological systems are related. To this end, we generated a number of reaction removal sequences or *trials*, each of which resulted in the cessation of growth of our metabolic system. Each step in a trial provided a snapshot of the metabolic system from which growth could be computed as well as topological metrics of the metabolite-reaction network G_B, the metabolite network G_M and the reaction network G_R. This allowed us to calculate a measure of correlation between growth and each of the metrics. In this section, we will summarize some of our findings.

Unambiguously linking robustness to topology is difficult The term "robustness" is meaningless without context. Since the context of an organism constitutes all its interactions with its environment, a precise definition may forever elude us. However, every organism engages in a (small) number of vital functions that dominate its struggle for

survival. By studying only these functions and their degradation in the face of perturbations, we may discover some of the principles that help organisms to achieve their resilience. However, an unambiguous connection between such functions/perturbations and the topology of the underlying biochemical network is usually hard to define.

As we studied microbial metabolism, we focused on the one function at which the cell must be successful before all else: biomass production, or growth. While this is a simple representation of cellular activity, it has the advantage of being based on a well-studied theoretical model of metabolism, flux balance analysis, that can easily be modified to work with a perturbation model of reaction removal. Still, although we were able to directly link metabolic networks, functions and perturbations, finding correlations between robustness and topology proved not to be trivial.

There is no obvious way to reduce a metabolic system to a network This is a consequence of the correspondence between metabolic systems and hyper-networks. Analyzing hyper-networks directly is the ideal approach but these general structures have resisted the theoretical analysis that has produced the useful tools of (classical) network theory. Therefore, conversion is an analytical necessity. We described three ways of converting a hyper-network to a network: the metabolite-reaction network G_B, the metabolite network G_M and the reaction network G_R. The multiplicity of representations is a well-known problem that Holme *et al.* [20] investigated by matching graph theoretical properties of the three network representations to biological data in order to discover the network representation that "best" captures biological knowledge. We found the correlations were strongest for the metabolite nodes in G_B and for G_M. These findings suggest that metabolite nodes are most important for studying the structure of a metabolic system. In line with this, Holme *et al.* found G_M to be the most favorable representation, although we favor G_B since it maintains most of the original metabolic information.

Topology correlates weakly with growth Many of the topological metrics we calculated did not correlate with growth. We classified those that did correlate into two groups: those caused by hub metabolite nodes and those caused by loner metabolite nodes. They point to the importance of (a) global connectivity (by hub nodes that tie the network together by connecting many reaction nodes); and (b) local connectivity (by metabolites that are produced and/or consumed by few reactions). The role of hub nodes was verified in an experiment where we removed currency metabolites, which led to a large shift in metrics correlated to growth.

6.6. OUTLOOK

In this work, we studied the relationship between topology and growth. Using our framework as a starting point, one can investigate whether other functions of the metabolic network are related to topology or whether topology plays a role in other biological networks (e.g., gene regulation or protein interaction networks).

Our approach can be refined in a number of ways. On the one hand, flux balance analysis can be done with more sophisticated methods, such as MOMA [21] (Minimization Of Metabolic Adjustment) and ROOM [22] (Regulatory On/Off Minimization), both

of which were designed to better approximate the metabolic behavior of systems from which reactions have been removed. With sufficient enzyme kinetic parameters, one could even attempt to solve the non-linear differential equations (6.1). On the other hand, we could remove genes rather than reactions, more in line with the biological perturbations we intend to model. In this case, removing a gene may lead to the removal of multiple reactions, or alternatively a reaction may only be removed if all genes coding for isoenzymes are lost. However, such refinements to the model are unlikely to paint a very different picture since, if there were an effect, a first-order approach (such as ours) would pick up some correlation if it were there.

This work was an exploration of how topology is related to robustness. Although whether topology confers robustness or vice versa remains an open question, a change of perspective points to a number of paths for future investigation.

In a more local approach, one could isolate a small, fixed sub-network such as the citric acid cycle (a central part of metabolism in many organisms). Then our framework could be applied almost unchanged. Metrics would still be computed for entire networks but only the values corresponding to the sub-network under consideration would be compared with growth.

On a more global level, one could consider a number of species related by evolution. The species cannot necessarily be directly compared to each other, since they are specialized for different environments (and thus different contexts). But these differences in specialization enable us to study the connection between robustness and topology, since differences in metabolism are the results of specialization and these differences will be reflected in metabolic networks.

An interesting related approach is the study of the metabolic networks of gene knockout mutants of a given organism. This is essentially our approach with *in silico* knockouts replaced by *in vivo* knockouts, giving actual flux measurements which are more reliable than fluxes computed by flux balance analysis approaches.

REFERENCES

[1] W. Winterbach, H. Wang, M. Reinders, P. V. Mieghem, and D. de Ridder, *Metabolic network destruction: Relating topology to robustness,* Nano Communication Networks **2**, 88 (2011).

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

[3] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, and R. Chen, *Topological structure analysis of the protein-protein interaction network in budding yeast,* Nucleic Acids Research **31**, 2443 (2003).

[4] S. Maslov and K. Sneppen, *Specificity and stability in topology of protein networks,* Science **296**, 910 (2002).

[5] N. Guelzim, S. Bottani, P. Bourgine, and F. Kepes, *Topological and causal structure of the yeast transcriptional regulatory network,* Nature Genetics **31**, 60 (2002).

[6] A. L. Barabasi and R. Albert, *Emergence of scaling in random networks*, Science **286**, 509 (1999).

[7] W. K. K. Kim and E. M. Marcotte, *Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence.* PLoS Computational Biology **4**, e1000232+ (2008).

[8] G. Lima-Mendez and J. Helden, *The powerful law of the power law and other myths in network biology, Mol. BioSyst.*, Molecular BioSystems **5**, 1482 (2009).

[9] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, *Network motifs: Simple building blocks of complex networks*, Science **298**, 824 (2002).

[10] R. J. Prill, P. A. Iglesias, and A. Levchenko, *Dynamic properties of network motifs contribute to biological network organization*, PLoS Biology **3**, e343+ (2005).

[11] P. J. Ingram, M. P. Stumpf, and J. Stark, *Network motifs: structure does not determine function.* BMC Genomics **7**, 108+ (2006).

[12] J. Stelling, U. Sauer, Z. Szallasi, F. Doyle 3rd, and J. Doyle, *Robustness of cellular functions*, Cell **118**, 675 (2004).

[13] H. Kitano, *Biological robustness*, Nature Reviews Genetics **5**, 826 (2004).

[14] H. Kitano, *Computational systems biology*, Nature **420**, 206 (2002).

[15] J. D. Orth, I. Thiele, and B. O. Palsson, *What is flux balance analysis?* Nature Biotechnology **28**, 245 (2010).

[16] T. Murata, *Petri nets: Properties, analysis and applications, Proceedings of the IEEE*, Proceedings of the IEEE **77**, 541 (1989).

[17] S. Hardy and P. N. Robillard, *Modeling and simulation of molecular biology systems using petri nets: modeling goals of various approaches.* Journal of Bioinformatics and Computational Biology **2**, 595 (2004).

[18] M. E. J. Newman, *Assortative mixing in networks*, Physical Review Letters **89**, 208701+ (2002).

[19] B. O. Palsson, *Organisms - Systems Biology Research Group*, (2009), `http://systemsbiology.ucsd.edu`.

