ISAS - INTERNATIONAL SCHOOL
FOR ADVANCED STUDIES

Degree correlations and clustering hierarchy in networks: measures, origin and consequences

Thesis submitted for the degree of
Doctor Philosophiæ

Candidate:
Alexei Vázquez Vázquez

Supervisor:
Prof. Amos Maritan
Prof. Alessandro Vespignani

October 2002

SISSA - SCUOLA
INTERNAZIONALE
SUPERIORE
I STUDI AVANZATI
TRIESTE
Via Beirut 2-4

TRIESTE

SISSA ISAS

Scuola Internazionale Superiore di Studi Avanzati
International School for Advanced Studies

Degree correlations and clustering hierarchy in networks: measures, origin and consequences

Thesis submitted for the degree of
Doctor Philosophiæ

Candidate:
Alexei Vázquez Vázquez

Supervisor:
Prof. Amos Maritan
Prof. Alessandro Vespignani

October 2002

To my parents

Acknowledgements

During the research leading to this Ph.D. thesis I have been involved with different colleagues. This have lead me the opportunity to learn different approaches to the study of graphs from a statistical mechanics point of view. A strong collaboration have been established with Alessandro Vespignani, Romualdo Pastor-Satorras and Yamir Moreno through the analysis of the Internet topology and the process running on top of it. Guided by Amos Maritan, and together with Alessandro Flammini and Alessandro Vespignani, I have been able to visualize and put in practice some applications to the study of proteins interactions and functions. Moreover, from the collaboration with Michele Leone, Yamir Moreno, Alessandro Vespignani and Martin Weigt I have get more inside into the general properties of statistical mechanics models on top of graphs. These fruitful collaborations would have not been possible without the charming and motivating environment at the International School for Advanced Studies and The Abdus Salam International Centre for Theoretical Physics. My gratitude is extended to all the staff of these scientific institutions for their help during my Ph.D. program.

I thank Alessandro Flammini, Amos Maritan, Yamir Moreno, and Alessandro Vespignani for a careful reading of this manuscript, whose quality increases considerably by their comments and suggestions. I thank Lada Adamic for connecting me to Clip2 Distribute Search Solutions to obtain the Gnutella network data. I thank Albert L. Barabási and collaborators for providing me the data of the co-autorship network of mathematical journals. I thank Wim van Criekinge for allowing me to reprint a very nice figure about two-hybrid experiments.

In all these scientific collaborations there is always a human side. In this respect I am very grateful to Alessandro Vespignani for teaching me more than science. His door has been always opened to me to discuss about anything from work to life. I also want to thanks Amos Maritan for his support during my Ph.D. In particular, for pushing me into the field of proteomics.

After all, we are not scientists with a human life but humans with a scientific life. Therefore, I have to thanks my parents for allowing and supporting me in the way to become a physicist and my girlfriend Gabriella for reminding me every time I see her that there is life beyond physics.

To all of you, thanks.

Contents

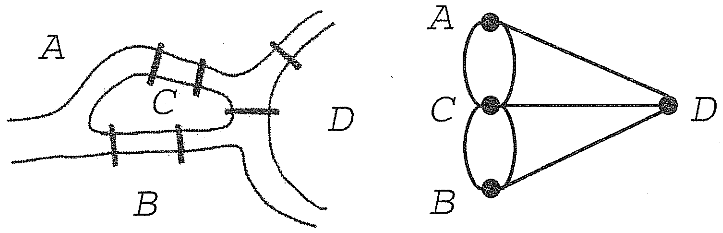
List of Figures

List of Tables

Chapter 1

Introduction

In the last few years there has been a great interest in the study of complex networks, where the term complex refers to one of the following properties: small world effect [147, 146], power law degree distribution [3, 45] and more recently degree correlations [115, 138, 103, 101]. This explosion has been possible due to the increase of available data and the technological capabilities to collect and process them. Now we count with the graph representation of a wide variety of systems with sizes ranging from hundred to billions of nodes. This includes technological networks like the physical Internet [102, 86, 69, 124, 53, 60, 30, 115, 138, 153], the World Wide Web [4, 12, 28], electronic mail web [48], and the map of electronic circuits [34]. Biological networks like the protein-protein interactions [132, 71, 70, 72, 144], metabolic paths [74, 145], and food webs [33, 126] have been recently studied. The social networks like the citation graph [82, 122, 134], the scientific collaboration graphs [105, 107, 57, 13], the sexual relations [85], among others, have been also analyzed.



The applicability of graph theory was already recognized in the eighteen century by the Swiss mathematician Leonard Euler. The system analyzed by Euler was the set of Königsberg bridges (see the figure above). The town of Königsberg in Prussia is divided by the river Pregel into four parts A, B,

C, and D. The problem under discussion was whether it is possible, starting from any part, to take a walk in such a way that every bridge is crossed exactly once. He notices that details like the size or shape of each part, among others, are irrelevant for this problem. Actually, all the information we need is contained in the graph representing the town. More precisely a multi-graph because there are multiple bridges between pairs of town parts. Using this representation and the fact that every town part can be reached from an odd number of bridges he concluded that a walk containing every bridge only once is not possible.

In general, as in the Euler problem, the graph theoretical approach to a system can be divided in: the graph representation of the system (graph representing the Königsberg bridges), the analysis of the topology of the resulting graph (all vertices have an odd number of incident edges), and its influence on processes running on top of that graph (the Euler walk). Moreover, in many cases it is important to understand the origin of the graph topology (Why the bridges were built in this way?) In the first step we represent the system in a set of units (the vertices) and the set of interactions among them (the edges). This representation is, in general, not unique and depends on the level of abstraction that we use. For instance, the World Wide Web can be represented as the set of web pages and hyper-links among them or, at a coarse grained level, it can be studied as the set of web sites and the hyper-links among them. Moreover, in this second representation the interactions can be weighted by the number of hyper-links going from a web page in one site to a web page in the other. Thus, the representation depends on the properties we want to study and on the level of simplification of our approach.

The graph representation gives us an abstract view of the system under study. Some of its topological properties may be deduced by just analyzing the representation itself but, in general, we should build the graph. For small systems, like in the Euler problem, the construction of the graph may be simple. However, for large systems like the World Wide Web or the Internet collecting the data to generate the graph is not an easy task at all. The construction of the graph is in many cases the subject of research projects requiring a large amount of resources, like the Internet mapping projects [102, 124, 69, 86], the two hybrid experiments to determine protein-protein interactions [132, 71, 70] and in social network research [97].

Once we have the graph we can characterize it using different topological measures. In particular the degree (the number of edges incident to a vertex), the minimum path distance between pairs of vertices and the clustering coefficient (the fraction of edges among the neighbors of a node) have attracted the attention of the physics community in the last few years. Watts and

Strogatz [147, 146] have shown that many real networks are characterized by a small average minimum path distance and a large clustering coefficient that together are named as the *the small world effect.* The name comes from the fact that we can reach every vertex in the graph crossing a small amount of edges. Moreover, Barabási and collaborators [12] have pointed out that many real networks are also characterized by power law degree distributions, giving an appreciable probability to observe high degree vertices. A more exhaustive analysis reveals that, in addition to power laws, truncated power laws and exponential distributions are also observed [6].

In order to explain these observations different models have been proposed. Before entering in their description we should mention the ancestor of all of them, the random graph model introduced by Erdös and Rényi in 1959 [50]. The random graph model is quite simple, in one of its variants one connects every pair of vertices with a probability p. Erdös and Rényi obtained that varying p there is a percolation transition from a graph made by disconnected clusters of vertices with size of the order 1 to a graph with a giant component containing a finite fraction of the vertices [51]. Moreover, in any of the phases the vertex degrees are distributed according to a Poisson distribution, the average minimum path distance between vertices in the largest cluster scales logarithmically with its size, and they have a small average clustering coefficient [23].

