SELF-SIMILARITY AND SCALING THEORY OF COMPLEX NETWORKS

by

Chaoming Song

A dissertation submitted to the Graduate Faculty in Physics in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2008

UMI Number: 3296953

UMI Microform 3296953

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

This manuscript has been read and accepted for the
Graduate Faculty in Engineering in satisfaction of the
dissertation requirement for the degree of Doctor of Philosophy.

Hernan Makse

[**required signature**]

Date — Chair of Examining Committee

Sultan Catto

[**required signature]**

Date — Executive Officer

Hernan Makse, Physics Department, CCNY

Joel Koplik, Physics Department, CCNY

Sergey Buldyrev, Physics Department, Yeshiva University

Sergey Vitkalov, Physics Department, CCNY

Carlos Meriles, Physics Department, CCNY

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Acknowledgements

I am deeply grateful to my supervisor, Professor Hernan Makse, Physics Department and Levich Institute, City College of New York, for his important support and constructive comments throughout this work.

I am grateful to Professor Shlomo Havlin, Physics Department, Bar-Ilan University, who contributed to most parts of this work.

I wish to express my thanks to Dr. Lazaros Gallos, who contributed to many parts of this work.

Abstract

SELF-SIMILARITY AND SCALING THEORY OF COMPLEX NETWORKS

by

Chaoming Song

Adviser: Professor Hernan Makse

Scale-free networks have been studied extensively due to their relevance to many real systems as diverse as the World Wide Web (WWW), the Internet, biological and social networks. We present a novel approach to the analysis of scale-free networks, revealing that their structure is self-similar. This result is achieved by the application of a renormalization procedure which coarse-grains the system into boxes containing nodes within a given "size". Concurrently, we identify a power-law relation between the number of boxes needed to cover the network and the size of the box defining a self-similar exponent, which classifies fractal and non-fractal networks. By using the concept of renormalization as a mechanism for the growth of fractal and non-fractal modular networks, we show that the key principle that gives rise to the fractal architecture of networks is a strong effective "repulsion" between the most connected nodes (hubs) on all length scales, rendering them very dispersed. We show that a robust network comprised of functional modules, such as a cellular network, necessitates a fractal topology, suggestive of a evolutionary drive for their existence. These fundamental properties help to understand the emergence of the scale-free property in complex networks.

Contents

List of Figures **viii**

1 Introduction **1**

1.1 Scale free networks 2

1.2 Small-world property 2

1.3 Fractality in real-world networks 3

2 Fractality of Complex Networks **5**

2.1 The box covering method 6

2.2 Fractal dimension 7

2.2.1 Cellular networks 10

2.2.2 Internet 21

2.2.3 Protein-protein interaction networks 22

2.2.4 Random scale-free network 25

2.2.5 The Barabási-Albert model and the Erdös-Rényi random graph 26

2.3 Paradox between fractality and small-world 27

2.4 Renormalization 32

2.5 Degree dimension and scaling relationship 35

3 Box Covering Algorithms **42**

3.1 The greedy coloring algorithm 42

3.2 Burning algorithms 49

3.2.1 Burning with the diameter ℓ_B, and the Compact-Box-Burning (CBB) algorithm 51

3.2.2 Burning with the radius r_B, and the Maximum-Excluded-Mass-Burning (MEMB) algorithm 55

3.2.3 Comparison between the different algorithms 60

3.3 Box-size correction 63

4 Scaling Theory of the Correlation **66**

4.1 Joint probability distribution 66

4.2 Degree correlation profile 67

4.3 Renormalization of degree correlation 70

4.4 Measurement of correlation exponent 73

4.5 Scaling theory 75

4.5.1 Hub-repulsion dimension 75

4.5.2 Relation between ϵ and d_e 78

4.6 Classification of real-world networks 79

5 Modelling Self-Similar Networks **83**

5.1 Growth mechanism 83

5.2 Mathematical model 88

5.3 Scaling theory 91

5.4 The minimal model 95

5.5 Additional supporting evidence for the fractal network model 99
5.6 Study of other scale-free models 99
5.7 Global small world: short cuts in the network 103
5.8 Resilience of fractal networks under intentional attack 107

6 Network Dynamics **109**
6.1 Introduction 109
6.1.1 Metabolism modelling 111
6.2 Modularity, diffusion and resistance 112
6.3 Resistance and Diffusion 116
6.3.1 Resistance measurements 117
6.3.2 Diffusion measurements 117
6.3.3 Scaling exponents of R and T 118
6.4 Renormalization and scaling theory 119
6.5 Influence of modularity on transport 123
6.6 Flow distribution across the network 125

Bibliography **131**

List of Figures

1.1 Representation of the Protein Interaction Network of Yeast. The colors show different subgroups of proteins that participate in different functionality classes [15]. 4

2.1 Details of the box covering method for $\ell_B = 2$ in a schematic network with 8 nodes. Left panel show a minimum covering with 4 boxes. Right panel is a different covering for the same network with $N_B = 5$. 6

2.2 Log-log plot of the N_B vs ℓ_B revealing the fractal dimension of the WWW. 8

2.3 Log-log plot of the N_B vs ℓ_B revealing the fractal dimension of the two protein interaction networks: *H. sapiens* and *E. coli* 8

2.4 Log-log plot of the N_B vs ℓ_B revealing the fractal dimension of the the cellular networks of *A. fulgidus*, *E. coli* and *C. elegans* according to Eq. (2.1). 9

2.5 Internet: Log-log plot of $N_B(\ell_B)$. The solid line represents the modified power law fit, Eq. (2.2). The inset shows a linear-log plot indicating that the decay is slower than exponential. 23

2.6 Log-log plot of N_B versus ℓ_B for different protein-protein interaction networks. While *E. coli* and *H. sapiens* show a clear power law behavior, the other protein networks show a modified power-law behaviour or a pure exponential decay. The inset shows a linear-log plot of $N_B(\ell_B)$. 24

2.7 Log-log plot of N_B versus ℓ_B for random scale-free network with $\gamma = 2.35$, which shows a pure exponential decay. The inset shows a linear-log plot of $N_B(\ell_B)$. 25

2.8 Barabási-Albert model of scale-free networks with preferential attachment for 150,000 nodes and $m = m_0 = 3$ and $m = m_0 = 5$. m_0 is the initial number of nodes in the system and m is the number of links of a newly created node in the dynamical growth of the network [19]. Log-log plot of N_B versus ℓ_B showing the lack of a power law behaviour. The inset shows a linear-log plot indicating that N_B decreases faster than exponential with ℓ_B. 26

2.9 Mean value of the box mass in the box covering method, $\langle M_B \rangle$, and the cluster mass in the cluster growing method, $\langle M_c \rangle$, for the WWW. The solid lines represent the power-law fit for $\langle M_B \rangle$ and the exponential fit for $\langle M_c \rangle$ according to Eqs. (2.4) and (2.5), respectively. 29

2.10 Probability distribution of M_B and M_c for $\ell_B = 4$ for the WWW. The curves are fitted by a power-law and a log-normal distribution, respectively. 31

2.11 Demonstration of the renormalization method to complex networks for different ℓ_B in a network demo. The first column depicts the original network. We tile the system with boxes of size ℓ_B (different colors correspond to different boxes). All nodes in a box are connected by a minimum distance smaller than the given ℓ_B. For instance, in the case of $\ell_B = 2$, we identify four boxes which contain the nodes depicted with colors red, orange, white, and blue, each containing 3, 2, 1, and 2 nodes, respectively. Then we replace each box by a single node; two renormalized nodes are connected if there is at least one link between the unrenormalized boxes. Thus we obtain the network shown in the second column. The resulting number of boxes used to tile the network $N_B(\ell_B)$ versus ℓ_B gives the fractal dimension as in Eq. (2.1). The renormalization procedure is applied again and repeated until the network is reduced to a single node (third and fourth columns for different ℓ_B). 33

2.12 Three stages in the renormalization scheme applied to the entire WWW for $\ell_B = 3$. Here we color the nodes in the web according to the boxes to which they belong. 34

2.13 The degree distribution of WWW [8] according to different box size ℓ_B. 36

2.14 The degree distribution of Protein interaction network of *yeast*, according to different box size ℓ_B 36

2.15 The degree distribution of the inward degree (a) and the outward degree (b) of metabolic network of E. *coli* , according to different box size ℓ_B. 37

2.16 The degree distribution of Internet [26], according to different box size ℓ_B. 38

2.17 Log-log plot of the s vs ℓ_B revealing the degree dimension of the WWW. 38

2.18 Log-log plot of the s vs ℓ_B revealing the degree dimension of the two protein interaction networks: *H. sapiens* and *E. coli*. 39

2.19 Log-log plot of the s vs ℓ_B revealing the degree dimension of the cellular networks of *A. fulgidus*, *E. coli* and *C. elegans* according to Eq. (2.1). 39

3.1 Illustration of the solution for the network covering problem via mapping to the graph coloring problem. Starting from G (upper left panel) we construct the dual network G' (upper right panel) for a given box size (here $\ell_B = 3$), where two nodes are connected if they are at a distance $\ell \geq \ell_B$. We use a greedy algorithm for vertex coloring in G', which is then used to determine the box covering in G, as shown in the plot. 43

3.2 (color online) Probability distribution function $P(N_B)$ of the number of boxes N_B for the greedy algorithm, applied to the cellular network of *E.coli*. Different box sizes ℓ_B are used as indicated in the plot. Inset: PDFs of the normalized quantity $N_B/\langle N_B\rangle$ in a semi-log plot for the greedy algorithm, suggesting that $P(N_B)$ follows a Gaussian distribution. 45

3.3 Normalized variance σ_B of the greedy algorithm for different box sizes ℓ_B, for (a) fractal and (b) non-fractal networks, where the box size ℓ_B is normalized by the maximum box size $\ell_B^{\max}$. The slopes for the fractal networks are (left to right): $\delta = 0.85$, 1.3, 2.2. For non-fractal networks: $\delta = 1.5$. 47

3.4 Comparison of the minimum $N_B^{\min}$ (line) and mean $\langle N_B\rangle$ (symbols) number of boxes for the greedy algorithm after 10,000 random reshuffles in real-world networks. 48

3.5 Our definitions for a box that is (a) non-compact for $\ell_B = 3$, i.e. could include more nodes, (b) compact, (c) connected, and (d) disconnected (the nodes in the right box are not connected in the box). (e) For this box, the values $\ell_B = 5$ and $r_B = 2$ verify the relation $\ell_B = 2r_B + 1$. (f) One of the pathological cases where this relation is not valid, since $\ell_B = 3$ and $r_B = 2$. 50

3.6 Two-dimensional geometrical analogue of the CBB algorithm. Initially we choose a random point and consider the circle with radius ℓ_B. We then choose another random point within this circle which serves as a new circle center and calculate the union of these two circles. We continue by iteratively selecting random centers for circles in the union of all the previous circles. 52

3.7 Illustration of the CBB algorithm for $\ell_B = 3$. (a) Initially, all nodes are candidates for the box. (b) A random node is chosen, and nodes at a distance further than ℓ_B from this node are no longer candidates. (c) The node chosen in (b) becomes part of the box and another candidate node is chosen. The above process is then repeated until the box is complete. 53

3.8 Comparison of (a) the mean number of boxes, $\langle N_B \rangle$, and (b) the normalized variance, σ_B, between the greedy algorithm and CBB. . 54

3.9 Burning with the radius r_B from (a) a hub node or (b) a non-hub node results in very different network coverage. In (a) we need just one box of $r_B = 1$ while in (b) 5 boxes are needed to cover the same network. This is an intrinsic problem when burning with the radius. (c) Burning with the maximum distance ℓ_B (in this case $\ell_B = 2r_B + 1 = 3$) we avoid this situation, since independently of the starting point we would still obtain $N_B = 1$. 56

3.10 Illustration of the MEMB algorithm for $r_B = 1$. *Upper row: Calculation of the box centers* (a) We calculate the excluded mass for each node. (b) The node with maximum mass becomes a center and the excluded masses are recalculated. (c) A new center is chosen. Now, the entire network is covered with these two centers. *Bottom row: Calculation of the boxes* (d) Each box includes initially only the center. Starting from the centers we calculate the distance of each network node to the closest center. (e) We assign each node to its nearest box. 57

3.11 Comparison between the number of boxes obtained using MEMB, N_B^{MEMB}, and the mean number of boxes, $\langle N_B \rangle^{\mathrm{greedy}}$, obtained from the greedy algorithm. 59

3.12 Comparison of the distribution of N_B for 10^4 realizations of the four network covering methods presented in this paper. Notice that three of these methods yield very similar results with narrow distributions and comparable minimum values, while the random burning algorithm fails to reach a value close to this minimum (and yields a broad distribution). 61

3.13 Comparison of the mean number of boxes $\langle N_B \rangle$ vs ℓ_B for the four presented algorithms. All methods yield the same value of the fractal dimension $d_B = 3.5$. 62

3.14 Comparison of the fractal dimension before and after applying the box-size correction in (a) an Erdos-Renyi model at criticality and (b) a fractal network model. The straight lines correspond to the analytical predictions. Insets: Improvement of the fractal dimension d_B calculation as we increase the number of boxes used for this calculation. 64

4.1 Empirical results on real complex networks. (a) Schematics showing that fractal networks are characterized by a power law dependence between N_B and ℓ_B while (a) non-fractal networks are characterized by an exponential dependence. (b) Plot of the correlation profile of the fractal metabolic network of E. *coli*, $R_{\mathrm{E.}coli}(k_1, k_2)/R_{\mathrm{WWW}}(k_1, k_2)$, and (d) the non-fractal Internet $R_{\mathrm{Int}}(k_1, k_2)/R_{\mathrm{WWW}}(k_1, k_2)$, compared with the profile of the WWW in search of a signature of fractality. 68

4.2 The joint degree distribution $P(k_1, k_2)$ of WWW (top row) and Internet at the router level (bottom row) before and after renormalization (with $\ell_B = 3$). 71

4.3 Factor $E_b(k)$ as defined in the text for three successive renormalization stages of the Internet at the router level, for $\ell_B = 3$ and $b = 3$. The data have been vertically shifted in order to show the invariance. Inset: Invariance of $E_b(k)$ vs k for different values of b. 75

4.4 The quantity $E_b(k)$ can distinguish between the fractal (WWW, protein homology network) and non-fractal (Internet, cond-mat coauthorship) topology of networks. The different topologies correspond to different scaling behavior with the degree k. 76

4.5 Scaling of $\mathcal{E}(\ell_{\mathcal{B}})$ as defined in Eq. (4.6) for the fractal topology of the WWW with $d_e = 1.5$, and the non-fractal topology of the Internet showing that fractal topologies are strongly anticorrelated at all length scales. 77

4.6 Classification of scale-free networks. The thin curves correspond to the prediction of the minimal model for varying e values. The line $\epsilon = \gamma - 1$ corresponds to a completely random network structure. The line $\epsilon = 2$ separates fractal ($\epsilon > 2$) from non-fractal networks ($\epsilon \leq 2$), while the line $\epsilon = \gamma$ describes a fractal tree. The schematics illustrate networks where hub correlations are stronger than in random networks (area I), weaker than random but non-fractal (area II), non-fractal according to the minimal model ($\epsilon = 2$), and fractal (area III). 80

5.1 Empirical results on real complex networks. (a) Schematics showing that fractal networks are characterized by a power law dependence between N_B and ℓ_B while (b) non-fractal networks are characterized by an exponential dependence. 84

5.2 The dynamical growth process can be seen as the inverse renormalization procedure with all the properties of the network being invariant under time evolution. In this example $\tilde{N}(t) = 16$ nodes are renormalized with $N_B(\ell_B) = 4$ boxes of size $\ell_B = 3$. 86

5.3 Analysis of Mode I, only: the boxes are connected directly leading to strong hub-hub attraction or assortativity. This mode produces a scale-free, small-world network but without the fractal topology. 87

5.4 Mode II alone produces a scale-free with a fractal topology but not the small-world effect. Here the boxes are connected via non-hubs . 87

5.5 Different modes of growth with $m = 2$. Starting with (a) five nodes at $t = 0$, the different connectivity modes lead to different topological structures, which are (b) Mode I, (c) Mode II and (d) combination of Mode I and II with probability $e = 0.5$. 96

5.6 Predictions of the model for $e = 0.5$ for the degree distribution showing the power law behavior with $\gamma = 3$ and its invariance under time evolution. 97

5.7 Resulting topology predicted by the minimal model for $e = 0.8$, $n = 5$, $a = 1.4$, $s = 3$ and $m = 2$. The colors of the nodes show the modular structure with each color representing a different box. We also include loops in the structure as discussed later on. 100

5.8 Ratio $R_{e=1}(k_1, k_2)/R_{e=0.8}(k_1, k_2)$ to compare the hub-hub correlation emerging from the model networks generated with $e = 1$ and $e = 0.8$, respectively. 101

5.9 Plot of N_B versus ℓ_B showing that Mode I is non-fractal (exponential decay) and $e = 0.8$ is fractal (power-law decay) according to Fig.5.8 and in agreement with the empirical results of Fig. 4.1. 102

5.10 Scaling of $\mathcal{E}(\ell_B)$ reproducing the behavior of fractal networks for $e = 0.8$ and non-fractal networks Mode I, $e = 1$, as found empirically in Fig. 4.5. 102

5.11 Lack of fractality in the BA model of preferential attachment [19] . 103

5.12 Lack of fractality in the hierarchical model [50] 104

5.13 Lack of fractality in the model of Jung, Kim, and Kahng (JKK model)[59] which is an example of pseudo fractal models as discussed by Dorogovtsev and Mendes [58]. 104

5.14 The scaling of $\langle M_B \rangle$ does not show finite size effects. The initial exponential dependence range of $\langle M_c \rangle$ increases as the size of the network increases. 105

5.15 N_B vs ℓ_B for the model with $e = 0.5$ shows that the fractality still holds in the presence of random noise. The straight line gives the theoretical prediction of the model $d_B = 2$. 105

5.16 Average of the shortest path between two nodes as a function of the system size showing the global small world for the model ($e = 0.5$). 106

6.1 Example of a network G that undergoes renormalization to a network G'. 112

6.2 Typical behavior of the probability distributions for the resistance R vs R' and the diffusion time T vs T', respectively, which verifies that the ratios of these quantities during a renormalization stage are roughly constant for all pairs of nodes. 113

6.3 Average value of this ratio for the resistance R/R' and the diffusion time T/T', respectively, as measured for different ℓ_B values. 114

6.4 Rescaling of (a) the resistance and (b) the diffusion time according to Eqs. (6.10) and (6.11) for the protein interaction network of yeast (upper symbols) and the fractal network generation model (lower filled symbols). The data for PIN have been vertically shifted upwards by one decade for clarity. Different symbols correspond to different ratios k_1/k_2 and different colors denote a different value for k_1. Inset: Resistance R as a function of distance ℓ, before rescaling, for constant ratio $k_1/k_2 = 1$ and different k_1 values. 122

6.5 Comparison of the random walk exponent d_w extracted numerically (symbols) with the theoretical prediction (Eq. (6.16), line) vs the modularity exponent d_M, for different values of m and x. Open circles correspond to the result for the PIN and metabolic networks. Inset: Direct (unscaled) numerical calculation of d_w as a function of m, for varying x values (shown in the plot). 126

6.6 Probability distribution $P(I)$ of current magnitudes I flowing through the links in PIN (solid triangles) and metabolic networks (solid circles). Empty symbols are the corresponding results for the randomly rewired networks. Inset: Invariance of $P(I)$ for the metabolic network under renormalization with different ℓ_B values. 127

6.7 Probability distribution $P(I)$ for the fractal model before and after randomly adding 1% of links or rewiring 10% of the network. Inset: $P(I)$ for the fractal model with varying d_M values, where $m = 2$ and x varies from 2 to 4. 128

6.8 (a) Current flow through the links of the yeast PIN network, for one random selection of the two nodes acting as current input/output. (b) Minimum spanning tree for the PIN. The thickness of a link corresponds to the current flowing through this link. Different node colors correspond to different protein functions. 129

Chapter 1

Introduction

Self-similarity is a property of fractal structures, a concept introduced by Mandelbrot and one of the fundamental mathematical results of the 20th century [1, 2, 3]. The importance of fractal geometry stems from the fact that these structures were recognized in numerous examples in Nature, from the coexistence of liquid/gas at the critical point of evaporation of water [4, 5, 6], to snowflakes, to the tortuous coastline of the Norwegian fjords, to the behavior of many complex systems such as economic data, or the complex patterns of human agglomeration [2, 3].

Typically, real world scale-free networks exhibit the small-world property [7], which implies that the number of nodes increases exponentially with the diameter of the network, rather than the power-law behavior expected for self-similar structures. For this reason complex networks were believed to *not* be length-scale invariant or self-similar.

1.1 Scale free networks

The study of real complex networks has revealed that many of them share some fundamental common properties. Of great importance is the form of the degree distribution for these networks, which is unexpectedly wide. This means that the degree of a node may assume values that span many decades. Thus, although the majority of nodes have a relatively small degree, there is a finite probability that a few nodes will have degree of the order of thousands or even millions. Networks that exhibit such a wide distribution $P(k)$ are known as *scale-free* networks, where the term refers to the absence of a characteristic scale in the degree k. This distribution very often obeys a power-law form with a degree exponent γ, usually in the range $2 < \gamma < 4$ [8],

$$P(k) \sim k^{-\gamma} \,. \tag{1.1}$$

1.2 Small-world property

A more generic property, that is usually inherent in scale-free networks but applies equally well to other types of networks, such as in Erdős-Rényi random graphs, is the *small-world* feature. Originally discovered in sociological studies [9], it is the generalization of the famous 'six degrees of separation' and refers to the very small network diameter. Indeed, in small-world networks a very small number of steps is required to reach a given node starting from any other node. Mathematically this is expressed by the slow (logarithmic) increase of the average diameter of the network, $\bar{\ell}$, with the total number of nodes N, $\bar{\ell} \sim \ln N$, where ℓ is the *shortest* distance between two nodes and defines the distance metric in complex networks

[10, 11, 12, 8], namely,

$$N \sim e^{\bar{\ell}/\ell_0}, \tag{1.2}$$

where ℓ_0 is a characteristic length.

1.3 Fractality in real-world networks

These network characteristics have been shown to apply in many empirical studies of diverse systems [7, 13, 8]. The simple knowledge that a network has the scale-free and/or small-world property already enables us to qualitatively recognize many of its basic properties. However, structures that have the same degree exponents may still differ in other aspects [14]. For example, a question of fundamental importance is whether scale-free networks are also self-similar or fractals. The illustrations of scale-free networks (see e.g. Figs. 1.1 and 2.12b) seem to resemble traditional fractal objects. Despite this similarity, Eq. (1.2) definitely appears to contradict a basic property of fractality: fast increase of the diameter with the system size. Moreover, a fractal object should be self-similar or invariant under a scale transformation, which is again not clear in the case of scale-free networks where the scale has necessarily limited range. So, how is it even possible that fractal scale-free networks exist? In the following, we will see how these seemingly contradictory aspects can be reconciled.

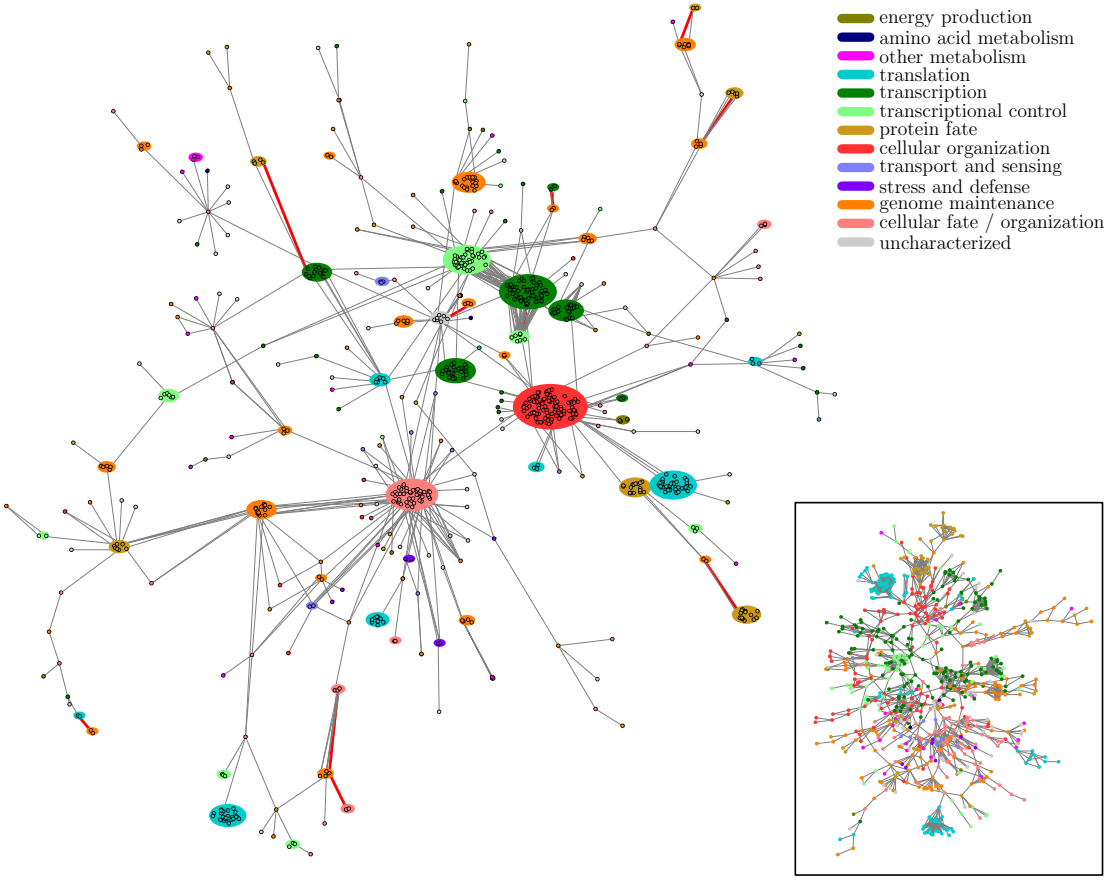


Figure 1.1: Representation of the Protein Interaction Network of Yeast. The colors show different subgroups of proteins that participate in different functionality classes [15].

Chapter 2

Fractality of Complex Networks

The classical theory of self-similarity requires a power-law relation between the number of nodes N and the diameter of a fractal object ℓ [16, 17]. The fractal dimension can be calculated using either *box-counting* or *cluster-growing* techniques [2]. In the first method the network is covered with N_B boxes of linear size ℓ_B. The fractal dimension or box dimension d_B is then given by [3]:

$$N_B \sim \ell_B^{-d_B} \,. \tag{2.1}$$

In order to demonstrate this concept we first consider a self-similar network embedded in Euclidean space, of which a classical example would be a fractal percolation cluster at criticality [16]. In order to unfold the self-similar properties of such clusters we calculate the fractal dimension using a "box covering" method and a "cluster growing" method [2].

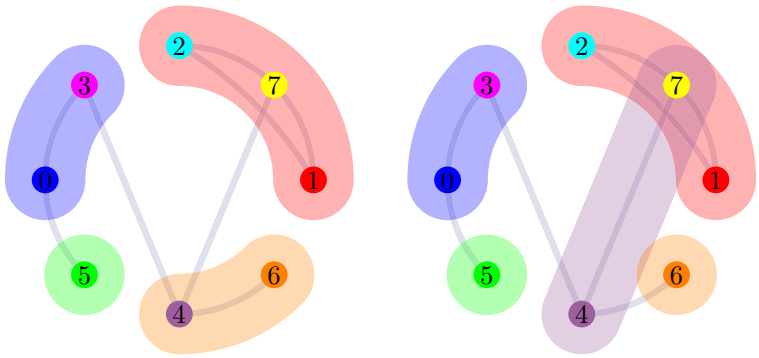


Figure 2.1: Details of the box covering method for $\ell_B = 2$ in a schematic network with 8 nodes. Left panel show a minimum covering with 4 boxes. Right panel is a different covering for the same network with $N_B = 5$.

2.1 The box covering method

The box covering method is central to the understanding of the scale-invariant properties of networks.

As a sample, we build a schematic network to demonstrate the box covering method in Fig. 2.1a. We tile the system by first assigning nodes 1 and 2 to the box colored in blue. Notice that the maximum distance between the nodes of a given box is $\ell_B - 1$. Thus, node 8 would not be in the blue box since its distance from node 2 is $\ell = 2$ (even though its distance from 1 is $\ell = 1$). Then we cover the nodes 6 and 7 with the orange box, and the nodes 3, 4, and 5 with the red box. Finally, the last node 8 is assigned to the green box. The number of boxes to cover the network is then $N_B = 4$.

From the above explanation it should be clear that there are many ways to tile the network. For instance in Fig. 2.1b we show another tiling. In this case we assign nodes 4 and 7 together in a single box instead of nodes 6 and 7 as in Fig. 2.1a. This tiling results in an extra box needed to cover node 6 and therefore in a

larger number of nodes to tile the system, $N_B = 5$.

While there are many ways to assign nodes to the boxes, we notice that the rigorous mathematical definition of Eq. (2.1) corresponds to the *minimum* number of boxes needed to cover the network [3]. This minimization does not have any consequence for the determination of the fractal dimension in homogeneous clusters. However, it may become relevant when calculating the fractal dimension of a complex network with a *widely* distributed number of links. Finding the minimum number of boxes to cover the network is a hard optimization problem to solve, analogous to the graph coloring problem in the NP-complete complexity class. This minimization problem has to be solved by an exhaustive numerical search since there is no numerical algorithm to solve this kind of problems.

We have performed the search over a limited part of the phase-space for the WWW to obtain an estimation of the average and the minimum number of boxes needed to tile the network for every value of ℓ_B. We find that the average value of the boxes is very close to the estimated minimum number of boxes. Moreover, we find that the minimization is not relevant and any covering gives rise to the same fractal dimension. Systematic study of optimal covering algorithms will be discussed in next chapter.

2.2 Fractal dimension

This procedure is applied to several different real networks: *(i)* a part of the WWW composed of 325,729 web-pages which are connected if there is a URL link from one page to another [8, 18], *(ii)* a social network where the nodes are 392,340

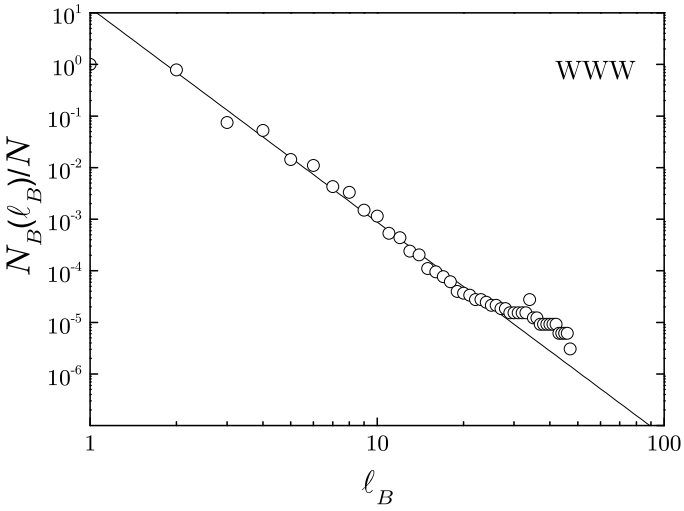


Figure 2.2: Log-log plot of the N_B vs ℓ_B revealing the fractal dimension of the WWW.

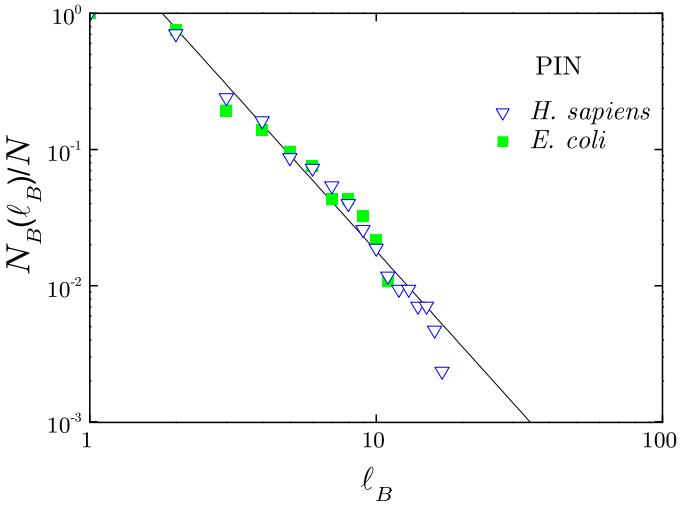


Figure 2.3: Log-log plot of the N_B vs ℓ_B revealing the fractal dimension of the two protein interaction networks: *H. sapiens* and *E. coli*

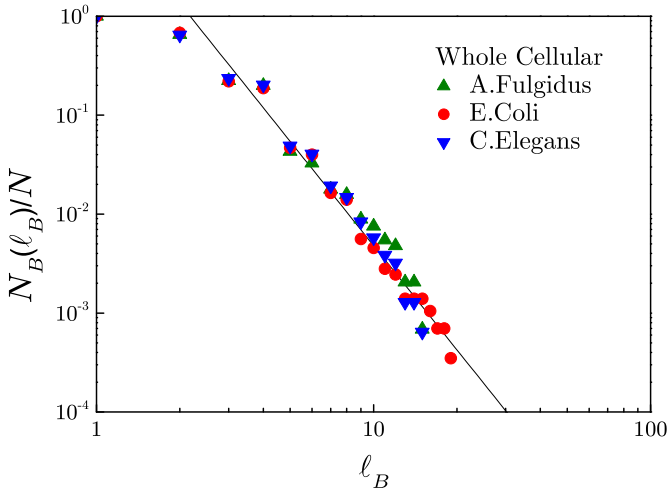


Figure 2.4: Log-log plot of the N_B vs ℓ_B revealing the fractal dimension of the the cellular networks of *A. fulgidus*, *E. coli* and *C. elegans* according to Eq. (2.1).

actors linked if they were cast together in at least one movie [19, 18], *(iii)* the biological networks of protein-protein interactions found in *E. coli* (429 proteins) and *H. sapiens* (human) (946 proteins) linked if there is a physical binding between them (database available via the Database of Interacting Proteins [20, 21], other PINs are discussed later on), and *(iv)* the cellular networks compiled by [22] using a graph-theoretical representation of the whole biochemical pathways based on the WIT [23] integrated-pathway genome database of 43 species from Archaea, Bacteria, and Eukarya. Here we show the results for *A. fulgidus, E. coli* and *C. elegans* [22], while the full database will be analyzed later on. It has been previously determined that the WWW and actors networks are small-world and scale-free, characterized by Eq. (1.1) with $\gamma = 2.6$ and 2.2, respectively [24]. For the PINs of *E. coli* and *H. sapiens* we find $\gamma = 2.2$ and 2.1, respectively. All cellular networks

are scale-free with average exponent $\gamma = 2.2$ [22]. We confirm these values and show the results for the WWW in Fig. 2.2.