[20] P. Holme and M. Huss, *Substance graphs are optimal simple-graph representations of metabolism*, Chinese Science Bulletin **55**, 3161 (2010), for a list of the currency metabolites, see the arXiv version at `http://arxiv.org/abs/0806.2763`.

[21] D. Segrè, D. Vitkup, and G. M. Church, *Analysis of optimality in natural and perturbed metabolic networks*, Proceedings of the National Academy of Sciences of the United States of America **99**, 15112 (2002).

[22] T. Shlomi, O. Berkman, and E. Ruppin, *Regulatory on/off minimization of metabolic flux changes after genetic perturbations*, Proceedings of the National Academy of Sciences of the United States of America **102**, 7695 (2005).

7

LOCAL TOPOLOGICAL SIGNATURES FOR NETWORK-BASED PREDICTION OF BIOLOGICAL FUNCTION

Wynand WINTERBACH, Piet VAN MIEGHEM, Marcel REINDERS, Huijuan WANG, Dick DE RIDDER

7.1. ABSTRACT

In biology, similarity in structure or sequence between molecules is often used as evidence of functional similarity. In protein interaction networks, structural similarity of nodes (i.e., proteins) is often captured by comparing node signatures (vectors of topological properties of neighborhoods surrounding the nodes).

In this paper, we ask how well such topological signatures predict protein function, using protein interaction networks of the organism *Saccharomyces cerevisiae*. To this end, we compare two node signatures from the literature – the graphlet degree vector and a signature based on the graph spectrum – and our own simple node signature based on basic topological properties.

We find the connection between topology and protein function to be weak but statistically significant. Surprisingly, our node signature, despite its simplicity, performs on par with the other more sophisticated node signatures. In fact, we show that just two metrics, the link count and transitivity, are enough to classify protein function at a level on par with the other signatures suggesting that detailed topological characteristics are unlikely to aid in protein function prediction based on protein interaction networks.

This chapter was published in Pattern Recognition in Bioinformatics, Lecture Notes in Computer Science, *Springer Berlin Heidelberg* (2013) [1].

7.2. INTRODUCTION

To what extent does structure determine function in biology? Evolutionary principles have shown function and structure to be well correlated in genes with common evolutionary ancestors, allowing biologists to infer functions of proteins or genes based on their sequence *homology* (i.e., similarity) with other proteins or genes. With the arrival of network biology [2], homology was extended to take not only sequence similarity into account but also similarity of molecular interactions. These interactions can be either direct (physical) or indirect (functional). In other words, the manner in which a protein (or gene) is connected to other proteins in interaction networks matters. These other connecting proteins can be chosen in many ways, although the most common approach is to consider a network neighborhood centered around a protein in question, including all proteins and links within a fixed number of hops. Structural similarity of network neighborhoods is determined by comparing their *topological* properties. Typically, these properties are represented as a vector, known as a *topological signature*.

Topological signature similarity has been used as a measure of functional similarity between proteins in several algorithms aimed at the discovery of homology relations between proteins [3–5]. Although topological similarity and amino acid sequence similarity are typically both used to determine homology [3, 5], some of these algorithms perform well using only topological similarity [4, 5]. Researchers have also used topological similarity to predict relations other than homology, in effect assuming that structural similarity implies similarity of biological traits in proteins not necessarily related by evolution. Involvement in cancer (a phenotype) was found to be encoded in topological similarity [6] and even general protein function appears to be encoded in topology [7]. Given this predictive quality, the key question is thus: how exactly does local topology reflect function, and what signatures best capture local topology?

In this paper, we set out to answer these questions in a specific context, i.e.the prediction of protein function by means of node signatures in various protein interaction networks of the organism *Saccharomyces cerevisiae*. Topological signatures in the literature capture a lot of topological detail; in this paper we investigate the extent to which this detail improves protein function prediction (if at all). To this end, we study two such signatures – the graphlet signature of Milenković and Pržulj [7] and a signature based on the normalized Laplacian spectrum of a network [5] – as well as a simple node signature of our design. Predictive power of the signatures is determined by how well they discriminate between proteins with a given biological function and those without the function. To this end we use support vector machines, treating topological signatures as feature vectors and biological labels as classifier labels. Note that our aim is not the construction of an optimal protein function classifier, as for that purpose one would include many other types of data; rather, we use prediction accuracy as a measure to explore the relation between local topology and function.

7.3. METHODS

7.3.1. TOPOLOGICAL SIGNATURES

In the remainder of the text, G refers to a network (usually an interaction network), n to an arbitrary node of G and N the number of nodes in G. A k-neighborhood G_n^k of a

(a) (b)

Figure 7.1: Two neighborhoods of n: (a) G_n^1 and (b) G_n^2.

node n is an induced subnetwork of G on the set of nodes encompassing n and all nodes within k hops of n (a subnetwork is induced when two nodes in the subnetwork are connected by a link if, and only if they are connected in G). The subnetwork G_n^1 spanned by the gray nodes and bold links in Figure 7.1a is a 1-neighborhood of n, whilst the subnetwork G_n^2 spanned by the gray nodes and bold lines in Figure 7.1b is a 2-neighborhood of n.

GRAPHLET SIGNATURE:

Graphlets are small, connected, induced subnetworks, as illustrated in Figure 7.2 (labeled $X_1, X_2, \ldots, X_{30}$). The graphlet degree of a node n for a given subnetwork X_i can be regarded as a generalization of its degree: the number of X_i graphlets that contains n. In the special case where X_1 (i.e. two nodes connected by a link) is considered, the number of of X_1 subgraphs containing n is just the degree of n. A graphlet signature (also graphlet degree sequence [7]) generalizes the graphlet degree by including counts for all of the subnetworks in Figure 7.2.

To simplify exposition, we first construct a graphlet signature containing only the numbers of subnetworks X_1, X_2 and X_3 (Figure 7.2) that contain n. Such a signature can be represented as a vector of three integers. However, X_2 is not symmetrical, as the white node is structurally different from the two black nodes (which are interchangeable). We distinguish cases in which n takes the role of the white node from cases in which n takes the role of the black nodes. Thus, two counts for X_2 are maintained (one for each kind of node), leading to a signature vector of four integer components: one for X_1, two for X_2 and one for X_3 (vector indices are shown next to one node of each color).

The full graphlet signature is constructed by extending the construction above to the rest of the subnetworks in Figure 7.2. In total, the signature vector has 73 components (vector indices appear next to nodes). The largest subnetworks in Figure 7.2 have five nodes and therefore the graphlet signature is computed on 4-neighborhoods. The larger subnetworks in Figure 7.2 contain induced copies of smaller subnetworks (e.g., X_{30} contains X_9, X_3 and X_1), so that the components of the graphlet signature are not independent. Milenković and Pržulj [7] devised a weighting scheme to reduce this effect. We reweigh graphlets according to their method. Graphlet signatures were computed using code adapted from the original version of GraphCrunch [8].