Random graphs thus exhibit one of the properties of real networks, the logarithmic scaling of average minimum path distance with the number of vertices. However they do not explain the other two properties: high clustering and broad degree distribution. Taking into account that a large clustering coefficient is characteristic of some regular lattices Watts and Strogtz [147] introduced the small world model, that is an interpolation between a regular lattice and a random graph. In this model one starts with a regular lattice and rewire each edge with a probability p, choosing one of the ending vertices at random. Numerical simulations and analytical studies [15, 111, 112, 40, 14, 43] have shown that there is a crossover graph size $N_c(p)$ separating the region $N < N_c(p)$ where the average minimum path distance scales linearly with N as in a regular lattice to $N > N_c(p)$ where it scales with $\ln N$ as for random graphs. Moreover, the clustering coefficient remains relatively large provided p is not to close too 1 [14]. However, as in the random graph model, the degree distribution is peaked around its mean value [14].

The random graph model can be easily generalized to obtain graphs with power law degree distributions, or any arbitrary degree distribution, generating random graphs with a given degree sequence [2, 98]. These graphs can be used to study the influence of the degree distribution but not their origin.

Barabási and collaborators proposed a mechanism that explains the origin of power law degree distributions [9, 10]. This mechanism is based on two fundamental properties of many real networks, their growing nature and the existence of a preferential attachment: new vertices added to the graph are attached preferentially to high degree vertices. In particular a linear preferential attachment, where the probability to get connected to a vertex is proportional to its degree, leads to power law degree distributions [12, 11]. The preferential attachment mechanism can be generalized in many ways. A sub-linear preferential attachment leads to bounded degree distributions while a super-linear one leads to a graph with a single hub connected to almost any other vertex [81, 79]. The power laws can be also truncated after the introduction of other ingredients like aging [42], bounded capacity [6] or limited information [100]. Moreover, the introduction of quenched [20] and annealed [47, 135] disorder leads to logarithmic corrections and multi-fractal scaling, respectively.

Finally, once we know the topology of the graph representing our system we can investigate the influence of this topology on the different processes that can be performed on top of this graph. Analytical and numerical studies of percolation [5, 37, 32, 38], spreading phenomena [117, 116, 90, 99, 108], the Ising model [41, 84] on top of random graphs with power law degree distributions reveal that the relevant parameter is in this case the ratio between the second and the first moments of the degree distribution. If the second moment diverges then the system will not exhibit a phase transition, *i.e.* the graph has always a giant component that is robust under random vertex or node removal, the spreading of a disease has a finite prevalence, and the system is always ferromagnetic.

The topology of real networks is also characterized by degree correlations [115, 103] and clustering hierarchy [138, 121]. Moreover, it has been shown that growing network models with [79] and without [31] preferential attachment lead to non-trivial degree correlations. Therefore, the extension of previous results for uncorrelated graphs is of utmost importance. The study of models on graphs with degree correlations is quite recent [103, 101, 18, 140]. Some expressions for the size of the giant component and related quantities have been obtained in Ref. [103] whereas an equation for the epidemic threshold has been provided in [101, 21]. General statistical mechanics approaches for models on correlated graphs has also been developed in Refs. [18, 140]. Moreover, it is not clear how degree correlations can affect the performance of optimization algorithms.

In this work we will present our contribution in the study of complex networks. In the 2^{nd} Chapter we introduce some basic concepts of graph theory and some of the most significant graph models. In the 3^{rd} Chapter we introduce our proposal to study degree correlations and clustering hierarchy in real networks. Most of our findings in this direction have been obtained from a detailed analysis of the Internet topology and, therefore, we use it as a case study. We investigate the scale-free properties of the Internet maps, focusing on the degree and betweenness distributions. Furthermore, we propose two metrics based on the degree and the clustering correlation functions, that appear to sharply characterize the hierarchical properties of Internet maps. Finally, we extent the use of these metrics to discriminate between other real networks that appear similar simply on the basis of their degree distribution.

In the 4^{th} Chapter, we study a hypothesis for the origin of an effective preferential attachment, based on growing graph models with local rules. We investigate three different models with applications to different real graphs. In all of them we obtain an effective preferential attachment and an inverse proportionality between the clustering coefficient and the vertex degree. Moreover, using numerical simulations we also show that these models lead to degree correlations.

In the 5^{th} Chapter we introduce a general statistical mechanics approach to investigate the influence of degree correlations. Particular attention is devoted to the problem of percolation and vertex covering. The study of the percolation problem is related to the robustness of graphs upon removal of its vertices or edges. The resilience to damage has a great impact in the performance of communication and biological networks. Besides, we have chosen the vertex covering problem for two reasons: It belongs to the basic NP-hard optimization problems over graphs [55], and has found applications in monitoring Internet traffic [26] and denial of service attack prevention [120].

Finally, using our main results, topology analysis, graph generators, and statistical mechanics models, in the 6^{th} Chapter we propose a method for protein function assignment using protein-protein interaction data. This example constitutes a case study where, as in the Euler problem, we can see the different parts and potentiality of a graph theoretical approach.

The results presented in this thesis and related works have appeared in the following papers:

- A. Vázquez, *Scale free networks generated by recursive searches*, Europhys. Lett. **54**, 430–435 (2001).
- R. Pastor-Satorras, A. Vázquez and A. Vespignani, *Dynamical and correlation properties of the Internet*, Phys. Rev. Lett. **87**, 258701–258704 (2001).
- A. Vázquez, R. Pastor-Satorras and A. Vespignani, *Large-scale topo-*

logical and dynamical properties of Internet, Phys. Rev. E **65**, 066130–066141 (2002).

- Y. Moreno and A. Vázquez, *The Bak-Sneppen model on Scale-Free Networks*, Europhys. Lett. **57**, 765 (2002).
- M. Leone, A. Vázquez, A. Vespignani, and R. Zecchina, *Ferromagnetic ordering in graphs with arbitrary degree distribution*, Eur. Phys. J B **28**, 191–197 (2002).
- A. Vázquez, A. Flammini, A. Maritan and A. Vespignani, *Modeling of protein interaction networks*, cond-mat/0108043.
- A. Vázquez, R. Pastor-Satorras and A. Vespignani, *Internet topology at the router and autonomous system level*, cond-mat/020608.
- A. Vázquez and M. Weigt, *Computational complexity arising from degree correlations in networks*, cond-mat/0207035.
- A. Vázquez, A. Flammini, A. Maritan and A. Vespignani, *Assignment of protein function from protein-protein interactions*, submitted to Nat. Biotech.
- A. Vázquez and Y. Moreno, *Resilience to damage of graphs with degree correlation*, cond-mat/0209182.
- A. Vázquez, M. Boguñá, Y. Moreno, R. Pastor-Satorràs, and A. Vespignani, *Topology and correlations in structured scale-free networks*, cond-mat/0209183.

Chapter 2

Preliminaries

2.1 Graph representation of real networks

Let us now introduce some basic definitions and concepts that will be used throughout this work. A *graph* $\mathcal{G}$ is defined by a pair of sets $(\mathcal{V}, \mathcal{E})$ such that $\mathcal{V} \neq \emptyset$ and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. $\mathcal{V}(\mathcal{G}) = \{1, \ldots, N\}$ is the set of vertices of $\mathcal{G}$ and N is the number of vertices or graph size. $\mathcal{E}(\mathcal{G}) = \{(i_1, j_1), \ldots, (i_E, j_E)\}$ is the set of edges and E the number of edges in the graph. If the edge (i, j) is not equivalent to (j, i) then the graph is said directed and (i, j) represents an edge from vertex i to vertex j. Otherwise the graph is said undirected and $(i, j) \equiv (j, i)$ represents an edge connecting vertex i and j, which are said adjacent. An example of an undirected graph is shown in Fig. 2.1. It is made of $N = 18$ vertices and $E = 42$ edges. In the following we will focus on undirected graphs, with some quotes to directed graphs when necessary.