Figures 2.2 and 2.3 show the results of $N_B(\ell_B)$ according to Eq. (2.1). They reveal the existence of self-similarity in the WWW, actors, and *E. coli* and *H. sapiens* protein-protein interaction networks with fractal dimensions $d_B = 4.1$, $d_B = 6.3$ and $d_B = 2.3$ and $d_B = 2.3$, respectively. The cellular networks are shown in Fig. 2.4 and have a fractal dimension $d_B = 3.5$.

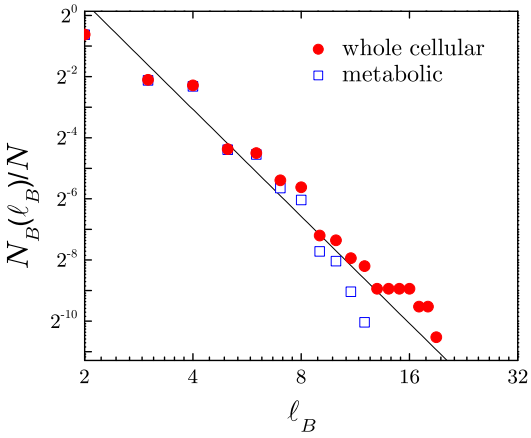
More detailed discussions about fractal dimension in real-world networks are given below.

2.2.1 Cellular networks

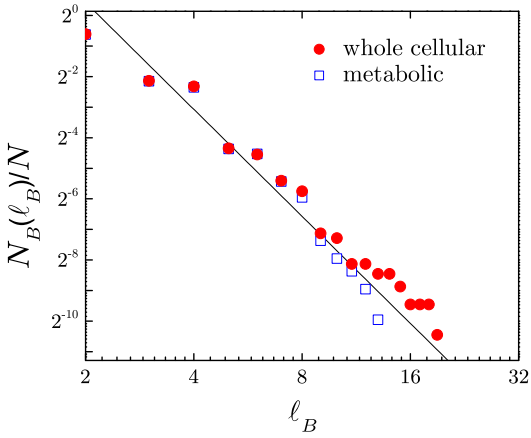
The WIT database [23, 22] of cellular networks considers the cellular functions divided according to bioengineering principles containing datasets for intermediate metabolism and bioenergetics (core metabolism), information pathways, electron transport, and transmembrane transport. The metabolic network is a subset of all reactions that take place in the cell. Since this is the largest part of the network we analyze it separately and compare it with the full biochemical reaction network. The data presented in Fig. 2.4 represents the full biochemical reaction networks of only three substrates. Here we present results of the 43 different substrates represented in the database for the metabolic and full networks. The following figures show the results of N_B vs ℓ_B. Both the metabolic and full networks display the power law relationship of self-similar networks with the same fractal dimension (within error bars) for all the organisms considered (the metabolic networks show a finite size effect due to their smaller size). We find an average fractal dimension

$d_B = 3.5$. The solid line in the figures represent the average fit.

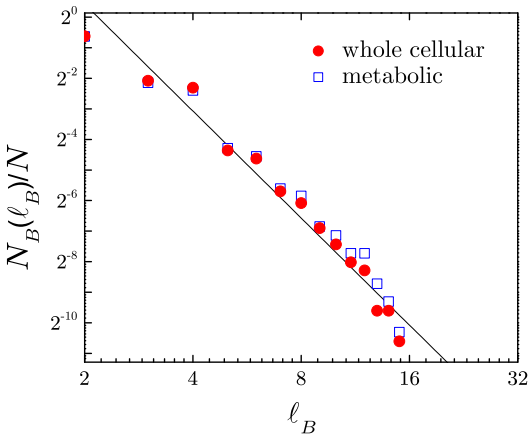
Aquifex aeolicus



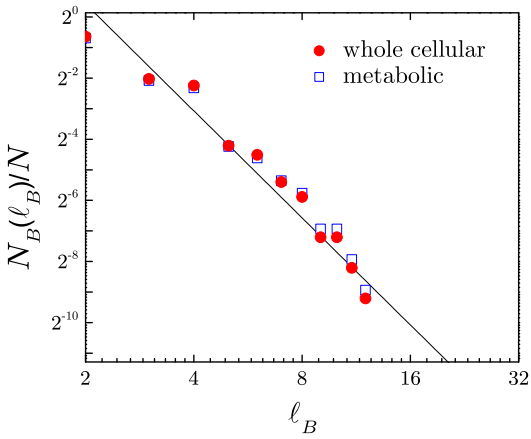
Actinobacillus actinomycetemcomitans



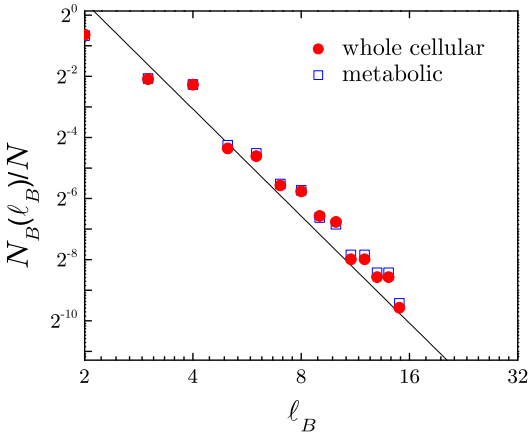
Archaeoglobus fulgidus



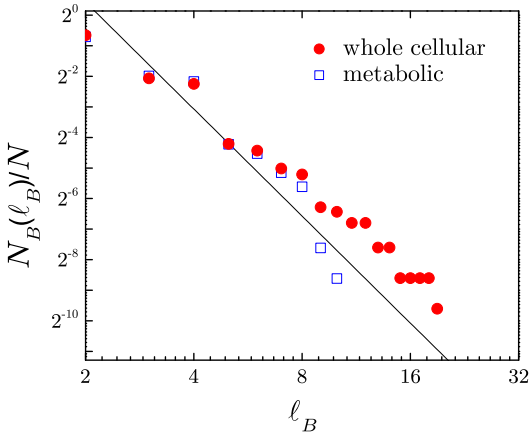
Aeropyrum pernix



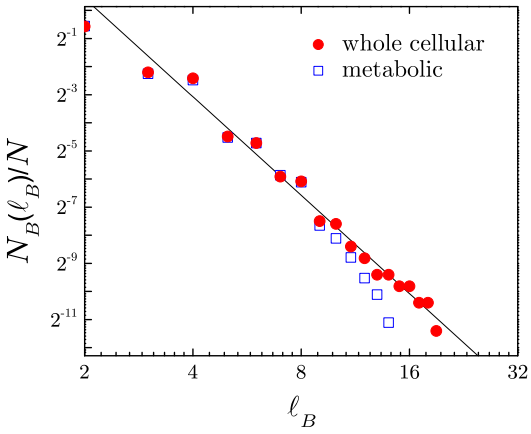
Arabidopsis thaliana



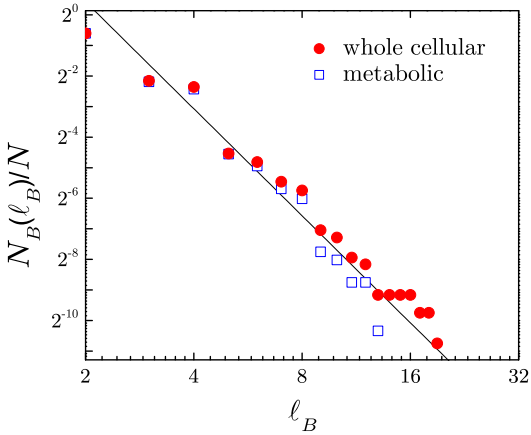
Borrelia burgdorferi



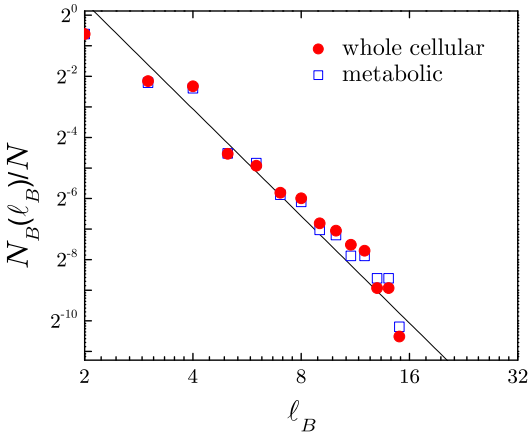
Bacillus subtilis



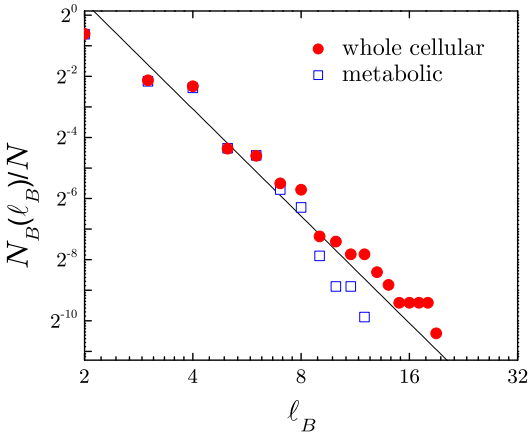
Clostridium acetobutylicum



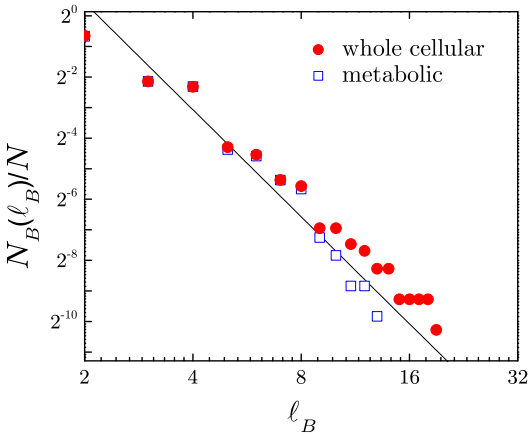
Caenorhabditis elegans



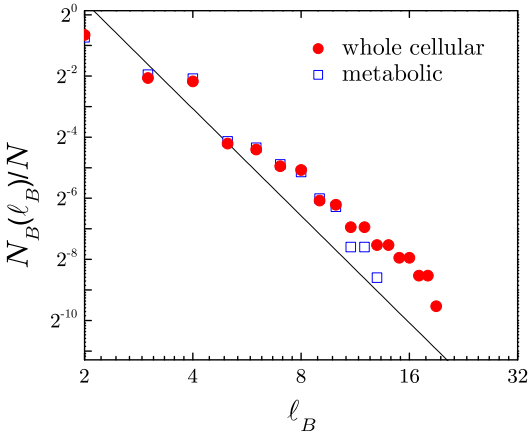
Campylobacter jejuni



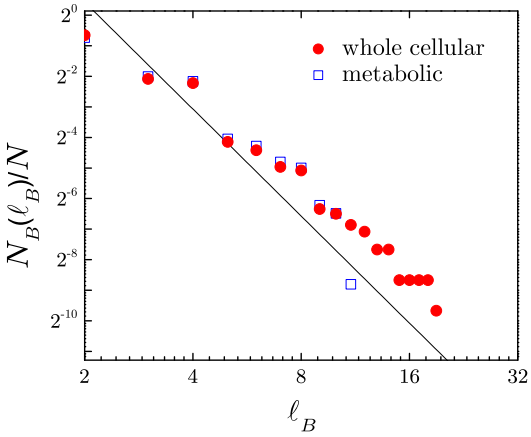
Chlorobium tepidum



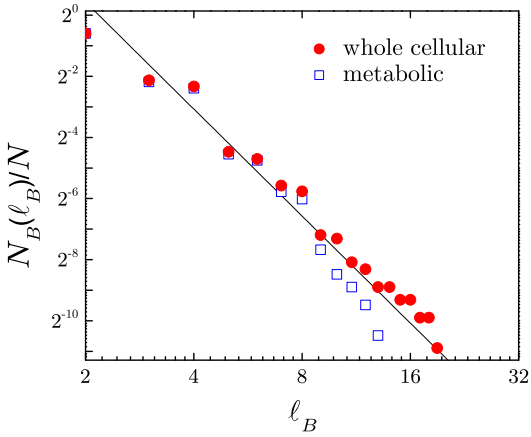
Chlamydia pneumoniae



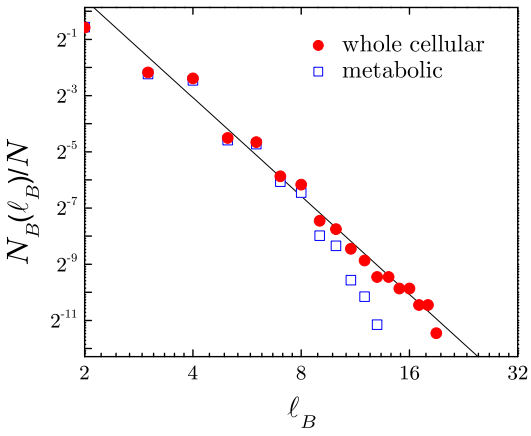
Chlamydia trachomatis



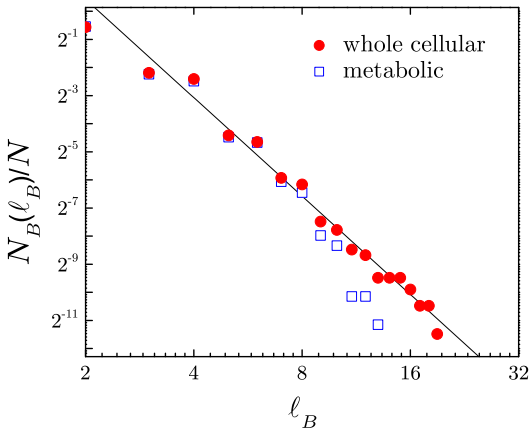
Synechocystis sp.



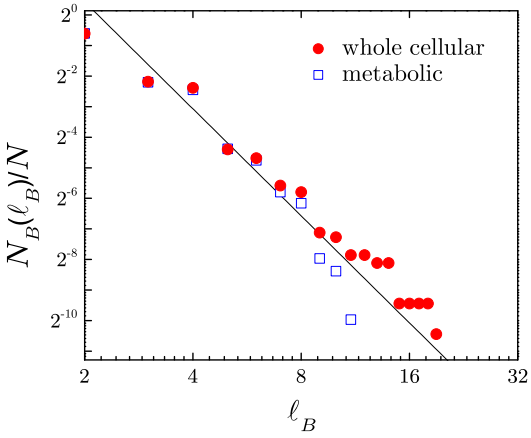
Deinococcus radiodurans



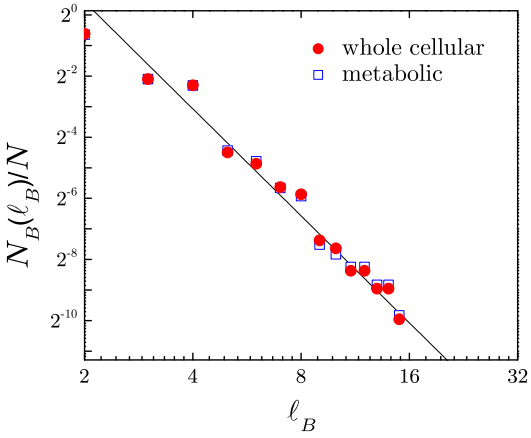
Escherichia coli



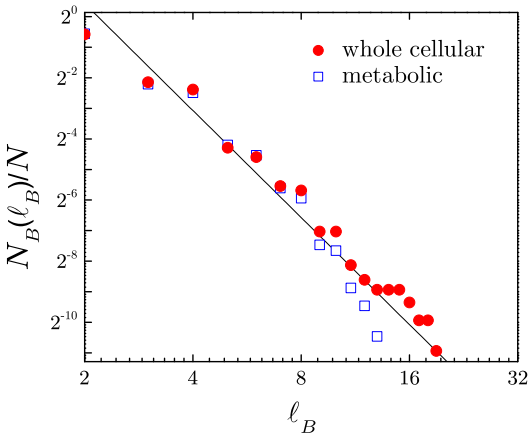
Enterococcus faecalis



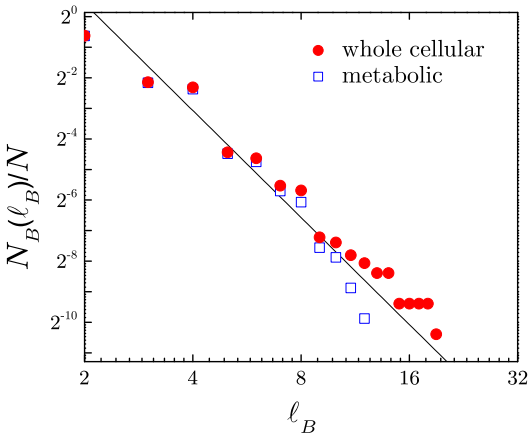
Emericella nidulans



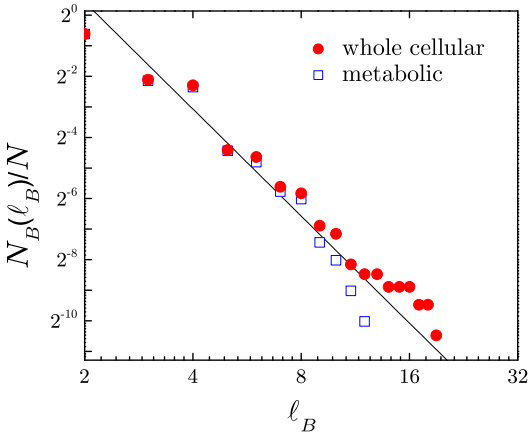
Haemophilus influenzae



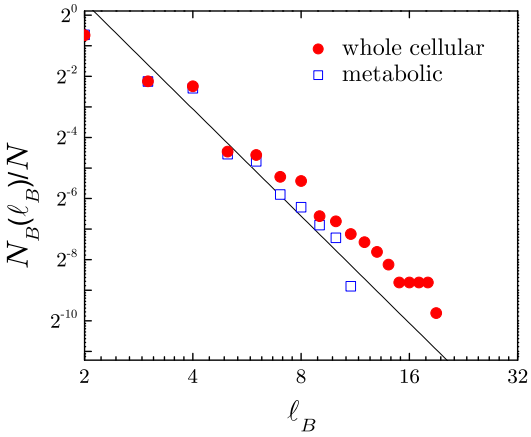
Helicobacter pylori



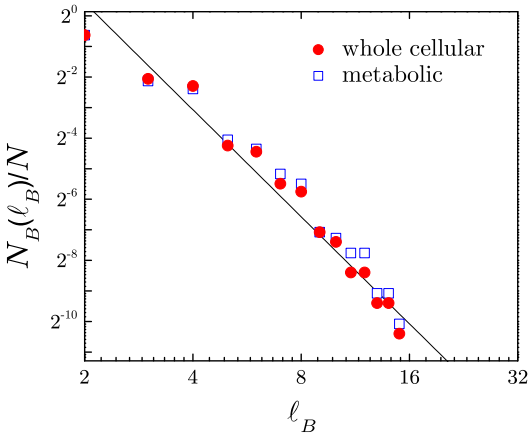
Mycobacterium bovis



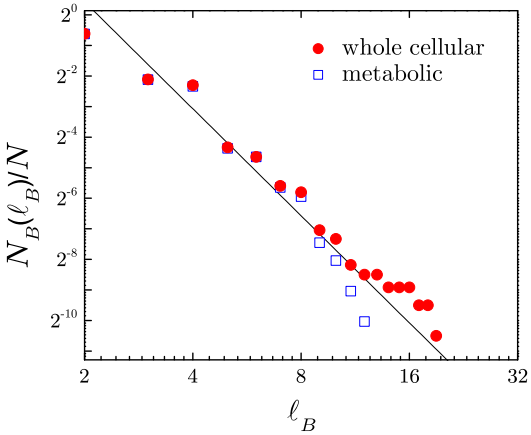
Mycoplasma genitalium



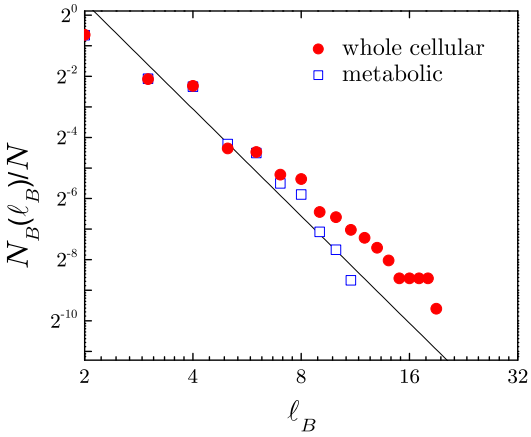
Methanococcus jannaschii



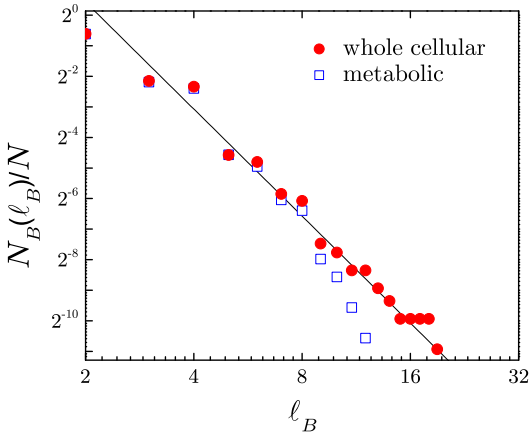
Mycobacterium leprae



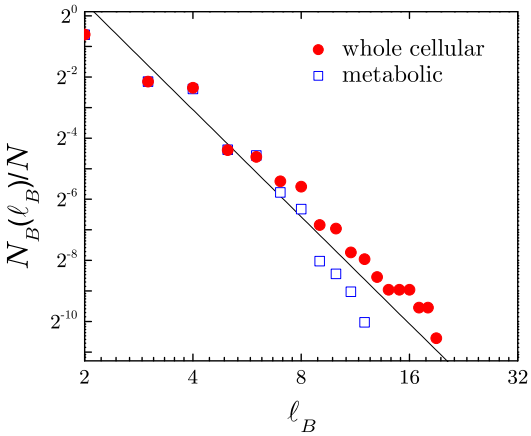
Mycoplasma pneumoniae



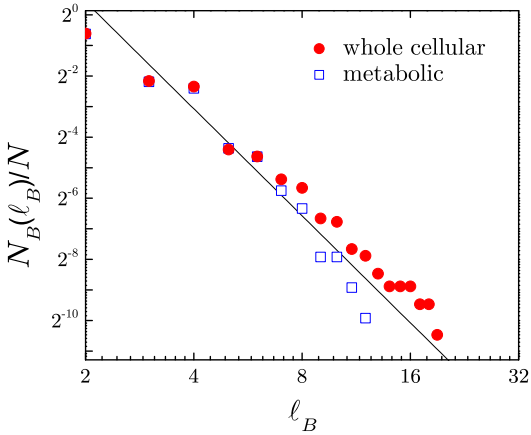
Mycobacterium tuberculosis



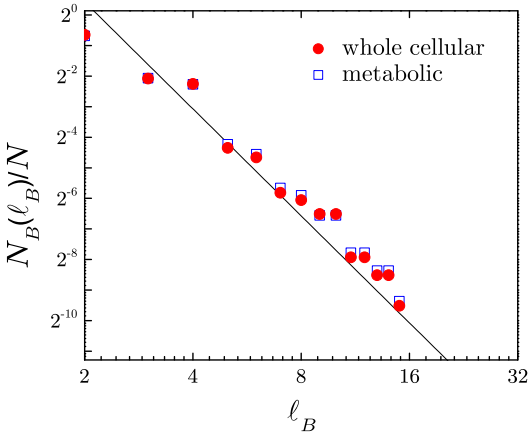
Neisseria gonorrhoeae



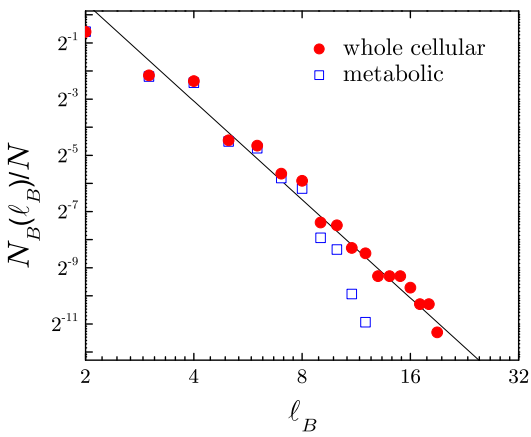
Neisseria meningitidis



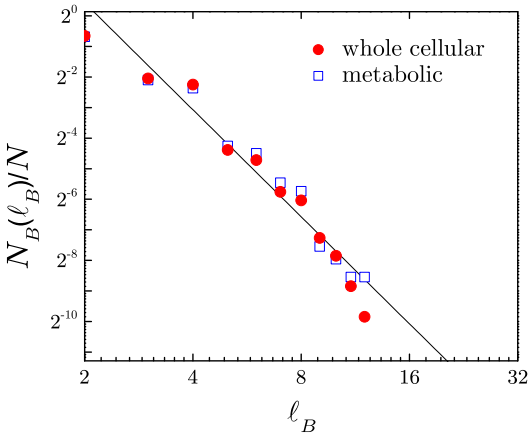
Oryza sativa



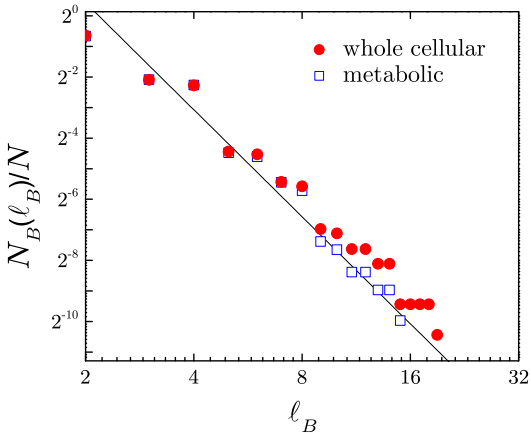
Pseudomonas aeruginosa



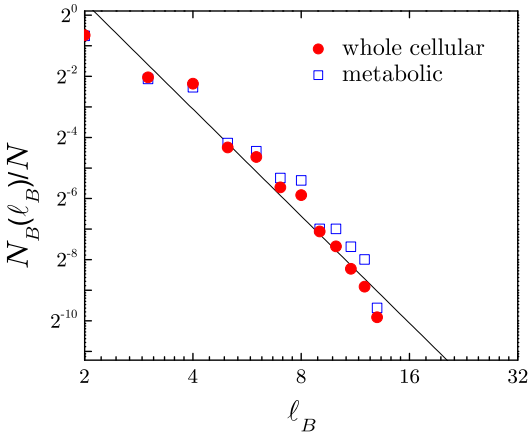
Pyrococcus furiosus



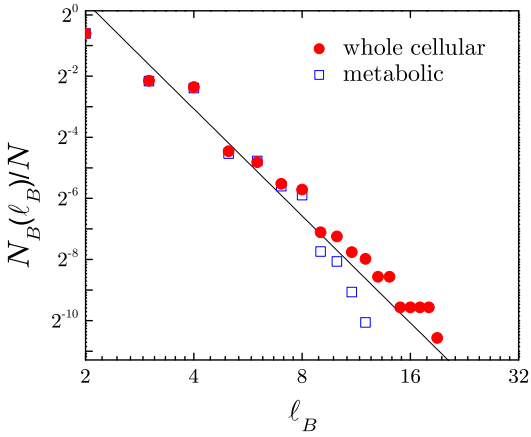
Porphyromonas gingivalis



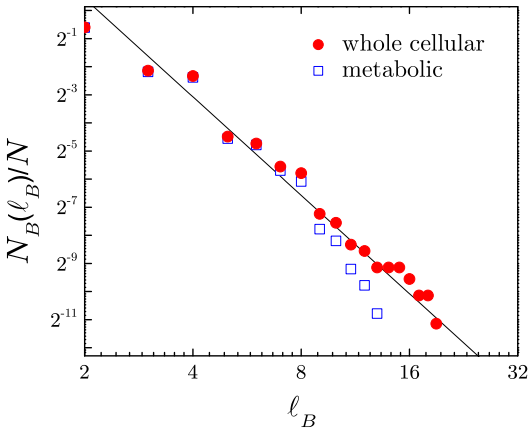
Pyrococcus horikoshii



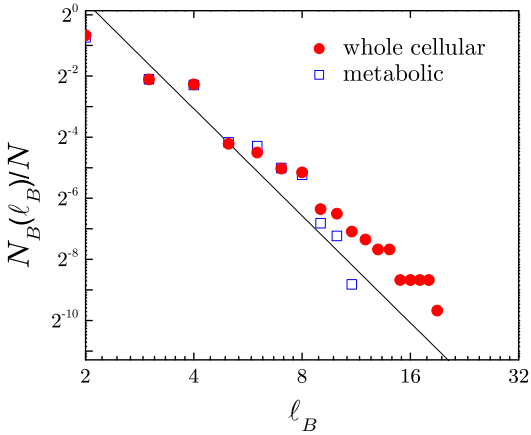
Streptococcus pneumoniae



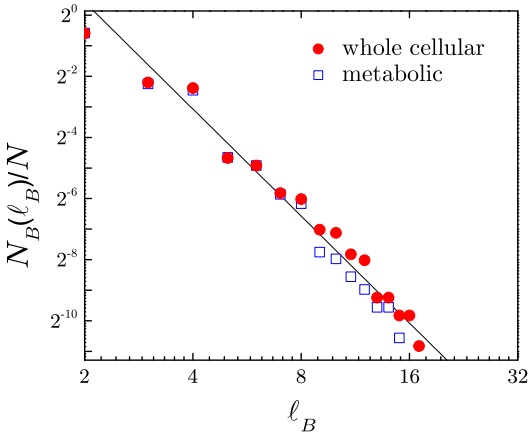
Rhodobacter capsulatus



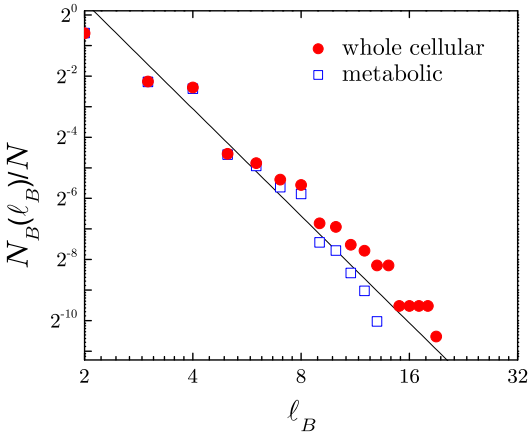
Rickettsia prowazekii



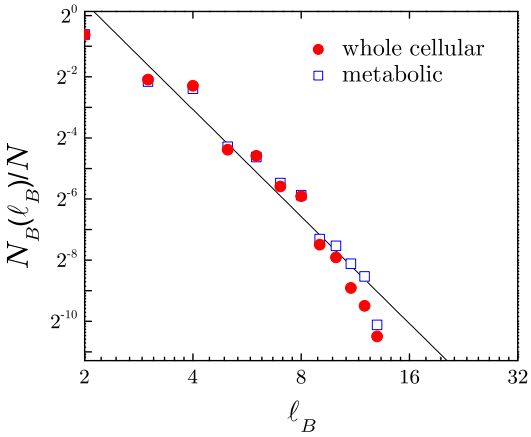
Saccharomyces cerevisiae



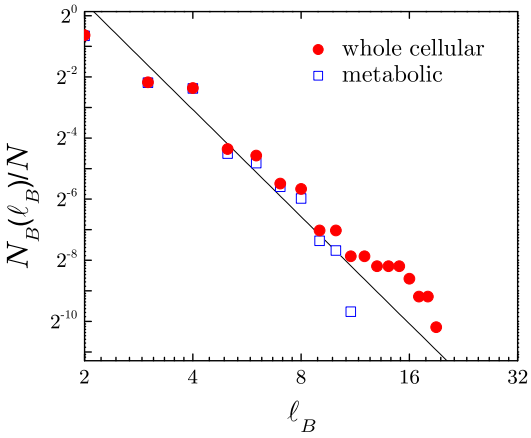
Streptococcus pyogenes



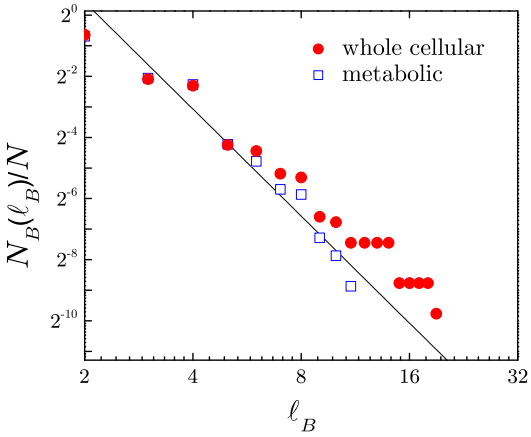
Methanobacterium thermoautotrophicum



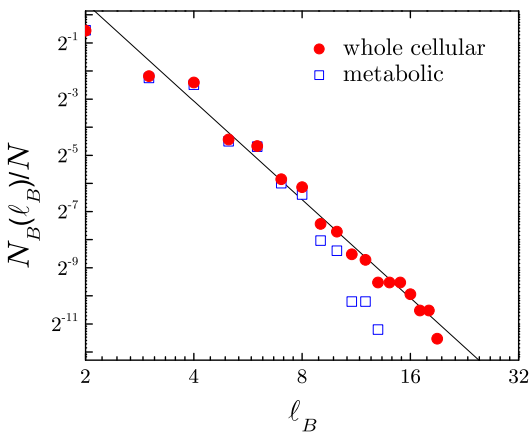
Thermotoga maritima



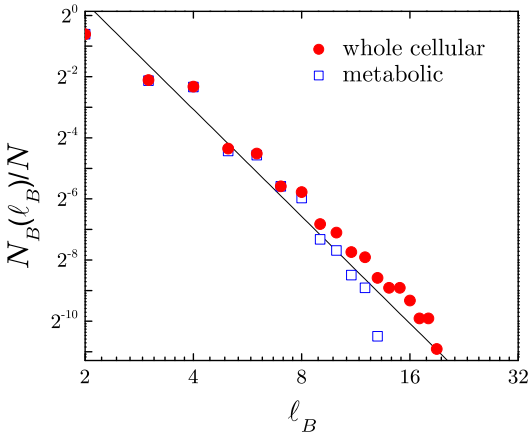
Treponema pallidum



Salmonella typhi



Yersinia pestis



2.2.2 Internet

It is interesting to note that not all complex networks show the clear self-similarity of the networks presented so far. We analyze the Internet composed of computers and routers linked by physical lines such as the database collected by the SCAN project (the "Mbone" [25], we also analyze the database of the Internet Mapping

Project [26] and found similar results). Figure 2.5 shows the result of $N_B(\ell_B)$. We fit the curve with a modified power-law

$$N_B(\ell_B) \sim (\ell_B + \ell_0)^{-d_B}, \tag{2.2}$$

with $\ell_0 = 14.9$ representing a cut-off and $d_B = 8.5$, suggesting a large fractal dimension. The decay of N_B with ℓ_B is faster than a power-law and slower than exponential as shown in the inset of Fig. 2.5.

