SISSA ISAS

SCUOLA INTERNAZIONALE SUPERIORE DI STUDI AVANZATI
INTERNATIONAL SCHOOL FOR ADVANCED STUDIES

Statistical mechanics approach to complex networks: from abstract to biological networks

Thesis submitted for the degree of

Doctor Philosophiæ

Candidate:
Vittoria Colizza

Supervisor:
Prof. Amos Maritan

October, 18th 2004

Contents

Introduction

The study of networks can be traced back to the eighteenth century with the solution of the Königsberg bridge problem by Euler (1735), often referred to as the first example of a network theory application. Developments in this field have led to the mathematical formulation of graph theory in the twentieth century with the study of random graphs [1–4] and to the extensive studies in the framework of social sciences [5, 6].

Recent years, however, have witnessed a substantially increased interest in network research, characterized by a shift in the focus from the study of single node and edge properties of very small graphs by direct visualization, to the analysis of the statistical features of large-scale networks. The possibility of gathering data on a global scale, thanks to the progress in electronics and technology in general, has created an unprecedented opportunity to develop comprehensive explanations for phenomena occurring in diverse real systems, including information networks [7–14], technological networks [15–19], transportation systems [20–25], biological systems [26–42], social [5, 6, 43–45] and financial [46–48] systems and many others. Despite the abstract view of real-world systems, completely ignoring details associated to individuals or interactions between them, network representation still accounts for the fundamental aspect of complexity characterizing these systems and provide a general framework to study and uncover organizational principles determining the formation and evolution of various complex systems [49, 50].

With this purpose in mind, we have concerned the development of a class of models to reproduce the common peculiar features displayed by natural and artificial networks, based on optimality criteria [51].

In physics, optimality has been recognized to be a key factor in determining the physical behaviour of systems. For example, Snell's law in optics can be derived from Fermat's principle and the current pattern in an electrical network can be deduced by minimizing the dissipated energy. Moreover, it has been found that optimal patterns in the framework of transportation networks, obtained from the minimization of energy dissipation, closely resemble those observed in Nature and are able to explain the emergence of scaling properties in the study of several features [20, 52]. Analogous results are obtained also in biological systems, such as cardiovascular networks or plant vascular systems, where allometric scaling is shown to originate from the general features of networks under the assumption of maximum efficiency in transportation of nutrients, regardless of specific dynamical or geometric assumptions [24,25]. Motivated by these results, we have investigated the role of selective pressure in determining the topological features observed in natural and artificial complex networks. Several mechanisms have been suggested to reproduce the striking features displayed by real-world networks [53–64]. However, also the role of optimization and its interplay with dynamical mechanisms of network growth might be crucial in the evolution of complex networks, as e.g. it seems to occur in biological networks.

A second issue addressed in this thesis consists in the investigation of large-scale networks through a real-space renormalization group treatment. The theory of renormalization group approach has been recognized to be very powerful in understanding the critical behaviour of a large variety of physical systems [65–71]. Applications range from statistical physics and the broad framework of condensed matter theory, to chemical physics, quantum field theory, and others. Few works have made use of renormalization as a tool to study processes occurring on networks, such as percolation [72], or to study phase transition emerging in the Watts and Strogatz small-world model [73]. Our work is an attempt to investigate scale-free properties characterizing real-world complex networks, by means of renormalization group approach. The focus is on elucidating the most fundamental properties, while ignoring less important details of the networks.

Finally, we have focused our attention on the study of biological networks, in particular on a prominent example in this area, provided by protein-protein interaction networks [26–32] (PIN) of various organisms, which can be mathematically described as graphs whose nodes represent proteins and edges connect pairs of interacting proteins. It is worth to notice that all interaction data exhibit a non-trivial topological structure characterized by high levels of heterogeneity. Moreover, these properties are shared by many biological networks that appear to have recurrent architectural principles that might point to common organizational mechanisms. The resulting network topology is clearly interwoven with its biological significance and analysis in this direction is thoroughly discussed. Results of this analysis, trying to extract information encoded in the network about correlations between topology and protein functionality, are of extreme importance for the prediction of unannotated proteins. Indeed, despite the enormous progresses in genomic biology occurred recently, the determination of the biological function of a protein in a cell is a costly task which requires extensive biochemical analysis. A great amount of proteins for each completely sequenced genome is still functionally uncharacterized [74]. Moreover, the concept of function itself has radically changed, moving from the idea of individual task for each protein to a cooperative participation in a biological activity during cell cycle [75,76]. For this reason, the functional annotation of uncharacterized proteins represent a crucial point in post-genomic biology, and the search for reliable methods designed for the functional assignment is of extreme importance. Our work has focused on the development of two distinct bioinformatics methods for the prediction of protein functional assignments, on the basis of the network of interactions [77].

The thesis is organized as follows. In the first chapter we briefly provide an introduction to the main concepts of graph theory and basic metrics used to analyze complex networks. In the following, we review some of the most important models developed within this framework, starting with a class of static graph models - including random graphs, generalized random graphs and small world model - designed to reproduce the observed network structure, and then introducing a different approach which focus on the evolutionary dynamics leading to the observed topological properties. Following a path which was found to be successful in explaining network features in the framework of transportation networks, we then propose an alternative mechanism to reproduce scale-free behaviour and small world networks observed in Nature, by the introduction of optimality criteria and selection principles. Comparison with previous works in this area are outlined in the discussion.

In the second chapter we present a real-space renormalization group approach to complex networks. We investigate the critical features and scale-free properties of several different models, developed to describe different real-world networks, under renormalization. We discuss the results of the topological analysis of renormalized networks obtained with various decimation criteria and propose other renormalization procedures for future investigations.

In the third chapter we focus our attention on biological networks and provide an overview of the main results obtained in the study of protein-protein interaction networks by means of statistical physics methods and the theory of complex networks. We start by reviewing the most recent and important experimental techniques used to detect protein interactions and the specific own advantages and drawbacks as well as the biases intrinsically present in each of them. In the following we focus on the analysis of the yeast *Saccharomyces cerevisiae* protein interaction network obtained with different experimental methods. We discuss topological properties and the emergence of non-trivial structural organization, far from the random paradigm. We also investigate critical features exhibited by the networks with the renormalization group approach presented in chapter 2. We then report on works concerning the development of biological evolutionary models that reproduce the architecture and the peculiar features characterizing protein interaction networks, in an attempt to explain their origin for a deeper understanding of the structure and function of a living cell. Finally, we discuss several results relative to the analysis of correlations between the pattern of interactions and the biological function of proteins.

The fourth chapter presents two different bioinformatics methods for the prediction of functional annotation of proteins for which we have few or no functional information at all. Taking full advantage of the functional information encoded in the connectivity pattern of the whole protein interaction network, the two methods are able to provide functional prediction for the entire set of unclassified proteins. We discuss the results obtained in the tests performed to assess the statistical reliability and the robustness of function predictions provided by both methods.

Finally, in the last section we draw our conclusions and present the main perspectives of this study.

Chapter 1

Network Optimization

Many complex systems can be described by networks of interacting elements. Examples, covering very different contexts, include the Internet [15–17], the World Wide Web [10–14], phone call networks [78, 79], biological systems as protein [26–32] or metabolic networks [33–36], social networks [5, 6], networks in finance [46–48], citation networks [7–9] and scientific collaboration networks [43–45], ecological networks as food webs [37–39], neural networks [40–42], transportation networks such as airplane networks [18, 19], river networks [20–23], circulatory and vascular networks [24, 25] and many others.

In the last few years, a great effort has been performed in order to study and analyze large-scale properties of real-world networks, emerging from empirical observations. Network research has then focused the attention on the development of a variety of techniques and models to investigate the meaning and origin of statistical properties of networked systems.

Within this framework, we developed a model for network optimization. Our work is an attempt to follow a track which proved to be very fruitful in understanding river networks. Based on physical considerations, the concept of optimal channel networks (see paragraph 1.2.1) was shown to lead to patterns in striking accord with observational data [20, 52]. Following earlier work by Solé and collaborators [80, 81], Fabrikant *et al.* [82], and Mathias and Gopal [83] (see paragraph 1.2.2), we attempt here to explore some of the patterns that arise from the considerations of optimality.

Starting from a review of general aspects and recent developments in the field of complex networks (section 1.1), we introduce the concept of optimality (section 1.2), presenting successful results obtained in a related context (channel networks) and previous works in the framework of complex networks. In section 1.3, we propose a class of optimal models evolved by local rules and chosen according to global properties of the system [51]. Our focus in on elucidating the behaviour of networks resulting from optimality criteria.

1.1 General framework: complex networks

A network is a set ot n items, called *nodes* or *vertices* or *sites*, connected by *links* or *edges*; n is known as the *size* of the system. In the following, we will consider only undirected networks, i.e. networks in which no direction is associated to the links; also self-loops and multiple edges are ignored. A network can be mathematically described by its $(n \times n)$ *adjacency matrix* $\mathbf{A}$, where $A_{ij} = 1$ only if the two nodes i and j are connected, otherwise $A_{ij} = 0$. Since the network is undirected, the matrix $\mathbf{A}$ is symmetrical, i.e. $A_{ij} = A_{ji}$; self-loops are absent, thus $A_{ii} = 0$.

Here we review the main concepts and definitions used for the analysis of complex networks, followed by a presentation of models developed.

Degree. The degree k_i of a vertex i (also known as *connectivity* or *coordination* of a node) is the number of edges connected to that vertex, or, in other words, the number of its interacting partners. It is the most basic topological feature of a network and can be expressed in terms of the adjacency matrix, as $k_i = \sum_j A_{ij}$. The average degree $\langle k \rangle$ is simply

$$\langle k \rangle = \frac{2l}{n} \tag{1.1}$$

where l is the total number of links; indeed each link contributes to the degree of its two ends.

Degree distribution. An important network feature to be analyzed is the degree distribution $P(k)$, which represents the probability that a randomly chosen node has degree k; thus $\langle k \rangle = \sum_k kP(k)$. Real networks display a highly *heterogeneous* degree distribution, whose behaviour follows a power-law, considerably differing from purely random networks (as those defined in the following paragraph), characterized by a bell-shaped, exponentially bounded distributions. The *heavy-tailed* distribution implies that there exist a finite probability of finding a node with very high degree, much larger than the average value $\langle k \rangle$, thus exhibiting very large degree fluctuations. The average degree $\langle k \rangle$ does not represent a characteristic scale for the system, so that the network is said to exhibit a *scale-free* behaviour in its connectivity properties.

Clustering coefficient. It is a local measure of the interconnectedness of the nodes, representing the probability that the neighbors of a given node are also connected (in social networks it is the probability that a friend of my friend is also my friend). Considering a node i, the clustering coefficient C_i is defined as

$$C_i = \frac{2e_i}{k_i(k_i-1)} \tag{1.2}$$

where e_i is the number of links connecting neighbors of i out of the total number of possible connections $k_i(k_i-1)/2$ (for peripheral nodes having $k_i = 1$, C_i is taken equal to zero). The mean clustering coefficient ($\langle C \rangle$ or simply C in the following) is defined as the average over all nodes in the system:

$$C = \frac{1}{n}\sum_i C_i \, . \tag{1.3}$$

The number of edges e_i can be expressed in terms of the adjacency matrix $\mathbf{A}$:

$$e_i = \frac{1}{2}\sum_{v,w} A_{iv}A_{vw}A_{vw} \tag{1.4}$$

revealing that C_i is a measure of correlations in the adjacency matrix. Natural and artificial networks display very high clustering coefficients, a clear deviation from random graph behaviour.

The behaviour of $C(k)$ as a function of vertex degree, averaged over all nodes with degree k, has also been investigated, in order to characterize hierarchy and structural organization of networks [84–87]:

$$C(k) = \frac{1}{nP(k)}\sum_i C_i \delta_{k_i,k} \, . \tag{1.5}$$

A decreasing behaviour of $C(k)$ with k has been empirically observed in some real-world networks [87–90].

Shortest path length. The distance on a graph is defined as the minimum number of links needed to go from one node to another in the system; if the two nodes, i and j, belong to disconnected components, their distance d_{ij} is set to infinity. The maximum distance between any two nodes in the network is usually addressed as the network *diameter*. However, another definition involves what is instead commonly known as the average distance or average shortest path length L, defined as the average value of d_{ij} over all possible pairs of nodes:

$$L = \frac{2}{n(n-1)} \sum_{i<j} d_{ij} \,. \tag{1.6}$$

The property of having the average path length L considerably smaller respect to network size n, first observed in social networks of acquaintances and then verified in a large number of different networks, is known as the *small world effect*.

Degree correlations. The simplest case of degree correlations are those between the degrees of interacting vertices. Several different quantities can be used to measure degree correlations in a network. For example, one could compute the joint degree-degree distribution $P(k, k')$, representing the probability that a link has, at its ends, two nodes with degrees k and k' [30]. However, because of the poor statistics of empirical data and consequent large fluctuations in the computed values, it is better to introduce a more coarse, but less fluctuating measure, i.e. the average degree of the nearest neighbors of vertices with degree k [17, 84]:

$$k_{nn}(k) = \sum_{k'} k' P(k'|k) \tag{1.7}$$

with $P(k'|k)$ being the conditional probability that a vertex having degree k is connected to a vertex with degree k'. The behaviour of $k_{nn}(k)$ as a function of vertex degree k can be used to detect a property known as *assortativity* in social networks, occurring when k_{nn} is an increasing function of k. In particular, it has been shown that real-world networks can be classified in two distinct classes [91]: one showing an assortative behaviour, in the sense that high degree vertices tend to be connected (all social networks), and the other displaying *disassortative mixing*, implying that vertices with high degrees mostly have neighbors with low degrees (information networks, biological networks, technological networks). Random graphs show a k_{nn} behaviour independent of k.

The measure of degree-degree correlations can be reduced to a single number, by calculating the Pearson correlation coefficient of the degrees at either ends of an edge [91,92]. This number should assume positive values in case of assortative networks, while negative values for disassortative ones.

The features illustrated so far will be used to characterize the behaviour of a generic network in a renormalization process (chapter 2), and to thoroughly analyze and investigate the topology of a specific real-world network, the protein-protein network (section 3.2).

1.1.1 Random graph model

The theory of random graphs was introduced by Erdös and Rényi in the early 1960's [1–3]. According to their original formulation of the model, a random graph is defined as a set of n distinct vertices connected by l edges, randomly chosen out of the $n(n-1)/2$ possible pairs of nodes. There exist a total number $\binom{n(n-1)/2}{l}$ of equivalent and equiprobable random graphs, composed by n nodes and l undirected links.

An alternative definition of random graph is the binomial model. Starting with n distinct vertices, a random wiring is performed with probability p, denoted as connection probability - each of the $n(n-1)/2$ possible pairs of nodes is connected with probability p (see fig. 1.1). In this case, the number of edges is not fixed *a priori*, and the probability of having l edges is given by:

$$P_{n,l} = p^l(1-p)^{n(n-1)/2-l} \tag{1.8}$$

The definition of random graph in terms of the connection probability p is the one adopted in the following.

The main properties of random graphs derive from very simple considerations. The average degree of the Erdös-Rényi model can be easily evaluated, by noticing that the expectation value of the number of edges is $\langle l \rangle = n(n-1)p/2$. Thus, following eq. (1.1), we obtain:

$$\langle k \rangle_{rand} = \frac{2\langle l \rangle}{n} = (n-1)p \simeq np \qquad \text{for large } n. \tag{1.9}$$

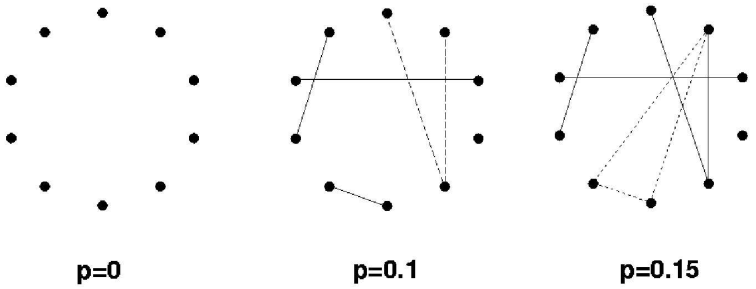


Figure 1.1: Schematic representation of a random graph model. Left: n isolated vertices ($p = 0$). Center and Right: two realizations corresponding to different values of the connection probability - respectively $p = 0.1$ and $p = 0.15$.

The degree distribution $P(k)$ is simply obtained from the binomial process underlying random graph generation. In fact, the probability that a node has degree k is equal to the probability that it is connected to other k vertices of the network (i.e. p^k), times the probability that it is not connected to the remaining $n - k - 1$ nodes (i.e. $(1-p)^{n-k-1}$). Hence,

$$P_{rand}(k) \;=\; \binom{n-k}{k}\, p^k\,(1-p)^{n-k-1} \tag{1.10}$$

which is approximated by a Poissonian distribution in the limit of large n:

$$P_{rand}(k) \;\simeq\; e^{-\langle k\rangle}\frac{\langle k\rangle^k}{k!} \tag{1.11}$$

The peculiar characteristic of random graph degree distribution is the exponential decay for large degrees, with very small degree fluctuations (see fig. 1.2).

The average clustering coefficient $\langle C_{rand}\rangle$ of the classical random graph is expressed as:

$$\langle C_{rand}\rangle \;=\; p \;=\; \frac{\langle k\rangle}{n}\,, \tag{1.12}$$

since the probability that two neighbours of a given node are connected is simply equal to the probability of two random nodes being linked, i.e. the connection

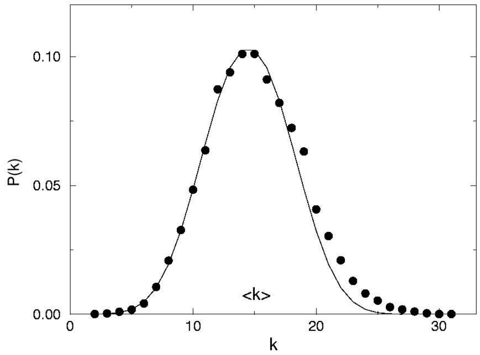


Figure 1.2: Degree distribution resulting from numerical simulations of a random graph with $n = 10^5$ nodes and connection probability $p = 0.0015$. The plot compares $P(k)$ with the expectation value of the Poisson distribution (eq. (1.11)): deviations are small.

probability p, because of the indipendence of wiring events. The average clustering coefficient thus decreases with network size, as n^{-1}, for fixed values of the average degree $\langle k \rangle$.

The structure of the random graph varies with the connection probability $p = \langle k \rangle / n$: for $p < n^{-1}$ (i.e. $\langle k \rangle < 1$) the network is composed of many disconnected subnetworks; for $p > n^{-1}$ (i.e. $\langle k \rangle > 1$) a giant connected component appears, with size $s \sim n$ (for a thorough discussion see [4]). For values $p > n^{-1}$, we can obtain an approximate expression for the average path length [50]:

$$L_{rand} \simeq \frac{\ln n}{\ln \langle k \rangle} \tag{1.13}$$

Since $L_{rand} \sim \ln n$, thus slowly increasing with network size, random graphs exhibit the small world behaviour observed in many real complex systems. However, for what concerns other features, they do not closely resemble networks in the real world. Indeed, they are characterized by a Poissonian degree distribution peaked around the average degree and have a low clustering coefficient which tends to zero in the limit of large network sizes, resulting inadequate for a good representation of real complex networks.

The random graph model can be easily generalized to incorporate arbitrary degree distributions [93, 94]; however, generalized random graphs cannot be used to investigate the origin of such distributions.

1.1.2 Small world model

Looking at the two extremes - a regular graph on one hand, being clusterized but having very large average distances, and a classical random graph on the other, displaying small world effect but low clustering - Watts and Strogatz [95–97] proposed a model, called *small world model*, which interpolates between the two graphs (see fig. 1.3).

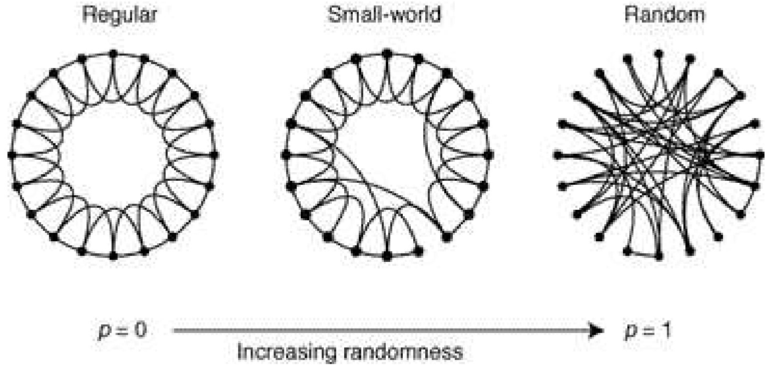


Figure 1.3: Construction of the small world model, which interpolates between a completely ordered graph ($p = 0$) and a classical random graph ($p = 1$), with the constraints of fixed numbers of nodes and edges. In the picture, an ordered ring lattice with $n = 20$ nodes and degree $k = 4$ is shown as the starting step of the algorithm (left, $p = 0$). By increasing p, an increasing number of edges is rewired, with the emergence of short-cuts in the system. For $p = 1$ all edges are rewired randomly, thus recovering a random network.

Starting from a completely ordered graph (for the sake of simplicity, a one-dimensional model, i.e. a ring lattice, is considered) composed of n nodes each connected to k nearest neighbors, each link of the system is randomly rewired according to probability p. The rewiring connects the starting node to a randomly chosen node of the network, avoiding self-loops and duplicate connections. The rewiring probability p is the parameter tuning the transition from ordered lattices to random graphs.

The key result reached by the model is represented by the behaviour it displays in terms of clustering coefficient and average path length, on varying the rewiring probability p. They are both shown in fig. 1.4 as a function of the parameter p. Starting with a $k-$regular network for $p = 0$, displaying high values of both $\langle C\rangle$ and $\langle L\rangle$, with increasing p the average clustering coefficient remains almost unchanged, while the average path length shows a dramatic decrease, reaching almost the value corresponding to a random graph. This transition is due to the emergence of short-cuts (see fig. 1.3) connecting nodes of the systems which otherwise would be far away in the network, thus considerably decreasing the average distance.

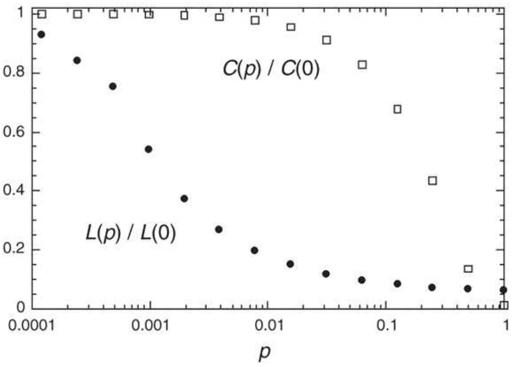


Figure 1.4: Average clustering coefficient and average shortest path length, normalized to corresponding values obtained for regular lattices, as a function of the rewiring probability p. The dramatic drop in $\langle L\rangle$ is due to the emergence of short-cuts in the system, and occurs when the clustering coefficient $\langle C\rangle$ remains almost constant.

Therefore there is a broad region in the parameter space in which networks with high clustering and very low typical distances (comparable to random ones) are obtained, in agreement with the characteristics observed in real systems; such networks are called *small world networks*. However, this model still misses some of the peculiar features emerging from empirical observations, leading to relatively homogeneous networks characterized by a Poissonian degree distribution.

1.1.3 Dynamical models

The models discussed so far belong to the class of *static models*, since networks are characterized by a fixed size n. Although they have provided a reference theoretical framework for a long time, they do not explain the origin of all the features encountered in natural and artificial complex networks. In this paragraph we will examine a class of models, based on network growth, whose aim is to capture and understand how these peculiar features emerge.

These models are based on what is now accepted as the possible explanation of the emergence of scale-free degree distributions in growing networks, i.e. the *preferential attachment* rule introduced by Barabási and Albert [53, 54], and related to the "rich-get-richer" idea formulated by Simon [98] in the 50's, later called *cumulative advantage* by Price [7] in his work to explain the appearance of power-law distributions in the network of scientific citations.

The Barabási-Albert (BA) model is based on two main ingredients - the growing nature of networks and a preferential attachment mechanism for which connections emerging from new nodes are more probably established towards more connected nodes ("rich-get-richer" phenomenon). Starting from m_0 nodes, the algorithm is as follows:

- *growth*: at each time step a new node enters the system;
- *linear preferential attachment*: the new node is connected to m ($m < m_0$) already existing nodes (i) with a probability which is linear in their connectivities (k_i):

$$\Pi_{BA}(k_i) = \frac{k_i}{\sum_i k_i} \tag{1.14}$$

The algorithm is repeated n times, resulting in a network characterized by a power-law degree distribution $P(k) \sim k^{-\gamma}$ with exponent $\gamma = 3$ (see fig. 1.5), and average degree $\langle k \rangle = 2m$.

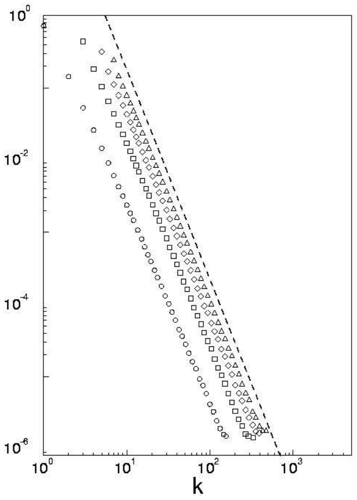


Figure 1.5: Degree ditribution obtained from numerical simulations of BA model with $n = 3 \cdot 10^5$ and different values of m and m_0 ($m = m_0 = 1, 3, 5, 7$). The slope of the dashed line is $\gamma = 2.9$.

The model can be solved exactly in the limit of large network size n, using a master-equation approach [53–56], thus confirming results obtained numerically. In fig. 1.6 results for the clustering coefficient and average path length, obtained from numerical simulations of BA model, are shown in comparison to the corresponding values displayed by a random graph. Recent analytical results [99, 100] indicate that the average path lentgh scales as

$$\langle L \rangle \sim \frac{\ln n}{\ln \ln n} \tag{1.15}$$

thus showing small world properties. However, as in random graphs, the clustering coefficient decreases with network sizes vanishing in the limit of infinite size.

The importance of the BA model lies in its ability to reproduce networks characterized by high heterogeneity in the degree distribution and small world effect. However, it is not able to capture other features, such as the behaviour displayed by the clustering coefficient, or higher order correlations (an issue which will be first addressed in chapter 2 and then thoroughly discussed in chapter 3 in the framework of protein interaction networks).

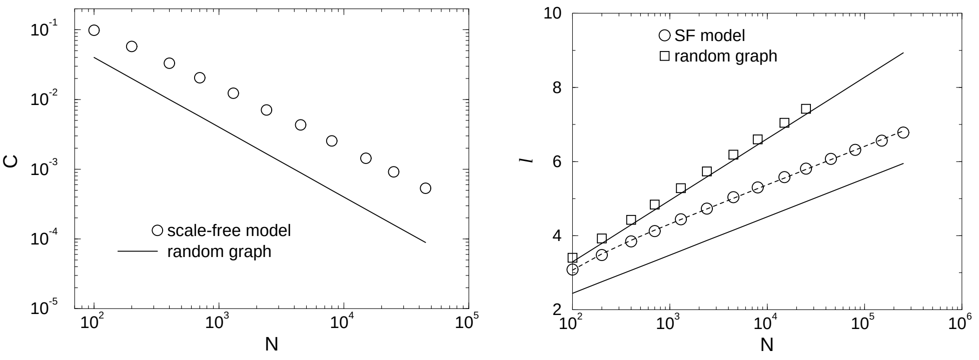


Figure 1.6: Left: clustering coefficient as a function of network size of BA model with $\langle k \rangle = 4$, compared with the clustering coefficient of a random graph, C_{rand}. Right: average path length as a function of network size of BA model with $\langle k \rangle = 4$, compared with the correpsonding values obtained in a random graph having same size and average degree.

Several generalizations of the model have been developed [56–63] in order to introduce more realistic mechanisms actually occurring in real processes and to extend variability of the power-law exponent. Preferential attachments different from linear ($\Pi(k) \sim k^{\alpha}$, $\alpha \neq 1$) were shown to result in networks not characterized by a scale-free behaviour [55, 57]. In order to address the absence of correlations between node degree and its age, as in the Internet, competition mechanisms have been introduced to enrich the growing dynamics of preferential attachment [61–63]. A stochastic parameter, called *fitness*, is associated to each node, representing all features, besides degree, that might contribute to node growth rate. The fitness of a node has been also introduced in another model [64] which belongs to a different class, being not based on preferential attachment related rules. Links are established between nodes which gain mutual advantage from the interaction, with fitness parameter embodying the intrinsic properties of each vertex. Networks with power-law degree distribution and connectivity correlations are obtained under certain circumstances (details will be discussed in the following chapter).

1.2 Optimal design

Progress in the theory of complex networks has focused on the development of the dynamical models discussed in the previous section. Only few works investigated the role of optimization in complex networks [80–83].

However, optimal design has been shown to play a crucial role in understanding the properties of transportation networks encountered in different contexts, such as branching structures in biology, whose scaling properties are obtained under the assumption of maximal efficience in transportation [24, 25], or drainage networks of river basins, whose scaling features emerge as the outcome of a minimization of energy dissipation [20, 52].

These results would support optimization as an alternative scenario for the explanation of peculiar features observed in complex networks.

1.2.1 River networks

Branching river networks are striking examples of natural fractal patterns. Experimental observations have shown the emergence of power-law behaviour in the probability distribution of several quantities describing the morphology of river basins, over a wide range of scales, despite the great diversities in geologic, lithologic, vegetational, climatic and hydrologic factors [20].

Experimental data on river networks are extracted from digital terrain maps, usually consisting of discretized elevation fileds z_i on a lattice, with i representing the pixel, i.e. unit area, on the lattice [20, 101]. A river network is represented as an oriented spanning tree, in which orientation of the links corresponds to drainage direction. A geomorphic quantity of interest is the total drainage area a_i associated to site i, defined as:

$$a_i = \sum_j w_{ji} a_j + r_i \tag{1.16}$$

where w_{ji} is an element of the adjacency matrix $\mathbf{W}$, with $w_{ji} = 1$ only if $j \to i$, otherwise it is equal to zero, and r_i represents the local injection at site i, commonly assumed to be homogeneous.

The distribution of total drainage areas a_i and of upstream lengths l_i are found to follow a power-law behaviour, with exponents in the narrow ranges $[1.40, 1.46]$

and $[1.67, 1.85]$, respectively [21, 22]. Moreover, scaling properties emerge when looking to other geomorphological quantities [25].

