# EVOLUTION, HIERARCHY AND MODULAR ORGANIZATION IN COMPLEX NETWORKS

A Dissertation

Submitted to the Graduate School
of the University of Notre Dame
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

by

Erzsébet Ravasz, M.S.

---

Albert-László Barabási, Director

Graduate Program in Physics
Notre Dame, Indiana
September 2004

# EVOLUTION, HIERARCHY AND MODULAR ORGANIZATION IN COMPLEX NETWORKS

Abstract

by

Erzsébet Ravasz

Large systems in nature and civilization share some important organizing principles uncovered in the framework of complex network research. Here we aim to present a few advances in understanding the generic topological characteristics of these systems. We start with an introduction to basic concepts of network research, continuing with a repertoire of well studied network examples and a brief history of previous modelling efforts. Next, we present a detailed investigation of scientific collaboration networks, with special focus on the role of internal links in determining the networks's scaling properties, and on limitations of certain measurements imposed by the database.

Many real networks in nature and society share two generic properties: they are scale free and they display a high degree of clustering. We show that the scale free nature and high clustering of real networks are the consequence of a hierarchical organization, implying that small groups of nodes form increasingly large groups in a hierarchical manner, while maintaining a scale free topology. In hierarchical networks the clustering coefficient follows a strict scaling law, which can be used to identify the presence of a hierarchical organization in real networks. We find that several real networks, such as the World Wide Web, actor network, the Internet at the domain level and the semantic web obey this scaling law, indicating that

hierarchy is a fundamental characteristic of many complex systems.

We then focus on the metabolic network of 43 distinct organisms and show that many small, highly connected topological modules combine in a hierarchical manner into larger, less cohesive units, their number and degree of clustering following a power law. Within *Escherichia coli* we find that the uncovered hierarchical modularity closely overlaps with known metabolic functions. We show that enzyme essentiality is not randomly distributed in the metabolic network, on the contrary, essential enzymes tend to cluster into a few small, well defined modules of the metabolism. Finally, we present an enzyme evolution-based model for metabolic network growth. This model reproduces the observed scale free and hierarchical organization of metabolic networks using local wiring rules.

I dedicate my thesis to my father, Ravasz József.

CONTENTS

FIGURES

# ACKNOWLEDGMENTS

I would like to first thank my advisor, Albert-László Barabási, for his constant support and guidance thorough my four years of graduate studies.

Further thanks are due to my undergraduate advisor, Zoltán Néda, for introducing me to research in physics and showing strong, continuing moral support and friendship.

Next is my research group, especially Zoltán Dezső, for being around to talk about research or the broader topic of doing research, and also for being there for me as a very old friend. I also benefited from kind guidance and constant help from Prof. Zolán Oltvai from Northwestern University and the postdocs in my group: Eivind Almaas, Stefan Wuchty, Marcio de Menezes and Alexei Vázquez. Last but not least I would like to mention my office mate Soon-Hyung Yook, our most helpful secretary, Suzanne Aleva, and my former office mates, Ginestra Bianconi and Réka Albert for their support and often welcome advice.

Completing a Ph.D. is not possible for me without other kinds of support. I would like to start with my fiancé, Peter Regan, in thanking him for bearing with me, helping me grow up some more and understand many aspects of the life I chose when I joined a Ph.D. program. He also deserves thanks for more practical assistance, from talking about research issues to proofreading my thesis.

I also wish to thank my family from home; my father, József, my mother, Erzsébet and my sister, Marek, for their support, their understanding, their power to arm me with new resources at each of my visits, and finally for their constant

interest in my research. I am further grateful to my friends from home: Kinga, Réka, Ana, Vetu, Levi, Ági and Ati, for making me feel like I was not really gone from home, and thus making it possible for me not to have regrets about my time here.

For all those who made these years enjoyable, who showed me many new worlds and offered friendships on many continents: warm thanks go to Hye-Young, Istán, Bhoopesh, Hilma, Igor, Lim and Smarajit, further to András, Sadia, Dylan, Mark, Audi, Claudio, Xochitl, Donny, Jason and Tatiana.

# CHAPTER 1

# INTRODUCTION TO NETWORKS

## 1.1 Motivation

Research into complex systems has had a long history of bottom-up approaches, which break the system into small or elementary constituents and map out interactions between these components. The development of science is marked by ever-deeper exploration of the constituent parts of the world around us, as well as the ways in which these parts assemble. The Standard Model describing elementary particles and the four types of interactions governing our world is perhaps the most successful example. Biology has developed a very detailed description of cellular components such as the DNA molecule or the various proteins and metabolites. Furthermore, many of the interactions that govern a cell's life have been investigated in great detail, including transcription of DNA, protein assembly and enzyme function. On the other hand, natural and social systems display characteristics that are fundamentally determined by their organization, emergent phenomena created by their interacting constituents. In many cases, if one takes a step back and does not focus on the variation in parts and interactions, a complex system as a whole is made up of an assemblage of generic elements and connections; in other words, it looks like a network [12, 58, 183]. For example, a cell's metabolism is maintained by a biochemical network, whose nodes are substrates and links are chemical reactions [89, 160, 109, 101, 98]. But equally complex webs describe human societies,

whose nodes are individuals and links represent social interactions [118, 208]; the World Wide Web (WWW) [13, 127, 115, 34], where nodes are Web documents connected by URL links; the scientific literature, whose nodes are publications and links are citations [175, 50, 176], or the language made of words and linked by various syntactic or grammatical relationships between them [189, 57, 73].

Due to the diversity and large number of the nodes and interactions, the topology of these evolving networks remained largely unknown and unexplored prior to the last decade. Yet, the inability of contemporary science to address the properties of complex networks limited advances in many disciplines, including molecular biology, computer science, ecology and the social sciences. The recent availability of system-level data on the network of interactions in large numbers of systems has opened the door for interdisciplinary research in fields where the behavior of the system as a whole is a central question. Recognizing generic organizational principles and order behind the diversity and apparent randomness of these different systems has certainly been a surprise along the way.

Two properties of real networks have generated considerable attention. First, many networks display a high degree of clustering, measured by the clustering coefficient [211] (the probability that a node's two first neighbors are also connected). Empirical results indicate that the clustering coefficient, averaged over all nodes, is significantly higher for many real networks than for a random network of similar size [211, 12, 58], and it is to a high degree independent of the number of nodes in the network [12]. At the same time, many networks of scientific or technological interest, ranging from the World Wide Web [13] to biological networks [101, 206, 98, 205] have been found to be scale free [20, 21]: there is no well-defined "connectivity scale" that approximates the degree (number of connections) of most nodes in the system. Instead, the distribution of degrees follows an inverse power law with expo-

nents between 2 and 3, indicating that these systems have very large connectivity fluctuations. Most nodes have one or two links, but there are a few hubs with very large degrees. The scale free property and strong clustering are not exclusive, but they coexist in a large number of real networks including metabolic webs [101, 206], the protein interaction network [98, 205], the world wide web [13] and some social networks [149, 147, 22].

As an introduction to the research that uncovered some of the generic properties of complex networks, we next summarize the main concepts used to characterize these systems. In particular, we offer a short account of specific real networks that the network community mostly focused on, some of which we examine in more detail in later chapters.

## 1.2 Properties of Complex Networks

The study of networks has its roots in graph theory in mathematics, going back to Euler, who studied some properties of a small graph defined by bridges in Königsberg.[1] Up to the landmark paper of Erdős and Rényi [64] in 1959, mathematics focused on the properties of large, ordered graphs. Erdős and Rényi introduced randomness for the first time to account for some of the properties of classes of networks. The community of physicists who started to investigate real networks such as the World Wide Web or scientific citations linking papers borrowed concepts from graph theory, helping them with the development of several new concepts that characterize network properties.

[1]The river Pregel crossing Königsberg has two islands caught between its branches, connected to the main land and each other by seven bridges. The puzzle of the town, asking if one can cross all the bridges without crossing the same one twice was resolved in 1736, in a short paper by Leonard Euler. He proved that such a walk does not exist, due to properties of the underlying graph [19].

*Degree Distribution*

Degree (or connectivity), the number of links $k$ a node has, is the most elementary property of a node. The overall graph is characterized by the *average degree*, $\langle k \rangle$. Yet, as noted above, the average degree does not capture the potential degree variations present in the network, which is better characterized by the degree distribution, $P(k)$, providing the probability that a node has exactly $k$ links.

*Clustering*

Nodes in many real systems exhibit a tendency to form tightly connected subgraphs. This property can be quantified by the clustering coefficient [211], a measure of the degree to which the neighbors of a particular node are connected to each other. For example, in a friendship network $C$ reflects the degree to which friends of a particular person are friends with each other as well. Formally, the clustering coefficient of node $i$ is defined as

$$C_i = \frac{2\, n_i}{k_i\,(k_i - 1)} \tag{1.1}$$

where $n_i$ denotes the number of links connecting the $k_i$ neighbors of node $i$ to each other. Accordingly, we can define the average clustering coefficient of a network as

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i. \tag{1.2}$$

*Distance Measures*

Processes taking place along the links of a network, such as package routing on the Internet, travelling via air or contacting a virus from an infected individual are often affected by the length of the paths between two nodes through the network. In most graphs, there are many paths connecting any two nodes $i$ and $j$, thus a useful distance measure is the length of the shortest path, $l_{ij}$. The mean shortest

path length, defined as

$$\langle l \rangle = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} l_{ij} \tag{1.3}$$

offers a measure of the network's navigability. Another distance measure of a network is its *diameter*, the largest distance between two nodes in the system. This quantity has similar scaling behavior to the average path length.

Networks that can be "crossed" by a small number of steps despite their often large size inspired the concept of *small world networks*, first illustrated on social networks.[2] This small world property characterizes most complex networks from actors in Hollywood [211] to coauthorship of scientific papers [149] or metabolites in a cell [206, 71].

*Degree Correlations*

Degree correlations describe some organizational properties of networks that the degree distribution does not address: Given a degree sequence of all the nodes, do high-degree vertices in a network preferentially associate with other high-degree vertices, or they are mainly connected to low-degree ones? This question has different types of answers depending on the level of detail one wishes to use to address it. The *degree correlation coefficient* [150] is a number between -1 and 1, representing the Pearson correlation coefficient[3] of the degrees at either ends of an edge. Networks in which hubs are preferentially connected to other hubs are called *assortative*, and have a positive degree correlation coefficient. Social networks tend to be assortative, while most of the networks in biology or communication tend to be *disassortatively*

[2] A surprising study conducted by social psychologist Stanley Milgram in 1967 [140] showed that two people in the United States have an average of *six degrees of separation* on the social acquaintance network, which makes our social universe a *small world*.

[3] The Pearson correlation coefficient of two arrays of numbers $X = \{x_1, x_2, \ldots, x_N\}$ and $Y = \{y_1, y_2, \ldots, y_N\}$ is defined as $r = \frac{\overline{(x-\overline{x})\cdot(y-\overline{y})}}{\sigma_x \cdot \sigma_y}$, where over-line stands for the average of the quantity under the line and $\sigma_x$ ( $\sigma_y$) is the standard deviation of $x_i$ ($y_i$).

mixed: hubs in these networks preferentially connect to smaller nodes [150, 151]. More detailed representations of degree correlations are given by the mean degree of the neighbors of a vertex as a function of its degree [164], or by two-dimensional histograms of the degrees of the vertices at the two ends of an edge [134, 135].

The degree distribution $P(k)$ and the $k$ dependence of the clustering coefficient, $C(k)$, often share generic features across different systems, allowing us to classify various networks. Parameters such as the average degree $\langle k \rangle$, average path length $\langle l \rangle$ and average clustering coefficient $\langle C \rangle$ characterize properties unique to the particular network under consideration, thus they are less generic than the degree distribution or the organization of clustering addressed in §4.

## 1.3 Networks Around Us

Advances in the field of complex networks facilitated by the increasing availability of electronic databases established a wide list of complex networks ranging over several disciplines. These networks serve as canonical test systems for new ideas in the field. A short account of some of them can help us lay groundwork for understanding the research that uncovered their generic features as well as ways to classify differences between them.

### 1.3.1 Social Networks

*Scientific Collaboration Networks*

Collaboration networks represent graphs of scientists (the nodes) who work(ed) together, coauthoring at least one publication (the link). Our best characterized databases are based on papers in biomedical research (from the MEDLINE archive), astrophysics, condensed matter and high-energy physics (from the Los Alamos E-print Archive and SPIRES) and computer science (from the NCSTRL Archive) over

a five-year window [149, 147, 148], as well as mathematics and neuroscience over a seven-year window [22]. All coauthorship networks display small world properties, high clustering and degree distributions consistent with power law tails.

*The Movie Actor Network*

The Internet Movie Database (`www.imdb.com`) is the source of one of the largest social networks open to study.[4] Based on all movies since the 1980's, the network has over $400,000$ actors as its nodes and movies that represent the links between them [211, 155]. The degree distribution of the actor network has a power law tail [20, 16, 11], and its clustering coefficient is much larger than that of a random network of similar size.

*The Web of Human Sexual Contacts*

Sexually transmitted diseases like AIDS spread on the subset of the social network described by sexual relationships [26, 102, 117, 131, 143, 169]. Although precise data about the links of this network is quite hard to collect, a few investigations have given us insights about its topology. Liljeros *et al.* [131] have estimated the degree distribution of the sex web using a survey about the number of sexual partners of 2810 Swedish individuals. Their investigation shows that the distribution of the number of sexual contacts of both both men and women follow power laws. This finding has been confirmed by a study based on data from the National Survey of Sexual Attitudes and Lifestyles in Great Britain, a sexual behavior study in rural Zimbabwe and a study targeting gay men in London [182].

This finding has a strong impact on epidemiological studies aimed at eradicating diseases spreading on sexual contact networks, as scale free networks with degree

[4]The number of actors from the two studies based on the content of the web page grew from $225,226$ nodes in 1998 [211] to $449,913$ nodes by 2000 [155].

exponents under 3 were found to allow diseases with arbitrarily low virulence to stay endemic and to show no improvement upon random immunization of their nodes [165, 166, 51]. The exponents of the degree distributions are found to be between 2.5 and 3.5, depending on the sex of the considered individuals as well as the investigated time spans. There is a striking exception in the case of gay men, where degree exponents of the sexual contact networks range between 1.5 and 2.

*Networks Describing the Business World*

The business world is a territory where the definition and characterization of networks created and governed by the laws of economics are a new, expanding topic of interest. For example, world trade defines a network of countries connected via trade relationships, a small network with scale free degree distribution, and small world characteristics [187]. An interesting study of business boards of Fortune 1000 companies (the 1000 US companies with the largest revenues) has shown that both the network of company directors (connected by boards they both are members of) and the dual network of boards (connected by shared directors) have small world properties [48].

The large-scale vulnerabilities of banking systems are naturally related to the topology of mutual credit-relation networks. An empirical study of the Austrian bank network shows that its architecture falls in line with general observations about most of real networks: it has a scale free degree distribution, very small path length and strong clustering [33].

### 1.3.2 Technological and Communication Networks

Information exchange between people, companies or social groups of any kind leads to quickly evolving networks of communication. These graphs, like the World Wide Web, the e-mail or the phone-call networks are supported by physical infras-

tructures that link the communication devices and cover most of the inhabited parts of the planet. We first mention these technological graphs, followed by examples of modern communication networks.

*The Internet*

The Internet, a network of physical cables between computers, routers and other telecommunication devices, is one of the favorite models of complex network studies. Its topology defined at two different levels of detail is continuously mapped, and the huge number of nodes and links offer very good statistical grounds for the measurement of many network features. At the most basic level the nodes are routers, while edges are the physical connections between them. The Autonomous System (AS) level of the Internet is a coarse-grained view of this system, where each autonomous Internet domain (defined by local data routing, such as the whole network domain of the University of Notre Dame) is represented by a single node. Any two autonomous systems connected by a physical link are considered connected in the AS level representation.

Maps at both levels have been publicly available since 1999 [69, 85, 35, 41], when Faloutsos *et al.* [69] measured the degree distribution at both levels and concluded that both follow power laws. Further studies of these networks showed that they also display small world behavior with very small average path length (around 9 for the router, 3 for the AS level Internet), along with high clustering coefficients [41, 222, 164].

*Electronic Circuits*

Electronic circuits define networks of logical gates connected by current-carrying wires or junctions. These networks were also found to have, perhaps surprisingly, scale free degree distributions [72]. They have been the focus of research on special

subgraphs characteristic of different networks [141], as well as on the dynamics of information flow [49].

*Power Grids*

Power grids are networks of generators, transformers and substations linked by high-voltage transmission lines spanning a whole country or region, distributing electric current. Statistical studies on the power grid covering western states in the U. S. indicate that they are small world networks with relatively high average clustering coefficient and an exponential degree distribution [211, 209, 16]. Recent interest in vulnerabilities of the power grid has been triggered by extensive electricity blackouts which affected large regions of the eastern United States [10].

Transportation systems such as airline route networks [16, 25], roads [106], railways [124, 186] and pedestrian traffic [42] or naturally occurring ones such as river networks [52, 133, 177, 178] or blood vessels are further examples of networks sharing some similarity with the Internet or the power grid: they all span a region of physical space. The constraints enforced on their topology by the physical extent of the links lead to distinct topological properties, as we discuss in §4.

*E-mail Networks*

The topology of e-mail networks with electronic addresses as nodes and e-mails as the links has been investigated based on data stored in server log files [60, 88]. The importance of this communication network comes from its ability to spread viruses, a process similar to natural virus spreading along social interactions. Thus, the finding that e-mail networks have scale free degree distribution explains the surprising prevalence of old viruses, in spite of easy-access anti-virus software [165, 166, 51].

*Phone-call Networks*

The phone-call network connecting people who had long-distance conversation via AT&T (in the course of one day), a large directed graph mapped out by Aiello *et al.* [7, 8], was also found to have a power law degree distribution both for incoming and outgoing calls.

### 1.3.3 Information Networks

In this section we discuss a few networks that in some sense capture the way humans store, organize or structure information.

*The World Wide Web*

The World Wide Web (WWW) [94], often incorrectly referred to as the "Internet" is a huge network of web-pages linked by directed URL links [126, 127, 34, 77]. It is the largest available network, with $2 \cdot 10^8$ nodes [34], yet is also very typical in many of its properties: strong clustering and small world behavior with an average path length estimated to be around 16 [34, 13, 6, 4]. Moreover, both distributions of the ingoing and the outgoing links are power laws with scaling over more than five orders of magnitude [13, 123, 115, 6, 4]. In a coarse-grained network representation of the World Wide Web, each web domain like the whole `www.nd.edu` page system is represented as a node, while any hyperlink from a document in this domain to another domain defines an edge between them. This bird-eye view of the WWW also gives us a scale free network, and an even smaller cyber-world: the average path length of this graph is 3.1 [6, 4].

*Citation Networks*

Citation networks of different scientific areas reflect the way research papers build on previous knowledge. They can be constructed using on-line databases of scientific

papers; links of these networks are the references between them [50, 175, 185, 176]. These references are directed links, and studies of their topology indicate that the in-degree distribution of these networks follow power laws [50, 175], while the out-degree distribution has a well-defined maximum and an exponential tail [197].

*Language Networks*

Words in a human language can be linked in several ways [57, 73, 194]. Word co-occurrence networks hint at methods used by people to organize concepts while choosing them for communication [73, 74]. Defined as graphs of words linked if they appear no more than two words apart with a frequency higher that a chosen threshold, co-occurrence networks based on the British National Corpus (`http://info.ox.ac.uk/bnc/`) were found to have a degree distribution with two distinct regimes of power law scaling [73]. Semantic relations like hyponyms, hypernyms[5] and meronyms[6] define a different linguistic network. A large database (WordNet) containing $66,025$ nouns along with their semantic relationships allowed for a study that showed this network to also have scale free and small world characteristics [189]. A language network of words linked if they are synonyms [221] reflects our way of building concepts of different levels of generality. Perhaps not surprisingly, this abstraction of human language into a network also has a degree distribution with power law tail, along with a very high clustering coefficient.

We end our discussion of example non-biological networks with Table 1.1, summarizing a few of their properties along with some data-source references. A somewhat more detailed introduction to the biological networks relevant to recent systems

[5] A noun A is the *hyponyms* of B if it describes a subset of things corresponding to B, its *hypernyms*: coat is a hyponym of clothing.

[6] A *meronym* of a word describes a physical part of what is described by this word: button is a meronym of coat.

biology research follows.

TABLE 1.1. A LIST OF NON-BIOLOGICAL NETWORKS

| **Network** | **Size** | $\langle k \rangle$ | $\gamma_{in}/\gamma_{out}$ | **Refs.** |
|---|---|---|---|---|
| Coauthors, MEDLINE | $1,520,251$ | 18.1 | 2.5 | [149] |
| Coauthors, LANL | $52,909$ | 9.7 | 1.3 | [149] |
| Coauthors, SPIRES | $56,627$ | 173 | 1.03 | [149] |
| Coauthors, computer sci. | $11,994$ | 3.59 | 1.3 | [149] |
| Coauthors, neuroscience | $209,293$ | 11.54 | 2.1 | [22] |
| Coauthors, mathematics | $70,975$ | 3.9 | 2.5 | [22] |
| Movie actors | $212,250$ | 28.78 | 2.3 | [21] |
| Sexual contacts, Sweden | $2,810$ | – | 3.4 | [131] |
| World trade | 179 | 43 | 2.6 | [187] |
| Company directors | $7,673$ | 14.44 | – | [48] |
| Banks in Austria | 883 | 35.7 | 1.72/3.1 | [33] |
| Internet, AS | 10697 | 5.98 | 2.4 | [41] |
| Internet, Faloutsos | $3,888$ | 2.57 | 2.48 | [69] |
| Internet, Govindan | $150,000$ | 2.66 | 2.4 | [85] |
| Electronic circuits | $20,000$ | 2.0 | 2.1 | [72] |
| Power grid | $4,941$ | 2.67 | – | [211] |
| E-mail | $59,912$ | 2.88 | 1.49/2.03 | [60] |
| Phone-call | $53,000,000$ | 3.16 | 2.1/2.1 | [7] |
| WWW, Notre Dame | $325,729$ | 4.51 | 2.1/2.45 | [13] |
| WWW, Kumar | $4 \times 10^7$ | 7 | 2.1/2.38 | [123] |
| WWW, Broder | $2 \times 10^8$ | 7.5 | 2.1/2.72 | [34] |
| WWW, domains | $260,000$ | – | 1.94/– | [6] |
| Citation | $783,339$ | 8.57 | 3/– | [175] |
| Words, co-occurrence | $460,902$ | 70.13 | 2.7 | [73] |
| Words, synonyms | $22,311$ | 13.48 | 2.8/2.8 | [221] |

Note: A list of real networks along with a few of their properties, such as size, average degree $\langle k \rangle$, in- ($\gamma_{in}$) and out-degree ($\gamma_{out}$) exponents of the connectivity distribution. Expanded after [12, 153, 154].

## 1.4 Networks in Living Systems

Networks emerge in many disguises in biological systems, from food webs in ecology to various biochemical nets in molecular biology. In particular, the wide range of interactions between genes, proteins and metabolites in a cell are best suited for a network representation. During the last decade, genomics has produced an incredible quantity of molecular interaction data, contributing to maps of specific cellular networks. The emerging fields of transcriptomics and proteomics have the potential to contribute to the already extensive data sources provided by the genome wide analysis of gene expression at the mRNA and protein level [161, 40, 37].

Indeed, extensive protein–protein interaction maps have been generated for a variety of organisms including viruses [75, 137], prokaryotes, like *H. pylori* [171] and eukaryotes, like *S. cerevisiae* (baker's yeast) [97, 96, 184, 196, 79, 91, 98], *C. elegans* (worm) [130] and *D. melanogaster* (fruit fly)[83]. Beyond the current focus on uncovering the structure of genomes, proteomes and interactomes of various organisms, some of the most extensive data sets are the metabolic maps [160, 109], catalyzing an increasing number of studies focusing on the architecture of the metabolism [101, 71, 206].

### 1.4.1 Metabolic Networks

The metabolism of the cell is a collection of all of its chemical reactions, responsible for synthesizing all of its building blocks and for obtaining its energy. The structure of metabolic networks was addressed by two independent studies by Fell and Wagner [71, 206] and Jeong *et al.* [101]. Fell and Wagner assembled a list of chemical reactions representing the central routes of the energy metabolism and small-molecule building block synthesis in *E. coli* [71, 206]. A substrate graph was defined by the nodes representing all metabolites, two substrates being considered

linked if they occurred in the same reaction. They found the substrate graph to be scale free with glutamate, coenzyme A, 2-oxoglutarate, pyruvate and glutamine having the highest degree; substrates viewed as an evolutionary core of the *E. coli*. At the same time, Jeong *et al.* analyzed the metabolic networks of 43 organisms representing all three domains of life [101], finding that the power law degree distribution for both incoming and outgoing edges holds for organisms of all kingdoms. Furthermore, the average separation [101] between nodes as well as the average clustering coefficient [206, 71] has roughly the same value for all organisms under consideration, regardless of the number of substrates found in the given species. Interestingly, the ranking of the most connected substrates is largely identical for all organisms. A recent study comparing the system-level properties of metabolic networks in various organisms indicates that the structural features of these networks are more conserved than the components themselves [168, 215].

### 1.4.2 Protein Interaction Networks

Protein interactions offer another opportunity to study cellular networks, considering proteins as nodes and physical interactions (binding) as links. It has been shown that interaction networks of *S. cerevisiae* (baker's yeast) [196, 220, 219], *H. pylori* [171], *C. elegans* (worm) [207] and *D. melanogaster* (fruit fly) [83] proteins exhibit distinct scale free behavior [98, 205, 171, 83]. Although protein interaction data is derived from different sources and is retrieved by different methods, the emergence of the scale free property appears to be a robust feature. Scale free networks are vulnerable upon targeted attack on their highly connected nodes [14]. Therefore, mutations of highly interacting proteins are expected to be lethal for the cell, a prediction supported by explicit measurements [100].

### 1.4.3 Protein Domain Networks

Proteins are made of long amino acid chains that fold in a particular three-dimensional structure, which allows them to perform different functions. These three-dimensional structures are not completely different from protein to protein. Certain regions of the amino acid chain fold into tight helical structures called $\alpha$ *helices*. Another characteristic structure is the $\beta$ *sheet*, a planar arrangement in which consecutive amino acids are parallel to each other, with opposite orientation. These two main structural building blocks define the *secondary structure* or a protein. The $\alpha$ helices and $\beta$ sheets linked by short *turns* or *random coils* (regions of the chain which lack secondary structure) often assemble into structural subunits called *domains*, stable folds shared among many proteins.[7] These subunits are the functional building blocks of proteins: often each domain has a separate, well defined function such as binding a small metabolite, spanning the plasma membrane, containing the catalytic or the DNA-binding site, or providing a surface to bind specifically to another protein.

The domain architecture of proteins was studied by considering protein domains as nodes and their co-occurrence in proteins as links [216, 17, 217], documenting again the emergence of a scale free architecture. Domains which appear in cellular functions crucial for the maintenance of multi-cellular organisms, such as signal transduction and cell-cell contacts, were found to be the most connected. Interestingly, as one looks at organisms of increasing complexity, the slope of the degree distribution of their domain networks is found to decrease. Similarly, interactions of domain families generated from sequence and structural data [162, 217] revealed that highly connected domains on sequence level appear to be the most frequently

[7] Representative examples are the *zinc finger*, a protein domain that binds to DNA, made of two $\beta$ sheets and an $\alpha$ helix, or the *helix-turn-helix* domain that also binds to DNA.

interacting as well.

### 1.4.4 Genetic Regulatory Networks

A living cell receives signals and reacts to information received from its environment, such as presence of food or toxins. Similarly, different developmental processes ask for a different set of proteins performing different functions during a cell cycle. The machinery at the heart of a living system that integrates information and governs the cellular processes is the genetic regulatory network. A few genes (typically a few hundred out of a few thousands in a simple organism) are responsible for coding proteins with regulatory functions. These genes are copied into messenger RNA and translated into transcription factor proteins. These regulatory proteins attach to DNA upstream another gene, where their presence helps or represses the binding of the protein complex (RNA polymerase) which copies the gene code into messenger RNA. By thus increasing or decreasing the mRNA level of this regulated gene, the transcription factor is responsible for increase or decrease in the number of proteins encoded by the regulated gene. In other words, one gene is able to turn the biological function encoded by another gene on or off. The network made of genes as nodes and genetical regulatory interactions as links is a dense information processing network, the coordinator of a cell's life. Many of these interactions are known in several organisms, but no precise method yet exists for a system-level mapping of the full network [128]. Nonetheless, some on-line databases summarize the information collected by individual experiments [180, 46], allowing the reconstruction of partially complete networks in *E. coli* and *S. cerevisiae* (baker's yeast) regulation [141, 188]. Generic statistical features we have seen in the previous cellular networks, such as scale free degree distribution and high clustering, characterize these networks as well.

### 1.4.5 Neural Networks

The worm *C. elegans* is the only organism with a completely mapped neural network. It has 282 neurons and close to 2000 connections (synapses or a gap junctions) [212]. This small but quite dense network has an exponential degree distribution and quite high clustering coefficient [211, 16].

### 1.4.6 Functional Network of the Brain

Functional magnetic resonance imaging techniques can be used to measure the activity of different regions of the human brain. Correlations between these regions can define a functional network of brain sites connected by common patterns of activity. These networks are dynamic, and the details of their architecture is interesting for functional studies of the brain. Nonetheless, their large-scale organization is scale free, with high clustering coefficients [62].

### 1.4.7 Ecological Networks

Food webs are networks of species linked by predator–prey interactions. These networks have been mapped out in a few habitats by ecologists who use them to investigate interactions between different species [167]. A few independent studies on food webs of different sizes have shown that they are highly clustered, and the average path length between species is below 3 [213, 142, 38]. The nature of their degree distribution is unclear, mostly due to the small size of these systems. Some studies found power law [142], others exponential behavior [38, 39].

We summarize the properties of the aforementioned biological networks with Table 1.2.

Networks offer us a new way to categorize systems of very different origin under a single framework. This approach has uncovered unexpected similarities between the organization of various complex systems such as scale free degree distribution, small

TABLE 1.2. A LIST OF BIOLOGICAL NETWORKS

| **Network** | **Size** | $\langle k \rangle$ | $\gamma_{in}/\gamma_{out}$ | **Reference** |
|---|---|---|---|---|
| Metabolic, *E. coli* | 778 | 7.4 | 2.2/2.2 | [101] |
| Protein, *S. cerevisiae* (DIP) | 1870 | 2.39 | 2.4 | [98] |
| Protein, *S. cerevisiae* (UETZ) | 985 | 1.83 | 2.5 | [205] |
| Protein, *C. elegans* | 3024 | 3.66 | – | [130] |
| Protein, *D. melanogaster* | 4679 | 2.04 | 1.6 | [83] |
| Protein Domain Families | 876 | 9.32 | 1.6 | [162] |
| Protein Domain (PromDom) | 5995 | 2.33 | 2.5 | [216] |
| Protein Domain (Pform) | 2478 | 1.12 | 1.7 | [216] |
| Protein Domain (Prosite) | 13.60 | 0.77 | 1.7 | [216] |
| Genetic reg., *E. coli* | 423 | 2.59 | 1.3 | [188] |
| Neural, *C. elegans* | 282 | 14 | – | [212] |
| Brain sites, *H. sapiens* | 4891-31530 | 4.12-13.41 | 2 | [62] |
| Food web, Ythan estuary | 134 | 8.7 | 1.05 | [142] |
| Food web, Silwood park | 154 | 4.75 | 1.13 | [142] |

Note: A list of biological networks along with a few of their uncovered properties. We indicate the size of the network, its average degree $\langle k \rangle$. For directed networks we list both the in-degree ($\gamma_{in}$) and out-degree ($\gamma_{out}$) exponents, while for the undirected networks these values are identical. Expanded after [12, 154].

world behavior and high clustering coefficient. These common features indicate that the networks describing them are governed by generic organization principles and mechanisms. Understanding the driving forces which invest different networks with similar topological features enables statistical physics as well as systems biology to combine the numerous details about various complex systems into a single framework, offering means to address their structure as a whole.

# CHAPTER 2

# MODELLING REAL NETWORKS

In the past few years a series of network models have been developed to explain nontrivial generic properties of real-world networks, such as the small world property, scale free degree distribution or high degree of clustering. There is a rich literature of models aimed at both capturing properties common to most networks and explaining features particular to specific real networks. In this chapter we review two of the most influential models, followed by a detailed analysis of scientific collaboration networks.

## 2.1 The Erdős–Rényi Model

Following the founding work of Erdős and Rényi in the 1950's [64, 65, 67, 66], large networks with no apparent design principles were described as random graphs [29], a model proposed as the simplest and most straightforward realization of a complex network.[1]

According to the binomial model[2] of an Erdős–Rényi (ER) random graph [29, 64, 12], one starts with $N$ nodes and connect each pair of nodes with probability $p$, creating a graph with approximately $p\,N(N-1)/2$ randomly distributed links

[1] Detailed review of random networks is available in the books of Bollobás [29], Cohen [44] (focusing on the similarity between phase transitions and random graph theory) and Karoński and Rućinski [108] (focusing on the history of the topic).

[2] The original formulation of the Erdős–Rényi model in their original paper [64] is an ensemble of random graphs, all of which are equally probable realizations of the following algorithm: one randomly chooses $n$ edges of the $N(N-1)/2$ possible edges between $N$ nodes, and connects those only. This formulation is equivalent to the binomial ER model.

(Fig. 2.1*a,b*). The majority of graphs generated in this manner have a few distinctive properties, some of which are characteristic to real networks as well.

### 2.1.1 Emergence of a Giant Component

Erdős and Rényi showed that there is a a threshold probability value, $p_c$, which separates two classes of random networks. If $p < p_c$ where $p_c = 1/N$, the graph consists of small, isolated groups of nodes with roughly the same size. As $p$ reaches the threshold value $p_c$, suddenly a giant connected component emerges containing most nodes. In fact, the ER model is equivalent to infinite-dimensional percolation, from the same universality class as mean-field percolation [193].

### 2.1.2 Degree Distribution

A fairly accurate estimate of the degree distribution can be easily calculated if one assumes that the nodes of the graph are independent: node $i$ having the exact degree $k_i$ does not affect the possible degrees of any other node. This assumption is not strictly true due to the finite ways links can be arranged between $N$ nodes, nonetheless the full degree distribution derived by Bollobás [29] does not significantly differ from the independent-node approximation [12].

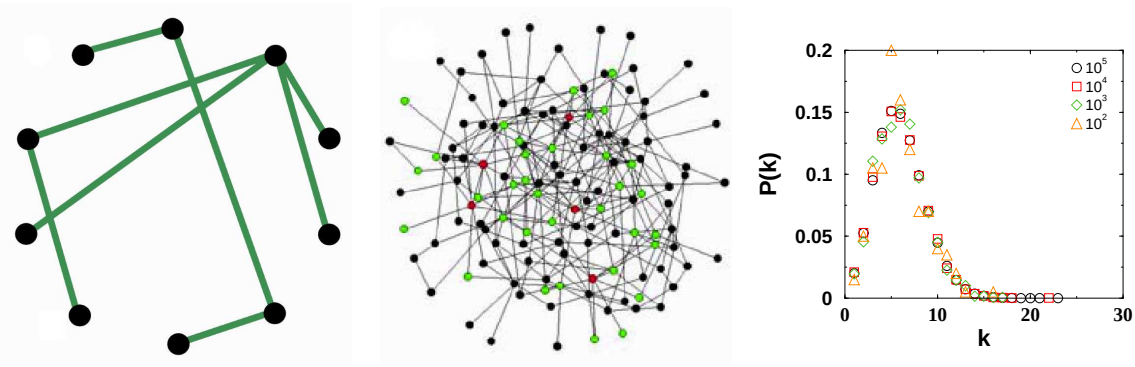
The probability that a node $i$ in the graph has $k$ links is the product of probabilities of having these $k$ links ($p^k$) and not having the rest of the possible $N-k-1$ links ($(1-p)^{N-k-1}$) weighted by the ways these $k$ links can be placed ($C^k_{N-1}$):

$$P_i(k) = C^k_{N-1}\, p^k\, (1-p)^{N-k-1}. \tag{2.1}$$

Assuming that all nodes have the same probability of having exactly $k$ links, the degree distribution of a random graph is described by equation (2.1), a binomial distribution with a Poisson distribution limit as $N \to \infty$:

$$P_{\mathrm{ER}}(k) = e^{-pN}\, \frac{(pN)^k}{k!} = e^{-\langle k \rangle}\, \frac{\langle k \rangle^k}{k!}. \tag{2.2}$$

The Poisson degree distribution indicates that random networks are fairly homogeneous (Fig. 2.1*b,c*): it was proved that even for high values of $p$ the maximum degree is the same order of magnitude as the average degree [12]. Almost all networks described in §1, however, are characterized by power law degree distributions, a property the ER model does not account for. Random connectedness does not allow the emergence of hubs, nor the very large number of low-connectivity nodes.