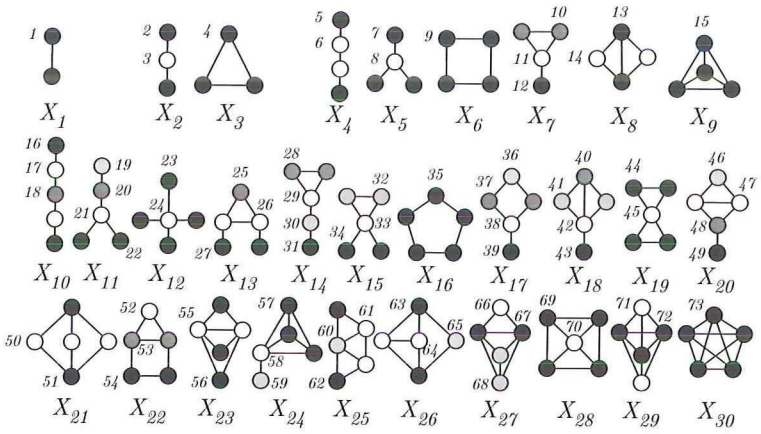


Figure 7.2: All non-isomorphic undirected networks (graphlets) with up to five nodes. For a given node n in a network G, Milenković & Pržulj [7] count how many times each of these networks includes n and appears as an induced subnetwork in G in order to construct a graphlet signature for n.

SPECTRAL SIGNATURE:

We assume that the nodes in G are labeled with numbers 1 through N. The *adjacency matrix* A of G is an $N \times N$ matrix in which $a_{i,j} = 1$ if the nodes i and j are connected by a link and $a_{i,j} = 0$ otherwise. The degree matrix Δ of G is a matrix in which $a_{i,i}$ equals the degree of node i and $a_{i,j} = 0$ if $i \neq j$. The *normalized Laplacian* is defined as $Q_{\text{norm}} = I - \Delta^{-1/2} A \Delta^{-1/2}$. The *spectrum* of Q_{norm} is its set of N eigenvalues. All eigenvalues of Q_{norm} fall within the range of $[0,2]$.

In general, two different neighborhoods have different numbers of nodes and therefore spectra of different sizes, making spectra unsuitable as feature vectors. We derive feature vectors by computing histograms of the spectra [5]. Histograms with 20 bins are computed on the range $[0,2]$, showing why the normalized Laplacian spectrum is preferred over the non-normalized version.

SIMPLE METRIC SIGNATURE:

Our own simple metric signature serves as a baseline. It contains four very simple topological properties of neighborhoods: 1) number of nodes, 2) number of links, 3) link density and 4) transitivity (the ratio of triangles to connected node triplets).

MULTI-RESOLUTION SIGNATURES:

One way to compute the spectral and simple metric signatures is to choose a fixed k and to compute the signatures on all k-neighborhoods. By focusing on fixed k, one may miss topologically distinguishing features at other "resolutions", i.e., other values of k. We construct "multi-resolution" versions of the spectral and simple metric signatures respectively by concatenating signatures of G_n^1, G_n^2 and G_n^3 for a given node n; henceforth we shall only consider these "multi-resolution" versions of the signatures. The graphlet

signature is already "multi-resolution" in the sense that its component graphlets span G_n^1, G_n^2, G_n^3 and G_n^4.

A COMBINED SIGNATURE:
Finally, we consider a signature that combines the previous signatures by simply concatenating the 1) graphlet signature, 2) the multi-resolution spectral signature and 3) the multi-resolution simple metric signature.

7.3.2. DATASETS

MOLECULAR NETWORKS:
All of the networks considered in this paper are protein interaction networks for the organism *Saccharomyces cerevisiae*. We have collected seven such networks, derived from four primary sources. Kim & Marcotte [9] provide two protein interaction networks, the first a high-quality literature-curated network and the second a high-throughput network. Yeastnet [10] provides several datasets with yeast protein interactions of which we downloaded the literature-curated dataset (denoted "LC" on the website) and the yeast 2-hybrid high-throughput dataset ("HC"). These two pairs of networks were selected because each pair contains a literature curated network and a high-throughput network, thereby providing insight into the impact of network quality on classification performance.

Our remaining two datasets are due to Krogan [11] and von Mering [12]. Both of these were used by Milenković & Pržulj [7] to test how well their graphlet signature approach fared in predicting protein function. We used the same two subsets of the von Mering dataset: "von Mering" contains the first 11000 protein interactions (of high-, medium- and low-confidence), whilst "von Mering core" contains all high-confidence interactions of the original dataset.

BIOLOGICAL LABELS:
Like Milenković and Pržulj [7], we used the MIPS protein annotations [13] as biological labels. MIPS annotations are hierarchical and have the form "xx.yy.zz..." where the letters denote two-digit biological categories. A protein may be annotated with multiple such annotations. The left-most category ("xx") gives the general protein function; each following two-digit category is a refinement ("yy" and "zz"). In this paper, we consider only general protein functions, of which there are 27 in the MIPS database.

7.3.3. CLASSIFICATION

Classification is performed using support vector machines (SVMs). There are numerous biological categories in the MIPS database and a protein may be annotated with any number of these categories. Since SVMs are binary classifiers, we use a one-versus-all strategy whereby we train a classifier for each biological category. Classifier performance is measured using the area under the curve (AUC) of the receiver operator curve (ROC) of a classifier. All classifier-related work was performed using Scikit-learn [14].

The radial basis function (RBF) kernel was used to train all SVMs. To reduce the impact of experimental omissions and noise, we only compute signatures on nodes whose degrees are at least 3 and that have at least one MIPS annotation. Furthermore, to ensure

the presence of enough positive instances in both testing and training sets, biological labels that appear in less than 20 nodes are not considered for classification training.

TRAINING REGIME:
For each topological signature type, for each network, for each biological function, a double cross validation training loop is performed [15]. The "outer" loop is a four-fold loop in which the training set contains 75% of the dataset whilst the testing set contains 25% of the dataset. For a given network and biological function, the folds are fixed, meaning that classifiers are trained on the same training samples for all topological signatures. Classifier performance is expressed as a combination of the mean and standard deviation of the four AUC values associated with the four outer folds.

The "inner" loop is responsible for finding the classifier with the best classification performance on the training set received from the "outer" loop. SVM classifiers using the RBF kernel require two parameters: a cost C (for penalizing incorrectly classified instances) and the RBF radius γ. These are optimized by walking along a grid of parameter pairs and training a classifier for each pair. Each grid point (i.e., parameter pair) is evaluated using the average AUC of a five-fold cross-validation loop. The parameters with the best AUC score are thus considered optimal. At the start of the "inner" loop, both the training and testing sets are centered and scaled using the center and variance of the training set. The graphlet signature is reweighed after this point using the weighting scheme of Milenković and Pržulj [7] as mentioned earlier in the paper (if reweighing is applied beforehand, it would be removed by the scaling step).

As grid searches are expensive, we first perform a parameter search on a coarse grid, followed by a second search on a fine grid around the optimal parameters found in the first search. The coarse grid is given by the Cartesian product $\mathscr{C} \times \Gamma$ of costs $\mathscr{C} = \{2^{-5}, 2^{-3}, 2^{-1}, \ldots, 2^{15}\}$ and RBF radii $\Gamma = \{2^{-15}, 2^{-13}, 2^{-11}, \ldots, 2^{3}\}$. The optimal parameter pair (C, γ) discovered on $\mathscr{C} \times \Gamma$ is then used to specify a fine grid $\mathscr{C}' \times \Gamma'$ where $\mathscr{C}' = \left\{2^{\log_2 C - 2 + i/2} \mid i \in \{0, 1, \ldots 8\}\right\}$ and $\Gamma' = \left\{2^{\log_2 \gamma - 2 + i/2} \mid i \in \{0, 1, \ldots 8\}\right\}$.