Thus these networks lack the clear fractal structure found for the WWW, actors and the biological networks. However, we find that the distribution of $P(M_B)$ remains a power law and the degree distribution $P(k)$ is invariant under the renormalization suggesting that some self-similar properties might still be valid for the Internet. We notice that Internet maps are made by programs that use the IP protocol to trace the connections between each registered node in the Internet. These maps are incomplete since they map a few routers from each domain and also due to the existence of firewalls. Thus, the apparent lack of self-similarity might be due to incomplete information of the network.

2.2.3 Protein-protein interaction networks

We also analyze the protein interaction networks of the fruit fly *D. melanogaster* as given in [27], the bacterium *H. pylori* [28], the baker's yeast *S. cerevisiae* [29], and the nematode worm *C. elegans* [30], which are all available via the DIP database [21]. Figure 2.6 shows the results of N_B versus ℓ_B indicating that their behaviour is in between a pure power-law decay and a pure exponential. As with the Internet

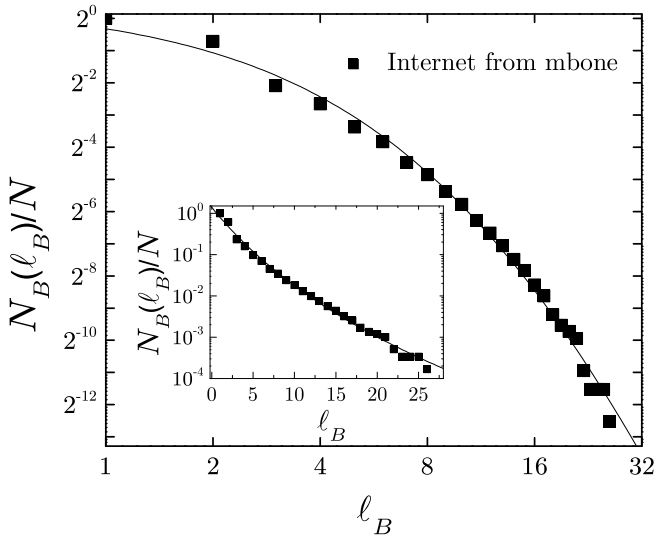


Figure 2.5: Internet: Log-log plot of $N_B(\ell_B)$. The solid line represents the modified power law fit, Eq. (2.2). The inset shows a linear-log plot indicating that the decay is slower than exponential.

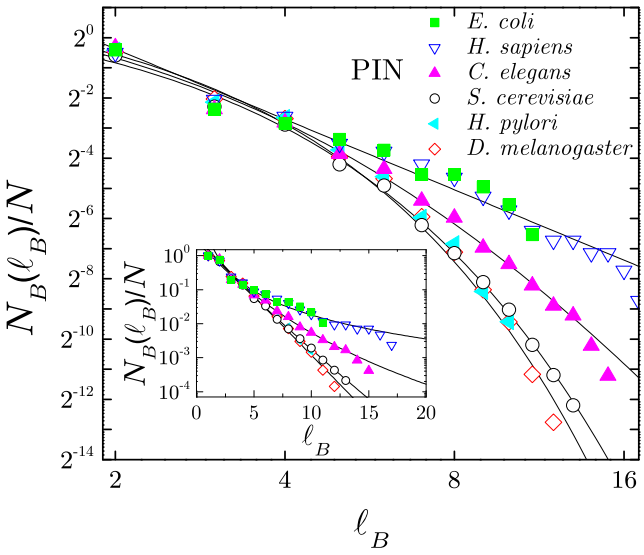


Figure 2.6: Log-log plot of N_B versus ℓ_B for different protein-protein interaction networks. While *E. coli* and *H. sapiens* show a clear power law behavior, the other protein networks show a modified power-law behaviour or a pure exponential decay. The inset shows a linear-log plot of $N_B(\ell_B)$.

data, we are able to fit the results with Eq. (2.2) with $\ell_c = 7.2$ and $d_B = 7.6$ for *C. elegans*. For *H. pylori* and *D. melanogaster* the fit is a pure exponential $N_B(\ell_B) \sim \exp(-\ell_B/\ell_e)$ with $\ell_e \approx 1$, while for *S. cerevisiae* the data could be fitted either by an exponential or by large values of ℓ_c and d_B (note that the exponential is the limit of Eq. (2.2) for $\ell_c \to \infty$, $d_B \to \infty$ and $\ell_c/d_B =$ constant). On the other hand, we observe that for small scales, N_B seems to display the same power law found for *E. coli* and *H. sapiens*. The lack of clear self-similarity in these networks might be due to the incompleteness of these databases which are continuously being updated with newly discovered physical interactions [20].

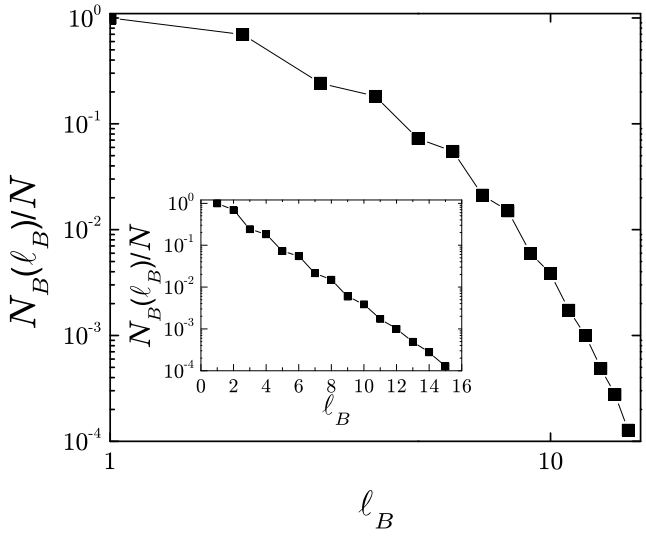


Figure 2.7: Log-log plot of N_B versus ℓ_B for random scale-free network with $\gamma = 2.35$, which shows a pure exponential decay. The inset shows a linear-log plot of $N_B(\ell_B)$.

2.2.4 Random scale-free network

Next we introduce an example of a model lacking self-similarity: the random scale-free model. This model consists of nodes to which a number of links are assigned with a power-law degree distribution and then connected randomly. Such a network shows a small world effect and a scale-free property but is not self-similar. We numerically find that the number of boxes decays exponentially with the box size (see Fig. 2.7). Moreover, while Eq. (2.7) is still valid in this case, the power law relation in Eq. (2.8) is replaced by an exponential law. We conjecture that the reason for this is a clustering of hubs; by assigning randomly the connections between the nodes, two nodes with a large number of links will have a large prob-

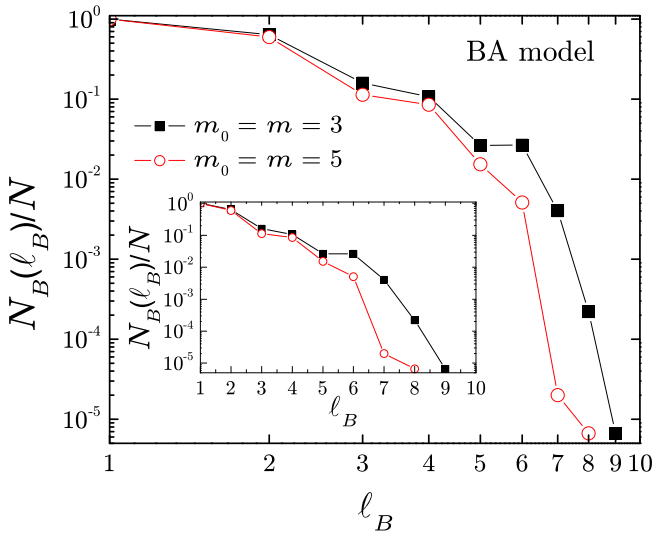


Figure 2.8: Barabási-Albert model of scale-free networks with preferential attachment for 150,000 nodes and $m = m_0 = 3$ and $m = m_0 = 5$. m_0 is the initial number of nodes in the system and m is the number of links of a newly created node in the dynamical growth of the network [19]. Log-log plot of N_B versus ℓ_B showing the lack of a power law behaviour. The inset shows a linear-log plot indicating that N_B decreases faster than exponential with ℓ_B.

ability to be connected. This induces spatial correlations in the values of k which may explain the breakdown of self-similarity. In contrast, the simple tree-structure proposed above does not cluster the hubs by construction.

2.2.5 The Barabási-Albert model and the Erdös-Rényi random graph

We also analyzed the Barabási-Albert model of complex networks [19] (which introduces the concepts of preferential attachment to describe the dynamics of scale-free networks). The results of $N_B(\ell_B)$ are shown in Fig. 2.8 for different parameters in

the model (see [19] for details) reveling that the structure is not self-similar; N_B seems to decrease faster than exponential with ℓ_B.

The random graph of Erdös-Rényi [10, 11] (in which nodes are connected at random with a given probability) lacks self-similarity as well, although in this case it is expected since this model is not scale-free but presents a Poissonian degree distribution.

2.3 Paradox between fractality and small-world

Instead of covering the network with boxes, a random seed node is chosen and nodes centered at the seed are grown so that they are separated by a maximum distance ℓ. The procedure is then repeated by choosing many seed nodes at random and the average "mass" of the resulting clusters, $\langle M_c \rangle$ (defined as the number of nodes in the cluster) is calculated as a function of ℓ to obtain the following scaling:

$$\langle M_c \rangle \sim \ell^{d_f}, \tag{2.3}$$

defining the fractal cluster dimension d_f [3]. If we use Eq. (2.3) for a small-world network, then Eq. (1.2) readily implies that $d_f = \infty$. In other words, these networks cannot be characterized by a finite fractal dimension, and should be regarded as infinite-dimensional objects. If this were true, though, local properties in a part of the network would not be able to represent the whole system. Still, it is also well established that the scale-free nature is similar in different parts of the network. Moreover, a graphical representation of real-world networks allows us to see that those systems seem to be built by attaching (following some rule) copies

of itself.

The answer lies in the inherent inhomogeneity of the network. In the classical case of a *homogeneous* system (such as a fractal percolation cluster) the degree distribution is very narrow and the two methods described above are fully equivalent, because of this local neighborhood invariance. Indeed, all boxes in the box-covering method are statistically similar with each other as well as with the boxes grown when using the cluster-growing technique, so that Eq. (2.3) can be derived from Eq. (2.1) and $d_B = d_f$.

The crux of the matter is to understand how one can calculate the fractal dimension in complex *inhomogeneous* networks with a *broad* degree distribution such as Eq. (1.1). Under these conditions Eqs. (2.1) and (2.3) are not equivalent as will be shown below. The application of the proper covering procedure in the box covering method, Eq. (2.1), for complex networks unveils a set of self-similar properties such as a finite fractal dimension and a new set of critical exponents for the scale-invariant topology.

We now elaborate on the apparent contradiction between the two definitions of the fractal dimension in complex networks. After performing a renormalization at a given ℓ_B, we calculate the mean mass of the boxes covering the network, $\langle M_B(\ell_B)\rangle$, to obtain

$$\langle M_B(\ell_B)\rangle \equiv N/N_B(\ell_B) \sim \ell_B^{d_B}, \tag{2.4}$$

which is corroborated by direct measurements for all the networks and shown in Fig. 2.9 for the WWW.

On the other hand, the average performed in the cluster growing method (for this calculation we average over single boxes without tiling the system) gives rise

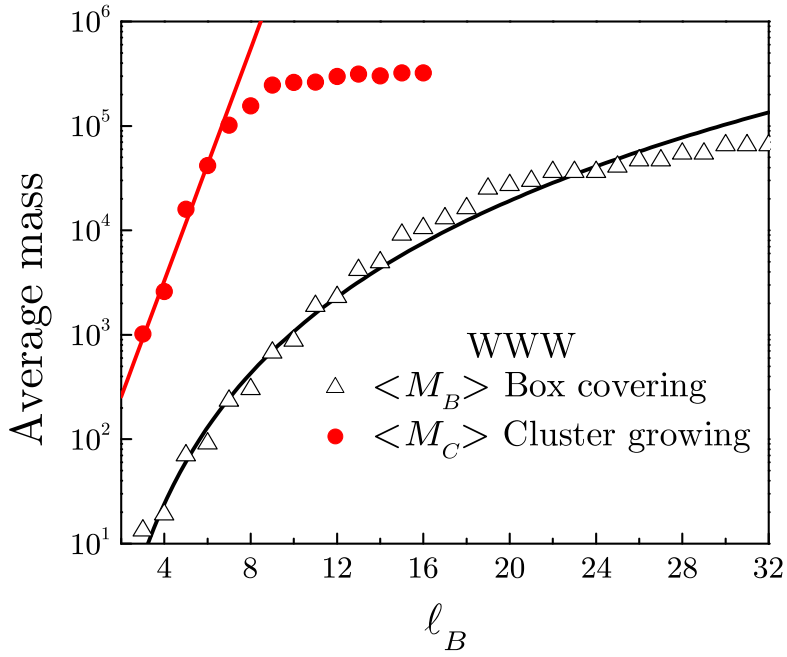


Figure 2.9: Mean value of the box mass in the box covering method, $\langle M_B \rangle$, and the cluster mass in the cluster growing method, $\langle M_c \rangle$, for the WWW. The solid lines represent the power-law fit for $\langle M_B \rangle$ and the exponential fit for $\langle M_c \rangle$ according to Eqs. (2.4) and (2.5), respectively.

to an exponential growth of the mass

$$\langle M_c(\ell_B)\rangle \sim e^{\ell_B/\ell_1}, \tag{2.5}$$

with $\ell_1 \approx 0.78$ in accordance with the small-world effect Eq. (1.2), as seen in Fig. 2.9.

These results also contrast with the case of Euclidean homogeneous networks (percolation), in which, for a given size of the box ℓ_B, the boxes needed to cover the fractal structure have on average the same number of sites or nodes and therefore both methods to calculate the fractal dimension are equivalent.

The topology of scale-free networks is dominated by several highly connected hubs— the nodes with the largest degree— implying that most of the nodes are connected to the hubs via one or very few steps. Therefore the average performed in the cluster growing method is biased; the hubs are overrepresented in Eq. (2.5) since almost every node is a neighbor of a hub. By choosing the seed of the clusters at random, there is a very large probability of including the hubs in the clusters. On the other hand the box covering method is a global tiling of the system providing a flat average over all the nodes, i.e. each part of the network is covered with an equal probability. Once a hub (or any node) is covered, it cannot be covered again. We conclude that Eqs. (2.1) and (2.3) are not equivalent for inhomogeneous networks with topologies dominated by hubs with a large degree.

The biased sampling of the randomly chosen nodes is clearly demonstrated in Fig. 2.10. We find that the probability distribution of the mass of the boxes for a given ℓ_B is very broad and can be approximated by a power-law: $P_{\ell_B}(M_B) \sim M_B^{-2.2}$

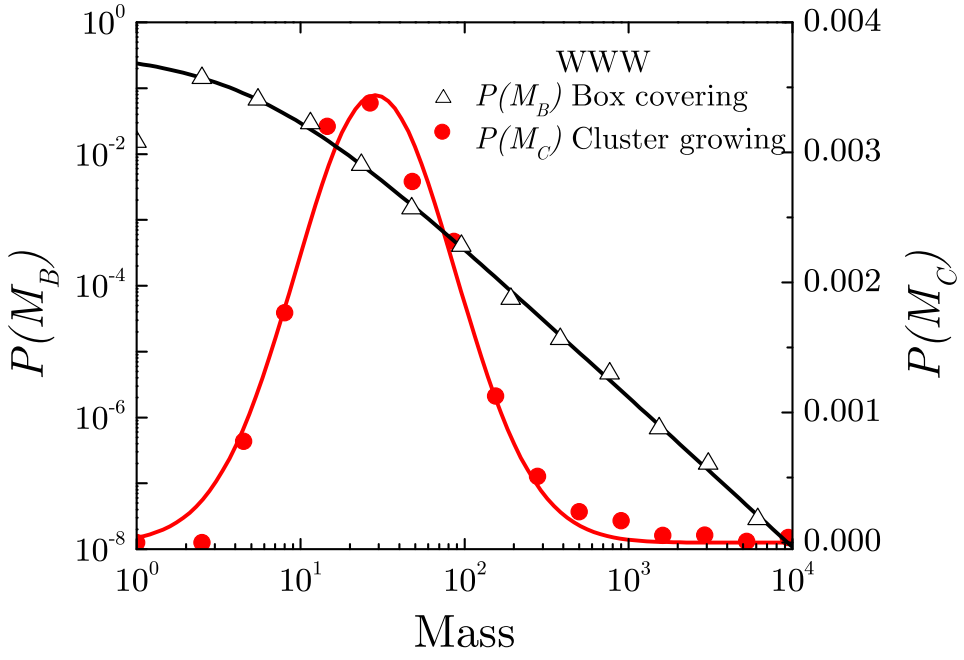


Figure 2.10: Probability distribution of M_B and M_c for $\ell_B = 4$ for the WWW. The curves are fitted by a power-law and a log-normal distribution, respectively.

in the case of WWW and $\ell_B = 4$. On the other hand, the probability distribution of M_c is very narrow and can be fitted by a log-normal distribution (see Fig. 2.10). In the box covering method there are many boxes with very large and very small masses in contrast to the peaked distribution in the cluster growing method, thus showing the biased nature of the latter method in inhomogeneous networks. This biased average leads to the exponential growth of the mass in Eq. (2.5) and it also explains why the average distance is logarithmic with N as in Eq. (1.2).

2.4 Renormalization

The box covering method provides a powerful tool for further investigations of the network properties as it enables a renormalization procedure, revealing that the fractal properties and the scale-free degree distribution persists irrespective of the amount of coarse-graining of the network.

Subsequent to the first step of assigning the nodes to the boxes we create a new renormalized network by replacing each box by a single node. Two boxes are then connected, provided that there was at least one link between their constituent nodes.

Figure 2.11 shows the same network as in Fig. 2.1 for the case $\ell_B = 2$. The second column of the panels shows this step in the renormalization procedure for the schematic network.

This procedure is applied to the WWW in Fig. 2.12. The main panel corresponds to the second stage in the renormalization of the web for $\ell_B = 3$. The procedure is applied again obtaining the remaining panels in Fig. 2.12 until the

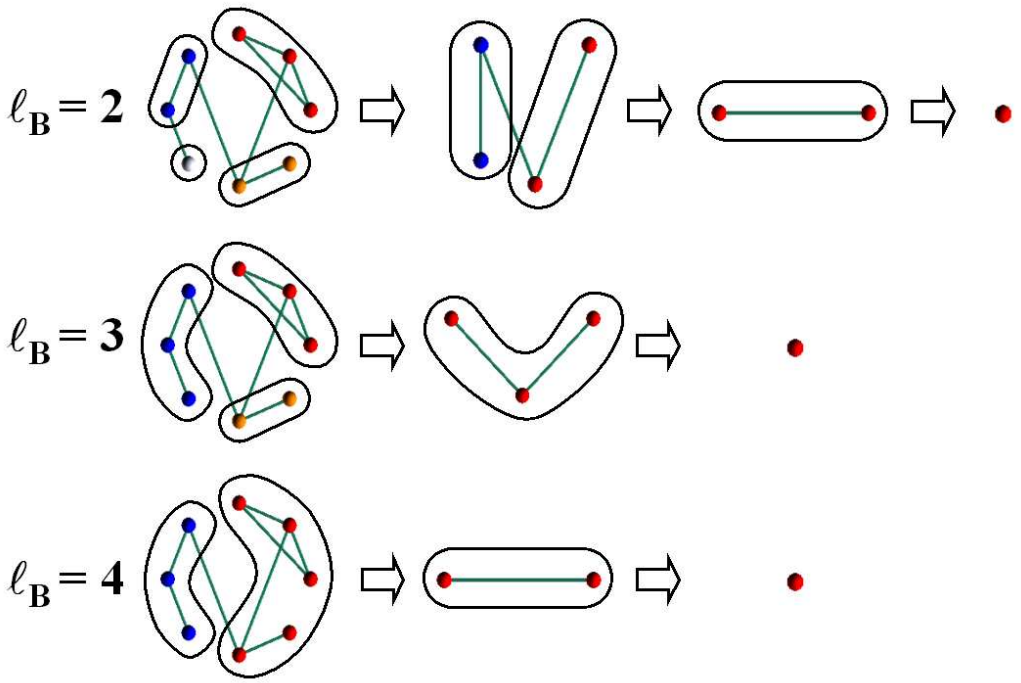


Figure 2.11: Demonstration of the renormalization method to complex networks for different ℓ_B in a network demo. The first column depicts the original network. We tile the system with boxes of size ℓ_B (different colors correspond to different boxes). All nodes in a box are connected by a minimum distance smaller than the given ℓ_B. For instance, in the case of $\ell_B = 2$, we identify four boxes which contain the nodes depicted with colors red, orange, white, and blue, each containing 3, 2, 1, and 2 nodes, respectively. Then we replace each box by a single node; two renormalized nodes are connected if there is at least one link between the unrenormalized boxes. Thus we obtain the network shown in the second column. The resulting number of boxes used to tile the network $N_B(\ell_B)$ versus ℓ_B gives the fractal dimension as in Eq. (2.1). The renormalization procedure is applied again and repeated until the network is reduced to a single node (third and fourth columns for different ℓ_B).

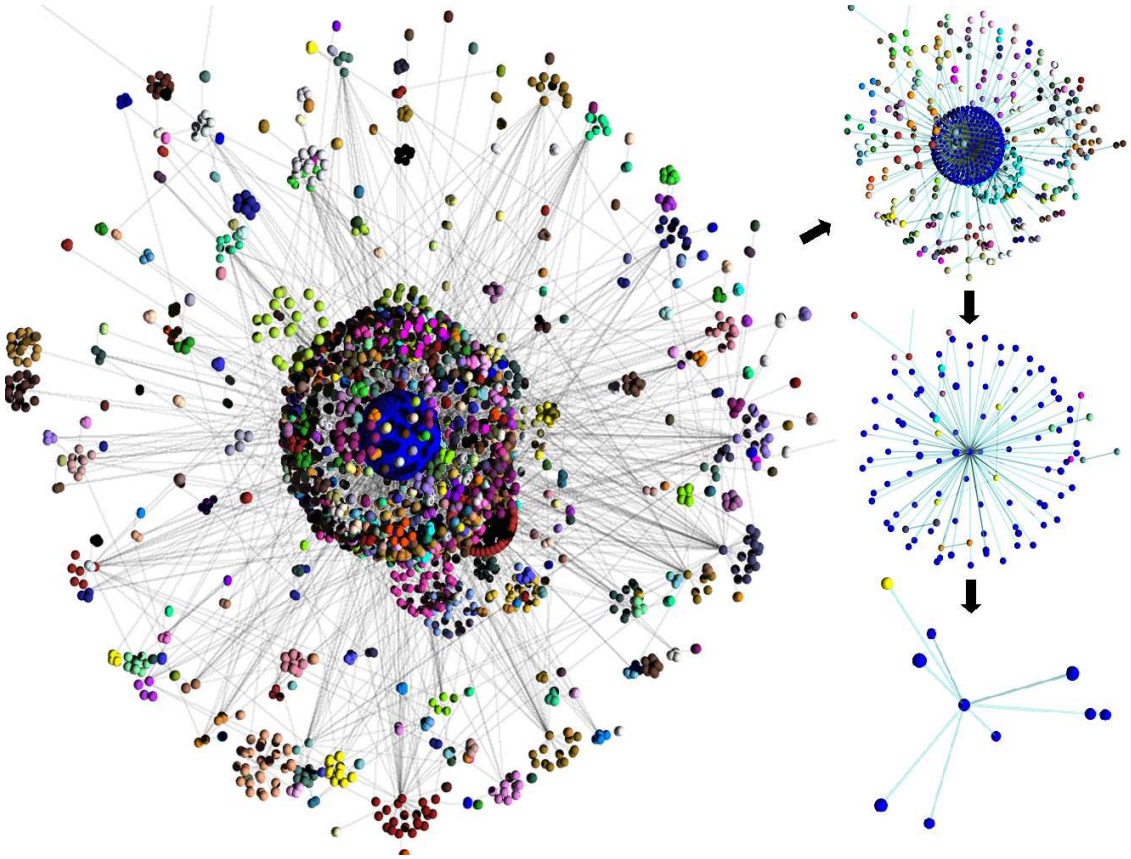


Figure 2.12: Three stages in the renormalization scheme applied to the entire WWW for $\ell_B = 3$. Here we color the nodes in the web according to the boxes to which they belong.

web is reduced to a single box in the last panel. The colors of the nodes corresponds to the boxes to which they belong.

The renormalized network gives rise to a new probability distribution of links, $P(k')$, which is invariant under the renormalization:

$$P(k) \to P(k') \sim (k')^{-\gamma}. \tag{2.6}$$

The validity of this scale transformation is supported for a wide range of real-world networks including both fractal networks (WWW, protein interaction network of *yeast* and metabolic network of E. *coli*) and non-fractal networks (Internet), by showing a data collapse of all distributions with the same γ according to (2.6) in Fig. 2.13, Fig. 2.14, Fig. 2.15 and Fig. 2.16, respectively. It is important to note that even though the Internet is not fractal, the degree distribution is still invariant under renormalization.

2.5 Degree dimension and scaling relationship

Further insight arises from relating the scale-invariant properties (2.1) to the scale-free degree distribution (1.1). The number of links k' of each node in the renormalized network versus the maximum number of links k in each box of the unrenormalized network exhibits a scaling law

$$k \to k' = s(\ell_B)k. \tag{2.7}$$

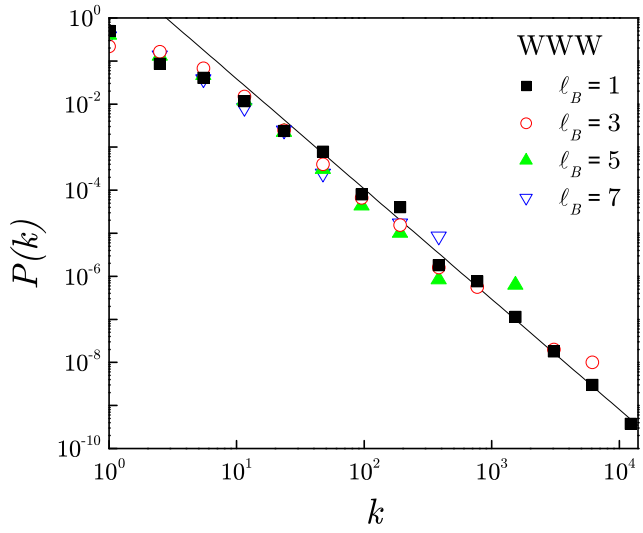


Figure 2.13: The degree distribution of WWW [8] according to different box size ℓ_B.

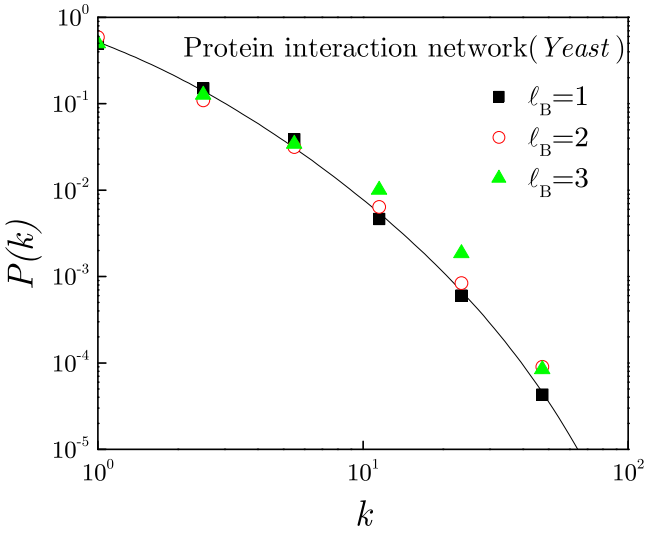


Figure 2.14: The degree distribution of Protein interaction network of *yeast*, according to different box size ℓ_B

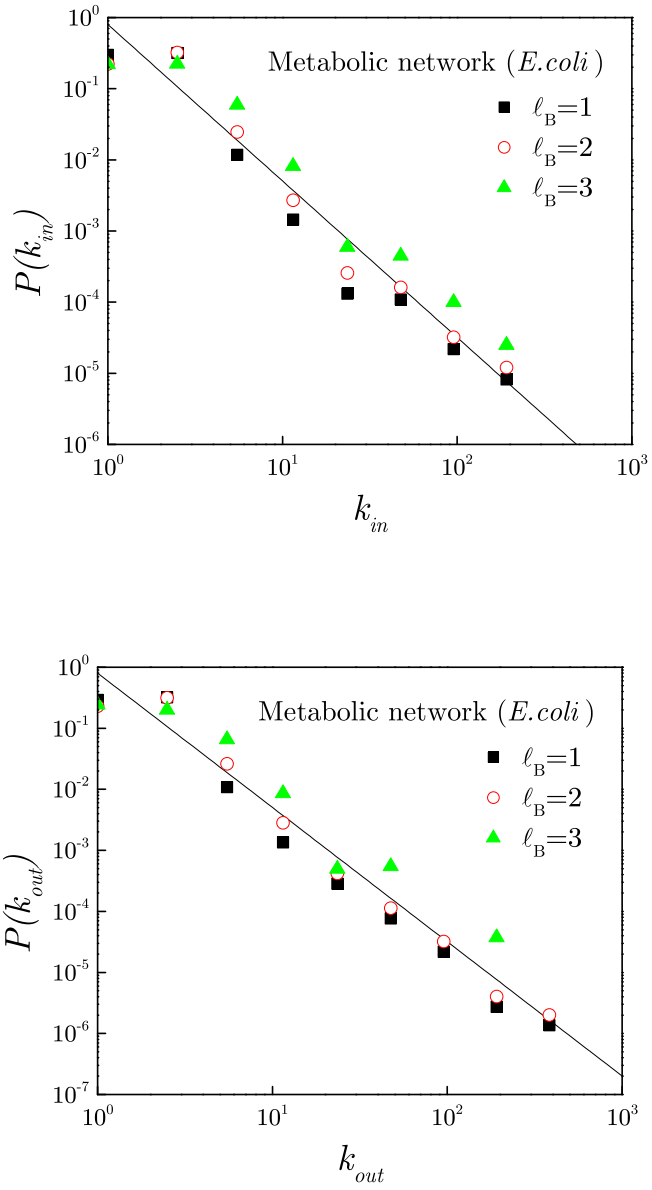


Figure 2.15: The degree distribution of the inward degree (a) and the outward degree (b) of metabolic network of E. *coli* , according to different box size ℓ_B.

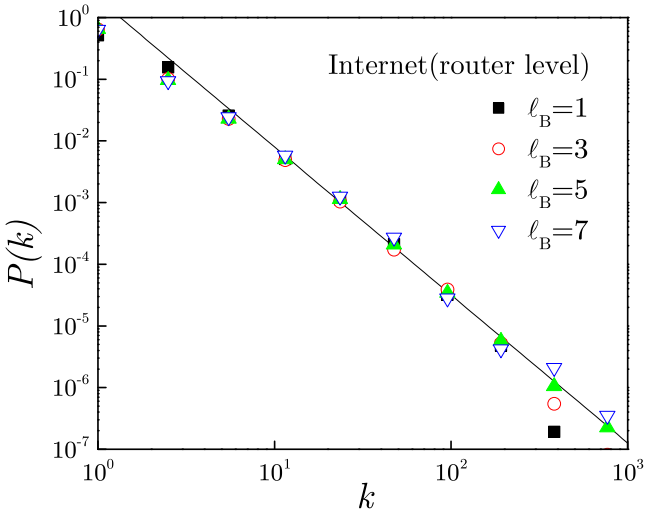


Figure 2.16: The degree distribution of Internet [26], according to different box size ℓ_B.

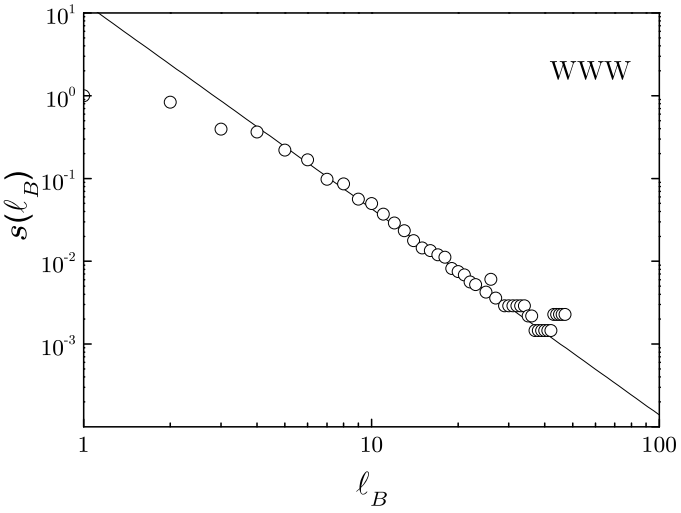


Figure 2.17: Log-log plot of the s vs ℓ_B revealing the degree dimension of the WWW.