**(a)** Construction **(b)** A small ER network **(c)** Degree distribution

Figure 2.1. Erdős–Rényi random networks. **(a)** The model is constructed by laying down $N$ nodes and connecting each pair of nodes with probability $p$. The figure shows a particular realization of such a network for $N = 10$ and $p = 0.2$. **(b)** The random network generated by the Erdős–Rényi model is rather homogeneous, i.e. most nodes have approximately the same number of links. (From [14]) **(c)** The degree distribution $P(k)$ is strongly peaked at $k = \langle k \rangle$ and decays exponentially for large $k$.

### 2.1.3 Average Path Length

The average path length of a random network can be estimated using the approximation that all nodes in the network have $\langle k \rangle$ connections. Thus, the number of nodes one step away from any node is $\langle k \rangle$, two steps away there are roughly $\langle k \rangle^2$ nodes and so on. We can reach all $N$ nodes in $\langle l \rangle$ steps on average, leading to $N = \langle k \rangle^{\langle l \rangle}$. Thus, the average path length of a random network is:

$$\langle l \rangle_{\mathrm{ER}} \simeq \frac{\ln N}{\ln \langle k \rangle}, \tag{2.3}$$

with only logarithmical dependence on the system size.

The diameter of a random graph [43] has similar properties as the average path length, furthermore for most values of $p$ almost all possible graphs have a diameter very close to the value estimated in equation (2.3).

We have seen in §1 that most real-world networks display small world behavior: their average path length turns out to be very close to the average path length of random graphs with the same size (same number of nodes and links) [12]. The ER model elegantly accounts for the small world property of real networks.

### 2.1.4 Clustering Coefficient

Almost all complex networks mentioned in §1 display strong clustering of their nodes. However, in a random graph any two first neighbors of a node are linked with probability $p$, leading to

$$\langle C \rangle_{\mathrm{ER}} = p = \frac{\langle k \rangle}{N}. \tag{2.4}$$

This value is typically much smaller than the actual clustering coefficients estimated for real networks (see Fig. 9 in [12]). Random networks are not just homogeneous in their node degree values, they also lack tight subgraphs. On the other hand, real networks are very heterogeneous in both the degree values and clustering of their nodes.

The Erdős–Rényi model has guided our thinking about complex networks for decades. The growing interest in complex systems prompted many scientists to ask whether these systems share some organizing principles apart from randomness. We have seen that the topology of the networks underlying many complex systems systematically deviates from a random graph, a good indication of similar organization.

## 2.2 scale free Network Models

An important starting point in the scientific quest of organizing principles that govern different real-world networks is a 1999 paper by Barabási and Albert [20]. They were the first to notice what is today considered a trivial property of many real networks: they have scale free degree distributions [19]. Aiming to explain the emergence of these topologies observed both in the World Wide Web map [13], the movie actor network and the neural network of the *C. elegans* worm, Barabási and Albert point out the importance of the evolution of real networks. The model they proposed was fundamentally different from the Erdős–Rényi approach or other existing network models,[3] due to the fact that they looked at the emergence of network connectivity patters as a result of the way real networks grow. This shift of approach not only allowed them to construct a model that shows how scale free graphs can emerge, but it also opened the door for understanding the specific mechanisms that governed the growth of many of the investigated real networks.

### 2.2.1 The Barabási–Albert Model

The first principle behind the philosophy of the Barabási–Albert (BA) scale free model is that real networks grow constantly through addition of new nodes that link to nodes already present in the system. Second, in most real networks there is a higher probability for a new-coming node to link to the existing nodes that already have large number of connections, a property called *preferential attachment* [20, 21].

[3]The observation that real-world networks show strong clustering, unlike the complete lack of local organization of random networks, have inspired Watts and Strogatz to introduce a network model which reconciles clustering properties typical of regular lattices with the small world property of random networks [211]. They start with a circle of nodes connected to their first $k$ neighbors, then they pick a random fraction $p$ of the links and reconnect one end to any randomly chosen node. For a wide parameter range these networks become small world (i.e. their average path length is small) while still retaining a high clustering coefficient. Nonetheless, they are homogeneous in their connectivity: they have a Poisson-like degree distribution, failing to reproduce the hubs characteristic to real networks.

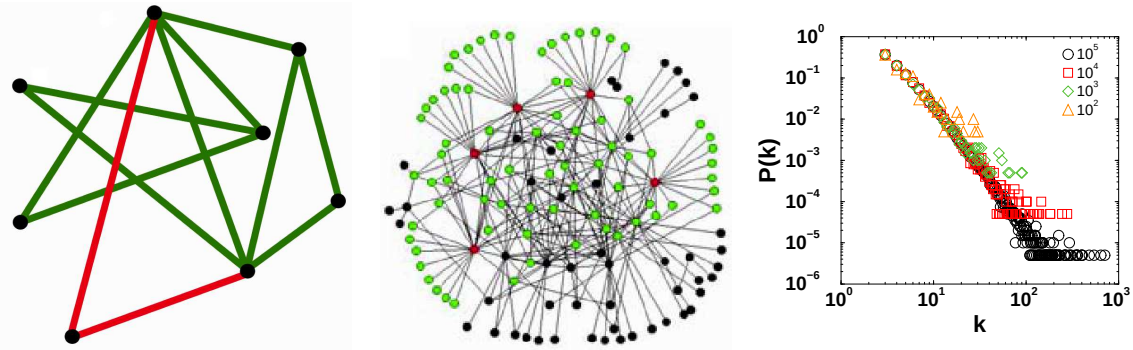
Indeed, we link with higher probability to a more connected (thus better known) document on the WWW, or we tend to repeatedly cite much cited papers. These two ingredients, growth and preferential attachment, inspired the BA model which leads to networks with power law degree distribution.

The algorithm of the model goes as follows [20, 21] (see Fig. 2.2*a*):

- **Growth.** Starting with a small core of $m_0$ nodes, at every time step a new node is added with $m \leq m_0$ edges, that connect the new node to $m$ different nodes already present in the system.

- **Preferential attachment.** The $m$ nodes to which the new one connects to are chosen with probabilities $\Pi_i$ proportional to their degree, $k_i$:

$$\Pi_i = \frac{k_i}{\sum_j k_j}. \tag{2.5}$$




**(a)** Construction **(b)** A small BA network **(c)** Degree distribution

Figure 2.2. Barabási–Albert scale free network. **(a)** The model is constructed from $m_0$ initial nodes by subsequent addition of new nodes that connect to existing ones with probabilities proportional to their degree. **(b)** The scale free network generated by the BA model is not homogeneous, hubs with degrees orders of magnitude larger than the average coexist with many very low-connectivity nodes. (From [14]) **(c)** The $P(k)$ degree distribution follows a power law, $P(k) \sim k^{-\gamma}$, with degree exponent $\gamma = 3$.

Networks generated by this growth process have a few hubs with degrees orders of

magnitude larger than the average degree (Fig. 2.2*b,c*), and their degree distribution is a power law with degree exponent $\gamma = 3$.

### 2.2.2 Mean-field Calculation of the Degree Distribution

The continuum approach used in [20, 21] to calculate the degree distribution of a BA network follows the time-dependence of individual node degrees during network growth. It approximates the degree of a node with a continuous variable changing in time at a rate proportional to $\Pi(k_i)$ times the number of newly created links:

$$\frac{\partial k_i}{\partial t} = m\,\Pi(k_i) = m\,\frac{k_i}{\sum_{j=1}^{N-1} k_j}. \tag{2.6}$$

The sum in the denominator is twice the number of edges at time $t$: $\sum_{j=1}^{N-1} k_j = 2mt$, thus

$$\frac{\partial k_i}{\partial t} = \frac{k_i}{2t}. \tag{2.7}$$

Using the initial conditions that each node $i$ is introduced at time $t_i$ with $k_i(t_i) = m$ links, we have

$$k_i(t) = m\sqrt{\frac{t}{t_i}}. \tag{2.8}$$

Further, the probability that a node has a degree $k_i(t) < k$ is given by:

$$P\left(k_i(t) < k\right) = P\left(t_i > \overline{t}\,\right) \qquad \text{where} \qquad \overline{t}\, = \frac{m^2 t}{k^2}. \tag{2.9}$$

The probability that a node entered the system at the time-step $t_i$ is the same for all nodes,

$$P(t_i) = \frac{1}{m_0 + t}, \tag{2.10}$$

thus the probability that $t_i$ falls in the $(\overline{t}, t)$ interval is:

$$P\left(t_i > \overline{t}\,\right) = P\left(k_i(t) < k\right) = 1 - \frac{\overline{t}}{m_0 + t} = 1 - \frac{m^2 t}{k^2(m_0 + t)}. \tag{2.11}$$

The degree distribution $P(k)$ can be obtained from equation (2.11) using

$$P(k) = \frac{\partial P(k_i(t) < k)}{\partial k} \quad \Rightarrow \tag{2.12}$$

$$P_{\mathrm{BA}}(k) = \frac{2m^2 t}{m_0 + t}\,\frac{1}{k^3}, \tag{2.13}$$

in good agreement with measurements from large simulated networks [21, 12].

There are other mean-field approaches which lead to similar formulas of the degree distribution of the BA model [58], such as the master-equation approach proposed by Dorogovtsev, Mendes and Samukhin [59] or the rate-equation approach introduced by Krapivsky, Redner and Leyvraz [121]. These approaches are equivalent, and offer the same asymptotic results as the continuum theory. The exact solution of the degree distribution for the Barabási–Albert model has been worked out by graph theorists Bollobás and Riordan [31].

### 2.2.3 Properties of the Barabási–Albert Model

*Average Path Length*

In §2.1.3 we have seen that random networks display small world property and their average path length scales as the logarithm of $N$. Analytical results have also been obtained for the Barabási–Albert network [32, 45, 43] showing that they are "ultra-small:" the average path length scales as

$$\langle l \rangle_{\mathrm{BA}} \sim \ln(\ln N). \tag{2.14}$$

This result actually extends beyond the Barabási–Albert model and it was shown to hold for any large scale free network with a degree exponent between 2 and 3, a range of exponents covering most of the studied real-world networks.

*Clustering Coefficient*

We have seen that most real-world networks have large average clustering coefficients due to the abundance of tight subgraphs in them. However, the construction of the Barabási–Albert model is very homogeneous when it comes to subgraphs: there is no mechanism other than chance by which a network built using the BA

model develops many tightly connected neighborhoods. Indeed, analytical calculation of the clustering coefficient [116, 30] shows that

$$\langle C \rangle_{\rm BA} \sim \frac{(\ln N)^2}{N}. \tag{2.15}$$

Thus the clustering coefficient of the BA model decreases with the system size, although this decrease is slower than that of a random network. The model was not aimed to capture the inhomogeneity of the tight subgraphs existent in real networks.

### 2.2.4 Comments on Preferential Attachment

The Barabási–Albert model uses a very simple, linear preferential attachment rule for network evolution: $\Pi(k) \sim k$. Real systems have many different phenomena affecting this rule, from aging (or saturation) of nodes [16, 54], to internal edge addition,[4] removal [56, 22], rewiring [11], or existence of nodes with different inherent abilities (fitness) to compete for links [27, 28, 68]. Nonlinearities in the attachment rule result in deviations from power law degree distribution [121]. Krapivsky, Redner, and Leyvraz [121] showed that sub-linear preferential attachment (as observed in the movie actor network [11] or in the neuroscience coauthorship graph [22, 99]) can lead to a stretched exponential degree distribution, while a super-linear attachment rule leads to "winner takes all" scenario, with a central node taking a significant fraction of the edges. Strictly scale free topologies only emerge as a result of linear preferential attachment.

Networks for which the time each node joined the network is known (coauthorship networks, the citation network, the actor collaboration network and the AS level Internet [99, 146, 164]) allow a direct measurement of the $k$–dependence of the $\Pi(k)$ attachment rule. The $\Pi(k_i)$ probability that a node $i$ with $k_i(t)$ degrees at

[4] For a detailed account of the dynamics of constant internal link addition in scientific collaboration networks see §3 [22].

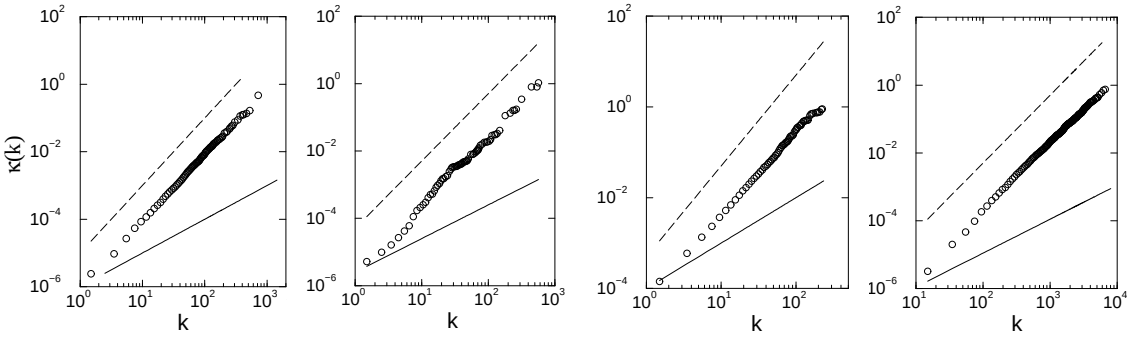
time $t$ acquires the next incoming link can be approximated by the $k_i(t+\Delta t)-k_i(t)$ change of its degree in a time window $\Delta t$ much shorter than the total lifetime of the network, divided by the total number of links added to the system in this $\Delta t$ time window, $\Delta k$:

$$\Pi(k_i) \sim \frac{k_i(t+\Delta t)-k_i(t)}{\Delta k}. \tag{2.16}$$

To reduce the fluctuations on the data from the four above mentioned networks, Jeong *et al.* calculated the cumulative preferential attachment, defined as

$$\kappa(k) = \sum_{k_i=0}^{k} \Pi(k_i). \tag{2.17}$$

Figure 2.3 indicates that real networks show behavior close to the linear preferential attachment assumed by the BA model. However, while for the Internet [99, 164] and the citation network [99], $\Pi(k)$ depends linearly on $k$, for the neuroscience collaborations and the movie actor network this dependence is sub-linear with $\alpha = 0.8 \pm 0.1$ [99].




**(a)** Citations **(b)** Internet **(c)** Neuroscience **(d)** Movie actors

Figure 2.3. Measurement of the cumulative preferential attachment. Dashed line corresponds to linear preferential attachment, the continuous line to no preferential attachment. (From [99])

### 2.2.5 Consequences of the scale free Topology

An important consequence of the existence of hubs in scale free networks is that these systems exhibit high tolerance to random perturbations but are sensi-

tive to targeted attack on the highly connected nodes [14]. Accordingly, failure of randomly selected nodes cannot destroy the network's integrity. However, the systematic removal of the hubs will rapidly fragment the network. This feature is of particular importance for biological systems, since it reflects the biochemical network's resilience against random mutations. Therefore, highly connected nodes in biochemical networks might be potential candidates for drug targets. The presence of hubs in a scale free network has a fundamental impact on virus spreading as well. Classical epidemiological models predict that infectious diseases with transmission probability under an epidemic threshold will inevitably die out. However, in scale free networks the epidemic threshold is reduced to zero [165]. Thus, as some social and sexual networks are known to exhibit a scale free topology [131], even extremely weakly infectious viruses can spread and prevail, making random immunization ineffective.

The scale free model uncovers some of the origins of the inhomogeneities observed in real networks, shedding light on the mechanisms by which hubs appear and play central role in the navigability and robustness of these systems. The next step in our journey towards a more complete understanding of real network topology is presented in more detail in chapter 4, preceded by a detailed investigation of the evolution of scientific collaboration networks.

# CHAPTER 3

# MODELLING SCIENTIFIC COLLABORATION NETWORKS

## 3.1 Motivation

One of the most prolific mathematicians of all time, Paul Erdős has written over 1400 papers with over 500 coauthors. This unparalleled productivity inspired the concept of the Erdős number, which is defined to be one for his many coauthors, two for the coauthors of his coauthors and so on. The tightly interconnected nature of the scientific community is reflected by the conjecture that all publishing mathematicians, as well as many physicists and economists have rather small Erdős numbers [1]. The coauthorship networks is of general interest for understanding the topological and dynamical laws governing complex networks, as it represents the largest publicly available computerized social network.

Social networks have been much studied in social sciences [208, 118]. A general feature of these studies is that they are restricted to rather small systems, and often view networks as static graphs, whose nodes are individuals and links represent various quantifiable social interactions. In contrast, recent approaches with methodology rooted in statistical physics focus on large networks, searching for universality both in the topology of the graph and in the dynamics governing its evolution. In addition to uncovering generic properties of real networks, these studies signal the emergence of a new set of modelling tools that considerably enhance our ability to characterize and model complex interactive systems. To illustrate the power of these

advances we choose to investigate the collaboration network of scientists in detail.

Newman has taken an important step towards applying modern network ideas to collaboration networks [149, 147, 148] by studying several large databases, focusing on several fields of research over a five-year period. He established that collaboration networks have all the general ingredients of small world networks: they have a surprisingly short node-to-node distance and a large clustering coefficient [149], much larger than the one expected from a random Erdős-Rényi type network of similar size and average connectivity. Furthermore, the degree distribution appears to follow a power law [147, 148].

Our study takes a different, but complementary approach to collaboration networks than that followed by Newman. We view collaboration networks as prototype of *evolving* networks, where the accent is on dynamics and evolution. Indeed, the coauthorship network constantly expands by the addition of new authors to the database, as well as the addition of new internal links representing papers coauthored by authors that were already part of the database. The topological properties of these networks are determined by these dynamical and growth processes. Consequently, in order to understand their topology, we first need to understand the dynamical process that determines their evolution. In this aspect Newman's study focuses on the static properties of the collaboration graph, while our work investigates the dynamical properties of these networks. We show that such dynamical approach can explain many of the static topological features seen in the collaboration graph.

It is important to emphasize that the properties of the coauthorship network are not unique. The WWW is also a complex evolving network, where nodes and links are added (and removed) at a very high rate, the network topology being profoundly determined by these dynamical features [20, 13, 127, 126]. The actor

network of Hollywood is very similar to the coauthorship network, because it grows through the addition of new nodes (actors) and new links (movies linking existing actors) [209, 211, 11]. Similarly, the nontrivial scaling properties of many cellular, ecological or business networks are all determined by dynamical processes that contributed to the emergence of these networks. So why single out the collaboration network as a case study? A number of factors have contributed to this choice. First we needed a network for which the dynamical evolution is explicitly available. That is, in addition to a map of the network topology, it is important to know the time at which the nodes and links have been added to the network, crucial for revealing the network dynamics. This requirement reduces the currently available databases to two systems: the actor network, where we can follow the dynamics by recording the year of the movie release, and the collaboration network for which the paper publication year allows us to track the time evolution. Of these two, the coauthorship data is closer to a prototypical evolving network than the Hollywood actor database for the following reasons: in the science collaboration network the coauthorship decision is made entirely by the authors, i.e. decision making is delegated to the level of individual nodes. In contrast, for actors the decision often lies with the casting director, a level higher than the node. While in the long run this difference is not particularly important, the collaboration network is still closer in spirit to a prototypical evolving network such as social systems or the WWW.

Our work stands on three pillars. First, we use direct measurements on the available data to uncover the mechanism of network evolution. This implies determining the different parameters and uncovering the various competing processes present in the system. Second, building on the mechanisms and parameters revealed by the measurements we construct a model that allows us to investigate the large-scale topology of the system, as well as its dynamical features. The predictions offered by

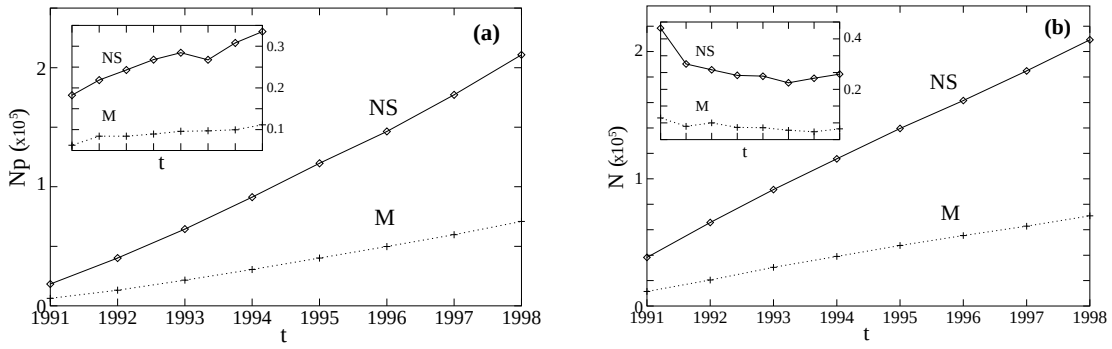
a continuum theory of the model allow us to explain some of the results that were uncovered by our as well Newman's measurements. The third and final step will involve computer simulations of the model, serving several purposes: (i) It allows us to investigate quantities that could not be extracted from the continuum theory; (ii) Verifies the predictions of the continuum theory; (iii) Allows us to understand the nature of the measurements we can perform on the network, explaining some apparent discrepancies between the theoretical and the experimental results.

## 3.2 Databases: Coauthorship in Mathematics and Neuroscience

In order to get information on the topology of a scientific coauthorship web one needs a complete data-set of the published papers, ideally from the birth of the discipline until today. However, computer databases cover at most the past several decades. Thus any study of this kind needs to be limited to only a recent segment of the database. This will impose unexpected challenges that need to be addressed, since such limited data availability is a general feature of most networks.

The databases considered by us contain article titles and authors of all relevant journals in the field of mathematics (M) and neuroscience (NS), published in the period 1991–98. We have chosen these two fields for several reasons. A first factor was the size of the database: biological sciences or physics are orders of magnitude larger, too large to address their properties with reasonable computing resources. Second, the selected two fields offer sufficient diversity by displaying different publishing patterns: in NS collaboration is intense, while mathematics, although there is increasing tendency towards collaboration [87], is still a basically single-investigator field.

In mathematics our database contains 70,975 different authors and 70,901 papers for an interval spanning eight years. In NS the number of different authors is




**(a)** Number of papers **(b)** Number of authors

Figure 3.1. **(a)** Cumulative number of papers for the M and NS databases in the period 1991–98. The inset shows the number of papers published each year. **(b)** Cumulative number of authors (nodes) for the M and NS databases in the period 1991–98. The inset shows the number of new authors added each year.

209,293 and the number of published papers is 210,750. A complete statistics for the two considered database is summarized in figure 3.1, where we plot the cumulative number of papers and authors for the period 1991–98. We consider "new author" an author who was not present in the database from 1991 up to a given year.

Before proceeding we need to clarify a few methodological issues that affect the data analysis. First, in the database the authors are represented by their surname and initials of first and middle name, thus there is a source of error in distinguishing some of them. Two different authors with the same initials and surname will appear to be the same node in the database. This error is important mainly for scientists of Chinese and Japanese descent. Second, a given author will occasionally use one or two initials in different publications, and in such cases he/she will appear as separate nodes. Newman [149] showed that the error introduced by those problems is of the order of a few percent. Our results are also affected by these methodological limitations, but we do not expect that it will have a significant impact on our results.

## 3.3 Data Analysis

In this section we investigate the topology and dynamics of the two databases, M and NS. Our goal is to extract the parameters that are crucial to the understanding of the processes which determine the network topology, offering input for the construction of an appropriate model.

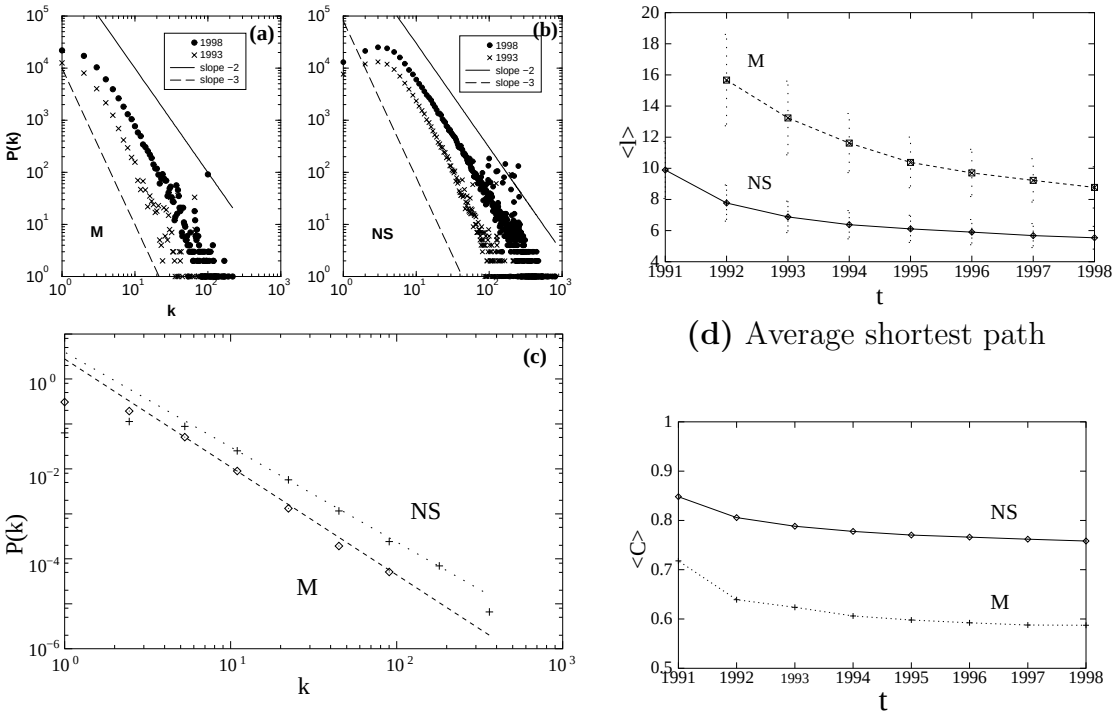
### 3.3.1 Degree Distribution Follows a Power Law

The degree distributions of both the M and NS data indicate that collaboration networks are scale free. The power law tail is evident from the raw, uniformly binned data (Fig. 3.2*a,b*), but the scaling regime is better seen on the plot that uses logarithmic binning, reducing the noise in the tail (Fig. 3.2*c*). The cumulative data with logarithmic binning indicates $\gamma_M = 2.4$ and $\gamma_{NS} = 2.1$ for the two databases. We will see in the coming sections that the data indicates the existence of two scaling regimes with two different scaling exponents. The combination of these two regimes could easily give the impression of an exponential cutoff in the $P(k)$ for large $k$. Further analysis, offered in §3.4 indicates that a consideration of two scaling regimes offers a more accurate description.

### 3.3.2 Average Shortest Path Length Decreases in Time

Determining the average shortest path length in a large network is a rather time-consuming procedure. Usually sampling a fraction of all nodes and determining their distance from all other points gives reasonable results. The results for the cumulative database are shown in figure 3.2*d*.

The figure indicates that $\langle l \rangle$ decreases in time, which is highly surprising because all network models so far predict that the average shortest path length should increase with system size [13, 29, 32]. The decreasing trend observed by us could




**(a–c)** Degree distribution

**(d)** Average shortest path

**(e)** Average clustering coefficient

Figure 3.2. Degree distribution of the **(a)** M and **(b)** NS coauthorship networks, showing data based on the cumulative results up to years 1993 (×) and 1998 (•). **(c)** Degree distribution shown with logarithmic binning computed from the full dataset cumulative up to 1998. The lines correspond do the best fits, and have the slope 2.1 (NS, dotted) and 2.4 (M, dashed). **(d)** Average shortest path length in the M and NS databases. $\langle l \rangle$ is computed on the cumulative data up to the indicated year. The error bars indicate the standard deviation of the distances between all pairs of nodes. **(e)** Clustering coefficient of the M and NS database, determined for the cumulative data up to the year indicated on the $t$ axis.

have two different origins. First, it is possible that indeed, shortest paths do shrink as new internal links are added, i.e. papers written by authors that were previously part of the database. They increase interconnectedness, thus decreasing the average path length. Second, the decreasing path length could be a consequence of the fact that we do not have access to the full database, but only starting from year 1991. As we demonstrate in §3.4.2, such incomplete data sets could result in an apparently decreasing average path length even if this length actually increases for the full sys-

tem. The slow convergence indicates that an even longer time interval is perhaps needed to reach the asymptotic limit, in which different relevant quantities take on a stationary value. The smaller average path length for the NS field is expected, since mathematicians tend to work in smaller groups and write papers with fewer coauthors.

### 3.3.3 Clustering Coefficient Decays with Time

The clustering coefficient of a node in the coauthorship network tells us how much a node's collaborators are willing to collaborate with each other, and it represents the probability that two of its collaborators wrote a paper together. The clustering coefficient for the cumulative network as a function of time is shown in figure 3.2*e*.

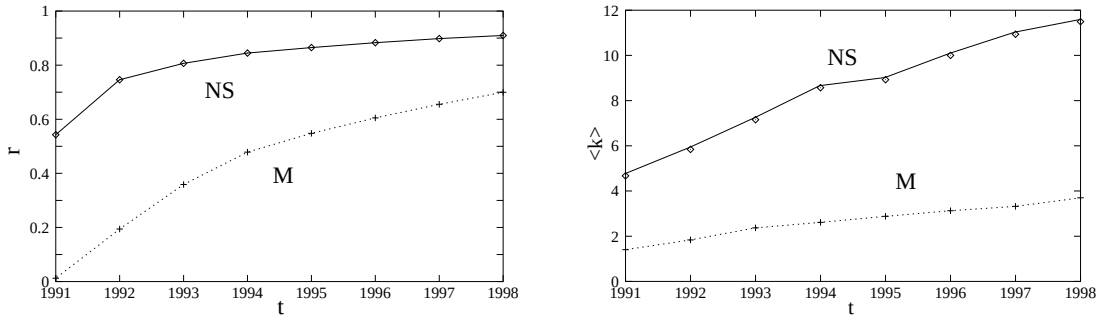
The results, in agreement with average path-length measurements, suggest a stronger interconnectedness for the NS compared with M, and a slow convergence in time to an asymptotic value.

### 3.3.4 Relative Size of the Largest Cluster Increases

It is important to realize that the collaboration network is fragmented in many clusters. There are several reasons for this. First, in every field there are scientists who do not collaborate at all, that is they are the only authors of all papers on which their name appears. This is more frequent in mathematics, which despite an increasing tendency toward collaboration [87], is still more fragmented than physics or neural science. Second, and most important, the database contains papers published only after 1990. Thus there is a possibility that two authors coauthored a paper before 1990, but in our database they appear as disconnected.

If we look only at a single year, we see many isolated clusters of authors. The cumulative data-set containing several years develops a giant cluster that contains a large fraction of the authors. To investigate the emergence of this giant connected

component we measured the relative size of the largest cluster, $r$, giving the ratio between the number of nodes in the largest cluster and the total number of nodes in the system. A cluster is defined as a subset of nodes interconnected by links. Results from our cumulative coauthorship networks are presented in figure 3.3*a*. As expected, in M the fraction of clustered researchers is considerably smaller than in NS.




**(a)** Largest cluster **(b)** Average connectivity

Figure 3.3. **(a)** Relative size of the largest cluster for the M and NS database. **(b)** Average number of links per node, $\langle k \rangle$. Results are computed on the cumulative data up to the given year.

The continuous increase in $r$ may appear as the scenario commonly described as percolation [36] or the much studied emergence of the giant component in random networks [29, 32]. However, the process leading to this giant cluster is fundamentally different from these much studied phenomena. In most research fields, apart from a very small fraction of authors that do not collaborate, all authors belong to a single giant cluster from the very early stages of the field. That is, the system is almost fully connected from the very first moment. The only reason why the giant cluster in our case grows so dramatically in the first several years is that we are missing the information on the network topology before 1991. A good example is the actor network, where the huge majority of the actors are part of the large cluster at any

stage of the network, starting from early 1900's until today. However, if we would start recording collaborations only after 1990 for example, the data would indicate, incorrectly, that many actors are disconnected. The increasing $r$ indicates only the fact that we are reconstructing the already existing giant cluster, and it is only a partial measure of its emergence.

Finally, the fast convergence of the NS cluster size to an approximately stationary value around 0.9 indicates that after 1994 the network reached a roughly stationary topology, i.e. the basic alliances are uncovered. This does not seems to be the case for M, where after ten years $r$ still increases, perhaps due to smaller publication and collaboration rate in the field.

### 3.3.5 Average Degree Increases

With time the number of nodes in our coauthorship network increases due to arrival of new authors. The total number of links also increases through the connections made by new authors with old ones and by new connections between old authors. A quantity characterizing the network's interconnectedness is the average degree $\langle k \rangle$, giving the average number of links per author. The time dependence of $\langle k \rangle$ for the cumulative network is shown in figure 3.3*b*, indicating an approximately linear increase of $\langle k \rangle$ with time. This is a rather important deviation from the majority of currently existing evolving network models, that assume a constant $\langle k \rangle$ as the network expands. As expected, the average degree for M is much smaller than for NS.

### 3.3.6 Node Selection is Governed by Preferential Attachment

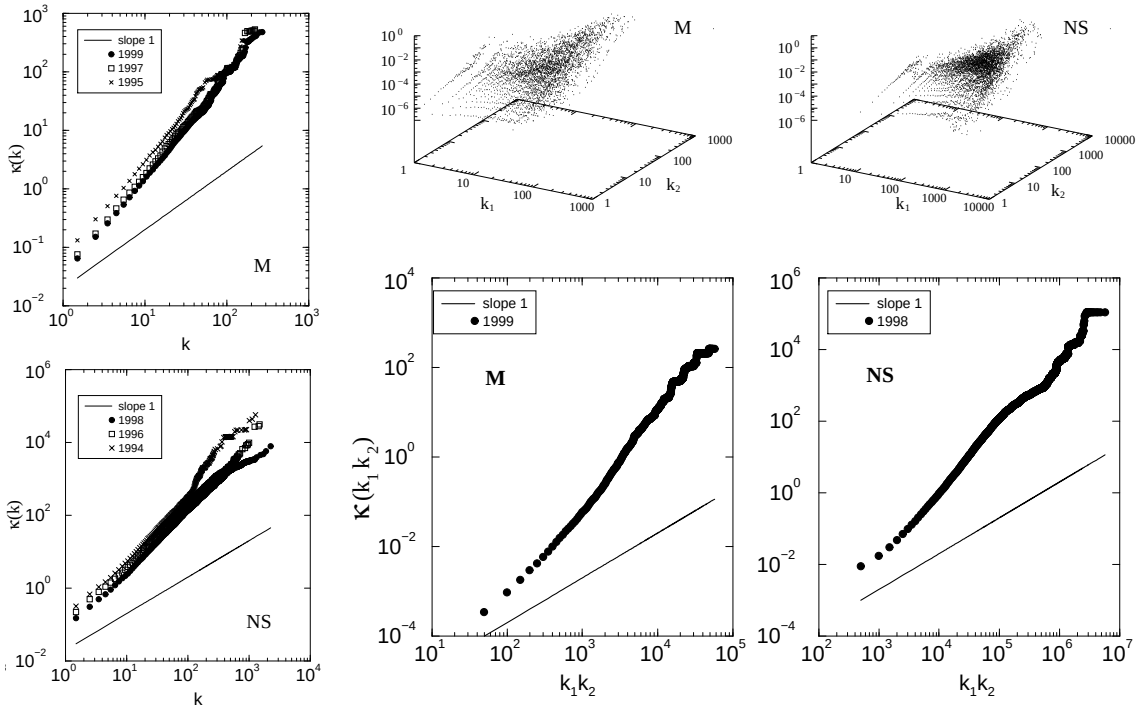
The availability of dynamic data on the network development allows us to investigate the presence of preferential attachment in the coauthorship network at two levels.