7.4. RESULTS AND DISCUSSION

Using the training regime described in the Methods section, we have computed, for each topological signature, for each network, for each biological function, the average classifier performance as well as its standard deviation. As this is a large amount of data, we have condensed the results into Figure 7.3a which shows, for a given topological signature and biological function, classification performance averaged over all networks, except for the high-throughput Yeastnet network. This dataset proved to be too small and gave poor, noisy classification results for all topological signatures. Figure 7.3a contains only those biological functions that appear in all the datasets. We also plotted the classification results for one high-quality dataset, the literature-curated Yeastnet dataset, in Figure 7.3b. The trends in Figure 7.3a are broadly similar in all of the networks although classification performance is generally lower than in Figure 7.3b.

What stands out most from both Figure 7.3a and Figure 7.3b, is that topology is, in general, a weak predictor of biological function. However, the mean AUC values are all above 0.5, showing that topology does encode a certain amount of information about biological function (the statistical significance of the mean AUC values being larger than

	Metabolism	Energy	Cell cycle & DNA processing	Transcription	Protein synthesis	Protein fate	Protein w. binding function	Regulation of protein function	Cellular transport	Cellular communication	Cell rescue & defense	Environment interaction	Cell fate	Development	Biogenesis	Cell type differentiation	Number of unique nodes
Kim & Marcotte, HT	271	63	377	481	130	347	399	62	207	46	123	94	80		220	128	1123
Kim & Marcotte, LC	452	84	674	676	149	655	652	157	469	134	235	221	192	35	458	288	1933
Krogan	321	81	423	483	183	378	405	70	205	61	148	115	87		277	134	1281
von Mering core	102	25	75	158	102	88	130		54		22	29	26		84	48	371
von Mering	471	120	231	382	289	295	369	49	193	39	114	99	68		222	104	1307
Yeastnet, HT	96	22	110	90		112	97	26	96	24	52	44	52		99	63	353
Yeastnet, LC	442	82	618	630	207	637	645	142	580	124	222	239	187	39	444	281	2006

Table 7.1: The number of positive instances for various combinations of network and biological function (i.e., proteins having given biological functions).

0.5 was tested using the t-test; in the majority of cases – and in all cases involving the biological categories "metabolism", "transcription", "protein synthesis" and "protein fate" – the associated p-values are below 0.05). The overall differences between Figure 7.3a and Figure 7.3b can be explained by differences in network quality and network size: quality affects classifier performance whilst network size affects its variance (network sizes are given in Table 7.1). The high-throughput networks contain the most noise and are therefore associated with worse classification performance.

At the level of biological categories both Figure 7.3a and Figure 7.3b show big differences in classification performance. The number of positive instances associated with a biological category (see Table 7.1) is weakly correlated with classifier performance, partly explaining the differences. Biology offers a possible explanation for the high AUC values associated with the labels "Transcription" and "Protein Synthesis": transcription and synthesis are both processes driven by permanent protein complexes rather than temporary groups of proteins (as found in many other processes). Thus, nodes with these functions tend to find themselves in densely connected clusters more often than other nodes.

Both overall classification performance, as well as performance associated with individual biological categories are dependent on the way in which biological categories are defined. Some categories are more general than others (for example, "Development" includes proteins engaged in diverse functions, whereas "Transcription" is a more specific function), contributing to differences in classification performance between categories. When the categories are too general, overall classification performance suffers as classifier inputs become difficult to distinguish. We have performed experiments (data not shown) in which we used two levels of the MIPS labels (labels of the form "xx.yy" rather than just "xx", i.e., more specific categories). Two-level categories led to better classification performance in some cases (notably those associated with transcription) and worse performance in other cases. The culprit is likely a paucity of positive instances

	Metabolism	Energy	Cell cycle & DNA processing	Transcription	Protein synthesis	Protein fate	Protein w. binding function	Regulation of protein function	Cellular transport	Cellular communication	Cell rescue & defense	Environment interaction	Cell fate	Development	Biogenesis	Cell type differentiation
Kim & Marcotte, HT	.39	.95	**.04**	.17	.39	.77	.69	.14	.10	.15	.19	.23	.35		.85	.56
Kim & Marcotte, LC	.42	.91	.10	.06	**.05**	**.00**	.27	.64	**.01**	.61	.74	**.01**	.31	.76	**.05**	.70
Krogan	.94	.08	.34	.13	.26	.20	.07	.12	.47	.90	.91	.07	.43		.18	.32
von Mering core	.75	.55	.14	.08	.26	.82	.56		.79		.92	.53	.87		.53	.97
von Mering	.19	.32	.49	.12	.59	.24	.26	.14	.24	.06	.50	.43	.60		.17	**.04**
Yeastnet, HT	.44	.22	.12	.36		.68	.19	.07	.18	**.04**	.12	**.00**	.45		.69	.70
Yeastnet, LC	.80	.42	.84	.55	.60	.11	.91	.85	**.04**	.23	.93	.62	.63	**.05**	**.01**	.12

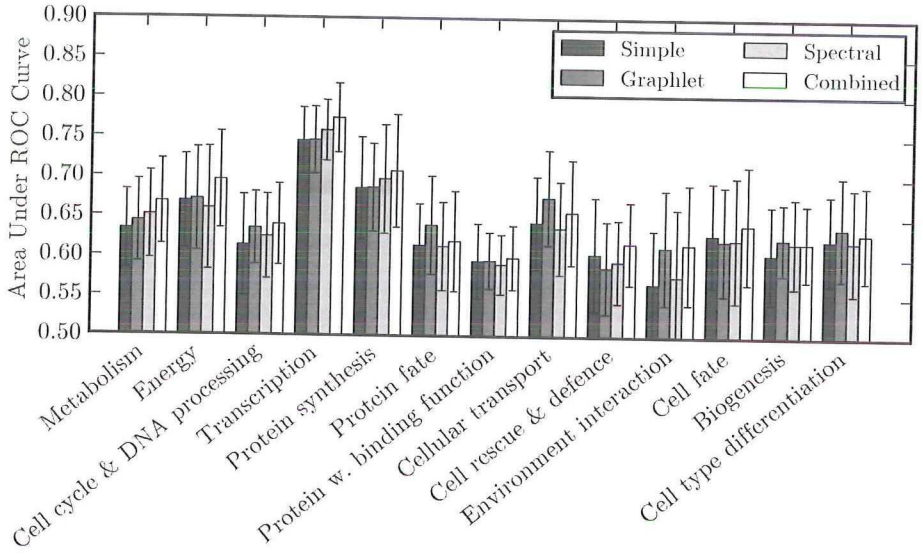
Table 7.2: p-values of one-way ANOVA tests applied to the AUC values of the three topological signatures (graphlet, spectral and simple) for each network and biological function combination. We consider p-values of 0.05 and below to be significant (shown in bold text).

associated with many of the two-level labels.