The explanation of the dynamical origin of the fractal character of river basins represents a fundamental task within this framework. A lattice model based on an energy minimization principle, was shown to reproduce, despite its simplicity, many features of natural river networks [52, 102, 103]. Optimal configurations, called optimal channel networks (OCNs), are obtained by minimizing a dissipated energy, expressed as

$$E = \sum_i k_i a_i^\gamma \tag{1.17}$$

where k_i characterizes the local soil properties and γ is defined by an empirical relationship between the local topographic drop in elevation, Δz_i, and the flow rate J_i, which is found to be proportional to the drainage area in the case of uniform rainfall in time and space ($J_i \sim a_i$):

$$\Delta z_i \sim a_i^{\gamma-1} \tag{1.18}$$

with a numerical value of $\gamma \sim 1/2$.

The shape of the cost function, i.e. its concavity or convexity, was shown to directly impact on the topology of optimal networks [20, 52, 104–106], leading to spanning, loopless structure configurations for the value of interest $\gamma \sim 1/2$. Indeed, it was proved [106] that when the exponent γ of an overall cost for the local transportation of material is smaller than 1, it is cheaper to send material from a given site to only one of its neighbors, rather than to more than one, thus leading to the emergence of tree structures.

OCNs have been thoroughly discussed and analyzed [104], through analytical and numerical approaches. Results in the scaling exponents obtained for the statistics of the global minimum were found to confirm analytical predictions, which differ from values measured in natural river networks. Statistical exponents characterizing local minima were instead found to be in completely agreement with those found in Nature (see, e.g. fig. 1.7), suggesting that real rivers do not explore the whole set of configurations during their evolution, often remaining trapped in some metastable configurations.

It is worth to mention that OCNs have been also used to address the question on whether a liquid material (likely water, given the current knowledge) ever

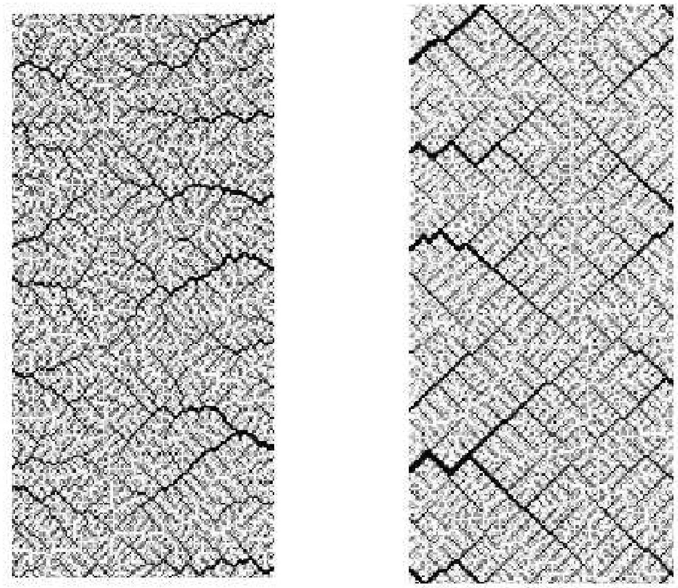


Figure 1.7: Examples of OCN configurations minimizing energy dissipation (eq. (1.17)). Left: local minimum showing features perfectly matching those observed in Nature. Right: global minimum, displaying too regular and ordered structure respect to irregularities observed in Nature.

flowed on Mars, by the analysis of the present martian landscape [23].

1.2.2 Complex networks

Optimization introduced in a model for network growth [82] has been suggested to represent a possible explanation of the degree distribution observed in the Internet topology. The simultaneous minimization of conflicting objectives, proposed by Fabrikant *et al.* [82], toghether with growing mechanisms, lead to power-law degree distributions for certain values of the parameters introduced.

Starting with a set of n points in the unit square, distributed uniformly at random, a tree is built by the introduction, at each time step i, of the node i which connect to one of the already existing nodes $j < i$, chosen with an optimization process. The function to be minimized is a linear combination of the

Euclidean distance d_{ij}^E between the two nodes, representing “last mile” costs, and of a measure h_j of the “centrality” of node j, representing operation costs due to communication delays:

$$f_i(j) = \alpha\, d_{ij}^E + h_j \tag{1.19}$$

where the “centrality” h_j might correspond to (i) the average graph distance from other nodes; (ii) the maximum graph distance from another node; (iii) the graph distance from a fixed centre of the tree. The parameter α tunes the relative importance of the two objectives, and is thought as a function of the final number n of points.

Resulting networks can be classified in three different types, according to the behaviour of the parameter α respect to n:

(1) a star network with $i = 0$ as its center, if α is less than a certain constant;

(2) a tree with an exponential degree distribution, if α grows at least as fast as $\sqrt{n}$;

(3) a tree with a power-law degree distribution, if α is in between the previous values.

In fig. 1.8 resulting cumulative degree distributions and associated trees of size $n = 10^5$ are shown, for values $\alpha = 4$ (top) and $\alpha = 20$ (bottom).

In the work by Mathias and Gopal [83], optimization has been introduced to investigate the origin of small world networks [97]. The model proposed is an attempt to understand the emergence of the small-world topology in networks where the physical distance is a criterion that cannot be ignored, such as, e.g., neural and transportation networks. Optimal structures arising as a consequence of a trade-off between maximal connectivity and minimal wiring are investigated. Networks considered in the model are composed of vertices arranged symmetrically along a ring, as in [97]. The size n of the system, as well as the number of links, is fixed during optimization; since physical distance is taken into account, the nodes are equally spaced on the ring and maintain their positions. The initial configuration is a $k-$regular network, in analogy with [97]. The energy function E to minimize is defined as a linear combination of the wiring cost W and the average degree of separation between the nodes, L:

$$E = \lambda L + (1 - \lambda) W \tag{1.20}$$

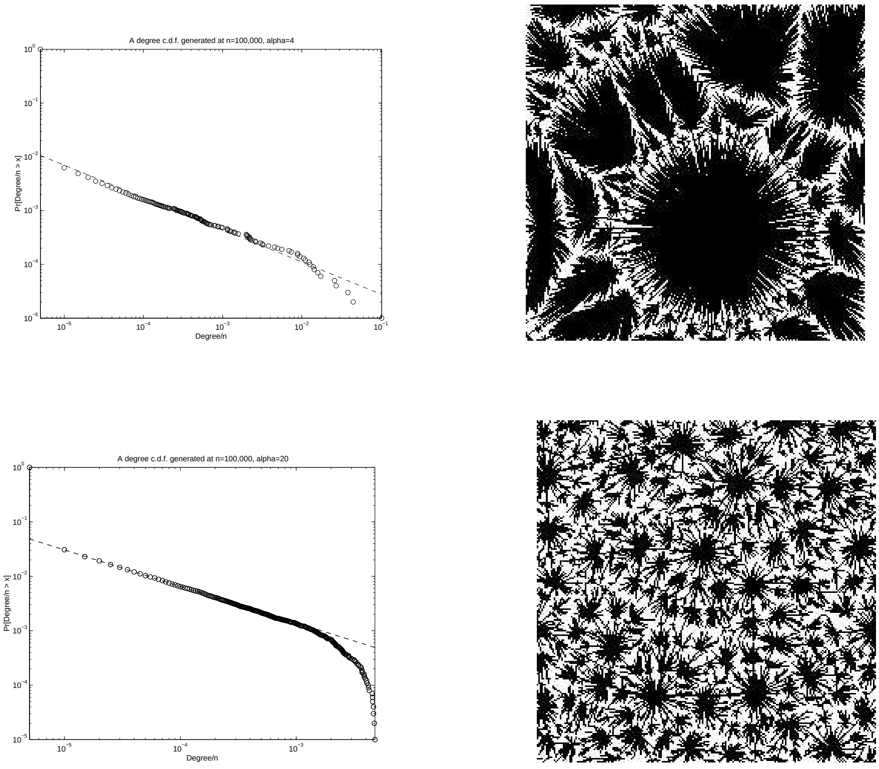


Figure 1.8: Cumulative degree distributions and associated trees of size $n = 10^5$ generated for $\alpha = 4$ (top) and $\alpha = 20$ (bottom).

where L is the average path length, as defined in eq. (1.6), normalized to the value $L(0)$ obtained in a regular network with degree k, and the cost of wiring is defined as the sum of the Euclidean distances between any pair of connected nodes:

$$W = \sum_{l_{ij}} \sqrt{(x - x_i)^2 + (y - y_i)^2} \tag{1.21}$$

with (x_i, y_i) being the fixed oordinates of the vertex i on the ring lattice.

As expected, for $\lambda = 0$ a regular graph with a high average path length ($L \sim n$) results, since the system tries to minimize the cost of wiring edges. On the other extrem, at $\lambda = 1$ the optimization results in a near random network ($L \sim \ln n$), since only the characteristic path length is to be minimized. Some examples of optimal networks obtained for different values of λ are shown in fig. 1.9, where toghether with the ring lattice representations (left), also corresponding 2-d graphs are reported (right).

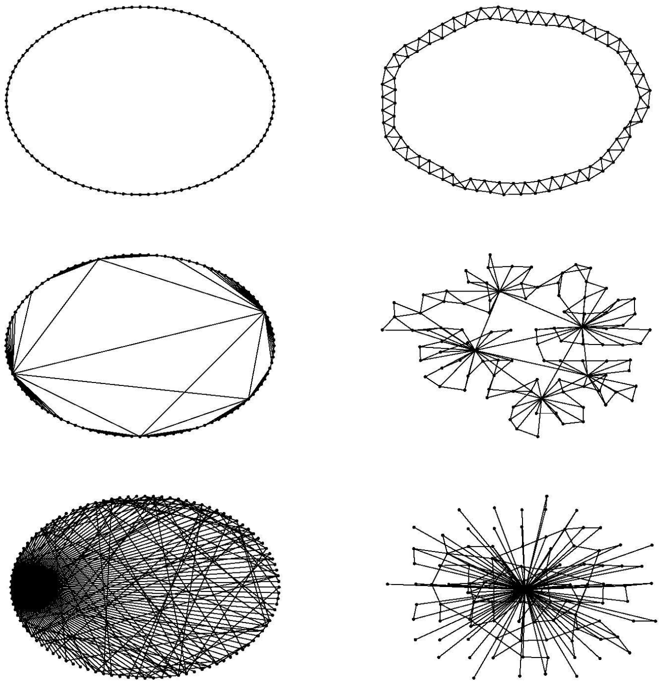


Figure 1.9: Left: Ring lattices corresponding to values $n = 100$, $k = 4$, as the parameter λ is varied over the $[0,1]$ range: $\lambda = 0$ (top), $\lambda = 0.5$ (center), $\lambda = 1$ (bottom). Right: The same networks are displayed as 2d-graphs using a graph generator with a spring embedder (from [83]).

Both the previous works take into account Euclidean distance between the nodes of a spatial network, which generally does not play any relevant role in real complex networks. In a model introduced by Solé and collaborators [80, 81], it was shown that the minimization of a linear combination of average degree and average graph distance, regardless of the physical distance, can lead to the emergence of a truncated power-law in the degree distribution.

Networks considered have a fixed number of nodes n, but no constraints on the number of links l. Starting from a classical random graph [1, 4], in which two given nodes are connected with some probability p, the energy function to be minimized in the optimization algorithm is defined as follows:

$$E(\lambda) = \lambda d + (1 - \lambda)\rho \tag{1.22}$$

where λ is the parameter controlling the relative importance of the two contributions: the network density ρ, defined as $\langle k \rangle/(n-1)$ and the normalized distance d, i.e. the average path length normalized to the maximum value D^{max} it can assume in a connected network ($D^{max} = (n+1)/3$).

As expected, the two extreme values $\lambda = 0$ and $\lambda = 1$ lead to, respectively, Poissonian and completely connected networks (i.e. cliques). Varying the value of λ, three types of optimal networks are found (see fig. 1.10):

(A) an exponential-like network;

(B) a network with truncated power-law in the degree distribution;

(C) a star-like network

In fig. 1.10 it is also reported the behaviour of the degree entropy $H(\lambda)$, defined as:

$$H(\lambda) = -\sum_{k=1}^{n-1} P_k \ln P_k \tag{1.23}$$

where P_k is the fraction of nodes having degree k.

1.3 The model

Our focus here is the proposal and analysis of a class of models in which the key selection criterion for network topology is optimality [51]. The goal is thus to understand the topology of networks which minimize a physically motivated cost function. Strikingly, we find a variety of distinct topologies and novel phase transitions between them on varying the number of links per node.

Suppose that some type of information has to be communicated between pairs of nodes of the network [107]. It is plausible that besides the average distance

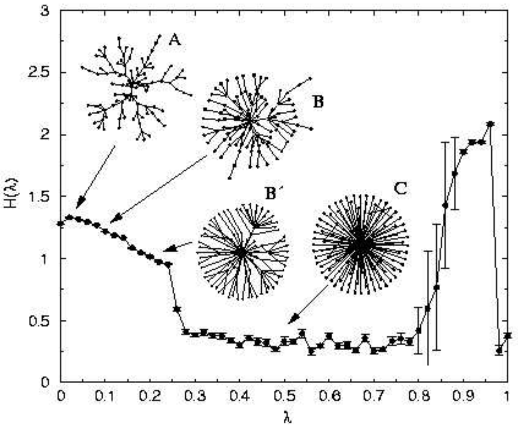


Figure 1.10: Degree entropy $H(\lambda)$ (averaged over 50 replicas) as a function of λ with $n = 100$ and $\rho = 0.2$. Optimal networks for selected values of the parameter λ are shown. A: an exponential-like network with $\lambda = 0.01$. B: a network with truncated power-law degree distribution with $\lambda = 0.08$. B': intermediate graph between B and C. C: a star network with $\lambda = 0.5$.

between any two nodes, the type of nodes encountered along the path(s) joining them may also matter in the optimization of the dynamics of communication taking place in the system. For example, selective pressure may operate so as to choose certain nodes because of their high connectedness - or else to avoid them. Associated with the type of node, is a local feature that depends only on its degree, namely, the number of edges rooted in the node. On a global scale, we will distinguish among structures that rewire local features at random selecting the changes if the new structure provides a selective advantage. It is well known that in many such optimization problems, the key factor that matters is the shape of the cost function [20,52,104–106]. The concavity or convexity of the cost function can be embodied by a power law form with scaling exponent α less than or greater than 1 respectively:

$$H_\alpha = \sum_{i<j} d_{ij}(\alpha), \tag{1.24}$$

where i and j are pairs of nodes of the network, and

$$d_{ij}(\alpha) = \min_P \sum_{p \in P: i \to j} k_p^\alpha . \tag{1.25}$$

Here P is any path connecting site i to site j of the system, p is any node belonging to such a path and k_p is the degree or connectivity of node p. The weighted distance $d_{ij}(\alpha)$ is a global quantity associated with the pair i, j and is the minimum of the sum of degrees k_p^α (a local property), evaluated along the path P from i to j, over all the paths connecting i and j. In the limiting case $\alpha \to 0$, eq. (1.25) becomes the standard definition of distance on a network [50].

The new definition of weighted graph distance introduced in eq. (1.25) captures the conflict between two competitive trends: (i) the avoidance of long paths so that the system minimizes distances regardless of "traffic" to simply reduce the graph distance between vertices; and (ii) the need to avoid heavy traffic arising from highly connected nodes (hubs) which behave as bottlenecks along the path from one vertex to another.

The networks minimizing the cost eq. (1.24) are searched for among the ensemble containing a fixed number of nodes n, as well as the number of links (edges) l. The optimization method used in the numerical simulations is a Metropolis scheme at zero temperature. The goal is to obtain the statistics of all local minima which are accessible topologies associated with the chosen dynamics [108].

The protocol of the simulation is as follows:

(i) *Generation of a random initial configuration* with fixed n and l. Starting with a single node, at each time step a new vertex is added and connected to an already existing node, extracted with uniform probability; the algorithm is repeated until a connected tree of size n is obtained; the remaining $l - (n-1)$ links are randomly added in the system, by linking pairs of vertices not already connected, extracted with uniform probability. This procedure ensures the generation of a connected random network with fixed n and l.

(ii) *Random rewiring.* Specifically, a link connecting the sites i and j is randomly chosen and substituted with a link from i to a site k, not already connected to i, extracted with uniform probability among the sites of the system. This ensures that the number of links l, as well as the size of the system n, remains constant during the minimization.

(iii) *Connectedness control.* If the graph is not connected after rewiring, step (ii) is repeated;

(iv) *Energetic control.* The new value of $H_\alpha(t+1)$ is calculated. The new configuration is accepted only if it is energetically favorable, i.e. only if $H_\alpha(t+1) < H_\alpha(t)$; otherwise the change is rejected and we return to step (ii).

Note that the zero-temperature setting ensures feasible optimality of the emerging network structure [104–106], a feature that is relevant for dynamical accessibility of complex optimal structures. The minimization algorithm stops after F consecutive failed changes on the network; we have chosen $F = n(n-1)$, so that, on average, each pair of vertices is allowed to change its state twice.

The resulting networks are analyzed in terms of the degree distribution $P(k)$, i.e. the fraction of nodes with connectivity k, the average distance between pairs of nodes, L, and the average clustering coefficient $\langle C \rangle$, a measure of the local interconnectivity of vertices in the system, as defined in eqs. (1.2) and (1.3).

1.4 Trees

In the special case of trees, i.e. loopless structures, the path $P : i \to j$ connecting the vertices i and j is unique and the weighted distance d_{ij} assumes the form:

$$d_{ij} = \sum_{p \in P: i \to j} k_p^\alpha . \qquad (1.26)$$

1.4.1 Results

Minimizing the cost function through Metropolis algorithm we obtain optimal networks with emerging topologies all displaying a precise structural organization or hierarchy (see e.g. fig. 1.11). They are characterized by a central node from which almost identical branching structures depart. Let g be the generation or level of the branching structure, with $g = 0$ at the central node and $g = G$ at the last level. Because of the symmetry encountered in optimal patterns, each node belonging to the same generation g has the same value of connectivity k_g, except for the last level due to finite size effects. Therefore, the cost function of eq. (1.24) can be rewritten as a sum over the generation levels:

$$H_\alpha = \sum_{g=0}^{G} N_g \, b_g \, k_g^\alpha \tag{1.27}$$

where N_g is the number of nodes belonging to the level g, b_g is the node betweenness for level g, defined as the total number of shortest paths between any two vertices in the network passing through a node belonging to that level, and k_g is the degree of each node at generation level g. Choosing a vertex v belonging to the level g, we indicate with $n_g\,(k_g - 1)$ the number of nodes contained in the branches departing from v, except v itself; n_g thus represents the number of nodes contained in the branch starting from level $g + 1$ to the final generation G (see fig. 1.11). Node betweenness b_g for level g can be expressed in terms of n_g and the level connectivity k_g:

$$b_g = n_g\,(k_g - 1)\,[n - n_g(k_g - 1) - 1] + \frac{(k_g - 1)(k_g - 2)}{2}\,n_g^2 + n - 1. \tag{1.28}$$

The first term on the right hand side of eq. (1.28) accounts for the paths connecting one of the $n_g\,(k_g - 1)$ nodes to one node belonging to the rest of the network (whose total number is $n - n_g(k_g - 1) - 1$, with 1 accounting for the node v) and passing through the vertex v at level g; the second term represents the number of paths passing through the given vertex and connecting nodes belonging to the same branch departing from that vertex; finally, the last term represents the number of paths starting from the node v to every other node in the network $(n - 1)$.

The number of nodes N_g, belonging to the generation level g, and the number of nodes n_g, contained in one of the branches departing from that level, can be expressed in terms of level connectivities k_g:

$$N_g = k_0 \prod_{i=1}^{g-1}(k_i - 1) \quad \text{for } g \geq 2 \tag{1.29}$$

with $N_0 = 1$ and $N_1 = k_0$, and

$$n_g = 1 + \sum_{i=g+1}^{G-1} \prod_{j=g+1}^{i} (k_j - 1) \quad \text{for } g \leq G - 2 \tag{1.30}$$

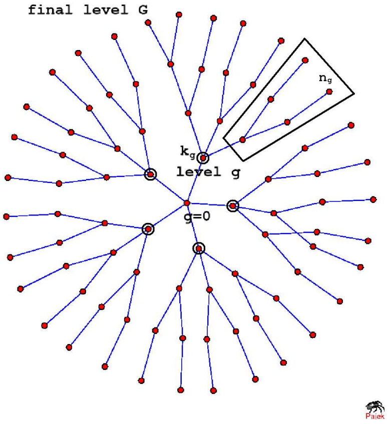


Figure 1.11: Graph representation of a typical tree with $n = 82$ and $\alpha = 0.9$. Centre $g = 0$ and final level G are shown. The N_g nodes of level g are those in black circles. The polygon shows the n_g vertices. The graph has been produced with the Pajek software [109].

with $n_G = 0$ and $n_{G-1} = 1$.

The expression of the cost function in terms of level contributions - eq. (1.27) - toghether with the expressions for b_g, N_g and n_g in terms of generation connectivities k_g - eqs. (1.28), (1.29) and (1.30) - allow us to exhaustively explore configuration space of hierarchical trees, being those found as minima in numerical simulations. For different system sizes, we explore thoroughly all the possible discrete values that coordination k_g can assume for each generation g, compatible with the size n, and evaluate the corresponding value of $H(\alpha)$. For a fixed value of n, the optimal pattern is the one corresponding to the minimum of $H(\alpha)$; the total number of generations G is automatically determined by the minimization of the cost function. The coordination of the last two levels, G and $G - 1$, might not be constant for all vertices belonging to those levels, due to finite-size effects and discreteness of the problem.

We have studied different sizes of the system, up to values $n = 700$ nodes, and investigated the role of concavity of the cost function, adopting values of α greater and smaller than 1. In fig. 1.12 we present some samples of optimal patterns obtained.

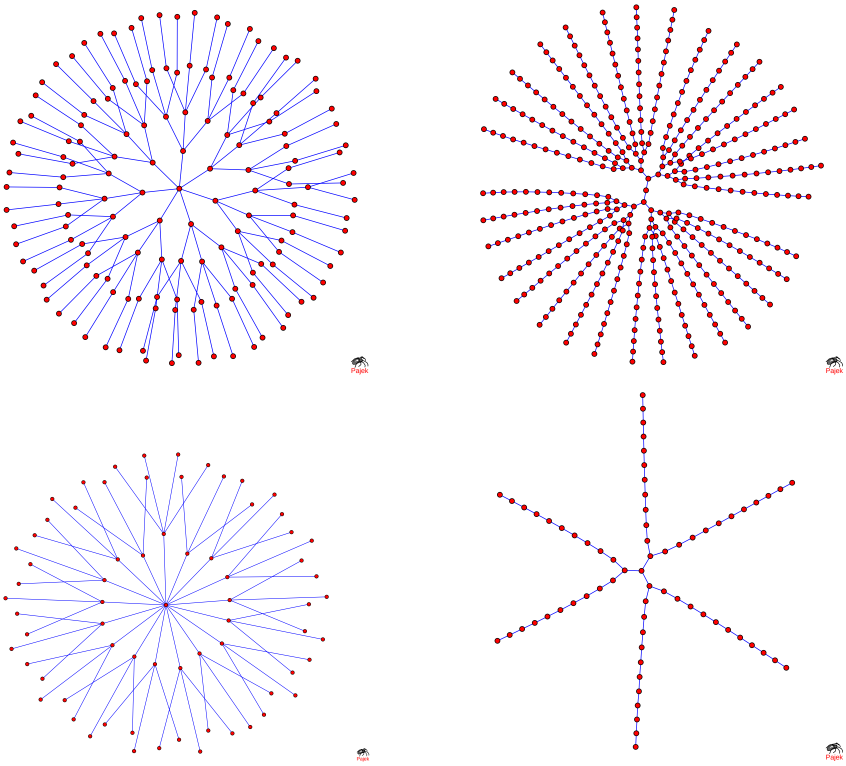


Figure 1.12: Graph representation of four optimal trees with: Top Left: $\alpha = 0.9$, $n = 156$; Top Right: $\alpha = 4.0$, $n = 424$; Bottom Left: $\alpha = 0.5$, $n = 70$; Bottom Right: $\alpha = 4.0$, $n = 68$.

The importance of the shape of the cost function is reflected in the behaviour displayed by optimal trees. When $\alpha < 1$ (graphs on the left in fig. 1.12), the minimization of the graph distance between any two nodes in the system dominates, leading to branching structures characterized by higher connectivities in order to decrease the total number of generations G, for fixed values of α and n, and reduce network diameter. When $\alpha > 1$, instead, the system tries to minimize node degree, thus leading to the emergence of several linear patterns (graphs on the right in fig. 1.12), i.e. sub-networks composed of nodes having connectivity $k = 2$, which are disadvantaged and, thus, almost completely absent for $\alpha < 1$.

We are not able to go further with our investigation of optimal trees, including e.g. the study of the degree distribution, since for the values of n analyzed we do not obtain a range of connectivities large enough for the analysis.

However, we have shown that optimization of loopless structures leads to the emergence of a feature commonly observed in natural and artificial complex networks - the hierarchical organization of network structures - which has been the object of several recent works [84–87,110,111]. In fig. 1.13 we show, as an example, one of the typical loopless networks we obtained from optimization, corresponding to the values $n = 85$ and $\alpha = 0.4$ (on the left), which very closely resembles a graph representation of a telephone network (on the right), where the terminals are telephone sets and a node is a switching center for routing telephone calls [112,113].

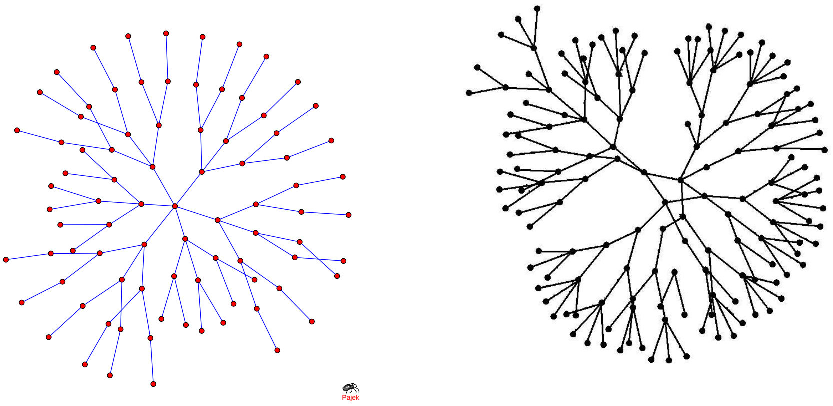


Figure 1.13: Left: graph representation of an optimal tree corresponding to the values $n = 85$ and $\alpha = 0.9$. Right: graph representation of a local telephone network (from [112]).

In the next section we will discuss topologies of generic networks, i.e. networks with loops, resulting from the optimization of the cost function, which display the emergence of strikingly interesting features.

1.5 Generic networks

Taking into account more generic networks, i.e. networks with loops, we must refer to the original expression of the weighted distance, eq. (1.25), and use Metropolis minimization algorithm to find optimal topologies.

For each optimization, we performed 200 independent simulations to average over initial random configurations. We have varied the size n of the system over the following values: n = 35, 50, 70, 100, 140, 200. For each size, the different values of the ratio $r = l/n$ investigated are: r = 1.05, 1.1, 1.2, 1.3, 2.0, 2.3, 3.0.

1.5.1 Results

On varying r, we observe two distinct behaviors. The first occurs for values of $r \sim 1$: the system displays a broad distribution of degrees for several values of α (see $P(k)$ in fig. (1.14), for $\alpha = 0.7$). However, the behavior does not seem to be a genuine power law because the sharp cut-off does not display the expected dependence on the system size n. Unfortunately, the high computational cost, due to the exponential growth of possible configurations and the nature of the global selection process pursued, prevents us to increase further the size of the system in order to quantify the weak dependence of the cut-off on n. As α increases, this heterogeneous region shrinks around the value $r = 1$ and is vanishingly small for $\alpha > 1$. The second behavior is obtained for larger values of the ratio r – the degree distribution obtained is strongly peaked around the average value of k, $\langle k \rangle$ (fig. (1.15)).

A sample of network topologies are illustrated in fig. (1.16), for different values of α and r. On increasing the value of the ratio r, one moves from networks characterized by the presence of some highly connected nodes together with many peripheral sites (Top Left and Right) to networks in which almost every node has the same degree $k = \langle k \rangle$ (Bottom Left and Right). In addition, a sharp transition is observed in terms of the average clustering coefficient $C = \langle C_i \rangle$, as defined in eq. (1.3).

For $\alpha > 1$ (fig. 1.17 Top), the system undergoes a clear phase transition as the value of the ratio r increases passing from a regime characterized by zero clustering to one in which the clustering coefficient becomes different from zero. The cost

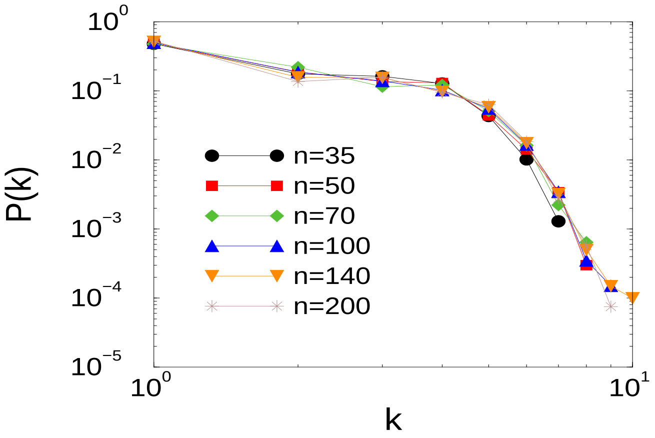


Figure 1.14: Degree distribution, averaged over 200 realizations, for several system sizes ($n = 35,\ 50,\ 70,\ 100,\ 140,\ 200$) for $\alpha = 0.7$ and $r = 1.05$. The system displays a range of degrees.