- *New nodes:* For a new author who appears for the first time on a publication, preferential attachment has a simple meaning: it is more likely that the first paper will be coauthored with somebody that already has a large number of coauthors (links) than with somebody less connected. As a result "old" authors with more links will increase their number of coauthors at a higher rate than those with fewer links. As figure 3.4*a* shows, we find that $\kappa(k)$ (see equation (2.17)) is nonlinear, increasing as $\kappa(k) \sim k^{\nu+1}$, where the best fit gives $\nu \simeq 0.8$ for M and $\nu \simeq 0.75$ for NS. This implies that $\Pi(k) \sim k^{\nu}$, where $\nu$ is different from 1 [21]. As simulations have shown, such nonlinear dependence generates deviations from a power law $P(k)$ [21]. This was supported by analytical calculations [121, 120], demonstrating that the degree distribution follows a power law only for $\nu = 1$. The consequence of this nonlinearity will be discussed in §3.4.

- *Internal links:* A large number of new links appear between old nodes as the network evolves, representing papers written by authors that were part of the network, but did not collaborate before. Such internal links are known to effect both the topology and dynamics of the network [56]. These internal links are also subject to preferential attachment. We studied the probability $\Pi(k_1, k_2)$ that an old author with $k_1$ links forms a new link with another old author with $k_2$ links. The $\Pi(k_1, k_2)$ probability map can be calculated by dividing $N(k_1, k_2)$, the number of new links between authors with $k_1$ and $k_2$ links, with the $D(k_1, k_2)$, number of pairs of nodes with degrees $k_1$ and $k_2$ present in the system:

$$\Pi(k_1, k_2) = \frac{N(k_1, k_2)}{D(k_1, k_2)}. \tag{3.1}$$

  The three-dimensional plot of $\Pi(k_1, k_2)$ is shown in figure 3.4*b*, the overall behavior indicating preferential attachment: $\Pi(k_1, k_2)$ increases as either $k_1$ or




**(a)** Incoming nodes **(b)** Internal nodes

Figure 3.4. **(a)** Cumulative preferential attachment of newly added nodes, $\kappa(k)$. In the absence of preferential attachment $\kappa(k) \sim k$, shown as continuous line on the figures. **(b)** Internal preferential attachment $\Pi(k_1, k_2)$ shown on the 3D plots (top); cumulative internal preferential attachment, $\kappa(k_1 k_2)$ as a function of the $k_1 k_2$ product (bottom plots). The straight lines have slope 1, expected if there would be no preferential attachment. All results were computed by considering the new nodes coming in the specified year, and the network formed by nodes already present up to this year.

$k_2$'s increase.

A natural hypothesis is to assume that $\Pi(k_1, k_2)$ factorizes into the product $k_1 k_2$. As figure 3.4*b* shows, we indeed find that

$$\kappa(k_1 k_2) = \int_1^{k_1 k_2} \Pi(k_1' k_2')\, d(k_1' k_2') \tag{3.2}$$

can be well approximated with a slope 2 as a function of $k_1 k_2$, indicating that for internal links the preferential attachment is linear in the degree.

## 3.4 Modelling the Web of Science

In this section we use the obtained numerical results to construct a simple model for the evolution of the coauthorship network. It is important to emphasize that the purpose of the model is to capture the main mechanisms that affect the evolution of the network, and not to incorporate every numerical detail of the measured web.

We denote by $k_i(t)$ the number of links node $i$ has at time $t$; by $T(t)$ and $N(t)$ the total number of links and total number of nodes at time $t$, respectively. We assume that all nodes present in the system are active, i.e. they can author further papers. This is a reasonable assumption as the time span over which data is available to us is shorter than the typical professional lifetime of a scientist. In agreement with figure 3.1, we consider that new researchers join the field at a constant rate, leading to

$$N(t) = \beta t. \tag{3.3}$$

The average number of links per node in the system at time $t$ is thus given by:

$$\langle k \rangle = \frac{T(t)}{N(t)}. \tag{3.4}$$

Figure 3.4*b* suggests, that the probability to create a new internal link between two existing nodes is proportional to the product of their degrees. Consequently, denoting by $a$ the number of newly created internal links per node in unit time, we write the probability that between node $i$ and $j$ a new internal link is created as

$$\Pi_{ij} = \frac{k_i k_j}{\sum'_{s,m} k_s k_m} N(t)\, a, \tag{3.5}$$

where the prime sign indicates that the summation is done for $s \neq m$ values.

Measurements also indicate that new nodes link to the existing nodes with preferential attachment (Fig. 3.4*a*), $\Pi(k)$ following $k^\nu$ with $\nu \simeq 0.75 - 0.8$. Aiming to obtain an analytically solvable model, at this point we neglect this nonlinearity

and we approximate $\Pi(k)$ with a linear $k$ dependence. The effect of the nonlinearities will be discussed at the end of this section. Thus, if node $i$ has $k_i$ links, the probability that an incoming node will connect to it is given by

$$\Pi_i = b \frac{k_i}{\sum_j k_j}, \tag{3.6}$$

where $b$ is the average number of new links that an incoming node creates.

We have thus formulated the dynamical rules that govern our evolving network model, capturing the basic mechanism governing the evolution of the coauthorship network:

- Nodes join the network at a constant rate $\beta$.
- Incoming nodes link to the already present nodes following preferential attachment (equation (3.6)).
- Nodes already present in the network form new internal links following preferential attachment (equation (3.5)).
- We neglect the aging of nodes, and assume that all nodes and links present in the system are active, able to initiate and receive new links.

In the model we assume that the number of authors on a paper is constant. In reality $m$ is a stochastic variable, as the number of authors varies from paper to paper. However, for the scale free model the exponent $\gamma$ is known to be independent of $m$, thus making $m$ a stochastic variable is not expected to change the scaling behavior.

### 3.4.1 Continuum Theory

Taking into account that new links join the system with a constant rate, $\beta$, the continuum equation for the evolution of the number of links node $i$ has can be

written as:

$$\frac{dk_i}{dt} = \frac{b\beta k_i}{\sum_j k_j} + N(t)\, a \sum_j{}' \frac{k_i k_j}{\sum'_{s,m} k_s k_m}. \tag{3.7}$$

The first term on the right hand side describes the contribution due to new nodes (equation (3.6)) and the second term gives the new links created with already existing nodes (equation (3.5)). The total number of links at time $t$ can be computed taking into account the internal and external preferential attachment rules:

$$\sum_i k_i = T(t) = \int_0^t 2\left[N(t')a + b\beta\right] dt' = t\beta(at + 2b). \tag{3.8}$$

Consequently the average number of links per node increases linearly in time,

$$\langle k \rangle = at + 2b, \tag{3.9}$$

in agreement with our measurements on the collaboration network (Fig. 3.3*b*). The master equation (3.7) can be solved if we approximate the double sum in the second term. Taking into account that we are interested in the asymptotic limit where the total number of nodes is large relative to the connectivity of the nodes, we can write:

$$\sum_{s,m}{}' k_s k_m = \sum_s k_s \sum_m k_m - \sum_m k_m^2 \approx \left(\sum_i k_i\right)^2. \tag{3.10}$$

We have used here the fact that $T(t)^2$ depends on $N^2$, while $\sum_i k_i^2$ depends only linearly on $N$ (we investigate the $N \to \infty$ limit). Using (3.10) equation (3.7) now becomes:

$$\frac{dk_i}{dt} = \frac{bk_i}{t(at + 2b)} + \frac{k_i a}{at + 2b}. \tag{3.11}$$

Introducing the notation $\alpha = a/b$, we obtain:

$$\frac{dk_i}{dt} = \frac{k_i}{t}\,\frac{t\alpha + 1}{t\alpha + 2}. \tag{3.12}$$

This differential equation is separable, the general solution having the form

$$k_i(t) = C_i \sqrt{t}\,\sqrt{2 + \alpha t}. \tag{3.13}$$

The $C_i$ integration constant can be determined from the initial conditions for node $i$. Since node $i$ joins the system at time $t_i$, we have $k_i(t_i) = b$, leading to

$$k_i(t) = b\sqrt{\frac{t}{t_i}}\sqrt{\frac{2+\alpha t}{2+\alpha t_i}}. \tag{3.14}$$

This implies that for large times ($t \to \infty$) the connectivity of the node scales linearly with time, i.e. $k(t) \sim t$ (Fig. 3.3*b*).

A quantity of major interest is the degree distribution, $P(k)$. The nodes join the system randomly at a constant rate, which implies that the $t_i$ values are uniformly distributed in time between 0 and t. The distribution function for the $t_i$ in the $[0, t]$ interval is simply $\rho(t) = 1/t$. $P(k)$ can be obtained in a similar manner to the calculation in §2.2.2 (equation (2.8) to (2.13) after determining the $t_i(k_i)$ dependence from equation (3.14):

$$P(k) = -\rho(t)\left.\frac{dt_i}{dk_i}\right|_k = \tag{3.15}$$

$$= b^2(2/\alpha + t)\frac{1}{k^2}\frac{1}{\sqrt{k^2 + b^2 t(2+\alpha t)}}. \tag{3.16}$$

An immediate consequence of this result is that the connectivity distribution depends both on the observation time $t$ and on the range of $k$ values we explore. In the asymptotic limit $t \to \infty$ we obtain

$$P(k) \propto \frac{1}{k^2}, \tag{3.17}$$

predicting a scale free behavior with exponent $\gamma = 2$. At short times, however, the exponent is different, the network exhibiting a scale free behavior similar to the scale free model [20, 21]:

$$P(k) \propto \frac{1}{k^3}. \tag{3.18}$$

Thus the model predicts that the degree distribution of the collaboration network displays a crossover between two scaling regimes. In general, scaling is controlled

by the time dependent crossover connectivity, given by

$$k_c = \sqrt{b^2 t(2 + \alpha t)}. \tag{3.19}$$

For $k \ll k_c$ the degree distribution scales with an exponent $\gamma = 2$, while for $k \gg k_c$ the degree distribution scales with $\gamma = 3$. The crossover connectivity, $k_c$, increases linearly in time for $t \gg 2/\alpha$, which implies that in the asymptotic limit ($t \to \infty$) only the $\gamma = 2$ exponent is observable.

Note that this result predicts that the degree distribution has two scaling regimes, one with $\gamma = 2$ for small $k$, followed by a crossover to $\gamma = 3$ for large $k$. This crossover towards a larger exponent can be easily approximated with an exponential cutoff, which is why we believe that in [149] the power law with an exponential cutoff gave a reasonable fit. However, as [147, 148] and our results show, for data sets with better statistics the scaling regimes can be distinguished. Indeed, the crossover is visible in figure 3.2*a,b* as well, in particular for the degree distribution of NS. The degree distribution taken in 1993 has a clear $\gamma = 3$ tail, as for the studied short time frame (3 years) $k_c$ is expected to be low. This $\gamma = 3$ tail all but disappears, however, in 1998, being replaced with a $\gamma = 2$ exponent, as predicted by equation (3.17) for the limit $t \to \infty$. The M database shows similar characteristics, albeit the crossover is masked by a higher spread in the data point due to weaker statistics.

Plotting two differently cumulated values instead of $P(k)$ the $\gamma = 2$ and $\gamma = 3$ scaling regimes become more evident. Let us denote by $F(k)$ the primitive function of $P(k)$, defining:

$$\Phi(k) = 1 - \int_1^k P(k')\, dk'. \tag{3.20}$$

$\Phi(k)$ can be determined numerically by integrating $P(k)$ between 1 and $k$. For small $k$ the function $\Phi(k)$ should scale as

$$\Phi(k) \propto k^{-1}, \tag{3.21}$$

assuming that $P(k)$ scales as given by equation (3.17). As figure 3.5$a$ shows, we indeed find that for large $t$ (1998) the measured $\Phi(k)$ function converges to a $k^{-1}$ behavior, which is less apparent on the small $t$ curves (1993 and 1995).

To investigate the large $k$ behavior of $P(k)$ we measured the $\tau(k)$ function defined as:

$$\tau(k) = \int_k^\infty P(k')\, dk', \tag{3.22}$$

which captures the scaling of the tail. According to equation (3.17) for large $k$ and small $t$ one should observe

$$\tau(k) \propto k^{-2}. \tag{3.23}$$

As figure 3.5 shows, we indeed find that for NS for small $t$ (1993) the large $k$ scaling follows the prediction (3.22), and, as predicted, the scaling increasingly deviates from it as time increases.

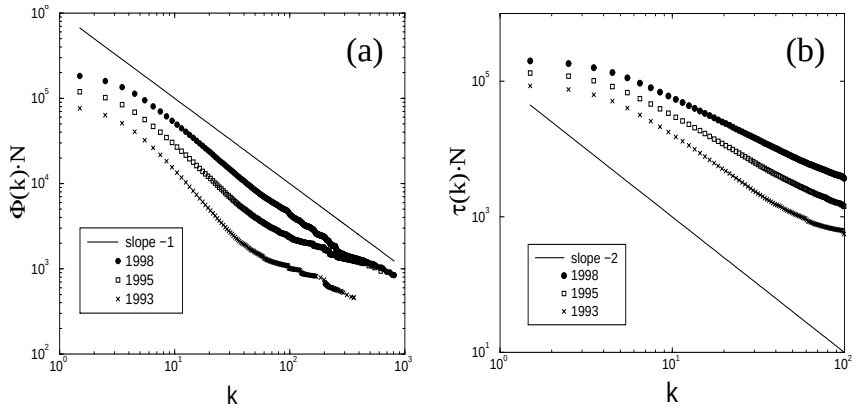



Figure 3.5. Scaling of $\Phi(k)$ **(a)** and of $\tau(k)$ **(b)** for the NS database, demonstrating the trends in the small and large $k$ behavior of the degree distribution.

### 3.4.2 Monte Carlo Simulations

While the continuum theory predicts the connectivity distribution in agreement with the empirical data, there are other quantities, such as the average path length

and clustering coefficient, that cannot be calculated using this method at this point. To investigate the behavior of these measures of the network topology next we study the proposed model using Monte Carlo simulations.

Due to memory and computing time limitations we investigated relatively small networks, with total number of nodes $N < 4000$. While these networks are considerably smaller than the real networks, their scaling and topological features should be representative. In order to form a reasonable number of internal links, we increased the parameter $a$ in equation (3.5). For comparison purposes we note that in the real system we have $a_{\rm M} = 0.31/\text{year} \simeq 10^{-4}/\text{simulation step}$ and $a_{\rm NS} = 0.98/\text{year} \simeq 3.684 \cdot 10^{-5}/\text{simulation step}$,[1] numbers that can be derived from the data shown in figure 3.3*b* and figure 3.1*b*.

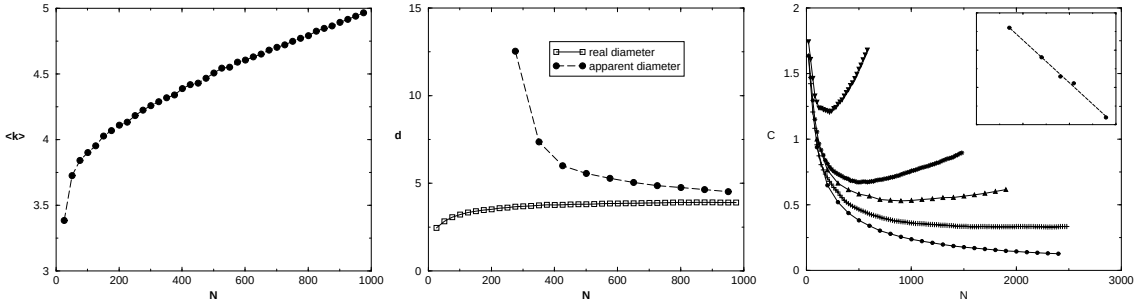
The advantage of the modelling efforts, including the Monte Carlo simulations, is that they reproduce the network dynamics from the very first node. In contrast, the database we studied records nodes and links only after 1991, when much of the networks structure was already in place. By collecting data over several years we gradually discovered the underlying structure. We expect that after a quite long measurement time the structure revealed by the collected data will be statistically indistinguishable from the full collaboration network. However, the dynamics we measure during this process for the relevant quantities (average path length, average connectivity, clustering coefficient) might differ from those characterizing the full network, since all of them are computed on the *incomplete* network (revealed by the available data). However, Monte Carlo simulations allow us to investigate the effect of the data incompleteness on the relevant network measures.

We investigated the time dependence of the average connectivity, the average path length and the clustering coefficient, using the parameters $N_{max} = 1000$, $a =$

[1] One simulation step corresponds to the addition of one new author connecting $b$ links to the system.

0.001, $\beta = 1$ and $b = 2$. In order to improve statistics, the results were averaged over 10 independent configurations.

*Average degree:* As figure 3.6*a* indicates, asymptotically the average connectivity increases linearly, in agreement with both our measurements (see Fig. 3.6*b*) and the continuum theory (see equation (3.9)).




**(a)** Average connectivity **(b)** Average shortest path **(c)** Clustering coefficient

Figure 3.6. Computer simulated dynamics for $N_{max} = 1000$, $a = 0.001$, $\beta = 1$ and $b = 2$. **(a)** Average connectivity. **(b)** Real and apparently measured average path length ($N_s = 200$). **(c)** Clustering coefficient for different values of the $a$ parameter as a function of the system size $N$ (values of $a$ are 0 ($\bullet$), 0.00025 ($+$), 0.0005 ($\triangle$), 0.00075 ($*$) and 0.002 ($\bigtriangledown$)). The inset shows the scaling of the minimum value of $C$ as a function of $a$, the line shows a fit $\ln N_{min} = -(1.9 + 1.14 \cdot \ln a)$.

*Average path length:* The empirical results indicated (see Fig. 3.2*c*) that the average path length decreases with time for both databases. In contrast, our simulations show a monotonically increasing $d$, in apparent disagreement with the real system.

Note that an increasing average path length agrees with measurements done on other models, including scale free and exponential networks, all predicting an approximately logarithmic increase with the number of nodes, $d \sim \ln(N)$ [13, 29]. This contradiction between the models and our empirical data is rooted in the incomplete data we have for the first years of our measurements. To show this we perform the following simulation. We construct a network of $N = 1000$ nodes,

however, we will record the apparent average path length of the network made of nodes that have been added only after a predefined time: we try to mimic the fact that the data available for us gives $d$ only for publications after 1991. We find that the average path length of this incomplete network has a decreasing tendency, slowly converging to the real value (Fig. 3.6*b*), in agreement with the decrease observed in the empirical measurements (Fig. 3.2*e*). This result underlies the importance of simulations in understanding the dynamics of complex networks, and resolves the conflict between the simulation and the empirical data. It also indicates that most likely the average path length of the M and NS database does increase in time, but such increase can be observed only if much longer time intervals will be available for study.

*Clustering coefficient.* The clustering coefficient predicted by our simulations is shown in figure 3.6*c*. $C$ depends strongly on the value of the parameter $a$. For $a = 0$ we have essentially the BA model [20] and the clustering coefficient decreases monotonically. For $a > 0$ however, the clustering coefficient decreases at the beginning, but after reaching a minimum at $N_{min}$ it changes its trend, now increasing in time. Thus we expect that for all $a > 0$ the clustering coefficient should increase in the asymptotic limit, in agreement with our measurements on the collaboration network (see Fig. 3.2*d*). The $N_{min}$ position where the clustering coefficient has a minimum scales as power of the $a$ parameter, as shown as the inset in figure 3.6*c*.

We conclude that the decreasing $C$ observed for our database, shown in figure 3.2*d*, does not represent the asymptotic behavior. The observed behavior also indicates that one should view the values for $C$ reported in the literature, and measured for finite time-frames (maximum 5 years) with caution, as they might not represent asymptotic values.

*Degree distribution:* The simulations provide $P(k)$ as well, allowing us to check

the validity of the predictions of the continuum theory. Although the considered system sizes are rather small ($N_{max} = 3500$) compared to the $N \to \infty$ approximation used in the analytical calculation and the $N_M = 70,975$, $N_{NS} = 209,750$ for the empirical data, the behavior of $P(k)$, shown in figure 3.7 agrees with our continuum model and measurements. For small $k$ we observe the $\gamma = 2$ scaling, while for large $k$ $P(k)$ converges to the predicted $\gamma = 3$ exponent.

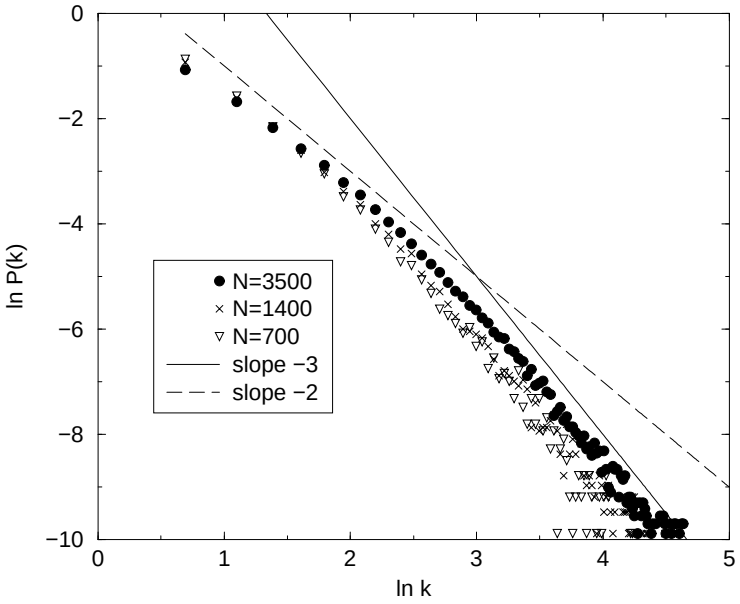



Figure 3.7. Connectivity distributions as predicted by numerical simulation for different stages of evolution of the network ($a = 0.001$, $\beta = 1$ and $b = 2$).

### 3.4.3 Nonlinear Effects

An issue that remained unresolved up to this point concerns the effect of the nonlinear preferential attachment. We have seen in §3.3.6 that the incoming links follow

$$\Pi_i = b \frac{k_i^\nu}{\sum_j k_j^\nu}, \tag{3.24}$$

with $\nu \approx 0.8$. On the other hand, for such preferential attachment Krapivsky *et al.* have shown that the degree distribution follows a stretched exponential;

i.e., the power law is absent [121, 120]. This would indicate that $P(k)$ for the coauthorship network should follow a stretched exponential, which disagrees with our and Newman's findings (we have explicitly checked that a stretched exponential is not a good fit for our data). What could then override the known effect of the $\nu < 1$ nonlinear behavior? Next we propose a possible explanation: the linearity of the internal preferential attachment can restore the power law nature of $P(k)$.

To test the potential effect of the nonlinearities in the preferential attachment of newly added nodes, we have simulated the model with $\nu = 0.75$, otherwise all parameters being unchanged. We show on figure 3.8 the degree distribution for the linear ($\nu = 1$) and the nonlinear ($\nu = 0.75$) case. As one can see, the $\nu = 1$ and

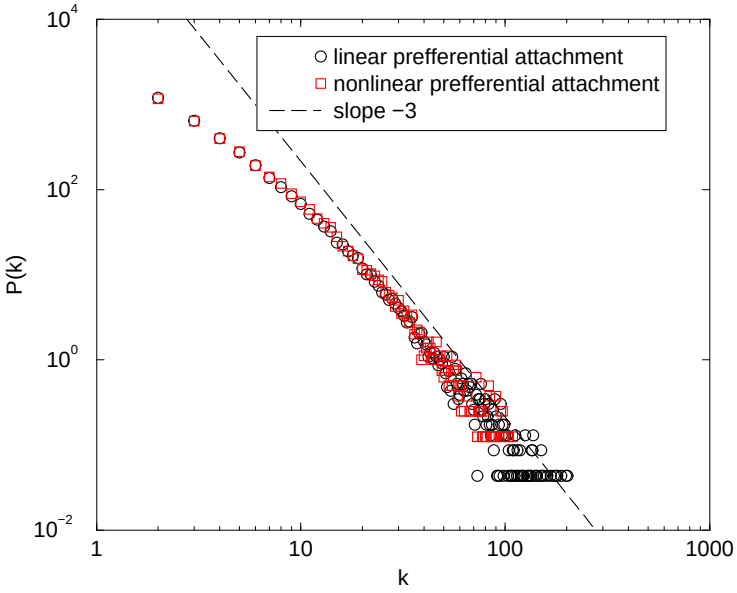



Figure 3.8. Connectivity distribution generated by the numerical simulations for linear ($\nu = 1$) and nonlinear ($\nu = 0.75$) preferential attachment ($N_{max} = 3500$, $a = 0.0005$, $\beta = 1$ and $b = 2$).

$\nu = 0.75$ case can be hardly distinguished. This could have two origins. First, the simulations are limited to $t = 3500$ simulation steps, due to the discussed running time limitations. Thus we are hardly in the asymptotic regime. On the other hand, the agreement indicates that the nonlinearity has a barely distinguishable effect on

$P(k)$, with the internal attachment dominating the system behavior.

In summary, the domination of the internal attachment affects are expected to be even more dominant for the real network. Indeed, in the collaboration network the fraction of the links created as internal links is much higher than those created by the incoming nodes, as an author qualifies for a new incoming link only on his first paper. Most scientists contribute for a considerable time to the same field, publishing numerous subsequent papers, and these later links will all appear as internal links. Thus typically the number of internal links is much higher than the number of new links, making the network's topology much more driven by the internal links than by the external ones. This is one possible reason why the effect of the nonlinear behavior, while clearly present, cannot be detected in the functional form of $P(k)$.

## 3.5 Discussion

In the past few years we have witnessed considerable advances in addressing the topology and dynamics of complex networks. Along this road a number of quantities have been measured and calculated, aiming to characterize the network topology. However most of these studies are fragmented, focusing on one or a few characteristics of the network at a time. Here, we have presented a detailed study of a network of high interest to the scientific community: the collaboration network of scientists, which also represents a prototypical example of a complex evolving network. This study allows us to investigate to what degree can we use various known measures to characterize a given network. An important result of our investigation is the understanding that we need to be careful in distinguishing between the asymptotic and the intermediate behavior. In particular, most quantities used to characterize the network are time dependent. For example, the average path length, the clus-

tering coefficient, as well as the average degree of the nodes are often used as basic time independent network characteristics. Our empirical results show that many of these key quantities are time dependent, without a tendency to saturate within the available time-frame. Thus their value at a given moment tells us little about the network. They can be used, however, at any moment, to show that the network has small world properties, i.e. it has a small average path length, and a clustering coefficient that is larger than one expected for a random network.

A quantity that is often believed to offer a stationary measure of the network is the degree distribution. Our empirical data, together with the analytic solution of the model shows that this is true only asymptotically for the coauthorship network: we uncover a crossover behavior between two different scaling regimes. We tend to believe that the model's predictions are not limited to the collaboration network: as on the WWW and for the actor collaboration network similar basic processes take place, chances are that similar crossovers could appear there as well.

A third important conclusion of the study regards the understanding that the measurements done on incomplete databases could offer trends that are opposite compared to that seen in the full system. An example is the average path length: we find that the empirically observed decreasing tendency is an artifact of the incomplete data. However, our simulations show that one can, with careful modelling, uncover such inconsistencies. But this also offers an important warning: for any network, before attempting to model it, we need to fully understand the limitations of the data-collection process, and test their effect on the quantities of interest for us.

In summary, the modelling efforts presented here are only the starting point for a systematic investigation of the evolution of social networks. It is important to note that such modelling is open ended: more details can be incorporated that could undoubtedly improve the agreement between the empirical data and theory.

# CHAPTER 4

# HIERARCHY IN NETWORKS

Many real networks are expected to be fundamentally modular, meaning that the network can be seamlessly partitioned into a collection of modules. Each module is expected to perform an identifiable task, separate from the function of other modules [89, 215, 125, 188]. On the other hand, most networks have a scale free connectivity distribution, a topology in which a hierarchy of hubs of all sizes links all parts of the network into a highly integrated web, making separable grouping of nodes apparently difficult. Therefore, there must be a way to reconcile the scale free property with the network's potential modularity. This dilemma was central to the community trying to understand the architecture of cellular networks. On our journey towards an answer, hierarchical network architecture has first served us as an intuitive principle that helped us construct deterministic scale free networks. Soon, however, it lead us to the conclusion that a large variety of real networks have an underlying hierarchical structure. After a short presentation of two deterministic scale free models only aimed to offer deterministic scale free graphs [24, 53], we present our hierarchical model in detail [174, 173, 23], followed by a comparison of some of its properties to real-world networks.

## 4.1 Deterministic scale free Models

The high interest in understanding the topology of complex networks has resulted in the development of a considerable number of network models [20,

21, 55, 54, 56, 59, 120, 121, 122, 27, 28]. Most of these are based on incremental growth and preferential attachment [20, 21], and stochasticity was a common feature of all network models that generate scale free topologies. That is, new nodes connect to nodes already present in the system using a probabilistic rule. This randomness present in the models, while in line with the major features of networks seen in nature, made it harder to gain a visual understanding of what makes them scale free, and how different nodes relate to each other. It was therefore of major theoretical interest to construct models that lead to scale free networks in a deterministic fashion. Here we present our first simple model, generating a deterministic scale free network using a hierarchical construction [24].

### 4.1.1 Description of the Model

The construction of the model follows a hierarchical rule commonly used in deterministic fractals [132, 202], as shown in figure 4.1. The network is built in an iterative fashion, each iteration repeating and reusing the elements generated in the previous steps as follows:

- *Step 0:* We start from a single node, the *root* of the graph.
- *Step 1:* We add two more nodes, and connect each of them to the root.
- *Step 2:* We add two units of three nodes, each unit identical to the network created in *Step 1*, and we connect each of the *bottom* nodes of these two units to the root.
- *Step 3:* We add two units of nine nodes each, identical to the units generated in the previous iteration, and connect all eight bottom nodes of the two new units to the root.

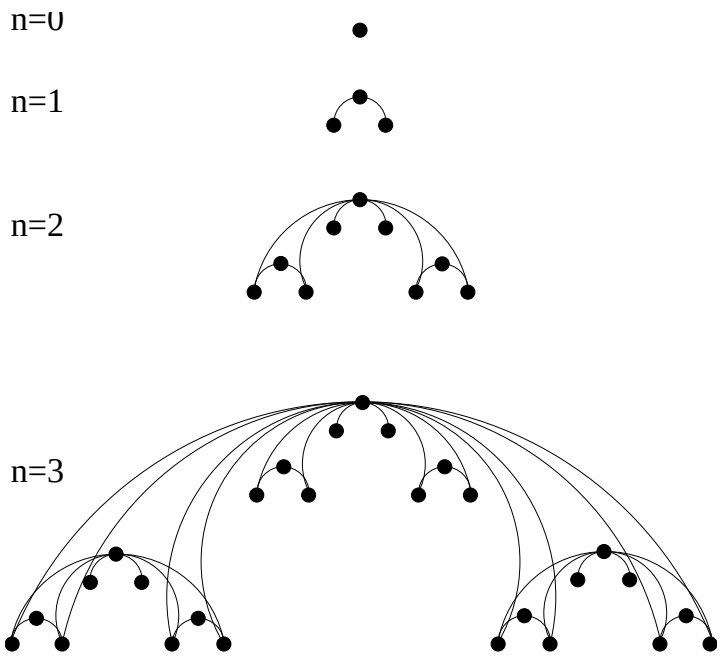



Figure 4.1. Construction of the deterministic scale free network.

These rules can be easily generalized. Indeed, step $n$ would involve the following operation:

- *Step n:* Add two units of $3^{n-1}$ nodes each, identical to the network created in the previous iteration (step $n-1$), and connect each of the $2^n$ bottom nodes of these two units to the root of the network.

Thanks to its deterministic and discrete nature, the degree distribution of this model can be calculated exactly.

The tail of the degree distribution is determined by the most connected nodes, or hubs. Clearly the biggest hub is the root, and the next two hubs are the roots of the two units added to the network in the last step. Therefore, in order to capture the tail of the distribution, it is sufficient to focus on the hubs.

In step $i$ the degree of the most connected hub, the root, is $2^{i+1}-2$. In the next iteration two copies of this hub will appear in the two newly added units. As

we iterate further, in the $n$th step $3^{n-i}$ copies of this hub will be present in the network. However, the two newly created copies will not increase their degree after further iterations. Therefore, after $n$ iterations there are $(2/3)3^{n-i}$ nodes with degree $2^{i+1}-2$. Since spaces between degrees the nodes grow with increasing $k$ values, the exponent of the degree distribution can be properly calculated using the cumulative degree distribution. The tail of the cumulative degree distribution, determined by the hubs, follows

$$P_{\rm cum}(k) \sim k^{1-\gamma} \sim k^{-\frac{\ln 3}{\ln 2}}. \tag{4.1}$$

Thus the degree exponent is

$$\gamma = 1 + \frac{\ln 3}{\ln 2}. \tag{4.2}$$

The origin of this scaling behavior can be understood by inspecting the model's construction. Indeed, at any moment we have a hierarchy of hubs, highly connected nodes which are a common component of scale free networks. The root is always the largest hub. However, at any step there are two hubs whose connectivity is roughly a half of the root's connectivity, corresponding to the roots of the two units added at step $n-1$. There are six even smaller hubs, with connectivity $2^{n-1}-2$, corresponding to the root of the units added at time $n-2$, and so on. This hierarchy of hubs is responsible for the network's scale free topology. As the number of hubs increases as powers of 3, while the number of links only as powers of two, the degree exponent is expected to be 1 plus a simple multiple of $\ln 3/\ln 2$.

The model introduced above offers a deterministic construction of a scale free network, with the interesting property of self similarity.

The proposed model generates a network with a fixed $\gamma = 1 + \ln 3/\ln 2$ degree exponent. However, one can easily modify the model to change the scaling exponent by varying the number of links connected to the root at each step. Similarly, by definition the model discussed here has a zero clustering coefficient [147, 147], as

it does not generate triangles of connected nodes. It is easy to change the rules, without changing the scaling exponent, to obtain a network that displays nonzero clustering, as we can see in the next section.

### 4.1.2 The Pseudofractal Graph

Dorogovtsev, Goltsev and Mendes constructed a deterministic scale free network which deserves our attention due to the clustering properties of its nodes.

The model network is grown as follows (see Fig. 4.1):

- *Time -1:* The growth starts from a single edge linking two nodes.
- *Time 0:* A new vertex is attached to both end vertices of the previous edge.

  ...
- *Time t:* At each time step a new vertex is attached to both ends of every edge of the graph (new vertices and edges are drawn with red on figure 4.1).