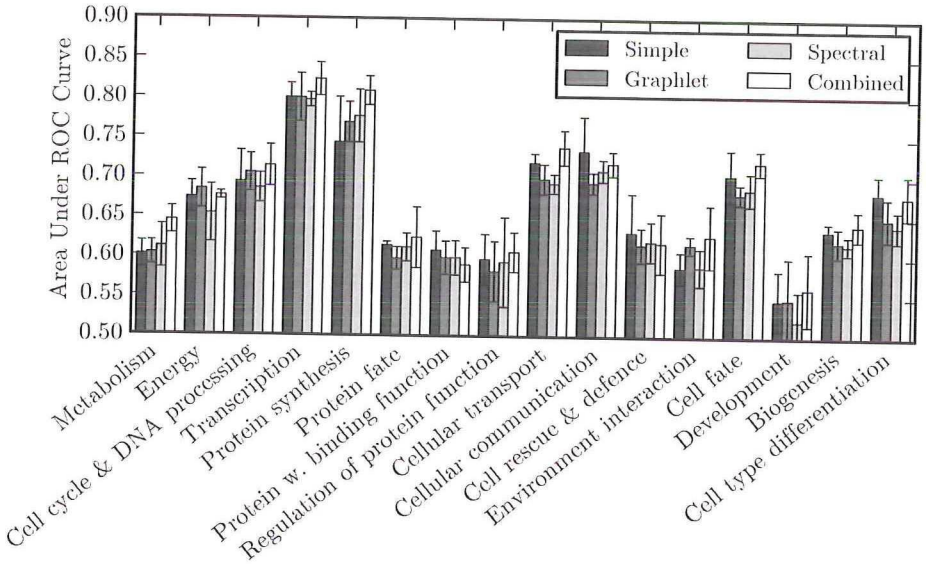
Another salient aspect of Figure 7.3a and Figure 7.3b is that the three topological signatures perform very similarly. We tested whether the AUC values of the individual signatures (i.e., not the combined signature) for each biological category were different, using a one-way ANOVA (Table 7.2). We consider p-values of 0.05 and below to be statistically significant and find only 10 dataset/function combinations that pass this threshold.

Although the three topological signatures lead to similar classification results, it may be possible that they nevertheless measure different (discriminative) topological characteristics. If this is true, combining the signatures should lead to improved classification performance. However, Figure 7.3a and Figure 7.3b do not support such a conclusion. Thus, in the context of our datasets and classifier, the topological signatures are not complementary.

Given that the simple metric signature is competitive with the graphlet and spectral signatures, it is natural to ask whether it cannot be further simplified. We investigated all possible combinations of the four metrics (number of nodes, number of links, density and transitivity) that make up the simple metric signature, constructing 14 simpler signatures: 4 signatures using only one metric each, 6 signatures using pairs of metrics and 4 signatures using triplets of metrics. The mean classification performance of these metrics, taken over all datasets and all biological categories, is shown in Figure 7.4. The link count L and transitivity T are sufficient for obtaining good classification performance. The implication is that what matters in function prediction in protein interaction networks, is the number of nodes and the "clusteredness" (transitivity). Since proteins of similar function tend to form clusters, their neighborhoods overlap and therefore they share topological characteristics. Apparently, "clusteredness" signatures are unique enough to distinguish similar proteins from other proteins.



(a)



(b)

Figure 7.3: Classification performance of the three topological signatures, as well as a signature that combines the three signatures. (a) Performance of our SVM classifiers averaged MIPS categories present in all datasets (excluding the high-throughput Yeastnet dataset; see text for explanation). Error bars show the standard deviation. (b) Classification performance of the three topological signatures on the literature-curated Yeastnet network [10].

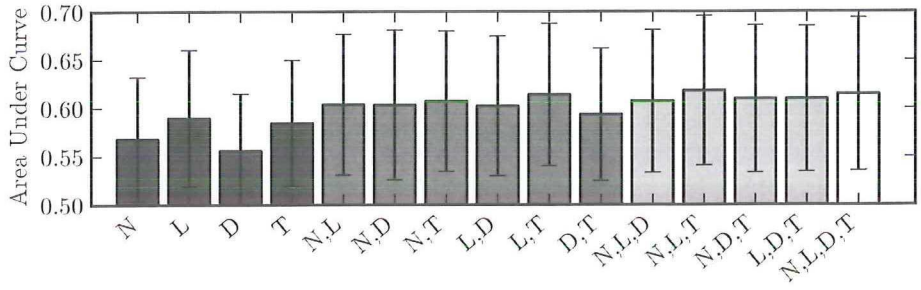


Figure 7.4: Classification performance of various combinations of the features used in the simple metric signature averaged over all datasets and all functions. Here, N is the number of nodes (in a neighborhood), L is the number of links, D is the density and T is the transitivity.

7.5. CONCLUSION

At the start of this paper, we asked to what extent structure – i.e., topology – determines function in biology. We focused on the use of signatures to express topological properties of neighborhoods surrounding nodes in molecular interaction networks. Our study is motivated by the use of topological signatures as a tool for discovering similar genes or proteins (under the assumption that topological similarity implies functional similarity). We specifically studied the use of such signatures to discriminate between proteins with a given biological function and those without it, using protein interaction networks derived from *Saccharomyces cerevisiae* and support vector machines.

Current node signatures, such as the graphlet signature [7] and signatures based on spectra [5] capture very detailed topological profiles. We compared these with our own topological signature, based on very simple network metrics. For all signatures, classifier performance tended to be weak, implying that topology is, at least for *Saccharomyces cerevisiae* protein interaction networks, a weak predictor of function. However, with the exception of one noisy protein interaction network classifiers performed better than random, showing that topology and function are linked. How much better depends on the functional category considered, with performance particularly strong for transcription and protein synthesis.

Our simple metric signature performed on par with the graphlet and spectral signatures. We also established that the signatures are not complementary for protein function prediction, as a combined signature incorporating all three signatures does not yield better accuracy. Since our simple metric signature captures less topological information than the other signatures, we conclude that fine topological detail is not very useful in the prediction of protein function. Strikingly, performance when using only the link count and transitivity, measures of "clusteredness", is as good as when using the more complex signatures. This is not simply a side-effect of dataset noise, as our simple metric signature performs equally well in the high quality networks.

Our work opens a number of paths for future research. For our conclusions to hold generally, the techniques used in this paper should be applied to other types of interaction networks (for example, co-expression networks and synthetic sick-or-lethal net-

works) and to networks derived from other organisms. It would be particularly interesting if link count and transitivity are found to be equally determinative in other interaction network types. Finally, it is not yet known how different "resolutions" contribute to signature performance and whether a particular resolution (i.e., k-neighborhoods of a particular k) dominates classification performance.

REFERENCES

[1] W. Winterbach, P. Mieghem, M. Reinders, H. Wang, and D. Ridder, *Local topological signatures for network-based prediction of biological function*, in *Pattern Recognition in Bioinformatics*, Lecture Notes in Computer Science, Vol. 7986, edited by A. Ngom, E. Formenti, J.-K. Hao, X.-M. Zhao, and T. Laarhoven (Springer Berlin Heidelberg, 2013) pp. 23–34.

[2] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, *Hierarchical organization of modularity in metabolic networks*, Science **297**, 1551 (2002).

[3] C.-S. Liao, K. Lu, M. Baym, R. Singh, and B. Berger, *IsoRankN: spectral methods for global alignment of multiple protein networks*, Bioinformatics **25**, i253 (2009).

[4] T. Milenković, W. L. L. Ng, W. Hayes, and N. Pržulj, *Optimal network alignment with graphlet degree vectors.* Cancer Informatics **9**, 121 (2010).

[5] R. Patro and C. Kingsford, *Global network alignment using multiscale spectral signatures*, Bioinformatics (2012), 10.1093/bioinformatics/bts592.

[6] T. Milenković, V. Memišević, A. K. Ganesan, and N. Pržulj, *Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data*, Journal of The Royal Society Interface **7**, 423 (2010), http://rsif.royalsocietypublishing.org/content/7/44/423.full.pdf+html .

[7] T. Milenković and N. Pržulj, *Uncovering biological network function via graphlet degree signatures.* Cancer informatics **6**, 257 (2008).