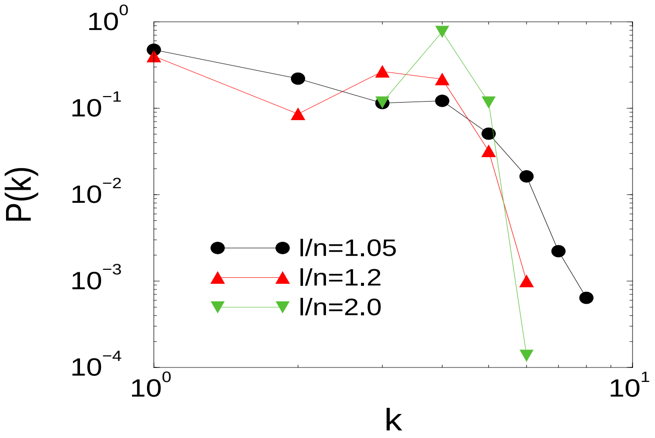


Figure 1.15: Crossover between the two distinct behaviors: the heterogeneous regime which exhibits a range of degrees and the homogeneous one characterized by a peaked distribution. Data are averaged over 200 realizations for $\alpha = 0.7$, $n = 70$ and for several values of $r = l/n$.

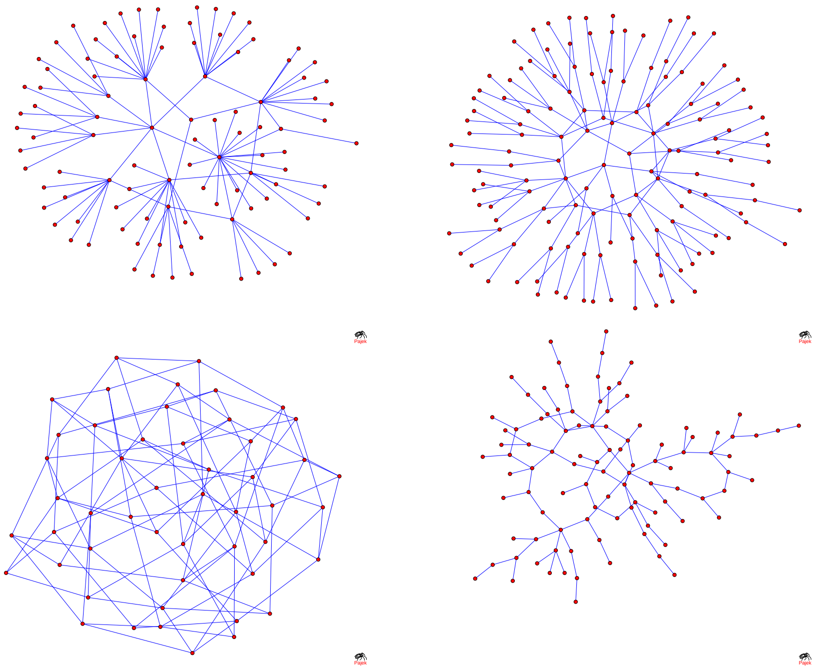


Figure 1.16: Graph representation of four typical networks with: Top Left: $\alpha = 0.4$, $r = 1.05$, $n = 100$; Top Right: $\alpha = 0.7$, $r = 1.05$, $n = 140$; Bottom Left: $\alpha = 0.5$, $r = 2.0$, $n = 50$; Bottom Right: $\alpha = 2.0$, $r = 1.05$, $n = 100$.

function in eq. 1.24 has two competing forces: the minimization of the graph diameter and the minimization of node degree. When $\alpha > 1$ the minimization of node degree dominates and the system attempts to minimize the degree of each node resulting in a peaked distribution around the mean value $\langle k \rangle$, with a non-trivial topology characterized by zero clustering and exhibiting the presence of long loops. (fig. 1.16 Bottom Right). When the ratio r reaches the critical value $r_c(\alpha)$, one obtains a non-zero clustering coefficient.

This transition also occurs for $\alpha < 1$. However, when $\alpha < 1$ one obtains an additional phase transition at $r'_c(\alpha)$, where the system passes from optimal networks exhibiting a non-zero clustering coefficient, to ones with no clustering at all. Indeed, when $\alpha < 1$, the tendency expressed by the cost function is to decrease the graph diameter, i.e. a measure of the mutual distance among pairs of nodes.

Starting from very small values of r, we observe topologies characterized by the presence of few interconnected hubs (i.e. sites with very high degree [49, 107]) linked to many peripheral sites (fig. 1.16 Top Left). The clustering is different from zero due to the mutual connnections between hubs.

The emergence of this extra phase transition underscores the importance of the concavity (convexity) of the cost function.

The limiting case $\alpha \to 0$ would correspond to the minimization of the standard graph distance, leading, in the region $r \sim 1$, to a single central hub connected to $n-1$ peripheral nodes, which share the remaining $l-n+1$ links. This situation leads to non-zero clustering. The minimization of the graph distance corresponds to a limiting case of [81] as well; however, in [81] there is no constraint on the number of links l, so that the optimal network they find is a clique, in which each node is connected to each other.

Increasing the ratio r does not favour adding other links among the hubs, because their already high degrees would only increase further, thus leading to a higher value of the cost function. Hence the system reorganizes by increasing the number of hubs, automatically reducing their degrees and distributing the peripheral sites among them, trying to avoid expensive triangles between hubs in order to further decrease their connectivities. When the transition occurs, at $r'_c(\alpha)$, the network does not exhibit hubs any more, but tends to become quite homogeneous in the sense that almost every node has connectivity close to the average value $\langle k \rangle$. Even in this regime the optimal topology is distinctly different from the random one. In fact, it displays a peaked degree distribution around the mean value $\langle k \rangle$ without significant clustering (fig. 1.16 Bottom Left), even though the network is not a tree. Emerging loops have the maximum possible length in order to reduce the energy function. Adding extra links to the network forces the loops to become smaller, still avoiding clustering up to a second critical value of r, $r_c(\alpha)$. Beyond this value, 'triangles' appear leading to a mean clustering coefficient different from zero. A similar transition occurs for $\alpha > 1$ (fig. 1.17 Bottom, inset), when the trend that dominates is the minimization of node connectivity.

The extent of the clustering phase for $r < r'_c(\alpha)$ and $\alpha < 1$ shrinks for increasing values of α; the critical value $r_c(\alpha)$ decreases as α increases, $\forall \alpha$. From fig. 1.14, 1.15 and fig. 1.17, one finds that several distinct topologies are obtained for different values of α and r:

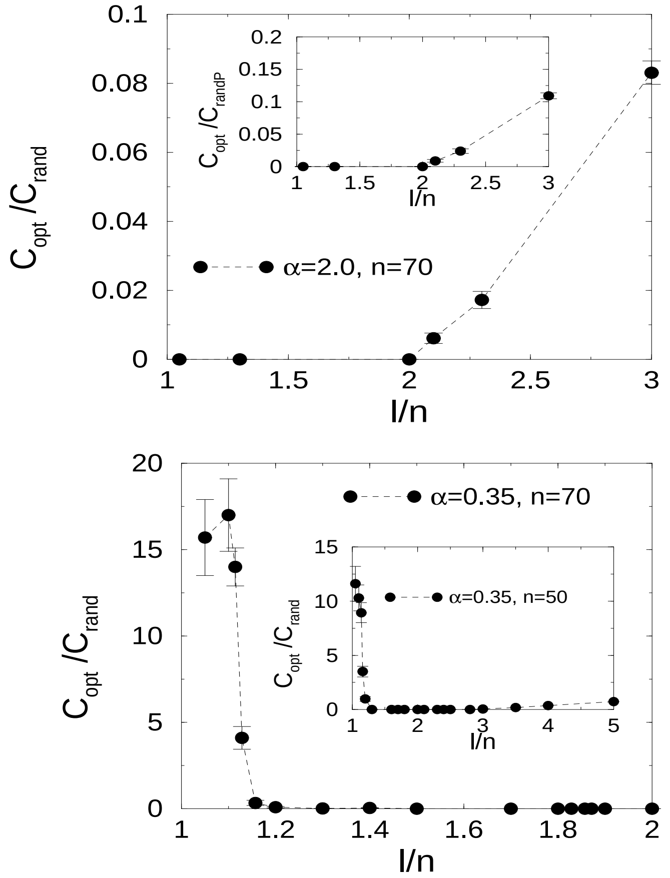


Figure 1.17: Mean clustering coefficient for the optimal configuration C_{opt} normalized to the mean clustering coefficient, C_{rand}, of the random configuration. Top: results for network size $n = 70$ and $\alpha = 2.0$; in the inset the behaviour of the ratio C_{opt}/C_{randP} is shown, where C_{randP} represents the mean clustering of a random graph with the same degree distribution $P(k)$ as the optimized network. Bottom: results for network size $n = 70$ and $\alpha = 0.35$; in the inset ($n = 50$, $\alpha = 0.35$) both the critical values, $r_c(\alpha)$ and $r'_c(\alpha)$, are shown.

- a **heterogeneous regime** exhibiting a broad distribution of degrees ($r \sim 1$, $\alpha < 1$) observable both in the **clustering** and **no clustering phase** depending on the value of α;

- a **homogeneous regime** for larger values of r with **clustering different from zero** ($r > r_c(\alpha)$ $\forall\alpha$, and $\alpha < 1$, $r < r'_c(\alpha)$ but not in the tree-like limit) or **zero clustering** ($\alpha < 1$, $r'_c(\alpha) < r < r_c(\alpha)$ and $\alpha > 1$, $r < r_c(\alpha)$).

Fig. 1.17 shows that in the region $\alpha < 1$ and $r \sim 1$ the average clustering coefficient of the optimal network is greater than the random one. Unfortunately, the limited number of nodes we can deal with (because of the high computational costs) precludes a direct comparison with large real networks. However, making the hypothesis of an optimal mean clustering, C_{opt}, independent of the system size n, we can compare the ratio C_{opt}/C_{rand} obtained in our simulations with those calculated for real networks having the same average degrees [49]: one obtains ratios C_{real}/C_{rand} in the range of those obtained numerically for several real networks.

Results presented have been compared to corresponding values obtained in a classical random graph [1, 4], C_{rand}, having same size n and number of links l. However, we have also studied system behaviour in comparison to a random graph characterized by the same degree distribution $P(k)$ of the optimized network [93, 94], C_{randP}. Both studies give similar results, as it is shown in the inset of fig. 1.17 (top).

We have also studied the characteristic path length, L, defined as the average, over all pairs in the system, of the graph distance between pairs of nodes. As shown in fig. 1.18, in the entire interval of α, the characteristic path length of the optimal configuration, L_{opt}, is comparable to or smaller than the random one, L_{rand}. Even though the small network sizes studied here do not allow us to reach definitive conclusions, there exist a range in α and l/n in which optimal networks seem to display a small-world effect [50].

1.6 Conclusions

Optimality leads to the emergence of several distinct network structures, showing hierarchical organization in loopless structures and including a heterogeneous regime characterized by a broad distribution of degrees in the tree-like topology limit. Besides the degree distribution, we have studied the clustering coefficient and the average path length of the selected networks which point to the existence of non-trivial phase transitions and to the features of the small-world effect. Our

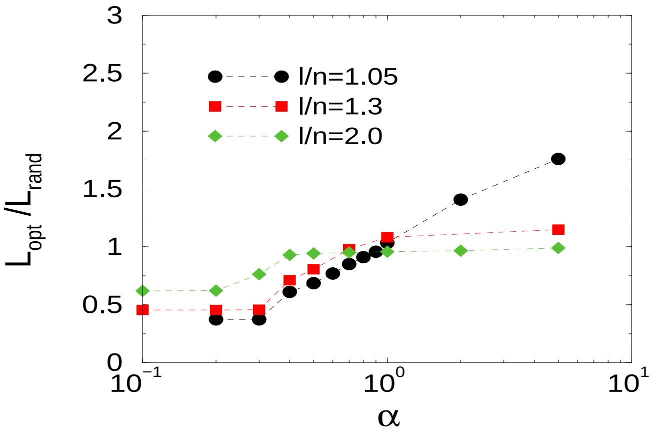


Figure 1.18: Characteristic path length L_{opt}, normalized to the classical random one L_{rand}, vs. α.

main result is that the emergence of the structural properties observed in natural network patterns may not be necessarily due to embedded growing mechanisms only, but may rather reflect the interplay of dynamical mechanisms with an evolutionary selective process.

Chapter 2

Network Renormalization

Renormalization Group (RG) Theory has proven to be extremely useful in the explanation of critical and multi-critical phenomena, and in other problems in the broad area of statistical physics and condensed matter theory, physical chemistry and beyond (for a review, see [71]). It was originally developed for applications to regular lattices. Here we propose a real-space renormalization group approach to complex networks.

2.1 Coarse-graining a generic network

Renormalization group techniques have been applied to stochastically growing networks, in order to study the critical behaviour of percolation [72], and to static networks, such as small world networks, in order to analyze the transition from a regular-lattice behaviour to a random-graph behaviour [73].

In the following, we investigate the critical behaviour observed in real-world complex networks - for what concerns, e.g., degree distribution and several measures of degree-degree correlations and hierarchical organization - through the application of real-space RG approaches. To our knowledge, no such investigation of criticality in complex networks has yet been performed.

Our focus is on elucidating the critical features of complex networks, by analyzing the robustness of these behaviours under renormalization, addressing the problem of distinguishing between critical and only apparently critical behaviours.

Moreover, through the application of renormalization techniques, we would like to “simplify” complex networks, providing more simple and understandable versions of large-scale networks, hopefully resulting in a powerful method for network visualization.

In the following sections, we introduce distinct RG techniques and the results obtained from their application to complex networks generated with several different models, designed for application to different real networks. An application to real-world networks is given in section 3.3, where protein-protein interaction networks are analyzed.

2.2 Derivation

We consider a generic network composed of n nodes and described by the adjacency matrix $\mathbf{A}$. The associated *laplacian* $\mathbf{\Gamma}$ is defined as

$$\Gamma_{ij} = A_{ij} - \delta_{ij} \sum_l A_{il} = \begin{cases} -k_i & \text{if } i \equiv j \\ A_{ij} & \text{otherwise} \end{cases} \tag{2.1}$$

where i and j represents two vertices of the newtork, δ_{ij} is the discrete delta function, and k_i is the degree of node i.

An energy function H can be associated to the network:

$$H = \frac{1}{4} \sum_{ij} A_{ij} (\phi_i - \phi_j)^2 = -\frac{1}{2} \vec{\boldsymbol{\phi}} \mathbf{\Gamma} \, \vec{\boldsymbol{\phi}} \tag{2.2}$$

which might be interpreted in at least two different ways. In the framework of electrical networks, eq. (2.2) represents the global dissipated *electrical energy* of the system, with $A_{ij}/2$ being the conductance between two nodes i and j, and ϕ_i representing the potential at vertex i. In the context of vibrational networks, ϕ_i represents the scalar displacement of vertex i from its equilibrium position and A_{ij} the strength of the link connecting i and j, thus leading to a mechanical interpretation of H as the *elastic energy* associated to a deformation $\vec{\phi} = (\phi_1, \phi_2, \ldots, \phi_n)$ of the network.

It is worth to notice that the laplacian $\mathbf{\Gamma}$ satisfies the following equation:

$$\sum_i \Gamma_{ij} = 0 \,, \tag{2.3}$$

having the trivial eigenmode $(1, 1, \ldots, 1)$. In order to be able to express the energy function H as in eq. (2.2) after renormalization, we want the constraint of eq. (2.3) to be satisfied by renormalized network.

After the application of renormalization procedures, illustrated in the next section, we will deal with a laplacian $\mathbf{\Gamma'}$ whose elements will in general not assume only values equal to 0 or 1 as in the initial network. Therefore, we will obtain a *weighted newtork*.

As discussed in the previous chapter, the theory of complex networks usually deal with networks characterized by binomial interactions, i.e. $A_{ij} = 0$ or 1, implicitly assuming that the "quality" or "strength" of each interaction is identical. However, there exist a lot of networks in which edge weight can be unambigously distinguished, since they are intrinsically weighted. Examples include scientific citation networks, where the strength is represented by the number of joint publications, transportation networks such as airline networks with passenger capacity, social networks where the strength of an interaction might be stronger or weaker, food webs, accounting for the diversity in predator-prey interactions, the Internet where the traffic passing through a link actually represent an important ingredient, and many others. Research attention has focused on this type of networks only very recently [19, 45, 48, 114–119].

Here we define some quantities of interest for the following analysis of weighted networks, emerging after renormalization. To each node is associated a *weighted degree* (also known as *strength* in the previously cited literature), defined as the sum of the weights associated to the links starting from node i:

$$k_i^w = \sum_j w_{ij} \tag{2.4}$$

where $w_{ij} \equiv \Gamma_{ij}$. In analogy with standard definitions (eq. (1.3) and eq. (1.7)), we define the *weighted clustering coefficient* for the node i as

$$C_i^w = \frac{\sum_{jl} w_{ij} w_{jl} w_{li}}{\sum_{jl} w_{ij} w_{il} \max(w)} \tag{2.5}$$

and the *weighted average neighbors degree* of the node i:

$$k_{nn,i}^w = \frac{1}{k^w} \sum_j w_{ij} k_j^w \,. \tag{2.6}$$

In eq. (2.5) we must divide C_i^w by the maximum weight $\max(w)$, to correctly normalize the clustering coefficient. Toghether with the *weighted degree distribution* $P(k^w)$, i.e. the probability that a randomly chosen node has weighted degree k^w, we also study the dependance of C^w and k_{nn}^w on the weighted degree k^w. The definitions just given differ from those found in the literature, since we are not interested in the relation between weighted networks and underlying binomial networks, characterized by the standard definitions of degree, clustering coefficient, etc., presented in the previous chapter. What we would like to investigate, instead, is the critical behaviour of netwoks under coarse-graining, which forces the appearance of weighted edges in the renormalized laplacian $\mathbf{\Gamma'}$.

2.3 Renormalization procedures

As a real-space renormalization technique, we adopted a decimation procedure in which a certain set of nodes $\mathbb{D}$ is removed from the original set of nodes $\mathbb{N}$, leaving a remaining ensemble $\mathbb{N}'$ of nodes, with $\mathbb{N} = \mathbb{N}' \cup \mathbb{D}$. Two different kinds of decimation procedures, explained in the following paragraph, have been explored.

In order to obtain the renormalized expression of $\mathbf{\Gamma'}$ associated to the network $\mathbb{N}'$, in terms of $\mathbf{\Gamma}$, we must integrate out the decimated sites, i.e. nodes belonging to the set $\mathbb{D}$:

$$\exp\left(-H'\right) \;=\; \exp\left(\frac{1}{2}\vec{\phi}\,\Gamma'\,\vec{\phi}\right) \propto \int \prod_{v\in\mathbb{D}} d\phi_v \;\exp(-H)\,. \tag{2.7}$$

We can rewrite the energy H as a sum of terms depending on decimated sites only ($v \in \mathbb{D}$), on remaining sites only ($i \in \mathbb{N}'$), and of mixed terms ($i \in \mathbb{N}'$ and $v \in \mathbb{D}$):

$$-H \;=\; \frac{1}{2}\left[\sum_{i,j\in\mathbb{N}'} \phi_i\Gamma_{ij}\phi_j \;+\; \sum_{v,w\in\mathbb{D}} \phi_v\Gamma_{vw}\phi_w \;+\; \sum_{i,v} \phi_i\Gamma_{iv}\phi_v\right]\;, \tag{2.8}$$

thus obtaining:

$$\exp\left(-H'\right) \propto \exp\left(\frac{1}{2}\sum_{i,j\in\mathbb{N}'} \phi_i\Gamma_{ij}\phi_j\right) \int \prod_{v\in\mathbb{D}} d\phi_v \;\exp\frac{1}{2}\left(\sum_{v,w\in\mathbb{D}} \phi_v\Gamma_{vw}\phi_w \;+\; \sum_{i,v}\phi_i\Gamma_{iv}\phi_v\right) \tag{2.9}$$

Making use of the following identity

$$\int \prod_i dx_i \, \exp\left(-\frac{1}{4}\sum_{ij} x_i V_{ij} x_j + \sum_i s_i x_i\right) = const \, \exp\left(\sum_{ij} s_i V_{ij}^{-1} s_j\right) \quad (2.10)$$

we finally obtain the result

$$\Gamma'_{ij} = \Gamma_{ij} - \sum_{v,w \in \mathbb{D}} \Gamma_{iv} \left(\Gamma^{(d)}\right)^{-1}_{vw} \Gamma_{wj} \quad (2.11)$$

where $\mathbf{\Gamma^{(d)}}$ is the matrix $\mathbf{\Gamma}$ restricted to the decimated ensemble $\mathbb{D}$ and $(\mathbf{\Gamma^{(d)}})^{-1}$ its inverse. It is easy to see that also the renormalized laplacian $\mathbf{\Gamma'}$ satisfies eq. (2.3). Since the field ϕ can be redefined by a multiplicative constant, also the elements of $\mathbf{\Gamma'}$ can be multiplied by a positive constant, leaving unchanged the properties of the newtork. This happens to be useful when the normalization to the maximum value of Γ_{ij} is required, as e.g. for the evaluation of the weighted clustering coefficient (see eq. 2.5).

2.3.1 'Minimum' and 'threshold' decimation

The decimated ensemble $\mathbb{D}$ may be determined according to different rules. The idea is to delete nodes which, in some sense, are *less important* in the network. Thus, one could think of removing nodes with smallest degree - known as dangling ends - since they are far from being the key vertices in the network, or those vertices having lowest betweenness (a quantity which measures the total number of shortest paths passing through the given node), since they represent the less visited nodes in the system.

Here we adopt a decimation procedure based on the elimination of nodes with lowest weighted degree. In particular, we distinguish between a 'minimum' decimation in which only the nodes having minimum weighted degree, $\min(k^w)$, are deleted and a 'threshold' decimation which remove the ensemble of nodes having k^w below a certain threshold. For example, the threshold might be chosen in such a way to decimate a certain percentage of nodes with lowest weighted degrees.

Removing nodes from the network, increases the connectivity k and results in a renormalization of already existing weighted edges ($\Gamma_{ij} \neq 0 \ \rightarrow \Gamma'_{ij} \neq 0$) and in

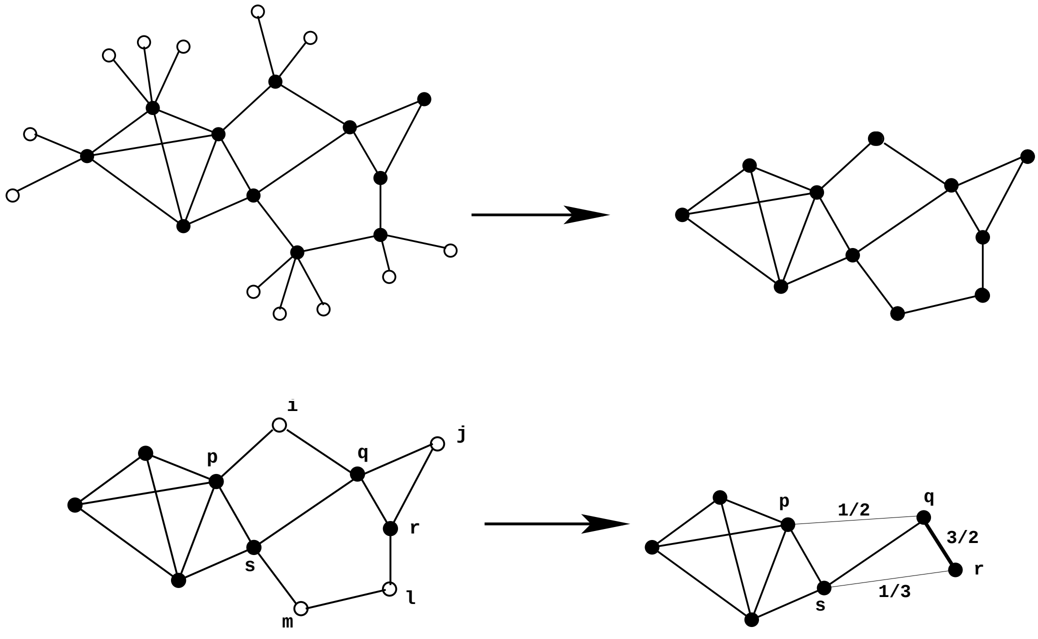


Figure 2.1: Graph representation of successive renormalizations through 'minimum' decimation. Top: Nodes with weighted degree $k^w = 1$, i.e. dangling ends, are selected for decimation (blank circles). No renormalization of links occurs after decimation of $k^w = 1$ nodes: all links have weight equal to 1 (in the picture, only weights $w_{ij} \neq 1$ report their own value). Bottom: Successive decimation involves nodes with $k^w = 2$, i.e. nodes i, j, l, m (blank circles). Renormalization affects already existing links (e.g. (q,r), going from $w_{qr} = 1$ to $w_{qr} = 3/2$, as illustrated in the picture), and also creates new edges (e.g. (p,q) with $w_{pq} = 1/2$ from the decimation of the node i, and (r,s) with $w_{rs} = 1/3$ from the decimation of two nodes, l and m). A thickness proportional to the weight has been used for visualization.

the creation of new weighted edges ($\Gamma_{ij} \equiv 0 \ \rightarrow \Gamma'_{ij} \neq 0$), thus producing short-cut connections. A simple representation of what occurs is depicted in fig. 2.1.

Decimation is then iterated many times and results obtained after each renormalization are compared and discussed.

2.4 Results

We have investigated renormalization group approach applied to several networks generated by the models presented in the previous chapter. We have studied the topological properties of the emerging weighted networks, after several decimations. Figures in this section show results obtained from a single realization of each model. Binning of the data is necessary to plot the behaviours of the quantities studied as a function of the weighted degree, since it assumes non-integer values. It also helps us in reducing the statistical noise inevitably arising from the consideration of a single realization. We used exponential bin length for the variable range, since plots are in logarithmic scale.

The models under considerations result in different critical features. They all display a power-law degree distribution, but are characterized by different behaviours when considering the clustering coefficient or higher degree correlations. For comparison, we have investigated also the Erdös-Renyi model [1], since it is not scale-free and does not display correlations between the nodes.

In the following we show the results - in terms of $P(k^w)$, $C^w(k^w)$ and $k^w_{nn}(k^w)$ - relative to the models:

Barabási-Albert network (BA) [53, 54], illustrated in detail in section 1.1. In the present simulations we use the values $m = m_0 = 2$.

"Good-get-Richer" network [64]. It introduces the concept of *fitness* assigned to each node, as a measure of the intrinsic properties of a vertex. It does not rely on a preferential attachment rule, differing from [61, 62], but proposes a generalization of the Erdös-Renyi model in which the probability of creating a connection between two vertices depends on a *mutual benefit* and is expressed in terms of their fitnesses. To each node i is assigned a fitness f_i extracted from a probability distribution $\rho(f)$. A link between nodes i and j is created with probability $F(f_i, f_j)$, dependent on the fitnesses of both vertices. In the following, we use an exponential distribution of fitnesses $\rho(f) = \exp(-f)$ and a step function $F(f_i, f_j) = \theta(f_i + f_j - z)$, where $z = z(n)$ is a given threshold; in our simulations $z = 6.2$. The rules chosen lead to a scale-free degree distribution and to non-trivial correlations resulting in power-law behaviours found in $C(k)$ and $k_{nn}(k)$.

Acquaintance network model We adopt a variant of the model for social networks developed in [120]. Starting with a single vertex, the algorithm consists in the iteration of the following steps:

- with probability $1-u$ a new node i enters the system and connect to an already existing node j extracted with uniform probability; *potential edges* between i and neighbors of j are created;
- with probability u a potential edge, selected randomly, is converted into an edge.

The concept of potential edge mimics in social networks the occurrence of potential acquaintances between acquaintances of a given person. Results presented are obtained using a probability $u = 0.8$.

Erdös-Renyi network (ER) [1], presented in section 1.1. Networks generated have average value $\langle k \rangle = 10$.

We have performed numerical simulations of these models using the parameters mentioned above and network size $n = 1000$. Although the size considered is relatively small, we must mention that we are not interested in the intrinsic properties of the models (and, thus, neither in the absolute values of the power-law exponents), but in the renormalization of several network types. Future developments will investigate larger networks, in order to limit finite-size effects. The study presented here, however, is useful for a possible application to real networks with similar sizes, such as, e.g., protein interaction networks (see section 3.3).

For each quantity investigated, we compare results obtained with 'minimum' and 'threshold' decimation techniques. The threshold is chosen in such a way to decimate at least the 50% nodes with lowest weighted degree.

We first analyze the weighted degree distribution $P(k^w)$. In fig. 2.2 we report on the left results obtained with minimum decimation and on the right those coming from the threshold decimation adopted. Number of decimations and corresponding network sizes are shown. For all models, the scale-free behaviour is recovered after successive renormalizations. Strikingly, also the slope is conserved, with power-law exponents within the deviations due to the fact that we are considering a single realization only. These observations are valid not only if we adopt minimum decimation, in which we are progressively reducing the number of nodes, but also if we apply threshold decimation, thus abruptly changing the size of the system ($n_{d+1} \simeq 50\% \, n_d$, with d indicating the decimation).

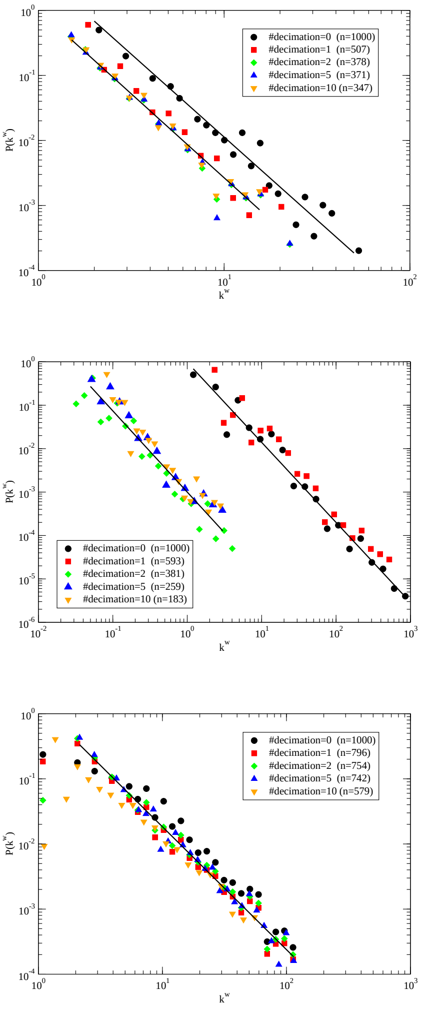


Figure 2.2: Results for $P(k^w)$ obtained with renormalization by 'minimum'(left) and 'threshold' decimation (right). Top: BA network. Center: “Good-get-Richer” network. Bottom: Acquaintance network.