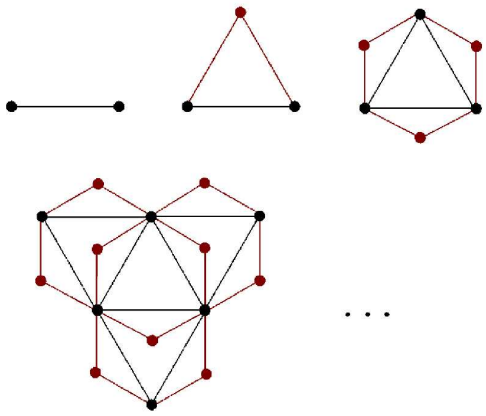


Figure 4.2. Construction of the pseudofractal network. (After [53])

The total number of vertices at time $t$ is $N_t = 3(3^t + 1)/2$, and the total number of edges is $L = 3^t + 1$, leading to an average degree $< k >= 4/(1 + 3^{-t})$.

*Degree distribution.* At time $t$, the number $n_t(k)$ of vertices of degree $k = 2, 2^2, \ldots, 2^{t-1}, 2^t$ is equal to $3^t, 3^{t-1}, \ldots, 3^2, 3$, respectively, while other degree values are absent. Calculating the cumulative distribution leads to a power law degree distribution with degree exponent $\gamma = 1 + \ln 3/\ln 2$.

*Distribution of clustering.* Usually only the average value of the clustering coefficient is considered. However, in this graph there is a one-to-one correspondence

between the clustering coefficient of a vertex and its degree:

$$C = 2/k. \tag{4.3}$$

In the next section we show that the scaling of the clustering coefficient of the vertices with their degrees is a signature of an underlying hierarchical architecture.

## 4.2 The Hierarchical Model

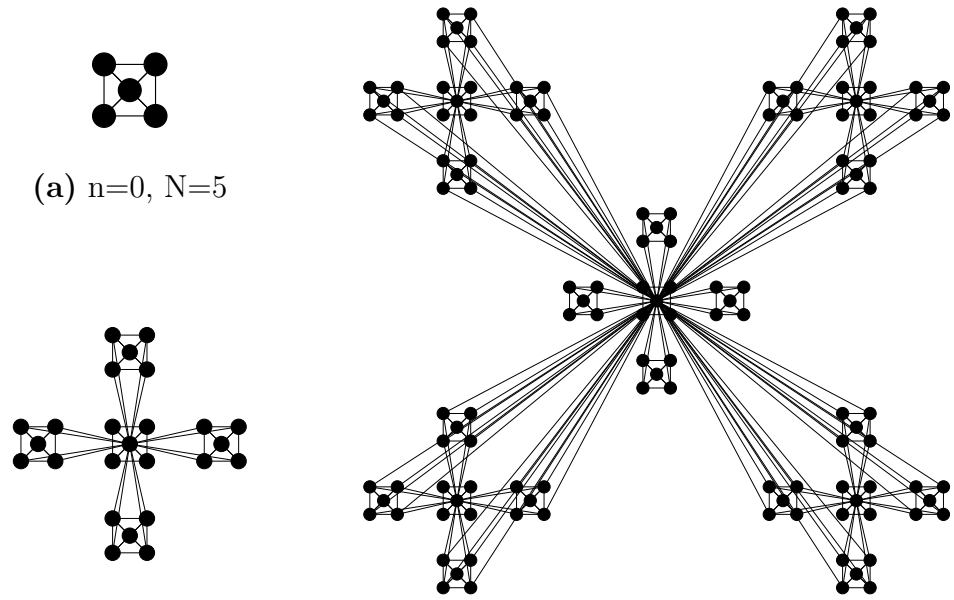
We have seen earlier that the scale free property and clustering are not exclusive: for a large number of real networks, including metabolic networks [101, 206], the protein interaction network [98, 205], the world wide web [13] and even some social networks [149, 147, 148, 22] the scale free topology and high clustering coexist. Most models proposed to describe the topology of complex networks have difficulty capturing these two features simultaneously. Here we show that the fundamental discrepancy between models and empirical measurements is rooted in a previously disregarded, yet generic feature of many real networks: their hierarchical topology. Indeed, many networks are fundamentally modular: one can easily identify groups of nodes that are highly interconnected with each other, but have only a few or no links to nodes outside of the group to which they belong to. In society such modules represent groups of friends or coworkers [86]; in the WWW denote communities with shared interests [76, 5]; in the actor network they characterize specific genres or simply individual movies. Some groups are small and tightly linked, others are larger and somewhat less interconnected. This clearly identifiable modular organization is at the origin of the high clustering coefficient seen in many real networks. Yet, most models reproducing the scale free property of real networks [12, 58] distinguish nodes based only on their degree, and are blind to node characteristics that could lead to a modular topology.

In order to bring modularity, the high degree of clustering and the scale free topology under a single roof, we need to assume that modules combine into each other in a hierarchical manner, generating what we call a *hierarchical network*. The presence of a hierarchy and the scale free property impose strict restrictions on the number and the degree of cohesiveness of the different groups present in a network, which can be captured in a quantitative manner using a scaling law, describing the dependence of the clustering coefficient on the node degree. We use this scaling law to identify the presence of a hierarchical architecture in several real networks, and the absence of such hierarchy in geographically organized webs.

### 4.2.1 Construction of the Model

Our starting point is a small cluster of five densely linked nodes (Fig. 4.3*a*). Next we generate four replicas of this hypothetical module and connect the four external nodes of the replicated clusters to the central node of the old cluster, obtaining a large 25-node module (Fig. 4.3*b*). Subsequently, we again generate four replicas of this 25-node module, and connect the 16 peripheral nodes to the central node of the old module (Fig. 4.3*c*), obtaining a new module of 125 nodes. These replication and connection steps can be repeated indefinitely, in each step increasing the number of nodes in the system by a factor five.

As mentioned in the previous section, precursors to this model have been proposed in Ref. [24] and extended and discussed in Ref. [53, 104] as a method of generating deterministic scale free networks. Yet, it was believed that aside from their deterministic structure, their statistical properties are equivalent with the stochastic models that are often used to generate scale free networks. In the following we argue that such hierarchical construction generates an architecture that is significantly different from the networks generated by traditional scale free models. Most



**(a)** n=0, N=5

**(b)** n=1, N=25

**(c)** n=2, N=125

Figure 4.3. The iterative construction leading to a hierarchical network. Starting from a connected cluster of five nodes shown in **(a)** we create four identical replicas, connecting the peripheral nodes of each cluster to the central node of the original cluster, obtaining a network of $N = 25$ nodes **(b)**. In the next step we create four replicas of the obtained cluster, and connect the peripheral nodes again, as shown in **(c)**, to the central node of the original module, obtaining a $N = 125$ node network. This process can be continued indefinitely.

important, we show that the new feature of the model, its hierarchical character, are shared by a significant number of real networks.

### 4.2.2 Properties of the Hierarchical Model

To analyze the scaling behavior of the degree distribution and clustering coefficient of this hierarchical network we first need to count the nodes with different degrees, then calculate their clustering coefficients. Starting with the first five nodes, we label the middle one a *hub* and we call the remaining four *peripheral*. All nodes that originate as copies of hubs are again called hubs, and we will continue calling copies of peripheral nodes peripheral. This distinction is useful since the rules

responsible for connecting these classes of nodes are somewhat different.

The central hub acquires $4^n$ links during the $n^{\text{th}}$ iteration. Let us call the central hub $H_n$, the four copies of this hub $H_{n-1}$. The $4 \cdot 5$ leftover module centers with sizes equal to the size of the network at the $n-2^{\text{th}}$ iteration are called $H_{n-2}$. At the $n^{\text{th}}$ iteration a hub $H_i$ has all the links the central hub had after the $i^{\text{th}}$ iteration:

$$k_n(H_i) = \sum_{l=1}^{i} 4^l = \frac{4}{3}\,(4^i - 1). \tag{4.4}$$

For any $i < n$ the number of $H_i$ modules:

$$N_n(H_i) = 4 \cdot 5^{n-1-i} \tag{4.5}$$

(there are four for $i = n-1$, $4 \cdot 5$ for $i = n-2$, ..., for $i = 1$ we have $4 \cdot 5^{n-2}$, or 4/5-th of the copies of the original five-node module). Since we have $4 \cdot 5^{n-1-i}$ $H_i$-type hubs of connectivity $k_n(H_i)$, we obtain $\ln N_n = c_n - i \ln 5$ and $\ln k_n \simeq i \ln 4 + \ln(4/3)$.

Let us now focus on the peripheral nodes. The largest possible connectivity of a peripheral node equals the number of iterations plus 2. These are the nodes at the "edge" of the drawn graph, the ones created in the last iteration and they are connected to one hub of each size (plus two other peripheral nodes in their small square). Thus peripheral nodes do not contribute to the tail of the degree distribution, they only influence the very small $k$ range.

Thus for all $k > n + 2$ we have $\ln N_n = c_n - k_n \frac{\ln 5}{\ln 4}$. The $k$ values contributing to the graph's degree distribution are not continuous (Fig. 4.4*a*): the gap between consecutive values grows as a power law. Thus the degree distribution function is a power law

$$P(k) \sim k^{-\gamma}, \qquad \text{where} \qquad \gamma = 1 + \frac{\ln 5}{\ln 4}. \tag{4.6}$$

The clustering coefficient of the $H_i$ hubs is easy to calculate. Their $\sum_{l=1}^{i} 4^l$ links come from nodes linked in a square, thus the number of connections between them

is equal to their number. Thus the number of links between the $H_i$ hub's neighbors is $\sum_{l=1}^{i} 4^l = k_n(H_i)$. This leads to

$$C_n(H_i) = \frac{k_i}{k_i\,(k_i-1)/2} = \frac{2}{k_i - 1}, \tag{4.7}$$

indicating that the $C(k)$ function scales as $k^{-1}$ (Fig. 4.4*b*) [53].

The average clustering coefficient of the hierarchical model asymptotically approaches 0.743, the correction to this asymptotic value decreases as a power law with the system size (Fig. 4.4*c*).

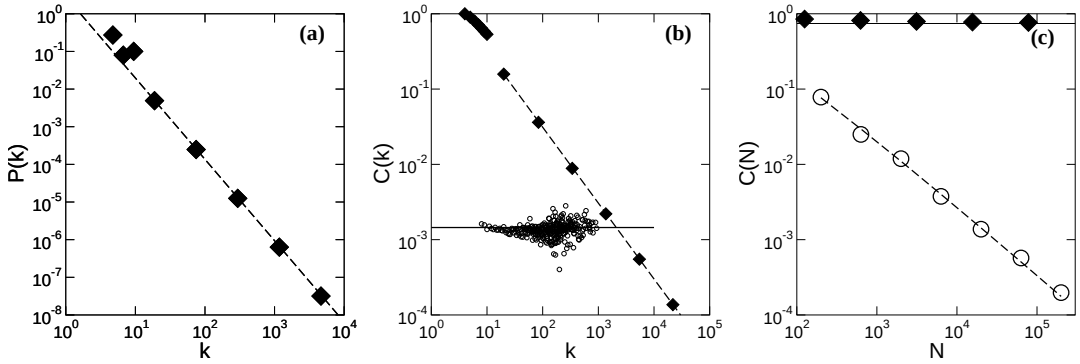



Figure 4.4. Scaling properties of the hierarchical model ($N = 5^7$). **(a)** The numerically determined degree distribution. The asymptotic scaling, with slope $\gamma - 1 = \ln 5/\ln 4$, is shown as a dashed line. **(b)** The $C(k)$ curve for the model, demonstrating that it follows equation (4.8). The open circles show $C(k)$ for a scale free model [20] of the same size, illustrating that it does not have a hierarchical architecture. **(c)** The dependence of the clustering coefficient, $C$, on the size of the network $N$. While for the hierarchical model $C$ is independent of $N$ (♦), for the scale free model $C(N)$ decreases rapidly (○).

### 4.2.3 Signature of Hierarchy

The most important feature of the network model of figure 4.3, not shared by either the scale free [20, 21] or random network models [64, 29], is its hierarchical architecture. The network is made of numerous small, highly integrated five-node modules (Fig. 4.3*a*), which are assembled into larger 25-node modules (Fig. 4.3*b*).

These 25-node modules are less integrated but each of them is clearly separated from the other 25-node modules when we combine them into the even larger 125-node modules (Fig. 4.3*c*). These 125-node modules are even less cohesive, but again will appear separable from their replicas if the network expands further.

This intrinsic hierarchy can be characterized in a quantitative manner using the finding of Dorogovtsev, Goltsev and Mendes [53] (See §4.1.2) that in their deterministic scale free network the clustering coefficient of a node with $k$ links follows the scaling law

$$C(k) \sim k^{-1}. \tag{4.8}$$

We argue that this scaling law quantifies the coexistence of a hierarchy of nodes with different degrees of clustering, and applies to the model of figure 4.3*a*–*c* as well. Indeed, the nodes at the center of the numerous 5-node modules have a clustering coefficient $C = 2/3$. Those at the center of a 25-node module have $k = 20$ and $C = 2/19$, while those at the center of the 125-node modules have $k = 80$ and $C = 2/79$, indicating that the higher a node's degree the smaller is its clustering coefficient, asymptotically following the $1/k$ law (Fig. 4.4*b*). In contrast, for the scale free model proposed in Ref. [20] the clustering coefficient is independent of $k$, i.e. the scaling law (4.8) does not apply (Fig. 4.4*b*). The same is true for the random [64, 29] or the various small world models [211, 145], for which the clustering coefficient is independent of the nodes' degree.

Therefore, the discrete model of figure 4.3 combines the two key properties of real networks within a single framework: their scale free topology and high modularity, which results in a system-size independent clustering coefficient. Yet, the hierarchical modularity of the model results in the scaling law (4.8), which is not shared by the traditional network models. The question is, could hierarchical modularity, as captured by this model, characterize real networks as well?

To investigate if such hierarchical organization is present in real networks we measured the $C(k)$ function for several networks for which large topological maps are available. Next we discuss each of these systems separately.

*Actor Network:* The `www.IMDB.com`-based Hollywood actor network mentioned in §1.3.1 consists of 392,340 nodes and 15,345,957 links [20, 11, 16]. As figure 4.5*a* indicates, we find that the high-$k$ range of $C(k)$ scales as $k^{-1}$, indicating that the network has a hierarchical topology. Indeed, the majority of actors with a few links (small $k$) appear only in one movie. Each such actor has a clustering coefficient equal to one, as all actors the actor has links to are part of the same cast, and are therefore connected to each other. The high $k$ nodes include many actors that acted in several movies, and thus their neighbors are not necessarily linked to each other, resulting in a smaller $C(k)$. At high $k$ the $C(k)$ curve splits into two branches, one of which continues to follow equation (4.8), while the other saturates. One explanation of this split is the decreasing amount of data-points available in this region. Indeed, in the high $k$ region the number of nodes having the same $k$ is rather small. If one of these nodes corresponds to an actor that played only in a few movies with hundreds in the cast, it will have both high $k$ and high $C$, considerably increasing the average value of $C(k)$. The $k$ values for which such a high $C$ nodes are absent continue to follow the $k^{-1}$ curve, resulting in jumps between the high and small $C$ values for large $k$. For small $k$ these anomalies are averaged out.

*Language network:* Here we study the network generated connecting two words to each other if they appear as synonyms in the Merriam-Webster dictionary [221] (See §1.3.3). The obtained semantic web has 182,853 nodes and 317,658 links and it is scale free with degree exponent $\gamma = 3.25$. The $C(k)$ curve for this language network is shown in figure 4.5*b*, indicating that it follows equation (4.8), suggesting

that the language has a hierarchical organization.

*World Wide Web:* The sample of the WWW we study, obtained by mapping out the `www.nd.edu` domain [13], has 325,729 nodes and 1,497,135 links, and it is scale free with degree exponents $\gamma_{\text{out}} = 2.45$ and $\gamma_{\text{in}} = 2.1$, characterizing the out and in-degree distribution, respectively (See §1.3.3). To measure the $C(k)$ curve we made the network undirected. While the obtained $C(k)$, shown in figure 4.5*c*, does not follow as closely the scaling law (4.8) as observed in the previous two examples, there is clear evidence that $C(k)$ decreases rapidly with $k$, supporting the coexistence of many highly interconnected small nodes with a few larger nodes, which have a much lower clustering coefficient.[1]

Indeed, the Web is full of groups of documents that all link to each other. For example, `www.nd.edu/~networks`, our network research dedicated site, has a high clustering coefficient, as the documents it links to have links to each other. The site is one of the several network-oriented sites, some of which point to each other. Therefore, the network research community still forms a relatively cohesive group, albeit less interconnected than the `www.nd.edu/~networks` site, thus having a smaller $C$. This network community is nested into the much larger community of documents devoted to statistical mechanics, that has an even smaller clustering coefficient. Therefore, the $k$-dependent $C(k)$ reflects the hierarchical nesting of the different interest groups present on the Web.

*Internet at the AS level:* As figure 4.5*d* shows, we find that at the domain level (see §1.3.2) the Internet [2] has a hierarchical topology as $C(k)$ is well approximated with equation (4.8). The scaling of the clustering coefficient with $k$ for the Internet was earlier noted by Vázquez, Pastor-Satorras and Vespignani (VPSV) [200, 200], who observed $C(k) \sim k^{-0.75}$. VPSV interpreted this finding, together with the

[1]Note that $C(k) \sim k^{-1}$ for the WWW was observed and briefly noted in Ref. [61].

observation that the average nearest-neighbor connectivity also follows a power law with the node's degree, as a natural consequence of the *stub* and *transit* domains, that partition the network in a hierarchical fashion into international connections, national backbones, regional networks and local area networks.

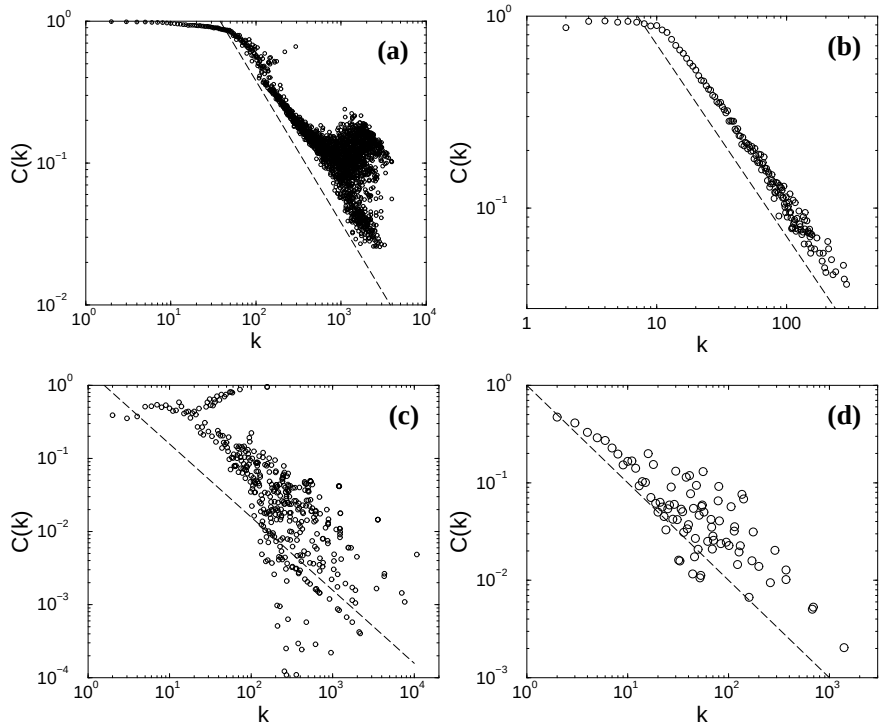



Figure 4.5. The scaling of $C(k)$ with $k$ for four large networks: **(a)** Actor network, two actors being connected if they acted in the same movie according to the `www.IMDB.com` database. **(b)** The semantic web, connecting two English words if they are listed as synonyms in the Merriam Webster dictionary [221]. **(c)** The World Wide Web, based on the data collected in Ref. [13]. **(d)** Internet at the Autonomous System level, each node representing a domain, connected if there is a communication link between them. The dashed line in each figure has slope $-1$, following equation (4.8).

Our measurements indicate, however, that some real networks lack a hierarchical architecture, and do not obey the scaling law (4.8). In particular, we find that the power grid and the router level Internet topology have a $k$ independent $C(k)$.

*Internet at the router level:* The router level Internet has 260,657 nodes con-

nected by 1,338,100 links [85] (see §1.3.2) . Measurements indicate that the network is scale free [69, 222] with degree exponent $\gamma = 2.23$. Yet, the $C(k)$ curve (Fig. 4.6*a*), apart from some fluctuations, is largely independent of $k$, in strong contrast with the $C(k)$ observed for the Internet's domain level topology (Fig. 4.5*d*), and in agreement with the results of VPSV [200, 201], who also note the absence of a hierarchy in router level maps.

*Power Grid:* The network studied by us represents the map of the Western United States, and has 4,941 nodes and 13,188 links [211] (see §1.3.2). The results again indicate that apart from fluctuations, $C(k)$ is independent of $k$.

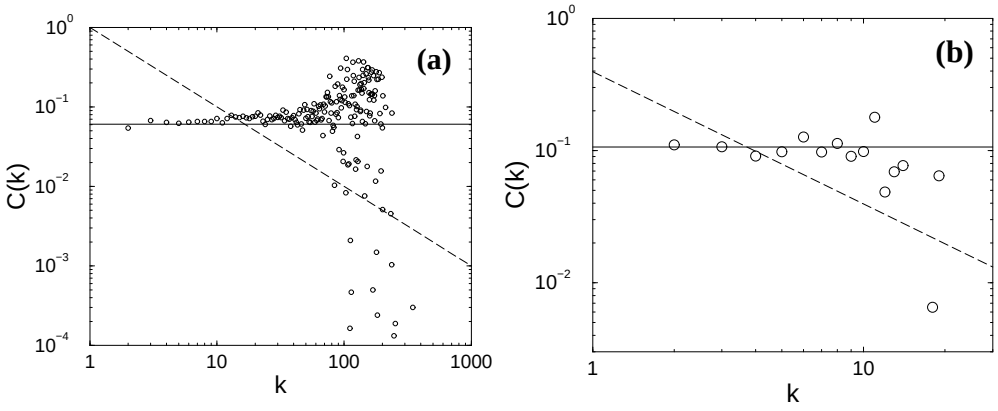



Figure 4.6. The behavior of $C(k)$ for two large, non-hierarchical networks: **(a)** Internet at router level [85]. **(b)** The power grid of Western United States. The dashed line in each figure has slope $-1$, while the solid line corresponds to the average clustering coefficient.

It is quite remarkable that these two networks share a common feature: a geographic organization. The routers of the Internet and the nodes of the power grid have a well defined spatial location, and the link between them represent physical links. In contrast, for the examples discussed in figure 4.5 the physical location of the nodes was either undefined or irrelevant, and the length of the link was not of major importance. For the router level Internet and the power grid the further are two nodes from each other, the more expensive it is to connect them [222]. Therefore,

in both systems the links are driven by cost considerations, generating a distance driven structure, apparently excluding the emergence of a hierarchical topology. In contrast, the domain level Internet is less distance driven, as many domains, such as the AT&T domain, span the whole United States.

In summary, we offered evidence that for four large networks $C(k)$ is well approximated by $C(k) \sim k^{-1}$, in contrast to the $k$-independent $C(k)$ predicted by both the scale free and random networks. In addition, there is evidence for similar scaling in the metabolism [174] (see next Chapter) and protein interaction networks [223]. This indicates that these networks have an inherently hierarchical organization. In contrast, hierarchy is absent in networks with strong geographical constraints, as the limitation on the link length strongly constraints the network topology.

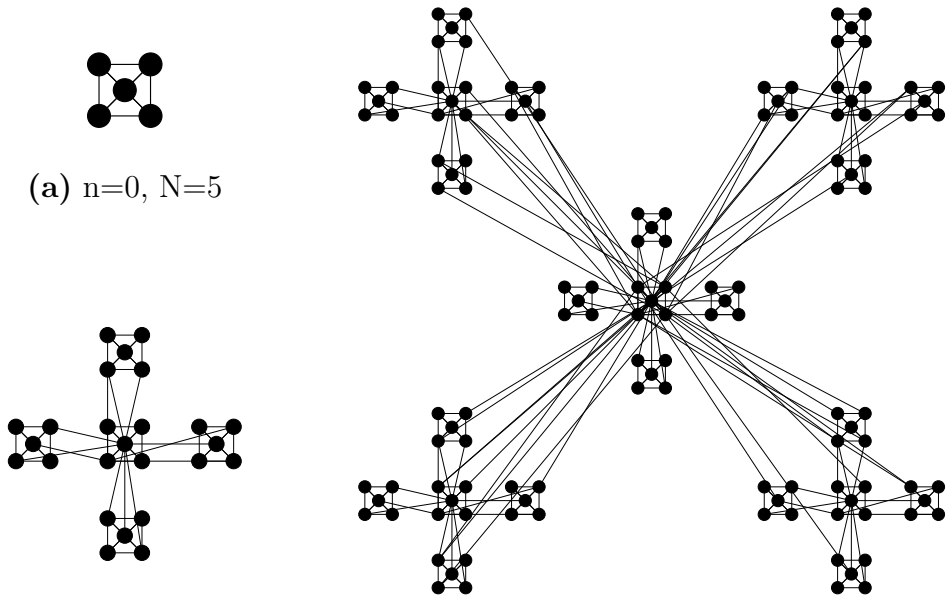
## 4.4 Stochastic Model and Universality

The hierarchical model described in figure 4.3 predicts $C(k) \sim k^{-1}$, which offers a rather good fit to three of the four $C(k)$ curves shown in figure 4.5. The question is, is this scaling law (4.8) universal, valid for all hierarchical networks, or could different scaling exponent characterize $C(k)$? Defining the hierarchical exponent, $\beta$, as

$$C(k) \sim k^{-\beta}, \tag{4.9}$$

is $\beta = 1$ a universal exponent, or can its value be changed together with $\gamma$? In the following we demonstrate that the hierarchical exponent $\beta$ can be tuned as we tune some of the network parameters. For this we propose a stochastic version of the model.

We start again with a small core of five nodes all connected to each other (Fig. 4.7*a*) and in step one ($n = 1$) we make four copies of the five-node module. Next, we randomly pick a fraction $p$ of the newly added nodes and connect each of



**(a)** n=0, N=5

**(b)** n=1, N=25

**(c)** n=2, N=125

Figure 4.7. Iterative construction of the stochastic hierarchical network ($p = 3/5$).

them independently to the nodes belonging to the central module (Fig. 4.7*b*). We use preferential attachment [20, 21] to decide to which central node the selected nodes link to. (That is, we assume that the probability that a selected node will connect to a node $i$ of the central module is $k_i / \sum_j k_j$, where $k_i$ is the degree of node $i$ and the sum goes over all nodes of the central module.) In the second step ($n = 2$, Fig. 4.7*c*) we again create four identical copies of the 25-node structure obtained thus far, but we connect only a $p^2$ fraction of the newly added nodes to the central module. Subsequently, in each iteration $n$ the central module of size $5^n$ is replicated four times, and in each new module a $p^n$ fraction will connect to the current central module, requiring the addition of $(5p)^n$ new links.

As figure 4.8 shows, changing $p$ alters the slope of both $P(k)$ and $C(k)$ on a log-log plot. In general, we find that increasing $p$ decreases the exponents $\gamma$ and $\beta$ (Fig. 4.8*b,d*). The exponent $\beta = 1$ is recovered for $p = 1$, i.e. when all nodes of

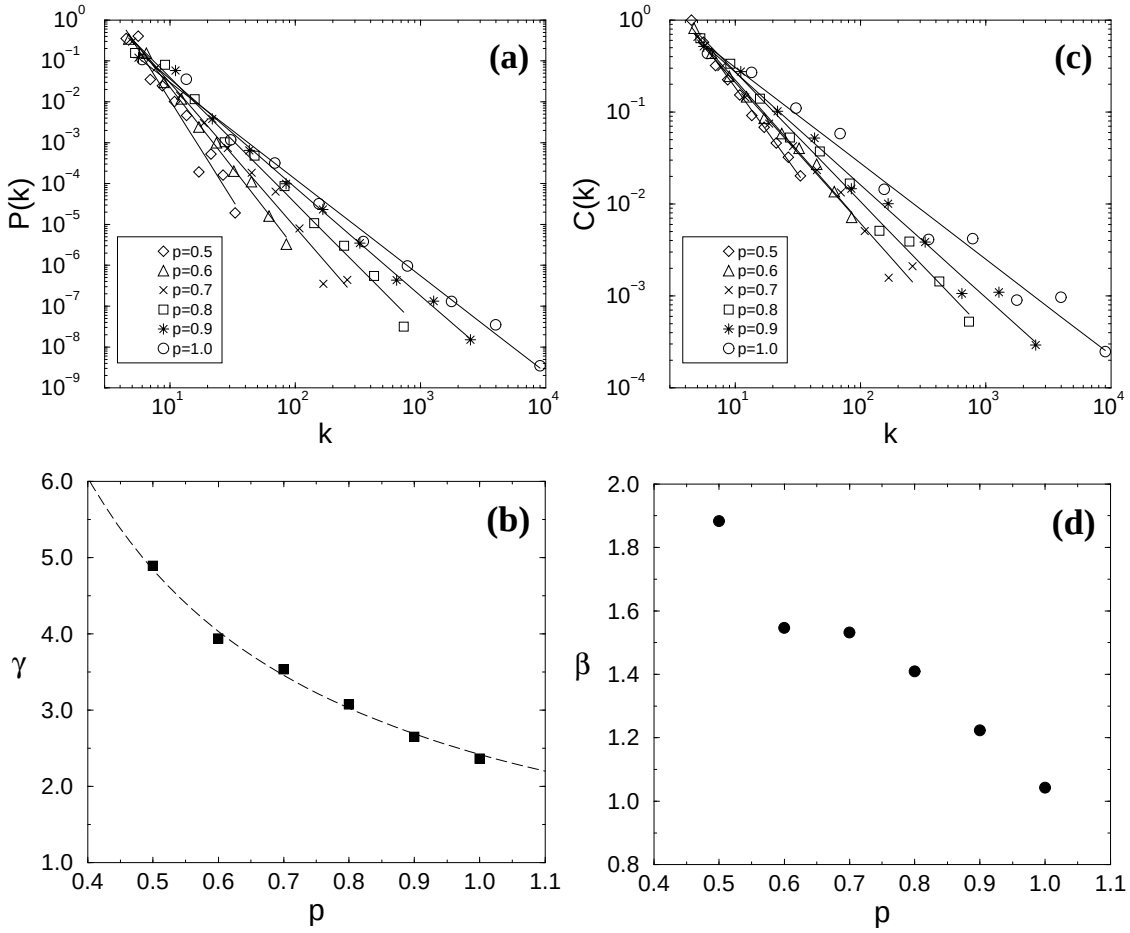



Figure 4.8. The scaling properties of the stochastic model. **(a)** The degree distribution for different $p$ values, indicating that $P(k)$ follows a power law with a $p$ dependent slope. **(b)** The dependence of the degree exponent $\gamma$ on $p$, determined by fitting power laws to the curves shown in **(a)**. The exponent $\gamma$ appears to follow approximately $\gamma(p) \sim 1/p$ (dashed line). **(c)** The $C(k)$ curve for different $p$ values, indicating that the hierarchical exponent $\beta$ depends on $p$. **(d)** The dependence of $\beta$ on the parameter $p$. The simulations were performed for $N = 5^7(78{,}125)$ nodes.

a module gain a link. While the number of links added to the network changes at each iteration, for any $p \leq 1$ the average degree of the infinitely large network is finite. Indeed, the average degree follows

$$\langle k \rangle_n = \frac{8}{5}\left(\frac{3}{2} + \frac{1-p^{n+1}}{1-p}\right), \tag{4.10}$$

which is finite for any $p \leq 1$.

## 4.5 Generality of the $C(k)$ Scaling

The scaling of $C(k)$ is not a unique property of the model discussed above. A version of the model, where we keep the fraction of selected nodes, $p$, constant from iteration to iteration, also generates $p$ dependent $\beta$ and $\gamma$ exponents. Furthermore, recently several results indicate that the scaling of $C(k)$ is an intrinsic feature of several existing growing network models. Indeed, aiming to explain the potential origin of the scaling in $C(k)$ observed for the Internet, VSPV note that the fitness model [27, 28] displays a $C(k)$ that appears to scale with $k$. While there is no analytical evidence for $C(k) \sim k^{-\beta}$ yet, numerical results [200, 201] suggest that the presence of fitness does generate a hierarchical network architecture. In contrast, in a recent model proposed by Klemm and Eguíluz there is analytical evidence that the network obeys the scaling law (4.8) [116]. In their model in each time step a new node joins the network, connecting to all *active* nodes in the system. At the same time an active node is deactivated with probability $p \sim k^{-1}$. The insights offered by the hierarchical model can help understand the origin of the observed $C(k) \sim k^{-1}$. By deactivating the less connected nodes a central core emerges to which all subsequent nodes tend to link to. New nodes have a large $C$ and small $k$, thus they are rapidly deactivated, freezing into a large $C$ state. The older, more connected, surviving nodes are in contact with a large number of nodes that have already disappeared from the active list, and they have small $C$.[2]

Finally, Szabó, Alava and Kertész have developed a rate equation method to systematically calculate $C(k)$ for evolving networks models [195]. Applying the method to a model proposed by Holme and Kim [92] to enhance the degree of clustering coefficient $C$ seen in the scale free model [20], they have shown that the

[2]Note, however, that as new nodes tend to connect to nodes that were added to the network shortly before them, the model generates a close to one-dimensional structure in time [199].

scaling of $C(k)$ depends on the parameter $p$, which governs the rate at which new nodes connect to the neighbors of selected nodes, bypassing preferential attachment. As for $p = 0$ the Holme-Kim model reduces to the scale free model, Szabó, Alava and Kertész find that in this limit the scaling of $C(k)$ vanishes. These models indicate that several microscopic mechanisms could generate a hierarchical topology, just as several models are able to create a scale free network [12, 58].

## 4.6 Discussion

The identified hierarchical architecture offers a new perspective on the topology of complex networks. Indeed, the fact that many large networks are scale free is now well established. It is also clear that most networks have a modular topology, quantified by the high clustering coefficient they display. Such modules have been proposed to be a fundamental feature of biological systems [89, 174], but have been discussed in the context of the WWW [77, 76, 127], and social networks as well [86, 210]. The hierarchical topology offers a new avenue for bringing under a single roof these two concepts, giving a precise and quantitative meaning for the network's modularity. It indicates that we should not think of modularity as the coexistence of relatively independent groups of nodes. Instead, we have many small clusters that are densely interconnected. These combine to form larger, but less cohesive groups, which combine again to form even larger and even less interconnected clusters. This self-similar nesting of different groups or modules into each other forces a strict fine structure on real networks.

Most interesting is, however, the fact that the hierarchical nature of these networks is well captured by a simple quantity, the $C(k)$ curve, offering us a relatively straightforward method to identify the presence of hierarchy in real networks. The law (4.8) indicates that the number and the size of the groups of different cohesive-

ness is not random, but follow rather strict scaling laws.

The presence of such a hierarchical architecture reinterprets the role of the hubs in complex networks. Hubs, the highly connected nodes at the tail of the power law degree distribution, are known to play a key role in keeping complex networks together, playing a crucial role from the robustness of the network [14, 44] to the spread of viruses in scale free networks [165]. Our measurements indicate that the clustering coefficient characterizing the hubs decreases linearly with the degree. This implies that while the small nodes are part of highly cohesive, densely interlinked clusters, the hubs are not, as their neighbors have a small chance of linking to each other. Therefore, the hubs play the important role of bridging the many small communities of clusters into a single, integrated network.