Degree: The degree d_i of a vertex i is defined as the number of edges incident to it, and $\langle d \rangle$ is the average of d_i over all vertices in the graph. In an undirected graph each edge contributes to the degree of two vertices and, therefore,

$$\langle d \rangle = \frac{2E}{N}. \tag{2.1}$$

For instance, the central vertex in Fig. 2.1 have degree $d = 18$ and the average degree is $\langle d \rangle = 2 \times 42/18 \approx 4.7$. The average degree for some real networks is shown in Table 2.1. In average each vertex has a small degree compared with that of a fully connected graph of the same size ($\langle d \rangle = N - 1$), *i.e.* they are sparse graphs. In directed graphs it is also convenient to divide the vertex degree into in-degree and out-degree, counting the number of edges going in and out from the vertex.

Clustering coefficient: The number of neighbors of a vertex i is given by its degree d_i. On their turn, these neighbors can be connected among them

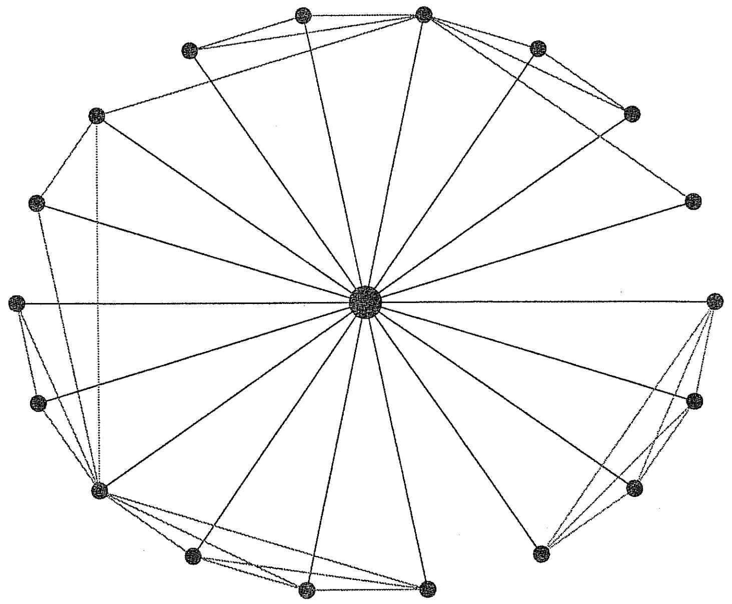


Figure 2.1: The co-authorship sub-graph induced by A. Vázquez. The central vertex corresponds to A. Vázquez and the neighbor vertices the scientists that have written a paper with him. Two scientists (vertices) are connected by an edge if they have been co-authors of at least one article.

forming a triangle with vertex i. The clustering coefficient c_i is then defined as the ratio between the number of edges e_i among the d_i neighbors of a given vertex i and its maximum possible value, $d_i(d_i-1)/2$, *i.e.*

$$c_i = \frac{2e_i}{d_i(d_i-1)}. \tag{2.2}$$

The average clustering coefficient $\langle c\rangle$ is the average of c_i over all vertices in the graph. The clustering coefficient provides a measure of how well are locally interconnected the neighbors of any vertex. For instance, the number of edges among the $d=18$ neighbors of the central vertex in Fig. 2.1 is $e=24$ resulting $c=2\times 24/18\times 17\approx 0.16$.

The maximum value of $\langle c\rangle$ is 1, corresponding to a fully connected graph. For random graphs [23] (see next section), which are constructed by connecting nodes at random with a fixed probability p, the clustering coefficient decreases with the network size N as

$$\langle c\rangle_{\rm rg} = \frac{\langle d\rangle}{N}. \tag{2.3}$$

On the contrary, it is independent of N for regular lattices. The average clustering coefficient for different real networks is shown in Table 2.1. As

Network	N	$\langle d\rangle$	$\langle l\rangle$	$\langle c\rangle$	$\langle d\rangle/N$
AS	10515	4.1	3.7	0.29	3.9×10^{-4}
router	228298	2.8	9.5	0.030	1.2×10^{-5}
WWW	325729	9.0	7.2	0.23	2.8×10^{-5}
Gnutella	727	3.6	4.3	0.014	5.0×10^{-3}
PIN	3278	2.4	7.6	0.11	7.3×10^{-4}
Math	70901	4.2	8.5	0.46	5.9×10^{-5}

Table 2.1: Average properties of some real networks. AS and router are the autonomous system [102] and router [69] level graph representations of the Internet, respectively. WWW a sub-graph of the WWW network, a data set collected by the Notre Dame group on Complex Networks (http://www.nd.edu/˜networks). Gnutella is the Gnutella peer to peer network, provided by Clip2 Distributed Search Solutions. PIN if the protein-protein interaction graph of *Saccharomices Cerevisiae* as obtained from two hybrid experiments [70]. Math is the co-authorship graph obtained from all relevant journals in the field of mathematics and published in the period 1991-1998 [13]. The average clustering coefficient of the corresponding random graph with the same number of vertices and edges ($\langle d\rangle/N$) is also shown for comparison.

it can be seen, it takes values orders of magnitude larger than that of a random graph with the same number of vertices and edges. Therefore, these networks are far from being random. The existence of a high clustering coefficient may have different origins. In the Internet graph representation the edges among vertices are considered as equivalent, but they are actually characterized by a real space length corresponding to the actual length of the physical connection between AS. The larger is this length, the higher the costs of installation and maintenance of the line, favoring therefore the connections between nearby nodes. It is thus likely that nodes within the same geographical region will have a large number of connection among them, increasing in this way the local clustering coefficient. Something similar takes place in the co-authorship graph, where the publications play the role of the geographical constraint in the Internet. In fact, a scientist being co-author of a paper with m authors will have the other $m-1$ authors as neighbors. At the same time, these $m-1$ authors will be mutually connected resulting in a large clustering coefficient. This is a general property of social graphs, that are actually bipartite graphs [110]. Scientists are connected by collaborations, actors by films, people by places or common activities, *etc.*

Minimum path distance: Two vertices i and j are said to be connected

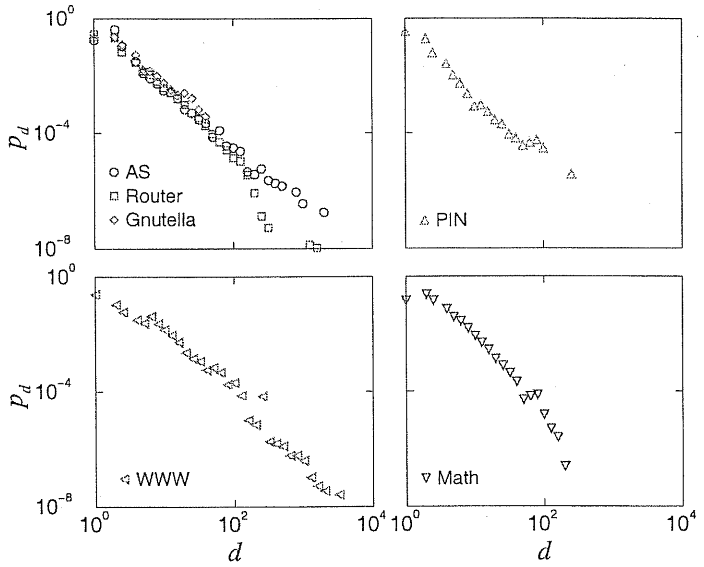


Figure 2.2: Degree distribution of the real graphs introduced in Tab. 2.1.

if one can go from node i to j following the edges in the graph. The path from i to j may be not unique and its distance is given by the number of vertices visited. The minimum path distance l_{ij} is defined as the shortest path distance between two nodes i and j, and $\langle l \rangle$ is its average over every pair of vertices in the graph. As it can be seen from Table 2.1 the average minimum path distance is in general small. These networks thus exhibit what is known as the *small-world* effect [147, 146]: in average one can go from one vertex to any other in the graph passing through a very small number of intermediate vertices. This necessarily implies that besides the local connections which contribute to the large clustering coefficient, there are some hubs which connect different parts of the graph, strongly decreasing the average minimum path distance.