Correlation properties seem to strengthen previous observations. In figs. 2.3

and 2.4, we report the average weighted clustering coefficient $C^w(k^w)$ and the average weighted connectivity of the neighbors $k^w_{nn}(k^w)$, respectively, as a function of the weighted degree k^w. As expected, while BA results are almost independent on k^w, the other models display non-trivial correlations between the degrees, resulting in scale-free properties. Also for these quantities, the emerging critical behaviour do not change after renormalization. Renormalized weighted clustering and weighted neighbor connectivity are in qualitative agreement with original ones, and seem also to provide correct results for the slope values, thus showing the robustness of the underlying connectivity correlations and hierarchical structure. Moreover, when the network display a lack of correlations, as in the case of BA model, renormalization do not affect the non-critical behaviour, reproducing almost flat average properties.

Finally, we have studied renormalization of a random graph to investigate networks far from criticality. Renormalized Erdös-Renyi networks display results similar to BA networks, for what concerns the absence of correlation - the flat behaviours of $C^w(k^w)$ and $k^w_{nn}(k^w)$ remain indeed unchanged. Also the behaviour displayed by renormalized $P(k^w)$ is preserved, with the range of weighted degrees shrinking due to decimation. In this case, threshold decimation only has been used, since minimum decimation leads to the deletion of a too small set of nodes, preventing a considerable change in the newtork size.

2.5 Discussion and perspectives

Renormalization group approach is a powerful tool in order to study and analyze critical phenomena. Our results on the application of RG techniques to complex networks suggest that it could represent an alternative way of investigating scale-free properties of real-world networks. For example, it could be useful in distinguishing between different behaviours, since critical properties - such as scale-free degree distribution, correlation properties and others - are preserved after successive renormalizations, performed with different techniques. Moreover, such properties seem to be *enforced* and more recognizable after renormalization, leading to power-laws which extend over a wider range of degrees, as observed, e.g., in the weighted clustering coefficient (fig. 2.3). On the other hand, if a net-

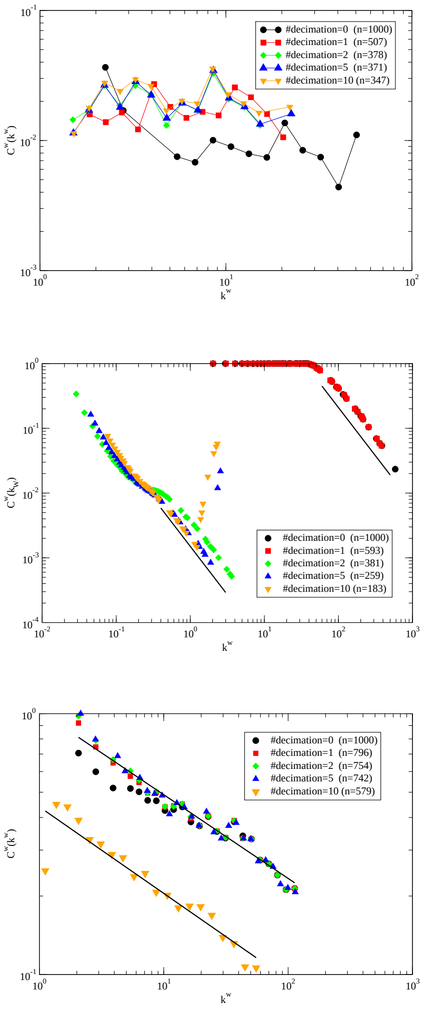


Figure 2.3: Results for $C^w(k^w)$ obtained with renormalization by 'minimum'(left) and 'threshold' decimation (right). Top: BA network. Center: "Good-get-Richer" network. Bottom: Acquaintance network.

work does not possess these features, renormalization leaves the system far from criticality.

However, this study is still at the beginning, and more work in this direction has to be done. For example, one could investigate renormalization of networks

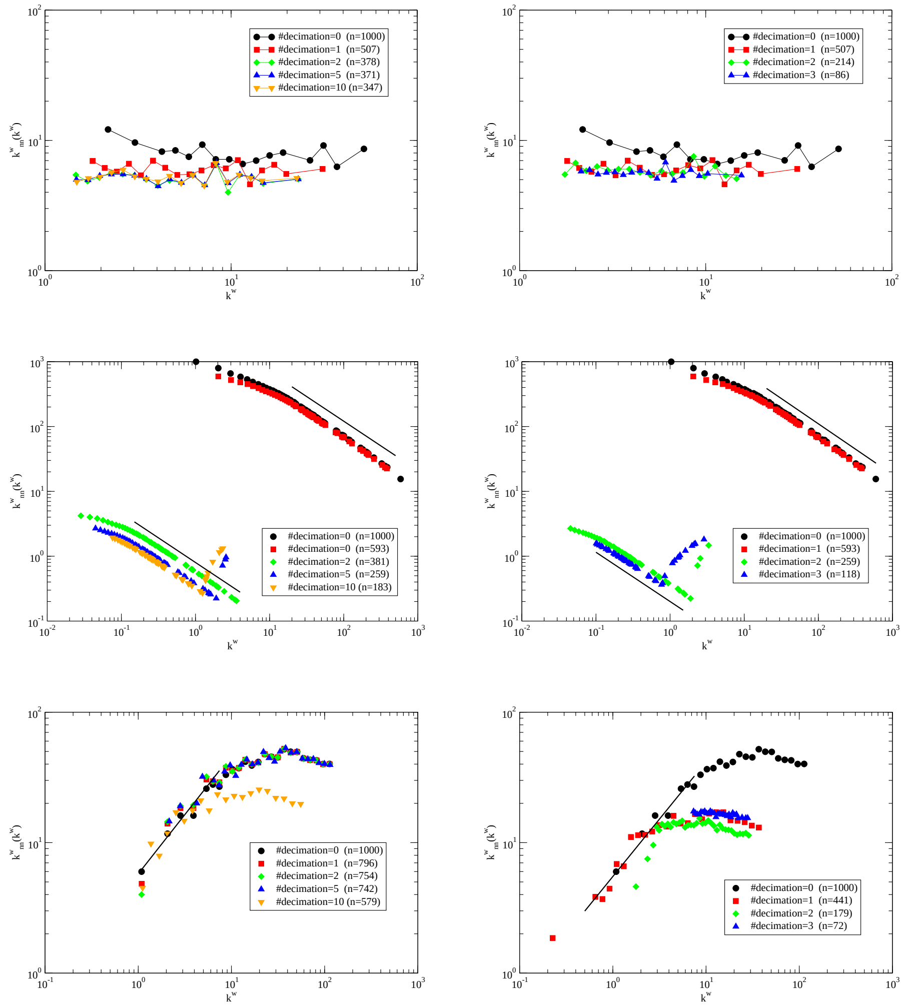


Figure 2.4: Results for $k^w_{nn}(k^w)$ obtained with renormalization by 'minimum'(left) and 'threshold' decimation (right). Top: BA network. Center: "Good-get-Richer" network. Bottom: Acquaintance network.

near criticality, studying a model which displays a phase transition from a non-critical behaviour to a critical one, depending on a parameter. Results might show that RG is able to distinguish between a critical behaviour and an apparently critical one.

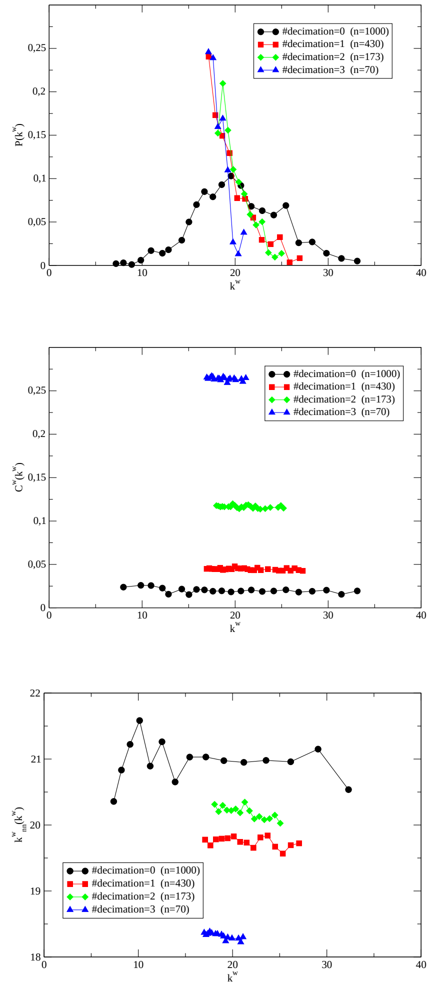


Figure 2.5: Results obtained with renormalization of Erdös-Renyi network through 'minimum' decimation. From top to bottom, results for $P(k^w)$, $C^w(k^w)$ and $k^w_{nn}(k^w)$ are shown.

Other decimation procedures can be studied, for example using weighted betweenness as decimation criteria.

A second, but not less important, aspect of the application of RG to complex networks deals with the problem of visualization. For networks of sizes up to some hundreds of nodes, a graph representation can still be used as a meaningful tool for analysis of network properties. Inspection by direct observation is an excellent method to study network structure, and it was indeed used in the first network studies. However, dealing with the large-scale networks encountered in real-world systems, consisting of a very large number of nodes, render direct visualization completely useless, since resulting pictures are noisy agglomerates of nodes and links where no structural property can be recognized. Thus, visualization of very large networks, even with the recent progresses in computer tools, still remains an open problem.

Renormalization applied to a complex network, since preserving its main features, could result in a smaller and more comprehensible network, suitable for direct meaningful visualization. This issue is still under investigation.

Chapter 3

Protein-protein Interaction Networks

Protein-protein interactions play crucial roles in virtually every cellular process, including DNA replication, transcription and translation, intracellular communication, cell cycle control and the workings of complex molecular motors. These biological functions rarely depend on single components, enlighting the fundamental importance of complex interactions between proteins. The recent possibility of collecting data on the global genomic and proteomic scale has created an unprecedented opportunity to develop comprehensive explanations for biological phenomena.

Global proteomic interaction data are synthetically represented as undirected networks exhibiting features far from the random paradigm which has dominated past effort in network theory. This evidence, along with the advances in the theory of complex networks, has triggered an intense research activity aimed at exploiting the evolutionary and biological significance of the resulting network topology.

Here we present a review of the results obtained in the characterization and modeling of the yeast *Saccharomyces cerevisiae* protein interaction networks obtained with different experimental techniques. We provide a comparative assessment of the topological properties and discuss possible biases in interaction networks obtained with different techniques. We report on dynamical models based on duplication mechanisms that cast the protein interaction networks in the family of dynamically growing complex networks. Finally we discuss various results and analysis correlating the network topology with the biological function of proteins.

3.1 Methods

The recent availability of complete genome sequences has strengthen the need for the development of new technologies such as high-throughput techniques to detect protein-protein interactions on a proteome-wide scale, interactions which were traditionally identified by small-scale experiments.

Here we will describe the current state of interaction-detection methods, making a distinction between the experimental techniques designed to identify physical bindings between proteins - such as yeast two-hybrid systems [26, 27] and mass spectrometry analysis of purified complexes of proteins [28,121] - and interaction prediction methods whose purpose is to detect functional associations between proteins, often underlying physical interactions [122, 123] - correlated mRNA expression profiles [124, 125], genetic interaction-detection [126, 127] and *in silico* approaches, such as gene fusion [128,129], gene neighborhood [122,130] and phylogenetic profiles [131, 132].

3.1.1 The two-hybrid system

The two-hybrid system is an experimental procedure able to detect pair-wise protein interactions. It exploits the modular properties typical of many eukaryotic transcription factors, which can be usually decomposed in two distinct modules, one directly binding to DNA (DB, DNA-binding domain) and the other activating transcription (AD, transcriptional activating domain) (fig. 3.1). The first component, DB, is able to bind to DNA even by itself, while the second module, AD, will activate transcription only if physically associated to a binding domain. This property is the result of a series of analysis made in the 80's by Ma and Ptashne [133] on transcription factors, while its use for the detection of protein interactions was first proposed by Fields and Song [134].

As it is illustrated in fig. 3.1, in any two-hybrid experiment two proteins are expressed as fusion proteins (*hybrids*) with a DNA-binding domain (DB, the bait) and a transcriptional activating domain (AD, the prey). Fusions partners are co-expressed in yeast nucleus where a protein-protein interaction is identified thanks to the activation of the reporter gene, which can be detected and measured.

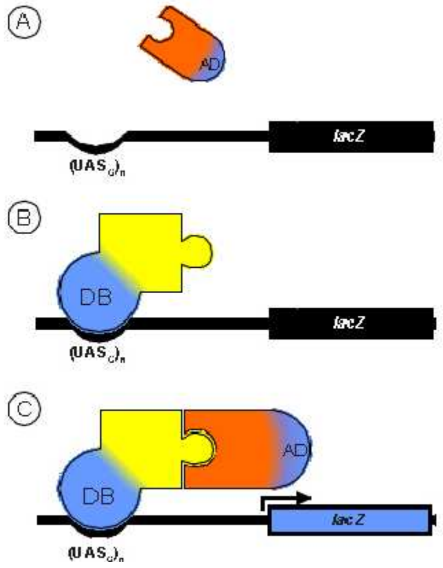


Figure 3.1: Two-hybrid assay. (A) A protein of interest is expressed as fusion protein to a transcriptional activating domain (AD). (B) A second protein of interest is fused to a DNA-binding domain (DB). (C) If the two proteins interact, transcription of the reporter gene is activated.

The two-hybrid system is able to identify virtually every protein-protein interaction. It is an *ex vivo* technique which can detect even transient, unstable or weak interaction, since the adopted strategy of the reporter gene implies a significant amplification. Moreover, it is simple, rapid, sensitive and inexpensive, because of the minimal requirements of a two-hybrid screen respect to, e.g., high quantities of purified proteins needed in traditional biochemical approaches. Indeed, it does not require any previous knowledge of the proteins to test and can be performed once the corresponding genes are known, thus being suitable for large-scale applications.

However, since it only detects binary interactions, it is not able to identify cooperative binding. Moreover, some kinds of proteins, such as transcription factors, cannot be used as fusions proteins to investigate their interactions, since they could activate transcription even in absence of any interaction. Also the extensive use of artificially made hybrids could represent a drawback, since it could lead to conformational changes in the proteins considered thus preventing transcriptional activation. This is one of the possible causes of false negative interactions, i.e. a true protein-protein interaction which is not detected by two-hybrid assays. This experimental procedure is also known to produce many false positives, identifying partners in screening procedures when no protein-protein interaction is present. Indeed, even if two proteins potentially interact into the nucleus, where this technique takes place, it could happen that they never find close to each other because of spatio-temporal constraints. For example, they could be localized in different cell types or in distinct compartments of the same cell, or even could be expressed at different times during the cell cycle. For this reasons, interactions detected by two-hybrid assays must be critically analyzed in order to assess their biological relevance.

3.1.2 Protein complex analysis

After the development of ultrasensitive mass spectrometric techniques for protein identification, new experimental procedures, besides two-hybrid screens, have been used to produce large-scale results for protein-protein interactions, such as purification of protein complexes. This procedure is made up of three main components: isolation of the bait or target protein, affinity purification of the complex and identification by mass spectrometry of proteins belonging to the complex. The protein of interest is isolated and fused to an affinity tag, by using one of the two protocols: tandem affinity purification (TAP) [28, 135] or high-throughput mass-spectrometric protein complex identification (HMS-PCI) [121]. TAP consists of two successive affinity purifications, using two tags fused with the bait and leading to the isolation of the target protein toghether with its associated proteins, as it is sketched in fig. 3.2. High-throughput mass-spectrometric protein complex identification, instead, employs a one-step immuno-affinity purification with transient overexpression of the target protein.

Comparison of results obtained through complex purification with yeast two-hybrid data shows a very small overlap [28]. A possible explanation could rely on the idea that cooperative binding embodied by complexes is not only the result

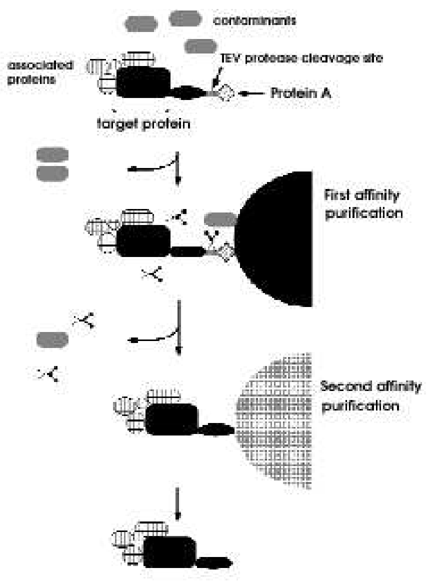

contaminants
associated proteins
TEV protease cleavage site
Protein A
target protein
First affinity purification
Second affinity purification

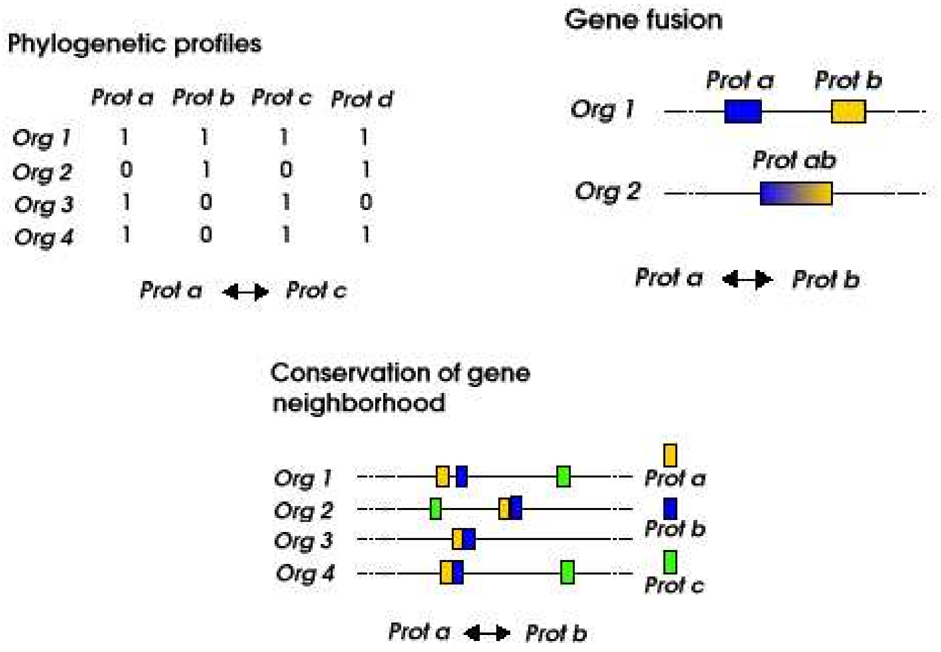
Figure 3.2: TAP procedure. The tagged bait undergoes two successive affinity purifications, leading to the isolation of its complex.

of a sum of pair-wise interactions, which completely lack any spatio-temporal information on proteins activity. Indeed, the main difference between complex purification methods and two-hybrid system relies in the identification of whole complexes isolated in a single step, thus detecting cooperative interactions between proteins which cannot result from two-hybrid screens, where the strategy adopted is based on the bi-modular properties of transcription factors. Moreover, it is an *in vivo* technique which employs only one artificially made protein (the bait), instead of two as in two-hybrid procedure, thus minimizing possible changes in conformational properties which could lead to steric interference. Complexes are found in physiological settings, since interactions take place in native environment. In order to test the validity of a complex identification, several components of the same complex can be used as tagged baits. However, the tagging procedure might interfere with complex formation and the purification process might loose weakly associated components of the complex.

3.1.3 Interaction prediction methods

Besides the physical interactions detected by the high-throughput experimental techniques described above, a complementary insight about protein-protein interactions is given by interaction prediction methods based on genomic information. From the analysis of genome sequences, these methods are able to identify functional associations between proteins (for a review, see [136]).

Prediction based on similar phylogenetic profiles look for the simultaneous presence or absence of two proteins in the genomes of different organisms (see fig. 3.3, top left). However, it requires complete sequencing of entire genomes, and is not suitable for essential genes.



	Prot a	Prot b	Prot c	Prot d
Org 1	1	1	1	1
Org 2	0	1	0	1
Org 3	1	0	1	0
Org 4	1	0	1	1

Figure 3.3: Interaction prediction methods. Top left: comparison of phylogenetic profiles. Top right: gene fusion events. Bottom: gene neighborhood conservation.

Gene fusion predicts an interaction between two proteins of a given organism, if they are part of the same polypeptide chain in another organism [128, 129] (fig. 3.3, top right).

The conservation of gene neighborhood in the genomes of different organisms (fig. 3.3, bottom) is interpreted as an indication of functional association between the proteins encoded by the two genes considered [122, 130, 137]. However, this occurs only in prokaryotic genomes, representing one of the main drawbacks of such approach.

Correlated mRNA expression predicts functional associations between proteins encoded by genes which show similar transcriptional responses to a change in the cellular status [124, 125]. Messenger RNA expression profiles can be measured under very different cellular conditions, thus representing an advantage respect to other techniques which can only take into account few settings.

A functional interaction between proteins can be detected also by synthetic genetic interactions [126, 127]: two non-essential genes show a synthetic lethal interaction if they cause cell death when simultaneously mutated [138, 139].

3.1.4 Data sets quality

Several databases have been recently compiled in order to collect and document the incredibly vast amount of large-scale data on protein interactions produced by high-throughput methods (for a review, see e.g. [140]). Their purpose is to analyze protein interactions in an attempt to comprehensively characterize the whole network of connections between proteins.

Data sets comparison should in principle take into account the different conditions under which interactions are detected, since they could lead to different results. Indeed, the intersection between different interaction data shows a surprisingly small overlap [141], underlining the need for a critical evaluation of the biological relevance of large-scale data.

The many discrepancies arising from data sets comparison could be due first of all to specific features of experimental methods, each of them characterized by its own advantages and drawbacks, so that results coming from one method may not largely overlap with those obtained with another technique because of specific restriction and different requirements. In this sense, different experimental techniques could be complementary, thus increasing our knowledge about the network. Secondly, these observations could be the result of low coverage of data sets, implying that a complete knowledge of the network of interactions has not been reached yet. However, it is also known that results produced by high-throughput methods are prone to *mistakes*. Indeed, although extensive, these data sets contain spurious interactions (false positives) and are missing many true interactions (false negatives) as well. For example, even referring to the same experimental technique, yeast two-hybrid assay, one can notice the incredibly small overlap among different datasets [142]; e.g. Ito's data and Uetz's data share only a very small percentage of interactions, the intersection of the two sets representing respectively about 4% and 14% of the total ensembles.

The assessment of the reliability of such data needs a comparison with a trusted reference set, in order to distinguish between validated interactions and background noise. Interactions detected by small-scale experiments could act as a benchmark, since they usually have been thoroughly investigated by multiple experiments and several checks. However, small-scale data set is not suitable to validate the majority of high-throughput data, because of the very limited number of high-confidence interactions it contains. The same problem is encountered when considering the intersection of different large-scale data sets of protein interactions. Indeed, it has been shown that connections detected by more than one method increase their accuracy respect to others, while however decreasing their coverage [143], resulting in a very small reference set.

For these reasons, the problem of investigating biological relevance and accuracy of protein interactions still represents a crucial step in analyzing protein-protein interaction data.

3.2 Topological analysis

Protein-protein interaction data, in the form of a list of binary interactions, is mathematically described as a network whose nodes represent proteins, connected by an edge if they directly interact. Only physical interactions detected by high-throughout methods are considered here.

In the following we study the topological properties of three distinct protein interaction networks of the yeast *Saccharomyces cerevisiae* obtained from different data sets:

network (I): a collection of binary interactions detected by two different two-hybrid assays [26,27], composed of a total of 2831 links among 2152 proteins;

network(II): interactions obtained from protein complex detection with tandem affinity purification techniques (TAP) [28]; it consists of 3221 interactions involving 1361 proteins;

network(III): a mixed collection of interactions obtained with different experimental techniques, documented at the Database of Interacting Proteins (DIP) [144]; it is composed of 4713 proteins and 14846 interactions.

It is worth to notice that, while (I) is composed of binary interactions between proteins directly detected by two-hybrid techniques, network (II) assigns hypothetical connections between proteins belonging to the same complex. Indeed, the topology inside a protein complex is not revealed by purification processes: not all associated proteins will in general interact with the bait, since the interaction could be mediated by other molecules, or interact with the bait at the same time, since interactions could occur under different physiological conditions. Therefore, for a direct comparison with pairwise interactions detected by other experiments, protein complex data have been assigned hypothetical interactions following two different models [141]: the *spoke* model, in which only interactions between the bait and associated proteins occur, and the *matrix* model, which assigns to a given complex all possible interactions between all proteins belonging to that complex, thus leading to cliques (i.e. fully connected sub-networks). In network (II) we have adopted the matrix model, since it has a higher coverage respect to the spoke model, although displays a lower accuracy when compared to a reference set [143].

In table 3.1 we report the size and the number of interactions of each network, toghether with the size of the giant component, i.e. the biggest connected sub-network. From a first topological insight, involving most basic features, we find that the three networks considered display different global properties. The small values of the average degree $\langle k \rangle$ of the three networks, compared to network sizes, point out that protein networks are almost sparse graphs. However, the values observed considerably differ one from the other, giving an indication of the differences in how well a protein is connected in each network. The average clustering coefficient $\langle C\rangle$ provides information about protein interconnectivity. We compare $\langle C\rangle$ computed on each network with the corresponding average clustering coefficient of a random network with the same degree distribution [145], expressed in terms of the first and second moment of the distribution:

$$\langle C_{rand}\rangle = \frac{1}{n}\frac{\left(\langle k^2\rangle - \langle k\rangle\right)^2}{\langle k\rangle^3} \tag{3.1}$$

The clustering coefficients for networks (I) and (II) are two orders of magnitude larger than the corresponding random ones, thus displaying a strong tendency to form 'triangles'. Also network (III) has a larger clustering coefficient respect to $\langle C_{rand}\rangle$, although it displays a lower ratio $\langle C\rangle/\langle C_{rand}\rangle$, meaning that it is less clustered than (I) and (II).

	(I)	(II)	(III)
# proteins	2152	1361	4713
# proteins *giant component*	1679 (78%)	1246 (91%)	4626 (98%)
# links	2831	3221	14846
$\langle k\rangle$	2.63	4.73	6.30
$\langle C\rangle$	0.10	0.22	0.09
$\langle C_{rand}\rangle$	0.0064	0.019	0.018

Table 3.1: Average global properties of networks (I), (II) and (III).

Differences from the random paradigm emerge also in the distribution of protein degrees, $P(k)$ (fig. 3.4), which shows a high level of heterogeneity in the connectivity properties of the three networks. Heavy-tailed degree distributions display a non-negligible probability of having proteins with very high degrees, much larger than $\langle k\rangle$, thus pointing out the existence of 'hubs', i.e. highly connected nodes, which play central roles in the connections among proteins with lower degrees.

Following Jeong *et al.* [29], we compare the observed $P(k)$ to a power-law with an exponential cut-off:

$$P(k) \simeq (k+k_0)^{-\gamma}e^{-k/k_c}\,. \tag{3.2}$$

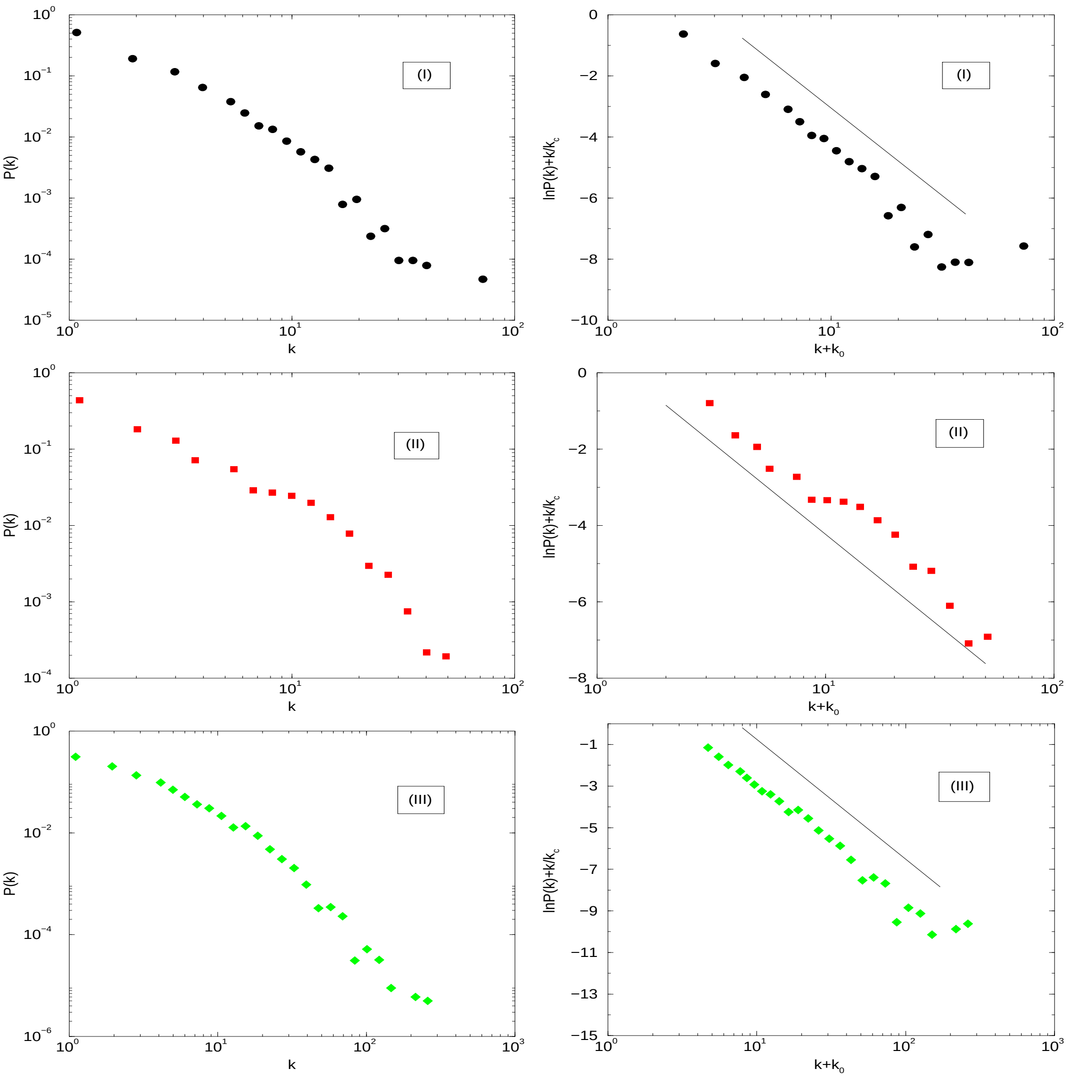


Figure 3.4: Degree distribution $P(k)$. On the left we report $P(k)$ in a double log-scale. On the right we plot $\ln P(k) + k/k_c$ (see text) as a function of $k + k_0$ on a single log-scale. From top to bottom: networks (I), (II) and (III). Lines plotted have slopes, respectively, 2.5, 2.1, 2.5.