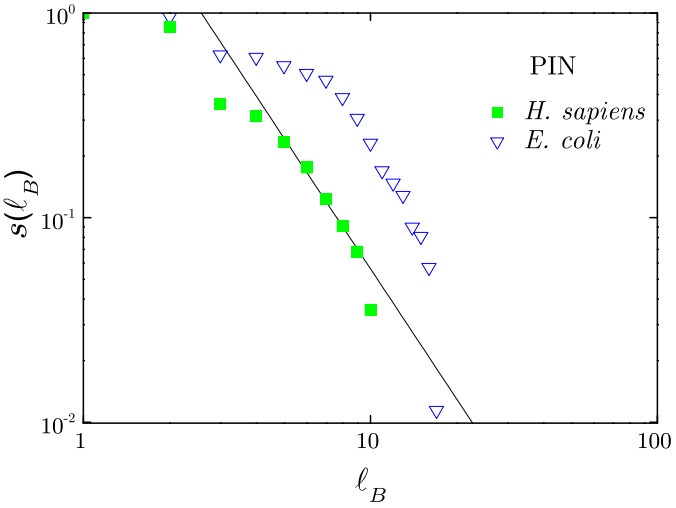


Figure 2.18: Log-log plot of the s vs ℓ_B revealing the degree dimension of the two protein interaction networks: *H. sapiens* and *E. coli*.

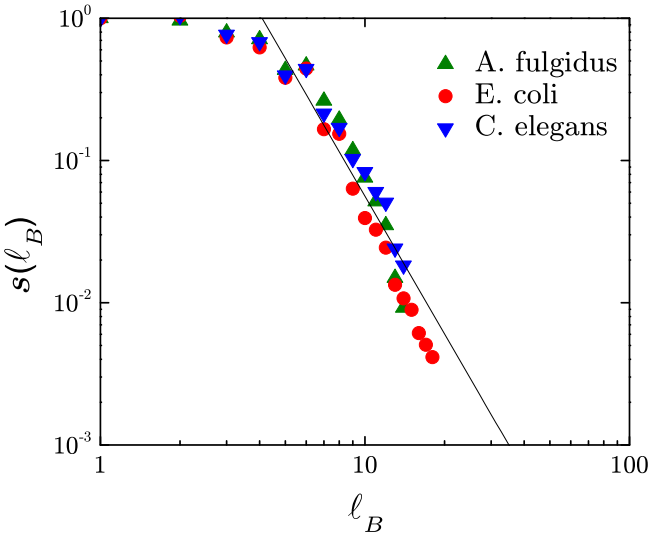


Figure 2.19: Log-log plot of the s vs ℓ_B revealing the degree dimension of the cellular networks of *A. fulgidus*, *E. coli* and *C. elegans* according to Eq. (2.1).

This equation defines the scaling transformation in the connectivity distribution. Empirically we find that the scaling factor $s(< 1)$ scales with ℓ_B with a new exponent d_k as

$$s(\ell_B) \sim \ell_B^{-d_k}, \tag{2.8}$$

shown in Fig. 2.17 for the WWW and actor networks (with $d_k = 2.5$ and $d_k = 5.3$, respectively), in Fig. 2.18 for the protein networks ($d_k = 2.1$ for *E. coli* and $d_k = 2.2$ for *H. sapiens*) and in Fig. 2.19 for the cellular networks with $d_k = 3.2$.

Equations (2.7) and (2.8) shed light on how families of hierarchical sizes are linked together. The larger the families, the fewer links exist. Surprisingly the same power-law relation exists for large and small families represented by Eq. (1.1).

Note that the degree dimension d_k reveal the degree invariance in complex networks, which didn't found in regular fractal before. Further insight arises from relating these two indexes to the exponent γ.

From Eq. (2.6) we obtain $n(k)dk = n'(k')dk'$, where $n(k) = NP(k)$ is the number of nodes with links k and $n'(k') = N'P(k')$ is the number of nodes with links k' after the renormalization (N' is the total number of nodes in the renormalized network). Using Eq. (2.7) we obtain $n(k) = s^{1-\gamma}n'(k)$. Then, upon renormalizing a network with N total nodes we obtain a smaller number of nodes N' according to $N' = s^{\gamma-1}N$. Since the total number of nodes in the renormalized network is the number of boxes needed to cover the unrenormalized network at any given ℓ_B, we have $N' = N_B(\ell_B)$. Then, from Eqs. (2.1) and (2.8) we obtain the relation between the three indexes

$$\gamma = 1 + d_B/d_k. \tag{2.9}$$

Network	d_B	d_k	$1 + d_B/d_k$ Eq. (2.9)	γ Eq. (1.1)
WWW	4.1	2.5	2.6	2.6
Actor	6.3	5.3	2.2	2.2
E. coli (PIN)	2.3	2.1	2.1	2.2
H. sapiens (PIN)	2.3	2.2	2.0	2.1
43 cellular networks	3.5	3.2	2.1	2.2
Scale-free tree	3.4	2.5	2.4	2.3

Table 2.1: Summary of the exponents obtained for the scale-invariant networks studied in the manuscript.

Equation (2.9) is confirmed for all the networks analyzed here in the table above. In all cases the scaling of these quantities gives rise to the same γ exponent as that obtained in the direct calculation of the degree distribution. The significance of this result is that the scale-free properties characterized by γ can be related to a more fundamental length-scale invariant property, characterized by the two new indexes d_B and d_k.

Chapter 3

Box Covering Algorithms

In last chapter, we applied a box covering technique which enabled us to demonstrate the existence of self-similarity in many real networks. Here we study and compare several possible box covering algorithms, by applying them to a number of model and real-world networks and we relate the box covering optimization to the vertex coloring algorithm. We also suggest a new definition for the box size ℓ_B, which seems to yield more accurate values for the fractal dimension d_B of a complex network.

3.1 The greedy coloring algorithm

We begin by recalling the original definition of box covering by Hausdorff [31, 32, 33]. For a given network G and box size ℓ_B, a box is a set of nodes where all distances ℓ_{ij} between any two nodes i and j in the box are smaller than ℓ_B. The minimum number of boxes required to cover the entire network G is denoted by N_B. For $\ell_B = 1$, N_B is obviously equal to the size of the network N, while $N_B = 1$

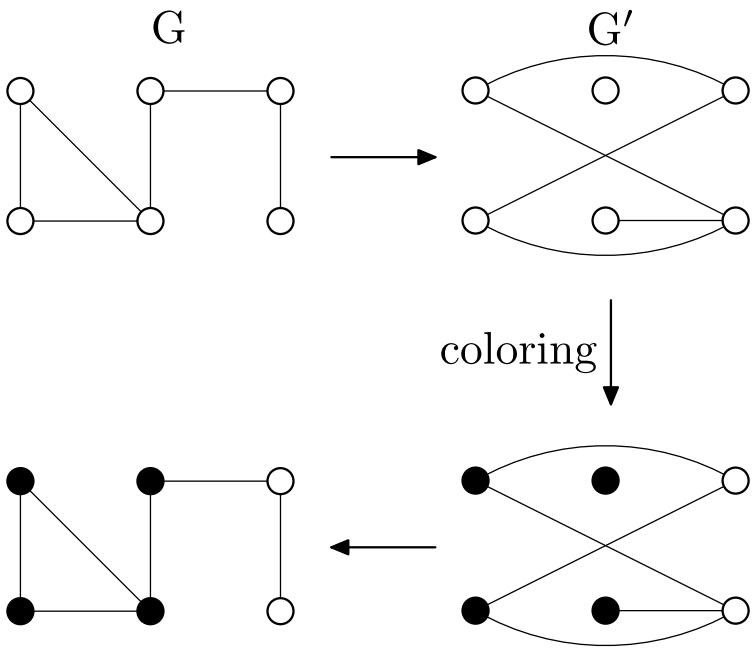


Figure 3.1: Illustration of the solution for the network covering problem via mapping to the graph coloring problem. Starting from G (upper left panel) we construct the dual network G' (upper right panel) for a given box size (here $\ell_B = 3$), where two nodes are connected if they are at a distance $\ell \geq \ell_B$. We use a greedy algorithm for vertex coloring in G', which is then used to determine the box covering in G, as shown in the plot.

for $\ell_B \geq \ell_B^{\max}$, where $\ell_B^{\max}$ is the diameter of the network (i.e. the maximum distance in the network) plus one.

The ultimate goal of all box-covering algorithms is to locate the optimum solution, i.e., to identify the minimum $N_B(\ell_B)$ value for any given box size ℓ_B. We first demonstrate that this problem can be mapped to the graph coloring problem, which is known to belong to the family of NP-hard problems [34]. This means that an algorithm that can provide an exact solution in a relatively short amount of time does not exist. This concept, though, enables us to treat the box covering problem using known optimization approximations. In order to find an approximation for

the optimal solution for an arbitrary value of ℓ_B we first construct a dual network G', in which two nodes are connected if the chemical distance between them in G (the original network) is greater or equal than ℓ_B. In Fig. 3.1 we demonstrate an example of a network G which yields such a dual network G' for $\ell_B = 3$ (upper row of the figure).

Vertex coloring is a well-known procedure, where labels (or colors) are assigned to each vertex of a network, so that no edge connects two identically colored vertices. It is clear that such a coloring in G' gives rise to a natural box covering in the original network G, in the sense that vertices of the same color will necessarily form a box since the distance between them must be less than ℓ_B. Accordingly, the minimum number of boxes $N_B(G)$ is equal to the minimum required number of colors (or the chromatic number) in the dual network G', $\chi(G')$, which is a famous problem in traditional graph theory.

In simpler terms, (a) if the distance between two nodes in G is greater than ℓ_B these two neighbors cannot belong in the same box. According to the construction of G', these two nodes will be connected in G' and thus they cannot have the same color. Since they have a different color they will not belong in the same box in G, which is our initial assumption. (b) On the contrary, if the distance between two nodes in G is less than ℓ_B it is possible that these nodes belong in the same box. In G' these two nodes will not be connected and it is allowed for these two nodes to carry the same color, i.e. they may belong to the same box in G, (whether these nodes will actually be connected depends on the exact implementation of the coloring algorithm, to be discussed later).

The exact solution for vertex coloring can only be achieved on small-size net-

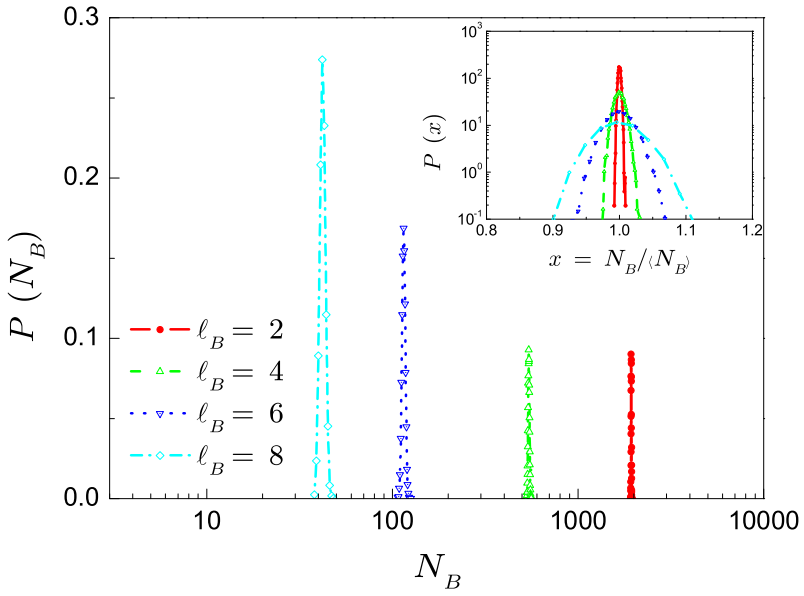


Figure 3.2: (color online) Probability distribution function $P(N_B)$ of the number of boxes N_B for the greedy algorithm, applied to the cellular network of *E.coli*. Different box sizes ℓ_B are used as indicated in the plot. Inset: PDFs of the normalized quantity $N_B/\langle N_B \rangle$ in a semi-log plot for the greedy algorithm, suggesting that $P(N_B)$ follows a Gaussian distribution.

works, since the optimal number of colors in an arbitrary graph is an NP-hard problem, as mentioned above, and in general should be solved by a brute-force approach [35, 36]. In practice, a greedy algorithm is widely adopted to obtain an approximate solution [37] and this also works very well for our case of box covering. We implement a simple version of the greedy algorithm as follows: 1) Rank the nodes in a sequence, 2) Mark each node with a free color, which is different from the colors of its nearest neighbors in G'. The algorithm that follows both constructs the dual network G' and assigns the proper node colors for all ℓ_B values in one pass. For this implementation we need a two-dimensional matrix $c_{i\ell}$ of size $N \times \ell_B^{\max}$, whose values represent the color of node i for a given box size $\ell = \ell_B$.

1. Assign a unique id from 1 to N to all network nodes, without assigning any colors yet.

2. For all ℓ_B values, assign a color value 0 to the node with id=1, i.e. $c_{1\ell} = 0$.

3. Set the id value $i = 2$. Repeat the following until $i = N$.

   (a) Calculate the distance ℓ_{ij} from i to all the nodes in the network with id j less than i.

   (b) Set $\ell_B = 1$

   (c) Select one of the unused colors $c_{j\ell_{ij}}$ from all nodes $j < i$ for which $\ell_{ij} \geq \ell_B$. This is the color $c_{i\ell_B}$ of node i for the given ℓ_B value.

   (d) Increase ℓ_B by one and repeat (c) until $\ell_B = \ell_B^{\max}$.

   (e) Increase i by 1.

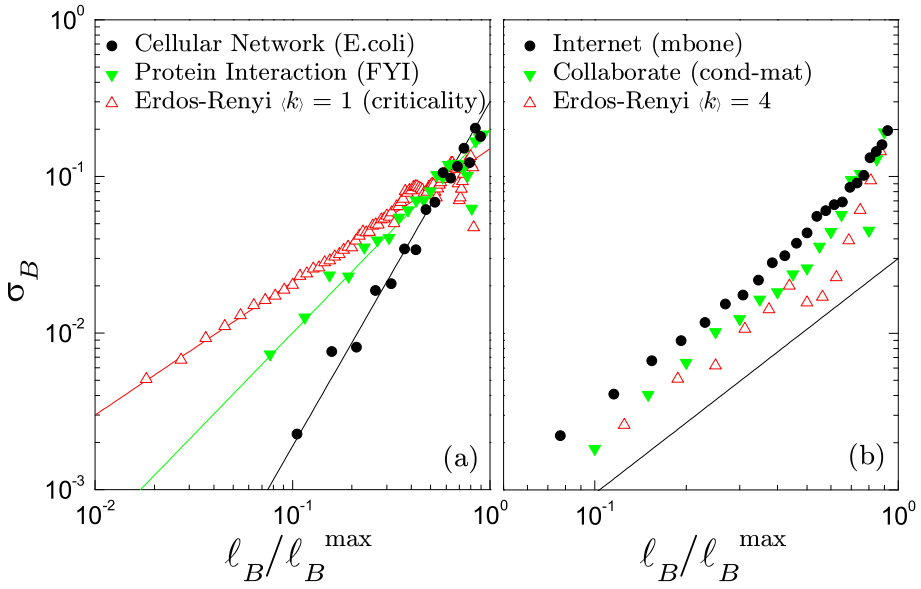


Figure 3.3: Normalized variance σ_B of the greedy algorithm for different box sizes ℓ_B, for (a) fractal and (b) non-fractal networks, where the box size ℓ_B is normalized by the maximum box size $\ell_B^{\max}$. The slopes for the fractal networks are (left to right): $\delta = 0.85$, 1.3, 2.2. For non-fractal networks: $\delta = 1.5$.

This greedy algorithm is very efficient, since we can cover the network with a sequence of box sizes ℓ_B performing only one network pass.

The results of the greedy algorithm may depend on the original coloring sequence. In order to investigate the quality of the algorithm, we randomly reshuffle the coloring sequence and apply the greedy algorithm for 10,000 times on several different models and real-world networks. In Fig. 3.2 we present a typical example for the PDFs of N_B for the cellular network of *E.coli*. The curves for all box sizes ℓ_B are narrow Gaussian distributions, indicating that almost any implementation of the algorithm yields a solution close to the optimal.

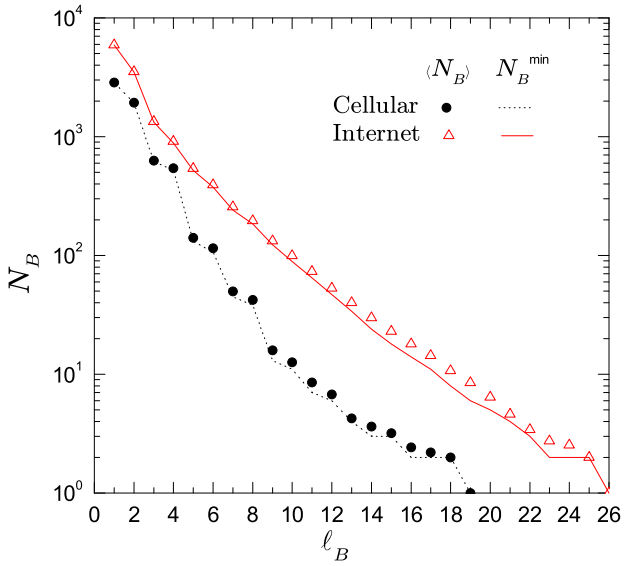


Figure 3.4: Comparison of the minimum $N_B^{\min}$ (line) and mean $\langle N_B \rangle$ (symbols) number of boxes for the greedy algorithm after 10,000 random reshuffles in real-world networks.

The uncertainty of the algorithm can be quantified via the normalized variances $\sigma_B \equiv \langle \Delta N_B^2 \rangle^{1/2} / \langle N_B \rangle$ of the PDFs. In Fig. 3.3, we present the σ_B dependence on the box size ℓ_B for both fractal (left panel) and non-fractal (right panel) networks. Surprisingly, when $\ell_B << \ell_{\max}$ all the networks seem to exhibit a power-law dependence

$$\sigma_B \sim \ell_B^{\delta} \,, \tag{3.1}$$

even for the case of non-fractal networks. In fractal networks the value of δ depends on the network structure, while for non-fractal networks δ seems to be constant with a value close to 1.5.

Strictly speaking, the calculation of the fractal dimension d_B through the rela-

tion $N_B \sim \ell_B^{-d_B}$ is valid only for the minimum possible value of N_B, for any given ℓ_B value, so an algorithm should aim to find this minimum N_B. For the greedy coloring algorithm it has been shown [37] that it can identify a coloring sequence which yields the optimal solution, i.e. the minimal value from the greedy algorithm coincides with the optimal value. Obviously, there is no rule as to when this minimum value has been actually reached. Yet, it is still meaningful to compare the mean value $\langle N_B \rangle$ with the minimum value N_B^{min} for our sample of 10,000 different realizations. We present such a comparison for the cellular network (fractal) and the Internet (non-fractal) in Fig. 3.4. For all ℓ_B values the difference between $\langle N_B \rangle$ and N_B^{min} is very small and the two values are almost indistiguishable from each other. This result is significant for implementation purposes, by pointing out that any realization of the above algorithm practically yields a quite accurate outcome.

The presented greedy algorithm is one of the simplest algorithms capable to solve the exact coloring problem. The coloring problem is very important in many fields, though, and consequently there is an enormous amount of studies on this subject. In principle, any one of the suggested algorithmic solutions in the literature can also be adopted for dealing with the box covering problem.

The form of the algorithm that was described above is the one that was used in Refs. [38, 39] for the calculation of N_B vs ℓ_B.

3.2 Burning algorithms

The presented greedy-coloring algorithm provides at the same time high efficiency and significant accuracy. A simpler approach, though, is to use more traditional

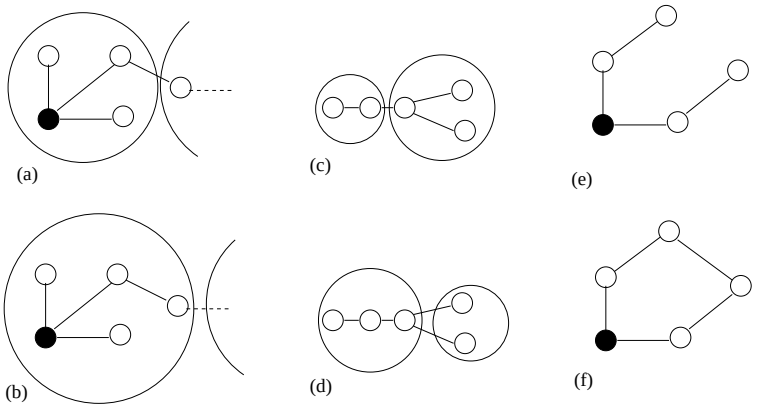


Figure 3.5: Our definitions for a box that is (a) non-compact for $\ell_B = 3$, i.e. could include more nodes, (b) compact, (c) connected, and (d) disconnected (the nodes in the right box are not connected in the box). (e) For this box, the values $\ell_B = 5$ and $r_B = 2$ verify the relation $\ell_B = 2r_B + 1$. (f) One of the pathological cases where this relation is not valid, since $\ell_B = 3$ and $r_B = 2$.

breadth-first algorithms. In the following sections we describe the basic simple burning algorithm and introduce two alternative (more sophisticated) methods based on similar ideas. We then proceed to compare these algorithms to the greedy-coloring algorithm.

In the following, we define a box to be 'compact' when it includes the maximum possible number of nodes, i.e. when there do not exist any other network nodes that could be included in this box. A 'connected' box means that any node in the box can be reached from any other node in this box, without having to leave this box. Equivalently, a 'disconnected' box denotes a box where certain nodes can be reached by other nodes in the box only by visiting nodes outside this box. For a demonstration of these definitions see Fig. 3.5.

A short note on the definition of the distances used. A box of size ℓ_B, according to our definition, includes nodes where the distance between any pair of nodes is

less than ℓ_B. It is possible, though, to grow a box from a given central node, so that all nodes in the box are within distance less than a given box radius r_B (the maximum distance from a central node). For the original definition of the box, ℓ_B corresponds to the box diameter (maximum distance between any two nodes in the box) plus one. Thus, these two measures are related through $\ell_B = 2r_B + 1$. In general this relation is valid for random configurations, but there may exist specific cases, such as e.g. nodes in a cycle, where this equation is not exact (Fig. 3.5).

3.2.1 Burning with the diameter ℓ_B, and the Compact-Box-Burning (CBB) algorithm

A traditional geometrical approach is the so-called 'burning' algorithm (breadth-first search). The basic idea is to generate a box by growing it from one randomly selected node towards its neighborhood until the box is compact, or equivalently that each box should include the maximum possible number of nodes. The algorithm is quite simple and can be summarized as follows:

1. Choose a random uncovered node as the seed for a new box.
2. All uncovered nodes connected to the current box are tested for being within distance ℓ_B from all the nodes currently in the box. Nodes that obey this criterion are included in the box.
3. Repeat (ii) until there are no more nodes that can be added in this box.
4. Repeat (i)-(iii) until all nodes are covered.

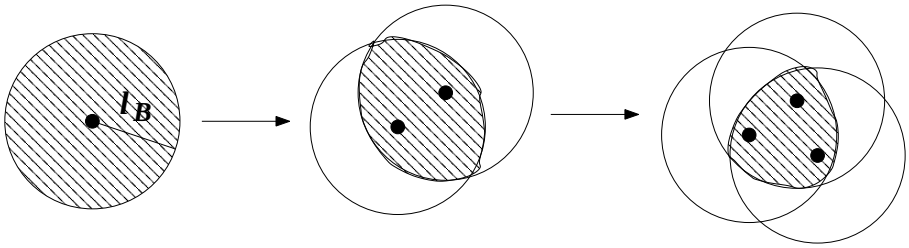


Figure 3.6: Two-dimensional geometrical analogue of the CBB algorithm. Initially we choose a random point and consider the circle with radius ℓ_B. We then choose another random point within this circle which serves as a new circle center and calculate the union of these two circles. We continue by iteratively selecting random centers for circles in the union of all the previous circles.

Although this algorithm is quite easy to implement, it requires a very long computational time. For this reason, we introduce a method that yields the exact same results as the above algorithm, but is computationally less intensive and can be executed much faster. We call this algorithm Compact-Box-Burning or CBB.

The method can be better understood in geometrical terms (Fig. 3.6). We start from a random point and draw a circle with radius ℓ_B. We then select a random point within this circle and draw a circle with radius ℓ_B using this new center. The union of the two circles includes all possible points that will eventually form the box. Iteratively adding points from the union of all previous circles and drawing new circles we eventually create a box where all the included points are within distance ℓ_B from each other. For the case of a complex network, we apply the following algorithm (see Fig. 3.7):

1. Construct the set C of all yet uncovered nodes.

2. Choose a random node p from the candidate set C and remove it from C.

3. Remove from C all nodes i whose distance from p is $\ell_{pi} \geq \ell_B$, since by

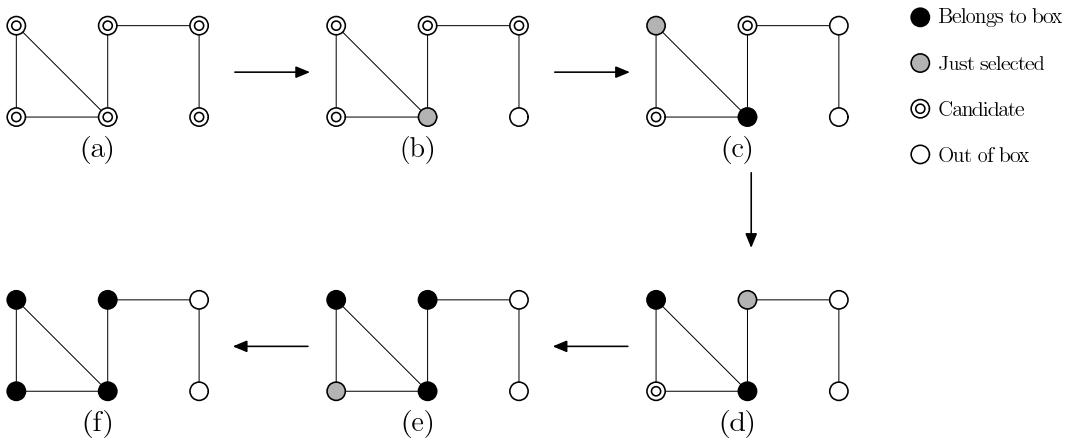


Figure 3.7: Illustration of the CBB algorithm for $\ell_B = 3$. (a) Initially, all nodes are candidates for the box. (b) A random node is chosen, and nodes at a distance further than ℓ_B from this node are no longer candidates. (c) The node chosen in (b) becomes part of the box and another candidate node is chosen. The above process is then repeated until the box is complete.

definition they will not belong in the same box.

4. Repeat steps (ii) and (iii) until the candidate set is empty.

The set of the chosen nodes p forms a compact box. We then repeat the above procedure until the entire network is covered.

We also performed 10,000 realizations for the CBB algorithm and calculated the mean value $\langle N_B \rangle$ and the normalized variance σ_B. In Fig. 3.8 we compare the greedy algorithm with CBB for both fractal and non-fractal networks. The value of $\langle N_B \rangle$ is roughly the same for both algorithms, with the value from CBB slightly larger (at most 2%) than the one from the greedy algorithm. More interestingly, the normalized variances are very close for these two algorithms. This suggests that CBB provides results comparable with the greedy algorithm, but CBB may be a bit simpler to implement.

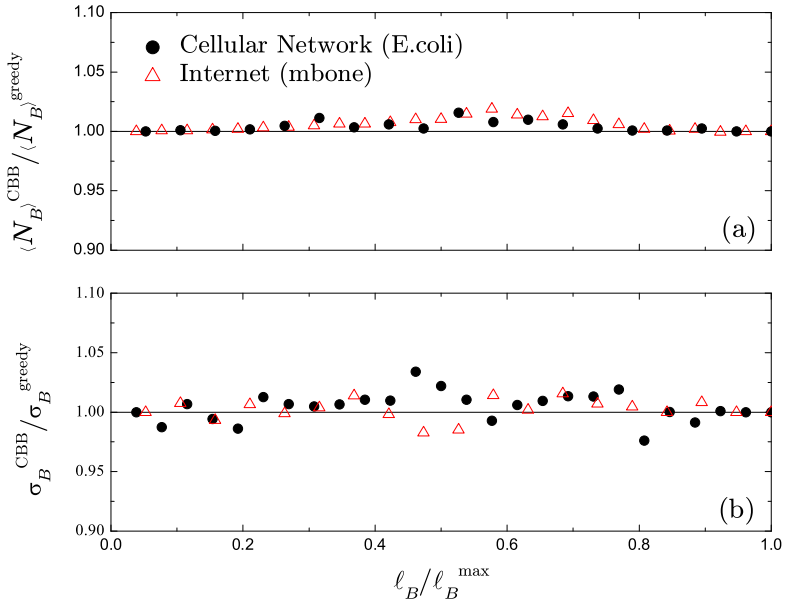


Figure 3.8: Comparison of (a) the mean number of boxes, $\langle N_B \rangle$, and (b) the normalized variance, σ_B, between the greedy algorithm and CBB.

3.2.2 Burning with the radius r_B, and the Maximum-Excluded-Mass-Burning (MEMB) algorithm

The formal definition of boxes includes the maximum separation ℓ_B between any two nodes in a box. However, it is possible to recover the same fractal properties of a network, where a box can be defined as nodes within a radius r_B from a central node. Using this box definition and random central nodes, this burning algorithm yields the optimal solution for non scale-free homogeneous networks, since the choice of the central node is not important. However, in inhomogeneous networks with wide-tailed degree distribution, such as the scale-free networks, this algorithm fails to achieve an optimal solution because of the hubs existence. For example, Fig. 3.9 demonstrates that burning with the radius from non-hubs is much worse than burning from hubs. In scale-free networks, when selecting a random node there is a high probability that this node will not be a hub, but a low-degree node instead, which leads the network tiling far from the optimal case. Additionally, a box burning originating from a non-hub node is not compact, in the sense that this box could contribute to a more efficient covering by incorporating more uncovered nodes without violating the maximum distance criterion. A variation of this algorithm for complex networks was presented in Ref. [40]. In general, this method cannot directly provide the optimum coverage, but it was shown that it finally yields the same fractal exponent d_B as the greedy coloring algorithm. Since the most important feature of similar studies is usually the calculation of the d_B exponent this algorithm can be very useful and, moreover, it is by far the easiest to implement.

To improve this completely random approach, we suggest an alternate strategy

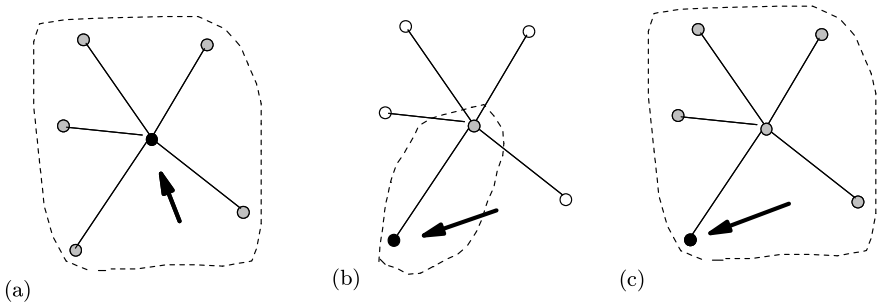


Figure 3.9: Burning with the radius r_B from (a) a hub node or (b) a non-hub node results in very different network coverage. In (a) we need just one box of $r_B = 1$ while in (b) 5 boxes are needed to cover the same network. This is an intrinsic problem when burning with the radius. (c) Burning with the maximum distance ℓ_B (in this case $\ell_B = 2r_B + 1 = 3$) we avoid this situation, since independently of the starting point we would still obtain $N_B = 1$.

that attempts to locate some optimal 'central' nodes which will act as the burning origins for the boxes. In principle one could use the hubs as box centers. However, depending on the nature of the network, choosing the hubs may not lead to the optimal solution because the hubs may be directly connected to each other or share a large number of common nodes, and this choice practically prohibits any low-degree node to be a box center which in some cases may be beneficial. Burning from the hubs represents a special case of the method that we will present, and it may emerge naturally from this algorithm if this is indeed the optimal way to cover the network. This is the case when hubs are not directly connected. In the following algorithm we use the basic idea of box optimization, where we require that each box should cover the maximum possible number of nodes. For a given burning radius r_B, we define the "excluded mass" of a node as the number of uncovered nodes within a chemical distance less than r_B. First, we calculate the excluded mass for all the uncovered nodes. Then we seek to cover the network

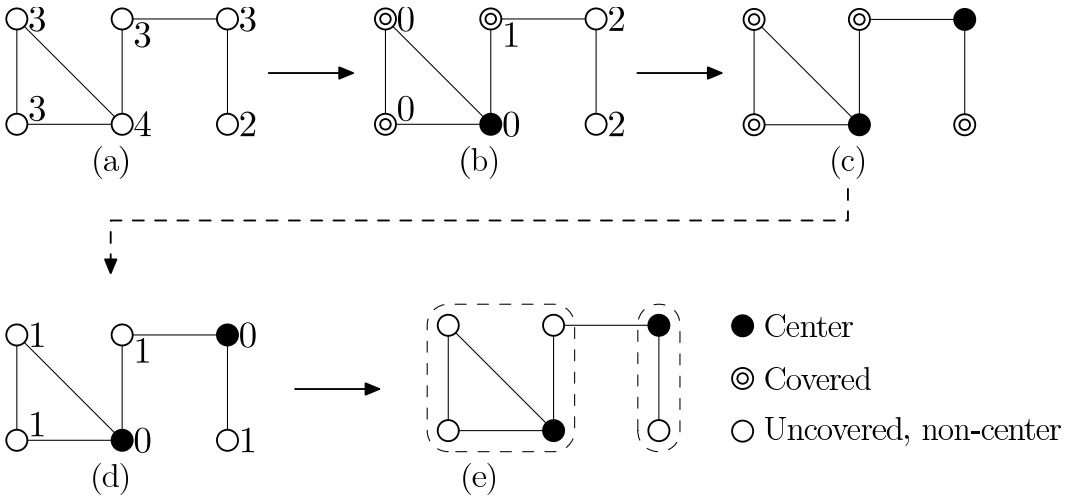


Figure 3.10: Illustration of the MEMB algorithm for $r_B = 1$. *Upper row: Calculation of the box centers* (a) We calculate the excluded mass for each node. (b) The node with maximum mass becomes a center and the excluded masses are recalculated. (c) A new center is chosen. Now, the entire network is covered with these two centers. *Bottom row: Calculation of the boxes* (d) Each box includes initially only the center. Starting from the centers we calculate the distance of each network node to the closest center. (e) We assign each node to its nearest box.

with boxes of maximum excluded mass. The details of this algorithm, which we call Maximum-Excluded-Mass-Burning or MEMB, are as follows (see Fig. 3.10):

1. Initially, all the nodes are marked as uncovered and non-centers.

2. For all non-center nodes (including the already covered nodes) calculate the excluded mass, and select the node p with the maximum excluded mass as the next center.