Degree distribution: The degree distribution $p_d = P(d_i = d)$ is the probability that a vertex has degree d. For random graphs [23] it is peaked around the average degree. However, Barabási and collaborators [12] have pointed out that many real networks are characterized by power law degree distributions, giving an appreciable probability to observe high degree vertices (see Fig. 2.2). A more exhaustive analysis reveals that, in addition to power laws, truncated power laws and exponential distributions are also observed [6].

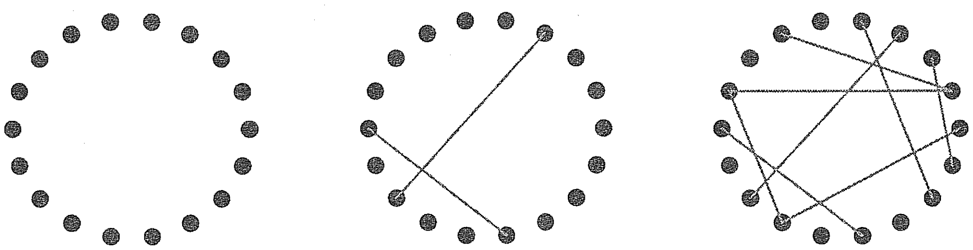


Figure 2.3: Schematic representation of the random graph model. Left: starting from a set of disconnected vertices, each pair of vertices is connected with a probability p. Middle and right: two realizations with a small p in the middle and a larger p in the right.

2.2 Random graph model

The random graph model introduced by Erdös and Rényi [50] is probably the simplest graph model including some heterogeinity. Starting from a set of N disconnected vertices, each pair of vertices is connected with a probability p (see Fig. 2.3). Due to is simplicity many of the properties of the random graph model can be computed exactly in the large N limit. The average number of edges in the graph is just a fraction p of the $N(N-1)/2$ possible edges yielding the averaged degree

$$\langle d \rangle = \frac{2E}{N} = pN, \tag{2.4}$$

where the last equality holds for $N \gg 1$. The degree distribution can be also computed taking into account that the degree of a vertex d_i is given by a binomial process [22]. In fact, the probability that vertex i is connected to another vertex is p. This process is repeated over the $N-1$ vertices excluding i resulting the binomial distribution

$$p_d = C^d_{N-1} p^d (1 - p^d), \tag{2.5}$$

where C^d_{N-1} is the number of ways in which a vertex is connected to d other vertices and not connected to $N-1-d$ vertices. In most of the cases of practical imterest the average degree is of the order of 1 and, therefore, from Eq. (2.4) it follows that $p \sim N^{-1}$. Taking the limit $N \gg 1$ with $\langle d \rangle = \text{const.}$ the binomial distribution is approximated by the Poisson distribution

$$p_d = e^{-\langle d \rangle} \frac{\langle d \rangle^d}{d!}. \tag{2.6}$$

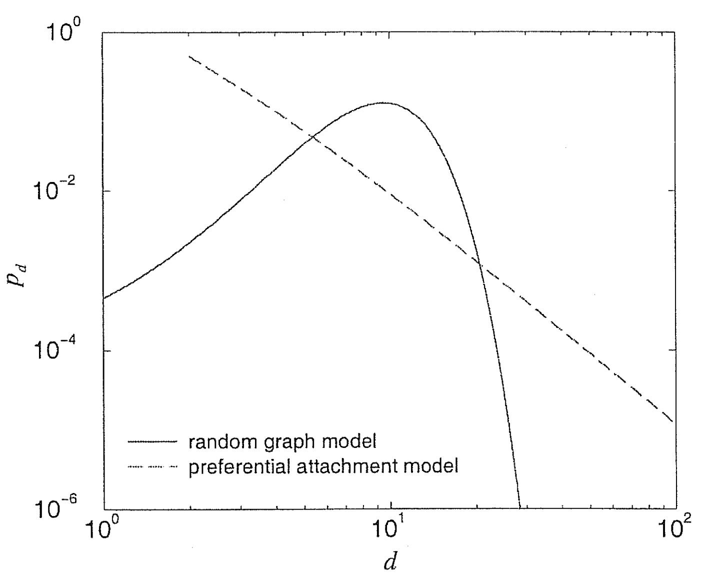


Figure 2.4: Degree distribution of the random graph with $pN = \langle d \rangle = 10$ and preferential attachment model with $m = 2$ analyzed in Sec. 2.4.

In Fig. 2.4 we plot the degree distribution obtained from this expression. It is characterized by a peak around the mean that is determined by the $d!$ in the denominator of Eq. (2.6).

Another magnitude that can be easily computed is the clustering coefficient. In the random graph model the probability that an edge between two vertices exist is independent of the existence of other edges and equal to p. Hence, in average, there will be $p\ d(d-1)/2$ edges among the $d(d-1)/2$ neighbors of a vertex with degree d. Then from Eqs. (2.2) and (2.4) it follows that

$$\langle c \rangle = p = \frac{\langle d \rangle}{N} \tag{2.7}$$

For sparse graphs with $\langle d \rangle \sim 1$ the clustering coefficient clearly decreases with increasing the graph size. Moreover, the clustering coefficient is independent of the vertex degree, *i.e.* if we restrict the average over vertices with a given degree d we get the same result in Eq. (2.7) independent of d.

The properties discussed so far do not exhibit any substantial change as a function of p. On the contrary, if we focus in global properties a qualitative change is observed when $p = 1/N$ ($\langle d \rangle = 1$). For instance, let us consider the fraction of vertices S in the largest cluster of connected vertices. For $p \approx 0$ all vertices are practically disconnected and, therefore, $S \sim 1/N \ll 1$. On the other hand, for $p \approx 1$ the graph is almost fully connected resulting

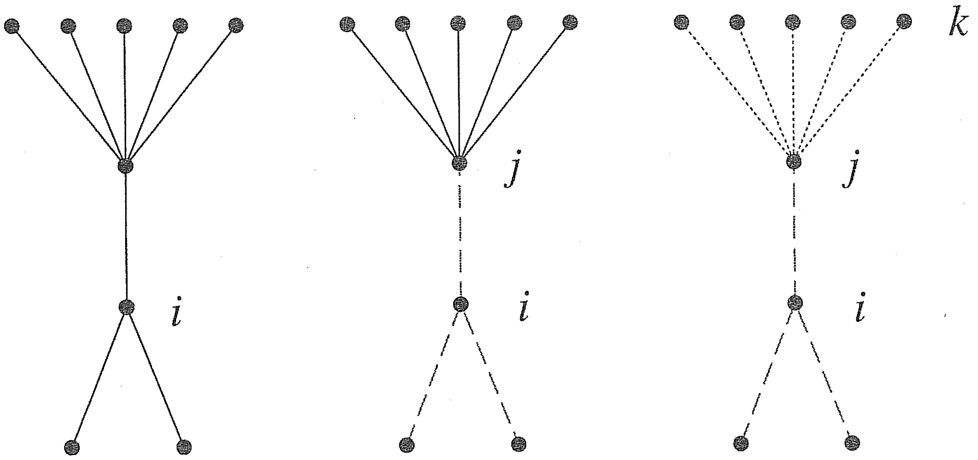


Figure 2.5: Schematic representation of the iterative approach to compute the size of the giant component.

$S \approx 1$.