# CHAPTER 5

# METABOLIC NETWORKS

## 5.1 Motivation

The identification and characterization of the system-level features of biological organization is a key issue in post-genomic biology [89, 114, 215]. The concept of modularity, the assumption that cellular functionality can be partitioned into a collection of well defined units [89, 125, 172, 93, 90, 188], is a very popular paradigm of this field, attempting to connect structural elements of living systems to functions they perform. Spatially and chemically isolated molecular machines or protein complexes (such as ribosomes and flagella) are prominent examples of such functional units, but more extended modules, such as those achieving their isolation through the initial binding of a signaling molecule [15], are also apparent. Simultaneously, it is recognized that the thousands of components of a living cell are dynamically interconnected, so that the cell's functional properties are ultimately encoded into a complex intracellular web of molecular interactions [114, 215, 125, 172, 93, 188]. This is perhaps most evident with cellular metabolism, a fully connected biochemical network in which hundreds of metabolic substrates are densely integrated through biochemical reactions. Within this network, however, modular organization (i.e., clear boundaries between subnetworks) is not immediately apparent. As we noted previously in §1.4.1, the degree distribution $P(k)$ of a metabolic network decays as a power law $P(k) \sim k^{-\gamma}$ with $\gamma \simeq 2.2$ in all studied organisms [101, 206], suggesting

that metabolic networks have a scale free topology. This implies the existence of a few highly connected nodes (e.g., pyruvate or coenzyme A), which participate in a very large number of metabolic reactions. With a large number of links, these hubs seem to integrate all substrates into a single, integrated web in which the existence of fully separated modules is prohibited by definition (Fig. 5.1*a*).

Nonetheless, a number of approaches for analyzing the functional capabilities of metabolic networks indicate the existence of separable functional elements [181, 183]. Also, from a purely topological perspective, metabolic networks are known to possess a high clustering coefficients [206], a property that is suggestive of a modular organization. In itself, this implies that the metabolism has a modular topology, potentially comprising several densely interconnected functional modules of varying sizes that are connected by few intermodule links (Fig. 5.1*b*). However, such clear-cut modularity imposes severe restrictions on the degree distribution, implying that most nodes have approximately the same number of links, which contrasts with the metabolic network's scale free nature [101, 206].

In the course of this chapter we show that hierarchical modularity (Fig. 5.1*c*), described in detail in the previous chapter, reconciles all the observed properties of metabolic networks within a single framework. Moreover, the hierarchical module structure can be easily uncovered and it corresponds to known functional classification of metabolic reactions. It also gives us a new perspective on the arrangement of lethal enzymes within the various parts of the metabolism. In the next chapter we propose a simple, biologically motivated mechanism which can explain the emergence of preferential attachment in metabolic networks. The network model based on this mechanism uses only local growth rules, it nevertheless captures the mentioned topological properties.

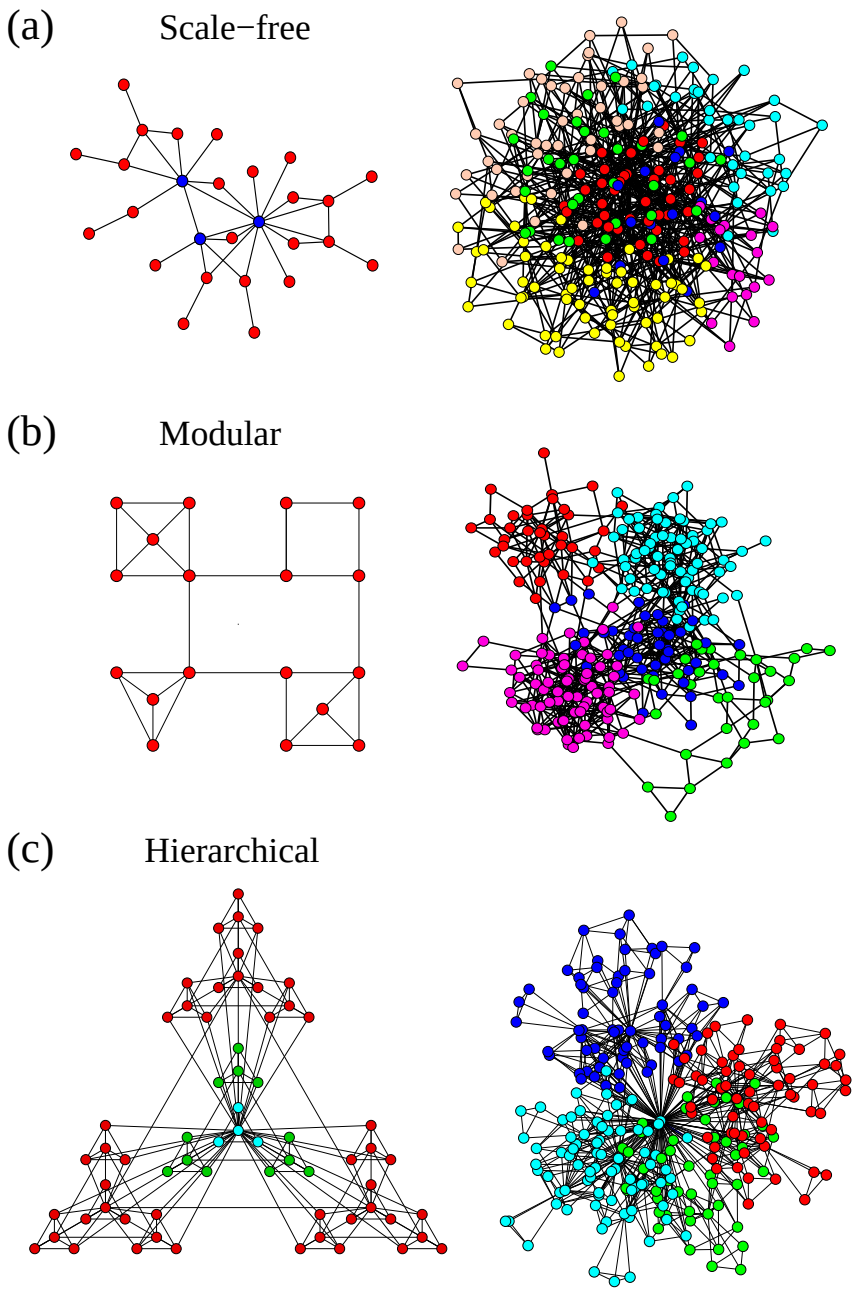



Figure 5.1. Complex network models. **(a)** Schematic illustration of a scale free network. **(b)** Schematic illustration of a manifestly modular network made of four highly interlinked modules connected to each other by a few links. **(c)** A hierarchical network with hierarchical levels represented in increasing order from blue to green to red. All 3D illustrations are networks with 256 nodes, arranged in space with a standard graph drawing algorithm [78].

## 5.2 Hierarchy in Cellular Metabolism

### 5.2.1 Definition of the Metabolic Network

Metabolic reactions can be mapped onto a network in a few different ways, for example one can represent the metabolism as a substrate graph where links are defined by reactions, or one can use the dual reaction graph where links are shared substrates of two different reactions.

In our metabolic network representation substrates are the nodes of the network, while the links connect all in-coming substrates (educts) of a reaction to all its outgoing substrates (products) [101] (Fig. 5.2*a*). The *E. coli* metabolic network defined this way has $N = 885$ nodes, and it can be visualized using a standard clustering graph drawing algorithm built into the Pajek software [78] (Fig. 5.2*b*)

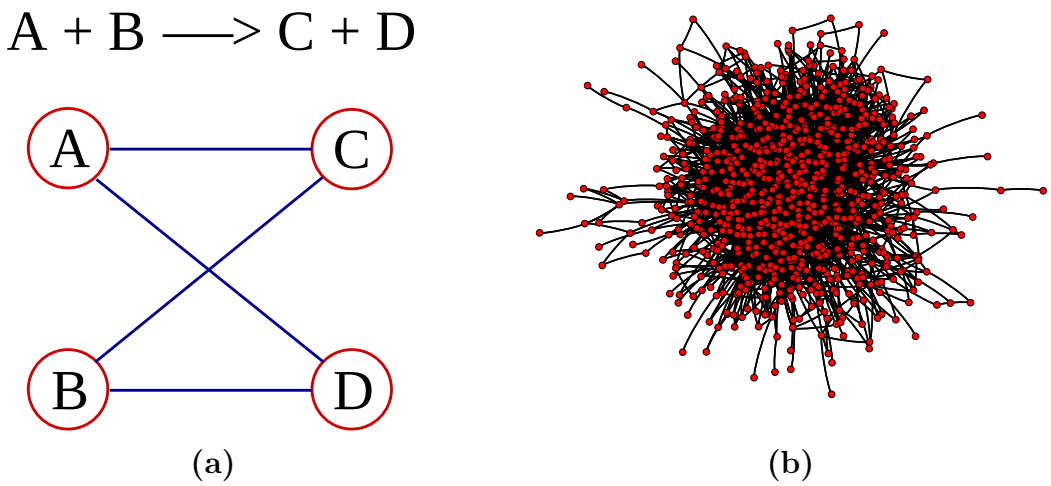



Figure 5.2. **(a)** Graph theoretic representation of a reaction in a metabolic network. **(b)** The complete *E. coli* metabolic network.

### 5.2.2 Clustering in Metabolic Networks

To determine whether strong clustering along with the known scale free topology is indeed a generic property of all metabolic networks, we first calculated the average

clustering coefficient for 43 different organisms [101, 160] as a function of the number of distinct substrates, $N$, present in their metabolism. We found that, for all 43 organisms, the average clustering coefficient is about an order of magnitude larger than that expected for a scale free network of similar size (Fig. 5.3), suggesting that metabolic networks in all organisms are characterized by a high intrinsic potential modularity. We also observed that, in contrast with the prediction of the scale free model, for which the clustering coefficient decreases as $(\ln N)^2/N$ [116, 30], the clustering coefficient of metabolic networks is independent of their size (Fig. 5.3).

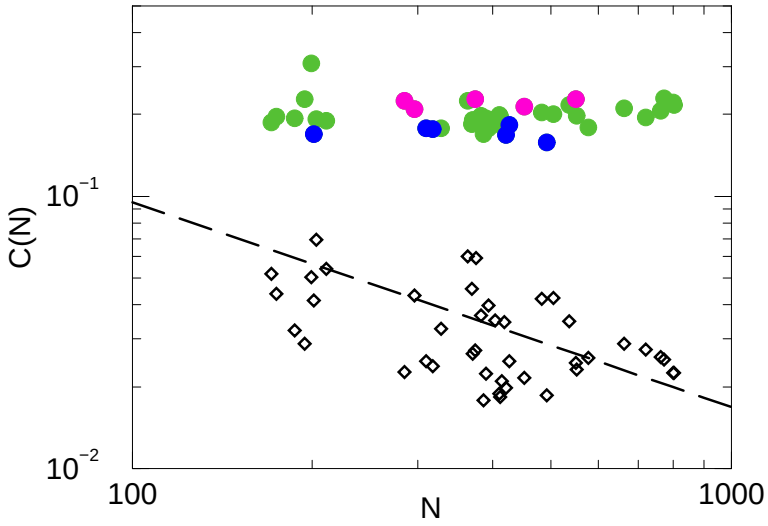



Figure 5.3. The average clustering coefficient for 43 organisms [101] is shown as a function of the number of substrates $N$ present in each of them. Species belonging to *archaea* (purple), *bacteria* (green), and *eukaryotes* (blue) are shown. The dashed line indicates the dependence of the clustering coefficient on the network size for a module-free scale free network, and the diamonds denote $C$ for a scale free network with the same parameters ($N$ and number of links) as observed in the 43 organisms.

These results demonstrate a fundamental conflict between the predictions of previous models of metabolic organization. The high, size-independent clustering coefficient offers strong evidence for modularity, whereas the power law degree distribution of all metabolic networks [101, 206] strongly supports the scale free model

and rules out a manifestly modular topology.

To investigate whether hierarchical organization is present in cellular metabolism we measured the $C(k)$ function for the metabolic networks of all 43 organisms.

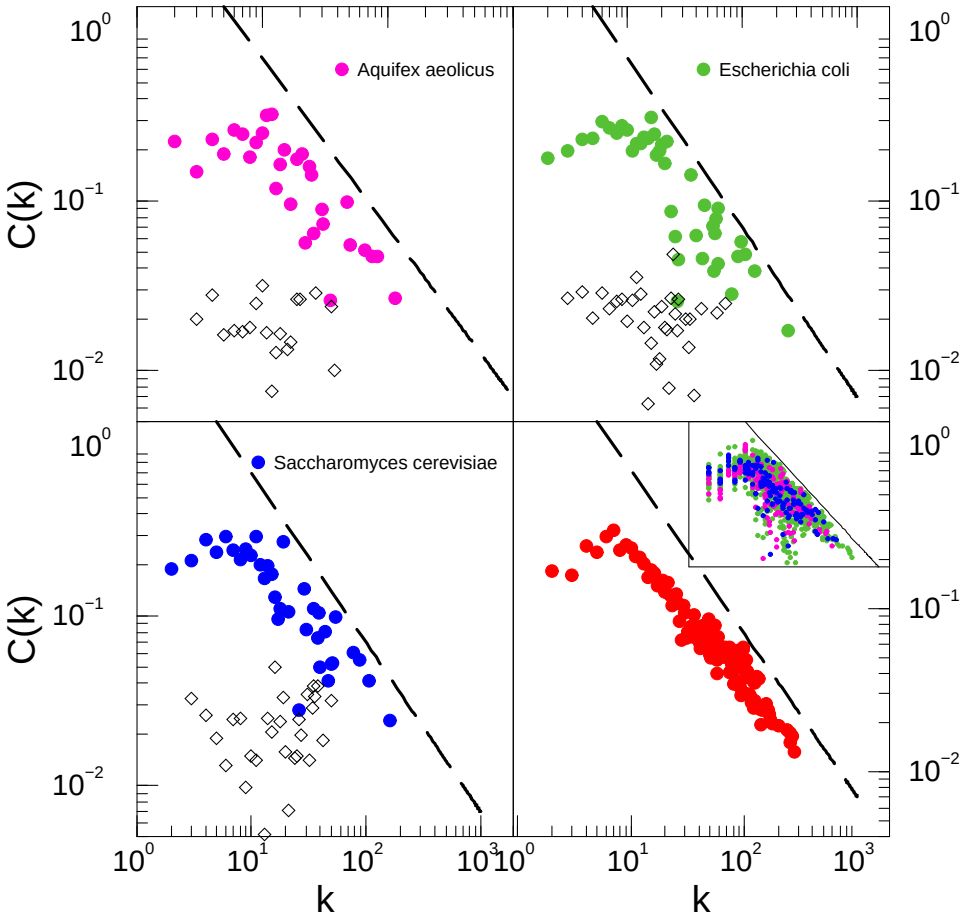



Figure 5.4. Dependence of the clustering coefficient on the node's degree in three organisms: **(a)** *Aquidex aeolicus* (archaea), **(b)** *Escherichia coli* (bacterium), **(c)** and *Saccharomices cerevisiae* (eukaryote). In **(d)** the $C(k)$ curves averaged over all 43 organisms are shown, while the inset displays all 43 species together. The dashed lines correspond to $C(k) \sim k^{-1}$, and in **(a–c)** the diamonds represent $C(k)$ expected for a scale free network (Fig. 5.2*a*) of similar size, indicating the absence of scaling. The wide fluctuations are due to the small size of the network.

As shown in figure 5.4, for each organism $C(k)$ is approximated by $C(k) \sim k^{-1}$, in contrast to the $k$–independent $C(k)$ predicted by both the scale free and modular networks. This provides direct evidence for an inherently hierarchical organization.

## 5.3 The *E. Coli* Metabolic Network

A key issue from a biological perspective is whether the identified hierarchical architecture reflects the true functional organization of cellular metabolism. To uncover potential relationships between topological modularity and the functional classification of different metabolites we concentrate on the metabolic network of *Escherichia coli*, whose metabolic reactions have been exhaustively mapped and studied, both biochemically and genetically [109].

### 5.3.1 Generating the Reduced *E. Coli* Metabolic Network

The *E. coli* network shown in figure 5.2 is very dense and it looks much like the module-free scale free network shown in figure 5.1*a*. A partial reason for this is the limitation of a three-dimensional spacial arrangement that is ill suited for illustration of dense webs. There is also a biological reason specific to the metabolic network. Namely, metabolism relies heavily on the usage of a few substrate pairs which undergo very generic chemical changes in a large number of reactions of all types. A representative example is the ATP–ADP pair, the cell's energy fuel molecules. As a phosphate group is broken off ATP (adenosine-triphosphate), the energy released from the chemical bond fuels the chemical change of the substrate(s) ATP reacts with. This mechanism is so generic that ATP and ADP are the greatest hubs of our network: they are linked to a significant fraction of all substrates.

*Biochemical Reduction*

Here we describe a method by which we account for the above mentioned peculiarity of metabolism. A link from ATP, ADP, water, etc. to a metabolite $A$ often carries little biologically relevant information about the function of $A$. There are many different reactions where other pairs of metabolites help some reactions to

take place: exchange of a proton or a methyl group, for example.

In order to focus on biologically relevant substrate transformations, we have performed a biochemical reduction of the metabolic network. Our guiding principle was to maintain the main line of substrate transformation on each pathway. In figure 5.5 we illustrate the reduction process, showing an original pathway map (left), the network corresponding to it (middle), and the network obtained after the reduction process (right).

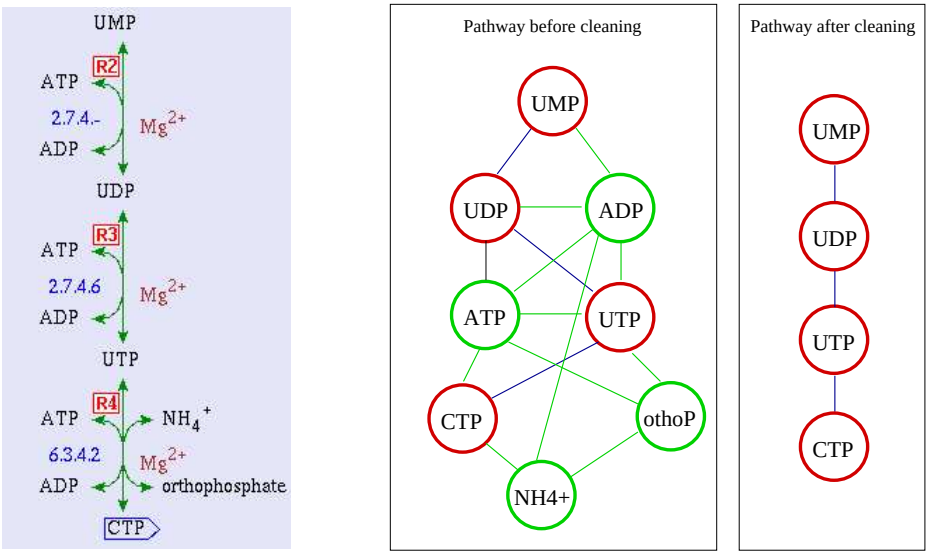



Figure 5.5. Biochemical reduction of the pathways of the metabolic network. The middle panel shows the full graph theoretic representation of the pathway shown in the left panel. The right panel displays the pathway after biochemical reduction.

It is important to note that the reduction process is completely local: it takes place at the level of each reaction, and does not result in the removal of metabolites, but only in the removal of links from the graph representation.

The resulting biochemically reduced metabolic network for *E. coli* is shown in figure 5.6*a*.

*Topological Reduction*

To further reduce the complexity of the metabolic graph we continue with a two-step topological reduction.

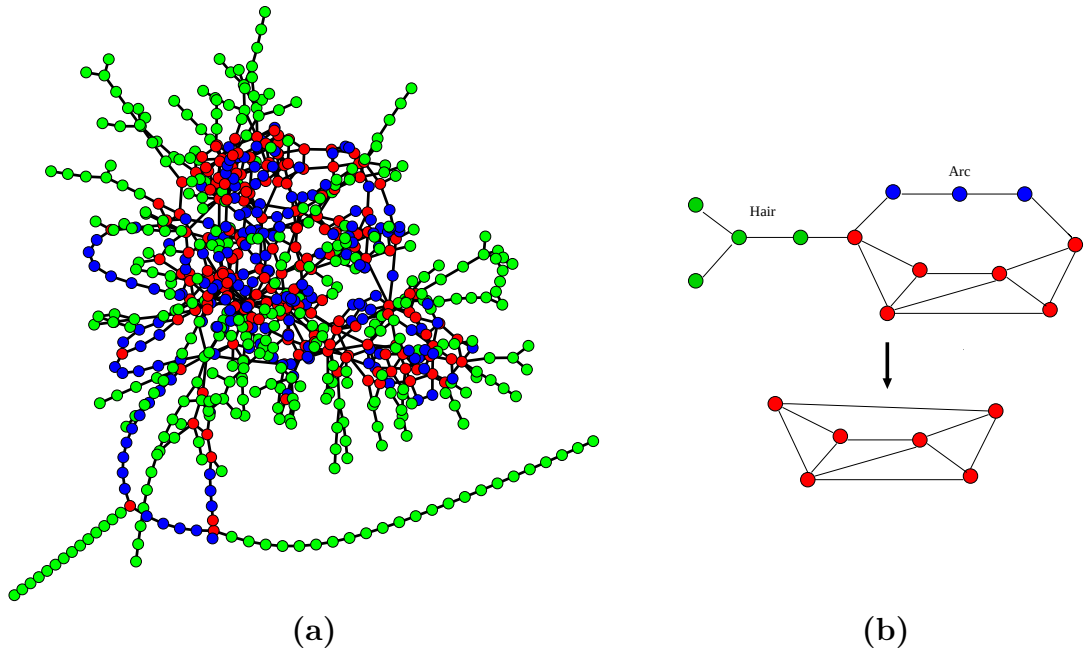
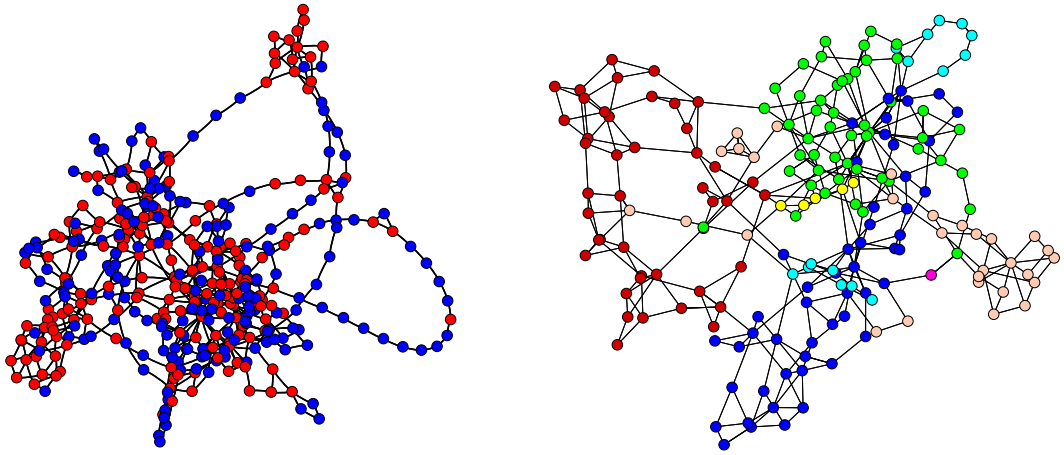



Figure 5.6. **(a)** The *E. coli* metabolic network after biochemical reduction. **(b)** Topological reduction, which implies temporally removing all *hair* (green), and replacing each *arc* (blue) with a single link.

As figure 5.6*a* shows, many pathways uncovered by the first reduction are connected to the rest of the metabolic network by a single substrate (green parts), or represent a long chain of consecutive substrates that appear as an arc between two substrates, and have no other side connections (blue arcs). Since the topological location of the strings of substrates depend only on one or two multiply connected terminal substrates (red), we can temporarily remove the elements of the long non-branching pathways without altering the topology of the core metabolism.

We define *hairs* (green on Fig. 5.6) as all sets of nodes that can be separated from the network by cutting one link. An *arc* (blue on Fig. 5.6) is an array of

nodes connected by only two links to the rest of the metabolism, leading from one well-connected substrate (red on Fig. 5.6) to another. To generate the reduced metabolic network we have removed all hairs from the network and replaced all arcs with a single link, directly connecting the substrates at the two ends of an arc.[1] A schematic topological reduction can be seen on figure 5.6.



**(a)** No hair. **(b)** No hair, shortened arcs.

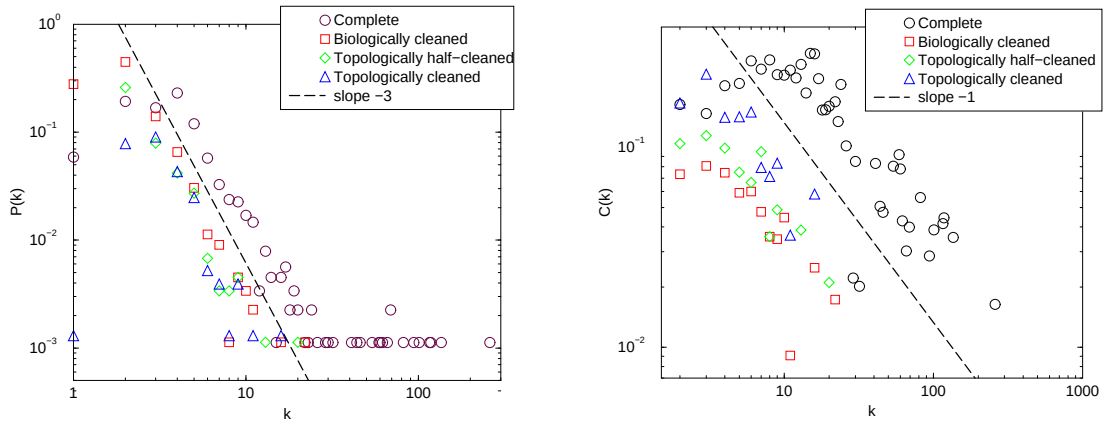
Figure 5.7. Topological reduction of the metabolic network. Starting from the biochemically reduced metabolic network shown in figure 5.6*a*, we removed all *hair* **(a)** and *arcs* **(b)** from the network. The color code of the nodes in the final figure denotes the corresponding substrate's functional class (See Fig. 5.11).

While the substrates removed during the topological reduction process are biologically important components of the network, their removal does not change the way subunits that they were removed from connect to other parts of the metabolism. In this sense, they are topologically irrelevant.[2] The result of this two step reduction of the *E. coli* metabolism is shown on figure 5.7.

[1]Note that we do not repeat the above described process on the reduced network. Thus, after the reduced network is ready, it can have arcs and hairs in it (see Fig. 5.7*b*, light blue arc in the top right corner). These appear, for example, when two linked "red" nodes both have hair on them, so they both have three links. After the reduction they are left with two links and thus are parts of a newly created arc.

[2]Removed substrates are later re-added for a final biological analysis (Fig. 5.13).

Both the biological and topological reduction process affects the connectivity and the clustering coefficient of the nodes, so it is important to note that these processes do not change the large-scale properties of the metabolic network. Figure 5.8 shows the degree distribution and the clustering coefficient of the metabolic network obtained during the different reduction stages. As the figure shows, the scaling of $P(k)$ and $C(k)$ remains largely unchanged during this process. This is not unexpected, as the reduction is purely a local process, which does not alter the networks's large-scale features.




**(a)** $P(k)$. **(b)** $C(k)$.

Figure 5.8. Degree distribution and clustering properties of all reduction stages of the *E. coli* metabolic network.

### 5.3.2 Finding the Hierarchically Embedded Modules

After reducing the metabolic network to a representative core, we proceed to break it up into clusters based on its wiring diagram.

*The Topological Overlap Matrix*

In order to quantify whether two nodes are closely linked into the same local cluster, we introduce the *topological overlap matrix*, $O_T(i, j)$. Topological overlap of 1 between substrates $i$ and $j$ implies that they are connected to the same substrates,

whereas a 0 value indicates that $i$ and $j$ do not share a link, nor links to common substrates among the metabolites they react with. Defining the adjacency matrix, $l_{i,j}$, of the network as

$$l_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected} \\ 0 & \text{if } i \text{ and } j \text{ are not connected} \end{cases}, \tag{5.1}$$

the elements of the overlap matrix are given by

$$O_T(i,j) = \frac{\sum_{l=1}^{N} l_{i,l} \cdot l_{j,l} + l_{i,j}}{\min(k_i, k_j) + 1 - l_{i,j}}. \tag{5.2}$$

The overlap matrix for the *E. coli* metabolic network, with alphabetically ordered substrates, is shown on figure 5.9*a*. There is some grouping of the overlap values of nearby nodes, due to the fact that similarly named metabolites often have related functions.

As the topological overlap matrix is expected to encode the comprehensive functional relatedness of the substrates forming the metabolic network, we investigated whether potential functional modules encoded in the network topology can be automatically uncovered.

*Hierarchical Clustering Algorithm*

We choose the un-weighted average linkage algorithm (or Un-weighted Pair Group Method with Arithmetic Mean) known as UPGMA [191, 63] for our hierarchical clustering method.[3] This algorithm first finds the largest overlap present in the matrix, joins the corresponding substrates $u$ and $v$ to a branching point on the tree, and substitutes them with a "new" cluster $\{u, v\}$. This new unit replaces the original $u$ and $v$ in the overlap matrix. It has an overlap with an arbitrary substrate

[3]Other approaches to discern modules in metabolic networks are based on the idea that edges along a large number of shortest paths are likely to link different modules of the network [92, 84]. Edges on the largest number of paths were iteratively removed, slowly breaking the network into its functional modules.

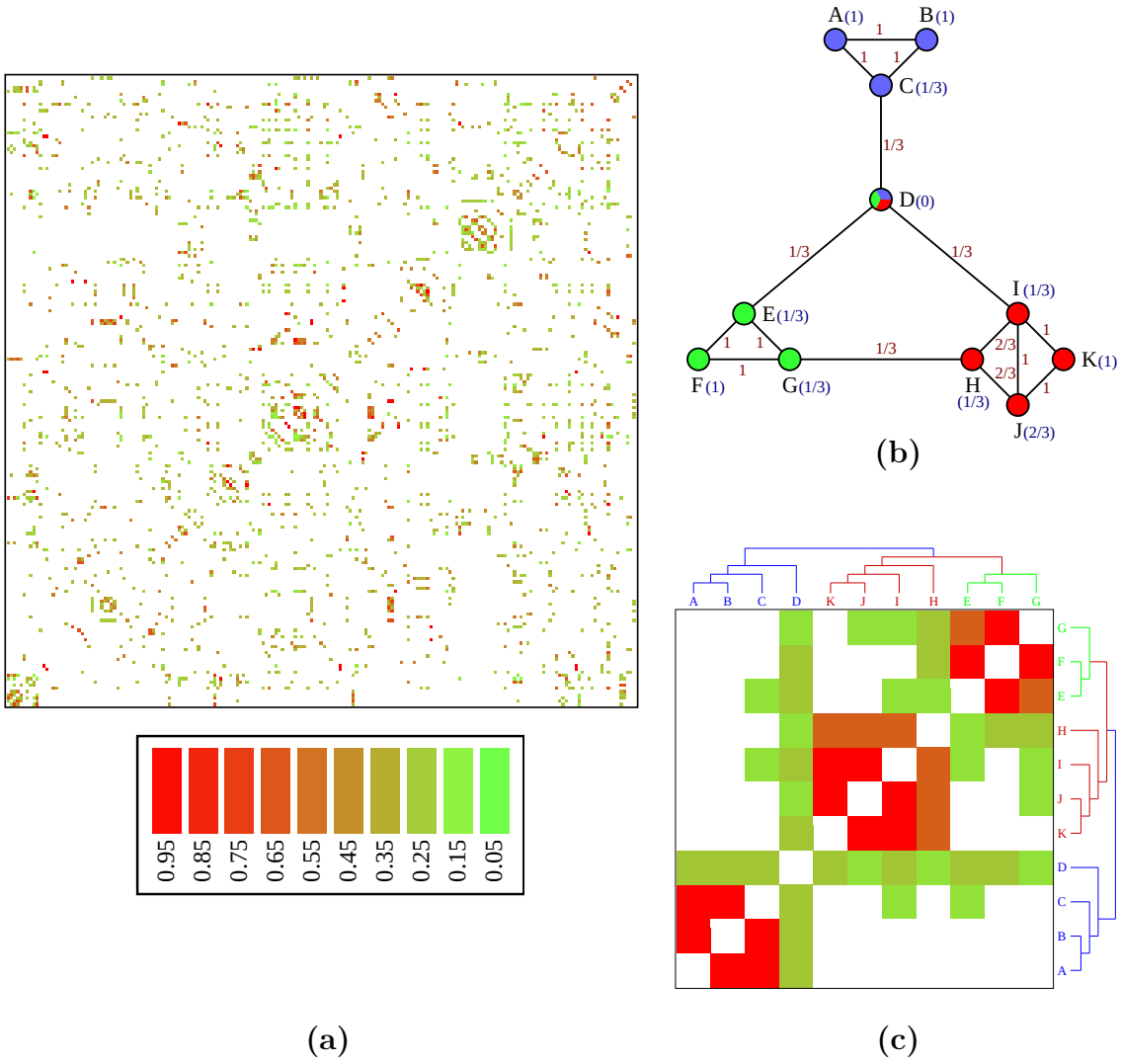



Figure 5.9. **(a)** Overlap matrix of alphabetical ordered substrates in the *E. coli* metabolic network. **(b)** Topological overlap illustrated on a small hypothetical network. On each link we indicate the topological overlap for the connected nodes and in parenthesis next to each node we indicate its clustering coefficient. **(c)** The topological overlap matrix corresponding to the small network shown in **(b)**. The rows and columns of the matrix were reordered by the UPGMA clustering method [191, 63], allowing us to identify and place close to each other those nodes that have high topological overlap. The matrix color code denotes the degree of topological overlap between the nodes (see side-bar on **(a)**).

(cluster) $w$ given by

$$O_T(\{u,v\},w) = \frac{n_u \cdot O_T(u,w) + n_v \cdot O_T(v,w)}{n_u + n_v}, \tag{5.3}$$

where $n_u$ is the number of components in cluster $u$. This definition ensures that all original overlap values are represented with the same weight in the overlap value of the joint cluster, hence the method's name, "un-weighted average linkage clustering." Repeating this rule eventually shrinks the overlap matrix to a single unit, corresponding to the root of the hierarchical tree. Thus, we obtain a tree with all the original substrates as its end-leafs, grouped naturally on branches reflecting their hierarchical overlap. When overlap values between clusters are redundant (i.e. there are at least two groups of clusters with the same overlap value) the program automatically joins the pair found first. The ordering of two branches under a junction is irrelevant, thus arbitrary. The distance between (height of) two junction levels is defined to be one.

First we tested the clustering algorithm on the small hypothetical network shown in figure 5.9*b*. The method placed those nodes that have a high topological overlap close to each other (Fig. 5.9*c*), correctly identifying the the three distinct modules built into the model of figure 5.9*b*. It also identified the relationship between the three modules, as EFG and HIJK are closer to each other in a topological sense than the ABC module (Fig. 5.9*b*).

### 5.3.3 Modules of the *E. coli* Metabolic Network

The clustering of the *E. coli* metabolic network, and thus the ordering of the overlap matrix according to a substrate's horizontal location on the hierarchical tree, lead to figure 5.10. This figure provided us a global topological representation of the metabolism.