Degree distributions of (I) and (III) are in good agreement with such functional form - a best fit of real data yelds power-law exponents $\gamma^{(\mathrm{I})} \simeq 2.5$ and $\gamma^{(\mathrm{III})} \simeq 2.5$

(slopes of solid lines in fig. 3.4, top and bottom), in good agreement with results of [29] concerning protein interaction data extracted from [26], and cut-offs $k_c^{(\mathrm{I})} \simeq 30$ and $k_c^{(\mathrm{III})} \simeq 100$. Interaction data derived from TAP experiments display a degree distribution which seems to deviate from the behaviour observed in (I) and (III), showing the presence of a 'bump' in the distribution for intermediate values of the degree. The solid line in fig. 3.4 (center) has a slope $\gamma^{(\mathrm{II})} \simeq 2.1$, representing the best fit to the data, using eq. 3.2.

The high level of heterogeneity, already encountered in other real-world networks, points out the emergence of a *scale-free* behaviour also in protein interaction networks, with large fluctuations in protein connectivity.

A deeper analysis of the topology of protein networks involves the study of the structural organization and of the degree correlations (see 1.1). The presence of a hierarchical organization of network structure can be characterized quantitatively by the clustering coefficient averaged over proteins with degree k. A non-trivial behaviour of $C(k)$ provides some hints on the presence of a hierarchy of proteins in the network, each characterized by a different degree of local interconnectivity, a fingerprint for modularity [84–89, 110, 111].

In fig. 3.5 (left) we report results for $C(k)$. Network (II) exhibits a clear heavy-tail which can be fitted to a power-law, $\sim k^{-0.48}$, while networks (I) and (III) do not display a scale-free behaviour. Two-hybrid data seem to remain almost constant for small degrees, exhibiting a drop for larger values of k, due to small network size, while the behaviour displayed by $C(k)$ for DIP data suggests the presence of a structural organization, although it is not a clear power-law $C(k) \sim k^{-\gamma_C}$. Results observed provide a strong and clear evidence for an inherent hierarchical organization only for network (II), but suggest the presence of a structural organization for the other networks, although characterized by weak and non-univocal signatures.

Degree-degree correlations are investigated by measuring the average degree of nearest neighbors of proteins with degree k, i.e. $k_{nn}(k)$ (fig. 3.5 on the right). Evidence of degree correlations are observed only in (III), which exhibits a power-law behaviour with exponent $\simeq 0.24$, whereas (I) and (II) display $k_{nn}(k)$ almost independent of k, thus displaying a lack of correlations.

Another interesting feature characterizing network architecture is the so-called *rich-club phenomenon* [146, 147], recently introduced as a quantitative metric to

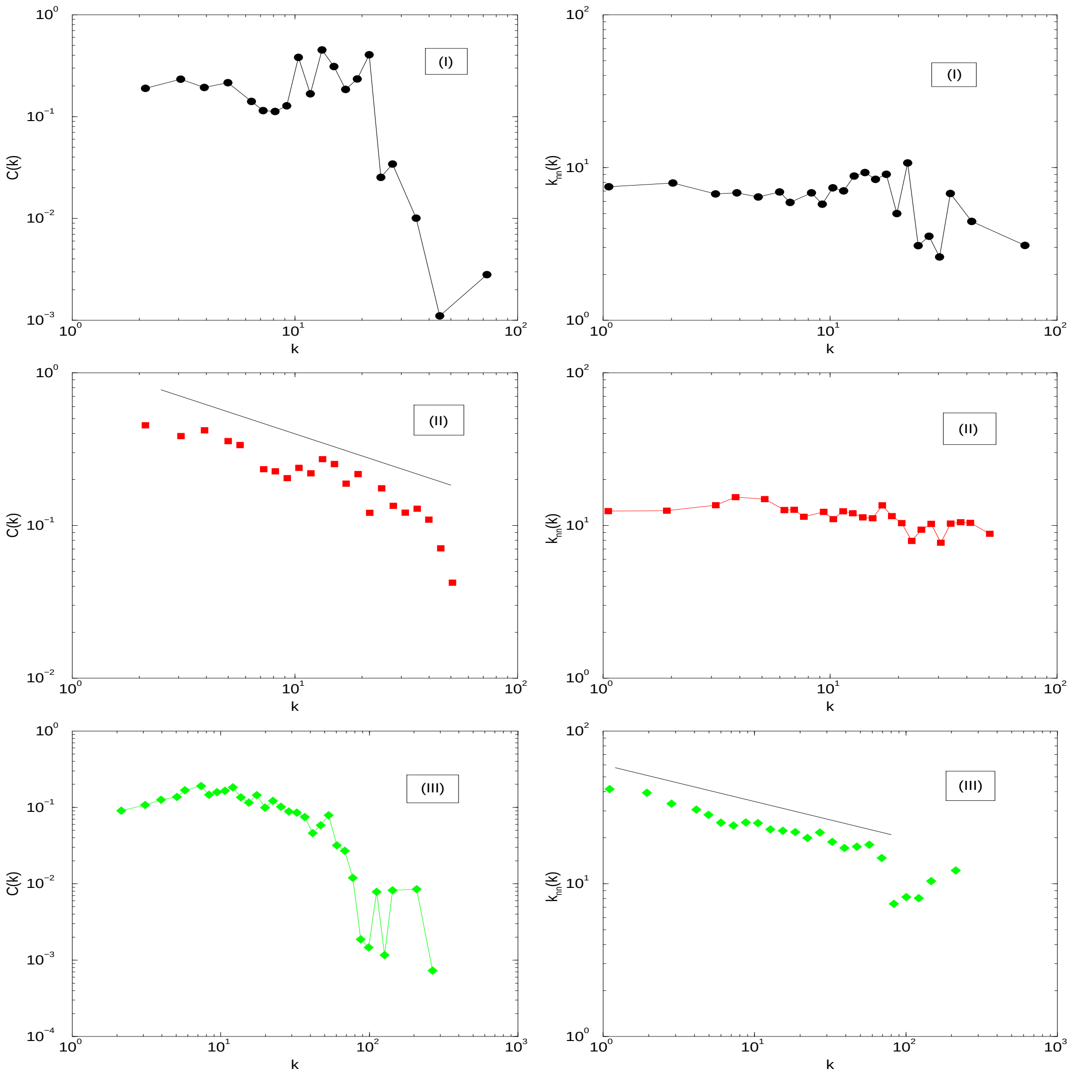


Figure 3.5: Average clustering coefficient $C(k)$ (left) and average neighbors degree $k_{nn}(k)$ (right) as a function of protein degree. From top to bottom: networks (I), (II) and (III). Clear power-law behaviours are observed for $C(k)$ in (II), with exponent $\simeq 0.48$, and for $k_{nn}(k)$ in (III), with exponent $\simeq 0.24$.

take into account the interconnectivity of highly connected nodes, also known as

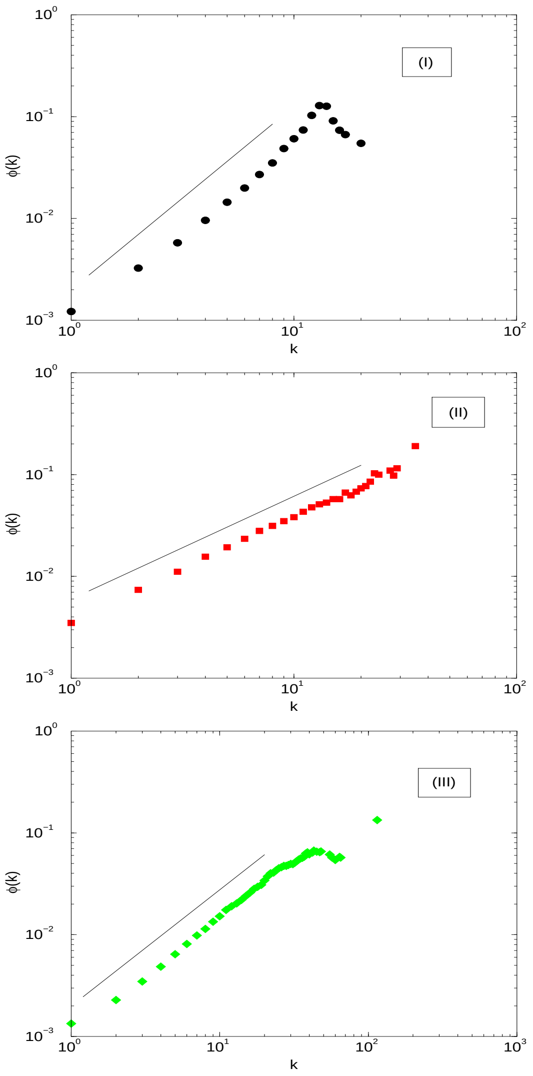


Figure 3.6: Rich-club phenomenon, measured by the quantity $\phi(k)$ (see eq. (3.3)) as a function of protein degree. From top to bottom: networks (I), (II) and (III).

'rich' nodes. It is defined as

$$\phi(k) = \frac{2e_{k>}}{n_{k>}(n_{k>}-1)} \tag{3.3}$$

where $n_{k>}$ represents the number of proteins having degree larger or equal to k

and $e_{k>}$ the number of links connecting proteins which belong to this ensemble. It thus measures the degree of interconnectedness between proteins having degree higher than a certain value k, and represents a relevant measure since the connectivity between rich nodes can be crucial for network properties and tasks, such as robustness and the performance of biological functions. Fig. 3.6 presents results obtained computing $\phi(k)$ in the three networks. Clear behaviours are observed, with networks (I) displaying power-laws with exponents $\simeq 1.96$, respectively, while (II) and (III) exhibit a linear dependance of the rich-club coefficient ϕ on the protein degree. The scale-free behaviour is an indication of the strong tendency of the system towards a higher cohesiveness among proteins with larger degree.

3.3 PIN renormalization

In this section we present the application of renormalization group (RG) approach, presented in chapter 2, to protein interaction networks. In particular, we show results relative to yeast networks (I) and (II), obtained with the 'minimum' decimation renormalization techniques (for details, see chapter 2). Starting with the giant component of (I) and (II), i.e. the biggest connected sub-graph of each protein network, of size, respectively, $n = 1679$ and $n = 1246$, we decimate iteratively the proteins with lowest weighted degree k^w and subsequently renormalize the laplacian $\mathbf{\Gamma}$, following eq. (2.11).

In fig. 3.7 we show results for both networks relative to renormalized weighted degree distribution $P(k^w)$, weighted clustering coefficient $C^w(k^w)$, and average neighbor weighted degree $k^w_{nn}(k^w)$, as functions of the weighted degree k^w.

Hierarchical organization properties and correlation features of renormalized protein networks, identified by the behaviours of $C^w(k^w)$ and $k^w_{nn}(k^w)$, confirm results already obtained in chapter 2 for abstract networks generated with different models. The power-law behaviour displayed by $C^w(k^w)$ in network (II) (fig. 3.7, center right) is recovered under successive application of the RG. The critical exponent of renormalized weighted clustering coefficient, as well as the drop observed for large degrees in the original network, is preserved even after the decimation of more than half of the proteins. The same occurs also for renormalized $C^w(k^w)$ in network (I) (center left), where the original behaviour, not identifiable with a

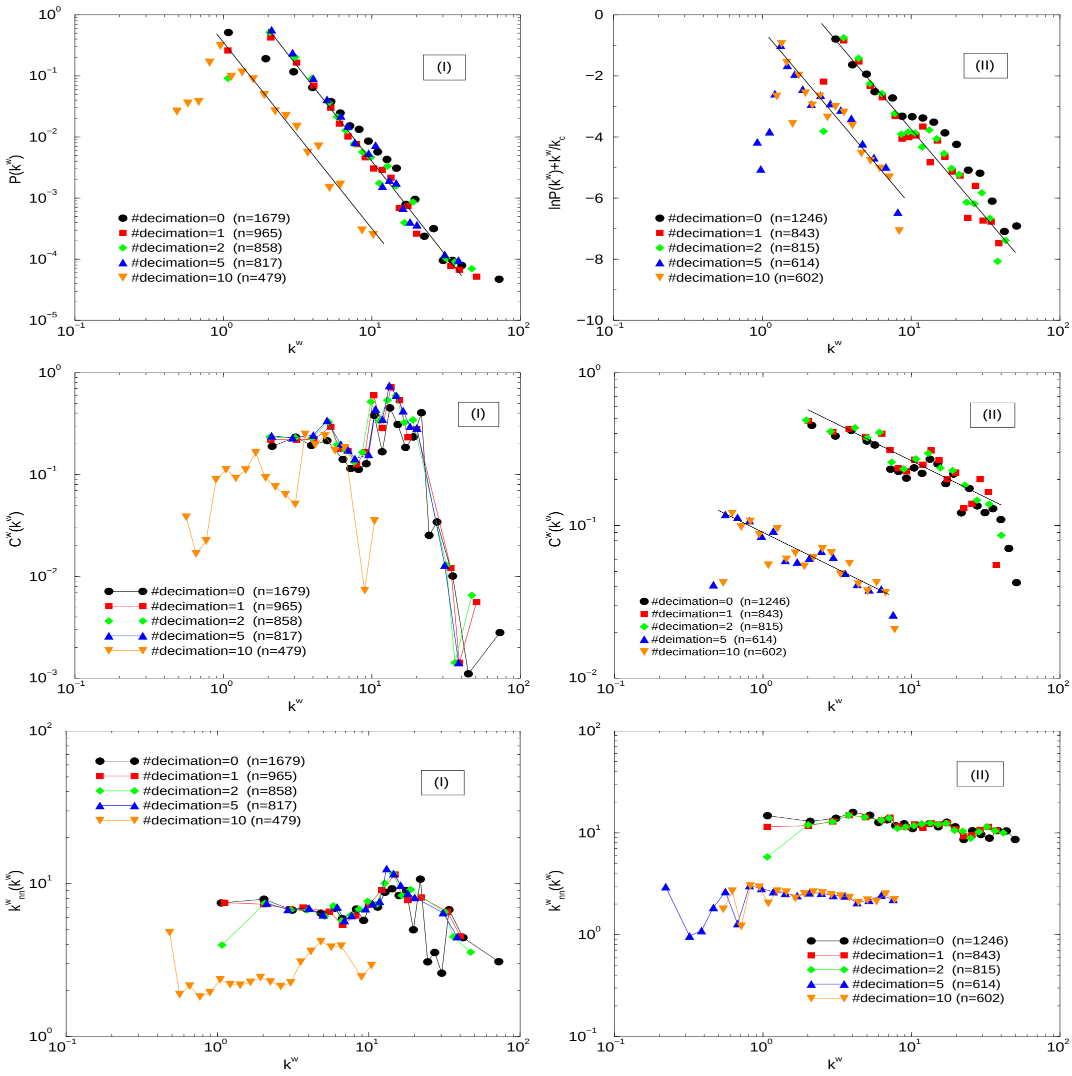


Figure 3.7: From top to bottom: results obtained for $P(k^w)$, $C^w(k^w)$, $k^w_{nn}(k^w)$ with the renormalization of network (I) (left) and (II) (right) by 'minimum' decimation. Number of decimations and relative network sizes are shown.

clear functional form, is preserved. The almost constant behaviours of $k^w_{nn}(k^w)$ observed in (I) and (II) (fig. 3.7, bottom) are preserved after renormalizations of the networks. The fluctuations observed for large values of the degree in the original network (I) lead to the emergence of a clear bump in the renormalized quantities.

A different behaviour is observed in the renormalized weighted degree distributions (fig. 3.7, top). While the original network (I) can be fitted with a power-law with an exponential cut-off (see eq. (3.2)), leading to a critical exponent of value ~ 2.5, all the renormalized distributions $P(k^w)$ corresponding to different decimations show a pure power-law behavior with exponent ~ 3, even at the tenth decimation which leaves the sytem composed of $n = 479$ proteins, i.e. less than 1/3 of the original size.

Also in network (II) renormalized $P(k^w)$ are all characterized by the same behaviour, which instead differs from the original one. Indeed, as already seen in the previous section, the functional form used to fit $P(k)$ in networks (I) and (III) - eq. (3.2) - is not in agreement with the degree distribution of network (II), which displays a change in its behaviour for intermediate values of k. Renormalized $P(k^w)$ instead follow a power-law behaviour with an exponential cut-off: a best fit of renormalized data, using eq. (3.2), yelds the same value ~ 2.5 for the critical exponents corresponding to successive decimations.

Results show that the RG applied to protein networks preserve its features for what concerns structural organization and connectivity correlations, but lead to different behaviour when considering the degree distribution. This could actually be a fingerprint of the ability of this technique to detect 'true' critical properties of the networks, making them clear after renormalization. Indeed, the degree distribution is more affected by small-scale details than the other quantities investigated. Thus, the presence of spurious effects which are known to exist in the protein interaction data sets, could result in a misleading behaviour of $P(k)$, while correct properties seem to emerge after a coarse-graining of less relevant details, thus eliminating also the *mistakes* contained.

3.4 Modeling

The development of models able to accurately reproduce the features observed in the topological analysis presented in the previous sections, could represent a fundamental improvement in explaining the origin of such features. Since protein-

protein interactions are involved in almost every biological process during cell life, the comprehension of the underlying mechanisms at the origin of the observed behaviours could lead to a better understanding of their role and significance in the cell.

In the following paragraphs, we first review two classes of models which take into account network growth as a fundamental ingredient [13,14,53,54,148] and explain some of the properties of PINs, such as scale-free degree distributions. Finally we present two more detailed models [31,32] based on specific mechanisms of protein networks that might have driven proteome evolution.

3.4.1 Dynamical models

In chapter 1 we have already examined a class of growing networks models which has been recently developed, aimed at capturing the origin of the heterogeneity typical of many artificial and natural networks and based on preferential-attachment related rule.

The preferential attachment assumption relies on the idea that new introduced nodes will more likely connect to existing nodes characterized by high connectivities. Indeed, this could be reasonable in some types of real networks, such as e.g. citation networks, where a new paper will more likely cite well-known, and thus highly citated works, considered as references for a particular topic. The actual applicability of preferential attachment as a realistic mechanism at work in network dynamics has been supported by studies in the evolution of real networks for which temporal data about the introduction of new nodes are available [17,149,150].

However, there are some real networks whose mechanisms for growth and evolution do not seem to be related to preferential attachment rule, although displaying power-law degree distributions. This is the case, e.g., of biochemical interaction networks [29,33–36], among which protein-protein interaction networks. Indeed, PINs are actually known to evolve on a very long time scale - so that a growing mechanism must be taken into account when developing models for proteome evolution - but the idea underlying BA model, i.e. the probability of new connections to a node dependent on its degree, does not find any reasonable applicability in this case.

Another class of models, known as vertex copying models [13,14,148], proposes a different mechanism which will be later on employed by more specific models designed for proteome evolution. Vertex copying models are based on the idea that some of the new nodes introduced into the network will acquire connections by copying already existing links of one node chosen randomly. They were developed in order to explain properties and structural phenomena observed in the Web, thus they consider directed networks.

Such mechanism of copying gives rise to an *effective* preferential attachment. In [148] the authors have proven that the distribution of the incoming degree follows a power-law with exponent $\gamma = (2-\beta)/(1-\beta)$, with $\gamma \in (2, +\infty)$, where β represents the probability of random edge addition and $(1-\beta)$ the probability of the copying mechanism.

The copying model has raised a new interest because of the possible applications of its processes in models for the evolutionary development of protein-protein interaction networks.

3.4.2 Duplication-Divergence models

Genomes of most organisms contain multiple copies of genes which are structurally and functionally closely related. According to a recent evolutionary theory [151, 152], genomes are thought to evolve through duplication of genes and subsequent diversification, occurring for the partitioning of the common ancestral functions (each gene of the duplicated pair undergoes functional degenerative mutation, while jointly retaining the entire functional annotation of the single ancestral gene), rather than for the evolution of new functions. Genome evolution corresponds, at a different level, to the evolution of the protein-protein interaction network whose nodes are represented by the proteins expressed by the genes. The process of gene duplication can be translated in the duplication of a protein sharing the same interacting partners as its ancestor, while divergence mechanisms lead to the loss or gain of interactions.

Two simple models for proteome evolution were first developed by Vazquez *et al.* [31] and by Solé *et al.* [32] to reproduce topological and large-scaling properties of protein-protein interaction networks. They are both based on the microscopic processes of duplication and complementary degenerative mutations. In [31] the algorithm for proteome evolution is as follows (see also fig. 3.8):

- *Duplication*: a protein i in the network is randomly chosen and duplicated, i.e. a new protein i' is created with links to each neighbor j of the protein i; an interaction between i and i' is created with probability p.
- *Divergence*: each neighbor j is considered; one of its two connections, with i or with i', is chosen and removed with probability q.

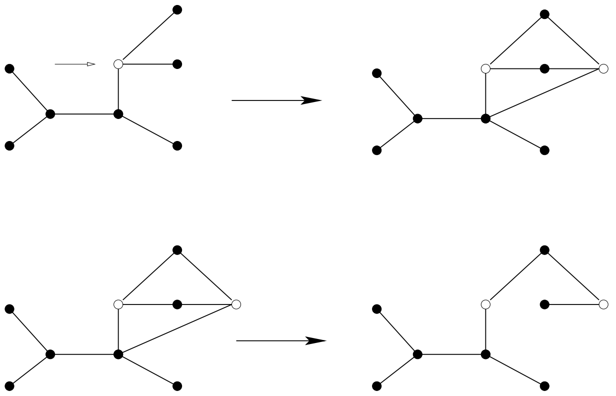


Figure 3.8: Duplication-Divergence (DD) model. Top: a node is duplicated (blank circle, indicated by the arrow). Bottom: some of the original or duplicate edges are removed with probability q.

The model by Solé *et al.* [32] takes into account also divergence due to addition of new connections in the network. The algorithm, sketched in fig. 3.9, is as follows:

(a) *Duplication*: a protein i in the network is randomly chosen and duplicated, so that the replicated protein i' share the same set of interactions as i.

(b) *Mutation by deletion of links*: links connecting the new protein i' to neighbors of i are deleted with probability δ.

(c) *Mutation by addition of links*: new links (not previously present) connecting the replicated protein i' with the rest of the nodes in the network are created with probability α.

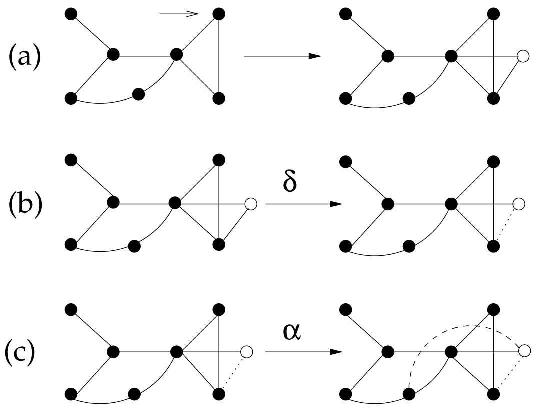


Figure 3.9: A model for proteome evolution: (a) duplication of a randomly chosen protein (indicated by the arrow); (b) mutation by deletion of links emerging from the new node (blank circle) with probability δ; (c) mutation by addition of random links starting from the new node, with probability α.

The process of mutation by addition of new links was found to have a probability much smaller than divergence due to deletion [153]. Vazquez *et al.* have tested the introduction of such mechanism in their model, without obtaining changes in the topological properties. Anyway, in order to have a finite average connectivity in Solé's model, the rate of addition of new links α must be inversely proportional to the network size, in agreement with the rates observed in [153].

Both models can be studied using a mean-field approximation. Here we follow [31], since the two approaches lead to analogous results in terms of the respective parameters introduced. The average degree $\langle k\rangle_{N+1}$ of the network with $N+1$ nodes can be expressed as:

$$\langle k\rangle_{N+1} = \frac{N\langle k\rangle_N + 2p + (1-2q)\langle k\rangle_N}{N+1} \tag{3.4}$$

where $2p$ represents the gain in average degree due to self-interacting link and $-2q\langle k_N\rangle$ the loss corresponding to removed connections for divergence process. In the continuum limit for large N, one obtains a differential equation whose solution shows two distinct behaviours, depending on the rate q. For $q > 1/2$, a finite average connectivity is reached, i.e. $\langle k\rangle = k_\infty = 2p/(2q-1)$, while for

$q < 1/2$, $\langle k \rangle$ diverges with N as N^{1-2q}. At $q = q_1 = 1/2$ a phase transition occurs. Networks obtained display multifractal connectivity properties, with a scale-free behaviour characterized by an infinite set of scaling exponents, a features that seems to be related to local inheritance mechanisms [154].

It is straightfroward to see that the biologically motivated local rules actually produce an effective preferential attachment, as in the copying model. Indeed, the probability that a node of the network with degree k gains one more link is given by the probability that one of its neighbors is duplicated (i.e. k/N) times the probability of its new link not to be removed (i.e. $1 - q$); thus, ignoring self-interactions:

$$\Pi(k) \sim (1-q)\, k/N \tag{3.5}$$

Because of degree changes due also to mutations in the duplicated links, a multifractal topology is obtained.

Other relevant quantities have been investigated in [31], such as the clustering coefficient which displays the correct behaviour, reaching a finite value for increasing network size. By optimizing the values of the two rates p and q in such a way to reproduce clustering coefficient and square coefficient values of the protein-protein interaction network of the yeast *Saccharomices cerevisiae*, the model is able to reproduce other quantities, such as average degree and degree distribution toghether with tolerance against random and selective deletion of nodes, which are in good agreement with experimental results (fig. 3.10).

In [32], approximate values of the rates δ and β are found by imposing the experimental value of the average degree of the yeast, toghether with estimations of the ratio α/δ from [153]. The degree distribution $P(k)$ obtained for networks of size comparable to yeast protein interaction networks (fig. 3.11) can be fitted by a power-law with an exponential cut-off, eq. 3.2, already used by Jeong *et al.* [29] to analyze the connectivity distribution of *Saccharomices cerevisiae*. The fit parameters, $\gamma = 2.5 \pm 0.1$ and $k_c \simeq 28$ are in good agreement with those found in [29]. Other quantities, such as clustering coefficient, average path length and size of the giant component, were quite well reproduced by the model.

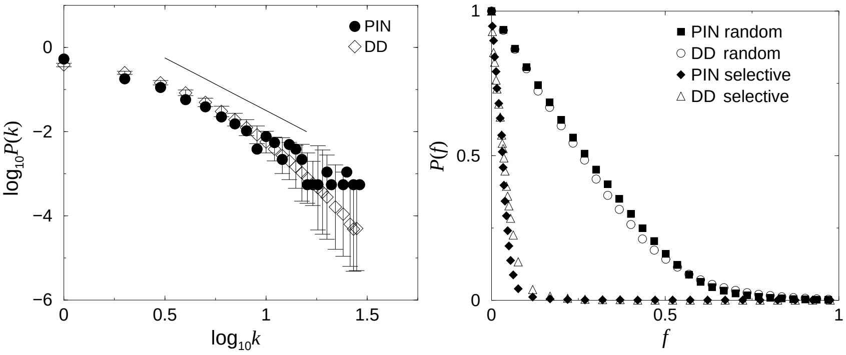


Figure 3.10: Duplication-Divergence model results. Left: Connectivity distribution of the protein interaction network (PIN) compared to DD model with optimized rates. The straight line is a power-law with exponent 2.5. Right: Fraction of nodes $P(f) = N(f)/N$ belonging to the giant component after a fraction f of nodes has been deleted. Comparison of DD model curves (averaged over 100 realizations) with experimental results.

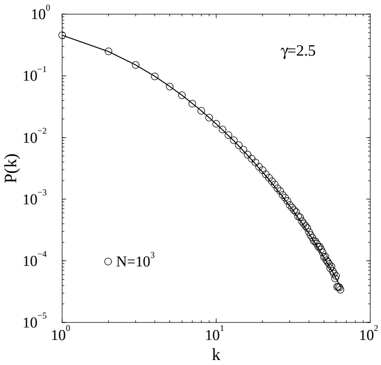


Figure 3.11: Degree distribution $P(k)$ for the model [32], averaged over 10^4 realizations of networks with size $N = 10^3$.

3.5 Functional characterization

Complete genome sequencing has not only accelerated the pace of discovery of new protein-protein interactions, through the development of high-throughput techniques like those described in section 3.1, but has also completely changed the view of protein function in biology [75, 76]. Indeed, it has led to a shift in biomedical research, from the study and analysis of single proteins to the investigation of the entire proteome. The traditional view of protein function as a task performed by a single protein independently from the others has been substituted by a more general context in which interactions between proteins play crucial roles when performing their activities. Several cellular processes are the outcome of complex interactions between proteins. Moreover, most proteins are not able to execute their tasks if they do not interact with other proteins, and alterations of protein interactions are shown to lead to many diseases.

The underlying network of interactions thus assumes a deeper meaning in terms of functional relationships between proteins, representing cooperative participation in performing functional tasks.

3.5.1 Topology/functionality correlations

In a work by von Mering *et al.* [143] about the quality of different protein interaction data sets, in terms of accuracy and coverage, it was shown that in highly accurate data sets (compared to a reference set, see section 3.1.4) functionally related proteins are more likely linked to each other, a feature which is usually exploited in function prediction models (see chapter 4) to infer functional annotation of unclassified proteins form classified neighbors. The authors computed the distribution of interactions according to functional categories and represented the results in terms of a matrix M whose generic element $M(\sigma_i, \sigma_j)$ represents the fraction of links between pairs of proteins performing, respectively, functions σ_i and σ_j. They found that the reference set adopted show considerably higher values along the matrix diagonal, thus in correspondance of shared functions between proteins.

Here we would like to go further in the investigation about the correlations between the pattern of interactions among proteins and their functionalities, with the purpose of reaching a deepened understanding of biological significance of network architecture.