The size of the giant component for intermediate values of p can be determined using different methods [51, 110]. Here we will use some iterative approach that is common to the generating function scheme [110] and to more general statistical mechanics approaches like the replica and cavity methods [96, 95]. Let u be the probability that a vertex does not belong to the giant component after we remove one of its edges. Now, take a given vertex i of the graph and remove all its edges (see Fig. 2.5). If the graph does not contain loops then the probability that one of the neighbors of i is not in the giant component is independent of the other neighbors and equal to u. Hence, the probability that a vertex of degree d does not belong to the giant component is given by u^d resulting

$$S = 1 - \sum_d p_d u^d. \tag{2.8}$$

This procedure can be iterated to compute u. To do that let us focus on one of the neighbors j of vertex i. First notice that the degree of that vertex is not distributed according to p_d. In the random graph there are no correlations among the different edges and, therefore, the probability that an edge coming from a randomly chosen vertex points to another vertex with degree d is

$$q_d = \frac{dp_d}{\langle d \rangle}. \tag{2.9}$$

Now, let $k \neq i$ an index denoting any neighbor of j, excluding i (see Fig. 2.5). The probability u that vertex j does not belong to the giant component,

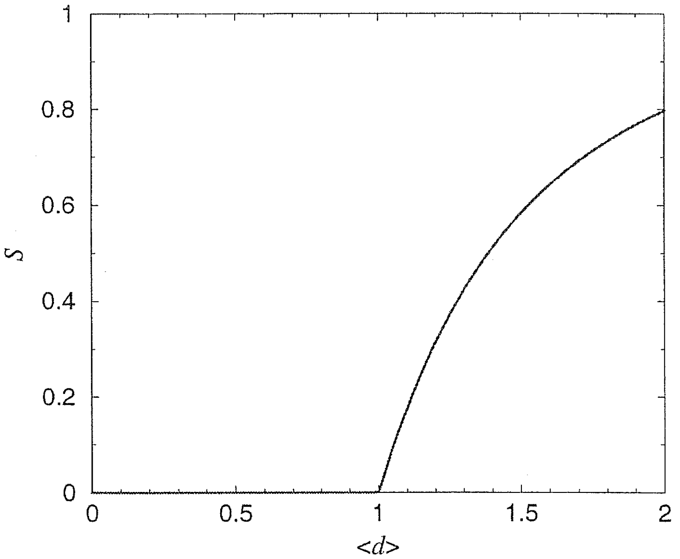


Figure 2.6: Size of the giant component of the random graph model as a function of the average degree.

given that the edge (i, j) has been removed, is equal to the probability u^{d-1} that any of its $d-1$ remaining neighbors is not in the giant component, provided the edges (j, k) have been removed. Taking the average over the degree distribution of vertex j we obtain

$$u = \sum_d q_d u^{d-1}. \tag{2.10}$$

Equations (2.8) and (2.10) can be rewritten as

$$S = 1 - G_0(u), \qquad u = G_1(u), \tag{2.11}$$

where

$$G_0(x) = \sum_d p_d x^d, \qquad G_1(x) = \sum_d q_d x^{d-1}. \tag{2.12}$$

$G_0(x)$ and $G_1(x)$ are the generating functions of p_d and q_d. The term generating function comes from the fact that p_d and q_d can be "generated" taking the derivatives of the functions $G_0(x)$ and $G_1(x)$, respectively [110].

For the random graph models the generating functions can be computed exactly resulting

$$G_0(x) = G_1(x) = e^{\langle d\rangle(x-1)}. \tag{2.13}$$

Then, the size of the giant component can be computed solving Eq. (2.11) with these generating functions. In Fig. 2.6 we plot S as a function of the average degree $\langle d\rangle$. For $\langle d\rangle < 1$ we obtain $S = 0$ which corresponds with a graph made of disconnected clusters of vertices with size of the order of 1. On the other hand, for $\langle d\rangle > 1$ there is a giant component containing a finite fraction of the vertices. Thus, at $\langle d\rangle = 1$ a percolation transition takes place. As the giant component, other global magnitudes have a qualitative change of behavior at $\langle d\rangle = 1$. For instance, let us consider the average minimum path distance between connected vertices $\langle l\rangle$. For $\langle d\rangle < 1$ it is finite because all clusters contain a fraction $1/N$ of vertices. However, for $\langle d\rangle > 1$ it scales logarithmically with the graph size according to

$$\langle l\rangle \sim \frac{\ln N}{\ln \langle d\rangle}. \tag{2.14}$$

This scaling is much slower than that of a D-dimensional regular lattice where $\langle l\rangle \sim N^{1/D}$.

In summary, the random graph model is characterized by a Poisson degree distribution, a clustering coefficient of the order N^{-1} and a small average minimum path distance. Nevertheless, in the previous section we have seen that real graphs are characterized by broad degree distributions that in many cases are given by a power-law decay. Moreover, as it is shown in Tab. 2.1 its clustering coefficient is order of magnitudes smaller than that of real graphs. Thus, the random graph model is not a good approximation of real graphs. In the following sections we show how some of these discrepancies can be partially solved.

2.3 Small world model

To solve the discrepancy between the small clustering coefficients of random graphs and that of real graphs Watts and Strogatz [147, 146] introduced the small world model that its an interpolation between regular lattices and the random graph model. They exploited the fact that many regular lattices, like the one despited in Fig. 2.7, exhibit a finite clustering coefficient independent of the lattice size. Their model is constructed starting from a regular lattice and them making a random reconnection of a certain fraction of the edges. Here we will consider a slightly different model introduced latter by Newman and Watts [111], with the same qualitative features but much easier for analytical treatment.

Consider a regular lattice with a finite clustering coefficient C and coordination K. An example is given by the K-ring in Fig. 2.7, where the

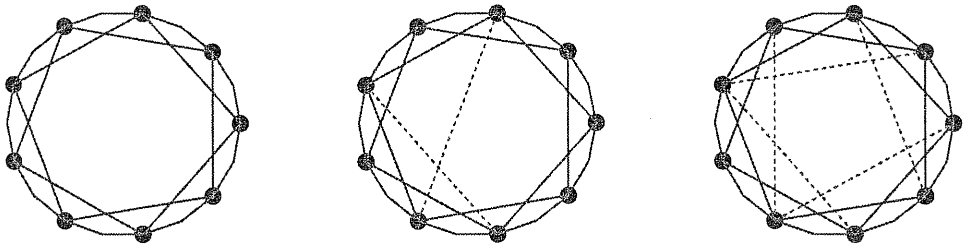


Figure 2.7: Schematic representation of the small world model. Left: starting from a regular lattice, each pair of vertices is connected with a probability p. Middle and right: to realizations with a small p in the middle and a larger p in the right.

vertices are put in a ring and they are connected to the K closest neighbors. For such a lattice the clustering coefficient is

$$C = \frac{2(K-2)}{4(K-1)}. \tag{2.15}$$

Now on top of this lattice generate a random graph. To be more precise, in addition to the existing edges given by the regular lattice, connect every pair of vertices with probability $p = z/N$. In this way we are actually making a multi-graph, where two vertices can be connected by more than one edge. However, for $p \sim 1/N$ the existence of these multi-edges is irrelevant. The average degree of the graph constructed in this way is

$$\langle d \rangle = z + K. \tag{2.16}$$

Moreover, since the only randomness is introduced by the random placement of edges then the degree distribution will be that of a random graph shifted by Z, the coordination number of the original regular lattice, *i.e.*

$$p_d = e^{-\langle d \rangle} \frac{\langle d \rangle^{d-Z}}{(d-Z)!}. \tag{2.17}$$

With respect to the average minimum path distance, numerical simulations and analytical studies [15, 111, 112, 40, 14, 43] have shown that there is a crossover graph size $N_c(p)$ separating the region $N < N_c(p)$ where the average minimum path distance scales linearly with N as in a regular lattice to $N > N_c(p)$ where it scales with $\ln N$ as for random graphs. The crossover size from one regime to the other scales as

$$N_c(p) \sim p^{-1/D}, \tag{2.18}$$

where D is the dimension of the original lattice. Now, we have already mentioned that to obtain an agreement with real graphs p should scale as $p = z/N$ with $z \sim 1$. Thus, from Eq. (2.18) it follows that

$$N_c(z) \sim z^{-1/(D-1)}. \tag{2.19}$$

Since $z \sim 1$ then $N_c(z) \sim 1$ and taking into account that $N \gg 1$ in most real graphs we can conclude that they are in the small world regime, where the average minimum path distance scales logarithmically with the graphs size.