3. Mark all the nodes with chemical distance less than r_B from p as covered.

4. Repeat steps (ii) and (iii) until all nodes are either covered or centers.

Notice that the excluded mass has to be updated in each step because it is possible that it has been modified during this step. A box center can also be an already covered node, since it may lead to a largest box mass. After the above procedure, the number of selected centers coincides with the number of boxes N_B that completely cover the network. However, the non-center nodes have not yet been assigned to a given box. This is performed in the next step:

1. Give a unique box id to every center node.

2. For all nodes calculate the "central distance", which is the chemical distance to its nearest center. The central distance has to be less than r_B, and the center identification algorithm above guarantees that there will always exist such a center. Obviously, all center nodes have a central distance equal to 0.

3. Sort the non-center nodes in a list according to increasing central distance.

4. For each non-center node i, at least one of its neighbors has a central distance less than its own. Assign to i the same id with this neighbor. If there exist several such neighbors, randomly select an id from these neighbors. Remove i from the list.

5. Repeat step (iv) according to the sequence from the list in step (iii) for all non-center nodes.

For both the greedy coloring and the CBB algorithm the connectivity of boxes is not guaranteed. That is, for some boxes there may not exist a path inside the box that connects two nodes belonging in this box. The reason is that some boxes may already include certain nodes that are crucial for the optimization of other

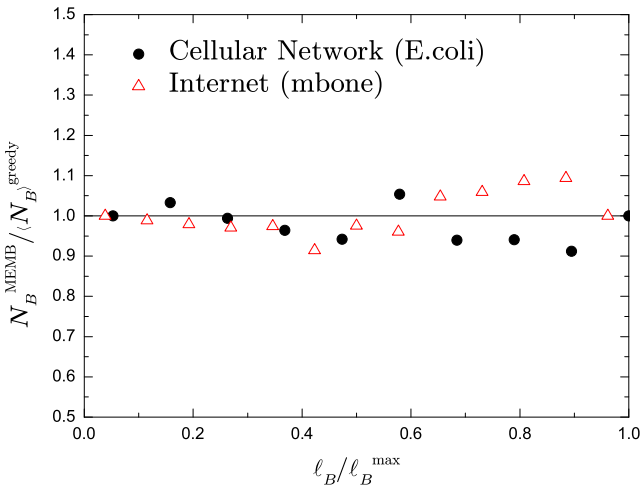


Figure 3.11: Comparison between the number of boxes obtained using MEMB, $N_B^{\rm MEMB}$, and the mean number of boxes, $\langle N_B\rangle^{\rm greedy}$, obtained from the greedy algorithm.

boxes. The MEMB algorithm, though, always yields connected boxes and this is the most appropriate method when this condition is required.

The MEMB algorithm is nearly deterministic, especially in the calculation of the N_B value. Randomness only enters in the order of choosing two nodes at equal distance from two centers. In order to directly compare the results with the greedy algorithm, we convert the radius r_B to the box-size ℓ_B, according to $\ell_B = 2r_B + 1$. Fig. 3.11 shows that the calculated number of boxes using MEMB, $N_B^{\rm MEMB}$, is also very similar to the mean value obtained from the greedy algorithm, $\langle N_B\rangle^{\rm greedy}$.

The MEMB algorithm was used in Figs. 2 and 3 of Ref. [39] for the calculation of hub-hub correlations, because in this case we want to isolate hubs in different boxes, a behavior similar to the model introduced in that paper (through the quantity $\mathcal{E}(\ell_{\mathcal{B}})$ defined in [39]). Also, we used this algorithm for studying the

evolution of conserved proteins in the yeast protein interaction network [41].

3.2.3 Comparison between the different algorithms

A comparison between the greedy coloring, the CBB and MEMB algorithms with the simple completely random burning with r_B (Fig. 3.12) shows that the three methods, except the random burning with r_B, are not sensitive to the specific realization used. This is manifested in the very narrow distributions of N_B and in the minimum value of the distribution which is very similar in all three cases (and very close to the average value, as well). On the contrary, when we use the random burning algorithm with r_B the corresponding distribution is significantly wider and the mean value $\langle N_B \rangle$ is much larger. Thus, a very large number of different realizations is required for achieving the optimal coverage in this case. Although the distributions in Fig. 3.12 correspond to a given value of ℓ_B (or equivalently r_B) the results are very similar for other ℓ_B values.

Despite these differences, the calculation of the fractal dimension d_B yields the same value for all the presented algorithms (Fig. 3.13), indicating that the scaling of the number of boxes is quite stable in all cases. Still, for the random burning it is not clear how many different realizations are needed in order for the average value to stabilize. Although from a practical point of view burning with r_B can still be used and give the correct dimension exponent d_B, it is not clear whether the properties of the boxes will be the same as in the optimal covering, e.g. whether applying renormalization to a network based on this covering will be similar to the renormalized network obtained from the optimal tiling.

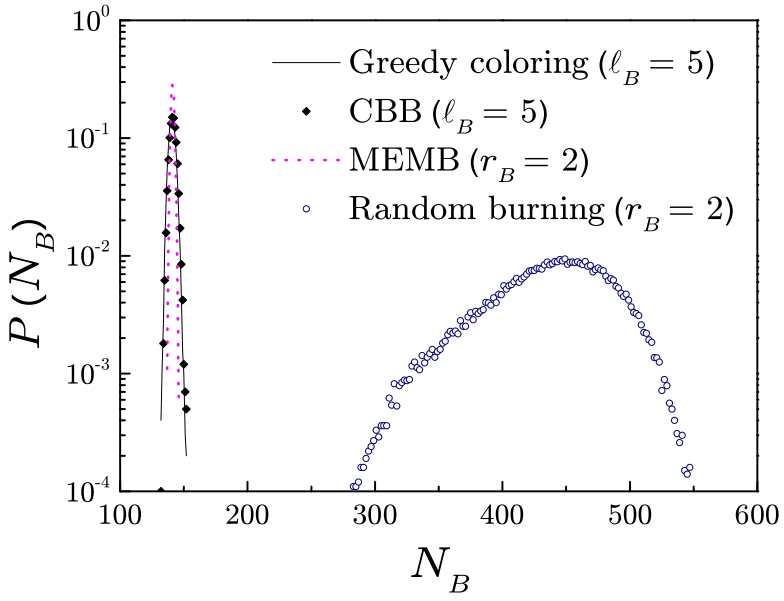


Figure 3.12: Comparison of the distribution of N_B for 10^4 realizations of the four network covering methods presented in this paper. Notice that three of these methods yield very similar results with narrow distributions and comparable minimum values, while the random burning algorithm fails to reach a value close to this minimum (and yields a broad distribution).

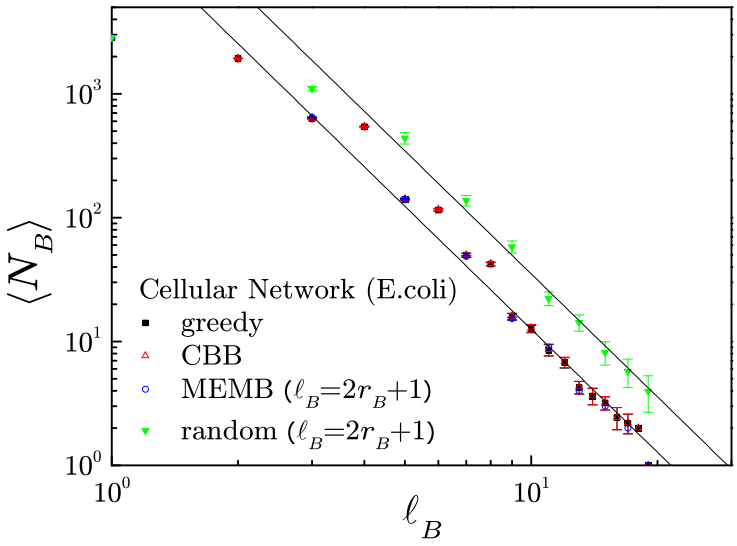


Figure 3.13: Comparison of the mean number of boxes $\langle N_B \rangle$ vs ℓ_B for the four presented algorithms. All methods yield the same value of the fractal dimension $d_B = 3.5$.

3.3 Box-size correction

In the usual box-covering techniques applied to regular fractals, as well as in all the methods described above, the box-size ℓ_B denotes the maximum possible distance within a box. Thus, it is always introduced as a cutoff value, rather than a direct measurement. Although in homogeneous systems, such as regular fractals, the difference may be indistinguishable, in many cases concerning inhomogeneous networks the actual size of boxes can be much smaller than this cutoff value ℓ_B. This difference is not expected to modify the asymptotic behavior of the scaling form $N_B \sim \ell_B^{-d_B}$. However, measurement of the fractal dimension d_B in real-world networks usually requires faster convergence, due to the small-world nature of many of them. Thus, we introduce an alternative definition for the box size ℓ_B^*. This parameter corresponds now to the actual box size (after we perform the network coverage in the usual way), and is defined as the maximum distance inside the particular box plus one, which is of course always smaller or equal to ℓ_B. The average box size ℓ_B^* over all boxes is used as a replacement of the previous cut-off size ℓ_B, and we replot the number of boxes $N_B(\ell_B^*)$ (whose maximum diameter is still ℓ_B) versus the average diameter ℓ_B^*. However, in order to obtain the correct box size and be consistent with the ℓ_B^* definition, the boxes have to be connected. Thus, we measure $N_B(\ell_B^*)$ via the MEMB algorithm, as described above.

We test the improvement of this modification by applying the measurement of ℓ_B^* to a couple of known examples. The fractal dimension of Erdos-Renyi networks at criticality ($\langle k \rangle = 1$) is known to be $d_B = 2$ (see e.g. [42]). In Fig. 3.14a we compare the numerical results before and after the size correction in such a network. The measurement of fractality after the correction seems to converge faster

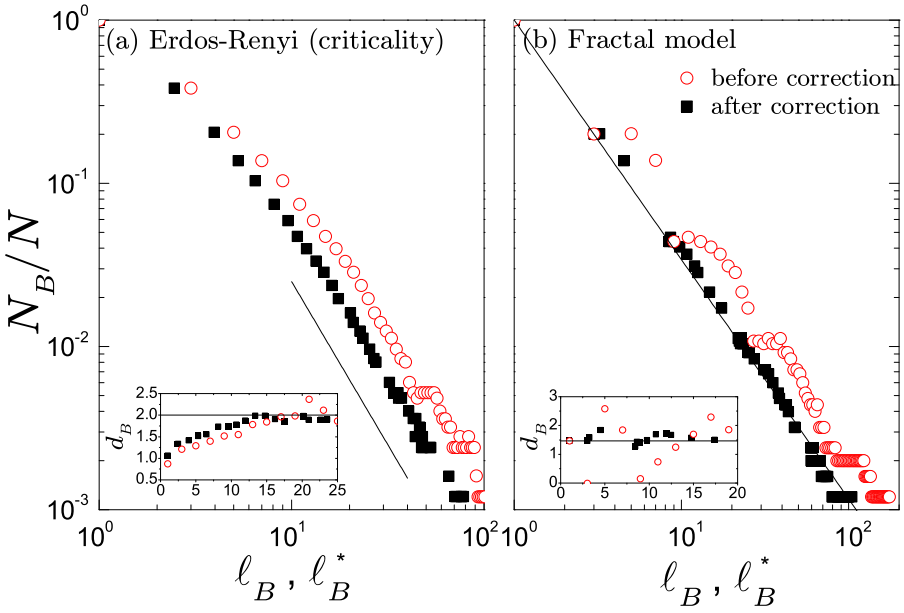


Figure 3.14: Comparison of the fractal dimension before and after applying the box-size correction in (a) an Erdos-Renyi model at criticality and (b) a fractal network model. The straight lines correspond to the analytical predictions. Insets: Improvement of the fractal dimension d_B calculation as we increase the number of boxes used for this calculation.

to the analytical prediction than the previous measurement. The improvement can be assessed by the inset plot where the use of ℓ_B^* is shown that the theoretically predicted value is achieved at smaller ℓ_B^*. Furthermore, the proposed correction has smoothened the tail in the plot, which may be crucial for the accurate determination of d_B, especially for the small box sizes considered in real networks.

The improvement achieved is more prominent in the case of the fractal network model proposed in [39]. Due to the construction process of this model, this network is highly modular with very inhomogeneous distribution of the links in the modules. As a result, the number of boxes for a given size ℓ_B fluctuates significantly and, as shown in Fig. 3.14b, it is very difficult to extract a reliable slope from the data. This discrete character has also been pointed in Ref. [43] where it is interpreted in terms of log-periodic oscillations in N_B. The use of ℓ_B^*, though, leads to a very robust slope which is exhibited over almost the entire range. As can be seen in the inset, d_B is practically always equal to its theoretical value when using the corrected value, in contrast to the uncorrected calculation where the value of d_B is more difficult to estimate.

Chapter 4

Scaling Theory of the Correlation

4.1 Joint probability distribution

Usually, the main topological features of networks are quantified through the degree distribution $P(k)$, where k is the number of links for a given node. In most cases this distribution has been found to exhibit a wide tail. Although the form of $P(k)$ has a direct influence on the network properties, it cannot convey all the information for the network structure, and it is possible that two networks with the same distribution $P(k)$ can have completely different topologies.

Therefore, the network structure is largely determined by the presence of degree correlations. This structure can be captured by the probability $P(k_1, k_2)$ that two nodes of degree k_1 and k_2 are connected to each other, and quantities derived from this, such as the Pearson coefficient r, the average degree of nearest neighbors k_{nn}, etc.

In next section, we will shown that hub anticorrelations, i.e. the tendency of the

hubs not to be directly connected with each other, give rise to fractal networks [38], such as the WWW, the protein homology network and many biological networks. On the contrary, when there is a large probability of direct hub connections the resulting networks, such as the Internet, the cond-mat coauthorship and most social networks, are non-fractals [39]. In this category fall also most of the available models for random scale-free networks, such as the Barabasi-Albert network [19] or the configuration model [48].

4.2 Degree correlation profile

The correlation profile [44] compares the joint probability distribution, $P(k_1, k_2)$, of finding a node with k_1 links connected to a node with k_2 links with their random uncorrelated counterpart, $P_r(k_1, k_2)$, which is obtained by random swapping of the links, yet preserving the degree distribution. A plot of the ratio $R(k_1, k_2) = P(k_1, k_2)/P_r(k_1, k_2)$ provides evidence of correlated topological structure that deviates from the random uncorrelated case. This latter model is obtained by, for instance, random swapping of the links in a given network [44], so that the degree distribution is preserved, but the correlation is completely lost.

The study of the ratio $R(k_1, k_2) = P(k_1, k_2)/P_r(k_1, k_2)$ reveals that most of the networks such as metabolic and protein interaction networks, the Internet and WWW are anticorrelated in comparison with the uncorrelated random case. This is because, even though this model is uncorrelated, there is still an effective attraction between the hubs since there is a large probability to randomly connect two nodes with large degrees. Thus, a plot of the ratio $R = P(k_1, k_2)/P_r(k_1, k_2)$

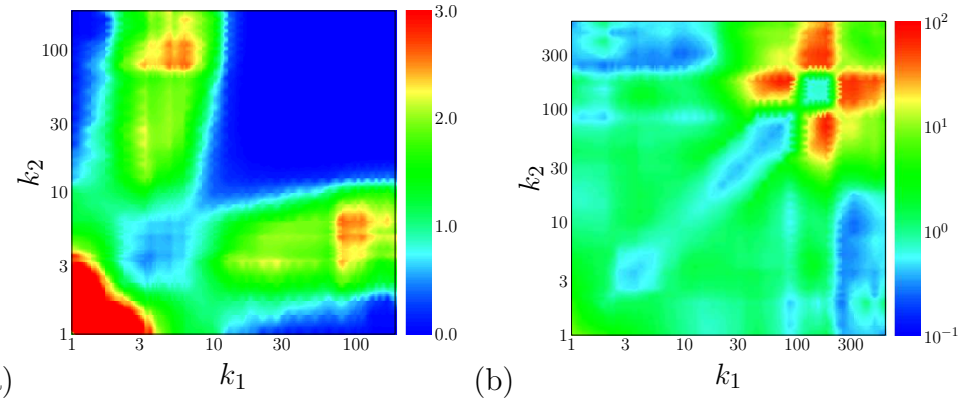


Figure 4.1: Empirical results on real complex networks. (a) Schematics showing that fractal networks are characterized by a power law dependence between N_B and ℓ_B while (a) non-fractal networks are characterized by an exponential dependence. (b) Plot of the correlation profile of the fractal metabolic network of E. *coli*, $R_{\text{E.}coli}(k_1, k_2)/R_{\text{WWW}}(k_1, k_2)$, and (d) the non-fractal Internet $R_{\text{Int}}(k_1, k_2)/R_{\text{WWW}}(k_1, k_2)$, compared with the profile of the WWW in search of a signature of fractality.

reveals that most of the real networks are anticorrelated in comparison with the uncorrelated model. Therefore this ratio does not allow to distinguish between fractal and non-fractal networks.

In search of uncovering the extent of anticorrelation that are needed to obtain fractals we study the ratios for different networks by using the WWW [8] as a reference (the use of any other network as a reference would lead to the same conclusions). This is done in Figs. 4.1a and 4.1b, by showing the correlation profiles of the cellular metabolic network of E. *coli* [22] (a fractal sample) and the Internet at the router level [26] (a non-fractal sample) respectively. These plots should be interpreted as follows: For instance, in Fig. 4.1a, let us take a large degree $k_1 = 100$ as an example. Then we see that the ratio $R_{E.coli}(k_1, k_2)/R_{WWW}(k_1, k_2)$ has a maximum for $k_2 \approx 5$ (red-yellow scale) for small k_2 ($k_2 < 10$), and a minimum

(blue scale) for large $k_2 > 10$. This means that the metabolic network has less probability to have hub-hub connection (two nodes with large degree connected) than a hub-non hub connection, when compared to the WWW. Therefore the metabolic network of E.coli is more anticorrelated than the WWW. In the same way, Fig. 4.1b shows that the hubs in the Internet have more probability to connect with other hubs than in the WWW, and therefore the Internet is less anticorrelated than the WWW. Therefore these patterns reveal that the fractal cellular networks are strongly anticorrelated (dissortative).

It is important to note that we can investigate the ratio between different networks such as $R_{\mathrm{E.}coli}(k_1, k_2)/R_{\mathrm{WWW}}(k_1, k_2)$ because $P(k_1, k_2)$ shows a power law behavior. Thus, even though the WWW and the metabolic network have different ranges of the values of k, power law scaling of $P(k_1, k_2)$ implies that the ratio is independent on the region of k_1 and k_2 used to plot this quantity.

The fractal network poses a higher degree of anticorrelation or disassortativity; nodes with a large degree tend to be connected with nodes of a small degree. On the other hand, the non-fractal Internet is less anticorrelated. Thus, fractal topologies seem to display a higher degree of hub repulsion in their structure than non-fractals. However, for this property to be the hallmark of fractality, it is required that the anticorrelation appears not only in the original network (captured by the correlation profiles of Fig. 4.1a and Fig. 4.1b), but also in the renormalized networks at all length scales. We note that other measures of anticorrelation, such as the Pearson coefficient r of the degrees at the end of an edge [45], cannot capture the difference between fractal and non-fractal network. We find that r is not invariant under renormalization.

4.3 Renormalization of degree correlation

Despite their importance, a unifying theoretical framework to fully describe and characterize degree correlations is still missing. Here we use renormalization theory to develop such a framework. We find that $P(k_1, k_2)$ develops a power-law form that is invariant under a length scale transformation. The degree correlations are fully characterized in terms of a new correlation exponent ϵ, which we calculate using a renormalization approach. We show how ϵ is given in terms of the interaction between hubs, and we find scaling relations, which are tested against real data and models. This allows us to propose a classification of all available networks according to the degree of correlations, where dissimilar networks are found to have structures that can be grouped into a small number of different classes. For example, the biological networks and the WWW are in the high anti-correlations part of the diagram, while the social networks and the Internet are clustered near the line of random networks.

We start by recalling the renormalization of a network under a scale transformation. The renormalization procedure tiles a network according to the box-covering algorithm [49], with the minimum number of boxes where the maximum distance in any box is less than ℓ_B. Each box is subsequently replaced by a node, and links are established between these new 'super-nodes' if at least one node included in a box was connected to any node of the other box. These boxes are treated as the nodes of the renormalized network and the process can be repeated iteratively, until the network reduces to a single node. Renormalization is a reliable method for determining how the network behaves at different length scales. Self-similarity is then obtained if the network structure remains invariant under the renormalization.

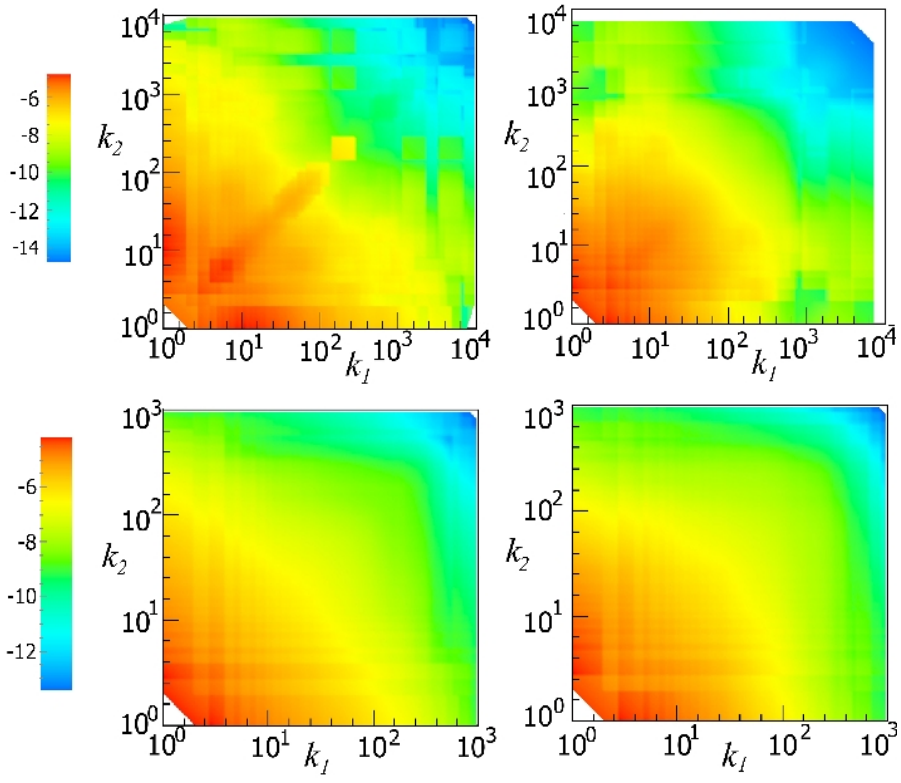


Figure 4.2: The joint degree distribution $P(k_1, k_2)$ of WWW (top row) and Internet at the router level (bottom row) before and after renormalization (with $\ell_B = 3$).

We use renormalization and scaling theory to determine the form of $P(k_1, k_2)$. Since the self-similarity of a scale-free network requires the invariance of the degree distribution $P(k)$, a power-law distribution of $P(k) \sim k^{-\gamma}$, where γ is the degree exponent, is the only form that can satisfy this condition [38]. Taking this idea one step further, it is interesting to clarify whether correlations between degrees, as expressed by the joint degree distribution $P(k_1, k_2)$, also remain invariant. In Fig. 4.2 we present an example of this distribution before and after renormalization for the WWW and the Internet at the router level (similar results are derived for other networks, as well). The statistical similarity of the two plots suggests the invariance of $P(k_1, k_2)$. Accordingly, this suggests that the k_1 and k_2 dependence can be separated and the behavior of the tail of the joint degree distribution is:

$$P(k_1, k_2) \sim k_1^{-(\gamma-1)} k_2^{-\epsilon} \ (k_1 > k_2). \tag{4.1}$$

The value of the first exponent $\gamma - 1$ in Eq. (4.1) is obtained from the density conservation law: $\int P(k_1, k_2) dk_2 = k_1 P(k_1) \sim k_1^{-(\gamma-1)}$. Equation (4.1) is also consistent with the known result for completely random networks

$$P(k_1, k_2) = k_1 P(k_1) k_2 P(k_2) = k_1^{1-\gamma} k_2^{1-\gamma}\,, \tag{4.2}$$

i.e. the exponent ϵ for these networks is $\epsilon = \gamma - 1$, as expected from the symmetry in this case.

4.4 Measurement of correlation exponent

One of the main results of this study is to confirm the form of Eq. (4.1) and provide scaling relations to determine the values of ϵ, as well as its relation with other exponents. In principle, a direct estimation of $P(k_1, k_2)$ could verify this scaling form, but such a calculation is very difficult because of large fluctuations, especially for real networks. It is clear, also, that ϵ, the correlation exponent, includes all the information on correlations in the network. Still, it is not a trivial task to directly calculate ϵ. For this reason, we introduce a new scale-invariant quantity $E_b(k)$ that simplifies the estimation of ϵ, even for small networks.

We are motivated to introduce this quantity by asking whether a node is significantly connected with more connected nodes, i.e. a node considers another node as a 'hub' if its degree is much larger than its own. For instance, we want to know if a node of low degree, e.g. $k = 5$, is connected to nodes with $k > 10$, but at the same time whether a node with $k = 500$ is connected to nodes with $k > 1000$. We, thus, define the ratio

$$E_b(k) \equiv \frac{\int_{bk}^{\infty} P(k|k_2)dk_2}{\int_{bk}^{\infty} P(k)dk}\,, \tag{4.3}$$

as the probability that a node of degree k has neighbors with degree larger than bk (b is an arbitrary positive number, and large b corresponds to the identification of the hubs). The conditional probability is $P(k_1|k_2) = P(k_1, k_2)/\int P(k_1, k_2)dk_1 = P(k_1, k_2)/k_2^{1-\gamma} = k_1^{-(\gamma-1)}k_2^{-(1+\epsilon-\gamma)}$. The denominator normalizes the probability $E_b(k)$ according to the number of nodes in the network with degree larger than bk. Thus, for a given b, a faster decay with k of $E_b(k)$ corresponds to stronger anticorrelations, i.e. hubs tend to avoid direct connections with each other. It is

straightforward to show that for a scale-free distribution:

$$E_b(k) \sim \frac{k^{1-\epsilon}}{k^{1-\gamma}} = k^{-(\epsilon-\gamma)} \,. \tag{4.4}$$

The calculation of $E_b(k)$ in a network can now be used to obtain the exponent ϵ, since usually it is straightforward to obtain the exponent γ from the degree distribution $P(k)$. Notice that we could have defined $E_b(k)$ without using the denominator in Eq. (4.3), i.e. $E_b(k) \equiv \int_{bk}^{\infty} P(k|k')dk'$, in which case we get $E_b(k) \sim k^{-(1-\epsilon)}$. However, due to the large fluctuations occuring at the tails of the degree distribution in real-life networks, the calculation of ϵ through Eq. (4.3) was proven to be more robust. First, we demonstrate in Fig. 4.3 that the scaling of $E_b(k)$ remains invariant under renormalization since the same scaling exponents are recovered for renormalized networks. The definition of $E_b(k)$ is also independent on the exact value of b, as shown in the inset of Fig. 4.3.

In Fig. 4.4 we present the behavior of $E_b(k)$ for a number of networks. The existence of a scaling relation over a large k range, combined with the invariance of this curve, is a direct confirmation of Eq. (4.4) and subsequently verifies the form used for $P(k_1, k_2)$ in Eq. (4.1). Thus, we have significant numerical support for the invariance of $P(k_1, k_2)$. The WWW and the protein homology network have been shown to have a fractal topology. The slope of $E_b(k)$ with k is small or negative in these cases with values of $\epsilon = 2.5$ and $\epsilon = 2.4$, respectively. This behavior is in striking contrast with the two non-fractal networks in the figure, i.e. the Internet at the router level and the cond-mat coauthorship network, where $E_b(k)$ increases almost linearly with increasing k. For these networks we find that $\epsilon = 1.2$ and

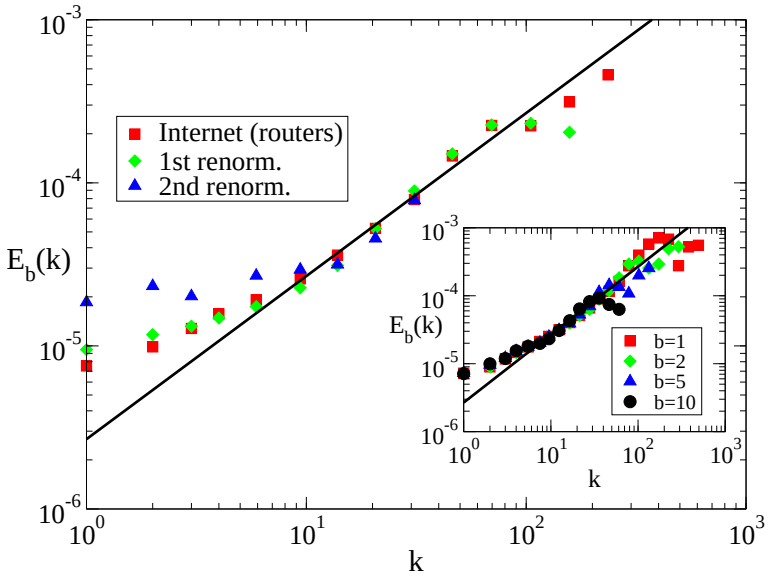


Figure 4.3: Factor $E_b(k)$ as defined in the text for three successive renormalization stages of the Internet at the router level, for $\ell_B = 3$ and $b = 3$. The data have been vertically shifted in order to show the invariance. Inset: Invariance of $E_b(k)$ vs k for different values of b.

$\epsilon = 1.6$, respectively.

4.5 Scaling theory

4.5.1 Hub-repulsion dimension

In order to develop a scaling theory for ϵ we now turn our focus on fractal networks, which are characterized by a finite fractal dimension [38]. After renormalization the number of nodes N in the network and the degree of a node k scale with ℓ_B as power laws with fractal exponent d_B and degree exponent d_k, respectively (we

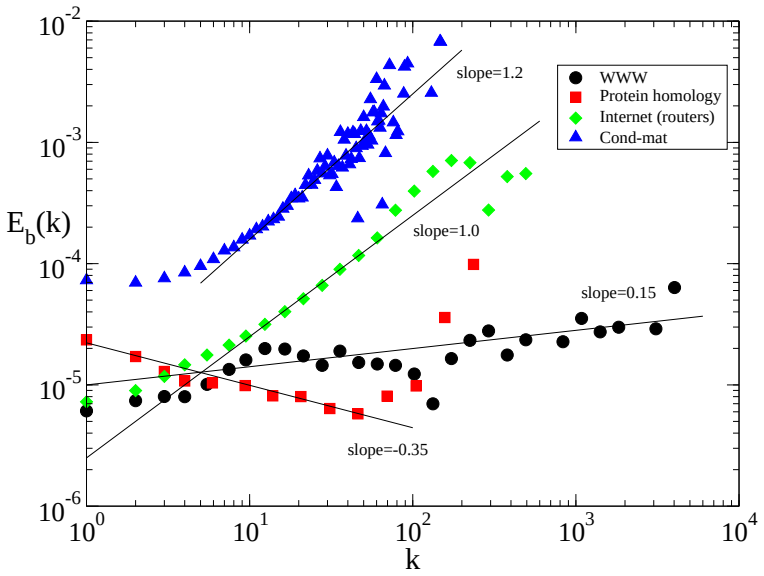


Figure 4.4: The quantity $E_b(k)$ can distinguish between the fractal (WWW, protein homology network) and non-fractal (Internet, cond-mat coauthorship) topology of networks. The different topologies correspond to different scaling behavior with the degree k.

use a prime to describe quantities measured in the renormalized network):

$$N \to N' \sim \ell_B^{-d_B} N \ , \ \ k \to k' \sim \ell_B^{-d_k} k \, . \tag{4.5}$$

If d_B and d_k are finite the network is fractal. If $d_B \to \infty$ and $d_k \to \infty$ (or equivalently the decay is exponential or faster) the network is not fractal.

After tiling the entire network with boxes of diameter ℓ_B, each of these boxes have one unique local hub (i.e. the largest degree node in the box). Considering all possible pairs of boxes, we introduce the probability $\mathcal{E}(\ell_B)$ that there exists a direct connection between the two hubs of any two boxes. A larger value of $\mathcal{E}(\ell_B)$ implies, thus, a higher probability of hub-hub correlations in the system and suggests a non-fractal structure. Furthermore, we found (see e.g. Figs. 2e, 3d of

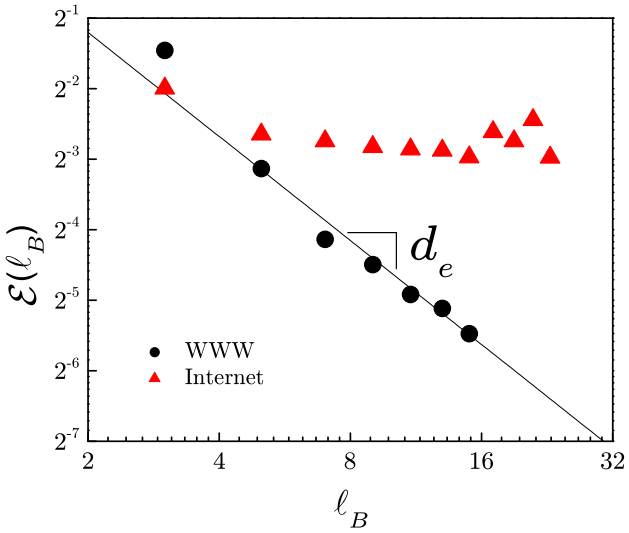


Figure 4.5: Scaling of $\mathcal{E}(\ell_B)$ as defined in Eq. (4.6) for the fractal topology of the WWW with $d_e = 1.5$, and the non-fractal topology of the Internet showing that fractal topologies are strongly anticorrelated at all length scales.