Protein-protein interaction networks of the *Saccharomyces cerevisiae* considered are those already investigated in section 3.2 from a topological point of view, i.e. (I) the two-hybrid data in [26, 27], (II) the data set obtained with an experiment of tandem-affinity purification (TAP) [28], (III) a heterogeneous collection of interactions detected by different techniques, documented at the Database of Interacting Proteins (DIP) [144]. The functional classification was extracted from the MIPS database [155]; the finest functional classification scheme is made up of 424 different functional classes, while the coarse-grained one contains only 18 functional categories. The number of proteins in each data set with no defined functional classification (i.e. belonging to the categories named *"classification not yet clear-cut"* and *"unclassified proteins"*) is, respectively: 638 out of 2152 in (I), 279 out of 1361 in (II) and 1665 out of 4713 in (III).

In order to study quantitatively the feature discussed above, i.e. the likelihood of functionally related proteins to be connected, we compute the rate of interacting protein pairs sharing at least one functional category, in all three data sets examined. Unclassified proteins are not considered, since they lack any functional annotation. Classified proteins are taken into account with the whole set of functions performed by each of them, thus preserving multi-functionality, while in [143] genes annotated with more than one function were manually assigned to one category.

The values obtained adopting the coarse-grained level of functional classification - 83% of interactions between proteins with at least one function in common in two-hybrid data, 83% also for TAP data and 72% for the mixed data set (III) - seem to confirm previous observations. However, the sensitive decrease observed in the functional rate measured in (III) respect to (I) and (II) highlight the need for caution when interpreting DIP data results, since it might indicate the presence of a large amount of false positives.

To determine the actual significance of these results, we compare them with the rates of shared functionalities obtained in two distinct null models, compatible with the constraints embodied by the number of proteins belonging to each functional category. The first null model (NM1) is simply obtained by performing a *functional rewiring* of the network, i.e. starting from the protein interaction networks considered we choose at random two proteins p_i and p_j and exchange their annotations. Unclassified proteins are also considered in the rewiring and the underlying network is not modified. The algorithm is repeated a certain number of times, large enough in order to obtain a network composed of proteins which have randomly acquired functional annotations, still preserving the composition of each multi-functional annotation and thus the total occurrence of each functional category.

The second null model (NM2), instead, is based on a *random functional assignment* on the empty network. Starting from the network of interactions with no functional annotation on it, we randomly assign functions to proteins extracted with uniform probability, following three constraints: (a) the number of proteins belonging to each functional category must be conserved; (b) a protein cannot be assigned the same function twice; (c) the number of unclassified proteins must be conserved, i.e. the number of proteins with no functions at the end of the algorithm cannot exceed the original number of unclassified proteins.

Performing 100 realizations of each null model, we obtain the average values of the rate of interactions between proteins having at least one function in common, toghether with their standard deviations (see table 3.2).

$\text{rate}_{\text{link}\rightarrow\text{f common}}$	(I)	(II)	(III)
exp	82.90%	82.89%	72.36%
NM1	$(60.55 \pm 0.19)\%$	$(65.35 \pm 0.22)\%$	$(49.28 \pm 0.15)\%$
NM2	$(60.64 \pm 0.20)\%$	$(64.05 \pm 0.20)\%$	$(49.62 \pm 0.16)\%$

Table 3.2: Rates of interacting protein pairs sharing at least one functional category. Results obtained from the three networks (exp) are compared with the values averaged over 100 realizations of the two null models - NM1 and NM2 - described in the text.

We notice that the random rates of shared functionalities between interacting proteins obtained in the two null models for each network, do not differ one from the other, but are both considerably lower than the corresponding real values (exp in table 3.2) computed on experimental data. These observations seem to be an indication of the emergence of a correlation between physical link and functional association in protein-protein interaction networks.

Results shown are obtained considering the whole set of classified proteins, independent of their degrees. In order to investigate a possible dependence of the shared functional rate on degree, we have computed the same quantity for low and high connectivity proteins, with the average connectivity $\langle k \rangle$ being the separation value: with k_{small} we indicate low degrees ($k_{small} < \langle k \rangle$), while k_{large} refers to high degrees ($k_{large} > \langle k \rangle$). In table 3.3 we report results obtained from the experimental data corresponding to the three networks of proteins - (I), (II) and (III) - compared with results from the null models - NM1 and NM2 - where no distinctions based on protein connectivity are considered since, in these randomized models, functional annotation is by definition uncorrelated with topology.

rate $_{\text{link}\rightarrow\text{f common}}$	(I)	(II)	(III)
link($\forall k$, $\forall k$)	82.90%	82.89%	72.36%
link(k_{small}, k_{small})	80.85%	81.16%	63.71%
link(k_{large}, k_{large})	88.50%	85.33%	78.76%
link(k_{small}, k_{large})	75.30%	79.70%	63.17%
NM1	$(60.55 \pm 0.19)\%$	$(65.35 \pm 0.22)\%$	$(49.28 \pm 0.15)\%$
NM2	$(60.64 \pm 0.20)\%$	$(64.05 \pm 0.20)\%$	$(49.62 \pm 0.16)\%$

Table 3.3: Comparison among the rates of interacting protein pairs sharing at least one functional category computed on: the whole set of links (link($\forall k$, $\forall k$)); link(k_{small}, k_{small}) between proteins with *small* degree; link(k_{large}, k_{large}) between proteins with *large* degree; link(k_{small}, k_{large}) between proteins having respectively *small* and *large* degree, with the average connectivity $\langle k \rangle$ being the separation value. For comparison, we report also the values obtained with the two null models - NM1 and NM2.

A common behaviour can be observed in all networks: the rate for functional sharing between interacting proteins increases when considering two proteins with large degree k_{large}, while it is considerably lower when the connected pair is composed at least by a protein with small degree k_{small}, with the lowest value assumed in correspondance of the type of links (k_{small}, k_{large}). We have also investigated the role of peripheral proteins, i.e. proteins with degree $k = 1$, since they could represent biases in the computation, being affected by false interactions with higher probability respect to other proteins. We have thus computed the same quantities as before without considering peripheral proteins among those with k_{small}.

However, the trend, already observed before, is still present, even in absence of potential biases introduced by proteins with $k = 1$, thus showing a deeper correlation of functional characterization with topology, which should be investigated in the next future.

Since many proteins are annotated in multiple categories, we asked if there exist a correlation between the number of functions performed by each protein and its topological properties, such as, e.g., its connectivity. We investigated the average value of the number of functions per protein, $\langle \# f \rangle$, first considering the entire set of classified proteins, then making distinctions between low and high connectivity proteins and finally as a function of protein degree. We do not observe significative changes in $\langle \# f \rangle$ when considering only proteins with k_{small} or k_{large} in the results reported in table 3.4, except for a slight increase of $\langle \# f \rangle$ with increasing degree for (I) and (III) data sets, as it is also confirmed by the behaviour of $\langle \# f \rangle$, as a function of protein degree, observed in fig. 3.12. Despite expected fluctuations, no deviation from the value averaged over all proteins (line at constant value) seems to occur in TAP data considering the whole interval of degrees. Two-hybrid and DIP data, instead, display deviations for degrees larger than the average value.

$\langle \# f \rangle$	(I)	(II)	(III)
whole set of proteins	2.23	2.20	2.08
proteins with k_{small}	2.13	2.18	1.96
proteins with k_{large}	2.43	2.22	2.32

Table 3.4: Number of functions per protein averaged over: the whole set of proteins; low connectivity proteins (k_{small}); high connectivity proteins (k_{large}).

We also studied the distribution of the number of functions performed by each protein. We compared these results with the distributions computed only on proteins having k_{small} and k_{large} (fig. 3.13, left) and also with those obtained from the second null model (NM2) (fig. 3.13, right). We considered only NM2, since it allows random changes in the number of functions (for instance a protein might be assigned the whole set of functional categories), while NM1 preserves multi-functionality as it is, by definition of the model, allowing only exchanges

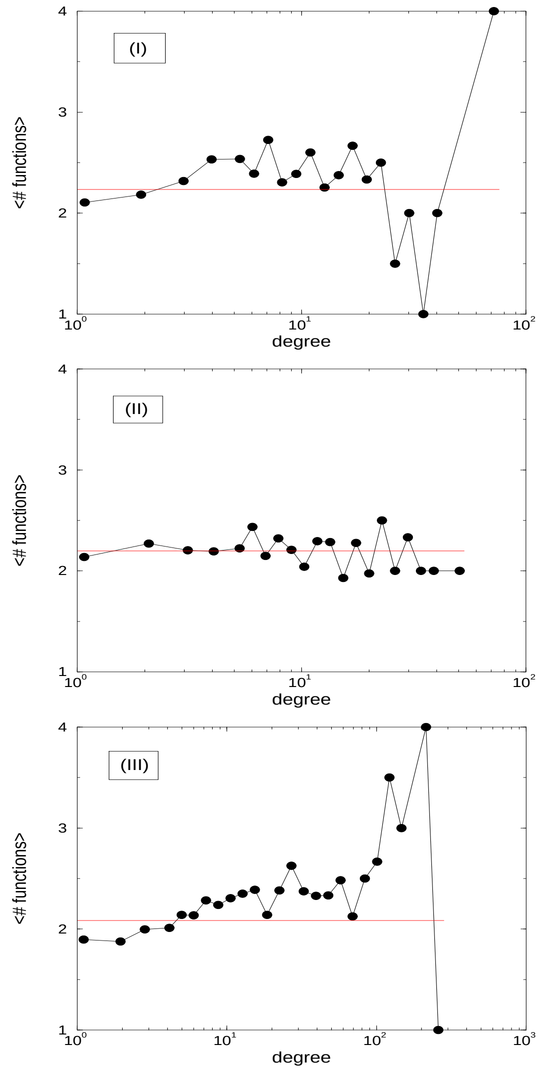


Figure 3.12: Average number of functions per protein, $\langle \#f \rangle$, as a function of protein degree. The line at constant value indicate the average of $\langle \#f \rangle$ over all proteins. From top to bottom, results relative to networks (I), (II) and (III) are shown.

between proteins.

Figure 3.13(right) does not show sensitive deviations of the NM2 distribution respect to the experimental one. A slightly different behaviour can be observed in fig. 3.13(left) where only small deviations seem to occur between the distribution computed on proteins with different degree, k_{small} or k_{large}, with networks (I) and (III) showing a common trend, opposite to the one displayed by network (II). In the first and third network, indeed, the distribution of function performed by proteins with large degree increase its values for larger values of the number of functions, while the opposite occurs for proteins with small degree, which enhance the probability of having only one function. An opposite behaviour can be observed for what concerns network (II).

These results reveal the need to go further in our investigations about topology and functionality correlations.

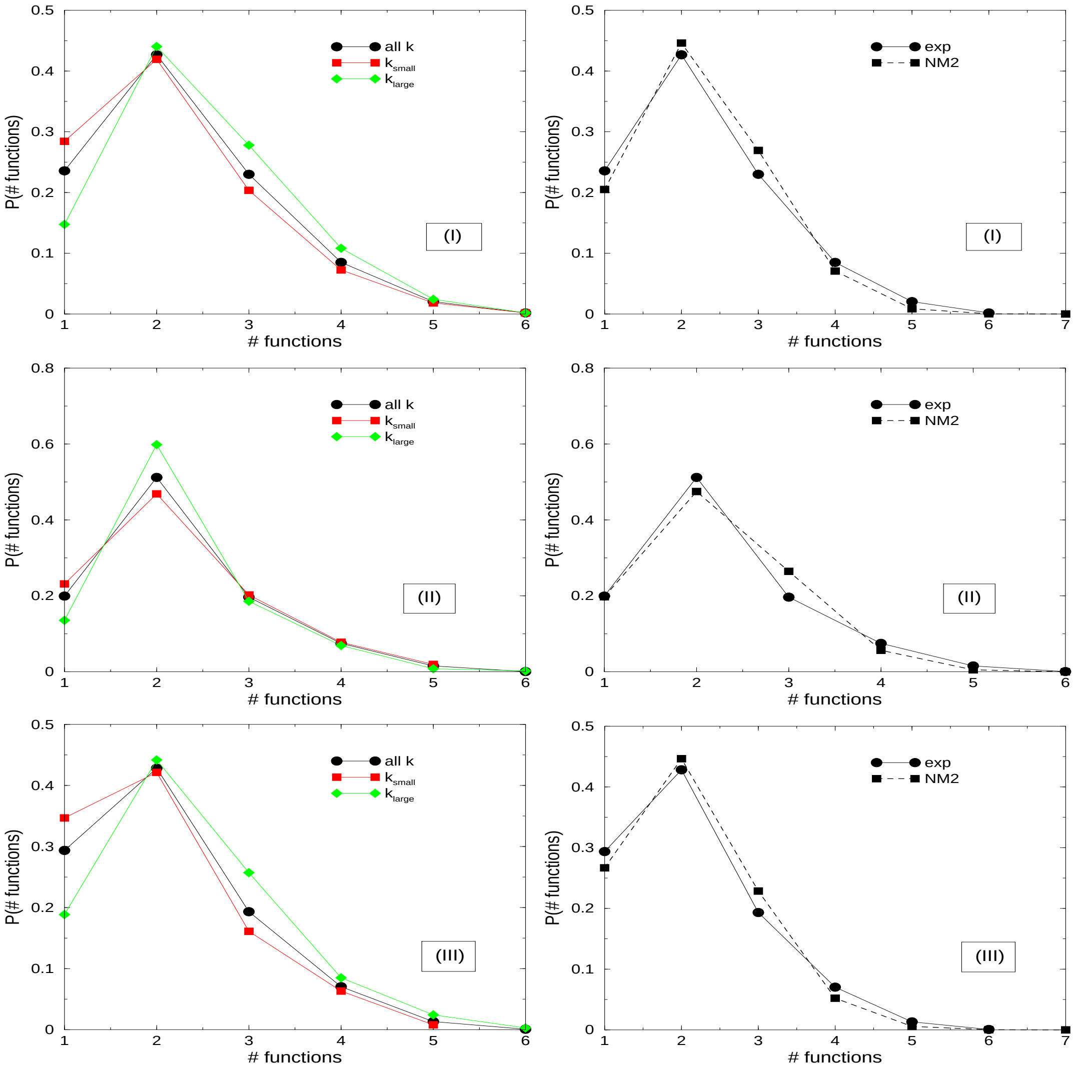


Figure 3.13: Distribution of the number of functions performed by each protein. Left: Comparison between the distribution obtained considering all classified proteins (all k) and those obtained taking into account only proteins with small degree (k_{small}) and large degree (k_{large}). Right: Comparison between values obtained from experimental data (exp) and from a randomized version (NM2), averaged over all classified proteins. From top to bottom, results relative to networks (I), (II) and (III) are shown.

Chapter 4

Protein Function Prediction

Despite the impressive progresses performed during the last years in genome sequencing and high-throughput proteomics techniques, a great amount (about 30%) of encoded proteins per completely sequenced genome are still functionally uncharacterized [74], revealing the need for new methods to deduce specific functional roles of these proteins, thus fully exploiting genome data.

Several approaches have been developed to facilitate the functional annotation of proteins for which we have few or no functional information at all [26–28, 121, 122, 128–132, 156–160]. In the following we present two distinct bioinformatics methods for function prediction, applied to the protein interaction network of the yeast *Saccharomyces cerevisiae*, and designed to take full advantage of the observed correlations between the pattern of interactions among proteins and their functionality. The two new approaches are compared with earlier network-based methods [26, 27, 156–160].

4.1 Optimization models

The strategy usually underlying function prediction models based on the network of protein interactions [26, 27, 156–160] relies on the assumption that the more two proteins are close to each other in the network, the more likely their functional annotations will be closely related. Indeed, as illustrated in the previous chapter (see table 3.2), the percentage of protein binding pairs sharing at least one function is about 83% (for the coarse-grained level of MIPS functional classification [155]) in two-hybrid data, 83% for TAP data, while it decreases to 72% for DIP data set. As already analyzed in section 3.5, these values represent a fingerprint of the functional relevance of connections between proteins, since they are considerably higher than the random rates obtained in two different null models considered.

This assumption has been exploited in the 'majority rule' method [156, 157] which assigns a putative function to proteins with no defined functional characterization on the basis of the functional annotation of its classified binding partners. It assigns to an unclassified protein the most common function(s) among the ones performed by its classified neighbors, without taking into account possible contributions coming from the unclassified ones. Therefore, the algorithm is not exploiting the whole network, since it is completely loosing possible information coming from those links which connect unclassified proteins. Indeed, unclassified proteins having partners with no defined functional classification should be influenced by the functional annotation assigned them by the method itself, leading to a final functional configuration which is consistent with the rules on which the method relies. Moreover, the approach can be applied only to a reduced number of proteins, since a great amount of proteins with unknown function interact only with unclassified partners.

The 'global optimization model' [158], instead, takes into account the whole set of interactions of each uncharacterized protein. The key point is that the functional annotation for proteins with unknown functions proposed by the model is obtained self-consistently, looking at shared functionalities between unclassified proteins and their classified and unclassified partners. To each functional assignment is associated a score which assigns the value -1 to each link between unclassified proteins or between classified and unclassified proteins when they share a common function:

$$E^{(1)} = -\sum_{i<j} J_{ij}\delta_{\sigma_i,\sigma_j} - \sum_i h_i^{(1)}(\sigma_i) \tag{4.1}$$

where $J_{ij} = 1$ only if i and j are both unclassified and directly connected (i.e. first neighbors), otherwise $J_{ij} = 0$, σ_i is the function of protein i, $\delta_{l,m}$ is the discrete delta function and $h_i^{(1)}(\sigma_i)$ represents the number of classified first neighbors of protein i sharing the same function σ_i. Majority rule assignment refers solely to unclassified proteins connected with at least one classified protein; therefore, its cost function is embodied by the second term only of the right hand side of eq. (4.1), completely neglecting pairs of interacting proteins with no functional annotation.

The functional assignment provided by the method is the one which minimizes the score function introduced. Since the problem cannot be solved anymore by taking into account each protein separately, a global optimization ought to be performed. If an unclassified protein interacts only with classified partners, the majority rule algorithm is recovered; if, instead, it has one or more partners with no defined classification, the functional annotations of those partners also matter and consistently influence its classification.

In the following sections we present two distinct computational methods for function prediction. They have been applied to three distinct protein-protein interaction networks of the *Saccharomyces cerevisiae*, obtained with different experimental techniques. The first one (I') is the same already studied in [158], which was obtained with a yeast two-hybrid assay [156]. It differs from protein network (I) investigated in chapter 3 because it does not contain Ito's "core" data set. It is composed by a total of 2238 identified interactions among 1826 proteins, yelding an average connectivity of 2.45 interactions per protein. Second and third data sets are the same already presented and analyzed in chapter 3.

The functional classification considered was extracted from the MIPS database [155], as in section 3.5; the level of functional classification adopted is the coarse-grained one containing 18 functional categories plus 2 categories indicating proteins with no defined functional classification, named *"classification not yet clear-cut"* and *"unclassified proteins"*. The number of uncharacterized proteins is, respectively: 461 out of 1826 in (I'), 279 out of 1361 in (II) and 1665 out of 4713 in (III).

The two bioinformatics methods developed are both based on a global optimization process involving the whole network, in which an opportunely introduced cost function, dependent on a functional assignment, is minimized through simulated annealing techniques [161, 162] (see Appendix A for further details). The predictions of both new methods, together with their robustness, have been statistically analyzed and evaluated through several tests. The introduction of new ingredients, such as information acquired on the network of interactions, respect to previous approaches based on protein-protein interaction data, demonstrated to be crucial in order to obtain higher quality results in the prediction of functional annotation of uncharacterized proteins. We extend the cost function of global optimization model (called GOM$_1$ in the following) to take into account more throughfully the topology of the network (mixed GOM method) and the observed correlations between the functions of interacting proteins (Maximum Entropy Estimate (MEE) method). We discuss separately the two methods.

4.1.1 Mixed Global Optimization Model

First, we developed a generalization of GOM$_1$ from a topological point of view, focusing the attention not only on directly interacting proteins, but even on proteins which are two edges away, i.e. *second neighbors.*

Analyzing the protein-protein interaction networks considered, we computed the fraction of second neighbors pairs having at least a functional category in common, analogous to what we have already done for binding partners. We obtained that in (I') about 75% of proteins two edges away share at least one function; similar results (77%) are observed in (II), while this percentage decreases to 63% considering (III), following the same trend already observed in connected proteins.

Experimental reasons support our assumption of not limiting the information we want to extract from the network to first neighbors only. Indeed, it has been recently demonstrated [163] that many experimental methods are not able to distinguish between a direct link connecting two proteins and an interaction mediated by at least a third protein.

Moreover, the importance of complex patterns of interactions which might embody functional associations could find an explanation in the framework of the genome evolutionary theory, introduced in section 3.4 to present the specific mechanisms thought to be responsible for proteome evolution. The processes of duplication and divergenge occurring at the proteome level lead to a pattern of interactions in which the products of duplicate genes consist of two functionally related proteins, being second neighbors if they retain at least one binding partner in common.

In a recent work by Samanta and Liang [159], the presence of topological redundancies in protein-protein interaction networks has been investigated in order to study a possible relation between shared partnerships observed and functional associations. It was shown that if two proteins share a significantly high number of common interactors, respect to what one would expect if interactions would be randomly rearranged, they display close functional associations. Analysis of such redundancies reveals that a great amount of protein pairs sharing an unexpected high number of common partners are not directly connected - about a 70% of the investigated pairs - showing that close functional associations can reasonably be found even between second neighbours.

To extract functional information also from proteins at distance 2, we introduce a second score function, analogous to the one introduced in [158] concerning first neighbors only. It counts the number of second neighbors pairs sharing at least a functional category.

$$E^{(2)} = -\sum_{i<j} S_{ij}\delta_{\sigma_i,\sigma_j} - \sum_i h_i^{(2)}(\sigma_i) \tag{4.2}$$

where $S_{ij} \neq 0$ only if i and j are both unclassified and connected through a path of length equal to 2 (i.e. second neighbors) and $h_i^{(2)}(\sigma_i)$ represents the number of classified second neighbors of protein i sharing the same function σ_i. We will refer to the global optimization of eq. (4.2) as GOM$_2$ in the following.

Starting from an initial random functional assignment for unclassified proteins, the two score functions, eq. (4.1) involving first neighbors and eq. (4.2) involving second ones, are optimized independently using simulated annealing techniques. Because of the frustration naturally arising in the system, the global optimization process leads to the presence of multiple optimal solutions, characterized by equal or very close values of their cost functions. Indeed, the system is not able to satisfy at the same time the requests for shared functionalities for all pairs of neighbouring proteins, due to the constraints imposed by classified proteins on their partners with unknown function. Therefore, the resulting computational problem is generally frustrated and the whole set of possible minimum configurations has to be taken into account. In order to do that, the optimization process is repeated 100 times, and for each unannotated protein, i, the frequency of occurrence, ν_i^f, of every function, f, is reported. The functional annotation proposed by each global optimization for each protein with no defined classification consists of the function(s) with the highest frequency of occurrence.

We have investigated the role of topological redundancies identified in [159], since they are represented by pairs of proteins which have an unexpected high number of common partners, i.e. second neighbors connected through several paths of length 2. We analyzed the predictive capacity of GOM$_2$ in two different cases, when the cost function assigns value -1 to each second neighbors pair (thus ignoring multiple connections) or to each path of length 2 connecting that pair. Results concerning GOM$_2$, reported in the following, have been obtained using the latter method since it was the most reliable in the tests we have performed. Thus, S_{ij} in eq. (4.2) differs from 0 only if proteins i and j are second neighbors and both unclassified, and measures the number of paths of length 2 connecting the given proteins.

Each global optimization process proposes a specific functional annotation for the whole set of unclassified proteins, corresponding to the minima of eqs. (4.1) and (4.2), respectively. The functional classification predicted by mixed GOM is obtained by merging the results coming from the two independent score functions used in GOM$_1$ and GOM$_2$.

4.1.2 Maximum Entropy Estimate Model

The second computational method for function prediction we propose depends on the k-point correlation functions evaluated on the network. Making use of a statistical inference criterion called maximum-entropy estimate, borrowed from information theory [164], we are able to determine the probability distribution consistent with the correlation functions computed on the basis of partial knowledge. The study reported here is restricted to $k = 1$ and $k = 2$, i.e. only to one-point and two-point correlation functions, but a straightforward extension can be done to include also k-body interactions with $k > 2$, such as for example functional correlations with second neighbors ($k = 3$).

The idea leading to the development of MEE is to make full use of the information which could be learned from the partial knowledge of functional characterization of the proteins belonging to the network of interactions. The interaction $J_{ij}\delta_{\sigma_i,\sigma_j}$ which appears in eq. (4.1) assumes values different from 0 only if two connected proteins, i and j, have at least a function in common and is justified by the observed rates of functional sharing between interacting proteins. A possible generalization of this interaction might include the correlations between different functional categories, σ and σ', instead of limiting only to functional interactions between identical functions underlying a physical connection.

We first consider a homogeneous network of proteins, each of them with degree z and characterized by a unique function in order to derive the cost function of MEE starting from the observed functional correlations, and then generalize to our case of an inhomogeneous network of proteins with multi-functional annotation.

The probability that two connected proteins have functions σ and σ' is:

$$\rho(\sigma,\sigma') = \frac{1}{L}\sum_{<i,j>} \delta_{\sigma_i,\sigma}\delta_{\sigma_j,\sigma'} \tag{4.3}$$

where L is the total number of links and $< i,j >$ represents the connected pair of proteins i and j. Summing over σ', we obtain the probability for a protein to have function σ:

$$\rho(\sigma) = \sum_{\sigma'} \rho(\sigma,\sigma') = \frac{1}{N}\sum_{i} \delta_{\sigma_i,\sigma} \tag{4.4}$$

with N being the total number of proteins in the network. Following the maximum-entropy estimate [164], we determine the total probability distribution $\rho_T(\{\sigma\})$ to find the network in the functional configuration $\{\sigma\} = (\sigma_1, \sigma_1, \ldots, \sigma_N)$, as the one which has maximum entropy subject to the information extracted from our partial knowledge. In other words, the probability distribution $\rho_T(\{\sigma\})$ is obtained from the maximization of the entropy

$$S = -\sum_{\{\sigma\}} \rho_T(\{\sigma\}) \ln \rho_T(\{\sigma\}) \tag{4.5}$$

subject to the constraints:

$$\rho(\sigma,\sigma') = \sum_{\{\sigma\}} \rho_T(\{\sigma\}) \frac{1}{L}\sum_{<i,j>} \delta_{\sigma_i,\sigma}\delta_{\sigma_j,\sigma'} \tag{4.6}$$

$$\rho(\sigma) = \sum_{\{\sigma\}} \rho_T(\{\sigma\}) \frac{1}{N}\sum_{i} \delta_{\sigma_i,\sigma} \tag{4.7}$$

$$1 = \sum_{\{\sigma\}} \rho_T(\{\sigma\}) \tag{4.8}$$

With the introduction of Lagrange multipliers, $\lambda(\sigma,\sigma')$, $\lambda(\sigma)$, μ, conjugate respectively to each constraint, the resulting probability distribution is of the Boltzmann form, $\rho_T(\{\sigma\}) \propto \exp(-H)$, with H being:

$$H(\{\sigma\}) = -\sum_{<i,j>} \lambda(\sigma_i,\sigma_j) - \sum_i z\lambda(\sigma_i) + \mu \tag{4.9}$$

which can be rewritten in terms of a single sum over all connected pairs of proteins:

$$H(\{\sigma\}) = -\sum_{<i,j>} \hat{\lambda}(\sigma_i,\sigma_j) + \mu \tag{4.10}$$

where

$$\hat{\lambda}(\sigma,\sigma') = \lambda(\sigma,\sigma') + \lambda(\sigma) + \lambda(\sigma') \tag{4.11}$$

The problem is now to determine $\hat{\lambda}(\sigma,\sigma')$ in terms of the observed quantities $\rho(\sigma,\sigma')$ and $\rho(\sigma)$. We consider a Bethe approximation following the cluster variation method (CVM) [165,166], which enables us to write an approximate expression of the entropy in terms of cluster entropies associated to clusters of proteins. We adopt a pair approximation, which takes into account single sites and pairs configurations of proteins and corresponds to the Bethe approximation. Thus, the entropy of the system can be written as

$$\begin{aligned} S = &-\sum_{<i,j>}\sum_{\sigma_i,\sigma_j} \rho(\sigma_i,\sigma_j)\ln\rho(\sigma_i,\sigma_j) + \\ &+\sum_i (z-1)\sum_{\sigma_i}\rho(\sigma_i)\ln\rho(\sigma_i) = \\ &= L\sum_{\sigma,\sigma'}\rho(\sigma,\sigma')\ln\rho(\sigma,\sigma') + \\ &+ (2L-N)\sum_{\sigma}\rho(\sigma)\ln\rho(\sigma) \end{aligned} \tag{4.12}$$

We can now minimize the free energy $F =< H > -S$, using eq. (4.12), with the following constraints

$$\sum_{\sigma,\sigma'}\rho(\sigma,\sigma') = 1 \tag{4.13}$$

$$\sum_{\sigma'}\rho(\sigma,\sigma') = \rho(\sigma) \tag{4.14}$$

and obtain the final expression of the cost function we were looking for:

$$\begin{aligned} E &= -\sum_{<i,j>} \hat{\lambda}(\sigma_i, \sigma_j) = \\ &= -\sum_{<i,j>} \ln \frac{\rho(\sigma_i, \sigma_j)}{(\rho(\sigma_i)\rho(\sigma_j))^{\frac{z-1}{z}}} \end{aligned} \tag{4.15}$$

where $z = 2L/N$ is the constant degree of the homogeneous network considered.

Going back to protein-protein interaction networks, we must face topological problems coming from the inhomogeneity of the network and problems derived from the multi-functionality associated to each protein. Indeed, going from a constant degree z to a variable degree z_i of protein i, the quantity defined in eq. (4.4) does not represent anymore the probability of finding the function σ performed by a protein in the network. We have, in fact:

$$\rho(\sigma) = \sum_{\sigma'} \rho(\sigma, \sigma') = \frac{\sum_i z_i \delta_{\sigma_i, \sigma}}{\sum_i z_i} \tag{4.16}$$

Thus, we must be cautious in deriving the cost function in case of a variable degree z_i with the same criteria and approximations used to obtain eq. (4.15), since eq. (4.16) has no more the meaning of a single-point probability, but is an average weighted with the coordination of each site.

Moreover, when evaluating $\rho(\sigma, \sigma`)$ and $\rho(\sigma)$, one has to take into account that often more than a single funtion is associated to each protein and that the statistics necessary to get good estimate of them is not always high enough.