Hence, with respect to the degree fluctuations and the average minimum path distance the small world model exhibits the same features of a random graph. On the contrary, an important difference is obtained for the clustering coefficient. The random placement of edges on top of the regular lattice increases each vertex degree to $d_i = K + z_i$. However, the correction to the number of edges among the neighbors of a vertex is of the order of z/N and, therefore, can be neglected. Hence, in the limit $N \gg 1$ the local clustering coefficient at a vertex i is given by

$$c_i = C\frac{K(k-1)}{(K+z_i)(K+z_i-1)}, \tag{2.20}$$

where C is the clustering coefficient of the original lattice and the second factor is just the ratio between the number of possible edges among the neighbors of vertex i, before and after the random placement of edges. Now, since the distribution of z_i is peaked around its mean z we obtain the following approximation

$$\langle c \rangle \sim C\frac{K(K-1)}{(K+z)(K+z-1)}. \tag{2.21}$$

In most real graphs the average degree is of the order of 1 and, therefore, from Eq. (2.16) it follows that z should be of the order of 1. In such a case, the clustering coefficient of the small world model will remain finite independent of the graph size and it takes values comparable to that of the original regular lattice. In this way we obtain a better agreement with the real graphs clustering coefficient. However, the degree distribution is essentially the same as in the random graph model and, therefore, it is in disagreement with that of real graphs.

2.4 Preferential attachment model

Barabási *et al* [12, 11] pointed out that to obtain the correct degree distribution one should take into account two fundamental properties of real graphs:

1- they are growing and 2- there is a preferential attachment. The growing nature of many real graphs is a fact. For instance the number of web pages and routers in the WWW and the Internet, respectively, is growing exponentially. The genetic, protein, and metabolic networks of living organisms grow in the course of evolution. Social networks becomes bigger in the course of history. Hence, this growing nature should be reflected in the graph topology.

However, the growing mechanism along does not lead to the power law degree distribution observed in real graphs [12, 11]. If the new incoming vertices are attached to old vertices selected at random then an exponential distribution is observed [11, 31]. As a different with the Poisson distribution the exponential distribution does not have a peak. Nevertheless, the decay for large degrees is still quite fast. A slower decay should be obtained if we take into account that the incoming vertices are not attached at random to old vertices. Actually, the existing vertices that are more "visible" will have a higher probability to receive a new connection to them. Barabási *et al* [12, 11] proposed the vertex degree as a measure of visibility. That is, vertices with a larger degree will have a larger probability to receive new edges from added vertices. To be more precise they defined the following model. Starting with a small set of m_0 of vertices and E_0 edges among them:

- *Growth*: At every time step we add a new vertex with $m \leq m_0$ edges;
- *Preferential attachment*: the vertex at the other end of these m new edges is chosen with probability

$$\Pi(d_i) = \frac{d_i}{\sum_j d_j}. \tag{2.22}$$

Different approaches have been proposed to compute the degree distribution of the preferential attachment model, including a continuum theory [11], a master equation scheme [46, 44] and a rate equation approach [81, 79]. Here we will use the rate equation approach. In this case one focuses on the average number $n_d(t)$ of vertices with degree d after t time steps. Now, when we add a new vertex with m new edges its quite improbable that for $N \gg 1$ two of these edges goes to the same vertex. Thus, we can assume that on each step the degree of a vertex can only increase by one. Under this approximation the evolution of $n_d(t)$ satisfy the rate equations

$$\frac{\partial n_d(t)}{\partial t} = A_{d-1}n_{d-1}(t) - A_d n_d(t) + A_{\text{new}}\delta_{dm}, \tag{2.23}$$

where A_d is the probability per unit time that a vertex with degree d increases its degree by one and A_{new} is the number of new vertices added per unit time.

The attachment and addition rate corresponding to the model introduced above reads

$$A_d = m\frac{d}{N\langle d\rangle}, \qquad A_{\text{new}} = 1. \tag{2.24}$$

The factor m comes from the fact that m new edges are added and $A_{\text{new}} = 1$ because only one vertex is added, on each step. Moreover, for $N \gg 1$ the average degree is given by

$$\langle d\rangle = \frac{2mN + E_0}{N + m_0} \approx 2m. \tag{2.25}$$

From Eqs. (2.23)-(2.25) it follows that

$$\frac{\partial n_d(t)}{\partial t} = \frac{d-1}{2}p_{d-1}(t) - \frac{d}{2}p_d(t) + \delta_{dm}, \tag{2.26}$$

where $p_d(t) = n_d(t)/N(t)$ is the degree distribution. The stationary solution to these equations can be obtained setting $\frac{\partial n_d(t)}{\partial t} = 0$, resulting

$$p_d = \begin{cases} \frac{2}{2+m} & \text{, if } d = m \\ \frac{d-1}{2+d}p_{d-1} & \text{, if } d > m \end{cases} \tag{2.27}$$

Iterating this equation we finally obtain

$$p_d = \frac{2m(m+1)}{d(d+1)(d+2)}. \tag{2.28}$$

Now, for $d \gg 1$

$$p_d \sim d^{-\gamma}, \qquad \gamma = 3. \tag{2.29}$$

Hence, the preferential attachment model of Barabási *et al* leads to a power law degree distribution with a exponent $\gamma = 3$. This exponent can be tuned to different values by considering more generals forms of the attachment rate [44, 79], of the type

$$A_d = \frac{a + bd}{\sum_{d'} a + bd'}, \tag{2.30}$$

and addition rates $A_{\text{new}} < 1$. In this way one obtains exponents in the range $2 < \gamma < \infty$. However, these results are only valid for attachment rates linear in d [79]. The use of a sub-linear form ($A_d \sim d^{\alpha}$, $\alpha < 1$) leads to stretched exponential distributions. On the other hand, super linear attachment rates ($A_d \sim d^{\alpha}$, $\alpha > 1$) have as outcome a "gelation" process, where there is a vertex connected to almost every other vertex in the graph.

Other properties of the preferential Barabási *et al* model are more difficult to compute analytically. The average minimum path distance growths logarithmically with the graph size with a double logarithm correction [25]

$$\langle l \rangle \sim \frac{\ln N}{\ln \ln N}. \tag{2.31}$$

Moreover, numerical simulations show that [3] the average clustering coefficient decreases with increasing the graph size as

$$\langle c \rangle \sim N^{-0.75}. \tag{2.32}$$

This decay is slower than the N^{-1} obtained for the random graph model but it is still to fast to explain the clustering coefficient observed in real graphs.

In summary, the Barabási *et al* model provide us a mechanism to obtain power law degree distributions in growing networks. If one consider other magnitudes like the clustering coefficient then one may conclude that this model is still insufficient to describe real graphs. However, we should not focus on the detailed properties of the model but on its philosophy. That is, if we assume that there is growth and an effective linear preferential attachment then we obtain a scale-free degree distribution. Actually, this effective preferential attachment have been measured in different real graphs, including the Internet [73, 115] and a variety of scientific collaboration graphs [73, 104, 13], supporting the hypothesis of a linear attachment rate. With regard to the other topological properties, we can construct many models with preferential attachment and different clustering coefficient, minimum path distances, and other metrics [24]. The interesting question is why the preferential is linear and it has not a definitive answer yet. In the fourth chapter we investigate one hypothesis.

Chapter 3

Hierarchy and correlations: the Internet and other real networks

We can go beyond the study of average measures and their distributions to detect correlations and hierarchy. Most of our findings in this direction have been obtained from a detailed analysis of the Internet topology and, therefore, we use it as a case study. We investigate the scale-free properties of the Internet maps, focusing on the degree and betweenness distributions. Furthermore, we propose two metrics based on the degree and the clustering correlation functions, that appear to sharply characterize the hierarchical properties of Internet maps. Finally, we extend the use of these metrics to discriminate between other real networks that appear similar according to their degree distribution.