Groups of metabolites forming tightly interconnected clusters are visually appar-

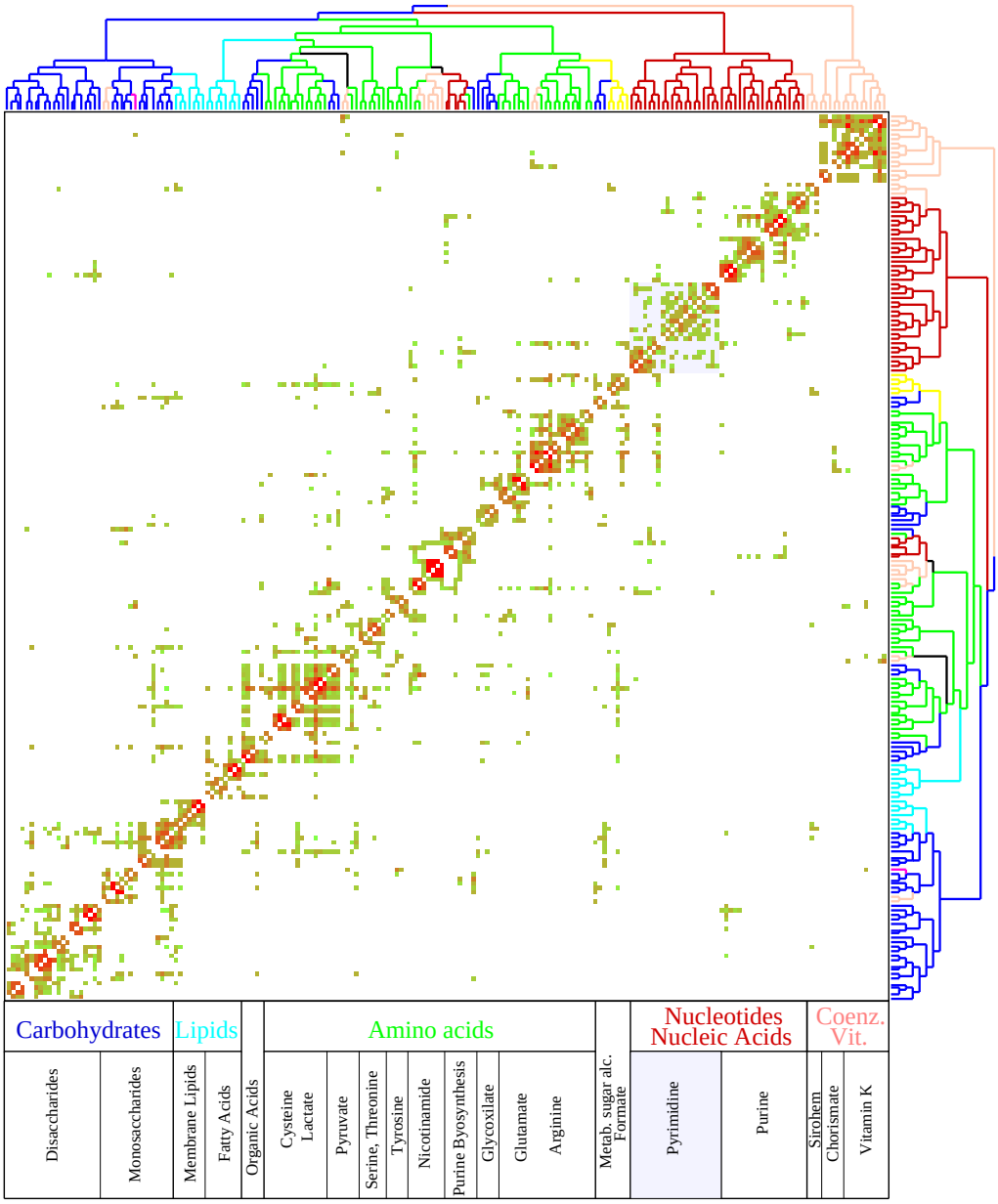



Figure 5.10. Topological overlap matrix corresponding to the *E. coli* metabolism, together with the corresponding hierarchical tree (*top and right side*) that quantifies the relationship between the different modules. The branches of the tree are color coded to reflect the functional classification of their substrates. The color code of the matrix denotes the degree of topological overlap shown in the matrix (See scale on Fig. 5.9*a*). On the bottom we show the large-scale functional map of the metabolism, as suggested by the hierarchical tree.

ent along the diagonal line of the matrix, and upon closer inspection the hierarchy of nested topological modules of increasing sizes and decreasing interconnectedness can also be seen.

To visualize the relationship between topological modules and the known functional properties of the metabolites, we color coded the branches of the derived hierarchical tree according to the predominant biochemical class of the substrates it produces, using a standard, small molecule biochemistry-based classification of metabolism [160]. The biochemical classes we used to group the metabolites represent carbohydrate metabolism (*blue*), nucleotide and nucleic acid metabolism (*red*), protein, peptide and amino acid metabolism (*green*), lipid metabolism (*cyan*), aromatic compound metabolism (*dark pink*), monocarbon compound metabolism (*yellow*) and coenzyme metabolism (*light orange*).

To our pleasant surprise, the color coding of the hierarchical tree according to biochemical classification of the metabolites proved a very good agreement between the uncovered modular hierarchy and the standard classes of the metabolism. As shown in figure 5.10, and in the three-dimensional representation of the core network in figure 5.11, we find that most substrates of a given small molecule class are distributed on the same branch of the tree (Fig. 5.10) and correspond to relatively well-delimited regions of the metabolic network (Fig. 5.11). Therefore, there are strong correlations between shared biochemical classification of metabolites and the global topological organization of *E. coli* metabolism (Fig. 5.10, bottom).

At the highest level, we find that the *E. coli* metabolic network is partitioned into three large classes, appearing as major branches on the tree.

1. The smallest of these branches consists of the **Coenzyme and Vitamin Metabolism** (light orange), its inner core is divided into *Vitamin K*- and *Terpene Metabolism*, while its outer part is specific to *Sirohem Anabolism*

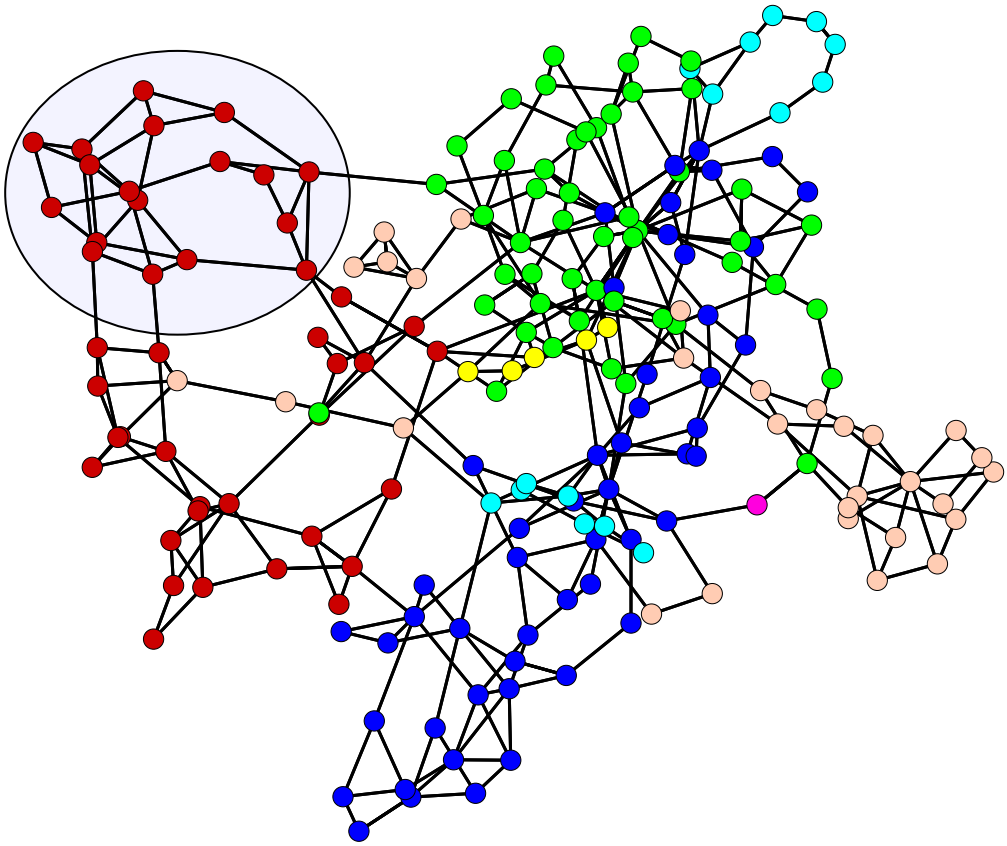


Figure 5.11. 3-D representation of the reduced *E. coli* metabolic network. Each node is color coded by the functional class to which it belongs, and is identical to the color code applied to the branches of the tree shown in figure 5.10. Note that the different functional classes are visibly segregated into topologically distinct regions of metabolism. The blue-shaded region denotes the nodes belonging to pyrimidine metabolism, discussed below.

(See Fig. 5.12 for a larger view of the hierarchical tree with the corresponding functional map).

2. The second major branch represents the **Nucleotide and Nucleic Acid Metabolism** (red). Its two major sub-branches are clearly divided into the *Pyrimidine* and *Purine Metabolism*. Interestingly, the Purine group has a small sub-branch representing Dihydrofolate Anabolism, a subgroup that is shared with the Coenzyme and Vitamin Metabolism (light orange). Its strong

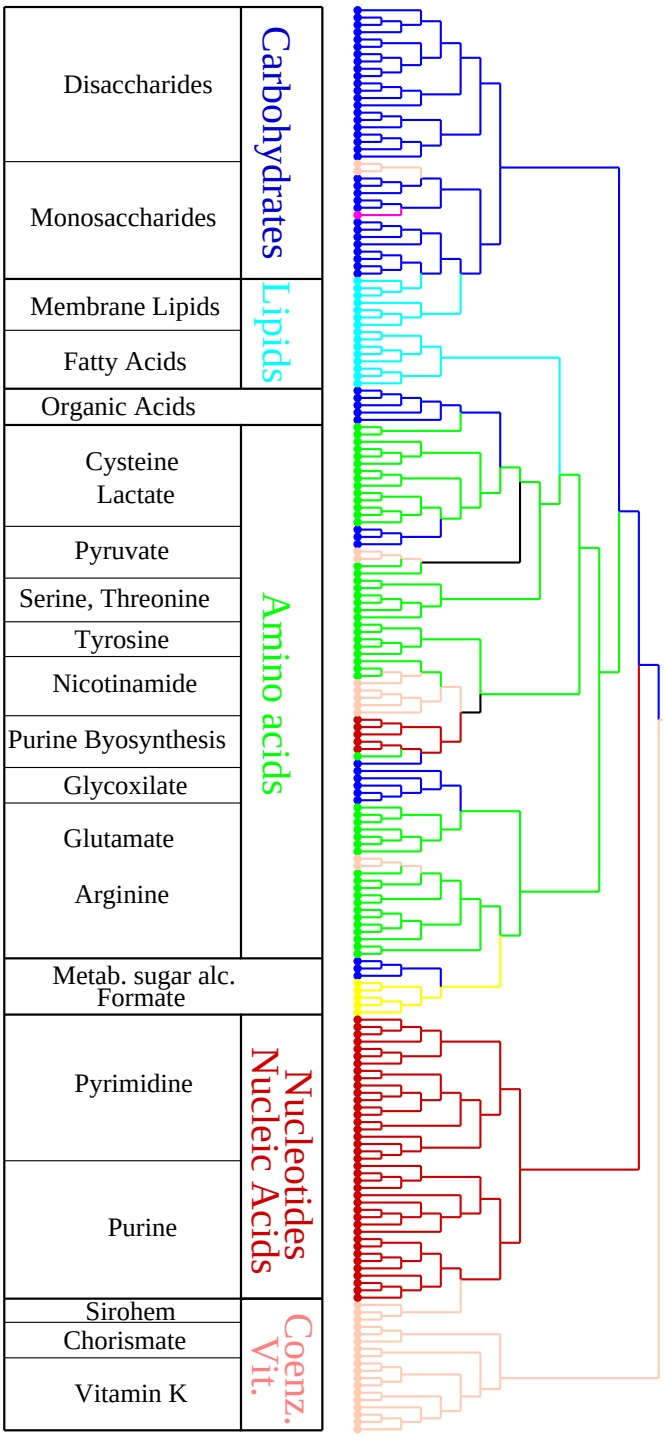



Figure 5.12. Hierarchical tree representing the reduced *E. coli* metabolic network.

link to Purine Metabolism is due to the dihydroneopterin-triphosphate synthesis pathway from guanosine-triphosphate.

3. The third and largest branch naturally breaks into a smaller branch containing largely **Carbohydrate Metabolism** (blue), and a second, less segregated one. This branch, in addition to **Proteins, Peptides and Amino Acids** group (PPA, green), also contains several apparently unrelated pathways.

   (a) The **Carbohydrate** branch contains most of the *Poly-* and *Disaccharides* on the sub-branch on the left; while *Monosaccharides* (some of which are also present on the left branch), *Sugar Alcohols*, and *Alcohol* metabolites dominate the right branch. *Membrane Lipid Metabolism* (cyan), which is fairly independent of the other **Lipid** group, the Fatty Acid Metabolism, is nested into the Carbohydrates branch due to shared glycerol metabolism pathways used in its biosynthesis. A small group representing Pyridoxine Anabolism (Vitamins: Vitamin 6B, light orange) is linked into this branch via biosynthesis from D-erythrose-4P. Another small nested group is the 3-phosphoshikimate biosynthesis from D-erythrose-4P, a part of Chorismate Metabolism shared by both the Aromatic Compounds Metabolism (dark pink), and the Coenzyme group.

   (b) On the **Proteins, Peptides and Amino Acids** group (PPA, green) a clear and separate sub-branch at the left side represents *Fatty Acid Metabolism* (part of **Lipid Metabolism**), strongly linked to the *Organic Acids* and the *Citrate Cycle* (Carbohydrates, blue). Since almost half of the Amino Acid class substrates are shared with Carbohydrates Metabolism, pathways belonging to *Pyruvate*, *Glyoxylate* and *Metabolism Sugar Alcohols* are naturally grouped within the PPA group, appearing

as small red branches on the figure. *Formate Metabolism*, which represents almost the complete **Monocarbon Compounds Metabolism** class (yellow) is linked to Metabolism Sugar Alcohols. The IMP anabolic pathway (part of Purine metabolism, blue) starts with 5-phospho-'alpha'-D-ribose-1-diphosphate and the substrates on this pathway diverge from Purine metabolism, and are grouped on the PPA branch. Similarly, *Nicotinamide Metabolism* (Coenzymes, light orange) is grouped into the PPA branch due to NAD(+) biosynthesis from L-aspartate. Enterobactin biosynthesis from the Chorismate pathway links parts of Chorismate metabolism (Coenzymes, light orange) to L-serine, the small (2 substrate) insert next to the Pyruvate group (a small blue group, shared by Carbohydrates and PPA). The pathway leading from L-Glutamate to L-glutamate-1-semialdehyde is part of Lipid, Aromatic Compounds and Coenzymes metabolism, its links anchor it into Glutamate metabolism. The PPA substrates on this large branch tend to group according to classifications based on the names of the amino acids, but not all of them show up on distinguishable sub-branches. They tend to group internally as well, for example most of glutamate and arginine metabolism substrates can be found on the same branch.

### 5.3.4 Biochemical Pathways in the Pyrimidine Module

To correlate modules obtained from our graph theory-based analysis to actual biochemical pathways, we concentrated on pathways involving the Pyrimidine Metabolites. Our method divided these pathways into four modules (Fig. 5.13, which represent a topologically well-limited area of *E. coli* metabolism (Fig. 5.11, *light-blue circle*).

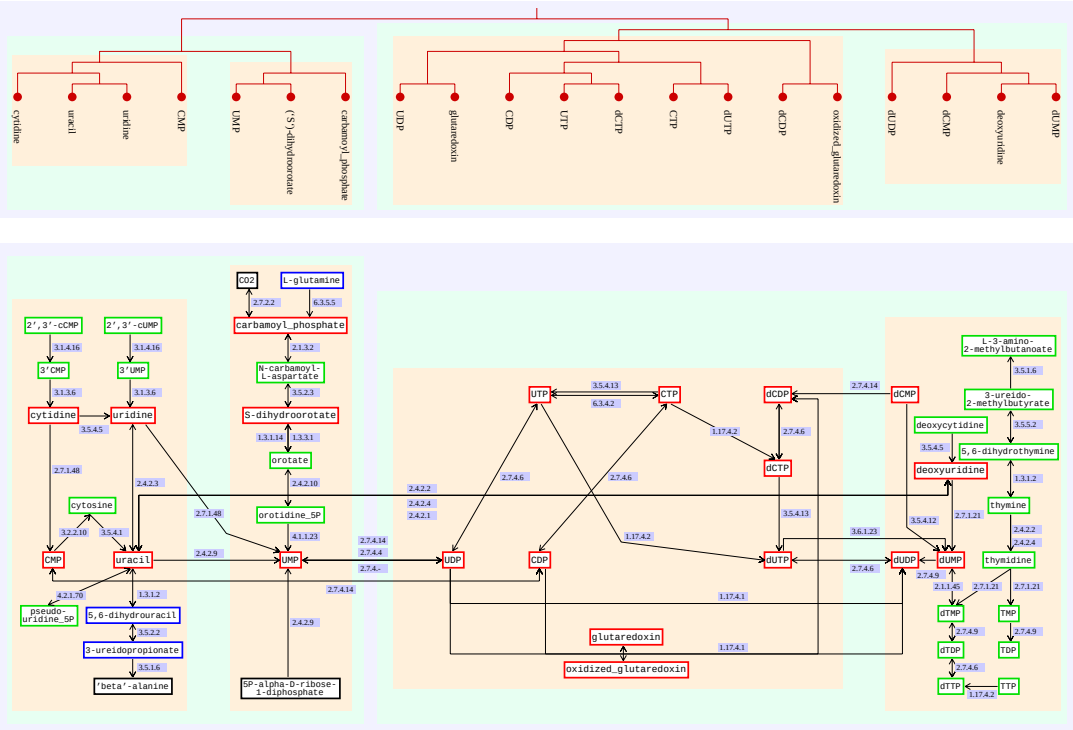



Figure 5.13. A detailed diagram of the metabolic reactions that surround and incorporate the pyrimidine metabolic module. Red boxes denote the substrates directly appearing in the reduced metabolism and the tree shown in figure 5.11. Substrates in green boxes are internal to pyrimidine metabolism, but represent members of non-branching pathways or end pathways branching from a metabolite with multiple connections. Blue and black boxes show the connections of pyrimidine metabolites to other parts of the metabolic network: Black boxes denote core substrates belonging to other branches of the metabolic tree figure 5.10, while blue boxes denote non-branching pathways (if present) leading to those substrates. Shaded boxes under reactions highlight modules suggested by the hierarchical tree. Shaded blue boxes along the links display the enzymes catalyzing the corresponding reactions, and the arrows show the direction of the reactions according to the WIT metabolic maps [160].

As shown in figure 5.13, all highly connected metabolites (red boxes) correspond to their respective biochemical reactions within pyrimidine metabolism, together with those substrates that were removed during the original network reduction procedure, and then re-added (Fig. 5.13, green boxes). However, it is also apparent that putative module boundaries do not always overlap with intuitive 'biochemistry-

based' boundaries. For instance, while the synthesis of UMP from L-glutamine is expected to fall within a single module based on a linear set of biochemical reactions, the synthesis of UDP from UMP leaps putative module boundaries.

### 5.3.5 Conclusions

The organization of metabolic networks is likely to combine a capacity for rapid flux reorganization with a dynamic integration with all other cellular function [206]. Our results indicate that the system-level structure of cellular metabolism is best approximated by a hierarchical network organization with seamlessly embedded modularity. In contrast to current, intuitive views of modularity (Fig. 5.1*b*) which assume the existence of a set of modules with a non-uniform size potentially separated from other modules, we find that the metabolic network has an inherent self-similar property: there are many highly integrated small modules, which group into a few larger modules, which in turn can be integrated into even larger modules. This is supported by visual inspection of the derived hierarchical tree (Fig. 5.10), which offers a natural breakdown of metabolism into several large modules, which are further partitioned into smaller, but more integrated sub-modules.

## 5.4 Lethality of the Metabolic Modules

Defining which gene products play an essential role and under what condition, is vital to understanding the complexity of living organisms. Although methods to rapidly and systematically determine genome-wide gene essentiality[4] are less advanced than other functional genomic techniques, a number of essentiality surveys involving different species have been reported. Many experimental approaches have been used to produce such data including individual gene knockouts in *S. cere-*

[4] *Essential* or *lethal* genes, defined relative to a certain condition the organism is placed in, are genes the absence of which causes the organism to die.

*visiae* [82, 214] and *C. elegans* [110], RNA interference in *C. elegans* [107] and whole-genome transposon mutagenesis studies in several microorganisms. In the latter group, complete or extensive lists of essential and expendable genes are available for *M. pneumoniae* and *M. genitalium* [95], *H. influenzae* [9], and *S. cerevisiae* [179]. However, relatively little effort has been committed to a system-level interpretation of these data in terms of cellular function or evolutionary relationships with other organisms [103]. *Escherichia coli* has historically been the focus of intense biochemical, genetic and physiologic scrutiny, but genomic essentiality data for this organism has remained incomplete.

Our close collaborators at the Northwestern University in Chicago, lead by Prof. Zoltán Oltvai, performed a genome-wide, comprehensive experimental assessment of the *E. coli MG1655* genes necessary for robust aerobic growth in rich medium [80]. Of the $4,291$ protein-coding genes known in *E. coli* they assessed the essentiality of $3,746$ genes ($\sim 87\%$). Using the data generated by the experimental group, first we demonstrated that essential genes have a significant tendency to be preserved by evolution throughout the bacterial kingdom, especially for a subset of genes representing key cellular processes such as DNA replication and protein synthesis. Next, we analyzed the essentiality of metabolic enzymes from the perspective of cellular system-level organization, demonstrating an enrichment of those enzymes that catalyze reactions within evolutionary conserved topological modules in the metabolic web of *E. coli* [80].

### 5.4.1 Experimental Procedure

The genetic footprinting technique, originally described for yeast [190], is an efficient experimental approach that allows the simultaneous study of thousands of

genes under various conditions[5] [81].

- **Transposome insertion.** The method begins with the introduction of a special short piece of DNA into a large number of living cells. Called the *transposome*, this piece of DNA has the ability to insert itself into the cellular DNA at a random position. Using very low concentrations of the transposome one can make sure that at most one insertion occurs per cell.

- **Selective Cell Growth.** The next step in the process is to study the behavior of cells affected by the insertions. Half of the treated cell culture is frozen and stored immediately after the transposome insertion, the other half is grown in rich medium for several generations. Cells hit by the insertion inside essential coding regions can not perform certain functions, transcribe certain proteins and die out of the population.

- **Polymerase Chain Reaction.** The surviving population is then subject to an experimental procedure that prepares the extracted cellular DNA for identification of the insertion position along the genome. This procedure is called *Polymerase Chain Reaction* (PCR), and it takes advantage of the natural ability of a single-strand DNA to assemble its complementary strand. In order for this process to occur, an initiation point is needed: a DNA region where the strand is already complemented (a short region on the single-strand DNA where base-pairing already occurred). After DNA is taken out of cells, it is inserted in a mix containing large quantities of free base molecules (A,T,G,C) and denatured (heated until it breaks into two strands). Then, two special single-strand DNA sequences are added: *primers* complementary to opposite strands of a duplex DNA. These are designed sequences that bind to the two

[5]For a more detailed description and experimental protocols, visit `http://www.umsl.edu/%7Ebalazsi/JBact2003/genetic.html`.

genomic strands in selected locations: they mark the beginning and the end of the region being amplified. In the present experiment one primer binds to the inserted transposome sequence, the other binds to a chosen *E. coli* gene (the experiment is performed separately with primers designed for all genes). After the binding of the two primers both DNA strands are complemented in the $5' \rightarrow 3'$ direction.[6] PCR consists of multiple cycles of denaturation of the synthesized DNA, annealing of the primers (lowering the temperature until base-pairing is favorable and the primers bind to the DNA) and synthesis of the complementary DNA chains. An exponential amplification of the fragment whose ends are defined by the $5'$ ends of the two primers occurs, while the longer, original template sequence is amplified at most linearly. In our case segments of the chosen gene between its starting point and all the insertions that are still present in the grown population are amplified.

- **Determining the insertion points.** DNA sequences resulted in the PCR were size-separated on agarose gels using a technique called *gel electrophoresis*. In this technique the charged DNA molecules are forced across a span of gel, driven by an electrical current. A molecule's properties determine how rapidly an electric field can move the molecule through a gelatinous medium. The band pattern yielded by the DNA mix in gel electrophoresis identifies the positions insertions occurred at, since smaller fragments have travelled the furthest.

The insertion frequency averaged over a 100,000 base-pair sliding window is shown in figure 5.14. Gaps in the data (chromosomal regions where transposition events could not be detected due to technical reasons) are indicated by short vertical

[6]DNA synthesis always proceeds in one direction along a single strand DNA, these directions are opposite on the two strands.

lines along the x-axis. The regions where distribution of transposition events significantly deviate from a Poisson process (P-values $< 0.01$) are marked by horizontal green lines. "OriC" shows the origin of chromosomal replication, "dif" denotes the replication termination area.[7]

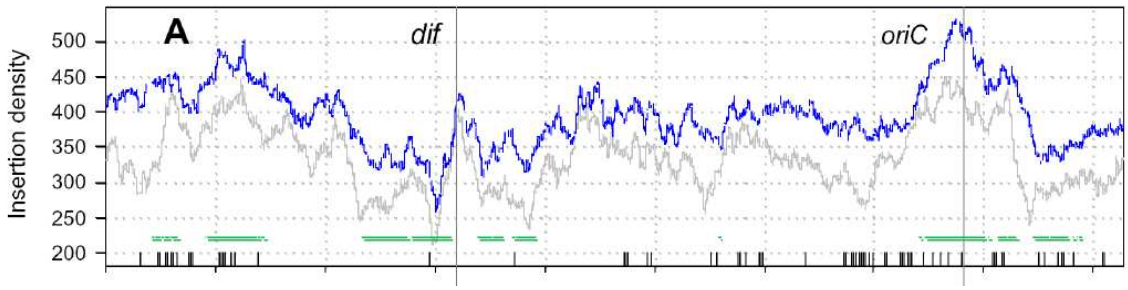



Figure 5.14. Distribution of transposon insertion density along the *E. coli* chromosome. Gray trace shows the transposon insertion density calculated as the number of transposition events per 100-kb sliding window over the entire *E. coli MG1655* chromosome. The blue trace was computed in a similar manner, except that all chromosomal regions corresponding to essential and ambiguous genes were excluded from the calculations in order to reconstruct insert distribution prior to selective outgrowth.

### 5.4.2 Evolutionary Preservation of Essential Genes

To assess the data set from an evolutionary perspective, we examined the distribution of conditionally essential and expendable *E. coli* genes with respect to the occurrence of orthologs[8] across a broad range of diverse bacterial genomes. Orthologs within a reference set of 32 complete bacterial genomes chosen to represent maximum phylogenic diversity were identified (see the Supplementary Material of [80]), and quantified by a simple parameter: the *Evolutionary Retention Index* (ERI). ERI is computed for each *E. coli* gene as the fraction of genomes from the

[7] *E. coli* has one circular chromosome and its replication starts at the *oriC* site, which moves in both directions towards the *dif* termination point.

[8] Two proteins from two different species that are thought to have a common evolutionary origin and thus are similar in amino acid sequence as well as in function are called *orthologs*.

reference set containing an ortholog of the gene, varying from 0 (for genes unique to *E. coli*) to 1.0 (for omnipresent genes).

The tendency of essential genes to be preserved by evolution is reflected in figure 5.15, showing the fraction of essential genes at different ERI values.[9] The relationship between the two parameters has the form $y = y_0 + a \cdot b^x$, implying that the essentiality fraction of genes with a given ERI is partly due to a very strong tendency of essential genes to be retained by evolution (the exponential behavior is dominant above ERI =0.6) and partly to a fraction of essential genes ($\sim 10\%$) that is present among genes within any ERI value group.

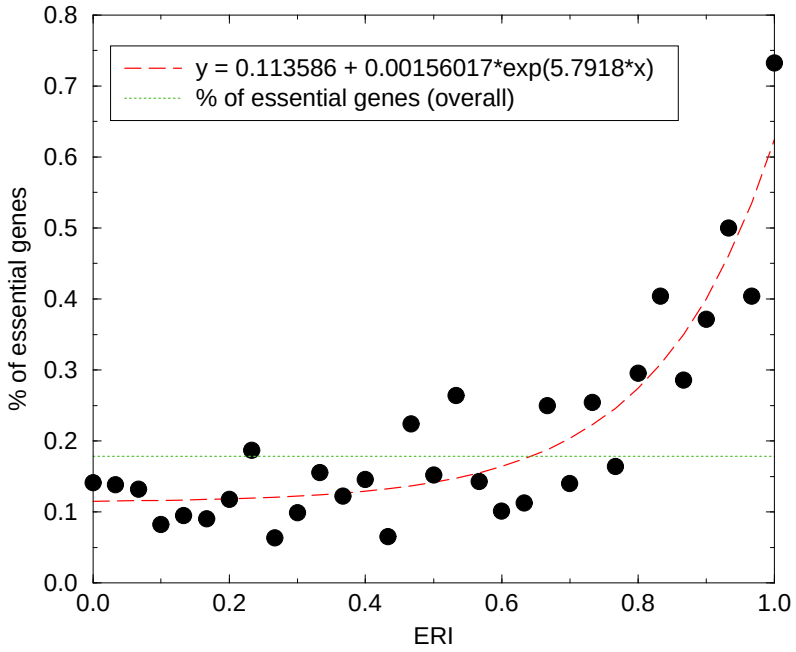



Figure 5.15. Fraction of essential genes at different ERI values. The data were fitted with $y = 12.0 + 0.023 \cdot (0.019)^x$ (dashed red line). The dotted line represents the fraction of essential genes for the whole genome.

[9]Unknown or ambiguous genes are not considered in the calculation of the fraction.

### 5.4.3 Essentiality of the Topological Modules

We have previously shown that hierarchical modularity in *E. coli* closely overlaps with known metabolic functions [174]. To comprehend the results of individual gene essentiality in the context of system level functional organization, we projected the essentiality of metabolic enzymes onto the global topological representation of the *E. coli* metabolic network.

As shown in figure 5.16 (top panel), the overall essentiality ratio of metabolic enzymes within the full metabolic network is relatively low (indicated by the green background of the whole tree), with essential enzymes limited to a subset of modules. Visual inspection of the figure indicates that while many metabolic modules are almost entirely nonessential, at the lowest hierarchical level several branches corresponding to small topological modules appear to be essential, i.e., they are composed of biochemical reactions catalyzed by predominantly essential enzymes. Of these, the largest fractions are within the nucleotide, coenzyme, and lipid metabolism. The pyrimidine metabolic module appears to contain the highest level of essential reactions.

A significant correlation between essentiality (tree on the top on Fig. 5.16) and ERI values (tree underneath) is apparent within metabolic modules, and many of the highly essential modules also contain metabolic enzymes with the highest ERI values.

### 5.4.4 Conclusions

The genetic footprinting technique used to assess gene essentiality in *E. coli* across the entire genome generated an internally coherent data set, which was examined at increasingly abstract levels to refine models of cellular organization. At the finest level, individual gene essentiality reveals basic physiologic information

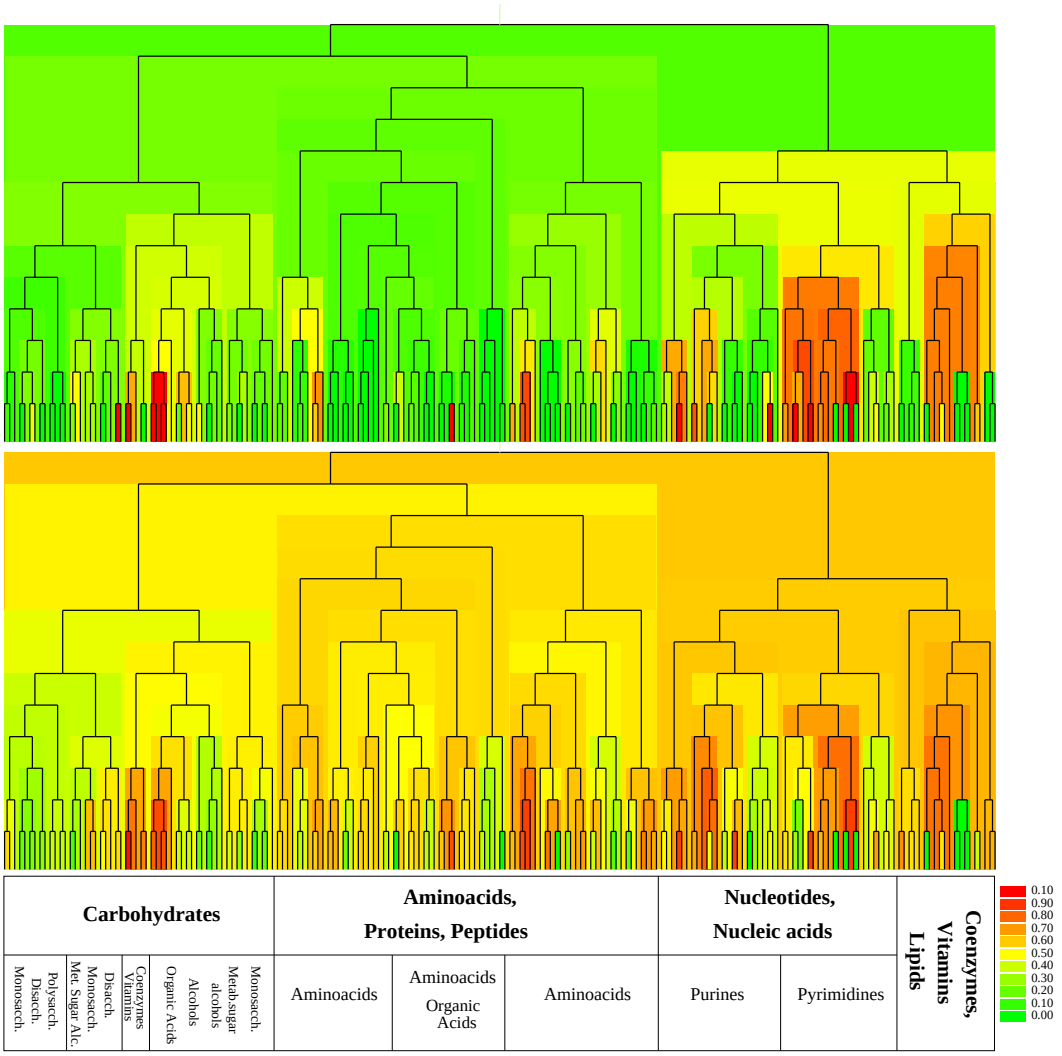



Figure 5.16. The evolutionary retention and essentiality ratio of enzymes in the topological modules of *E. coli* metabolism. The branches of the tree are color coded according to the fraction of essential enzymes (top panel) and the average ERI score of enzymes (bottom panel) catalyzing the biochemical reactions within a given topological module. Red indicates a 100% essentiality/conservation ratio within a module.

about cellular metabolism under specific growth conditions. At a more abstract level, the data can be used for focused comparative genomic analysis to define the core bacterial genetic repertoire, while at the highest level of abstraction, the data can be used to detect organizational principles of cellular networks.