Several tests have been performed in order to verify the validity of possible generalizations of eqs. (4.3), (4.4) and (4.15), consistent with the general approach outlined above, to include topological inhomogeneity and multi-functional characterization. Non-homogeneous features of our network can be simply introduced in the theoretical framework by defining $\rho(\sigma)$ as it is defined by the first equality in eq. (4.4), i.e. as $\sum_{\sigma'} \rho(\sigma, \sigma')$, ignoring it is no more a measure of the frequency of occurrence of function σ (second equality in eq. (4.4)). Concerning multi-functionality, instead, we can have limited confirms from the tests performed, because of our partial knowledge of spatio-temporal characterization of the functional tasks each protein performs.

In analogy with eq. (4.3), we thus define the two-points probability as

$$\rho(\sigma,\sigma') = \frac{1}{L}\sum_{<i,j>}\frac{1}{F_iF_j}\sum_{f_i=1}^{F_i}\sum_{f_j=1}^{F_j}\delta_{\sigma_i^{f_i},\sigma}\;\delta_{\sigma_j^{f_j},\sigma'} \tag{4.17}$$

where F_i (F_j) is the total number of functions performed by protein i (j), so that each functional interaction underlying a link between proteins i and j is normalized to 1. The function $\rho(\sigma)$ is defined as

$$\rho(\sigma) = \sum_{\sigma'}\rho(\sigma,\sigma') \tag{4.18}$$

The final expression for the cost function of MEE is:

$$E = -\sum_{<i,j>}\frac{1}{F_iF_j}\sum_{f_i=1}^{F_i}\sum_{f_j=1}^{F_j}\ln\frac{\rho(\sigma_i,\sigma_j)}{\rho(\sigma_i)^{\frac{z_i-1}{z_i}}\rho(\sigma_j)^{\frac{z_j-1}{z_j}}} \tag{4.19}$$

As for the mixed GOM, a global optimization of the score function introduced (eq. (4.19)) is performed through simulated annealing techniques starting from a random functional assignment, in order to reach a minimum which corresponds to the optimal functional assignment for the uncharacterized proteins. Frustration arising during optimization process leads to several nearly equivalent minima, which are taken into account by performing 100 minimizations and recording the frequency of occurrence ν_i^σ of function σ for protein i. The functions having highest frequency represent the functional prediction provided by the method.

4.2 Results

We have performed several checks in order to test the accuracy and quality of the functional predictions provided by the two methods developed. In the following sections, we show the results of such tests to assess the statistical reliability and the robustness of functional predictions.

4.2.1 Statistical reliability

In order to test the predictive power of mixed GOM and MEE, a certain fraction f of classified proteins has been set unclassified and both the methods have been applied to attempt to recover the correct functional annotation. The success of a functional prediction is then evaluated with the introduction of two different quantities, SR and SR_f. The rate of successful predictions obtained depends, of course, on the amount of available functional information on the network. The success rate SR is an overall measure which considers as successful a prediction that recovers at least one correct function of the test protein, as it is in [158], regardless of the number of correct predictions and of the total number of predictions. SR_f, instead, is a more specific measure introduced in order to analyze the effective predictive ability of our model, taking into account both the total number of functions predicted ($\#P$) and the total number of functions truly performed by the test protein t ($\#T$). In this way, we are able to distinguish among different cases of successful predictions, although leading to an equal contribution in terms of SR.

SR_f is defined as the average of $SR_f(t)$ over all test proteins t, $SR_f = \langle SR_f(t) \rangle$, where

$$SR_f(t) = 1 - \frac{\#(P \cup T) - \#(P \cap T)}{\#(P \cup T) + \#(P \cap T)} \tag{4.20}$$

with P representing the ensemble of predicted functions for the test protein t, T the list of actual functions performed by t, $P \cup T$ the total list of functions made up of the predicted and the true ones with no repetitions, $P \cap T$ the intersection between the two ensembles P and T, i.e. the list of correctly predicted functions.

This new measure for the success rate has been chosen because it is able to weight the ratio between the number of correctly predicted functions and the total number of functions we would like to recover, penalizing proposed functional annotations characterized by a great amount of predicted functions respect to the true ones. The maximum value for $SR_f(t)$ is obtained if $P \equiv T$, so that $SR_f(t) = 1$, while if the method is not able to predict any of the true functions, $SR_f(t)$ assumes its minimum value, $SR_f(t) = 0$. Moreover, two predictions leading to the same amount of correct results ($\#(P \cap T)$), can be distinguished looking at the number of functions predicted ($\#P$): eq. (4.20) will penalize the prediction with higher $\#P$. The new measure so introduced is an improvement of both the measures adopted in [158], i.e. the success rate SR, and adopted in [160] which is defined as the ratio between the number of correctly predicted functions and the total number of functions predicted, $\#(P \cap T)/\#P$, without taking into account the amount of true functions not predicted by the method.

In the following, we do not report results obtained for the function prediction made on network (III), since it displays a trivial behaviour in all tests performed. Indeed, after having set unclassified a certain fraction f of proteins with defined functional role, none of the global optimizations investigated here is more able to recover correct functions, leading, instead, to a trivial functional configuration that assigns the same functional category to each test protein. This trivial assignment is actually reached for every network considered whenever the available functional information left on the network, after having canceled test proteins classification, is too small to be exploited in order to make predictions. This usually happens for very large values of the fraction f - we have investigated $f = 0.7$ in networks (I') and (II) and the methods are still able to predict correct functions, although the success rate obviously decreases. For what concerns network (III), instead, even for very small values of the fraction f, the optimal state found for the system almost always predicts the same function for each protein, being unable to give significant results. This particular behaviour might find its origins in the topology of network (III), whose features seem to differ from those of networks (I') and (II), under the random deletion of protein functional classification. We remind that the protein-protein interaction network (III) is composed of interactions detected with very different experimental techniques - while (I') and (II) are the result of single-type experiments - without any critical assessment of their biological relevance and thus might be more affected by false interactions.

In fig. 4.1(top) we report the results of the rate of successful predictions obtained by the model mixed GOM applied to networks (I') (left) and (II) (right), and compare them with those of other methods, i.e. the global optimization model performed on first neighbors, GOM$_1$, or on second neighbors, GOM$_2$. Results are shown as a function of protein degree. We first observe that GOM$_1$ and GOM$_2$ strikingly give almost overlapping rates of success, although their functional predictions do not overlap in the whole interval of degrees, as fig. 4.1(bottom) shows. We expect therefore to obtain some extra correct functional information coming from predictions obtained exploiting second neighbors annotations. Indeed, this extra functional information leads to a rate of success, for mixed GOM, which is higher than the ones relative to independent global optimizations. Results for (II) show a higher accuracy in the predictions limited to low degrees, as we could expect by looking at the overlap between functional annotations given by GOM_1 and GOM_2 (fig. 4.1 bottom right). Indeed, GOM_1 and GOM_2 predictions differ only for degree values smaller than 5, so that the extra functional information gained by mixed GOM is found only in that region.

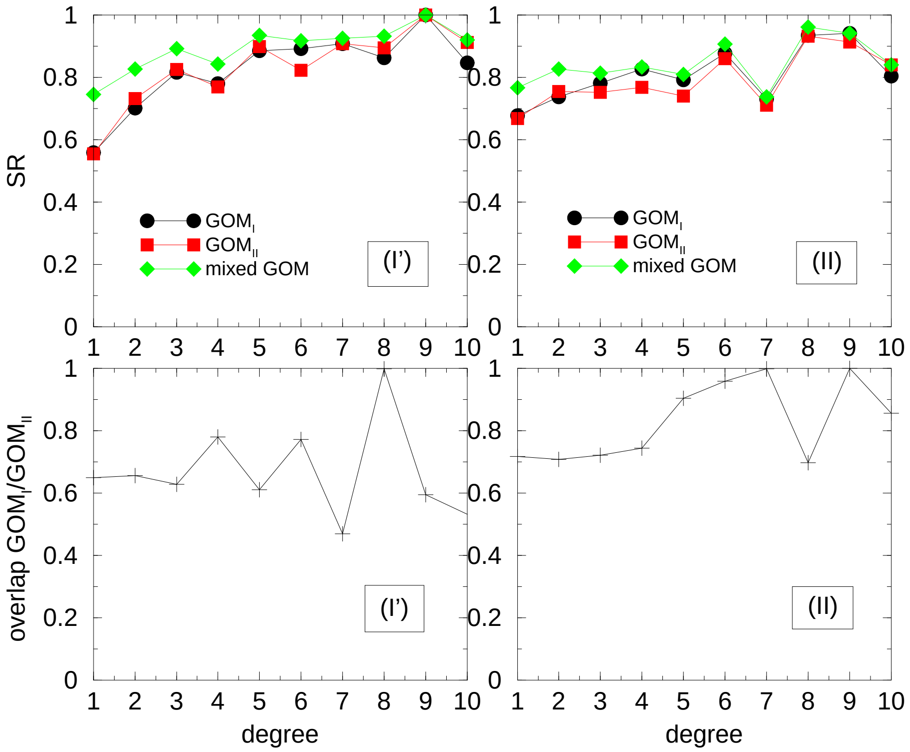


Figure 4.1: Top: Success rate SR of the functional predictions performed by the mixed GOM method for a fraction $f = 0.1$ of classified proteins set unclassified in networks (I') (left) and (II) (right), compared to the success rate obtained through the global optimization model applied to first neighbors (GOM_1) or to second neighbors (GOM_2); success rate is shown as a function of the number of interacting partners, i.e. the degree. Bottom: Overlap between the functional annotations predicted by GOM_1 and GOM_2 in networks (I') (left) and (II) (right), as a function of the protein degree.

By looking only at the rate of success SR, we could wonder if the increased success obtained with mixed GOM by merging the functional predictions of the two global optimizations performed separately is only a trivial result due to the consequent increase in the average number of predicted functions for each test protein. It could happen that we have improved only thanks to a reduction in the risk we undertake, i.e. by increasing the set of predicted functions. For this reason we have measured the new quantity introduced, SR_f, which enables us to better quantify the accuracy of functional prediction, as previously discussed.

Although the absolute value of the success rate so evaluated is decreased respect to SR, we still observe a higher accuracy in the predictions of mixed GOM respect to GOM$_1$ and GOM$_2$, as fig. 4.2(top) shows, revealing the importance of the functional information carried by second neighbors. In fig. 4.2(bottom) we compare SR_f with the probability of recovering correct functional annotation by randomly guessing. For each test protein t, we computed the probability $P^r_{\#P,c}(t)$ of correctly guessing $c = \#(P \cap T)$ functions out of $\#P$ functions extracted with uniform probability from the total number of functions $\#F$:

$$P^r_{\#P,c}(t) \;=\; \frac{C_{\#P,c}(t,P,T)}{\sum_c C_{\#P,c}(t,P,T)}, \tag{4.21}$$

where

$$C_{\#P,c}(t,P,T) \;=\; \binom{\#T}{c}\binom{\#F-\#T}{\#P-c} \tag{4.22}$$

The quality of the random predictions is almost independent from protein degree in the study of both networks, (I') and (II), and decreases to just $\sim$ 20% in (I') and $\sim$ 10% in (II), while mixed GOM give a success SR_f which is almost everywhere higher than, respectively, 60% in (I') and 55% in (II). However, the success of predictions obtained for poorly connected nodes (degree 1) in (I') by mixed GOM do not considerably differ from results obtained by random prediction; this is probably due to the presence of false interactions in the data set which may affect proteins with low degrees more than the others. A different behavior can be observed in (II), where mixed GOM predictions give higher quality results respect to random predictions even for peripheral proteins, underlying differences in data sets obtained with different experimental techniques. The rate of success obtained with the majority rule (MR) is also reported for comparison. For both data sets, two-hybrid and TAP data, it is somewhat lower than those obtained with global optimization methods, as we already expected.

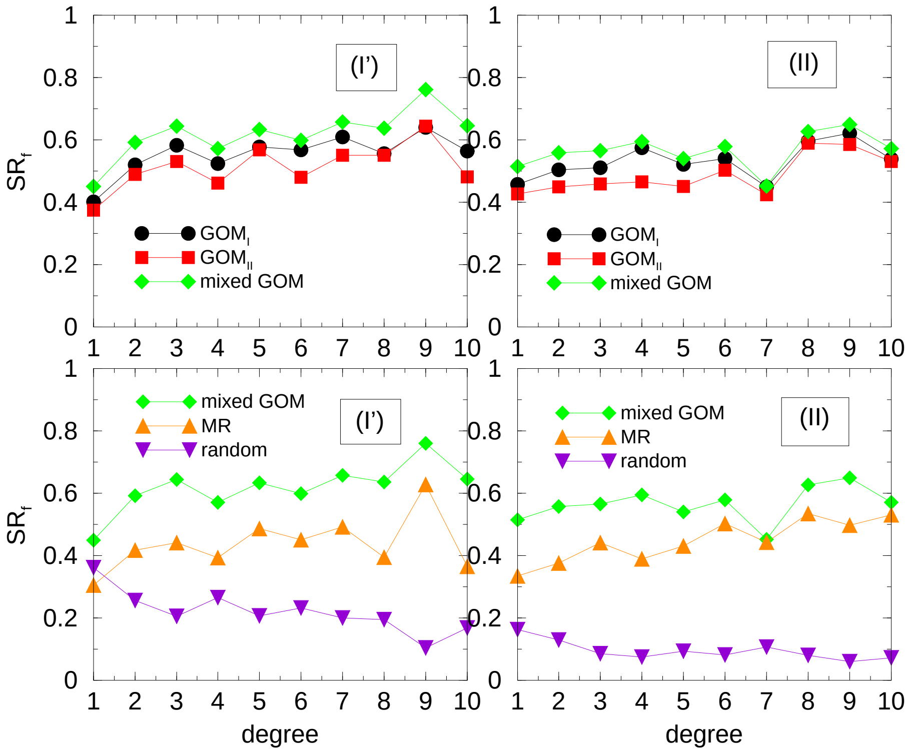


Figure 4.2: Top: Success rate SR_f of the functional predictions performed by the mixed GOM model for a fraction $f = 0.1$ of classified proteins set unclassified in networks (I') (left) and (II) (right), compared to the success rate obtained through GOM_1 and GOM_2. SR_f, as defined by eq. (4.20), represents a more accurate measure of the rate of successful predictions respect to SR. Results are shown as a function of protein degree. Bottom: Comparison of the predictive accuracy of mixed GOM, applied to networks (I') (left) and (II) (right), respect to the majority rule method (MR) and to the probability of randomly guessing the functional annotation of a protein, in terms of the success rate SR_f.

In fig. 4.3 we report results obtained with the second method introduced, MEE, in comparison with those already discussed. Although MEE is not able to improve prediction quality reached by mixed GOM, fig. 4.3 shows that making use of the correlations between functional annotations of interacting proteins leads to results very close to those given by GOM$_1$ and GOM$_2$ separately, which are based on considerations of different nature. The introduction in the cost function of information learned from the given knowledge of the protein-protein interaction network was shown to be relevant in the prediction of functional annotations of unclassified proteins.

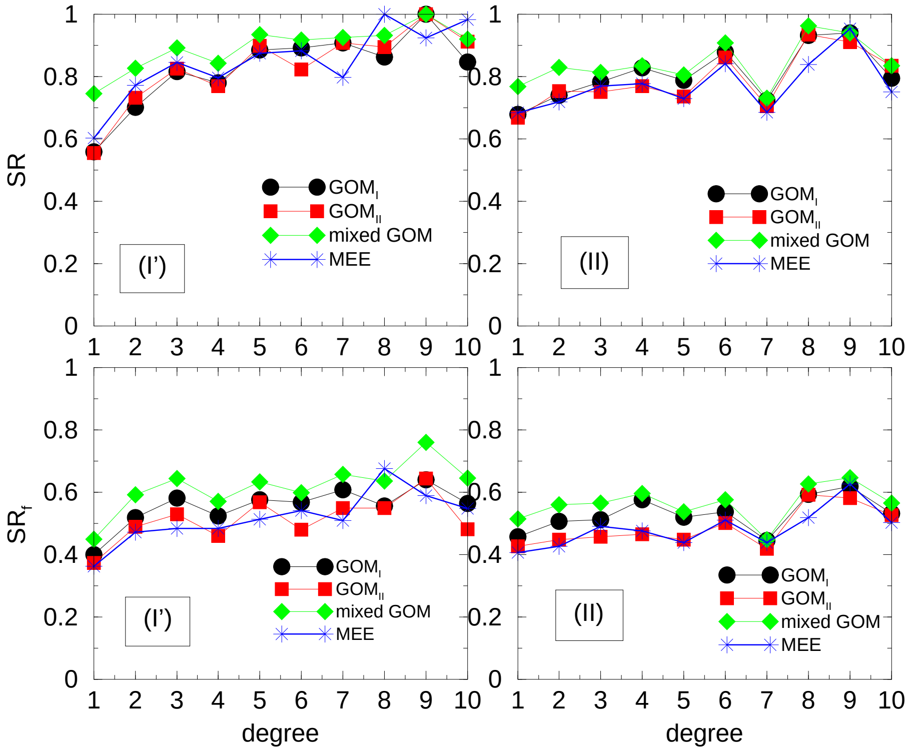


Figure 4.3: Comparison between the results obtained for networks (I') (left) and (II) (right) by the methods GOM$_1$, GOM$_2$, mixed GOM and MEE, in terms of two different measures of the rate of success, SR (top) and SR_f (bottom). Results are shown as a function of protein degree.

Future developments will consider higher order correlations to include also second neighbors, since we have shown they represent a source of supplementary information. This represents a more promising strategy than the one adopted in mixed GOM since it does not rely on some cost function with parameters assigned *a priori*. Indeed in eq. (4.1) $J_{i,j} = 1$ and in eq. (4.2) $S_{i,j}$ counts how many partners proteins i and j share wheras in eq. (4.19) there are no free parameters.

Table 4.1 reports a study of the functional predictions made by GOM$_1$ and GOM$_2$ in the same reliability test as before, analyzing their intersection and the relative rates of success due to each method only or to both of them. Results referring to both (I') and (II) are shown. Three cases are distinguished. (1) Totally overlapping predictions, i.e. identical GOM$_1$ and GOM$_2$ annotations, obtained in the 36% of the predictions made in the test for network (I') and in the 51% for network (II) (the higher percentage obtained for network (II) is in accordance with results obtained in the success rates). (2) Partly overlapping predictions, i.e. annotations sharing at least one function, which represent the 37% of the cases in network (I'), while the 26% in network (II). (3) Not overlapping predictions, meaning that the two annotations do not have any function in common, obtained in the 27% of the predictions made for network (I') and in the 23% of those made for network (II). For each case, we report the success rates obtained by each method only (*success GOM*$_1$ and *success GOM*$_2$) and those obtained when both methods predict the correct functional annotation (*success GOM*$_1$ *& GOM*$_2$). A success due to both methods could imply different predictions although both correct. This is of course the case of *not overlapping* predictions (where *success GOM*$_1$ *& GOM*$_2$ represents the 34% of the cases for (I') and the 42% for (II)), but could even occur with *partly overlapping* predictions (representing a great contribute to the total success, $\sim$ 75% for both networks), thus meaning that GOM$_2$ gives a different and non-trivial contribution for a correct functional annotation.

4.2.2 Robustness

Dealing with experimental results which are prone to false negative and false positive interactions (see section 3.1.4), we ought to assess the tolerance of our approaches with respect to the background noise present in the data sets. In particular, we want to test the robustness of functional predictions against changes in the topology of the network. To do that, we perform a second independent control experiment, which consists of a rewiring of the network by removing existing links with probability q and consequently inserting new interactions between pairs of proteins not already connected. We thus obtain a 'reshuffled' network character-

	(I')	(II)
totally overlapping	36%	51%
success	79%	80%
unsuccess	21%	20%
partly overlapping	37%	26%
success GOM_1	8%	9%
success GOM_2	4%	1%
success GOM_1 & GOM_2	74%	75%
unsuccess	14%	15%
not overlapping	27%	23%
success GOM_1	20%	20%
success GOM_2	23%	24%
success GOM_1 & GOM_2	34%	42%
unsuccess	23%	14%

Table 4.1: Intersection of functional predictions made by GOM_1 and GOM_2 for the reliability test in which a fraction $f = 0.1$ of classified proteins is set unclassified. Results referring to both (I') and (II) are shown.

ized by sensitive changes in the connectivity pattern but still having the same number of links as the original. We define a degree of dissimilarity f_l between the two networks to quantify such changes: f_l measures the fraction of links which assume different positions in the two networks. In general the degree of dissimilarity will not have a trivial dependence on the rewiring probability q, because of the possibility, in the rewiring process, that a link will connect again two proteins which actually interact in the original network, but are no more connected since their edge has already been removed.

Each method, mixed GOM and MEE, is applied to both the original and scrambled network, in order to obtain the corresponding functional predictions for the set of unclassified proteins. For every uncharacterized protein i, an overlap between the functional annotations proposed by each method is evaluated:

$$\Theta_i(f_l) = \sum_{\sigma} \left[\nu_i^{\sigma}(0) \cdot \nu_i^{\sigma}(f_l) \right]^{1/2} \tag{4.23}$$

where $\nu_i^\sigma(f_l)$ is the frequency of occurrence of function σ in the annotation of protein i obtained on the basis of the scrambled network with a degree of dissimilarity f_l with the original one, which corresponds to $f_l = 0$. The average overlap $\Theta(f_l) = \langle \Theta_i(f_l) \rangle$ is analyzed respect to f_l. In this section we monitor the overlap $\Theta(f_l)$ for the predictions given by GOM$_1$, GOM$_2$ and mixed GOM, using the original and the scrambled network, as a function of the degree of dissimilarity f_l.

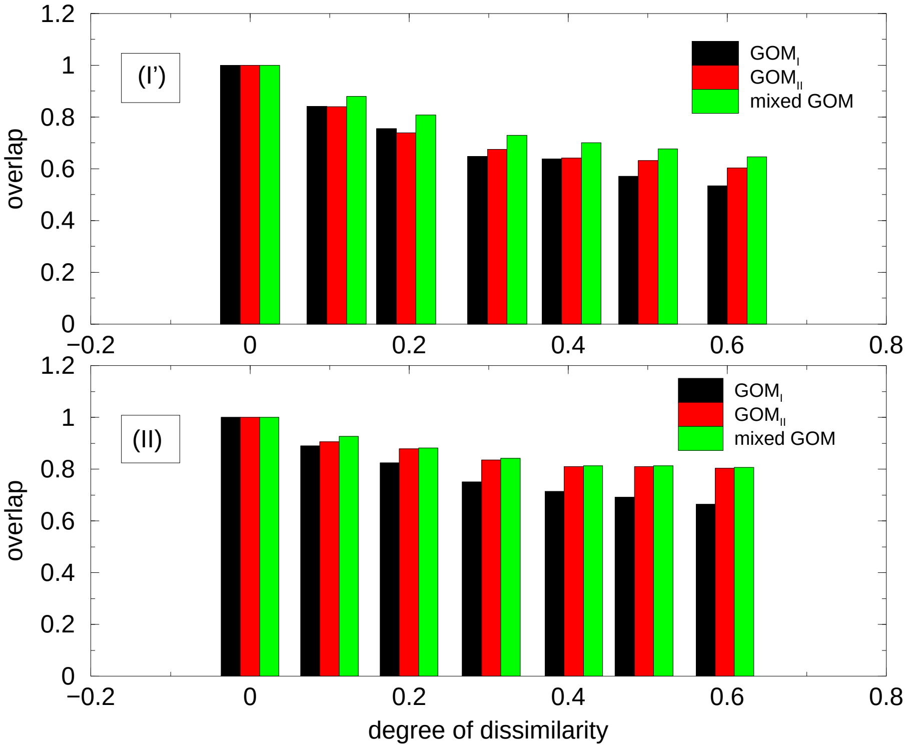


Figure 4.4: Overlap averaged over all unclassified proteins between functional predictions made by GOM$_1$, GOM$_2$ and mixed GOM, based on the original network and on the reshuffled one, having a degree of dissimilarity f_l. Top: network (I'). Bottom: network (II).

As shown in fig. 4.4, we find that the three methods analyzed (GOM$_1$, GOM$_2$ and mixed GOM) display a considerably high robustness against the presence of misplaced interactions in the dataset (for GOM$_1$ see also [158]). Indeed, for a dissimilarity of about 10% the three methods predict functional annotations overlapping with the original ones for more than 85% in both networks, (I') and (II). It is important to notice, however, that the rewiring of a link generally involves 3 to 4 proteins changing their connectivity configuration, thus implying that a 10% of dissimilarity actually means a connectivity pattern changed for about $30 - 40\%$ of the proteins present in the network.

The models GOM_1 and GOM_2 display different behaviours respect to network 'reshuffling'. While for low degree of dissimilarity optimizations involving first neighbors or second neighbors provide similar results, when f_l becomes increasing ($f_l > 40\%$) GOM_1 reveals to be more sensitive to the introduction of misplaced interactions in the network, clearly showing that the extraction of functional information from second neighbors is more robust against topology perturbations. Indeed, since in GOM_2 all paths of length 2 connecting pairs of proteins are considered, a change in one link only reduces the strength of second neighbors interaction quantified by cost function $E^{(2)}$ (see eq. (4.2)), without eliminating it, while it completely cancels first neighbors interaction.

Comparing results referring to different networks, (I') (top) and (II) (bottom), we observe that in both cases mixed GOM is able to obtain higher values of the overlap function, showing a behaviour which is more robust than that of GOM_1, since it is gaining higher degree of robustness coming from GOM_2. However, different values of the overlap function reached by mixed GOM for rather large degree of dissimilarity ($\Theta(f_l = 60\%) \simeq 65\%$ in (I'), while $\Theta(f_l = 60\%) \simeq 80\%$ in (II)) point out different degrees of robustness of such networks, probably due to distinct intrinsic features of data sets resulting from different experiments.

4.3 Discussion

Comparison of the proposed models with different approaches for functional prediction, other than the global optimization, reveal some important relative advantages. First of all, both our methods are able to offer a functional prediction for the *entire* set of unclassified proteins in the network, while others take into account only a fraction of it. Indeed, the statistical analysis of topological redundancies made by Samanta and Liang [159] focus the attention only on those protein pairs sharing a significantly high number of common partners, with no possibility of covering the entire network. An unannotated protein can eventually be classified only if it forms associations with other proteins, i.e. only if it belongs to the list of protein pairs showing topological redundancies.

Other approaches limit their predictions to proteins for which assignment is mostly unambiguous, thus strongly restricting the range of applicability of the method itself. In [160], for example, the authors choose to rule out proteins with degree less than 3, since they are expected to be more affected by the presence of false positive and negative interactions. Hence, such proteins are considered for computation but no functional prediction is possible. Starting from a data set composed of 2139 proteins, the classification tree constructed in [160] for the functional prediction contains only 602 proteins, thus dramatically reducing the ensemble of proteins for which a classification can be proposed. Our methods, instead, can not only be used for functional assignment on the whole set of unclassified proteins, but also show an increased rate of successful predictions (fig. 4.1), especially for low values of the degree.

The global optimization process is able to propose multi-functional annotations for each protein, because of the presence of multiple minima of the cost functions due to frustration (see Methods). Although we have already shown that our improvements reached with mixed GOM are not due to a trivial increase in the probability of guessing the correct functions, through the introduction of a new measure for the rate of success - SR_f - in the statistical reliability tests, here we would like to directly monitor how the number of functions predicted per protein varies in the three methods, GOM_1, GOM_2 and mixed GOM.

In fig. 4.5 we report the distribution of the number of functions per protein predicted by the two global optimizations, GOM_1 and GOM_2, and by mixed GOM, when applied to network (I')(top) and (II)(bottom). Each column represents the fraction of test proteins for which a certain number of functions has been predicted by using GOM_1, GOM_2 and mixed GOM. Merging functional annotations leads to an increase in the fraction of proteins having multi-functional classification respect to GOM_2. Indeed, referring to network (I'), the fraction of proteins with a predicted classification made up of only one function (first set of columns in the histogram) is greater in GOM_2 ($\sim 70\%$) than in mixed GOM ($\sim 53\%$), which instead tends to assign more than one function per protein respect to GOM_2 - e.g.

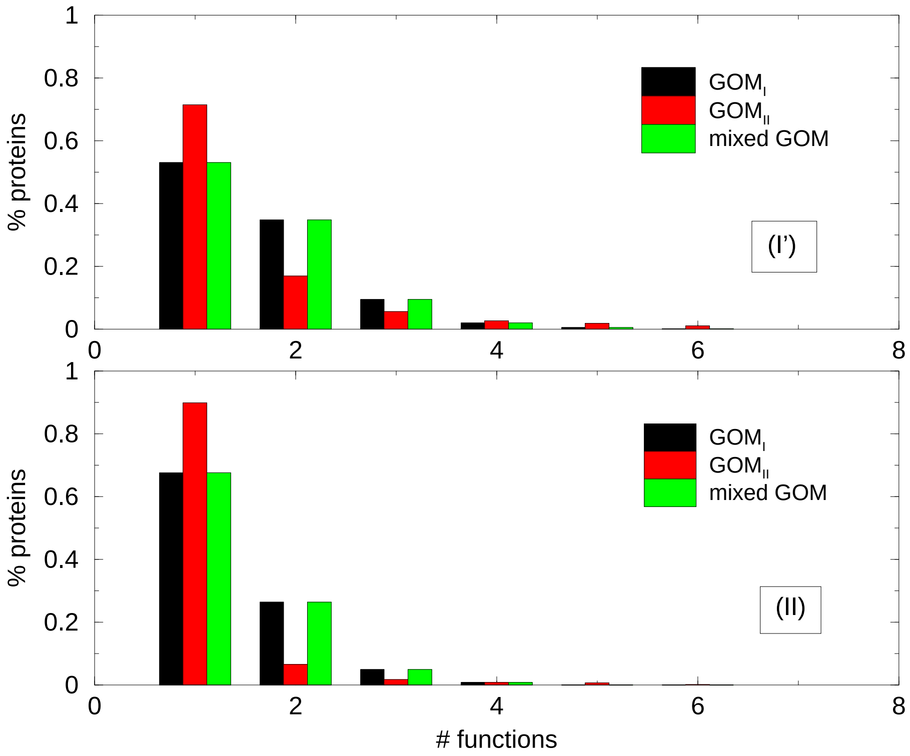


Figure 4.5: Distribution of the number of functions per protein predicted by the three methods, GOM$_1$ and GOM$_2$, and mixed GOM. Each column in the histogram represents the percentage of proteins having associated a certain number of functions by one of the methods - GOM$_1$, GOM$_2$ and mixed GOM. Top: network (I'). Bottom: network (II).

mixed GOM assigns 2 functions to $\sim$ 35% of the proteins, while GOM$_2$ only to $\sim$ 17% (second set of columns in fig. 4.5). However, the distribution of the number of functions per protein predicted with mixed GOM do not substantially differ from that obtained with GOM$_1$, showing again that merging functional classifications improves the succes rate without a substantial increase of the number of the predicted functions. Similar results are obtained considering network (II) (fig. 4.5, bottom).