The relentless growth of the Internet goes along with a wide range of internetworking problems related to routing protocols, resource allowances, and physical degree plans. The study and optimization of algorithms and policies related to such problems rely heavily on theoretical analysis and simulations that use model abstractions of the actual Internet. On the other hand, in order to extract the maximum benefit from these studies, it is necessary to work with reliable Internet topology generators. The basic priority at this respect is to best define the topology to use for the network being simulated. This implies the characterization of how routers, hosts, and physical links interconnect with each other in shaping the actual Internet.

In the last years, research groups started to deploy technologies and infrastructures in order to obtain a more detailed picture of the Internet. Several studies, aimed at tracking and visualizing the Internet large scale topology and/or performance, are leading to Internet mapping projects at different resolution scales. These projects typically collect data on Internet elements (routers, domains) and the connections among them (physical edges, peer connections), in order to create a graph-like representation of large parts of the Internet in which the vertices represent those elements and the edges represent the respective connections. Mapping projects focus essentially on two levels of topological description. First, by inferring router adjacencies it has been possible to measure the Internet router (IR) level topology. The second measured topology works at the autonomous system (AS) level and the degree obtained from AS routing path information. Although these two representations are related, it is clear that they describe the Internet at rather different scales. In fact, each AS groups a generally large number of routers, and therefore the AS maps are in some sense a coarse-grained view of the IR maps.

Internet maps exhibit an extremely large degree of heterogeneity and the use of statistical tools becomes mandatory to provide a proper mathematical characterization of this system. Statistical analysis of the Internet maps fabric have pointed out, to the surprise of many researchers, a very complex degree pattern with fluctuations extending over several orders of magnitude [53]. In particular, it has been observed a power-law behavior in metrics and statistical distributions of Internet maps at different levels [53, 60, 30, 138, 137, 151, 130, 35, 29]. This evidence makes the Internet an example of the so-called *scale-free* networks [3] and uncover a peculiar structure that cannot be satisfactorily modeled with traditional topology generators. Previous Internet topology generators, based in the classical Erdös and Rényi random graph model [50, 23] or in hierarchical models, yielded an exponentially bounded degree pattern, with very small fluctuations and in clear disagreement with the recent empirical findings. A theoretical framework for the origin of scale-free graphs has been put forward by Barabási and Albert [3] by devising a novel class of dynamical growing networks. Following these ideas, several Internet topology generators yielding power-law distributions have been subsequently proposed [92, 91, 75].

Data gathering projects [102, 124, 49, 69, 86] are progressively making available larger AS and IR level maps which are susceptible of more accurate statistical analysis and raise new and challenging questions about the Internet topology. For instance, statistical distributions show deviations from the pure power-law behavior and it is important to understand to which extent the Internet can be considered a scale-free graph. The way these scaling anomalies—usually signaled by the presence of cut-offs in the corresponding statistical distributions—are related to the Internet finite size and physical constraints is a capital issue in the characterization of the Internet and in the understanding of the dynamics underlying its growth. A further important issue concerns the fact that the Internet is organized on different hierarchical levels, with a set of backbone edges carrying the traffic between local area providers. This structure is reflected in a hierarchical arrangement of administrative domains and in a different usage of edges and degree of vertices. The interplay between the scale-free nature and the hierarchical properties of the Internet is still unclear, and it is an important task to find metrics that can exploit and characterize hierarchical features on the AS and IR level. Finally, although one would expect Internet AS and IR level maps to exhibit similar scale-free properties, the different resolution in both kinds of maps might lead to a diversity of metrics properties.

In this chapter we present a detailed statistical analysis of large AS and IR level maps [102, 49, 69]. We study the scale-free properties of these maps, focusing on the degree and betweenness distributions. While scale-free properties are confirmed for maps at both levels, IR level maps show also the presence of an exponential cut-off, that can be related to constraints acting on the physical degree and load of routers. Power-law distributions with a cut-off are a general feature of scale-free phenomena in real finite systems and we discuss their origin in the framework of growing networks. At the AS level we confirm the presence of a strong scale-free character for the large-scale degree and betweenness distributions. We also discuss that deviations from the pure power-law behavior found in recent maps [49] at intermediate degrees has a marginal impact on the resilience and information spreading properties of the Internet [5, 37, 117].

Furthermore, we propose two metrics based on the degree and the clustering correlation functions, that appear to sharply characterize the hierarchical properties of Internet maps. In particular, these metrics clearly distinguish between the AS and IR levels, which show a very different behavior at this respect. While IR level maps appear to possess almost no hierarchical structure, AS maps fully exploit the hierarchy of domains around which the Internet revolves. The differences highlighted between the two levels might be very important in the developing of faithful Internet topology generators. The testing of Internet protocols working at different levels might need of topology generators accounting for the different properties observed. Hierarchical features are also important to scrutinize theoretical models proposing new dynamical growth mechanisms for the Internet as a whole. Finally, we extend the use of these metrics to discriminate between other real graphs that appear similar according to their degree distribution.

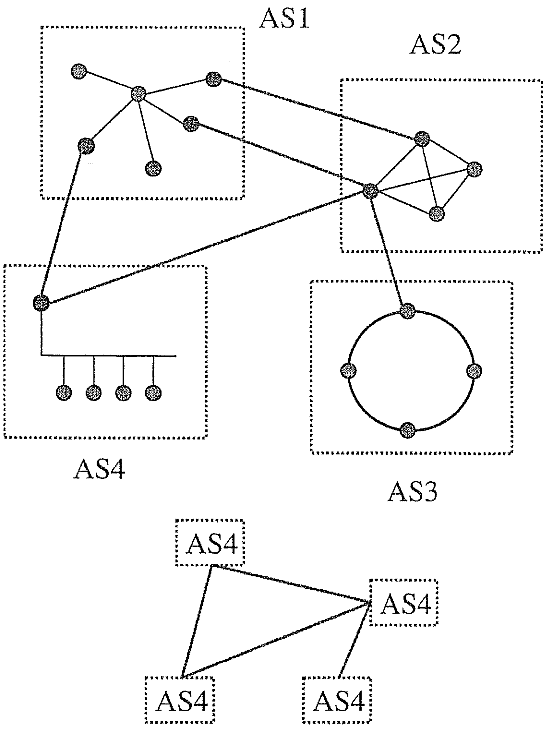


Figure 3.1: Schematic representation of the Internet maps at the Router (top) and Autonomous System (bottom) levels.

3.1 Internet maps

Nowadays the Internet can be partitioned in autonomously administered domains which vary in size, geographical extent, and function. Each domain may exercise traffic restrictions or preferences, and handle internal traffic according to particular autonomous policies. This fact has stimulated the separation of the inter-domain routing from the intra-domain routing, and the introduction of the Autonomous Systems Number (ASN). Each AS refers to one single administrative domain of the Internet. Within each AS, an Interior Gateway Protocol is used for routing purposes. Between ASs, an Exterior Gateway Protocol provides the inter-domain routing system. The Border Gateway Protocol (BGP) is the most widely used inter-domain protocol. In particular, it assigns a 16-bit ASN to identify, and refer to, each AS.

The Internet is usually portrayed as an undirected graph. Depending on the meaning assigned to the vertices and edges of the associated graph, we can obtain different levels of representation, each one corresponding to a different degree of coarse-graining respect to the physical Internet (see Fig. 3.1).

Internet Router level: In the IR level maps, vertices represents the routers, while edges represent the physical connections among them. In general, all mapping efforts at the IR level are based on computing router adjacencies from *traceroute* sequences sent to a list of networks in the Internet. The traceroute command performed from a single source provides a spanning tree from that source to every other (reachable) vertex in the network. By merging the information obtained from different sources it is possible to construct IR level maps of different portions of the Internet. In order to catch all the various cross-edges, however, a large number of source probes is needed. In addition, the instability of paths between routers and other technical problems—such as multiple alias interfaces—make the mapping a very difficult task. These difficulties have been diversely tackled by the different Internet mapping projects: the Lucent project at Bell Labs [86], the Cooperative Association for Internet Data Analysis [124], and the SCAN project at the Information Sciences Institute [69], that develop methods to obtain partial maps from a single source.