[8] T. Milenkovic, J. Lai, and N. Pržulj, *GraphCrunch: A tool for large network analyses*, BMC Bioinformatics **9**, 70 (2008).

[9] W. K. Kim and E. M. Marcotte, *Age-Dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence*, PLoS Computatinal Biology **4** (2008), 10.1371/journal.pcbi.1000232.

[10] K. McGary, I. Lee, and E. Marcotte, *Broad network-based predictability of saccharomyces cerevisiae gene loss-of-function phenotypes*, Genome Biology **8**, R258 (2007).

[11] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrín-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J.

Thompson, G. Musso, P. St Onge, S. Ghanny, M. H. Y. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt, *Global landscape of protein complexes in the yeast saccharomyces cerevisiae,* Nature **440**, 637 (2006).

[12] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, *Comparative assessment of large-scale data sets of protein-protein interactions.* Nature **417**, 399 (2002).

[13] A. Ruepp, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Güldener, G. Mannhaupt, M. Münsterkötter, and H. W. Mewes, *The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes,* Nucleic Acids Research **32**, 5539 (2004), http://nar.oxfordjournals.org/content/32/18/5539.full.pdf+html .

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine learning in Python,* Journal of Machine Learning Research **12**, 2825 (2011).

[15] L. F. A. Wessels, M. J. T. Reinders, A. A. M. Hart, C. J. Veenman, H. Dai, Y. D. He, and L. J. van 't Veer, *A protocol for building and evaluating predictors of disease state based on microarray data,* Bioinformatics **21**, 3755 (2005), http://bioinformatics.oxfordjournals.org/content/21/19/3755.full.pdf+html .

8

CONCLUSION

8.1. THESIS SUMMARY

The central question of this thesis is whether the topologies of molecular networks reflect the functions of the cellular systems they model. Since biological organisms are robust in that they continue functioning when damaged, we first asked whether molecular networks might provide clues about topological structures associated with robustness. Such structures would be useful in evaluating network robustness and could inform the design of more robust human-made networks such as communications and transport networks. In the context of human-made networks, robustness means *resilience against network disconnection and resilience against drastic increases in path length between arbitrary end-points.* Thus, if molecular network topology is to guide the design of communications or transport networks, it is implicitly assumed that transport (of signals or metabolites) are central to their biological roles. This is not true of all molecular networks but is partially true of protein interaction networks and metabolic networks, two classes of networks investigated in this thesis.

Degree assortativity is a relatively new and simple metric that we investigated as a measure of robustness. The metric is a correlation coefficient representing the tendency of nodes to connect to nodes of similar degree. We found that for dense networks, the range of values achievable by the metric is dependent on the density of the network and that this relationship is non-linear, limiting its use as a tool for comparing dense networks. On the other hand, for sparse networks such as molecular networks, the range of achievable degree assortativity values is larger and, crucially, not very dependent on network density. Thus, degree assortativities of sparse networks can be meaningfully compared.

Degree assortativity also lends itself well to greedy optimization: a network can be re-architected through a number of small, degree-preserving rewiring steps to have increased or decreased degree assortativity. Empirical studies suggest that this simple greedy approach can be used to achieve assortativity values close to the maximum or minimum values achievable through degree-preserving rewiring transformations. Coupled with our robustness envelope framework, we were able compare the robustness

of large networks with the same degree sequences but (arbitrarily) different assortativity values. Moderate increases in assortativity led to slightly increased resilience against targeted node attacks coupled with decreased resilience against random attacks; for moderate decreases in assortativity, the opposite was true. Greater changes in assortativity lead to network fragility. We find that although there is some connection between degree assortativity and robustness, the connection too weak to be practically useful.

We investigated the assortativity of metabolic networks, finding them to be neither assortative nor disassorative, a finding also corroborated for protein interaction networks. Perhaps degree assortativity is simply the wrong metric for studying molecular networks. In order to test this possibility, we studied robustness envelopes of unmodified molecular networks and compared them to robustness envelopes of other real-world networks. This showed that molecular networks are not distinguished by their robustness envelopes. In all of the previously-mentioned analyses, robustness meant *resilience against changes in general topological properties such as path length.* However, in the context of biological organisms, robustness normally means *the ability of an organism to continue functioning when part of the organism is damaged.* Therefore, a better way to discover the link between topology and biological robustness is to search for topological metrics that are correlated with biological function.

Accurately defining biological function is difficult, as we do not yet have many good models of molecular interaction (indeed, network biology came about as a tool to help analyze and understand molecular interactions). One of the best models of biological function of large networks available today is flux balance analysis, a model for computing metabolic reaction rates during metabolic steady state. From the flux rates, one can compute the rate of biomass production, a commonly used metric of cellular health for cells not subjected to stressors. In addition, high quality metabolic networks for a number of organisms are available. With these tools and data, we intentionally damaged a yeast metabolic network, correlating changes in the rate of biomass production to changes in a number of topological metrics in the metabolic network.

In general, correlations between biomass production and topological metrics are weak, providing further evidence that general topological metrics are unlikely to teach us much about biological robustness. We conclude that metrics summarizing network structure are insensitive to small changes in network structure whilst such changes may be decisive in organismal health. One way around this problem is to develop biologically-relevant topological metrics that would be sensitive to biologically significant changes. Developing biologically-relevant metrics for molecular networks will be difficult if one insists on whole-network analysis. But network biology need not focus only on whole-network analysis: at smaller scales (that is, smaller collections of interacting molecules), molecular interactions are better understood and at these scales, topological metrics are more sensitive to changes in network structure (since they are summarizing smaller networks).

Using a local approach, we correlated functions of proteins with topological characteristics of network regions surrounding these proteins. The correlations show a stronger connection between topology and biology than in the previous work. A number of topological metrics were correlated with protein function but no great differences in correlation were found between the metrics. This is surprising, given that one of the metrics

took only simple topological aspects into account suggesting that topological metrics need not be complex in order to be effective (at least given the current state of our understanding of biological networks).

8.2. FUTURE WORK

Network biology is still a young field and has the potential to develop in a number of directions. Here we identify a number of themes that we believe will advance the state of the art.

Our work started with a whole-network analysis and ended with with a local topological approach for the prediction of protein function. The global-to-local movement is a trend in the field of network biology but that does not mean that global and local approaches exclude one another. Surprisingly little attention is given to hierarchical analysis, a global-local technique that is commonly used in molecular biology. For example, interacting proteins are grouped into protein complexes which themselves are studied as units of interaction. One need not stop here since protein complexes can also be grouped into larger units that interact with one another. Network biology augments such analysis by studying interactions of complexes as networks. Hierarchical analysis of a system leads to multiple network descriptions at varying levels of detail. Networks at the highest level are small, capturing interactions between complex, large subsystems. Nodes of any network in the hierarchy (except those at the lowest level) represent subnetworks of the network one step lower in the hierarchy. Any given molecular interaction system has numerous hierarchical descriptions that reveal different aspects of the system.

Whilst existing topological metrics might be correlated with biological function at a local scale, our success rate at tying topology to biological function will improve if we focus on measuring biologically meaningful topological aspects. An example is a proposed topological metric for measuring biomass production in metabolic networks that counts the number of metabolites a metabolic network can produce, given a set of inputs. In a sense, the metric is a discrete approximation of the chemical process and, not surprisingly, correlates well with biomass production. On the other hand, it shows that cellular growth is somewhat insensitive to the exact distribution of flux rates in the network. Such biologically-inspired topological metrics could be designed for the various types of molecular network.