Instead of merging results in mixed GOM coming from two independent minimizations, GOM_1 and GOM_2, we have investigated the possibility of using a single cost function corresponding to a linear combination of the two terms relative, respectively, to first and second neighbors contributions. In other words, we optimized the following cost function:

$$E = E^{(1)} + \lambda E^{(2)}, \tag{4.24}$$

λ being the global weight of the linear combination, $E^{(1)}$ and $E^{(2)}$ the cost functions of eq. (4.1) and eq. (4.2), respectively. It is not easy to guess *a priori* a value of λ which could opportunely take into account new bioinformatics knowledge coming from second neighbors and enable us to improve the quality of our functional predictions. Making use of some topological and energetic informations acquired on the network, we have studied different values of the weight, among which:

- the energy ratio of the independent optimizations, GOM_1 and GOM_2, i.e.

  $$\lambda = \frac{E^{(1)}_{min}}{E^{(2)}_{min}}, \tag{4.25}$$

  where E_{min} represents the energy of the optimal configuration;

- the ratio of functional rates of protein couples (of first and second neighbors) with at least a function in common:

  $$\lambda = \frac{\mathrm{rate}^{(2)}_f}{\mathrm{rate}^{(1)}_f}, \tag{4.26}$$

  where $\mathrm{rate}^{(1)}_f$ ($\mathrm{rate}^{(2)}_f$) represents the observed percentage of first (second) neighbors sharing at least one function, computed on the classified proteins of the network; we have: $\lambda = 75\%/86\%$ in (I'), $\lambda = 77\%/83\%$ in (II) and $\lambda = 63\%/72\%$ in (III).

Results obtained do not differ from those relative to GOM_1 and GOM_2, so that a linear combination with a global weight does not seem to be able to distinguish between GOM_1 and GOM_2 contributions, at least for the values investigated. One of the two contributions seems to be negligible respect to the other; therefore, the cost function introduced does not capture the extra functional information obtained by mixed GOM.

Finally, we have investigated a linear combination of $E^{(1)}$ and $E^{(2)}$, with a local weight in the sums of eq. (4.2) dependent on the number of first and second neighbors, in order to locally normalize the number of interactions considered in each energy function:

$$\lambda = \frac{n_i^{(1)}}{n_i^{(2)}} = \frac{k_i}{n_i^{(2)}}, \tag{4.27}$$

where k_i represents the degree of protein i and $n_i^{(2)}$ the number of paths of length 2 starting from protein i. Also in this case we are not able to improve the rate of success already obtained by GOM$_1$ and GOM$_2$.

A possible solution would be to optimize the parameter λ in such a way to maximize the rate of success, but it would require extremely high computational costs.

4.4 Conclusions

We have proposed two general bioinformatics methods for predicting functional classification of uncharacterized proteins, based on protein-protein interaction networks. We have applied them to the interaction maps of *Saccharomyces cerevisiae* extracted from two-hybrid data [156], tandem-affinity purification data [28] and a mixed data set collected at the Database of Interacting Proteins [144].

The first method, called mixed GOM, relies on the assumption that a functional association could potentially exist not only between directly interacting proteins but even between proteins sharing common partner(s). It is based on two independent global optimization processes - GOM$_1$ and GOM$_2$ - respectively involving first and second neighbors and leading to separate functional predictions for the unclassified set of proteins. Merging predictions for each uncharacterized protein provides the annotation proposed by mixed GOM.

The second method is based on a general theoretical approach which extracts functional relevance underlying physical interactions between proteins from the available information about the system. Through a statistical inference criterion - the maximum entropy estimate - it determines an expression for the cost function in terms of the k-points functional correlations, computed on the classified part of the network of interactions. In this study $k \leq 2$. A global optimization is performed in order to reach the minimum of the cost function, which leads to the optimal functional assignment for proteins with no functional annotation.

Results obtained demonstrate the functional relevance of connections between proteins and extend this result even to interactions mediated by an intermediate protein, i.e. between second neighbors. Reliability tests have shown the robustness of presented methods in dealing with data sets which are incomplete and/or affected by misleading false interactions.

Conclusions and perspectives

In this thesis, we have presented models and theorethical tools to describe the non-trivial features of complex networks, and focused the attention on a specific real system - protein-protein interaction networks - to unravel underlying organizational principles and correlations with biological functions, in an attempt to understand cell's functional organization.

The class of models introduced relies on selection principles on the basis of optimality criteria. Our work is complementary to existing models that either rely on dynamical mechanisms, such as preferential attachment, or on topological and geometrical criteria. Tree-like structures, as well as networks with loops, have been investigated through numerical simulations and exhaustive exploration of hierarchical tree patterns, by means of an analytical expression of the cost function for loopless networks. In spite of its simplicity, the class of models proposed seems to capture several features of networks in Nature. Though by no means exhaustive, our results show that selective criteria blend chance and necessity as dynamic origins of recurrent network patterns.

Scale-free properties of networks have been investigated with theorethical tools usually employed to study critical phenomena, i.e. with a renormalization group (RG) treatment. Coarse-graining of less relevant details of the networks was found to lead to renormalized weighted networks which preserve original critical behaviours. RG has been applied to several network models, designed to reproduce different types of systems, and to two distinct protein-protein interaction networks of the yeast *Saccharomyces cerevisiae*, obtained from different experimental methods. Although we are firmly convinced that this work is at its first stage, results suggest that RG approach could represent a fruitful tool in the study of critical network properties, by 'simplifying' large-scale systems and thus uncovering 'true' critical behaviours from apparent ones. A possible application for the visualization of more understandable - though meaningful - networks will be the object of future developments, as well as other renormalization techniques.

In the last part of this thesis, we have focused the attention on a real example in the area of biological systems: protein-protein interaction networks. The study of the architecture and the dynamics of the complex interaction between the numerous constituents of the cell represents an extremely important issue in post-genomic biology, since it could lead to a comprehensive explanation of the behaviour of a cell and of the biological phenomena occurring.

Several non-trivial features were found in the investigation of three distinct protein interaction networks of the yeast *Saccharomyces cerevisiae*, obtained with different experimental techniques, regarding scale-free distributions of protein connectivity, clustering hierarchy and correlations. Results for the functional relevance of interactions emerged from the analysis correlating network topology with the biological function of proteins. Such results constitute the basic strategy of global optimization methods, developed to facilitate the functional annotation of proteins not yet classified. With the rapid recent developments in complete genome sequencing and the vast amount of data on protein-protein interactions becoming available, the functional classification of still uncharacterized proteins becomes of fundamental importance. The two bioinformatics approaches presented here are designed to take full advantage of the functional relationships underlying the complex pattern of interactions to propose a functional annotation for unclassified proteins, by introducing an opportune cost function which takes into account such information encoded in the network. In particular, we have considered a topological extension of a previous global optimization model, with the introduction of a new parameter in the cost function taking into account the role of second neighbors, and a second theorethical approach whose cost function depends on the k-point functional correlations evaluated on the network. While the first method, mixed GOM, relies on two parameters which are assigned specific values *a priori*, the second one, MEE, has no free parameters, since it extracts useful information from the given knowledge of the system, through maximum entropy estimates. Results point out that the introduction of these new ingredients are found to be crucial in the improvement of predictive ability, respect to previous works in this area. Future developments will consider higher order correlations.

Nevertheless we are aware that the collection of the increasing amount of experimental data must be followed by an estimation of the reliability of different data sets and of the coverage of distinct approaches, in order to obtain statistically significant protein-protein interaction networks and be able to develop new promising bioinformatics methods.

Appendix A

Simulated annealing

The method we have used in the global optimization processes, performed in mixed GOM and in MEE model, is *simulated annealing* [161, 162], in which a parameter T, analogous to the temperature, is introduced and lowered during the minimization process. More in detail, the minimization algorithm is as follows:

(i) Generation of a random initial functional configuration. To each unclassified protein i an initial function σ_i, randomly chosen among the total number of functions ($F = 18$ in our case), is assigned.

(ii) Random change of the functional configuration. An uncharacterized protein is randomly selected and a new associated function is extracted with uniform probability.

(iii) Energetic control. The change ΔE in the cost function is evaluated. If it is negative, the change is accepted and we go on to step (iv). Otherwise the quantity $\exp[-\Delta E/T]$ is compared to a random number p, uniformly extracted in the interval $[0, 1]$:

- if $\exp[-\Delta E/T] \leq p$, the change is accepted and we go on to step (iv);

- otherwise the change is rejected and we return to step (ii).

(iv) Updating of the new configuration. The functional configuration is updated by taking into account the change in the function of the unclassified protein extracted.

(v) Lowering of the parameter T. In each cycle, the parameter T is decreased by a factor R very close to 1, so that at the nth cycle we have $T(n) = R^n T(0)$, where $T(0)$ assumes a suitable chosen value.

After the initial condition is generated (step (i)), steps (ii)-(iv) are repeated several times until the system is thermalized at temperature T. Then, step (v) is performed and temperature T is decreased; the entire algorithm, from step (ii) to step (v), is repeated until the parameter T reaches a considerably low value.

Bibliography

[1] P. Erdös and A. Rényi. On random graphs. *Publicationes Mathematicae*, **6**:290, (1959).

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

[3] P. Erdös and A. Rényi. On the strength of connectedness of random graphs. *Acta. Math. Sci. Hung.*, **12**:261, (1961).

[4] B. Bollobás. *Random Graphs.* Cambridge Univ. Press, Cambridge, 2001.

[5] J. Scott. *Social network analysis: a handbook.* Sage, London, 2000.

[6] S. Wasserman and K. Faust. *Social network analysis.* Cambridge University Press, Cambridge, UK, 1994.

[7] D.J. de S. Price. A general theory of bibliometric and other cumulative advantage processes. *J. Amer. Soc. Inform. Sci.*, **27**:292, (1976).

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

[9] S. Redner. How popular is your paper? an empirical study of the citation distribution. *Eur. Phys. J. B*, **4**:131, (1998).

[10] B.A. Huberman and L.A. Adamic. Internet: Growth dynamics of the world-wide web. *Nature*, **401**:131, (1999).

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

[12] A.-L. Barabási, R. Albert, and H. Jeong. Scale-free characteristics of random networks: The topology of the world wide web. *Phys. A*, **281**:69, (2000).

[13] J.M Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A.S. Tomkins. The web as a graph: measurements, models, and methods. *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing*, :163, (2000).

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

[15] M Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationship of the internet topology. *Comput. Commun. Rev.*, **29**:251, (1999).

[16] G. Caldarelli, R. Marchetti, and L. Pietronero. The fractal properties of internet. *Europhys. Lett.*, **52**:386, (2000).

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

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

[19] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. USA*, **101**:3747, (2004).

[20] I. Rodriguez-Iturbe and A. Rinaldo. *Fractal River Basins. Chance and Self-Organization.* Cambridge Univ. Press, New York, 1997.

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

[22] A. Rinaldo, I. Rodriguez-Iturbe, and R. Rigon. Channel networks. *Ann. Rev. Earth Planet Sci.*, **26**:289, (1998).

[23] G. Caldarelli, P. De Los Rios, M. Montuori, and V.D.P. Servedio. The drainage basins trees in mars channel networks. *Eur. Phys. J. B*, **38**:387, (2004).

[24] G.B. West, J.H. Brown, and B.J. Enquist. A general model for the origin of allometric scaling laws in biology. *Science*, **276**:122, (1997).

[25] J.R. Banavar, A. Maritan, and A. Rinaldo. Size and form in efficient transportation networks. *Nature*, **399**:130, (1999).

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

[27] T. Ito *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA*, **98**:4569, (2001).

[28] A.C. Gavin *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**:141, (2002).

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

[30] S. Maslov and K. Sneppen. Specificity and stability in topology of protein newtorks. *Science*, **296**:210, (2002).

[31] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Modeling of protein interaction networks. *ComPlexUs*, **1**:38, (2003).

[32] R.V. Solé, R. Pastor-Satorras, E. Smith, and T.B. Kepler. A model of large-scale proteome evolution. *Advances in Complex Systems*, **5**:43, (2002).

[33] D.A. Fell and A. Wagner. The small world of metabolism. *Nature Biotech.*, **18**:1121, (2000).

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

[35] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E.D. Gilles. Metabolic network structures determines key aspects of functionality and regulation. *Nature*, **420**:190, (2002).

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

[37] R.V. Solé and J.M. Montoya. Complexity and fragility in ecological networks. *Proc. Roy. Soc. London Ser. B*, **268**:2039, (2001).

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

[39] D. Garlaschelli, G. Caldarelli, and L. Pietronero. Universal scaling relations in food webs. *Nature*, **423**:165, (2003).

[40] G. White, E. Southgate, J.N. Thompson, and S. Brenner. *The structure of the nervous system of the nematode C. Elegans.* Philos. Trans. Roy. Soc. London, 314, p.1, 1986.

[41] O. Sporns. Network analysis, complexity and brain function. *Complexity*, **8**:56, (2002).

[42] O. Sporns, G. Tononi, and G.M. Edelman. Theorethical neuroanatomy: relating anatomical and functional connectivity in graphs and cortical connection matrices. *Cerebral Cortex*, **10**:127, (2000).

[43] M.E.J. Newman. From the cover: The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA*, **98**:404, (2001).

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

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

[46] G. Bonanno, G. Caldarelli, F. Lillo, and R. N. Mantegna. Topology of correlation based minimal spanning trees in real and model markets. *Phys. Rev. E*, **68**:046130, (2003).

[47] G. Bonanno, G. Caldarelli, F. Lillo, S. Miccichè, N. Vandewalle, and R. N. Mantegna. Networks of equities in financial markets. *Eur. Phys. J. B*, **38**:363, (2004).

[48] D. Garlaschelli and S. Battiston and M. Castri and V.D.P. Servedio and G. Caldarelli. The scale-free topology of market investments. cond-mat/0310503.

[49] R. Albert and A.L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, **74**:47, (2002).

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

[51] V. Colizza, J.R. Banavar, A. Maritan, and A. Rinaldo. Network structures from selection principles. *Phys. Rev. Lett.*, **92**:198701, (2004).

[52] I. Rodriguez-Iturbe, A. Rinaldo, R. Rigon, E. Ijjasz-Vasques, and R.L. Bras. Energy-dissipation, runoff production, and the 3-dimensional structure of river basins. *Water Resour. Res.*, **28**:1095, (1992).

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

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

[55] P.L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Phys. Rev. Lett.*, **85**:4629, (2000).

[56] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Structure of growing networks with preferential linking. *Phys. Rev. Lett.*, **85**:4633, (2000).

[57] P.L. Krapivsky and S. Redner. Organization of growing random networks. *Phys. Rev. E*, **63**:art. no. 066123, (2001).

[58] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Principles of statistical mechanics of uncorrelated random networks. *Nucl. Phys. B*, **666**:396, (2003).

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

[60] P.L. Krapivsky and S. Redner. A statistical physics perspective on web growth. *Comput. Netw.*, **39**:261, (2002).

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

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

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

[64] G. Caldarelli, A. Capocci, P. De Los Rios, and M.A. Mu noz. Scale-free networks from varying vertex intrinsic fitness. *Phys. Rev. Lett.*, **89**:258702, (2002).

[65] L.P. Kadanoff. Variational principles and approximate renormalization group calculations. *Phys. Rev. Lett.*, **34**:1005, (1975).

[66] A.A. Migdal. Recursion equations in gauge theories. *Sov. Phys. JETP*, **42**:413, (1975).

[67] A.A. Migdal. Gauge transitions in gauge and spin lattice systems. *Sov. Phys. JETP*, **42**:743, (1975).

[68] C. Jayaprakash, E.K. Riedel, and M. Wortis. Critical and thermodynamic properties of the randomly dilute ising model. *Phys. Rev. B*, **18**:2244, (1978).

[69] J.M. Yeomans and R.B. Stinchcombe. Critical properties of site- and bond-diluted ising ferromagnets. *J. Phys. C*, **12**:347, (1979).

[70] M. Schwartz and S. Fishman. Real space renormalization goup - study of the random bond ising-model. *Physica A*, **104**:115, (1980).

[71] M.E. Fisher. Renormalization group theory: Its basis and formulation in statistical physics. *Rev. Mod. Phys.*, **70**:653, (1998).

[72] S.N. Dorogovtsev. Renormalization group for evolving networks. *Phys. Rev. E*, **67**:art. no. 045102, (2003).

[73] M.E.J. Newman and D.J. Watts. Renormalization group analysis of the small-world network model. *Phys. Lett. A*, **263**:341, (1999).

[74] H.W. Mewes, K. Albermann, K. Heumann, S. Liebl, and F. Pfeiffer. Mips: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res.*, **25**:28, (1997).

[75] T.C. Hodgman. A historical perspective on gene/protein functional assignment. *Bioinformatics*, **16**:10, (2000).

[76] D. Eisenberg, E.M. Marcotte, I. Xenarios, and T.O. Yeates. Protein function in post-genomic era. *Nature*, **405**:823, (2000).

[77] V. Colizza, P. De Los Rios, A. Flammini and A. Maritan. Protein function prediction based on protein-protein interaction network. *In preparation.*

[78] J. Abello, P.M. Pardalos and M.G.C. Resende. *External memory algorithms*, edited by J. Abello and J. Vitter, DIMACS series in Discrete Mathematics Theorethical Computer Science (American Mathematical Society), p.119 (1999).

[79] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, :171, (2000).

[80] S. Valverde and R.V. Solé R.F. Cancho. Scale-free networks from optimal design. *Europhys. Lett.*, **60**:512, (2002).

[81] R.F. Cancho and R.V. Solé. Optimization in complex networks. cond-mat/0111222.

[82] A. Fabrikant, E. Koutsoupias, and C.H. Papadimitriou. Heuristically optimized trade-offs: A new paradigm for power laws in the internet. *Lect. Notes Comput. Sci.*, **2380**:110, (2002).

[83] N. Mathias and V. Gopal. Small worlds: How and why. *Phys. Rev. E*, **63**:21117, (2001).

[84] A. Vazquez, R. Pastor-Satorras, and A. Vespignani. Large-scale topological and dynamical properties of internet. *Phys. Rev. E*, **65**:066130, (2002).

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

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

[87] E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *Phys. Rev. E*, **67**:026112, (2003).

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

[89] A. Vazquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Phys. Rev. E*, **67**:056104, (2003).

[90] A. Capocci, G. Caldarelli, and P. De Los Rios. Quantitative description and modeling of real networks. *Phys. Rev. E*, **68**:047101, (2003).

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

[92] M.E.J. Newman. Mixing patterns in networks. *Phys. Rev. E*, **67**:art. no. 026126, (2003).

[93] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Struct. Algor.*, **6**:161, (1995).

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

[95] D.J. Watts. *Small worlds.* Princeton University Press, Princeton NJ, 1999.

[96] D.J. Watts. Networks, dynamics, and the small world phenomenon. *Amer. J. Sociol.*, **105**:493, (1999).

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

[98] H.A. Simon. On a class of skew distribution functions. *Biometrika*, **42**:425, (1955).

[99] B. Bollobás and O. Riordan. The diameter of a scale-free random graph. *Combinatorica*, **24**:5, (2004).

[100] R. Cohen and S. Havlin. Scale-free networks are ultrasmall. *Phys. Rev. Lett.*, **90**:art. no. 058701, (2003).

[101] D.R. Montgomery and W.E. Dietrich. Where do channels begin. *Nature*, **336**:232, (1998).

[102] I. Rodriguez-Iturbe, A. Rinaldo, R. Rigon, E. Ijjasz-Vasques, and R.L. Bras. Fractal structures as least energy patterns - the case of river networks. *Geophys. Res. Lett.*, **19**:889, (1992).

[103] K. Sinclair and R.C. Ball. Mechanism for global optimization of river networks from local erosion rules. *Phys. Rev. Lett.*, **76**:3360, (1996).

[104] A. Maritan, F. Colaiori, A. Flammini, M. Cieplak, and J.R. Banavar. Universality classes of optimal channel networks. *Science*, **272**:984, (1996).

[105] J.R. Banavar, F. Colaiori, A. Flammini, A. Giacometti, A. Maritan, and A. Rinaldo. Sculpting of a fractal river basin. *Phys. Rev. Lett.*, **78**:4522, (1997).

[106] J.R. Banavar, F. Colaiori, A. Flammini, and A. Rinaldo. Topology of the fittest transportation network. *Phys. Rev. Lett.*, **84**:4745, (2000).

[107] S.H. Strogatz. Exploring complex networks. *Nature*, **410**:268, (2001).

[108] J.R. Banavar, F. Colaiori, A. Flammini, A. Maritan, and A. Rinaldo. Scaling, optimality, and landscape evolution. *J. Stat. Phys.*, **104**:1, (2001).

[109] The Pajek Software. *http://vlado.fmf.uni-lj.si/pub/networks/pajek/.*

[110] G. Caldarelli, R. Pastor-Satorras, and A. Vespignani. Structure of cycles and local ordering in complex networks. *Europ. Phys. J. B*, **36**:203, (2003).

[111] G. Bianconi and A. Capocci. Number of loops of size h in growing scale-free networks. *Phys. Rev. Lett.*, **90**:art. no. 078701, (2003).

[112] H. Inose. Communication networks. *Sci. Am.*, **3**:117, (1972).

[113] R.V. Solé, R.F. Cancho, J.M. Montoya, and S. Valverde. Selection, tinkering, and emergence in complex networks. *Complexity*, **8**:20, (2003).

[114] S.H. Yook, H. Jeong, A.-L. Barabási, and Y. Tu. Weighted evolving networks. *Phys. Rev. Lett.*, **86**:5835, (2001).

[115] J.D. Noh and H. Rieger. Stability of shortest paths in complex networks with random edge weights. *Phys. Rev. E*, **66**:066127, (2002).

[116] A. Barrat, M. Barthelemy, and A. Vespignani. Weighted evolving networks: Coupling topology and weight dynamics. *Phys. Rev. Lett.*, **92**:228701, (2004).

[117] G. Caldarelli, F. Coccetti and P. De Los Rios. Preferential Exchange: Strengthening Connections in Complex Networks. cond-mat/0312236.

[118] M.E.J Newman. Analysis of weighted networks. cond-mat/0407503.

[119] E. Almaas, P.L. Krapivsky and S. Redner. Statistics of weighted networks. cond-mat/0408295.

[120] J. Davidsen, H. Ebel, and S. Bornholdt. Emergence of a small world from local interactions: modeling acquaintance networks. *Phys. Rev. Lett.*, **88**:128701, (2002).

[121] Y. Ho *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**:180, (2002).

[122] T. Dandekar, B. Snel, M. Huynen, and P. Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **23**:324, (1998).

[123] H. Ge, Z. Liu, G.M. Church, and M. Vidal. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae. Nature Genet.*, **29**:482, (2001).

[124] R.J. Cho *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.*, **2**:65, (1998).

[125] T.R. Hughes *et al.* Functional discovery via a compendium of expression profiles. *Cell*, **102**:109, (2000).

[126] A.H. Tong *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, **294**:2364, (2001).

[127] H.W. Mewes *et al.* Mips: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**:31, (2002).

[128] A.J. Enright, I. Iliopoulos, N.C. Kyrpides, and C.A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**:86, (1999).

[129] E.M. Marcotte, M. Pellegrini, H.L. Ng, D.W. Rice, D.O. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**:751, (1999).

[130] R. Overbeek, M. Fonstein, M. D'Souza, G.D. Pusch, and N. Maltsev. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA*, **96**:2896, (1999).

[131] M. Pellegrini, E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA*, **96**:4285, (1999).

[132] M.A. Huynen and P. Bork. Measuring genome evolution. *Proc. Natl. Acad. Sci. USA*, **95**:5849, (1998).

[133] J. Ma and M. Ptashne. A new class of yeast transcriptional activators. *Cell*, **51**:113, (1987).

[134] S. Fields and O.-K. Song. A novel genetic system to detect protein-protein interactions. *Nature*, **340**:245, (1989).

[135] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Seraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotech.*, **17**:1030, (1999).

[136] A. Valencia and F. Pazos. Computational methods for the prediction of protein inetarctions. *Curr. Opin. Struct. Biol.*, **12**:368, (2002).

[137] J. Tamames, G. Casari, C. Ouzounis, and A. Valencia. Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.*, **44**:66, (1997).

[138] P. Novick, B.C. Osmond, and D. Botstein. Suppressors of yeast actin mutations. *Genetics*, **121**:659, (1989).

[139] L. Guarente. Synthetic enhancement in gene interaction - a genetic tool come of age. *Trends Genet.*, **9**:362, (1993).

[140] I. Xenarios and D. Eisenberg. Protein interaction databases. *Curr. Opin. Biotech.*, **12**:334, (2001).

[141] G.D. Bader and C.W.V. Hogue. Analyzing yeast protein-protein interaction data obtained from different sources. *Nature Biotech.*, **20**:991, (2002).

[142] C.M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg. Protein interactions - two methods for assessment of the reliability of high throughout observations. *Mol. Cell. Proteomics*, **1**:349, (2002).

[143] C. von Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**:399, (2002).

[144] Database of Interacting Proteins. *http://dip.doe-mbi.ucla.edu/.*

[145] M.E.J. Newman. Random graphs as models of networks. *in*, S. Bornholdt and H.G. Schuster, eds, 'Handbook of Graphs and Networks: From the Genome to the Internet', Wiley-VCH, Berlin:35–68, (2003).

[146] S. Zhou and R.J. Mondragon. The rich-club phenomenon in the internet topology. *IEEE Commun. Lett.*, **8**:180, (2004).

[147] R. Pastor-Satorras and A. Vespignani. *Evolution and Structure of the Internet : A Statistical Physics Approach.* Cambridge University Press, 2004.

[148] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A.S. Tomkins, and E. Upfal. Stochastic models for the web graph. *Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science*, :57, (2000).

[149] H. Jeong, Z. Néda, and A.-L. Barabási. Measuring preferential attachment in evolving networks. *Europhys. Lett.*, **61**:567, (2003).

[150] M.E.J. Newman. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, **64**:art. no. 025102, (2001).

[151] A. Force, M. Lynch, F.B. Pickett, A. Amores, Y. l. Yan, and J. Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, **151**:1531, (1999).

[152] M. Lynch and A. Force. The probability of duplicate gene preservation by subfunctionalization. *Genetics*, **154**:459, (2000).

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

[154] S.N. Dorogovtsev, A.N. Samukhin, and J.F.F. Mendes. Multifractal properties of growing networks. *Europhys. Lett.*, **57**:334, (2002).

[155] The MIPS Comprehensive Yeast Genome Database (CYGD). *http://mips.gsf.de/proj/yeast/CYGD/db/.*

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

[157] H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, and T. Tagaki. Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast*, **18**:523, (2001).

[158] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Global protein function prediction from protein-protein interaction network. *Nature Biotech.*, **21**:697, (2003).

[159] M.P. Samanta and S. Liang. Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc. Natl. Acad. Sci. USA*, **100**:12579, (2003).

[160] C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guenoche, and B. Jacq. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biol.*, **5**:art. no. R6, (2003).

[161] S. Kirkpatrick, C. D. Gelatt, and M.P. Vecchi. Optimization by simulated annealing. *Science*, **220**:671, (1983).

[162] S. Kirkpatrick. Optimization by simulated annealing - quantitative studies. *J. Stat. Phys.*, **34**:975, (1984).

[163] A. Edwards, B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt, and M. Gerstein. Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet.*, **18**:529, (2002).

[164] E.T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, **106**:620, (1957).

[165] R. Kikuchi. A theory of cooperative phenomena. *Phys. Rev.*, **81**:988, (1951).

[166] G. An. A note on the cluster variation method. *J. Stat. Phys.*, **52**:727, (1988).

Acknowledgments

I would like to thank Amos for giving me the opportunity to work on these fascinating topics, for his support and enthusiasm in proposing new ideas, and for the friendly and motivating atmosphere I found in his group of research. I would like to thank also Alessandro for interesting and stimulating discussions and for his constant presence (except for what concerns our 'divertimentini'....), as well as for his inextinguishable irony.

I am grateful to J. Banavar for the fruitful interactions we had and for his unusual kindness, and A. Rinaldo for his suggestions and, in particular, for his great and contagious enthusiasm. I must aknowledge also A. Vespignani for providing me protein-interaction data, for his hospitality in Paris and for useful discussions we had toghether with M. Vergassola. I thank P. De Los Rios for his suggestions and in particular for an unusual discussion we had in front of a beer in a very crowded pub in Bologna.

I would like to thank all the people in the Statistical and Biological Physics sector, as well as the Condensed Matter Physics sector. I must acknowledge my roommates - Simone, who shared with me an unbelievable hot summer without A/C, and Hamed - and all other lunchmates - Claudio, Giacomo, Lorenzo, Michele,.... A special thank goes to my friend Claudio, with whom I lived at the beginning and at the end of my present stay in Trieste, for his constant and persistent enthusiasm and optimism. He is one of the nicest and friendly person I met.

I warmly thank all the people who made my stay here more human and enjoyable, beginning with my friends from the gym, in particular Claudio, Marco, Max, 'il Greco' and all the others, with whom I spent really nice, though sweaty, evenings and very few mornings(!); my friends from the spanish course, in par-

ticular Riccardo and my prof.; Giacomo, Simone, Stefano and Federico, and the Roma Club of Trieste for giving us the opportunity of following (or suffering for??) soccer matches of our team; my friend and ski-mate Gaetano, for our many terrific week-ends on the snow (I will miss them!); Valerio for a totally crazy and unusual friendship; Lucio, 'cause he's always nice and positive; Tiziano for the nice discussions and useful suggestions; the staff of the Pizzeria Formula 1 who saved several of my (very) late dinners; my friends in Rome, who contributed to save my (in-)sanity, in particular Giuseppe and Floriana; my friends from Salerno - Johnny and Ale - for a very pleasant sharing of Trieste's joys and sorrows, fun and frustration, and for a crazy and unforgettable new year's eve in Rome.

Among all, Francesco; don't know how many kilometers we have done to meet each other in these three years.

I hope I didn't forget anyone.

Finally, a special thank goes to my family - mum and dad, Filippo and Tommaso - for their teachings and continuous encouragements, and for supporting me everytime with great enthusiasm.