Ref. [39]) that the probability $\mathcal{E}$ also scales with the length ℓ_B as

$$\mathcal{E}(\ell_B) \sim \ell_B^{-d_e} . \tag{4.6}$$

Fig. 4.5 shows $\mathcal{E}(\ell_B)$ for two real fractal and non-fractal networks: a map of the WWW domain (nd.edu) consisting of 352,728 web-sites [8] and a map of the Internet at the router level consisting of 284,771 nodes [26]. We find that for the fractal WWW, $d_e = 1.5$, indicating that it exhibits strong anticorrelation. On the other hand, the non-fractal Internet shows $\mathcal{E}(\ell_B) \sim$ constant.

These results confirm that fractal networks, including the protein interaction network [21] (with $d_e = 1.1$) and the metabolic network of E. *coli* [22] (with

$d_e = 4.5$), do have strong hub repulsion at all length scales and non-fractal networks have no or weak hub repulsion.

A general limitation when analyzing the scaling behavior of complex networks is the small range in which the scaling is valid. This is due to the small-world property that restricts the range of ℓ_B in Fig. 4.5. As an attempt to circumvent this limitation, we offer not only the empirical determination of the exponents but also scaling theory and models where the exponents can be further tested.

4.5.2 Relation between ϵ and d_e

Below we relate the exponent ϵ to the hub-hub repulsion through the hub correlation exponent d_e, which is crucial for fractality.

We consider that in the original network there exist $n(k_1, k_2)dk_1dk_2$ links for the hubs with degrees k_1 and k_2, which in the renormalized network becomes $n'(k_1', k_2')dk_1'k_2'$. The conservation of links means that $ndk_1dk_2 = n'dk_1'dk_2'$. The number of links is $n(k_1, k_2) \equiv LP(k_1, k_2) \sim NP(k_1, k_2)$, where L is the total number of links in the network $L = \langle k \rangle N$. In the renormalized network, the corresponding number of links will be $n' = \mathcal{E}(\ell_B)L'P'(k_1', k_2')$, since the probability that two nodes are connected in the renormalized network is measured by the quantity $\mathcal{E}$ for the hubs being connected in the original network. Thus,

$$NP(k_1, k_2)dk_1dk_2 = \mathcal{E}(\ell_B)N'P'(k_1', k_2')dk_1'dk_2' \,. \tag{4.7}$$

Using Eqs. (4.1), (4.5), (4.6), and (4.7) we get the relation $\ell_B^{d_B}\ell_B^{d_e}\ell_B^{(3-\gamma-\epsilon)d_k} = 1$

which finally leads to

$$\epsilon = 2 + d_e/d_k = 2 + (\gamma - 1)\frac{d_e}{d_B}\,, \tag{4.8}$$

where we have substituted the value $\gamma = 1 + d_B/d_k$. This relation of ϵ with d_e shows that correlations between the hubs of the boxes determine the correlations for all degrees, in accordance with the invariance under renormalization.

4.6 Classification of real-world networks

The direct determination of ϵ through the slope of $E_b(k)$ vs k leads us to construct a 'phase diagram' in the plane (ϵ, γ) that allows a comparison of the structure of scale-free networks using a common framework. The importance of such a plot is that we can classify these networks in three areas according to their degree of correlations, and we can use the same information to separate fractal from non-fractal networks.

In Fig. 4.6 we present such a plot, where many real-world networks can be compared in terms of their correlation properties in the plane (ϵ, γ), even though they correspond to dissimilar systems in biology, sociology or technology. As shown in Eq. (4.2), the exponent for a random network corresponds to the line $\epsilon = \gamma - 1$, as also verified in the plot for different γ values of the configuration model. In random network models, correlations arise because links are selected for connecting with each other equiprobably, so that the probability of two hubs being connected is large [51]. For example, in the process of building a network with the configuration model the probability of selecting a link from a node during the random attachment

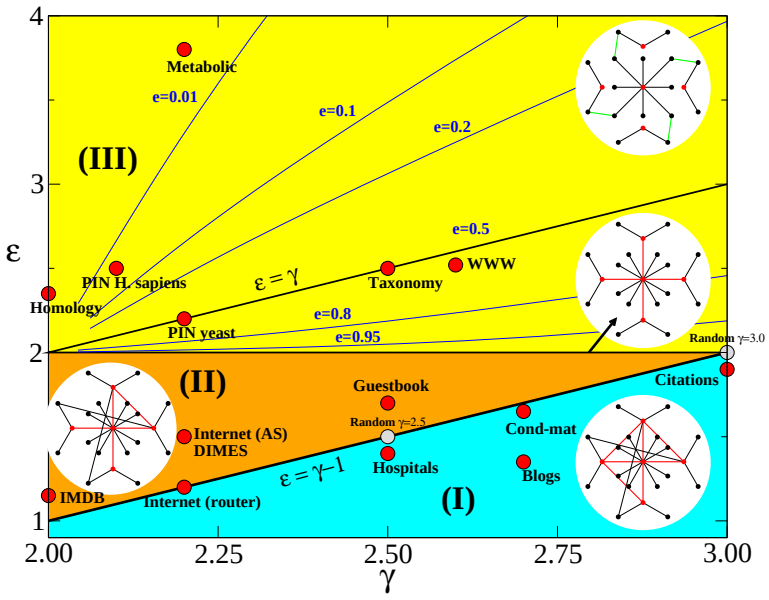


Figure 4.6: Classification of scale-free networks. The thin curves correspond to the prediction of the minimal model for varying e values. The line $\epsilon = \gamma - 1$ corresponds to a completely random network structure. The line $\epsilon = 2$ separates fractal ($\epsilon > 2$) from non-fractal networks ($\epsilon \leq 2$), while the line $\epsilon = \gamma$ describes a fractal tree. The schematics illustrate networks where hub correlations are stronger than in random networks (area I), weaker than random but non-fractal (area II), non-fractal according to the minimal model ($\epsilon = 2$), and fractal (area III).

is proportional to this node's degree, and so the hubs have a higher probability of being connected to each other. Thus, networks that are close to the line $\epsilon = \gamma - 1$ exhibit strong hub-hub correlations. This line separates the diagram in two main parts, (a) above the line where the hub correlations tend to become weaker, and (b) below the line where networks have even larger correlations (hubs are connected to each other with even higher probability than the one corresponding to a randomly created structure).

In the diagram, the social networks and the Internet at the router level are clustered around the line $\epsilon = \gamma - 1$. This is an indication that there is a strongly connected core of hubs in these systems, consistent with previous studies in such systems. The biological networks and the WWW, on the other side, are clearly quite far away from this line. This implies that there is a richer structure in these networks with hubs separated from each other, and the distance in the plot from the above line quantifies how different from randomness the network structure is, in terms of degree correlations.

An immediate important result from this diagram is the different placement of the Internet at the router level compared to the AS level [52]. Although the degree distribution of these two networks is the same ($\gamma = 2.2$), the correlation exponent ϵ reveals that there are more hub-hub connections at the router level, similar to the case of a random network. Contrary to that, the AS level exhibits a structure with less correlations that deviates from that of a simple random model.

We have already seen that a network with strong hub anticorrelations is a fractal network. This means that as ϵ increases and we move away from the $\epsilon = \gamma - 1$ line, we expect that at some point the networks will become fractal. The point of

emergent fractality can be easily calculated through Eq. (4.8), since in non-fractal networks we have $d_B \to \infty$, which in turn yields $\epsilon = 2$ for this borderline case. Indeed, we have verified via direct measurements of d_B that all the networks above the line $\epsilon = 2$ in Fig. 4.6 are fractals.

We can also study the importance of correlations by using the minimal fractal network model that we introduced in Ref. [39]. The main parameter in this model is the probability e that two hubs are directly connected. The behavior of this model for various e values is shown in the plot. Of particular importance is the line $\epsilon = \gamma$ which corresponds to the prediction of the minimal model for $e = 0.5$. This line also corresponds to a typical branching process that generates fractal tree networks. The case of $e = 0.5$ separates the 'strong' anticorrelations (above the $\epsilon = \gamma$ line) from weaker anticorrelations (below ($\epsilon = \gamma$) where, though, both structures are fractal as long as $e < 1$ ($e = 1$ leads to non-fractal networks).

The emerging picture from Fig. 4.6 is that the values of ϵ and γ can reveal the strength of correlations in a network and cluster the networks into well-defined areas whose properties we can easily identified. Starting from the random case line $\epsilon = \gamma - 1$ we can separate the phase space into areas where the hub correlations are stronger than in random models (area I) or weaker than that (areas II and III). The weak correlation areas II and III are further divided by the line $\epsilon = 2$ which determines whether the anticorrelations are strong enough to finally result in a fractal network (III) or not (II).

Chapter 5

Modelling Self-Similar Networks

The "democratic" rule of the seminal Erdös-Rényi model [10] (where the nodes in the network are connected at random) was first invoked to explain the small world effect. It was then replaced by the "rich-get-richer" principle of preferential attachment [19] to explain the scale-free property; a discovery carrying important implications on network vulnerability [53, 54]. However these rules do not capture the fractal topologies found in diverse complex networks. In the second section, we find that models of scale-free networks are not fractals. Here, we demonstrate a new view of network dynamics where the growth takes place *multiplicatively* in a correlated self-similar fashion, in contrast to the uncorrelated growth of models of preferential attachment [24, 19].

5.1 Growth mechanism

The way to distinguish between fractal and non-fractal networks is represented in their scaling properties as seen in Fig. 5.1a and 5.1b. Fractal networks can be

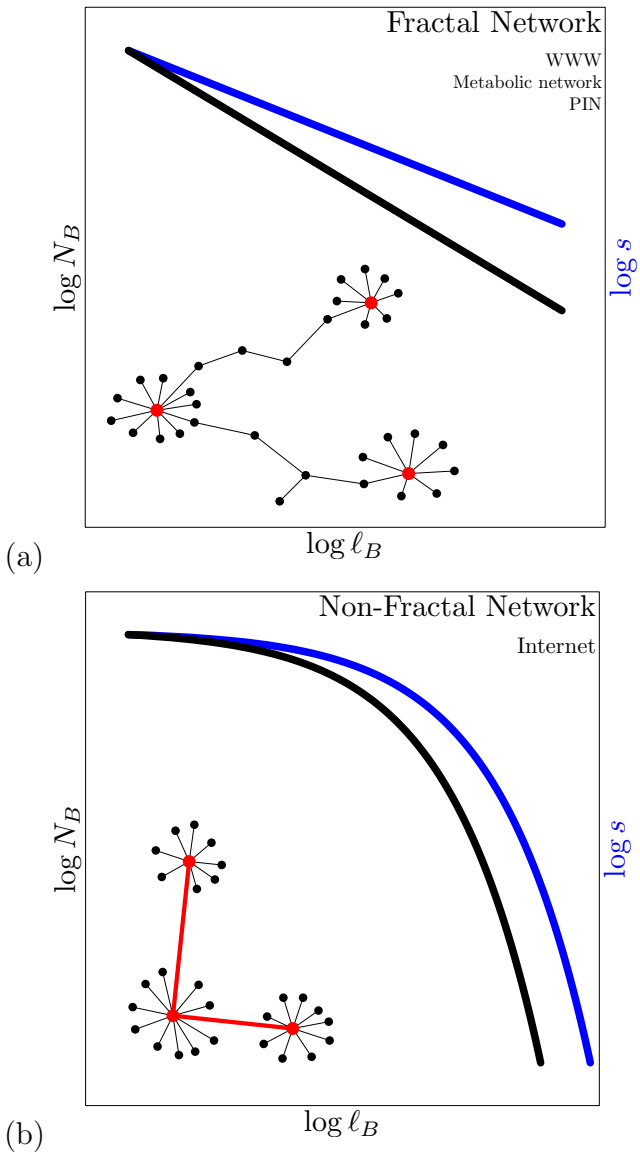


Figure 5.1: Empirical results on real complex networks. (a) Schematics showing that fractal networks are characterized by a power law dependence between N_B and ℓ_B while (b) non-fractal networks are characterized by an exponential dependence.

characterized by the following scaling relations (Fig. 5.1a):

$$N_B(\ell_B)/N \sim \ell_B^{-d_B} \quad \text{and} \quad k_B(\ell_B)/k_{hub} \sim \ell_B^{-d_k}, \tag{5.1}$$

where k_{hub} and $k_B(\ell_B)$ are the degree of the most connected node inside each box and that of each box respectively (Fig. 5.2). The exponents d_B and d_k are the fractal dimension and the degree dimension of network, respectively. While the term "fractal dimension" is usually reserved for geometrical self-similarity, here we relax the usage to include the topological self-similarity as well. For a non-fractal network like the Internet (Fig. 4.1b), we have $d_B \to \infty$ and $d_k \to \infty$; the scaling laws in Eq. (5.1) are replaced by exponential functions.

In last chapter, we showed that the emergence of self-similar fractal networks, such as cellular ones, is due to the strong repulsion (disassortativity [45]) between the hubs at all length scales. In other words, the hubs prefer to grow by connections to less-connected nodes rather than to other hubs, an effect that can be viewed as an effective hub repulsion. In this new paradigm, the "rich" still get richer, although at the expense of the "poor". In other words, the hubs grow by preferentially linking with less-connected nodes to generate a more robust fractal topology. In contrast, weakly anticorrelated or uncorrelated growth leads to non-fractal topologies such as the Internet.

Based on the results leading to Eq. (5.1), we propose a network growth dynamics as the inverse of the renormalization procedure. Thus, the coarse-grained networks of smaller size are network structures appearing earlier in time, as exemplified in Fig. 5.2. A present time network with $\tilde{N}(t)$ nodes is tiled with

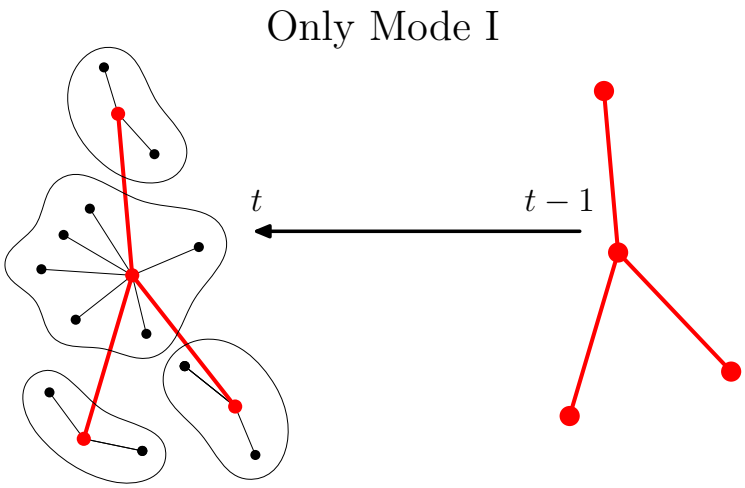


Figure 5.2: The dynamical growth process can be seen as the inverse renormalization procedure with all the properties of the network being invariant under time evolution. In this example $\tilde{N}(t) = 16$ nodes are renormalized with $N_B(\ell_B) = 4$ boxes of size $\ell_B = 3$.

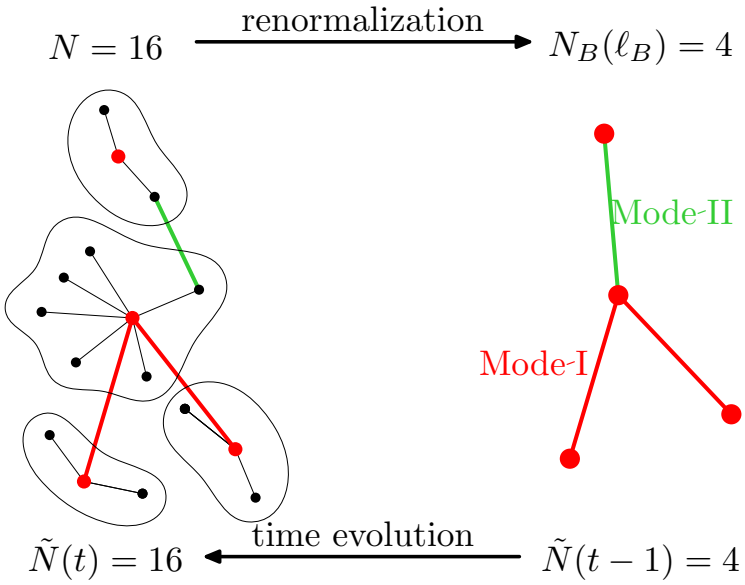


Figure 5.3: Analysis of Mode I, only: the boxes are connected directly leading to strong hub-hub attraction or assortativity. This mode produces a scale-free, small-world network but without the fractal topology.

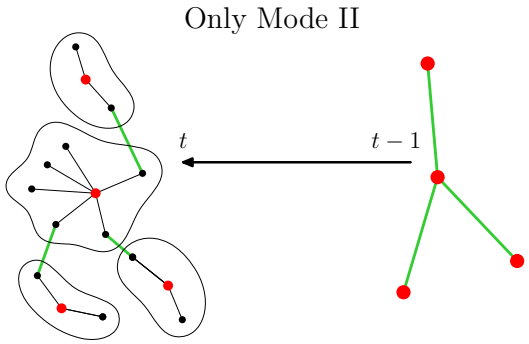


Figure 5.4: Mode II alone produces a scale-free with a fractal topology but not the small-world effect. Here the boxes are connected via non-hubs

$N_B(\ell_B)$ boxes of size ℓ_B. Each box represents a node in a previous time step, so that $\tilde{N}(t-1) = N_B(\ell_B)$. The maximum degree of the nodes inside a box corresponds to the present time degree: $\tilde{k}(t) = k_{hub}$, which is renormalized such that $\tilde{k}(t-1) = k_B(\ell_B)$. The tilde over the quantities are needed in order to differentiate the dynamical quantities, such as the number of nodes as a function of time, $\tilde{N}(t)$, from the static quantities, such the number of nodes of the present network, N, or the number of nodes of the renormalized network, N_B. The renormalization procedure applies to many complex networks in Nature [38]. These includes fractal networks such as WWW, protein interaction networks of E. *coli*, the yeast [41] and human, and metabolic networks of 43 different organisms from the three domains of life, and some sociological networks. The renormalization scheme can be applied to non-fractal networks, such as the Internet, as well. Below we will show that the main difference between these two groups is in the connectivity correlation. We also provide empirical, analytical and modeling evidences supporting this theoretical framework based on the validity of exponents, scaling theory, and statistical properties of the connectivity correlation.

5.2 Mathematical model

To quantitatively link the anticorrelation at all length scales to the emergence of fractality, we next develop a mathematical framework and demonstrate the mechanism for fractal network growth. In the case of modular networks, stemming

from Eqs. (5.1), we require that

$$\begin{aligned} \tilde{N}(t) &= n\tilde{N}(t-1), \\ \tilde{k}(t) &= s\tilde{k}(t-1), \\ \tilde{L}(t) + L_0 &= a(\tilde{L}(t-1) + L_0), \end{aligned} \tag{5.2}$$

where $n > 1$, $s > 1$ and $a > 1$ are time-independent constants and $\tilde{L}(t)$ is the diameter of the network defined by the largest distance between nodes. The first equation is analogous to the multiplicative process naturally found in many population growth systems [55]. The second relation is analogous to the preferential attachment rule [19]. It gives rise to the scale-free probability distribution of finding a node with degree k, $P(k) \sim k^{-\gamma}$. The third equation describes the growth of the diameter of the network and determines whether the network is small-world [12] and/or fractal. Here we introduce the characteristic size L_0, the importance of which lies in describing the non-fractal networks. Since every quantity increases by a factor of n, s and a, we first derive the scaling exponents in terms of the microscopic parameters: $d_B = \ln n / \ln a$, $d_k = \ln s / \ln a$. The exponent of the degree distribution satisfies $\gamma = 1 + \ln n / \ln s$. The dynamics represented by Eqs. (5.2) consequently leads to a modular structure where modules are represented by the boxes. While modularity has often been identified with the scaling of the clustering coefficient [50], here we propose an alternative definition of "modular network" as the one whose statistical properties remain invariant under renormalization.

In order to incorporate different growth modes in the dynamical Eqs. (5.2) we consider, without loss of generality, two modes of connectivity between boxes, whose relative frequencies of occurrence are controlled by the probability e repre-

senting the hub-hub attraction. *(i)* Mode I with probability e (Fig. 5.3): two boxes are connected through a direct link between their hubs leading to hub-hub attraction. *(ii)* Mode II with probability $1 - e$ (Fig. 5.4): two boxes are connected via non-hubs leading to hub-hub repulsion or anticorrelation. We will show that Mode I leads to non-fractal networks while Mode II leads to fractal networks. In practice, though Eqs. (5.2) are deterministic, we combine these two modes according to the probability e, which renders our model probabilistic.

Formally, for a node with $\tilde{k}(t-1)$ links at time $t-1$, we define $\tilde{n}_h(t)$ as the number of links which are connected to hubs in the next time step (see Fig. 5.2). Then the probability e satisfies:

$$\tilde{n}_h(t) = e\ \tilde{k}(t-1). \tag{5.3}$$

Using the analogy between time evolution and renormalization, we obtain the hub-repulsion exponent $d_e = -\ln e / \ln a$. This result suggest that the simultaneous appearance of both the small-world and fractal properties in scale-free networks is due to a combination of the growth modes. In general, the growth process is a stochastic combination of Mode I (with probability e) and Mode II (with probability $1 - e$). For the intermediate ($0 < e < 1$), the model predicts finite fractal exponents d_B and d_k and also bears the small-world property due to the presence of Mode I. Such a fractal small-world and scale-free network is visualized in Fig. 5.7 for $e = 0.8$. Supporting evidences are given by *(i)* Fig. 5.8, which shows that the model with $e = 0.8$ is more anticorrelated than the $e = 1$ model (Mode I), *(ii)* Fig. 5.9, which shows the power law dependence of N_B on ℓ_B for the fractal

structure ($e = 0.8$), and the exponential dependence of the non-fractal structure ($e = 1$), and *(iii)* Fig. 5.10 showing that Mode I reproduces $\mathcal{E}(\ell_B) \sim$ constant while the $e = 0.8$ model gives $\mathcal{E}(\ell_B) \sim \ell_B^{-d_e}$, which is in agreement with the empirical findings of Fig. 4.5 on real networks (the exponent $d_e = -\ln 0.8/\ln 1.4 = 0.66$ is predicted by the analytical formula according to Section 5.3). Furthermore, in the Section 5.4 we show that the predicted scale-free distribution is invariant under renormalization. Although simplistic, this minimal model clearly captures an essential property of networks: the relationship between anticorrelations and fractality (see Methods for more details). We have also considered the contribution of loops, which we find does not change the general conclusions of this study.

The same analysis is performed for the model in Fig. 5.8. For instance, in this figure, given a large degree $k_1 = 300$, the ratio $R_{e=1}(k_1, k_2)/R_{e=0.8}(k_1, k_2)$ is small (blue/green region) for small k_2 ($k_2 < 10$) but large (red/yellow region) for large k_2 ($k_2 > 10$). This means that the network with $e = 1$ is more likely to have hub-hub connections than the $e = 0.8$ case. Thus, the profile shows how $e = 0.8$ is more anticorrelated than Mode I. For the model we have: Mode I $<$ Model with intermediate $e <$ Mode II.

5.3 Scaling theory

Here we elaborate on several theoretical expressions presented in the main text. We fully develop the theoretical framework of renormalization and its analogy with the time evolution of networks.

Table 5.1: Relation between time evolution and renormalization.

Quantity	Time evolution	Renormalization
diameter	$\tilde{L}(t_1) + L_0$	$L/(\ell_B + L_0)$
	$\tilde{L}(t_2) + L_0$	L
number of nodes	$\tilde{N}(t_1)$	$N_B(\ell_B)$
	$\tilde{N}(t_2)$	N
degree	$\tilde{k}(t_1)$	$k(\ell_B)$
	$\tilde{k}(t_2)$	k_{hub}
hub-hub links	$\tilde{k}(t_1)$	$k(\ell_B)$
	$\tilde{n}_h(t_2\|t_1)$	$n_h(\ell_B)$

The multiplicative growth law is expressed as:

$$\begin{aligned} \tilde{L}(t+1) + L_0 &= a(\tilde{L}(t) + L_0), \\ \tilde{N}(t+1) &= n\tilde{N}(t), \\ \tilde{k}(t+1) &= s\tilde{k}(t), \\ \tilde{n}_h(t+1) &= e\tilde{k}(t), \end{aligned} \tag{5.4}$$

where all the quantities have been previously defined in the main text. In Fig. 5.2 of the main text we provide an example of these quantities in a hypothetical growth process. In this example $\tilde{N}(t) = 16$ nodes are renormalized with $N_B(\ell_B) = 4$ boxes of size $\ell_B = 3$ so that $\tilde{N}(t-1) = 4$ nodes existed in the previous time step. The box size is defined as the maximum chemical distance in the box plus one. The chemical distance is the number of links of the minimum path between two nodes. The central box has a hub with $k_{hub} = 8$ links, then $\tilde{k}(t) = 8$. After

renormalization, $k_B(\ell_B = 3) = 3$ for this central box, so that $\tilde{k}(t-1) = 3$. Out these three links, two are via a hub-hub connection (Mode I), thus $n_h(\ell_B) = 2$ and $\mathcal{E}(\ell_B) = 2/3$, for this case.

We obtain the relation between the quantities at two times $t_2 > t_1$ as

$$\begin{aligned} \tilde{L}(t_2) + L_0 &= a^{t_2-t_1}(\tilde{L}(t_1) + L_0), \\ \tilde{N}(t_2) &= n^{t_2-t_1}\tilde{N}(t_1), \\ \tilde{k}(t_2) &= s^{t_2-t_1}\tilde{k}(t_1), \\ \tilde{n}_h(t_2|t_1) &= e^{t_2-t_1}\tilde{k}(t_1). \end{aligned} \tag{5.5}$$

Notice that the quantity $\tilde{n}_h(t_2|t_1)$ represents a special case. This quantity indicates the number of links at time t_2, which are connected to hubs generated before time t_1. To avoid the confusion with the other quantities, we introduce a new notation $\tilde{n}_h(t_2|t_1)$ instead of $\tilde{n}_h(t_2)$ as used for the other quantities in Eq. (5.5). We also notice that the notation $\tilde{n}_h(t)$ in the main text Eq. (5.3) is then interpreted as $\tilde{n}_h(t|t-1)$ for short. We then obtain: $\tilde{n}_h(t_2|t_1) = e\ \tilde{n}_h(t_2-1|t_1) = \ldots = e^{t_2-t_1}\ n_h(t_1|t_1) = e^{t_2-t_1}\ k(t_1)$, where we have used that $n_h(t_1|t_1) = k(t_1)$.

The relationship between the quantities describing the time evolution and the renormalization is shown in Table 5.1. They are formalized as follows:

$$\begin{aligned} \ell_B + L_0 &= (\tilde{L}(t_2) + L_0)/(\tilde{L}(t_1) + L_0) = a^{t_2-t_1} \\ \mathcal{N}(\ell_B) &\equiv N_B(\ell_B)/N = \tilde{N}(t_1)/\tilde{N}(t_2) = n^{t_1-t_2}, \\ \mathcal{S}(\ell_B) &\equiv k_B(\ell_B)/k_{hub} = \tilde{k}_B(t_1)/\tilde{k}_B(t_2) = s^{t_1-t_2}, \\ \mathcal{E}(\ell_B) &\equiv n_h(\ell_B)/k_B(\ell_B) = \tilde{n}_h(t_2|t_1)/\tilde{k}(t_1) = e^{t_2-t_1}. \end{aligned} \tag{5.6}$$

Here we define the additional ratios, $\mathcal{N}$ and $\mathcal{S}$. Replacing the time interval $t_2 - t_1$ by $\ln(\ell_B + L_0)/\ln a$, as obtained from the first equation in (5.6), we obtain:

$$\begin{aligned} \mathcal{N}(\ell_B) &= (\ell_B + L_0)^{-\ln n/\ln a}, \\ \mathcal{S}(\ell_B) &= (\ell_B + L_0)^{-\ln s/\ln a}, \\ \mathcal{E}(\ell_B) &= (\ell_B + L_0)^{-\ln(1/e)/\ln a}, \end{aligned} \tag{5.7}$$

or

$$\begin{aligned} \mathcal{N}(\ell_B) &= (\ell_B + L_0)^{-d_B}, d_B \equiv \ln n/\ln a, \\ \mathcal{S}(\ell_B) &= (\ell_B + L_0)^{-d_k}, d_k \equiv \ln s/\ln a, \\ \mathcal{E}(\ell_B) &= (\ell_B + L_0)^{-d_e}, d_e \equiv \ln(1/e)/\ln a, \end{aligned} \tag{5.8}$$

which correspond to the equations described in the main text. Notice that we have considered $L_0 = 0$ in Eqs. (5.1) for simplicity. Equations (5.8) are more general and accommodate the case of non-fractal networks which are characterized by exponential functions:

$$\begin{aligned} \mathcal{N}(\ell_B) &\sim \exp(-\ell_B/\ell_0), \\ \mathcal{S}(\ell_B) &\sim \exp(-\ell_B/\ell_0'). \end{aligned} \tag{5.9}$$

These expressions arise from Eqs. (5.8) by taking the limit of $d_B \to \infty$, $d_k \to \infty$, and $L_0 \to \infty$ while $L_0/d_B \to \ell_0$ and $L_0/d_k \to \ell_0'$, where ℓ_0 and ℓ_0' are characteristic constants of the network.

5.4 The minimal model

In the framework of the minimal model, we start with a star structure at $t = 0$ as seen in Fig 5.5a. At each time step $mk(t)$ new nodes are generated for each node with degree $k(t)$, where m is an input parameter ($m = 2$ in Fig. 5.5). Accordingly, we have $\tilde{N}(t+1) = \tilde{N}(t) + 2m\tilde{K}(t)$, where $\tilde{K}(t)$ is the total number of links at time t. Since we do not consider the loop structure at the moment, we have $\tilde{K}(t) = \tilde{N}(t)$. Then we obtain $\tilde{N}(t+1) = (2m+1)\tilde{N}(t)$, or $n = 2m+1$. We find that the results of the model are independent on the initial configuration.

Then, two different connectivity modes are chosen as follows: Mode I, we keep all the old connections generated multiplicatively at time t (the red links in Fig. 5.5b). Mode II, all the old connections generated in the previous time time step are replaced by links between new generated nodes (see the green links in Fig. 5.5c).

Mode I implies $s = m + 1$, since $\tilde{k}(t+1) = m\tilde{k}(t) + \tilde{k}(t)$, where the term mk comes from newly generated nodes, and the term k comes from the links at previous time steps. The diameter of the network grows additively as: $\tilde{L}(t+1) = \tilde{L}(t) + 2$ ($a = 1$, $L_0 \to \infty$ and $(a-1)L_0 \to 2$ in Eqs. (5.4)) because at each step we generate one extra node at both sides of the network and therefore the size of the network is increase by 2, as seen in Fig. 5.5b. This implies $\tilde{L}(t) \sim 2t$. For this mode we obtain a non-fractal topology: $N_B(\ell_B)/N \sim \exp(-\frac{\ln n}{2}\ell_B)$ and $k(\ell_B)/k_{hub} \sim \exp(-\frac{\ln s}{2}\ell_B)$; a direct consequence of the linear growth of the diameter $\tilde{L}(t)$ which implies that the network is small-world. Moreover, $a = 1$ leads to $d_B = \ln n / \ln a \to \infty$ in this case (non-fractal).

Mode II alone gives rise to a fractal topology but with a breakdown of the small-

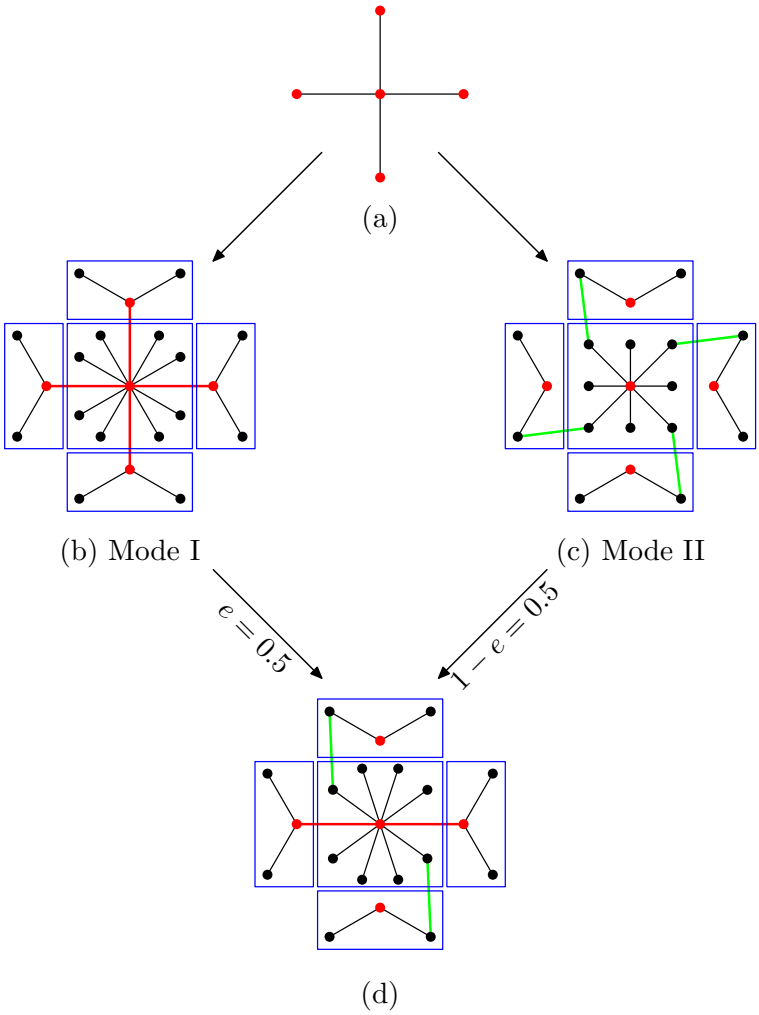


Figure 5.5: Different modes of growth with $m = 2$. Starting with (a) five nodes at $t = 0$, the different connectivity modes lead to different topological structures, which are (b) Mode I, (c) Mode II and (d) combination of Mode I and II with probability $e = 0.5$.