With the design of biologically-relevant network metrics, we no longer need to consider only highly homogeneous networks (in which nodes all represent the same object and in which links all represent the same kind of relation). Network theory has traditionally mainly dealt with homogeneous networks, partly due to the fact that transport networks and communications networks are relatively homogeneous (at least in comparison to molecular networks) and partly to ease mathematical analysis. Breaking homogeneity by considering molecular networks with multiple types of nodes and links will complicate mathematical analysis but it is a price worth paying for the additional biological detail captured in such networks. Decreased mathematical tractability can also be partly offset by software simulations: molecular networks are generally small enough for millions of instances to be tested against a given metric.

Related to the study of heterogeneous networks and hierarchical analysis is the study

of interlinked networks (that is, networks of networks). For example, consider metabolic networks and protein interaction networks: faulty interactions in the protein interaction network may lead to failure of membrane transporters necessary for transporting metabolites needed by the metabolic network; if the failure of the metabolic network is such that it can no longer produce amino acids, the protein interaction network will quickly disintegrate. These kinds of interactions cannot be understood by studying the metabolic network or protein interaction network in isolation. The strength of interlinked network analysis is that it leverages analytic techniques applicable to isolated networks (for example, flux balance analysis of metabolic networks) whilst putting interactions between networks on an equal footing with other interactions.

8.3. CLOSING REMARKS

This thesis has investigated the connection between molecular network structure and function. Network biology sparked a considerable amount of interest because it seemed to promise a new paradigm within which to study complex systems. Its impact has not been as radical – complex systems remain complex – but it is a relatively simple analytic tool which has broadened our understanding of cellular interaction systems. That alone is a good reason to continue developing this young field.

SUMMARY

During the second half of the 20th century, the field of molecular biology greatly improved our understanding of the cell. Molecular biology is a reductionist science – molecules and interactions are studied in isolation from the rest of the cell. However, as our understanding of the cell increased, it became apparent that cellular interactions are often decisive in biological function and that more holistic analysis techniques were required to understand certain biological phenomena.

The field of network biology emerged from the need to analyze cellular interactions at a scale beyond what was possible using a reductionist approach. Network biology, a synthesis of molecular biology and graph theory, treats sets of molecular interactions as molecular networks that can be analyzed using graph-theoretical techniques.

This thesis deal with three themes in network biology: 1) the correlation of structural properties of molecular networks with the robustness of the organisms they model, 2) the correlation of structural properties of molecular networks with biological properties, 3) the scale at which structural properties of molecular networks should be considered in order to make biologically meaningful inferences (that is, should entire networks or small regions of networks be considered).

The first theme was motivated by the hypothesis that, because biological organisms are robust (that is, they maintain function in the face of damage), molecular networks might contain structural features that are associated with robustness. Such knowledge could both be used to analyze networks for robustness and in the design of more robust networks (such as communications networks). The connection between robustness and structure was first studied from a purely structural perspective: various networks were compared to one another based on whether they maintained certain structural properties when damaged. The purely structural perspective showed no discernible differences between the structural robustness of molecular networks and other real-world networks. If the connection between biological robustness and network structure were to be understood, changes in biological function had to be tied to changes in molecular network structure. This was investigated by damaging metabolic networks and correlating changes in their structures to changes in their predicted ability to produce biomass. Although correlations were found, they proved to be weak, suggesting that structural properties computed using entire networks are too insensitive to capture biologically significant changes.

The second theme, correlations between structural properties of molecular networks and biological properties, at first glance appears to be similar to the first theme. But whereas the connection between structure and robustness focused on relating system-wide structural properties to system-wide biological behavior, this work focused on the relation between structural properties of individual molecules and their biological properties. Stated otherwise, the first theme deals with global analysis techniques whereas the second theme with (relatively) local analysis techniques. This shift towards the local

yielded stronger correlations between structure and biological function than achievable by more global approaches.

The original research in this thesis suggests that local analysis techniques have better predictive power than global analysis techniques. This observation forms the basis of the third theme: at which scale are molecular networks best analyzed in order to make biologically meaningful predictions? This theme was thoroughly examined in a literature survey of the field of network biology. The literature survey shows a slow shift from global approaches to more local approaches, mirroring the findings of this thesis. The success of local approaches over global approaches is explained by the fact that 1) it is simpler to associate biological interpretations with structural properties of small sets of molecules (rather than large networks of molecules) and 2) structural properties pertaining to entire networks tend to be insensitive to small changes in structure that may be biologically significant.

Besides the network biology-focused research, this thesis also examines the effect of changes in degree assortativity on the robustness of molecular networks and other real-world networks. The metric itself was also studied to better understand its limits and applicability to molecular networks.

SAMENVATTING

Gedurende de 2e helft van de 20e eeuw heeft moleculaire biologie ons begrip van de cel vergroot. Moleculaire biologie is een reductionistische wetenschap – moleculen en interacties worden geïsoleerd van de rest van de cel bestudeerd. Echter, sinds ons begrip van de cel toe is genomen, is het gebleken dat cellulaire interacties vaak bepalend zijn voor de biologische functies en dat er meer holistische analysetechnieken nodig zijn om bepaalde biologische verschijnselen te begrijpen.

Netwerkbiologie is ontstaan uit de noodzaak om cellulaire interacties te analyseren op een schaal groter dan wat er mogelijk was met een reductionistische benadering. Netwerkbiologie, een synthese van moleculaire biologie en graaftheorie, benadert verzamelingen van moleculaire interacties als moleculaire netwerken die geanalyseerd kunnen worden met graaf-theoretische technieken.

Dit proefschrift behandelt drie thema's binnen netwerkbiologie: 1) De correlatie van structuureigenschappen van moleculaire netwerken met de robuustheid van de organismen die gemodelleerd worden, 2) de correlatie van structuureigenschappen van moleculaire netwerken met biologische eigenschappen, 3) de schaal waarmee structuureigenschappen van moleculaire netwerken moeten worden beschouwd om biologisch zinvolle gevolgtrekkingen te maken (dat wil zeggen, moet het hele netwerk of kleine gebieden van netwerken worden beschouwd).

Het eerste thema werd ingegeven door de veronderstelling dat, omdat de biologische organismen robuust zijn (dit wil zeggen dat zij hun functie behouden bij schade), moleculaire netwerken mogelijk structuureigenschappen zullen bevatten die geassocieerd zijn met robuustheid. Dergelijke kennis kan zowel worden gebruikt voor het analyseren van netwerken op robuustheid en in het ontwerp van meer robuuste netwerken (bijvoorbeeld communicatienetwerken). De connectie tussen robuustheid en structuur werden eerst bestudeerd vanuit een structuurperspectief: verschillende netwerken werden met elkaar vergeleken gebaseerd op de vraag of ze bepaalde structurele eigenschappen zouden behouden bij schade. Dit structuurperspectief liet zien dat er geen onderscheidbaar verschil is tussen de structurele robuustheid van moleculaire netwerken en andere real-world netwerken. Als de verbinding tussen biologische robuustheid en netwerkstructuur zou worden begrepen, dan zouden veranderingen in biologische functie verbonden moeten zijn met de moleculaire netwerkstructuur. Dit werd onderzocht door schade toe te brengen aan metabolische netwerken en het correleren van de veranderingen in structuur met veranderingen in voorspelde mogelijkheid om biomassa te produceren. Hoewel er correlaties werden gevonden, bleken deze zwak, wat suggereert dat structuureigenschappen die berekend zijn met netwerken in hun geheel te ongevoelig zijn om biologisch significante veranderingen op te vangen.