# CHAPTER 6

# MODELLING THE *E. COLI* METABOLIC NETWORK

## 6.1 Motivation

The iterative duplication and integration of clustered nodes in the hierarchical model seamlessly combines scale free topology with an inherent modular structure [174]. However, the growth and evolution of this model, in a manner similar to many scale free models, is based on predefined global organization rules. Our understanding of biological systems suggests that local rules govern the growth dynamics of the underlying networks. The emergence of preferential attachment is expected to be a consequence of these local events, unlike in some human-made systems, where visibility of a node (correlated to its connectivity) often explains its increased ability to gain new links.[1] Indeed, modelling proteome evolution as a series of gene duplications leads to the experimentally observed scale free topology of protein-protein interaction networks [205, 198, 192, 163]. In duplication/diversification models for the evolution of the proteome, a node which represents a protein is randomly chosen and duplicated in each time step. While all interactions (edges) of the original protein are initially retained, subsequent mutations of genes might lead to the loss and gain of interactions (See Fig. 6.1). These models offer an elegant explanation for preferential attachment in protein networks: a well connected protein is more

[1]For example, a new webpage is quite likely to connect to a few well known hub pages everybody looks at: Google, CNN or Yahoo. Well cited papers are examples from citation networks: they have large impact on a research area and the more known (cited) they get, the more the community cites them.

likely to have a duplicating neighbor than less connected ones.

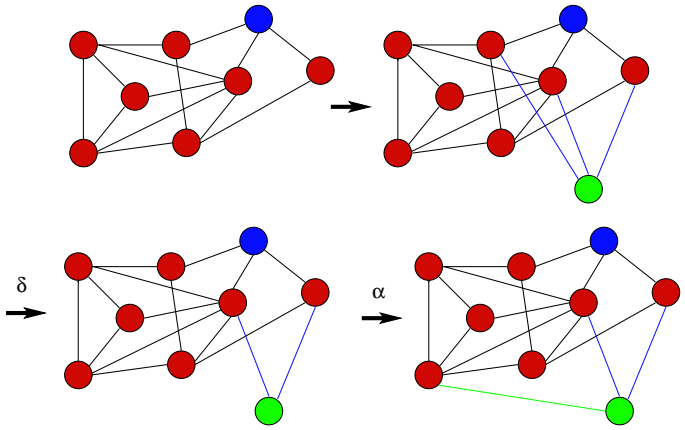



Figure 6.1. A duplication/diversification model for the evolution of protein interaction networks. A randomly chosen node (blue) is duplicated with all its interactions (green). Then these interactions may be lost with probability $\delta$, new ones can be gained with probability $\alpha$ (green interaction).

Metabolic reactions and the growth of metabolic networks are governed by the strict rules of chemistry. There is no good intuitive reason for a substrate to become a hub, nor for a well connected molecule (participating in many reactions) to become part of new reactions more easily than other molecules (preferential attachment). Furthermore, modelling of the evolution of metabolic networks encounters a few difficulties. First, while protein-protein interactions unambiguously define a network, chemistry suggests many ways to determine network structures in metabolic reactions. Second, since metabolic reactions are catalyzed by enzymes, we have to account for their impact on the evolution of a metabolic network. Here we propose a simple model with *local* growth and rewiring rules based on enzyme evolution, able to capture both the scale free and hierarchical nature of metabolic networks [129].

## 6.2 Definition of the Metabolic Network

We propose a network abstraction of metabolic reactions which is based on the structural similarity of molecules. In metabolic reactions an enzyme catalyzes the

transfer of molecular substructures from educts to products. Enzymatically catalyzed reactions are often relatively low in energy, thus the structural changes are usually small in each individual reaction. This results in strong chemical similarity of products and educts. As an example we show the hexokinase reaction of glycolysis (see Fig. 6.2), catalyzed by the *hexokinase* enzyme. *Hexokinase* transfers a phosphate group from adenosin triphosphate (ATP) to glucose (Glc), resulting in adenosin diphosphate (ADP) and glucose-6-phosphate (G6P). This catalytic activity is facilitated by an *active center* (a "cavity") where the educts adenosin-triphosphate (ATP) and glucose (Glc) fit perfectly. Due to the spatial arrangement, the transfer of a phosphate group from ATP to Glc is catalyzed.

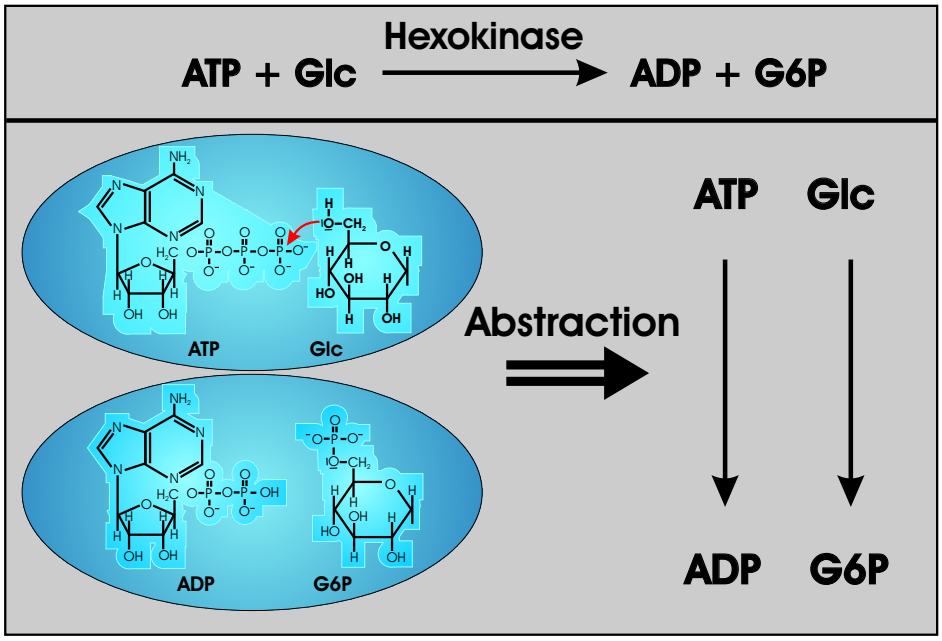



Figure 6.2. Network representation of a typical metabolic reaction, the hexokinase reaction of the glykolysis.

Glc and G6P resemble each other by sharing a large part of their structure, thus in our network representation these substrates are connected. Analogously, ATP and ADP are connected since ADP is the remainder of ATP after the enzymatically

catalyzed phosphate-group transfer. In this abstraction, based on structural similarity, edges reflect the activity of enzymes which facilitate the transfer of significant parts of the chemical substructures between connected substances. Although the majority of biochemical reactions are reversible, we only consider the most probable course of each reaction, an approximation which allows us to superimpose a direction on each edge. Finally, we obtain a directed graph reflecting enzymatically catalyzed exchanges of substructures between metabolites. Using all known metabolic reactions of *E.coli* retrieved from the WIT database [160] we set up a network consisting of 663 substrates embedded in a web by $1,010$ links (Fig. 6.3).[2]

Superimposing chemical similarity on the connection between substrates preservers the major characteristics of metabolic networks, scale free topology and hierarchical clustering (see Fig. 6.6).

## 6.3 Modelling Metabolic Network Evolution

Biochemical reactions are almost exclusively catalyzed by enzymes, which feature active sites in their spatial structure where substrates fit in. The quality of this fit is a decisive factor for the specificity and activity of the enzymes. Thus links in our network represent enzymatic activity which might be lost if the corresponding protein suffers a mutation. However, a mutation in the active center might destroy the old and render a new activity which does not entirely change the enzyme's specificity (see Fig. 6.4 for a schematic example).

### 6.3.1 Experimental Basis of the Proposed Model

Early theories of the metabolism's evolution gave rise to diverging perspectives on the emergence of enzymatic functions. One classical evolution theory [158] cred-

[2]We note that this network abstraction is very similar, although not always equivalent to the biochemical reduction in §5.3.1.

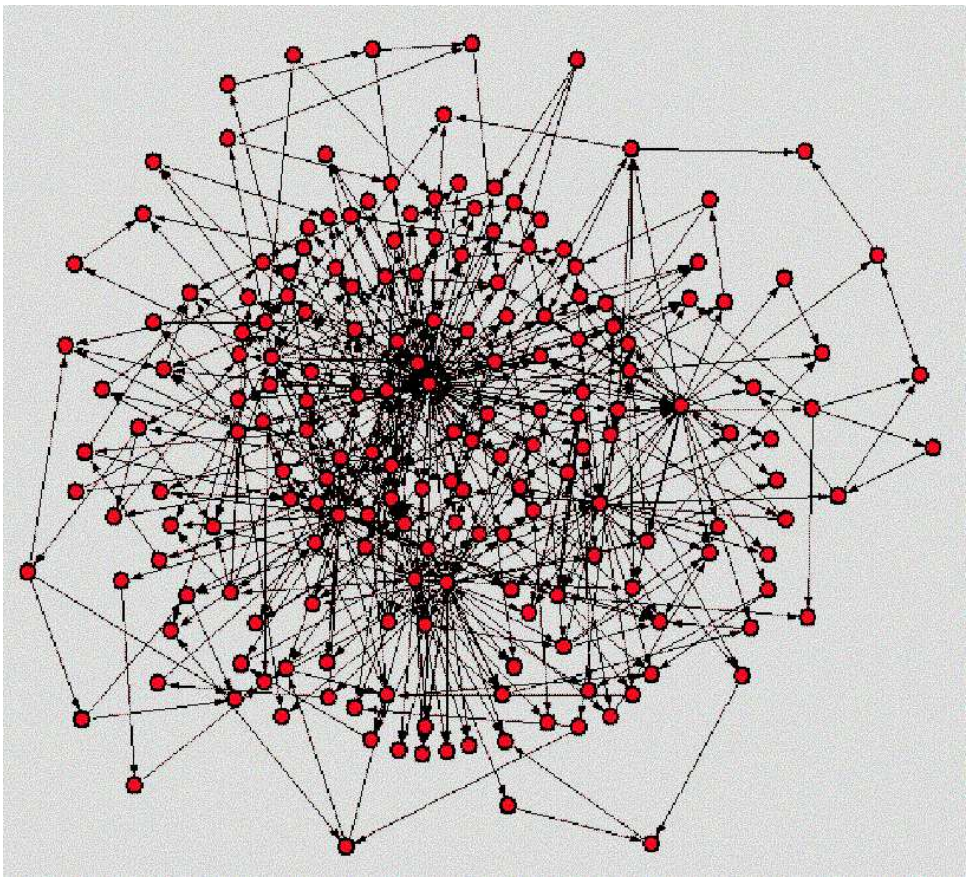


Figure 6.3. *E. coli* metabolic network based on similarity of reaction educts and products.

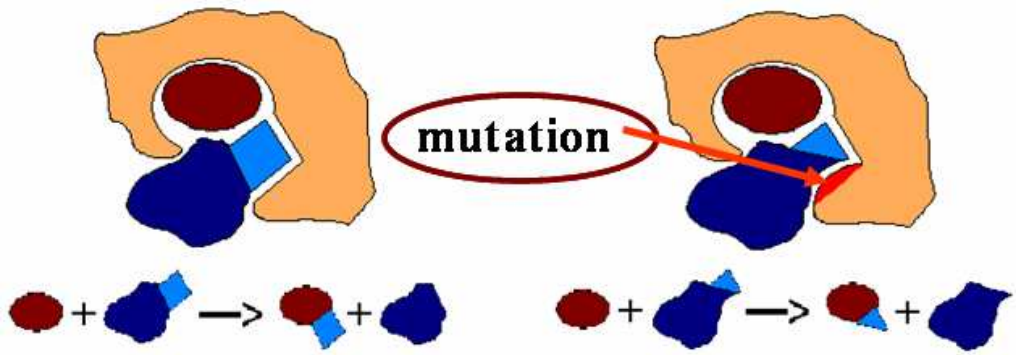



Figure 6.4. Schematic example of a mutation in the active site of an enzyme. The new enzyme can catalyze a reaction between substrates physically similar to the ones in the original reaction.

ited gene duplication for two copies of a particular protein, which can mutually backup their functions if mutations disable one of the proteins. To a certain extent, these relaxed selective constraints support the protein's capacity to develop new and dispose old functions. A different hypothesis [105] assumes that very early enzymes covered a broad band of specificity superposed by low activity. A series of gene duplications partitioned the initial set of functions, assigning each of the enzyme copies more specific tasks. However, not every mutation impacts a molecule's propensity to be evolutionary selected, a theory which is called the neutral theory of evolution [111, 112] or Non-Darwinian Evolution [113]. Although these theories provide different perspectives on the evolution of the enzyme's functionality, they share the common assumption that a mutation in the active site does not utterly change the structure of the enzyme's active site. The mutated and original enzyme should accept similar structures as educts, and lead to substrates similar to the original products.

A rough preservation of specificity for the educt-structure still allowing for functional diversity appears in a few different ways:

1. The enzyme might accept a slightly different educt, but otherwise lead to the same product. An example can be found in *Pseudomonas* strains: two closely related enzymes, *atzA* and *triA*, catalyze the hydrolysation of two triazines which differ only in one of three side chains. The hydrolysis generates the same product [170]. The enzymes differ in only nine amino acids: this small sequential change triggers the enzyme's susceptibility to an enlarged set of potential educts that structurally resemble each other on a large scale.

2. Analogously, the opposite case applies as well, since a mutation in an enzyme's active site might leave the specificity for the main educt, while generating a new product. The emergence of a new product out of the same educt is well

reflected by some experiments in artificial evolution. *Cytochrome P450 BM-3* normally catalyzes the oxidation of alkanes, however, with a directed mutation the enzyme showed alkane hydroxylation activity [70]

3. In a third scenario, a mutation affecting the enzyme's active site might have shifted the specificity for both the educts and products. For example the active site protein sequences of *L-galactosidase* and *L-glucuronidase*, which evolved from a common ancestor, resemble each other to only 25%. However, the educts and products involved in these reactions differ only to a slight extent in two distinct molecular positions. Furthermore, by changing four amino acids of the wild-type *L-glucuronidase*, its strong preference for L-glucuronids shifts to a clear affinity for L-galactosides [136].

These experiments show that a small sequential change can give rise to enzymes which are susceptible to chemically similar educts, while they feature products that resemble the initial product's structure.

### 6.3.2 Definition of the Model

Based on the presented evolutionary experiments, our model network grows as follows (see Fig. 6.5 for illustration):

- Start with a small seed network, connected randomly via directed links ($N_0 = 10$, $L_0 = 15$).
- At each time-step:
  - Randomly choose an arc, let $E$ (educt) be the starting, $P$ (product) the ending substrate of the link.[3]

[3]If $E$ and $P$ are linked in both directions, the two arcs are accounted for separately, so the random sampling picks them independently.

– Randomly choose between three cases:

1. $\underline{E' \to P \text{ case:}}$ With probability $1 - P_{\text{new}}$, choose a node $E'$ based on *similarity* to $E$ (do nothing otherwise): Pick $E'$ from the first and second neighbors of $E$, using probability $p$ for all the first neighbors and $p^2$ for all its second neighbors (neighbors are counted regardless of the directions of the links). The value of $p$ is obtained from the normalization: $N_{\text{first neighbors}} \cdot p + N_{\text{second neighbors}} \cdot p^2 = 1$. After $E'$ is chosen (and if there is no arc from $E'$ to $P$ and $E' \neq P$), put an arc from $E'$ to $P$.
2. $\underline{E \to P' \text{ case:}}$ Introduce a new node $P'$ into the graph with probability $P_{\text{new}}$, otherwise (probability $1 - P_{\text{new}}$) choose an existing node $P'$ based on its *similarity* to $P$. The similarity choice works the same way as in the previous case. After $P'$ is chosen (and if there is no arc from $E$ to $P'$ and $S \neq P'$) put an arc from $E$ to $P'$.
3. $\underline{E' \to P' \text{ case:}}$ Choose $E'$ from the existing nodes of the graph based on the similarity to $E$. Then, with probability $P_{\text{new}}$, introduce a new node $P'$ into the graph, or choose an existing node $P'$ with probability $1 - P_{\text{new}}$, based on its *similarity* to $P$. After $E'$ and $P'$ are chosen (and if there is no arc from $E'$ to $P'$ and $E' \neq P'$) put an arc from $E'$ to $P'$.

– If there was a new link created in this time-step, then remove the $S \to P$ arc with probability $1 - P_{\text{dup}}$.

- Due to the presence of $P_{\text{dup}}$, links are being removed and parts of the giant cluster can fall off, thus the emerging graph can have many disconnected clusters. We are interested in the giant cluster only (if it forms), so at each

time-step we disregard the nodes that are not part of the largest cluster. (This is computationally very important for low $P_{\text{dup}}$ values, where the largest cluster can be a small fraction of all the nodes once introduced into the graph.)

A new educt entering the cell is a rather rare event in comparison to the frequency that a mutated enzyme accepts a new, but already present educt. Therefore, we neglected this explicit albeit infrequent case without jeopardizing the overall validity of our approach. Since we rest on the assumption that a mutation of the active site renders an enzyme able to adopt roughly similar chemical structures, the choice of these substrates is limited to the immediate neighborhood of the affected substrate. Our abstraction of metabolic reactions implies a direct correspondence between the structural difference of two substrates and their distance in the network. As an approximation, we focus on the first two layers around a node, using the above given definition of "similarity." In figure 6.5 we give a schematic representation of the algorithm.

### 6.3.3 Properties of the Model

For almost all $\{P_{\text{dup}}, P_{\text{new}}\}$ parameter pairs for which a giant cluster emerges,[4] our model leads to scale free, hierarchical networks. Exceptions are low $P_{\text{new}}$ values where the network does not grow sufficiently fast to form large enough hubs to avoid an exponential cutoff in its degree distribution. We have searched the parameter space for pairs that best mimic the behavior of the metabolic map. The chosen parameter pair $P_{\text{dup}} = 0.85$ and $P_{\text{new}} = 0.6$ leads to good agreement of the average quantities as well as the scaling properties of the model and the *E. coli* metabolic network (see Table 6.1 and Fig. 6.6). An exception is the average clustering coeffi-

[4]Very high rates of new substrate incoming combined with almost no duplication naturally leads to a fragmented network in which the number of small fragments grows instead of one giant component.

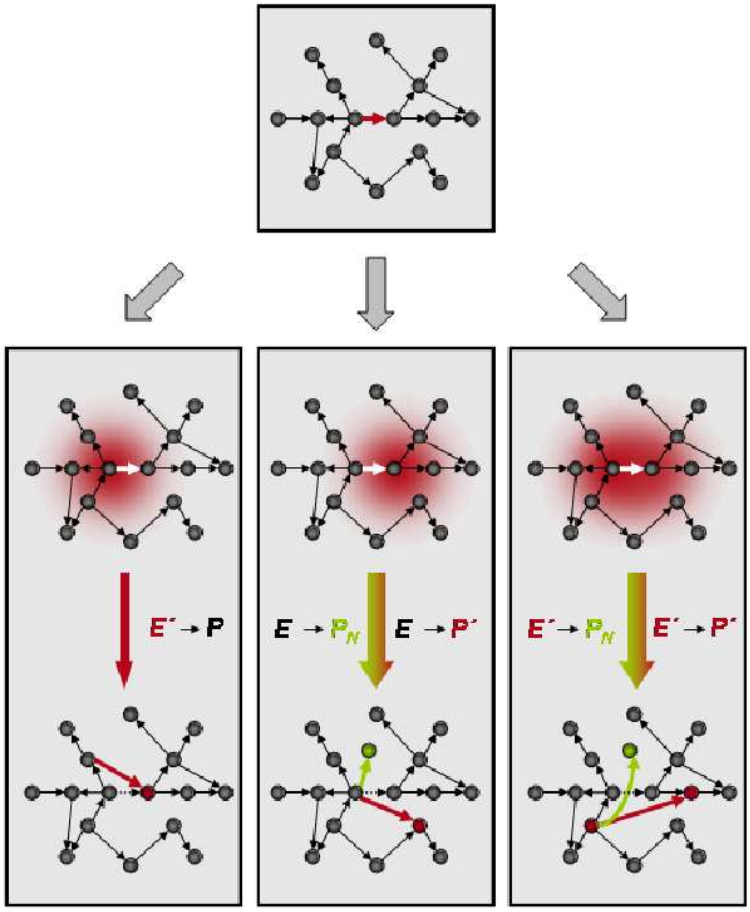



Figure 6.5. Schematic illustration of the metabolic network model. The three panels correspond to the $E' \to P$, $E \to P'$ and $E' \to P$ cases.

cient, which is quite low in our similarity-based metabolic network. This is largely due to substrates on "arcs" with $k = 2$ and $C = 0$. Arcs are very frequent in the metabolic network, since larger chemical changes of a substrate into another one are often facilitated by a series of enzymes forming a well-defined, non-branching pathway. This is a feature of the metabolic network that out model was not meant to capture. However, the scaling properties of the clustering coefficient do not change much under the topological reduction of the metabolic network (see §5.3.1), in which these arcs are all shortened to a link. The average clustering coefficient of this network is 0.27 (in parentheses in Table 6.1), and its scaling properties match the ones

of the model network better as well (Fig. 6.6).

TABLE 6.1. PROPERTIES OF THE METABOLIC AND MODEL NETWORKS

| | | *E. coli* | model |
|---|---|---|---|
| connectivity | $\langle k \rangle$ | 3.18 | 3.10 |
| clustering coefficient | $\langle C \rangle$ | 0.04 (0.27) | 0.26 |
| shortest path length | $\langle L \rangle$ | 4.25 | 4.31 |
| assortativity | $r$ | -0.12 | -0.08 |

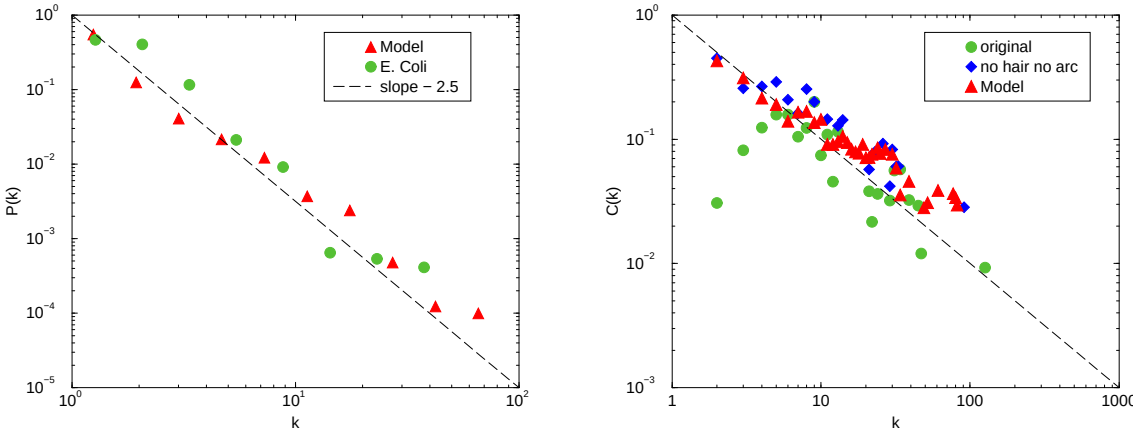



Figure 6.6. Log-binned distributions of the node's degree $P(k)$ and clustering coefficient $C(k)$ of the best fitting model and the metabolic network of *E. coli*.

## 6.4 Conclusions

Unlike other contemporary models which credit global organization rules for the incorporation of newly added nodes our model indicates that molecular evolution governed by local, decentralized mechanisms can also lead to the metabolic network topology observed in the cell. Introduction of chemical similarity, capturing structural resemblances which arise from the exchange of molecular substructures, enables us to define a new network abstraction of metabolic reactions. Moreover, this perspective allows us to model evolutionary events which act locally, in the

neighborhood of substrates affected by an altered enzyme. We also showed that rules governing enzyme network evolution can lead to preferential attachment not only in protein interaction networks, but also in the metabolism. Substrates that take part in many chemical reactions, the hubs, are more likely to be picked as new educts or products of a mutated enzyme: There are more enzymes with active centers specific to a substrate similar to them, enzymes which can then mutate to catalyze a new reaction, which the hub will be part of.

# CHAPTER 7

# OUTLOOK

The hierarchical architecture offers a new perspective on the topology of complex networks. Our study offers only a starting point for understanding the interplay between the scale free, hierarchical and modular nature of real systems. While the $C(k)$ curves offer a tool to unearth the presence of a hierarchy, the necessary ingredients of a theory to model the emergence of hierarchy is still open to question. Finally, the role of the geometrical factor, which appears to remove the hierarchy, needs to be elucidated. Further modelling and empirical studies should allow us to address these questions.

Progress has already been made both towards modelling the possible origin of the hierarchy [152, 195, 18] and towards the understanding of dynamic systems interacting along the links of an underlying hierarchical topology. These dynamical studies focus mostly on diffusion-related phenomena, like random walk [157, 156] or epidemic spreading [152]. An interesting finding of the latter study is that strong clustering in a network can lower the epidemic threshold,[1] and thus diseases with even smaller spreading rates can reach a finite size of the network. On the other hand, high clustering also limits the size of outbreaks. The fraction of the network a given disease can infect is smaller in hierarchical networks.

[1]The *epidemic threshold* of a disease is the spreading rate (the mean probability that an infected individual will transmit the disease to a susceptible network neighbor in unit time) above which the disease reaches a finite fraction of the population instead of dying out.

The fact that the hierarchical nature of networks is captured by a simple quantity, the $C(k)$ curve, offers us a relatively straightforward method to identify the presence of hierarchy in real networks. The law (4.8) indicates that the number and the size of the groups of different cohesiveness is not random, but follows rather strict scaling laws.

The method proposed in the thesis for identifying hierarchy became part of the standard tool kit one uses to analyze the topology of networks under consideration. Many studies have found that networks, either new or already inspected by the field, show hierarchical $C(k)$ scaling.

An interesting example is *software systems* analyzed by Myers [144], who represented six large computer software packages by directed networks of two kinds. Three of these systems were written in `C++`, object oriented software that lent itself to a *class-collaboration graph* representation. Classes describe the form of objects in these systems (the nodes of the graph), while collaboration is the process by which more complex, multi-functional classes are built from simpler ones (defining the links of the graph). Another group of three procedural systems written in `C` were represented as *static-call graphs*. Nodes of this network are procedures and possible calls between procedures represent directed links. The study found that all six graphs are scale free, with degree exponents of 2.5 for the call graphs, $\gamma_{\text{in}} \simeq 2$ and $\gamma_{\text{out}} \simeq 3$ for class-collaboration graphs. All six networks show power law $C(k)$ scaling, indicating hierarchical software organization.

A very different group of hierarchical networks of great interest for the community are social systems. The World Trade network mentioned in §1.3.1 shows a clear power law tail in its clustering function, with an exponent of 0.7 [187]. A social acquaintance network based on the `www.wiw.hu` website shows that this social sys-

tem also shows strong hierarchy with an exponent $\beta = 0.33$ of the $C(k)$ clustering function [47]. The e-mail network of the University at Rovia i Virgili in Tarragora, Spain, is also hierarchically structured, as demonstrated in a study by Guimerà *et al.* [88].

A large research group in Los Alamos has built a software system capable of tracing $10^6$ agents in a simulation of different aspects of human society in Portland, Oregon [42]. The simulation uses a variety of statistical data available about the city to generate a virtual population "living" its daily life in Portland. One subsystem of this simulation is TRANSIMS, which captures how individuals move around the city on a reality-based transportation infrastructure. The network of interest to us, generated by TRANSIMS, is a graph of over $18,000$ locations on the city, connected by directed links representing people moving from one location to another during one simulation day. The program can also generate subgraphs specific to different activities such as work, recreation or school. The study finds that the out-degree distribution of the location graph is a power law, and the clustering of locations with different out degrees falls as a power law. What is more interesting, however, is that subgraphs of the location network, such as graphs defined by links pointing to work-related, recreational or school locations only, show a significantly different topology. While the recreational and school-related graphs are hierarchical, links pointing to work locations define a clustered, but non-hierarchical graph similar to other spatially constrained networks like the router-level Internet or the power grid.

A study by Barrat *et al.* [25] focused on weighted networks such as the world-wide airport network [3] and the scientific collaboration network of condensed-matter physicists based on the Los Alamos Archive *cond-mat* [149, 147, 148]. They define the *weighted clustering coefficient* of a node, a generalization of the quantity which takes the weighted nature of links into account. They found both networks to be

hierarchical, based on both the traditional and generalized measure of clustering. Interestingly, weighted clustering coefficients of airports were generally larger than the standard measure, especially for larger nodes, indicating that although airport hubs have a wide variety of connections, their high-traffic links form a more interconnected set (the *rich-club phenomenon*). The two clustering measures of the collaboration graph are almost identical up to about 20 collaborators, indicating the presence of stable research groups with a well-defined average intensity of collaboration. Hubs of the scientific world, however, show similar signs of forming clubs as airports: collaborators of a hub are much more frequent among a more connected subset of researchers who, taking into account the networks's assortative nature, tend to be hubs as well.

## 7.2 Hierarchical Modularity as a Paradigm for Biological Organization

Modularity is not an exclusive property of the metabolism. Indeed, the protein interaction network of *S. cerevisae* [223], based on four independent databases [220, 219, 138, 139, 196, 96], the conformational spaces of RNA [218] and the genetic regulatory network also reflect a modular architecture.

The appearance of hierarchical modularity in biological networks supports the assumption that evolution acts on many levels. The accumulation of local changes, affecting the small highly integrated modules, slowly impacts the larger, less integrated modules as well. Thus, evolution might act in self-similar fashion, copying and reusing existing modules to further increase the organism's complexity. Especially in the face of eukaryotic evolution, this network based framework might be suitable to describe the explosion of complexity in the development of the single-celled *S. cerevisiae* toward the multicellular *H. sapiens*. Cellular functions like information storage, processing and execution are carried out by the genome, tran-

scriptome, proteome and metabolome. All these cellular functions can be described by networks of various heterogeneous components. One way to visualize the complex relationships between these components is to organize them into a simple complexity pyramid shown in figure 7.1 [159], in which various molecular components - genes, RNAs, proteins and metabolites - organize themselves into recurrent patterns such as metabolic pathways and genetic regulatory motifs. In turn, motifs and pathways are seamlessly integrated to form functional modules which are responsible for distinct cellular functions [89]. These modules are nested in a hierarchical fashion and define the cell's large-scale organization.

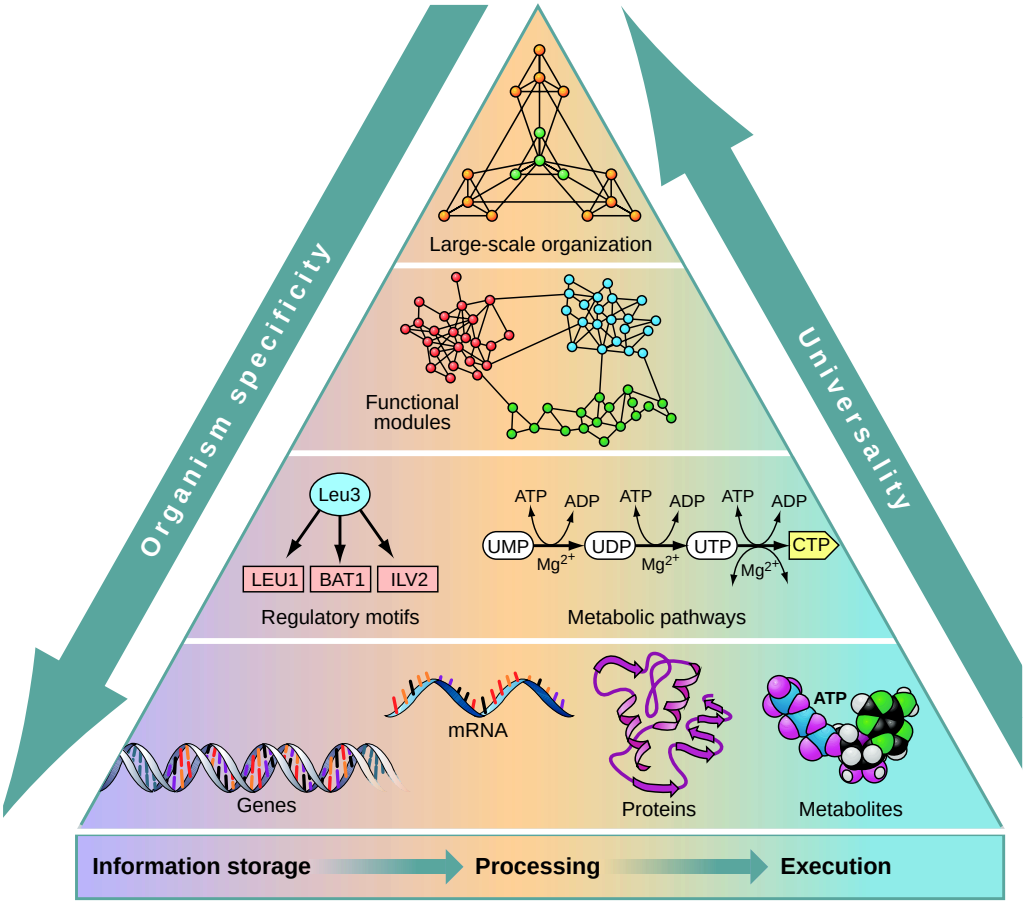



Figure 7.1. Life's complexity pyramid. (From [159])

Our present knowledge about the architecture of biological networks emphasizes two major aspects:

1. Discrete cellular functions are mediated with the aid of distinct albeit often blurred modules;

2. Network integrity is assured by a handful of highly connected nodes, making networks robust against random failures but exceedingly vulnerable to targeted attacks.

These features explain the observation that many mutations have little or no phenotypic effect [204], which appears to be consistent with the presence of genes that either cannot propagate their failure or whose function can be replaced by other components of the network. The presence of genes that integrate multiple signals and trigger widespread changes upon their failure proves the crucial role of highly connected genes. For example, the tumor-suppressor gene *p53* has been identified as a highly connected and thus crucial node which, once mutated, severely jeopardizes genome stability and integration of signals related to the control of cell-cycle and cell death [203, 119]. Emphasizing its crucial role, dysfunctional p53 proteins are involved in more than half of all human cancer phenotypes. With the increasing ability to identify and collect protein-protein interactions the determination of modules and highly connected proteins will become a major issue in the fast and effective identification of potential drug targets. The recent progress in biological networks has successively uncovered the skeleton and organization of networks, offering important insights about the assembly and functionality of components and subnetworks. In future, we will need to go several steps further addressing the dynamic aspects of various cellular networks.

## 7.3 Conclusions

The fact that many large networks are scale free is now well established. It is also clear that most networks have a modular topology, quantified by the high clustering coefficient they display. Such modules have been proposed to be a fundamental feature of biological systems [89, 174], but have been discussed in the context of the WWW [127, 76] and social networks as well [86, 210]. Hierarchical topology offers a new avenue for bringing these two concepts under a single roof, giving a precise and quantitative meaning for the network's modularity. It indicates that we should not think of modularity as the coexistence of relatively independent groups of nodes. Instead, we have many small clusters, that are densely interconnected. These combine to form larger, but less cohesive groups, which combine again to form even larger and even less interconnected clusters. This self-similar nesting of different groups or modules into each other forces a strict fine structure on real networks.

The presence of such a hierarchical architecture reinterprets the role of the hubs in complex networks. Hubs, the highly connected nodes at the tail of the power law degree distribution, are known to play a key role in keeping complex networks together, playing a crucial role from the robustness of the network [14, 44] to the spread of viruses in scale free networks [165]. Measurements on many real networks indicate that the clustering coefficient characterizing each node decreases with the degree. This implies that while the small nodes are part of highly cohesive, densely interlinked clusters, the hubs are not, as their neighbors have a small chance of linking to each other. Therefore, the hubs play the important role of bridging the many small communities of clusters into a single, integrated network.

While it is difficult to identify universal characteristics from single examples, once they are uncovered, they offer strong support for an emerging theme: networks

in nature are far from random, but evolve following robust self-organizing principles and evolutionary laws that cross disciplinary boundaries. The results reviewed here represent only one chapter in their story; systematic data driven studies focusing on the topology and evolution of real networks could fundamentally change how we approach the complex world around us.

# BIBLIOGRAPHY


[1] http://www.oakland.edu/∼grossman/erdoshp.html.