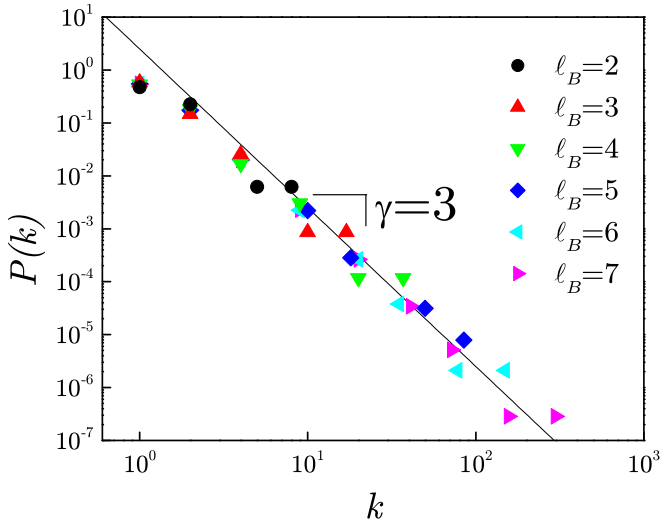


Figure 5.6: Predictions of the model for $e = 0.5$ for the degree distribution showing the power law behavior with $\gamma = 3$ and its invariance under time evolution.

world property. The diameter increases multiplicatively $\tilde{L}(t+1) = 3\tilde{L}(t)$ ($a = 3$ and $L_0 = 0$ in Eqs. (5.4)), because we replace all the links at previous time step by the paths with chemical distance 3. The degrees grow as $\tilde{k}(t+1) = m\tilde{k}(t)$ according to our generation protocol, which leads to $s = m$. The multiplicative nature of $\tilde{L}(t)$ leads to an exponential growth in the diameter with time, $\tilde{L}(t) \sim e^{t \ln 3}$, and consequently to a fractal topology with finite d_B and d_k according to Eqs. (5.6)-(5.7). This is seen because for this mode we have $a = 3$. We obtain $N_B(\ell_B) \sim \ell_B^{-d_B}$ with finite $d_B = \ln(2m+1)/\ln 3$ and also $k(\ell_B)/k_{hub} \sim \ell_B^{-d_k}$ with $d_k = \ln m/\ln 3$. However, the multiplicative growth of $\tilde{L}$ leads to the disappearance of the small-world effect, which is replaced by a power-law dependence.

The general growth process is a stochastic combination of Mode I (with probability e) and Mode II (with probability $1-e$, see Fig. 5.5d). We obtain $\tilde{L}(t+1) =$

$(3-2e)\tilde{L}(t)+2e$, and $a=3-2e$ and $L_0=e/(1-e)$. Then, when $e\to 1$ (Mode I) $a=1$ and $d_B\to\infty$, $L_0\to\infty$ and we obtain a non-fractal topology. On the other hand, Mode II has $e=0$, then $L_0=0$ and $a>1$ and d_B is finite, following the fractal scaling. For an intermediate $0<e<1$ this model predicts finite fractal exponents d_B and d_k and also predicts the small-world effect due to the presence of Mode I, as shown in the main text. This is seen in the exponential behavior of $\langle M_c\rangle$ versus ℓ_B.

In Fig. 5.6 we show further evidence that this model reproduces the self-similar properties found in fractal networks by plotting $P(k)$. We find that the model is modular since $P(k)$ is invariant under renormalization with $\gamma=1+\ln n/\ln s=3$, which is in agreement with the empirical findings of Fig. 5.6. Consistent with our predictions we find that $d_k=\ln s/\ln a=3.3$.

Finally we summarize the predictions of the model. The fractal dimension is $d_B=\ln n/\ln a$, and in the framework of the minimal model with probability e, we find $a=3-2e$ and $L_0=e/(1-e)$. Then, when $e\to 1$ (Mode I) $a=1$ and $d_B\to\infty$, $L_0\to\infty$ and $\ell_0=L_0/d_B=2/\ln n$ giving a non-fractal topology as in Eqs. (5.9). On the other hand, Mode II has $e=0$, then $L_0=0$ and $a>1$ and d_B is finite, following the fractal scaling of Eqs. (5.8), as long as the growth of the number of nodes is multiplicative with a well-defined value of n. The fact that $a=1$ implies a linear growth of the diameter $\tilde{L}(t)\sim 2t$, which produces the small-world property. An intermediate model with, for instance $e=0.8$ gives rise to a fractal network with the small world effect, as shown by the scaling of $N_B(\ell_B)$ in Fig. 5.9 and $\mathcal{E}(\ell_B)$ in Fig. 5.10, in the main text.

5.5 Additional supporting evidence for the fractal network model

Evidence is given in Fig. 5.14 for the minimal model with parameter $e = 0.8$. We calculate *(i)* the mean number of nodes (mass) of the boxes tiling the network, $\langle M_B(\ell_B)\rangle$ ($\equiv N/N_B(\ell_B)$), by using the box covering methods, and *(ii)* the local mass $\langle M_c(\ell_c)\rangle$ by averaging over boxes of size ℓ_c around a randomly chosen node (the cluster growing method, see [38, 2]). The results show how the minimal model reproduces one of the main properties of fractal networks [38]: the power-law relation for the global average mass $\langle M_B(\ell_B)\rangle \sim \ell_B^{d_B}$ with $d_B = \ln 5/\ln 1.4 = 4.8$ as a signature of fractality consistent with Eq. (5.1), and the exponential dependence of the local mass $\langle M_c(\ell_c)\rangle \sim e^{\ell_c/\ell_0}$ as a signature of the small-world effect: $\ell_c \sim \ln\langle M_c\rangle$. Note that the cluster growing method is actually a way to measure the distance, while the box covering method measures the fractality [38]. The model leads also to a smooth monotonic scaling in the size distribution of modules as observed in [38]. The global small-world properties are treated next.

5.6 Study of other scale-free models

While the origin of the scale-free property can be reduced to two basic mechanisms: growth and preferential attachment, as exemplified by the Barabási-Albert model (BA model [19]), the empirical result of fractality cannot be explained only in those terms. Notice that the term "scale-free" coined by Barabási-Albert [19] refers to the absence of a typical number of links, as exemplified by a power-law distribution

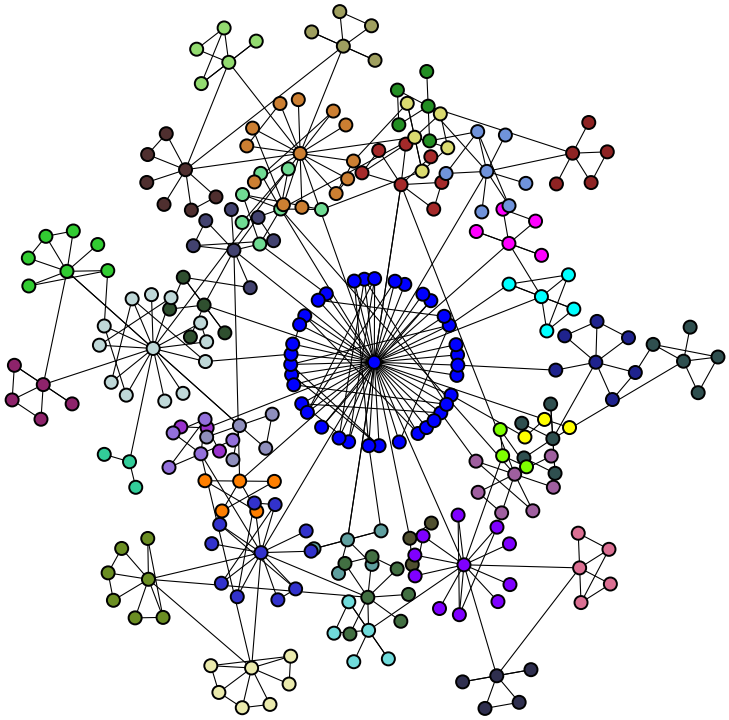


Figure 5.7: Resulting topology predicted by the minimal model for $e = 0.8$, $n = 5$, $a = 1.4$, $s = 3$ and $m = 2$. The colors of the nodes show the modular structure with each color representing a different box. We also include loops in the structure as discussed later on.

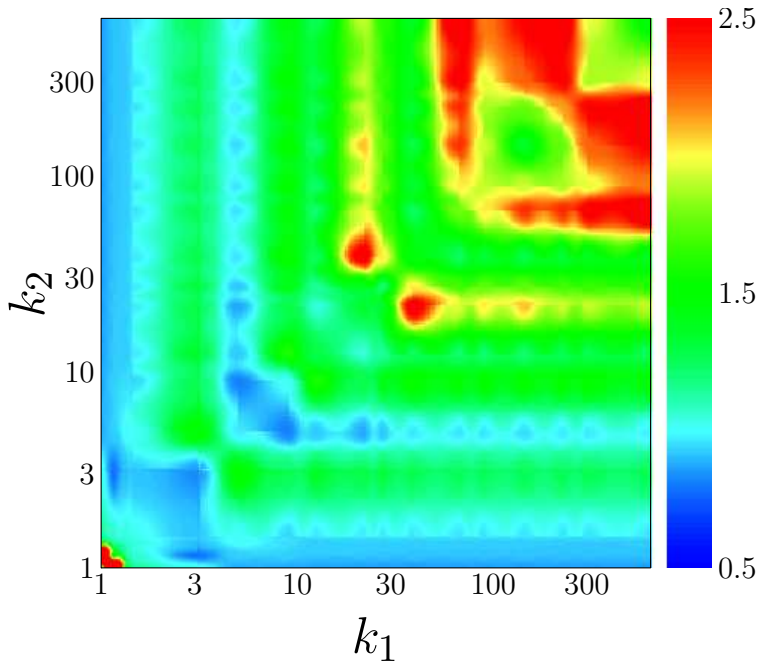


Figure 5.8: Ratio $R_{e=1}(k_1, k_2)/R_{e=0.8}(k_1, k_2)$ to compare the hub-hub correlation emerging from the model networks generated with $e = 1$ and $e = 0.8$, respectively.

of degree connectivity, but it does not refer to the length scale invariance found in [38].

We find that all models of scale-free networks such as the BA model of preferential attachment [19], the hierarchical model [50], and the so-called pseudo fractal models and trees [58, 59] are non-fractals. In Fig. 5.11, Fig. 5.12 and Fig. 5.13, we plot the number of boxes N_B versus ℓ_B for the models showing that in all the cases the decay of $N_B(\ell_B)$ is exponential or faster, indicating either an infinite d_B or not a well-defined fractal dimension.

In the present study we find the relation $\gamma = 1 + \ln n/\ln s$, by using $d_B = \ln n/\ln a$, and $d_k = \ln s/\ln a$, as explained in the text. However, non-fractal networks satisfy this relation as well despite the infinite fractal dimension $d_B \to \infty$.

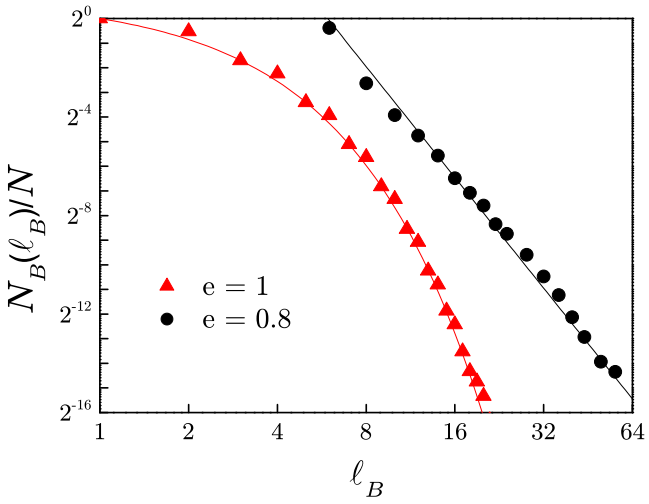


Figure 5.9: Plot of N_B versus ℓ_B showing that Mode I is non-fractal (exponential decay) and $e = 0.8$ is fractal (power-law decay) according to Fig.5.8 and in agreement with the empirical results of Fig. 4.1.

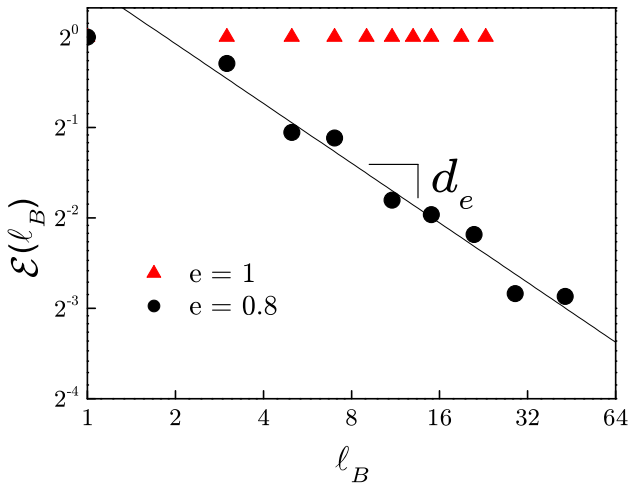


Figure 5.10: Scaling of $\mathcal{E}(\ell_B)$ reproducing the behavior of fractal networks for $e = 0.8$ and non-fractal networks Mode I, $e = 1$, as found empirically in Fig. 4.5.

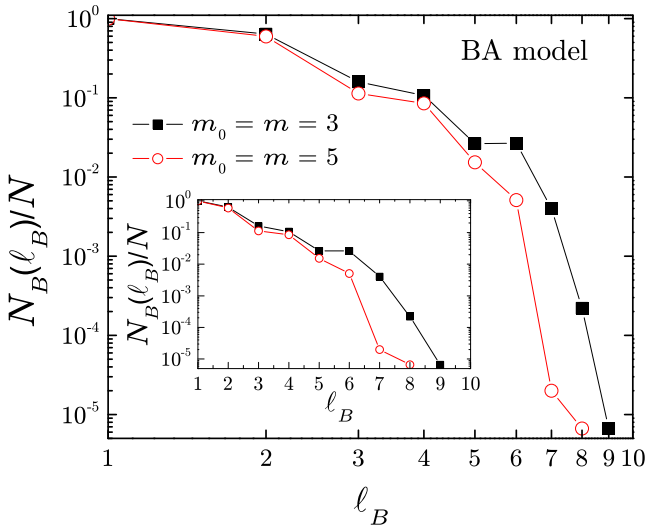


Figure 5.11: Lack of fractality in the BA model of preferential attachment [19]

Thus in general we say that when $\gamma = 1 + \ln n / \ln s$ is satisfied, then the degree distribution is invariant under renormalization.

5.7 Global small world: short cuts in the network

An important factor in the dynamics of real-world networks is the existence of randomness or noise in the growth process. The simplest type of noise is the appearance of random connections between nodes as exemplified in the Watts-Strogatz model of small world networks [12]. To investigate how noise affects the fractality of networks, we modify the dynamical law of the model as follows: at each time step, $p\tilde{K}(t)$ number of links are added in at random, here $\tilde{K}(t)$ is the

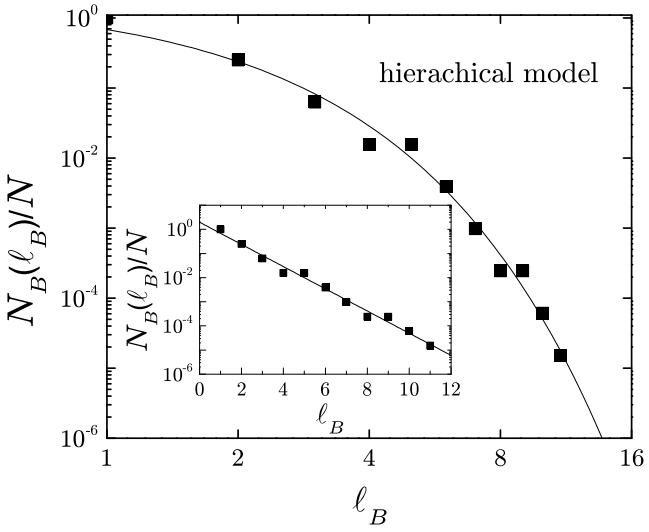


Figure 5.12: Lack of fractality in the hierarchical model [50] .

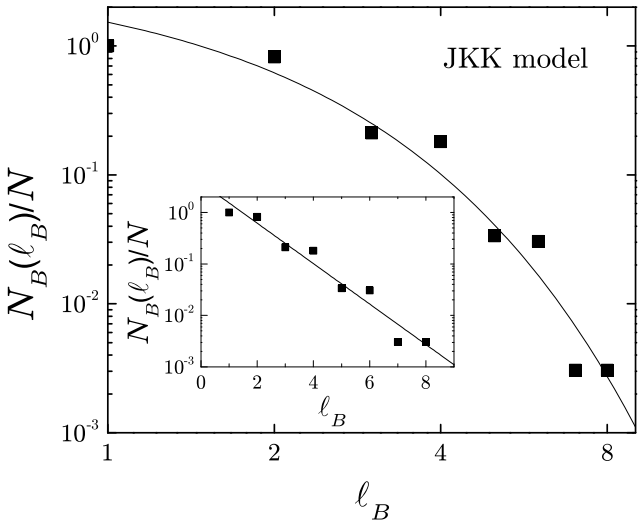


Figure 5.13: Lack of fractality in the model of Jung, Kim, and Kahng (JKK model)[59] which is an example of pseudo fractal models as discussed by Dorogovtsev and Mendes [58].

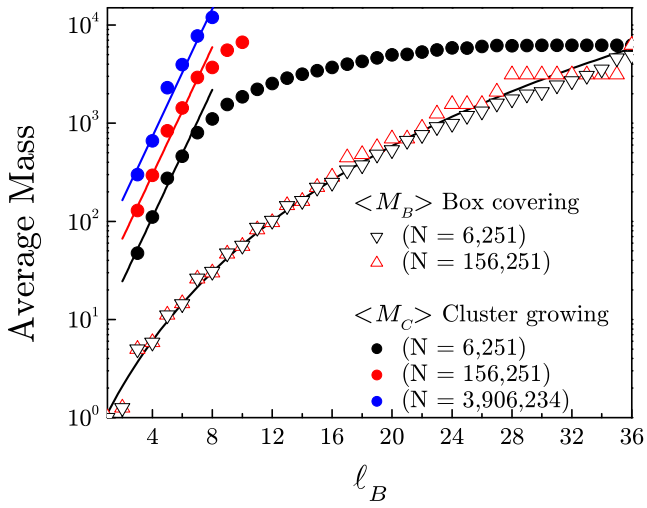


Figure 5.14: The scaling of $\langle M_B \rangle$ does not show finite size effects. The initial exponential dependence range of $\langle M_c \rangle$ increases as the size of the network increases.

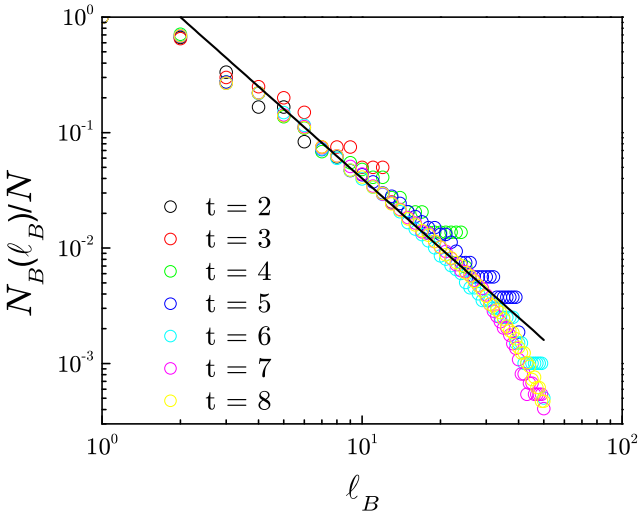


Figure 5.15: N_B vs ℓ_B for the model with $e = 0.5$ shows that the fractality still holds in the presence of random noise. The straight line gives the theoretical prediction of the model $d_B = 2$.

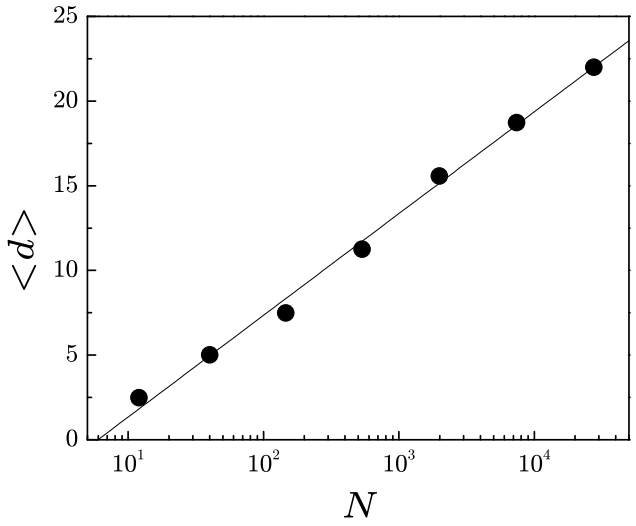


Figure 5.16: Average of the shortest path between two nodes as a function of the system size showing the global small world for the model ($e = 0.5$).

total number of links at time t, and p is a constant that controls the fraction of noise. We build a fractal, small-world and scale-free topology with parameters $m = 1.5$, $e = 0.5$, and add $p = 1\%$ random connections at each time step. Our analytical considerations predict a box dimension $d_B = 2$ in the absence of noise. The numerical simulation (see Fig. 5.15) shows that this prediction of d_B still fits well to the simulated data, except for a small deviation at large box sizes, i.e. the added noise appears as an approximate exponential tail at large distances. Interestingly, this method could be used to test the appearance of noise in real complex networks, or to asses the quality of the data in, for instance, protein interaction networks obtained by yeast two-hybrid methods which are known to suffer from many false positives.

Most interestingly, the addition of noise leads to the small-world effect at the

global level. In principle the existence of fractality seems to be at odds with the small-world effect. Fractality implies a power-law dependence on the distance, while the small-world effect implies an exponential dependence [38]. In Fig. 5.14 we show how the combination of Mode I and Mode II of growth leads to the global fractal property and the local small-world effect. In Fig. 5.16 we show that by adding a small fraction of short-cuts in a fractal complex network, we reproduce also the small-world effect at the global level. Using the algorithm explained above we add noise to the system and we find that the average distance $< d >$ over all pairs of vertices is

$$< d > \sim 2.61 \ln N, \tag{5.10}$$

(Fig. 5.16), indicating that the fractal model also predicts the global small-world. We notice that the fraction of short cuts needed to obtain the global small-world is very small, around 1%.

5.8 Resilience of fractal networks under intentional attack

To compare the stability of fractal and non-fractal networks under intentional attack (by removing hubs one by one from the largest to the smallest one), we generate two networks with $e = 0$ and $e = 1$. In general the threshold of collapse under attack depends on several parameters and not only on the correlated properties of the network. Since we wish to asses only the effects of anticorrelation for the vulnerability of the network, we set all the other parameters to be equal.

Thus, we use the same N, $< k >$, γ, and also the same number of loops in the structure. For this purpose we consider the number of intraloops inside the boxes and the number of interloops between boxes.

In practice, inside the box there are $mk(t)$ newly generated nodes. We add ymk extra links between them to generate triangles (y is a given constant) to obtain loops inside the boxes. For this case, we can rewrite $n = 2(1 + y)m + 1$, and the clustering coefficient $C(k) = (2ym/s)k^{-1}$. Thus, this kind of loops give rise to the known scaling of the clustering coefficient with k [50].

Another type of loops appears when more than one link connects two boxes (interloops). We find empirically that these kind of loops are also arranged in a self-similar way and are characterized by a new scaling exponent. In the framework of the minimal model, this type of loops can be introduced by adding x number of links between boxes at each time step, instead of keeping one link between boxes. These links could be of type Mode II (i.e., links between non-hubs), or otherwise could be between a hub from one box to a non-hub in the other node. In fact, this last mode of growth is a third mode that can be considered in the minimal model. We have not included it so far for simplicity, since it does not give rise to any new result. In general this mode could be thought of as a modified Mode II, and does not change the general conclusions of this study.

Combining the loop structure inside the boxes (intraloops characterized by y) and between boxes (interloops characterized by x) we obtain a general formula for the average degree $< k >= 2(1 + y) + (x - 1)/m$. In the case of the minimal tree structure discussed in the main text we have $y = 0$ and $x = 1$, which leads to $< k >= 2$, consistent with our previous arguments.

Chapter 6

Network Dynamics

6.1 Introduction

Transport in complex networks is a problem of much interest in many aspects of biology, sociology and other disciplines. For example, the study of metabolic fluxes in organisms is crucial for a deeper understanding of how the cell carries its metabolic cycle [60]. The use of metabolic flux analysis can provide important cellular physiological characteristics using the network stoichiometry and predict optimal flux distributions that satisfy a defined metabolic objective. Similarly, information flow between the molecules of a biological network provides insight for both the network structure and the functions performed by the network. Such an example is the concept of the 'diffusion distance' in a protein-protein interaction network which is used to predict possible interactions between proteins, simply by studying diffusion in the existing network [61]. In food webs, energy transfer between different levels of the web is crucial for the organism survival, while spread-

ing of a disease between different organisms may affect the regular operation of the equilibrated system. Moreover, applications of transport in complex networks extend to a plethora of other systems, such as, for instance, transport of information in the Internet, spreading of diseases and/or rumours in social networks, etc. Despite its significance, the laws of transport in such a complex substrate are yet unclear compared to transport in random media [62]. This is due to the complexity added by the heterogeneous degree distribution in such networks.

We study transport in real-world biological networks and via a model, which possess both self-similar properties and the scale-free character in their degree distribution. We explain our results with theoretical arguments and simulation analysis. We use approaches from renormalization theory in statistical physics that enable us to exploit the self-similar characteristics of the fractal networks and develop a scaling theory of transport, which we use to address the effects of the modularity and the degree inhomogeneity of the substrate.

Due to the existence of a broad degree distribution, transport on a network is different when it is between two hubs with a large number of connections k or between low-degree nodes [63]. We therefore characterize the transport coefficients by their explicit dependence on k_1, k_2, and ℓ, where k_1 and k_2 denote the degree of two nodes ($k_1 > k_2$), separated by a distance ℓ (distance is measured by the minimum number of links, i.e. it is the chemical distance). We study the diffusion time $T(\ell; k_1, k_2)$ and the resistance $R(\ell; k_1, k_2)$ between any two nodes in the system. The dependence on k_1 and k_2 is not significant in homogeneous systems, but is important in networks where the node degree spans a wide range of values, such as in biological complex networks. In fact, this dependence is critical for many other

properties as has been already shown for e.g. fractality, where traditional methods of measuring the fractal dimension may fail because they do not take into account this inhomogeneity [39].

Modularity is one of the most important aspect of these networks with direct implications to transport properties. Here, we quantify the modular character of complex networks according to our box-covering algorithm and reveal a connection between modularity and flow. Our results are consistent with recent experiments and metabolic flux studies, and provide a theoretical framework to analyze transport in a wide variety of network systems.

6.1.1 Metabolism modelling

In metabolism modelling, there exist three main approaches [64]. (i) The most detailed analysis includes dynamic *mechanism-based* models [65], but in general it is very difficult to incorporate experimental values for the needed kinetic parameters. (ii) In the second approach one simplifies the above models, and calculates the fluxes in a metabolic network via flux balance analysis, which includes a family of static *constraint-based* models [66]. The limiting factor in this analysis is that the problem is underconstrained (the number of unknown parameters, i.e. the fluxes, is larger than the number of metabolite conservation equations) and cannot be solved uniquely. (iii) Finally, a third approach that is widely used in metabolism modelling is to ignore stoichiometry, and focus only on the metabolites interactions without any thermodynamic aspects, which leads to *interaction-based* models [67], i.e. undirected networks where a link connects two nodes that participate in a metabolic reaction. In this paper we follow this third approach and we

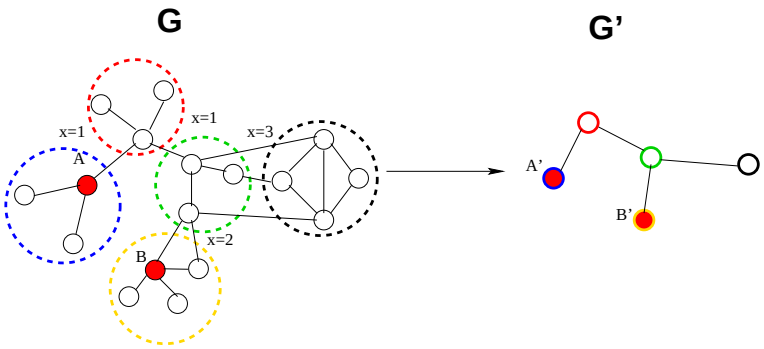


Figure 6.1: Example of a network G that undergoes renormalization to a network G'.

use this interaction-based network to study transport on this network, by drawing an analogy between the metabolic network and a resistance network. If we represent the metabolites as nodes that are linked through electrical resistances and the current flow represents the flux we can solve this problem without additional constraints, and this solution may shed additional light on the involved processes. The advantage of our approach is that it can isolate the topological effect and we can address a broader aspect of transport in biological networks, such as whether the observed flux inhomogeneity is because of the network topology itself or due to the adopted flux constraints. Moreover, this approach enables us to carry similar studies for diffusion on such networks.

6.2 Modularity, diffusion and resistance

In our work, we focus on two different examples of biological networks, namely the *E.coli* metabolic network [67] and the yeast protein interaction network (PIN) [15]. We analyze the filtered yeast interactome developed by Han *et al.* which removes a large number of false positives in high-throughput yeast two hybrid methods

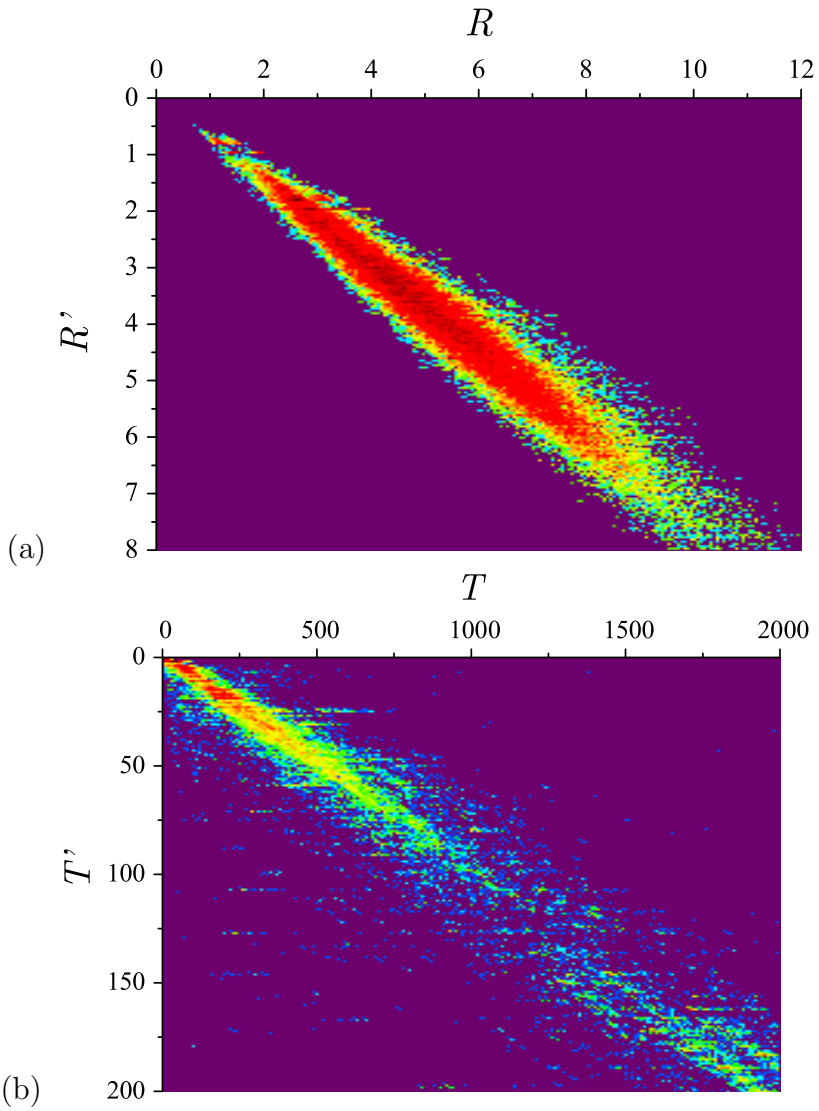


Figure 6.2: Typical behavior of the probability distributions for the resistance R vs R' and the diffusion time T vs T', respectively, which verifies that the ratios of these quantities during a renormalization stage are roughly constant for all pairs of nodes.

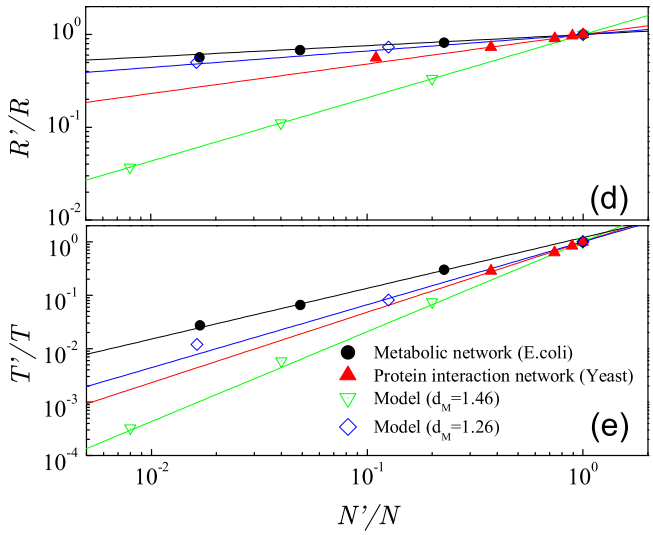


Figure 6.3: Average value of this ratio for the resistance R/R' and the diffusion time T/T', respectively, as measured for different ℓ_B values.

(see Supporting Information). Both networks have been shown to have fractal properties and can be covered with $N_B(\ell_B)$ non-overlapping boxes, where in each box the maximum distance between any two nodes is less than ℓ_B, the maximum distance in a box [39]. For a fractal network of N nodes, N_B follows a power-law dependence on ℓ_B,

$$N_B(\ell_B)/N \sim \ell_B^{-d_B} \tag{6.1}$$

and defines d_B as the fractal (or box) dimension of a network. These networks are also self-similar, i.e. their main properties, such as the degree distribution, remain invariant under a renormalization scheme where each box is replaced by a (super) node and links between boxes are transferred to the nodes of the renormalized network (see e.g. the example in Fig. 6.1 for a network G, tiled with $\ell_B = 3$

boxes, that yields the network G'). Many biological networks in the intermediate renormalized stages were shown to have similar properties as the original network.