Het tweede thema, correlaties tussen structuureigenschappen van moleculaire netwerken en biologische eigenschappen, lijkt op het eerste gezicht vergelijkbaar te zijn met het eerste thema. Maar waar de verbinding tussen structuur en robuustheid is gericht op

systeem brede structuureigenschappen te relateren aan systeem brede biologische eigenschappen, focussen we hier op de relatie tussen structurele eigenschappen van individuele moleculen en hun biologische eigenschappen. Anders gezegd, het eerste thema betreft de globale analysetechnieken, terwijl het tweede thema (relatief) lokale analysetechnieken betreft. Deze verschuiving naar het lokale leidde in vergelijking tot sterkere correlaties tussen structuur en biologisch gedrag dan bij de globale aanpak.

Het oorspronkelijke onderzoek in dit proefschrift suggereert dat lokale analysetechnieken een grotere voorspellende waarde hebben dan globale analysetechnieken. Deze observatie vormt de basis voor het derde thema: op welke schaal dien je moleculaire netwerken te analyseren om voorspellingen te doen die biologisch van betekenis zijn? Dit thema werd grondig bestudeerd in een literatuurstudie met betrekking tot het veld netwerkbiologie. Deze literatuurstudie laat een langzame verschuiving zien van globale benaderingen naar meer lokale benaderingen, een weerspiegeling van de resultaten van dit proefschrift. Het succes van lokale benaderingen in vergelijking tot globale benaderingen laat zich het beste verklaren doordat het 1) eenvoudiger is om biologische interpretaties met structuureigenschappen van kleine verzamelingen moleculen te associëren (in plaats van met grote verzamelingen van moleculen) en 2) structuureigenschappen die betrekking hebben op een heel netwerk kunnen ongevoelig zijn voor kleine biologisch significante veranderingen in de structuur.

Naast het onderzoek gericht op netwerkbiologie, beschrijft dit proefschrift ook het effect van veranderingen in de graad van assortativiteit op robuustheid van moleculaire netwerken en andere real-world netwerken. Deze metriek werd bestudeerd om beter te begrijpen hoe toepasbaarheid deze is op moleculaire netwerken en wat zijn beperkingen zijn.

ACKNOWLEDGMENTS

FRIENDS AND FAMILY

This dissertation would not have been possible without the constant love and support of my father and mother. You did much more than anyone could have expected of you. *Dankie, Mamma en Pappa.* And to my brothers Hendrik and Christian: thank you for always supporting me.

To my good friends Peter van der Merwe and Lisa Ramsay Spangehl, even though you are both very, very busy people, you have always been there for me when I really needed you.

Here in the Netherlands, I was fortunate to have Francois Malan and Linda Klumpers, two dear friends that helped me numerous times with practical as well as emotional matters.

I was happy to have Norbert Blenn as an office mate and a house mate. His zen-like approach to life definitely rubbed off on me. As for the other members of the NAS group, past and present: I miss you already. I was also a member of the bio-informatics group and I have fond memories of Delft bio-informaticians.

Last but most certainly not least, a warm acknowledgement to the friends I made in Delft, London and elsewhere in Europe since moving to the Netherlands in 2009.

ACADEMIA

I would like to highlight the roll that Dr. Dick de Ridder, my co-promotor, played during my PhD. He spent an immense amount of time guiding me, brainstorming with me and in co-drafting papers.

I am thankful for the guidance of my promotor Prof. Piet Van Mieghem, who tirelessly steered me towards a successful PhD even though the road was sometimes bumpy.

Dr. Huijuan Wang and Prof. Marcel Reinders also contributed in no small part to my supervision. Thank you both.

I would like to thank my co-authors, Stojan Trajanovski, Javier Martín-Hernández and Jurgen Nijkamp. It was a pleasure to work with peach of you.

Finally, I would like to thank Peter Krekel for translating my propositions and summary into Dutch.

Curriculum Vitæ

Wynand Winterbach

23-10-1979 Born in King Williams Town, South Africa.

Education

1993–1997 High School
Durban North College, Durban

1998–2001 Undergraduate in Mathematical Sciences
Stellenbosch University

2002–2004 MSc. in Applied Mathematics
Stellenbosch University
Thesis: The crossing number of a graph in the plane
Supervisor: Prof. dr. J. H. van Vuuren

2009–2013 PhD. Network Theory
Delft University of Technology
Thesis: Topology of Molecular Networks
Promotor: Prof. dr. ir. P. F. A. Van Mieghem
Co-promotor: Dr. ir. D. de Ridder

Work

Wynand Winterbach has worked as a programmer at a variety of institutions. His programming career started at Flextronics (2005) and iThemba Labs (2006). In 2007, he took a break from programming and lectured at a community college but soon returned to the world of software at Translate.org.za (2008). He currently works at Clinical Graphics (2013) where he develops software for hip motion simulation.

List of Publications

10. **W. Winterbach, P. Van Mieghem, M. Reinders, H. Wang, D. de Ridder**, *Topology of Molecular Interaction Networks*, BMC Systems Biology **7**, 90 (2013).

9. **W. Winterbach, P. Van Mieghem, M. Reinders, H. Wang, D. de Ridder**, *Local Topological Signatures for Network-Based Prediction of Biological Function*, *Pattern Recognition in Bioinformatics*, Lecture Notes in Computer Science **7986** (Springer Berlin Heidelberg, 2013).

8. **S. Trajanovski, J. Martín-Hernández, W. Winterbach, P. Van Mieghem**, *Robustness envelopes of networks*, Journal of Complex Networks **1**, 44 (2013).

7. **R. Van Berlo, W. Winterbach, M. de Groot, A. Bender, P. Verheijen, M. Reinders, D. de Ridder**, *Efficient calculation of compound similarity based on maximum common subgraphs and its application to prediction of gene transcript levels*, International Journal of Bioinformatics Research and Applications **9**, 4 (2013).

6. **W. Winterbach, D. de Ridder, H. Wang, M. Reinders, P. Van Mieghem**, *Do greedy assortativity optimization algorithms produce good results?*, The European Physical Journal B **85**, 5 (2012).

5. **H. Wang, W. Winterbach, P. Van Mieghem**, *Assortativity of complementary graphs*, The European Physical Journal B **83**, 2 (2011).

4. **W. Winterbach, H. Wang, M. Reinders, P. Van Mieghem, D. de Ridder**, *Metabolic network destruction: Relating topology to robustness*, Nano Communication Networks **2**, 2–3 (2011).

3. **J. Nijkamp, W. Winterbach, M. van den Broek, J-M. Daran, M. Reinders, D. de Ridder**, *Integrating genome assemblies with MAIA*, Bioinformatics **26**, 18 (2010).

2. **A. Burger, E. Cockayne, W. Gründlingh, C. Mynhardt, J. van Vuuren, and W. Winterbach**, *Infinite Order Domination in Graphs*, Journal of Combinatorial Mathematics and Combinatorial Computing **50**, (2004).

1. **A. Burger, E. Cockayne, W. Gruündlingh, C. Mynhardt, J. van Vuuren, and W. Winterbach**, *Finite Order Domination in Graphs*, Journal of Combinatorial Mathematics and Combinatorial Computing **49**, (2004).