[2] http://moat.nlanr.net/infrastructure.html.

[3] International Air Transportation Association Database, http://www.iata.org.

[4] L. A. Adamic, The small world web. In *Lecture Notes in Computer Science*, volume 1696, pages 443–454, Springer, New York, NY (1999).

[5] L. A. Adamic and E. Adar, Friends and neighbors on the web. *Preprint*, http://hpl.hp.com/shl/papers/web10/index.html.

[6] L. A. Adamic and B. A. Huberman, Power-law distribution of the World Wide Web. *Science*, 287: 2115 (2000).

[7] W. Aiello, F. Chung and L. Lu, A random graph model for massive graphs. In *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing*, pages 171–180, ACM, New York (2000).

[8] W. Aiello, F. Chung and L. Lu, Random evolution of massive graphs. In J. Abello, P. M. Pardalos and M. G. C. Resende, editors, *Handbook of Massive Data Sets*, pages 97–122, Kluwer, Dordrecht (2002).

[9] B. J. Akerley, E. J. Rubin, V. L. Novick, K. Amaya, N. Judson and J. J. Mekalanos, A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc. Natl. Acad. Sci. U.S.A.*, 99: 966–971 (2002).

[10] R. Albert, I. Albert and G. L. Nakarado, Structural vulnerability of the North American power grid. *Phys. Rev. E*, 69: 025103 (2004).

[11] R. Albert and A.-L. Barabási, Topology of evolving networks: local events and universality. *Phys. Rev. Lett.*, 85: 5234 (2000).

[12] R. Albert and A.-L. Barabási, Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74: 67–97 (2002).

[13] R. Albert, H. Jeong and A.-L. Barabási, Diameter of the World-Wide Web. *Nature*, 401: 130–131 (1999).

[14] R. Albert, H. Jeong and A.-L. Barabási, Attack and error tolerance of complex networks. *Nature*, 406: 378 (2000).

[15] U. Alon, M. G. Surette, N. Barkai and S. Leinler, Robustness in bacterial chemotaxis. *Nature*, 397: 168–171 (1999).

[16] L. A. N. Amaral, A. Scala, M. Barthélémy and H. E. Stanley, Classes of small-world networks. *Proc. Nact. Acad. Sci.*, 97: 11149 (2000).

[17] G. Apic, J. Gough and S. Teichmann, Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.*, 310: 311–325 (2001).

[18] M. Baiesi and S. S. Manna, Scale-free networks from a hamiltonian dynamics. *Phys. Rev. E*, 68: 047103 (2003).

[19] A.-L. Barabási, *Linked: The New Science of Networks*. Perseus Publishing, Cambridge, MA (2002).

[20] A.-L. Barabási and R. Albert, Emergence of scaling in random networks. *Science*, 286: 509–512 (1999).

[21] A.-L. Barabási, R. Albert and H. Jeong, Mean-field theory for scale-free random networks. *Physica A*, 272: 173–187 (1999).

[22] A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert and T. Vicsek, Evolution of the social network of scientific collaborations. *Physica A*, 311: 590 (2002).

[23] A.-L. Barabási, E. Ravasz and Z. N. Oltvai, Hierarchical organization of modularity in complex networks. In R. Pastor-Satorras, J. M. Rubi and A. Diaz-Guilera, editors, *Proc. of the XVIII Sitges Conference on Statistical Mechanics, Sitges, Barcelona, Spain, 2002*, volume 625, page 46, Springer, Berlin (2003).

[24] A.-L. Barabási, E. Ravasz and T. Vicsek, Deterministic scale-free networks. *Physica A*, 299: 559–564 (2001).

[25] A. Barrat, M. Barthélemy, R. Pastor-Satorras and A. Vespignani, The architecture of complex weighted networks. *Proc. Nat. Acad. Sci.*, 101: 3747–3752 (2004).

[26] P. S. Bearman, J. Moody and K. Stovel, Chains of affection: The structure of adolescent romantic and sexual networks. *Preprint,* Department of Sociology, Columbia University (2002).

[27] G. Bianconi and A.-L. Barabási, Competition and multiscaling in evolving networks. *Europhys. Lett.*, 54: 436 (2001).

[28] G. Bianconi and A.-L. Barabási, Bose-Einstein condensation in complex networks. *Phys. Rev. Lett.*, 86: 5632 (2001).

[29] B. Bollobás, *Random Graphs*. Academic Press, London (1985).

[30] B. Bollobás and O. Riordan, Mathematical results on scale-free random graphs. In S. Bornholdt and H. G. Schuster, editors, *Handbook of Graphs and Networks*, Wiley-VCH, Berlin (2002).

[31] B. Bollobás, O. Riordan, J. Spencer and G. Tusnády, The degree sequence of a scale-free random process. *Random Structures and Algorithms*, 18: 279–290 (2001).

[32] B. Bollobás and O. M. Riordan, The diameter of a scale-free random graph. *Preprint* (2002), `http://www.dpmms.cam.ac.uk/~omr10`.

[33] M. Boss, H. Elsinger, M. Summer and S. Thurner, The network topology of the interbank market. *Los Alamos Archive,* cond-mat/0309582 (2003).

[34] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajalopagan, R. Stata, A. Tomkins and J. Wiener, Graph structure in the web. *Comput. Netw.*, 33: 309–320 (2000).

[35] A. Broida and K. C. Claffy, Internet topology: Connectivity of IP graphs. In S. Fahmy and K. Park, editors, *Scalability and Traffic Control in IP Networks,* in *Proc. SPIE*, volume 4526, pages 172–187, International Society for Optical Engineering, Bellingham, WA (2001).

[36] A. Bunde and S. Havlin, *Fractals and Disordered Systems*. Springer-Verlag, Berlin, Heidelberg, second edition (1996).

[37] C. Burge, Chipping away at the transcriptome. *Nature Genet.*, 27: 232–234 (2001).

[38] J. Camacho, R. Guimera and L. A. N. Amaral, Analytical solution of a model for complex food webs. *Phys. Rev. E*, 65: 030901 (2002).

[39] J. Camacho, R. Guimera and L. A. N. Amaral, Robust patterns in food web structure. *Phys. Rev. Lett*, 88: 228102 (2002).

[40] H. Caron, B. van Schaik, M. van der Mee, F. Baas, G. Riggins, P. van Sluis, M.-C. Hermus, R. van Asperen, K. Boon and P. A. Voute, The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. *Science*, 291: 1289–1292 (2001).

[41] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. J. Shenker and W. Willinger, The origin of power laws in internet topologies revisited. In *Proceedings of the 1st Annual Joint Conference of the IEEE Computer and Communications Societies*, IEEE Computer Society (2002).

[42] G. Chowell, J. M. Hyman and S. Eubank, Analysis of a real world network: The city of Portland. In *Technical Report BU-1604-M*, Department of Biological Statistics and Computational Biology, Cornell University (2002).

[43] F. Chung and L. Lu, The diameter of random sparse graphs. *Adv. Appl. Math.*, 26: 257–279 (2001).

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

[45] R. Cohen and S. Havlin, Scale-free networks are ultra small. *Phys. Rev. Lett.*, 90: 058701 (2003).

[46] M. C. Costanzo, M. E. Crawford, J. E. Hirschman, J. E. Kranz, P. Olsen, L. S. Robertson, M. S. Skrzypek, B. R. Braun, K. L. Hopkins, P. Kondu, C. Lengieza, J. E. Lew-Smith, M. Tillberg, and J. I. Garrels, YPD$^{TM}$, PombePD$^{TM}$ and WormPD$^{TM}$: model organism volumes of the BioKnowledge$^{TM}$ Library, an integrated resource for protein information. *Nucleic Acids Res.*, 29: 75–79 (2001).

[47] G. Csányi and B. Szendrői, Structure of a large social network. *Phys. Rev. E*, 69: 036131 (2004).

[48] G. F. Davis, M. Yoo and W. E. Baker, The small world of the american corporate elite, 1982-2001. *Strategic Organization*, 1: 301–326 (2003).

[49] M. A. de Menezes and A.-L. Barabási, Fluctuations in network dynamics. *Phys. Rev. Lett.*, 92: 028701 (2004).

[50] D. J. de Solla Price, Networks of scientific papers. *Science*, 149: 510–515 (1965).

[51] Z. Dezső and A.-L. Barabási, Halting viruses in scale-free networks. *Phys. Rev. E.*, 65: 055103 (2002).

[52] P. S. Dodds and D. H. Rothman, Geometry of river networks. *Phys. Rev. E*, 63: 01611–016117 (2001).

[53] S. N. Dorogovtsev, A. V. Goltsev and J. F. F. Mendes, Pseudofractal scale-free web. *Phys. Rev. E*, 65: 066122 (2002).

[54] S. N. Dorogovtsev and J. F. F. Mendes, Evolution of networks with aging of sites. *Phys. Rev. E*, 62: 1842–1845 (2000).

[55] S. N. Dorogovtsev and J. F. F. Mendes, Exactly solvable small-world network. *Europhys. Lett.*, 50: 1–7 (2000).

[56] S. N. Dorogovtsev and J. F. F. Mendes, Scaling behaviour of developing and decaying networks. *Europhys. Lett.*, 52: 33 (2000).

[57] S. N. Dorogovtsev and J. F. F. Mendes, Language as an evolving word web. *Proc. Royal Soc. London B*, 268: 2603–2606 (2001).

[58] S. N. Dorogovtsev and J. F. F. Mendes, Evolution of networks. *Adv. Phys.*, 51: 1079 (2002).

[59] S. N. Dorogovtsev, J. F. F. Mendes and A. N. Samukhin, Structure of growing networks: Exact solution of the Barabási-Albert model. *Phys. Rev. Lett.*, 85: 6633 (2000).

[60] H. Ebel, L. I. Mielsch and S. Bormholdt, Scale-free topology of e-mail networks. *Phys. Rev. E*, 66: 035103 (2002).

[61] J.-P. Eckmann and E. Moses, Curvature of co-links uncovers hidden thematic layers in the World Wide Web. *Proc. Nact. Acad. Sci.*, 99: 5825 (2002).

[62] V. M. Eguiluz, D. R. Chialvo, G. Cecchi, M. Baliki and A. V. Apkarian, Scale-free brain functional networks. *Los Alamos Archive,* cond-mat/0309092 (2003).

[63] M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein, Cluster analysis and display of genome-wide expression patterns. *Proc. Nact. Acad. Sci.*, 95: 14863–14868 (1998).

[64] P. Erdős and A. Rényi, On random graphs I. *Publ. Math. (Debrecen)*, 6: 290–297 (1959).

[65] P. Erdős and A. Rényi, On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5: 17–61 (1960).

[66] P. Erdős and A. Rényi, On the evolution of random graphs. *Bull. Inst. Internat. Statist., Tokyo*, 38: 343–347 (1961).

[67] P. Erdős and A. Rényi, On the strength of connectedness of a random graph. *Acta Math. Acad. Sci. Hungar.*, 1: 261–267 (1961).

[68] G. Ergün and G. J. Rodgers, Growing random networks with fitness. *Physica A*, 303: 261–272 (2002).

[69] M. Faloutsos, P. Faloutsos and C. Faloutsos, On power-law relationships of the Internet topology. *Comput. Commun. Rev.*, 29: 251–262 (1999).

[70] E. Farinas, T. Bulter and F. Arnold, Directed enzyme evolution. *Curr. Op. Biotech.*, 12: 545–551 (2001).

[71] D. Fell and A. Wagner, The small world of metabolism. *Nature Biotech.*, 189: 1121–1122 (2000).

[72] R. Ferrer i Cancho, C. Janssen and R. V. Solé, Topology of technology graphs: Small world patterns in electronic circuits. *Phys. Rev. E*, 64: 046119 (2001).

[73] R. Ferrer i Cancho and R. V. Solé, The small-world of human language. *Proc. Roy. Soc. London B*, 268: 2261–2266 (2001).

[74] R. Ferrer i Cancho and R. V. Solé, Least effort and the origins of scaling in human language. *Procs. Natl. Acad. Sci. USA*, 100: 788–791 (2003).

[75] M. Flajolet, G. Rotondo, L. Daviet, F. Bergametti, G. Inchauspe, P. Tiollais, C. Transy and P. Legrain, A genomic approach to the *Hepatitis C* virus. *Gene*, 242: 369–379 (2000).

[76] G. W. Flake, S. Lawrence and C. L. Giles, Efficient identification of web communities. In *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining*, pages 150–160, ACM, Boston (2000).

[77] G. W. Flake, S. R. Lawrence, C. L. Giles and F. M. Coetzee, Self-organization and identification of web communities. *IEEE Computer*, 35: 66–71 (2002).

[78] T. M. J. Fruchterman and E. M. Reingold, Graph drawing by force-directed placement. *Software - Practice and Experience*, 21: 1129–1164 (1991), A version of the Fruchterman–Reingold algorithm has been implemented into the free `Pajek` software, available at `http://vlado.fmf.uni-lj.si/pub/networks/pajek/`.

[79] A. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. Rick and A.-M. Michon, Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415: 141–147 (2002).

[80] S. Y. Gerdes, M. D. Scholle, J. W. Campbell, G. Balázsi, E. Ravasz, M. D. Daugherty, A. L. Somera, N. C. Kyrpides, I. Anderson, M. S. Gelfand, A. Bhattacharya, V. Kapatral, M. DSouza, M. V. Baev, Y. Grechkin, F. Mseeh, M. Y. Fonstein, R. Overbeek, A.-L. Barabási, Z. N. Oltvai and A. L. Osterman, Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.*, 185: 5673–5684 (2003).

[81] S. Y. Gerdes, M. D. Scholle, M. D'Souza, A. Bernal, M. V. Baev, M. Farrell, O. V. Kurnasov, M. D. Daugherty, F. Mseeh and B. M. Polanuyer, From genetic footprinting to antimicrobial drug targets: examples in cofactor biosynthetic pathways. *J. Bacteriol.*, 184: 4555–4572 (2002).

[82] G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson and B. Andre, Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, 418: 387–391 (2002).

[83] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin and E. Vitols, A protein interaction map of *Drosophila melanogaster*. *Science*, 302: 1727–1736 (2003).

[84] M. Girvan and M. E. J. Newman, Community structure in social and biological networks. *Proc. Natl. Acad. Sci.*, 99: 7821–7826 (2002).

[85] R. Govindan and H. Tangmunarunkit, Heuristics for Internet map discovery. In *Proceedings of IEEE INFOCOM*, page 1371, IEEE, Piscataway, New Jersey (March 2000), Tel Aviv, Israel.

[86] M. S. Granovetter, The strength of weak ties. *American Journal of Sociology*, 78: 1360–1380 (1973).

[87] J. W. Grossman and P. D. F. Ion, On a portion of the well-known collaboration graph. *Congressus Numerantium*, 108: 129–131 (1995).

[88] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt and A. Arenas, Self-similar community structure in a network of human interactions. *Phys. Rev. E*, 68: 065103 (2003).

[89] L. H. Hartwell, J. J. Hopfield, S. Leibler and A. W. Murray, From molecular to modular cell biology. *Nature*, 402: C47–C52 (1999).

[90] J. Hasty, D. McMillen, F. Isaacs and J. J. Collins, Computational studies of gene regulatory networks: in numero molecular biology. *Nature Rev. Genet.*, 2: 268 (2001).

[91] Y. Ho, A. Gruhler, A. Heilbut, G. Bader, L. Moore, S.-L. Adams, A. Millar, P. Taylor, K. Bennett and K. Boutillier, Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415: 180–183 (2002).

[92] P. Holme, M. Huss and H. Jeong, Subnetwork hierarchies in biochemical pathways. *Bioinformatics*, 19: 532–538 (2003).

[93] N. S. Holter, A. Maritan, M. Cieplak, N. V. Fedoroff and J. R. Banavar, Dynamic modeling of gene expression data. *Proc. Natl. Acad. Sci. U.S.A.*, 98: 1693 (2001).

[94] B. A. Huberman, *The Laws of the Web*. MIT Press, Cambridge, MA (2001).

[95] C. A. Hutchison, S. N. Peterson, S. R. Gill, R. T. Cline, O. White, C. M. Fraser, H. O. Smith and J. C. Venter, Global transposon mutagenesis and a minimal mycoplasma. *Science*, 286: 2165–2169 (1999).

[96] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori and Y. Sakaki, A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Nat. Acad. Sci.*, 98: 4569–4574 (2001).

[97] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara and Y. Sakaki, Towards a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Nat. Acad. Sci.*, 97: 1143–1147 (2000).

[98] H. Jeong, S. Mason, A.-L. Barabási and Z. N. Oltvai, Lethality and centrality in protein networks. *Nature*, 411: 41–42 (2001).

[99] H. Jeong, Z. Néda and A.-L. Barabási, Measuring preferential attachment for evolving networks. *Euro. Phys. Lett.*, 61: 567 (2003).

[100] H. Jeong, Z. N. Oltvai and A.-L. Barabási, Prediction of protein essentiality based on genomic data. *ComPlexUs*, 1: 19–28 (2003).

[101] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai and A.-L. Barabási, The large-scale organization of metabolic networks. *Nature*, 407: 651–654 (2000).

[102] J. H. Jones and M. S. Handcock, An assessment of preferential attachment as a mechanism for human sexual network formation. *Preprint,* University of Washington (2003).

[103] I. K. Jordan, I. B. Rogozin, Y. I. Wolf and E. V. Koonin, Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.*, 12: 962–968 (2002).

[104] S. Jung, S. Kim and B. Kahng, A geometric fractal growth model for scale free networks. *Phys. Rev. E*, 65: 056101 (2002).

[105] H. Kacser and R. Beeby, Evolution of catalytic proteins: on the origin of enzyme species by means of natural selection. *J. Mol. Evol.*, 20: 38–51 (1984).

[106] V. K. Kalapala, V. Sanwalani and C. Moore, The structure of the United States road network. *Preprint,* University of New Mexico (2003).

[107] R. S. Kamath, A. G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapin, N. Le Bot, S. Moreno and M. Sohrmann, Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature*, 421: 231–237 (2003).

[108] M. Karoński and A. Ruciński, The origins of theory of random graphs. In R. L. Graham and J. Nešetřil, editors, *The Mathematics of Paul Erdős*, Springer, Berlin (1997).

[109] P. D. Karp, M. Riley, M. Saier, I. Paulsen, S. Paley and A. Pellegrini-Toole, The EcoCyc and MetaCyc databases. *Nucl. Acids Res.*, 28: 56–59 (2000).

[110] S. K. Kim, J. Lund, M. Kiraly, K. Duke, M. Jiang, J. M. Stuart, A. Eizinger, B. N. Wylie and G. S. Davidson, A gene expression map for *Caenorhabditis elegans*. *Science*, 293: 2087–2092 (2001).

[111] M. Kimura, Evolutionary rate at the molecular level. *Nature*, 217: 624–626 (1968).

[112] M. Kimura, *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge, England (1983).

[113] J. King and T. Jukes, Non-darwinian evolution. *Science*, 194: 788–798 (1969).

[114] H. Kitano, Systems biology: a brief overview. *Science*, 295: 1662–1664 (2002).

[115] J. Kleinberg, S. R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, The web as a graph: Measurements, models and methods. In *Proc. of the Int. Conf. on Combinatorics and Computing, COCOON'99*, page 1, Springer-Verlag, Berlin (July 1999), Tokyo.

[116] K. Klemm and V. M. Eguíluz, Growing scale-free networks with small-world behavior. *Phys. Rev. E*, 65: 057102 (2002).

[117] A. S. Klovdahl, J. J. Potterat, D. E. Woodhouse, J. B. Muth, S. Q. Muth and W. W. Darrow, Social networks and infectious disease: The Colorado Springs study. *Soc. Sci. Med.*, 38: 79–88 (1994).

[118] M. Kochen, editor, *The Small World*. Ablex, Norwood, NJ (1989).

[119] K. Kohn, Molecular interaction map of mammalian cell-cycle control and DNA repair systems. *Mol. Biol. Cell*, 10: 2703–2734 (1999).

[120] P. L. Krapivsky and S. Redner, Organization of growing random networks. *Phys. Rev. E*, 63 (2001).

[121] P. L. Krapivsky, S. Redner and F. Leyvraz, Connectivity of growing random networks. *Phys. Rev. Lett.*, 85 (2000).

[122] P. L. Krapivsky, G. J. Rodgers and S. Redner, Degree distributions of growing networks. *Phys. Rev. Lett.*, 86 (2001).

[123] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, Trawling the web for emerging cyber-communities. *Computer Networks*, 31: 1481–1493 (1999).

[124] V. Latora and M. Marchiori, Is the Boston subway a small-world network? *Physica A*, 314: 109–113 (2002).

[125] D. Lauffenburger, Cell signaling pathways as control modules: Complexity for simplicity. *Proc. Natl. Acad. Sci.*, 97: 5031–5033 (2000).

[126] S. Lawrence and C. L. Giles, Searching the World Wide Web. *Science*, 280: 98–100 (1998).

[127] S. Lawrence and C. L. Giles, Accessibility of information on the web. *Nature*, 400: 107–109 (1999).

[128] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford and R. A. Young, Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298: 799–804 (2002).

[129] N. K. Lehmann, S. Wuchty and E. Ravasz, The evolution of metabolic networks. *Preprint* (2003).

[130] S. Li, C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.-O. Vidalain, J.-D. J. Han, A. Chesneau and M. Vidal, A map of the interactome network of the metazoan *C. elegans*. *Science*, 303: 540–543 (2004).

[131] F. Liljeros, C. Edling, L. Amaral and Y. Aberg, The web of human sexual contacts. *Nature*, 411: 907–908 (2001).

[132] B. B. Mandelbrot, *The Fractal Geometry of Nature*. Freeman, San Fransisco (1982).

[133] A. Maritan, A. Rinaldo, R. Rigon, A. Giacometti and I. Rodríguez-Iturbe, Scaling laws for river networks. *Phys. Rev. E*, 53: 1510–1515 (1996).

[134] S. Maslov and K. Sneppen, Specificity and stability in topology of protein networks. *Science*, 296: 910–913 (2002).

[135] S. Maslov, K. Sneppen and A. Zaliznyak, Pattern detection in complex networks: Correlation profile of the Internet. *Los Alamos Archive,* cond-mat/0205379 (2002).

[136] I. Matsumura and A. Ellington, In vitro evolution of beta-glucuronidase into a beta-galactosidase proceeds through non-specific intermediates. *J. Mol. Biol.*, 305: 331–339 (2000).

[137] S. McGraith, T. Holtzman, B. Moss and S. Fields, Genome-wide analysis of vaccinia virus protein-protein interactions. *Proc. Natl. Acad. Sci.*, 97: 4879–4884 (2000).

[138] H. W. Mewes, D. Frishman, C. Gruber, B. Geier, D. Haase, A. Kaps, K. Lemcke, G. Mannhaupt and F. Pfeiffer, MIPS: a database for genomes and protein sequences. *Nucl. Acids Res.*, 28: 37–40 (2000).

[139] H. W. Mewes, D. Frishman, U. Güldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Mnsterktter, S. Rudd and B. Weil, MIPS: a database for genomes and protein sequences. *Nucl. Acids. Res.*, 30: 31–34 (2002).

[140] S. Milgram, The small-world problem. *Psychology Today*, 2: 60–67 (1967).

[141] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon, Network motifs: simple building blocks of complex networks. *Science*, 298: 824–827 (2002).

[142] J. M. Montoya and R. V. Solé, Small world patterns in food webs. *J. Theor. Biol.*, 214: 405–412 (2002).

[143] M. Morris, Sexual networks and HIV. *AIDS 97: Year in Review*, 11: 209–216 (1997).

[144] C. R. Myers, Software systems as complex networks: Structure, function, and evolvability of software collaboration graphs. *Phys. Rev. E*, 68: 046116 (2003).

[145] M. E. J. Newman, Models of the small world. *J. Stat. Phys.*, 101: 819–841 (2000).

[146] M. E. J. Newman, Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64: 025102(R) (2001).

[147] M. E. J. Newman, Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E*, 64: 016131 (2001).

[148] M. E. J. Newman, Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64: 016132 (2001).

[149] M. E. J. Newman, The structure of scientific collaboration networks. *Proc. Nat. Acad. Sci.*, 98: 404–409 (2001).

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

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

[152] M. E. J. Newman, Properties of highly clustered networks. *Phys. Rev. E*, 68: 026121 (2003).

[153] M. E. J. Newman, The structure and function of complex networks. *SIAM Review*, 45: 167–256 (2003).

[154] M. E. J. Newman, A. L. Barabási and D. J. Watts, editors, *The Structure and Dynamics of Complex Networks*. Princeton University Press, Princeton (2003).

[155] M. E. J. Newman, S. H. Strogatz and D. J. Watts, Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64: 026118 (2001).

[156] J. D. Noh and H. Rieger, Constrained spin-dynamics description of random walks on hierarchical scale-free networks. *Phys. Rev. E*, 69: 036111 (2004).

[157] J. D. Noh and H. Rieger, Random walks on complex networks. *Phys. Rev. Lett.*, 92: 118701 (2004).

[158] I. Ohno, *Evolution by gene duplication*. Springer, New York (1970).

[159] Z. N. Oltvai and A.-L. Barabási, Life's complexity pyramid. *Science*, 298: 763–764 (2001).

[160] R. Overbeek, N. Larsen, G. Pusch, M. D'Souza, E. S. Jr., N. Kyrpides, M. Fonstein, N. Maltsev and E. Selkov, WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Science*, 28: 123–125 (2000).

[161] A. Pandey and M. Mann, Proteomics to study genes and genomes. *Nature*, 405: 837–846 (2000).

[162] J. Park, M. Lappe and A. Teichmann, Mapping protein family interations: Intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J. Mol. Biol.*, 307: 929–938 (2001).

[163] R. Pastor-Satorras, E. Smith and R. Solé, Evolving protein interaction networks through gene duplication. *Santa Fe Working paper*, pages 02–02–008 (2002).

[164] R. Pastor-Satorras, A. Vázquez and A. Vespignani, Dynamical and correlation properties of the Internet. *Phys. Rev. Lett.*, 87: 258701 (2001).

[165] R. Pastor-Satorras and A. Vespignani, Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86: 3200–3203 (2001).

[166] R. Pastor-Satorras and A. Vespignani, Immunization of complex networkss. *Phys. Rev. E.*, 65: 036104 (2002).

[167] S. L. Pimm, *The Balance of Nature*. University of Chicago, Chicago (1991).

[168] J. Podani, Z. N. Oltvai, H. Jeong, B. Tombor, A.-L. Barabaási and E. Szathmáry, Comparable system-level organization of archae and eukaryotes. *Nature Genet.*, 29: 54–56 (2001).

[169] J. J. Potterat, L. Phillips-Plummer, S. Q. Muth, R. B. Rothenberg, D. E. Woodhouse, T. S. Maldonado-Long, H. P. Zimmerman and J. B. Muth, Risk network structure in the early epidemic phase of HIV transmission in Colorado Springs. *Sexually Transmitted Infections*, 78: i159–i163 (2002).

[170] S. Raillard, A. Krebber, Y. Chen, J. Ness, E. Bermudez, R. Trinidad, R. Fullem, C. Davis, M. Welch, J. Seffernick, L. Wackett, W. Stemmer and J. Minshull, Novel enzyme activities and functional plasticity revealed by recombining highly homologous enzymes. *Chem. Biol.*, 8: 891–898 (2001).

[171] J. C. Rain, L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik and V. Schachter, The protein-protein interaction map of *Helicobacter pylori*. *Nature*, 409: 211–215 (2001).

[172] C. V. Rao and A. P. Arkin, Control motifs for intracellular regulatory networks. *Annu. Rev. Biomed. Eng.*, 3: 391–419 (2000).

[173] E. Ravasz and A.-L. Barabási, Hierarchical organization in complex networks. *Phys. Rev. E*, 67: 026122 (2002).

[174] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai and A.-L. Barabási, Hierarchical organization of modularity in metabolic networks. *Science*, 297: 1551–1555 (2002).

[175] S. Redner, How popular is your paper? An empirical study of the citation distribution. *Euro. Phys. Journal B*, 4: 131–135 (1998).

[176] S. Redner, Citation statistics from more than a century of physical review. *Los Alamos Archive,* physics/0407137 (2004).

[177] A. Rinaldo, I. Rodríguez-Iturbe and R. Rigon, Channel networks. *Annual Review of Earth and Planetary Science*, 26: 289–327 (1998).

[178] I. Rodríguez-Iturbe and A. Rinaldo, *Fractal River Basins: Chance and Self-Organization*. Cambridge University Press, Cambridge (1997).

[179] P. Ross-Macdonald, P. S. Coelho, T. Roemer, S. Agarwal, A. Kumar, R. Jansen, K. H. Cheung, A. Sheehan, D. Symoniatis and N. Umansky, Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature*, 402: 413–418 (1999).

[180] H. Salgado, A. Santos-Zavaleta, S. Gama-Castro, D. Millán-Zárate, E. Díaz-Peredo, F. Sánchez-Solano, E. Pérez-Rueda, C. Bonavides-Martínez and J. Collado-Vides, RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.*, 29: 72–74 (2001).

[181] C. H. Schilling, D. Letscher and B. O. Palsson, Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.*, 203: 229 (2000).

[182] A. Schneeberger, C. H. Mercer, S. A. J. Gregson, N. M. Ferguson, C. A. Nyamukapa, R. M. Anderson, A. M. Johnson and G. P. Garnett, Scale-free networks and sexually transmitted diseases – A description of observed patterns of sexual contacts in Britain and Zimbabwe. *Sexually Transmitted Diseases*, 31: 380–387 (2004).

[183] H. G. Schuster, *Complex Adaptive Systems*. Scator Verlag, Saarbruskey, Germany (2002).

[184] B. Schwikowski, P. Uetz and S. Fields, A network of protein-protein interactions in yeast. *Nature Biotechn.*, 18: 257–1261 (2000).

[185] P. O. Seglen, The skewness of science. *J. Amer. Soc. Inform. Sci.*, 43: 628–638 (1992).

[186] P. Sen, S. Dasgupta, A. Chatterjee, P. A. Sreeram, G. Mukherjee and S. S. Manna, Small-world properties of the indian railway network. *Los Alamos Archive,* cond-mat/0208535 (2002).

[187] M. A. Serrano and M. Boguñá, Topology of the world trade web. *Phys. Rev. E*, 68: 015101 (2003).

[188] S. Shen-Orr, R. Milo, S. Mangan and U. Alon, Network motifs in the transcriptional regulation network of *E. coli*. *Nature Genet.*, 31: 64–68 (2002).

[189] M. Sigman and G. Cecchi, Global organization of the Wordnet lexicon. *Proc. Nat. Acad. Sci.*, 99: 1742–1747 (2002).

[190] V. Smith, D. Botstein and P. O. Brown, Genetic footprinting: a genomic strategy for determining a gene's function given its sequence. *Proc. Natl. Acad. Sci. U.S.A.*, 92: 64796483 (1995).

[191] J. Sokal, *Numerical Taxonomy*. Freeman, San Francisco (1973).

[192] R. Solé, R. Pastor-Satorras, E. Smith and T. Kepler, A model of large-scale proteome evolution. *Adv. Compl. Sys.*, 5: 43–54 (2002).

[193] D. Stauffer and A. Aharony, *Percolation Theory*. Taylor & Francis, London (1992).

[194] M. Steyvers and J. B. Tenenbaum, The largescale structure of semantic networks: Statistical analyses and a model for semantic growth. *Los Alamos Archive,* cond-mat/0110012 (2001).

[195] G. Szabó, M. Alava and J. Kertész, Structural transitions in scale-free networks. *Phys. Rev. E*, 67: 056102 (2003).

[196] P. Uetz, L. Giot, G. Cagney, T. Mansfield, R. Judson, J. Knight, D. Lockshorn, V. Narayan, M. Srinivasan and P. Pochart, A comprehensive analysis of protein-protein interactions of *Saccharomyces cerevisiae*. *Nature*, 403: 623–627 (2000).

[197] A. Vázquez, Statistics of citation networks. *Los Alamos Archive,* cond-mat/0105031 (2001).

[198] A. Vázquez, A. Flammini, A. Maritan and A. Vespignani, Modelling of protein interaction networks. *ComPlexUs*, 1: 38–44 (2003).

[199] A. Vázquez, Y. Moreno, M. Boguñá, R. Pastor-Satorras and A. Vespignani, Topology and correlations in structured scale-free networks. *Phys. Rev. E*, 67: 046111 (2003).

[200] A. Vázquez, R. Pastor-Satorras and A. Vespignani, Large-scale topological and dynamical properties of the Internet. *Phys. Rev. E*, 65: 066130 (2002).

[201] A. Vázquez, R. Pastor-Satorras and A. Vespignani, Internet topology at the router and autonomous system level. *Los Alamos Archive,* cond-mat/0206084 (2002).

[202] T. Vicsek, *Fractal Growth Phenomena*. World Scientific, Singapore (1989).

[203] B. Vogelstein, D. Lane and A. Levine, Surfing the p53 network. *Nature*, 408: 307–310 (2000).

[204] A. Wagner, Mutational robustness in genetic networks of yeast. *Nat. Genet.*, 24: 355–361 (2000).

[205] A. Wagner, The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.*, 18: 1283–1292 (2001).

[206] A. Wagner and D. A. Fell, The small world inside large metabolic networks. *Proc. Roy. Soc. London Series B*, 268: 1803–1810 (2001).

[207] A. Walhout, R. Sordella, X. Lu, J. Hartley, G. Temple, M. Brasch, N. Thierry-Mieg and M. Vidal, Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science*, 287: 116–122 (2000).

[208] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge University, Cambridge (1994).

[209] D. J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, Princeton (1999).

[210] D. J. Watts, P. S. Dodds and M. E. J. Newman, Identity and search in social networks. *Science*, 296: 130 (2002).

[211] D. J. Watts and S. H. Strogatz, Collective dynamics of small-world networks. *Nature*, 393: 440–442 (1998).

[212] J. G. White, E. Southgate, J. N. Thompson and S. Brenner, The structure of the nervous system of the nematode *C. elegans*. *Phil. Trans. R. Soc. London*, 314: 1340 (1986).

[213] R. J. Williams, E. L. Berlow, J. A. Dunne, A.-L. Barabási and N. D. Martinez, Two degrees of separation in complex food webs. *Proc. Nact. Acad. Sci.*, 99: 12913–12916 (2002).

[214] E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J. D. Boeke and H. Bussey, Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, 285: 901–906 (1999).

[215] Y. Wolf, G. Karev and E. Koonin, Scale-free networks in biology: new insights into the fundamentals of evolution? *Bioessays*, 24: 105–109 (2002).

[216] S. Wuchty, Scale-free behavior in protein domain networks. *Mol. Biol. Evol.*, 18: 1694–1702 (2001).

[217] S. Wuchty, Interaction and domain networks of yeast. *Proteomics*, 2: 1715–1723 (2002).

[218] S. Wuchty, Small-worlds in RNA. *Nucl. Acids Res.*, 31: 1108–1117 (2003).

[219] I. Xenarios, E. Fernandez, L. Salwinski, X. Duan, M. Thompson, E. Marcotte and D. Eisenberg, DIP: the database of interacting proteins: 2001 update. *Nucl. Acids Res.*, 29: 239–241 (2001).

[220] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte and D. M. S. Eisenberg, DIP: the database of interacting proteins. *Nucl. Acids. Res.*, 28: 289 (2000).

[221] S.-H. Yook, H. Jeong and A. L. Barabási. *Preprint* (2003).

[222] S.-H. Yook, H. Jeong and A. L. Barabási, Modelling the Internet's large-scale topology. *Proc. Nact. Acad. Sci.*, 99: 13382–13386 (2003).

[223] S.-H. Yook, Z. N. Oltvai and A. L. Barabási, Functional and topological characterization of protein-protein interaction networks. *Proteomics*, 4: 928–942 (2004).