This renormalization procedure also implies the presence of self-similar modularity in all length-scales, which is a central feature of these networks. The term modularity refers to the existence of sets of nodes whose links are connected preferably within this set rather than to the rest of the network. Thus, after tiling a network for a given value of ℓ_B, we introduce a measure of modularity for a network as

$$M(\ell_B) = \frac{1}{N_B} \sum_{i=1}^{N_B} \frac{L_i^{\rm in}}{L_i^{\rm out}} \,, \tag{6.2}$$

where $L_i^{\rm in}$ and $L_i^{\rm out}$ represent the number of links that start in a given box i and end either within or outside i, respectively. Large values of M correspond, thus, to a higher degree of modularity. Since the numerical value of $M(\ell_B)$ varies, though, with ℓ_B, a more reliable measure is the modularity fractal exponent d_M which we define through:

$$M(\ell_B) \sim \ell_B^{d_M} \,. \tag{6.3}$$

The value of $d_M = 1$ represents the borderline case that separates modular ($d_M > 1$) from random non-modular ($d_M < 1$) networks. For a lattice structure, the value of d_M is exactly equal to $d_M = 1$.

In Ref. [39] we had introduced a fractal network model where a network grows by adding m new offspring nodes to each existing network node, resulting in well-defined modules. In that version of the model, modules are connected to each other through $x = 1$ links, which leads to a tree structure. A generalization of this model (presented in detail in the Supporting Information) allows us to tune the

degree of modularity in the network by assigning a larger number of links $x > 1$ between modules. While for $x = 1$ all the modules are well-defined, increasing x leads to the presence of loops and to a progressive merging of modules, so that for large x values a node cannot be assigned unambiguously in a given module and modularity is destroyed. A straightforward analytical calculation in this case leads to (see Supporting Information)

$$d_M = \frac{\ln\left(2\frac{m}{x} + 1\right)}{\ln 3}\,. \tag{6.4}$$

In this paper, we use this value of d_M for the model and calculate d_M for real networks in order to study the influence of modularity on network transport.

6.3 Resistance and Diffusion

In general, the problem of transport is expressed in terms of $T(\ell; k_1, k_2)$ and $R(\ell; k_1, k_2)$, where T is the average first passage time needed by a random walker to cover the distance ℓ between two nodes with degrees k_1 and k_2, respectively, and R represents the resistance between these two nodes. For homogeneous systems (with very narrow degree distribution $P(k)$) such as lattices and regular fractals, there is no dependence on k_1 and k_2 and the average is only over the distance ℓ. One of the goals of this paper is to find the scaling of T and R in heterogeneous networks with a broad degree distribution and self-similar properties.

6.3.1 Resistance measurements

In order to measure the conductivity between two nodes A and B we consider that all links in the underlying network between any two neighbor nodes i and j have unit resistances $R_{ij} = 1$. By fixing the input current to $I_A = -1$ and the output to $I_B = 1$ we can solve the Kirchhoff equations and compute the voltages in the system. The measured resistance is then $R_{A\to B} = V_A - V_B$. However, due to the required inversion of the relevant matrices we are limited by the computer resources to networks of relatively small size, i.e. $N < 10^4$ nodes.

In principle, the magnitude of I_{ij} depends on the selection of the current input/output nodes. Upon closer inspection, though, we found that the distribution of the current magnitudes in the network links is not very sensitive to the selection of the current source and the current sink. Comparison of the result of averaging over one input and many outputs and over twenty different input and output pairs of nodes for the metabolic network showed that within statistical error these two results are almost identical.

6.3.2 Diffusion measurements

In many cases (and especially those including real-life networks) direct measurements of diffusion on complex networks exhibiting the small-world property may present significant difficulties, due to the limited time-range where diffusion takes place before settling quickly to a distance equal to the typical (very short) network diameter. The rising part of the mean-squared displacement as a function of the time is very small and reliable measurement of the diffusion exponent is very hard to do. Moreover, we need additional information in order to quantify the k and

ℓ dependence of the diffusion time between two nodes. For this purpose we used the peak of the first passage time distribution as a typical diffusion time $T(A, B)$ between two points A and B in the network. Since this quantity may be assymetric depending on which node we consider as origin, our diffusion time T represents the average of $(T_{A\to B} + T_{B\to A})/2$.

6.3.3 Scaling exponents of R and T

In the general case of a renormalizable network, T and R scale in the renormalized network G' (the primes always denote a quantity for the renormalized network) as

$$T'/T = \ell_B^{-d_w} \, , \; R'/R = \ell_B^{-\zeta} \, . \tag{6.5}$$

The exponents d_w and ζ are the random walk exponent and the resistance exponent, respectively. This equation is valid as an average of T and R over the entire system, applying for example in different generations when growing or renormalizing a fractal object. Thus, this relation holds true for both homogeneous and inhomogeneous systems.

For homogeneous systems, the above exponents d_w and ζ are related through the Einstein relation [62]

$$d_w = \zeta + d_B \, , \tag{6.6}$$

where d_B is the fractal dimension of the substrate on which diffusion takes place. This relation is a result of the fluctuation–dissipation theorem relating spontaneous fluctuations (diffusion) with transport (resistivity) and the underlying structure (dimensionality) [62]. Although the validity of this relation for scale-free networks

is not yet clear, our following analysis shows that it also applies for these systems, as well.

6.4 Renormalization and scaling theory

The renormalization procedure on self-similar biological networks provides a yet unexplored method for estimation of the dynamical exponents in these systems. Since such a network is left invariant after substituting all nodes in a box with a single node, we can calculate the transport properties on the networks during successive renormalization stages. With this method we can also study transport in biological networks before (R and T) and after renormalization (R' and T'). The results are presented in Fig. 6.2 for the yeast PIN, the E.coli metabolic network and the fractal network model with $d_M = 1.46$ (tree, highly modular) and $d_M = 1.26$ (network with loops and lower modularity). Figures 6.2 suggest a linear relation between R' and R (T' and T) for a given value of ℓ_B, so that the ratio R'/R (and T'/T) is almost constant for all boxes in the system for this ℓ_B value. For a different value of the box diameter ℓ_B this ratio is again constant for all boxes in the network, but assumes a different value. We can plot the values of this ratio as a function of the network size ratio N'/N for different values of ℓ_B (Figs. 6.3. The data indicate the existence of a power-law dependence, and a comparison with the model networks shows that the results are consistent with PIN exhibiting a more modular structure compared to the metabolic network.

Although in principle we can use our numerical results to directly calculate the exponents d_w and ζ through Eq. 6.5, this method is not practical because the

Table 6.1: Values of the exponents calculated from Fig. 6.3

Network	d_B	ζ/d_B	d_w/d_B
Metabolic network (E.coli)	3.3	0.08 ± 0.1	0.98 ± 0.1
PIN (yeast)	2.2	0.3 ± 0.04	1.3 ± 0.04
Model ($d_M = 1.46$, $m/x = 2/1$)	1.46	$\frac{\ln 3}{\ln 5}$	$1 + \frac{\ln 3}{\ln 5}$
Model ($d_M = 1.26$, $m/x = 3/2$)	1.89	$\frac{1}{3}\left(\frac{\ln 3}{\ln 2} - 1\right)$	$\frac{1}{3}\left(\frac{\ln 3}{\ln 2} + 2\right)$

variation of ℓ_B is very small. We can overcome this difficulty by using the system size N instead, where we combine Eqs. (6.1) and (6.5) to get

$$\frac{T'}{T} = \left(\frac{N'}{N}\right)^{d_w/d_B} , \frac{R'}{R} = \left(\frac{N'}{N}\right)^{\zeta/d_B} . \tag{6.7}$$

Thus, the slopes in Figs. 6.3 correspond to the exponent ratios d_w/d_B and ζ/d_B, respectively.

Notice also that the verification of the above equation through Fig. 6.3 validates the relation in Eq. (6.5) for inhomogeneous systems. The numerical values for the calculated exponents are shown in Table 6.1. These ratios are consistent in all cases, within statistical error, with the Einstein relation, Eq. (6.6).

Using these scaling arguments and the renormalization property of these networks we next predict the dependence of both R and T on the distance ℓ between two nodes and their corresponding degrees k_1 and k_2. After renormalization the network becomes smaller, so that both the degrees and the distances in the network decrease. A distance ℓ in G is scaled by a factor ℓ_B in G' so that $\ell' = \ell/\ell_B$, while in earlier work [39] it has been shown that the degree k of the largest hub in a box transforms to a degree k' for the renormalized box, where $k' = \ell_B^{-d_k} k$, and d_k is an exponent describing the scaling of the degree. According to the result of

Fig. 6.3 and Eq. (6.5) it follows,

$$R'(\ell'; k_1', k_2') = \ell_B^{-\zeta} R(\ell; k_1, k_2) \tag{6.8}$$

$$T'(\ell'; k_1', k_2') = \ell_B^{-d_w} T(\ell; k_1, k_2)\,. \tag{6.9}$$

Using dimensional analysis (see Supporting Information) we can show that

$$R(\ell; k_1, k_2) = k_2^{\zeta/d_k} f_R\left(\frac{\ell}{k_2^{1/d_k}}, \frac{k_1}{k_2}\right) \tag{6.10}$$

$$T(\ell; k_1, k_2) = k_2^{d_w/d_k} f_T\left(\frac{\ell}{k_2^{1/d_k}}, \frac{k_1}{k_2}\right)\,, \tag{6.11}$$

where $f_R()$ and $f_T()$ are undetermined functions. In the case of homogeneous networks where there is almost no k-dependence, these functions reduce to the forms $f_R(x, 1) = x^{\zeta}$, $f_T(x, 1) = x^{d_w}$, leading to the classical relations $R \sim \ell^{\zeta}$ and $T \sim \ell^{d_w}$.

The scaling in Eqs. (6.10) and (6.11) is supported by the numerical data collapse shown in Fig. 6.4. For the data collapse we used the values of the exponents ζ and d_w as obtained from the renormalization method above (Table 6.1) confirming the scaling in Eqs. (6.10) and (6.11).

The functions f_R and f_T introduced in Eqs. (6.10) and (6.11) have two arguments, so we first need to fix the ratio k_1/k_2 and in the plot (Fig. 6.4) we present different ratios using different symbols. We observe that the differences among varying ratios are small, so that the k_1/k_2 dependence in Eqs. (6.10) and (6.11)

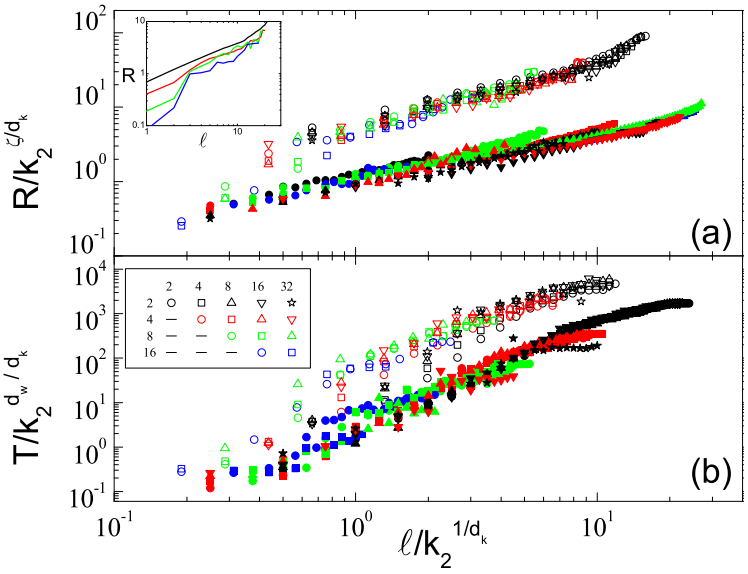


Figure 6.4: Rescaling of (a) the resistance and (b) the diffusion time according to Eqs. (6.10) and (6.11) for the protein interaction network of yeast (upper symbols) and the fractal network generation model (lower filled symbols). The data for PIN have been vertically shifted upwards by one decade for clarity. Different symbols correspond to different ratios k_1/k_2 and different colors denote a different value for k_1. Inset: Resistance R as a function of distance ℓ, before rescaling, for constant ratio $k_1/k_2 = 1$ and different k_1 values.

can be neglected,

$$R(\ell; k_1, k_2) = k_2^{\zeta/d_k} f_R\left(\frac{\ell}{k_2^{1/d_k}}\right) , \tag{6.12}$$

with an analogous form for $T(\ell; k_1, k_2)$. We are, thus, led to a simpler approximate form where the values of diffusion and resistance between two nodes depend only on the lowest degree node k_2 and the nodes distance ℓ. The same scaling form can be obtained as a function of k_1. However, it was proven in Ref. [68] for random networks that R depends on the smaller degree k_2, thus we arrive to Eq. (6.12). This equation is then a generalization to real networks of the result in Ref. [68] for random model networks where, without taking into account the distance, the resistance was found to depend solely on k_2.

6.5 Influence of modularity on transport

Modularity is a central feature of biological networks which contributes to a more efficient use of resources in the network, with yet unclear consequences for transport in these systems. In this section we use the fractal network model, which reproduces the main features of real networks, to better understand the influence of modularity on transport.

A direct calculation of ζ for the fractal network model is as follows. We consider the growth model where the distance ℓ between two nodes increases by a factor of 3, i.e. $\ell' = 3\ell$, and the resistance R between two neighbor nodes in G increases by a factor $3/x$ in G', since the linear distance between the two nodes has increased by a factor of 3 and there are x parallel paths connecting these nodes, i.e. $R' = 3R/x$. Combining these equations with Eq. (6.5) we find that the exponent ζ for this

model is given by:

$$\zeta = \frac{\ln(3/x)}{\ln 3} = 1 - \frac{\ln x}{\ln 3}\,. \tag{6.13}$$

Notice that for a tree structure ($x = 1$) we get $\zeta = 1$ as expected. This result is also an important step in linking statics with dynamics (a long standing problem in percolation theory [62]). Using only the value of x that describes how self-similar modules are connected to each other we can directly obtain the dynamic exponent ζ, i.e. how the structural property of modularity affects dynamics.

For the fractal dimension of model networks we already know that [39]

$$d_B = \frac{\ln(2m + x)}{\ln 3}\,. \tag{6.14}$$

If we assume that the Einstein relation in Eq. (6.6) is valid, then we can also calculate the value for the random walk exponent d_w:

$$d_w = \frac{\ln\left(\frac{6m}{x} + 3\right)}{\ln 3}\,. \tag{6.15}$$

A comparison of Eq. (6.15) with Eq. (6.4) yields:

$$d_w = 1 + d_M\,. \tag{6.16}$$

This is a very simple, yet powerful result. It manifests that for the fractal network model the degree of modularity directly affects the efficiency of transport and is the main feature that controls the type of diffusion.

The above relation, Eq. (6.16), is verified in Fig. 6.5 for the fractal network model. We generate a number of model networks where we vary both the number

of loop-forming links x and the number of offsprings m. For each pair of m and x we calculate numerically the exponent d_w from the slope of figures similar to Figs. 6.3 and use Eq. (6.4) for the value of d_M. The results are fully consistent with Eq. (6.16) and all the points lie on the predicted line. Sub-diffusion ($d_w > 2$) is observed for $d_M > 1$, in accordance with our observation that modularity slows down diffusion. On the contrary, for non-modular networks diffusion is accelerated remarkably ($d_w < 2$) which is also in agreeement with previous work on random networks. When $d_M = 1$ we recover classical diffusion ($d_w = 2$), even though the structure is still that of a scale-free network. For the biological networks, we find that the PIN network follows very closely the proposed scaling relation while the metabolic does not. This indicates that the model captures very well the modular structure of the PIN, while more structure is found in the metabolic compared to the above model.

6.6 Flow distribution across the network

In our scaling theory above we derived results for the average values of the current flowing in a complex network. The inherent inhomogeneity and modularity of biological networks is expected, though, to strongly influence the distribution of flow throughout the network. Using flux balance analysis, it was recently shown that the distribution of fluxes in the metabolic network is highly uneven and a small number of reactions have the largest contribution to the overall metabolic flux activity [60]. To study the influence of the complex substrate on the flow distribution we calculate the probability $P(I)$ of current I_{ij} flowing between all

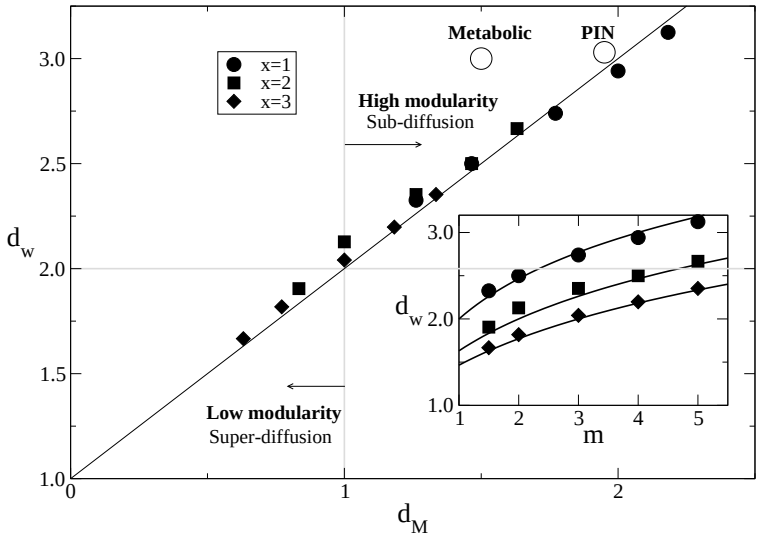


Figure 6.5: Comparison of the random walk exponent d_w extracted numerically (symbols) with the theoretical prediction (Eq. (6.16), line) vs the modularity exponent d_M, for different values of m and x. Open circles correspond to the result for the PIN and metabolic networks. Inset: Direct (unscaled) numerical calculation of d_w as a function of m, for varying x values (shown in the plot).

two neighboring nodes i and j.

The probability distribution $P(I)$ for the magnitude of the current I across a link of the metabolic network decays according to a power law (see Fig. 6.6). The form of the curve and the exponent in the range 1.0-1.5 of the decay is very similar to those found in previous studies of the metabolic flux, both experimental [69] and theoretical [60]. The decay suggests that only a small fraction of link carries high current. For the yeast PIN the distribution is even broader, and its form is different from the metabolic network. The variation of $P(I)$ is smaller in PIN compared to the metabolic network, meaning that in PIN there is a larger number of important links that carry large currents. The self-similar character of the biological networks is also verified for the distribution $P(I)$, as well, which

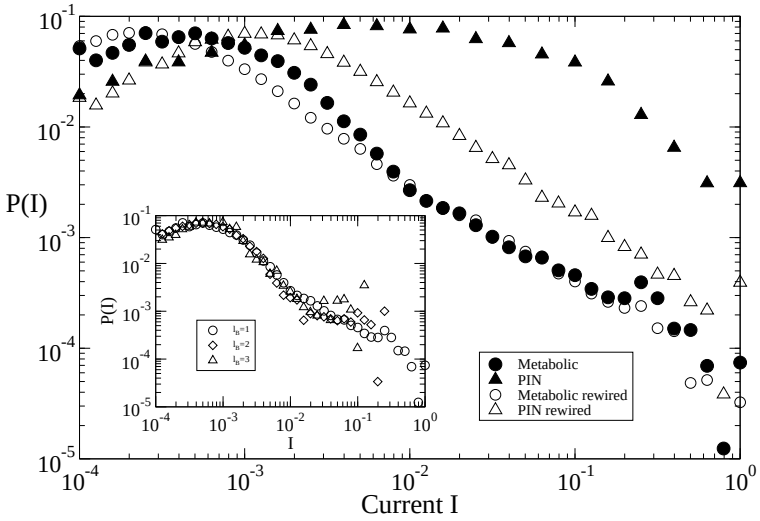


Figure 6.6: Probability distribution $P(I)$ of current magnitudes I flowing through the links in PIN (solid triangles) and metabolic networks (solid circles). Empty symbols are the corresponding results for the randomly rewired networks. Inset: Invariance of $P(I)$ for the metabolic network under renormalization with different ℓ_B values.

remains invariant after renormalizing the network, as seen in the inset of Fig. 6.6.

Next we explore the connection between $P(I)$ and topology. The information contained in $P(I)$ can be better understood if we compare these results with a surrogate random case. The random case is obtained by rewiring the original networks, preserving the degree distribution $P(k)$, but destroying any correlations between neighboring nodes. Thus, we remove all traces of the initial network organization. The distribution $P(I)$ for the metabolic network remains almost the same under this rewiring, indicating that, despite the modular character of the network, the original structure behaves similar to uncorrelated networks, since the degree correlations do not affect $P(I)$ in the metabolic network. In contrast, the distribution $P(I)$ for the rewired PIN is very different than the original distribution and is similar to that of the metabolic network. We can, thus, conclude that

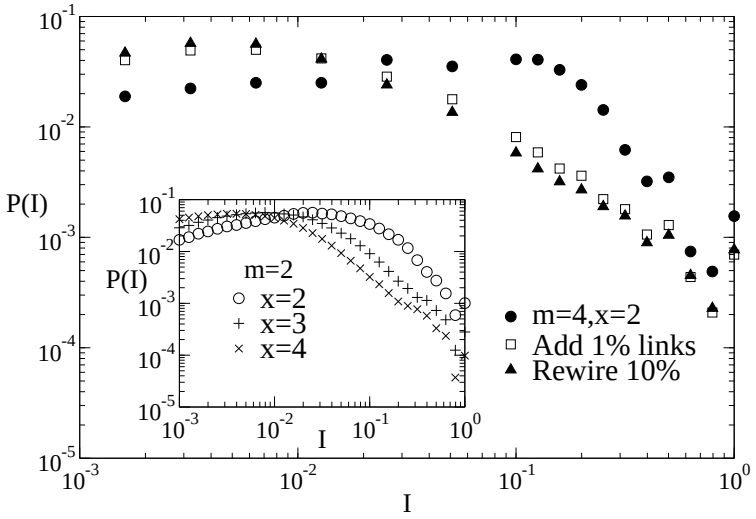


Figure 6.7: Probability distribution $P(I)$ for the fractal model before and after randomly adding 1% of links or rewiring 10% of the network. Inset: $P(I)$ for the fractal model with varying d_M values, where $m = 2$ and x varies from 2 to 4.

the original PIN has a much richer structure that deviates from the random case, corroborating thus the results on modularity from the previous section on diffusion.

The above results for the biological networks can be understood in terms of the fractal network model, where we can control the number of links that form loop structures in a network. In Fig. 6.7 we calculate $P(I)$ for the model network with $m = 4$ and $x = 2$, which is a highly modular structure ($d_M = 1.46$). The form of the $P(I)$ distribution is similar to that of the PIN, but if we add a small number of random links (or equivalently rewire a small part of the network) this distribution is significantly influenced in a similar way as observed in random PIN rewiring. This suggests again that in the case of PIN modularity is high. In the inset of this Figure we can also see that as x increases, i.e. more loops appear in the structure, the distribution has a longer tail, which shows that there is a smaller number of high-current links.

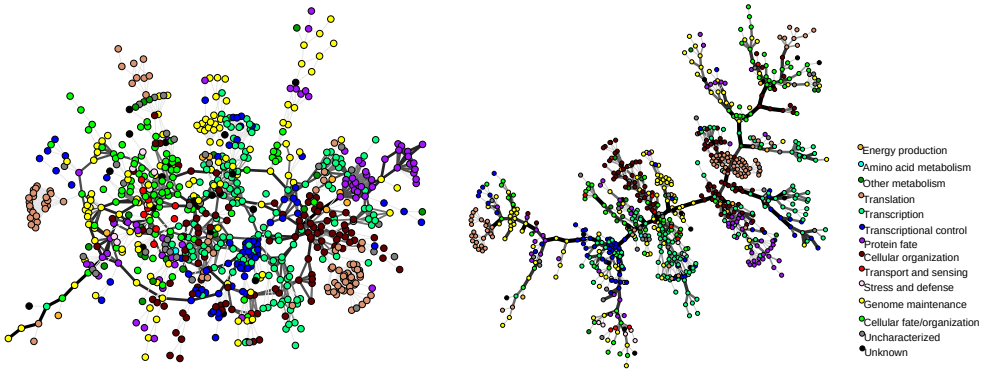


Figure 6.8: (a) Current flow through the links of the yeast PIN network, for one random selection of the two nodes acting as current input/output. (b) Minimum spanning tree for the PIN. The thickness of a link corresponds to the current flowing through this link. Different node colors correspond to different protein functions.

Since the number of added links in Fig. 6.7 is small, the modularity is preserved. We verified that the main factor that influences $P(I)$ is the number of loops in the network, rather than modularity itself, by fixing d_M and only varying x. In this case (described in Supporting Information) the $P(I)$ distribution is different for networks with the same d_M exponent. This can also be seen through Eq. (6.13), where the resistance exponent depends only on x.

Using the information of I_{ij} we can also construct the 'backbone' of the network in the form of the minimum spanning tree (MST) [70]. The importance of such a tree is that it identifies the substructure of the network that is dominant for transport. Starting from a completely empty network we insert links in decreasing order of current magnitude, provided that they do not form a loop. The resulting MST tree for the PIN is presented in Fig. 6.8, where the thickness of the links in the drawing increases logarithmically with increasing current through a link. The color of a node corresponds to the function performed by a protein in the network.

This tree (created solely on the base of current flowing through a network) exhibits a large degree of modularity where nodes that perform similar functions are close to each other. It is also possible through this construction to identify the most critical links in the network in terms of the largest current flowing through them.

Thus, the emerging picture from the above analysis for the PIN is one of a network with a strong backbone that carries most part of the flow combined with loops organized mainly within modules, so that flow through this backbone is not really influenced. This result highlights the strong modularity in the PIN structure. If a structure has a smaller degree of internal organization, as is the case in the metabolic network, then flow is more uniformly distributed.

Bibliography

[1] Mandelbrot, B. B. *The Fractal Geometry of Nature* (Freeman Co., San Francisco, 1982).

[2] Vicsek, T. *Fractal Growth Phenomena*, 2nd ed., Part IV (World Scientific, Singapore, 1992).

[3] Feder, J. *Fractals* (Plenum Press, New York, 1988).

[4] L. P. Kadanoff, *Statistical Physics: Static, Dynamics and Renormalization* (World Scientific, 2000).

[5] Stanley, H. E. *Introduction to Phase Transitions and Critical Phenomena* (Oxford University Press, Oxford, 1971).

[6] J. J. Binney, N. J. Dowrick, A. J. Fisher, and M. E. J. Newman. *The Theory of Critical Phenomena: An Introduction to the Renormalization Group* (Oxford University Press, Oxford, 1992).

[7] R. Albert and A.-L. Barabási, Rev. of Mod. Phys. **74**, 47 (2002); A.-L. Barabási, *Linked: How Everything Is Connected to Everything Else and What It Means*, (Plume, 2003); M. E. J. Newman, SIAM Review, **45**, 167 (2003);

S. N. Dorogovtsev, J. F. F. Mendes, Advances in Physics **51** 1079 (2002); S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW*, (Oxford University Press, Oxford, 2003); S. Bornholdt and H. G. Schuster, *Handbook of Graphs and Networks*, (Wiley-VCH, Berlin, 2003); R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet*, (Cambridge University Press, Cambridge, UK, 2004); Amaral, L. A. N. & Ottino, J. M. Complex networks - augmenting the framework for the study of complex systems. *Eur. Phys. J. B* **38**, 147-162 (2004).

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

[9] Milgram, S. *Psychol. Today* **2**, 60 (1967).

[10] Erdös, P. & Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17-61 (1960).

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

[12] Watts, D. J. & Strogatz, S. H. Collective dynamics of "small-world" networks. *Nature* **393**, 440-442 (1998).

[13] Faloutsos, M., Faloutsos, P. & Faloutsos, C. *Computer Communications Review* **29**, 251-262 (1999).

[14] J. P. Bagrow, E. M. Bollt, J. D. Skufca, D. ben-Avraham arXiv:cond-mat/0703470v1 [cond-mat.dis-nn]

[15] Han, J.-D. J., *et al. Nature* **430**, 88-93 (2004).

[16] Bunde, A. & Havlin, S. *Fractals and Disordered Systems*, edited by A. Bunde and S. Havlin, 2nd edition (Springer-Verlag, Heidelberg, 1996).

[17] D. ben-Avraham and S. Havlin, *Diffusion and Reactions in Fractals and Disordered Systems*, (Cambridge University Press, Cambridge, 2000).

[18] http://www.nd.edu/∼networks

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

[20] Xenarios, I. *et al.* DIP: the database of interacting proteins. *Nucleic Acids Res.* **28**, 289-291 (2000).

[21] *Database of Interacting Proteins* (DIP). http://dip.doe-mbi.ucla.edu

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

[23] http://igweb.integratedgenomics.com/IGwit

[24] Albert R. & Barabási, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys* **74**, 47-97 (2002).

[25] www.isi.edu/scan/scan.html

[26] Burch, H. & Cheswick, W. Mapping and Visualizing the Internet. *IEEE Computer* **32**, 4 (1999). http://research.lumeta.com/ches/map

[27] Giot, L. *et al.* A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727-1736 (2003).

[28] Rain, J.-C. *et al.* The protein-protein interaction map of *Helicobacter pylori*. *Nature* **409**, 211-215 (2001).

[29] Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623-627 (2000).

[30] Li, S. *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540-543 (2004).

[31] Peitgen H O, Jurgens H, and Saupe D 1993 *Chaos and Fractals: New Frontiers of Science* (Springer).

[32] Feder, J. *Fractals* (Plenum Press, New York, 1988).

[33] Bunde A and Havlin S (Eds.) 1995 *Fractals in Science* (Berlin: Springer-Verlag)

[34] Garey M. and Johnson D 1979 *Computers and Intractability; A Guide to the Theory of NP-Completeness* (New York: W.H. Freeman)

[35] Christofides N 1971 *Computer J.* **14** 38

[36] Wilf H 1984 *Info. Proc. Lett.* **18** 119

[37] Cormen T H, Leiserson C E, Rivest R L and Stein C 2001 *Introduction to Algorithms* (MIT Press)

[38] Song, C., Havlin, S., Makse, H. A. Self-similarity of complex networks. *Nature* **433**, 392-395 (2005).

[39] Song C, Havlin S and Makse H A 2006 *Nature Physics* **2** 275

[40] Kim J S, Goh K I, Salvi G, Oh E, Kahng B and Kim D 2006 *Preprint* cond-mat/0605324

[41] Song S, Rozenfeld, H and Makse H 2007 unpublished.

[42] Braunstein L A, Buldyrev S V, Cohen R, Havlin S, and Stanley H E 2003 *Phys. Rev. Lett.* **91** 168701

[43] Zhou W X, Jiang Z Q and Sornette D 2006 *Preprint* cond-mat/0605676

[44] Maslov, S., Sneppen, K. Specificity and Stability in Topology of Protein Networks. *Science* **296**, 910-913 (2002).

[45] Newman, M. E. J. Assortative Mixing in Networks. *Phys. Rev. Lett.* **89**, 208701 (2002).

[46] Pastor-Satorras, R., Vázquez, A., Vespegnani, A. Dynamical and Correlation Properties of the Internet. *Phys. Rev. Lett.* **87**, 258701 (2001).

[47] V. Colizza, A. Flammini, M.A. Serrano, and A. Vespignani, *Nature Physics* **2**, 110 (2006).

[48] M. Molloy and B.A. Reed, *Random Struct. Algorithms* **6** 161 (1995).

[49] C. Song, L. K. Gallos, S. Havlin, and H. A. Makse, *J. Stat. Mech.* P03006 (2007).

[50] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., Barabasi, A.-L. Hierarchical Organization of Modularity in Metabolic Networks. *Science* **297**, 1551-1555 (2002).

[51] M. Catanzaro, M. Boguna, and R. Pastor-Satorras, *Phys. Rev. E* **71** 027103 (2005).

[52] A. Vazquez, R. Pastor-Satorras, and A. Vespignani, cond-mat/0206084.

[53] Albert, R., Jeong, H., Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378-382 (2000).

[54] Cohen, R., Erez, K., ben-Avraham, D., Havlin, S. Resilience of the Internet to Random Breakdowns. *Phys. Rev. Lett.* **85**, 4626-4628 (2000).

[55] van Kampen, N. G. *Stochastic Processes in Physics and Chemistry* (North Holland, Amsterdam, 1981).

[56] Palla, G., Derenyi, I., Farkas, I., Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814-818 (2005).

[57] Kitano, H. Systems Biology: A Brief Overview. *Science* **295**, 1662-1664 (2002).

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

[59] Jung, S., Kim, S., Kahng, B. Geometric fractal growth model for scale-free networks. *Phys Rev E* **65**, 056101 (2002).

[60] Almaas, E., Kovacs, B., Vicsek, T., Oltvai, Z.N., & Barabasi, A.-L., (2004) *Nature* **427**, 839.

[61] Paccanaro, A., Trifonov, V., Yu, H., & Gerstein, M., (2005) *International Joint Conference on Neural Networks IJCNN, (Jul. 31-Aug. 4, Montreal, Canada).*

[62] Havlin, S., & ben-Avraham, D., (1987) *Adv. Phys.* **36**, 695.

[63] Rozenfeld, H.D., Havlin, S., & ben-Avraham, D., (2006) *preprint* cond-mat/0612330.

[64] Stelling, J., (2004) *Curr. Opin. Microbiol.* **7**, 513.

[65] Kitano, H. (2002) *Nature* **420**, 206.

[66] Reed, J.L., & Palsson, B.O., (2003) *J. Bacteriol.* **185**, 2692.

[67] Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., & Barabasi, A.-L., (2000) *Nature* **407**, 651.

[68] Lopez, E., Buldyrev, S., Havlin, S., & Stanley, H.E., (2005) *Phys. Rev. Lett.* **94**, 248701.

[69] Emmerling, M., et al. (2002) *J. Bacteriol.* **184**, 152.

[70] Kruskal, J. B., (1956) *Proc. Amer. Math. Soc.* **7**, 48.

[71] Hartwell, L. H., Hopfield, L. H., Leibler, S., Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47-C52 (1999).

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