Autonomous System level: In the AS level graphs each vertex represents an AS, while each edge between two vertices represents the existence of a BGP peer connection among the corresponding ASs. It is important to stress that each AS groups many routers together and the traffic carried by a edge is the aggregation of all the individual end-host flows between the corresponding ASs. The AS map can be constructed by looking at the BGP routing tables. In fact, the BGP routing tables of each AS contains a spanning tree from that vertex to every other (reachable) AS. We can then try to reconstruct the complete AS map by merging the degree information coming from a certain fraction of these spanning trees. This method has been actually used by the National Laboratory for Applied Network Research (NLANR) [102], using the BGP routing tables collected at the Oregon route server, that gathers BGP-related information since 1997. Enriched maps can be obtained from some other public sources, such as Looking Glass sites and the Reseaux IP Europeens (RIPE) [35], getting about 40% of new AS-AS connections.

These graph representations do not model individual hosts, too numerous, and neglect edge properties such as bandwidth, actual data load, or geographical distance. For these reasons, the graph-like representation must be considered as an overlay of the basic topological structure: the skeleton of the Internet. Moreover, the data collected for the two levels are different, and both representations may be incomplete or partial to different degrees. In particular, measurements may not capture all the vertices present in the actual network and, more often, they do not include all the edges among vertices. It is not our purpose here to argue about the reliability of the different maps. However, the conclusions we shall present in this paper seem rather stable in time for the different maps. Hopefully, this fact means that, despite the different degrees of completeness, the present maps represent a fairly good statistical sampling of the Internet as a whole. In particular, we shall use the map collected during October/November 1999 by the SCAN project with the Mercator software as representative of the Internet router level. At the autonomous system level we consider the (AS) map collected at Oregon route server and the enriched (AS+) map (available at [49]), both dated May 25, 2001.

3.2 Average properties

We start our study by analyzing some standard metrics: the total number of vertices N and edges E, the vertex degree d_i, the minimum path distance between pairs of vertices l_{ij}, the clustering coefficient c_i, and the betweenness b_i. The degree d_i of a vertex is defined as the number of edges incident to that vertex, *i.e.* the number of connections of that vertex with other vertices in the network. If vertices i and j are connected we will say that they are nearest neighbors. The minimum path distance d_{ij} between a pair of vertices i and j is defined as the minimum number of vertices traversed by a path that goes from one vertex to the other. The clustering coefficient c_i [147] of the vertex i is defined as the ratio between the number of edges e_i in the sub-graph identified by its nearest neighbors and its maximum possible value $d_i(d_i-1)/2$, corresponding to a complete sub-graph, *i.e.* $c_i = 2e_i/d_i(d_i-1)$. This magnitude quantifies the tendency that two vertices connected to the same vertex are also connected to each other. The clustering coefficient c_i takes values of order $\mathcal{O}(1)$ for grid networks. On the other hand, for random graphs [50, 23], which are constructed by connecting vertices at random with a fixed probability p, the clustering coefficient is of order $\mathcal{O}(N^{-1})$. Finally, the betweenness b_i of a vertex i is defined as the total number of minimum paths that pass through that vertex. It gives an measure of the amount of traffic that goes through a vertex, if the minimum path distance is considered as the metric defining the optimal path between pairs of vertices. The average values of these metrics over every vertex (or pair of vertices for d_{ij}) in the AS, AS+, and IR maps is given in Table 3.1.

The average degree for the three maps is of order $\mathcal{O}(1)$; therefore, they can be considered as *sparse* graphs. Despite the small average degree, however, the average minimum path distance is also very small, compared to the size

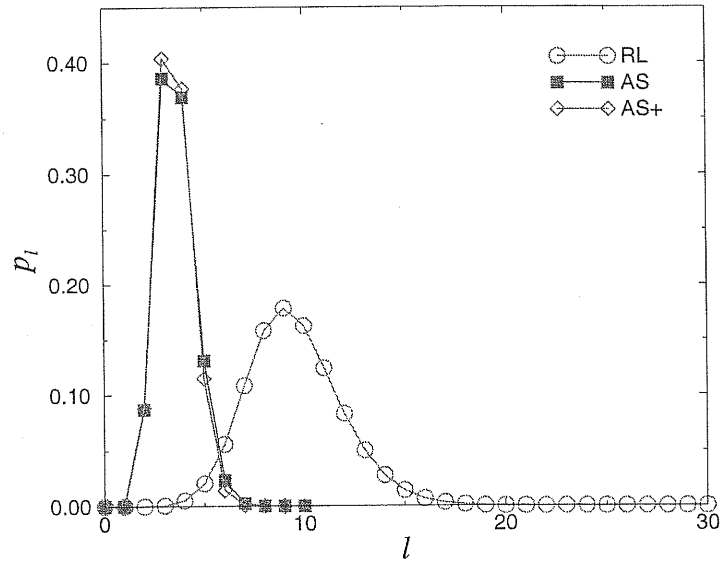


Figure 3.2: Probability distribution $p_l = \text{Prob}[l_{ij} = l]$ of the minimum path distance between vertices, for the AS, AS+, and IR maps.

of the maps. The probability distribution of the minimum path distance, $p_l = \text{Prob}[l_{ij} = l]$, is shown in Fig. 3.2. For all maps this distribution is sharply peaked around the average value $\langle l \rangle$; therefore, we can take $\langle l \rangle$ as the characteristic minimum path distance. In the next section we will show that this is not the case for the degree, that is characterized by large fluctuations from vertex to vertex. Thus, the Internet strikingly exhibits what is known as the "small-world" effect [147]: in average one can go from one vertex to any other in the system passing through a very small number of intermediate vertices. Since the network is sparse this necessarily implies that there are some hubs and backbones which connect different regional networks, strongly decreasing the value of $\langle l \rangle$. The small world evidence is strengthened by the empirical finding of clustering coefficients for the AS, AS+, and IR

Map	N	E	$\langle d \rangle$	$\langle l \rangle$	$\langle c \rangle$	$\langle b \rangle / N$
IR	228298	320105	2.80	9.51	0.03	4.14
AS	11174	23367	4.18	3.62	0.22	3.61
AS+	11461	32711	5.71	3.56	0.24	3.56

Table 3.1: Average metrics of the AS, AS+, and IR maps. See text for the metrics' definitions.

four orders of magnitude larger than the corresponding value for a random graph of the same size, $\mathcal{O}(N^{-1})$. As discussed above, this fact implies that neighbors of the same vertex are very likely on their turn connected among themselves. The high clustering coefficient of the Internet maps is probably due to geographical constraint. In Internet graphs, all edges are equivalent. Yet, the physical connections are characterized by a real space length. The larger is this length, the higher the cost of installation and maintenance of the physical line, favoring therefore the preferential connection between nearby vertices. It is likely that vertices within the same geographical region will have a large number of connections among them, increasing in this way the clustering coefficient.

Another measure of interest is given by the number of minimal paths that pass by each vertex. To go from one vertex in the network to another following the minimum path, a sequence of vertices is visited. If we do this for every pair of vertices in the network, there will be a certain number of key vertices that will be visited more often than others. Such vertices will be of great importance for the transmission of information along the network. This evidence can be quantitatively measured by means of the betweenness b_i; *i.e.* the number of minimum paths that go through each vertex i. This magnitude has been introduced in the analysis of social networks in Ref. [105, 106] and more recently it has been studied for the AS maps, with the name of load [58]. An algorithm to compute the betweenness has been described in Ref. [106]. For a star network the betweenness takes its maximum value $N(N-1)/2$ at the central vertex and its minimum value $N-1$ at the vertices of the star. The average betweenness of the AS, AS+, and IR maps analyzed here is $\mathcal{O}(N)$, as shown in Table 3.1. In the case of the AS and AS+ maps, despite the enriched map has a much larger number of edges, the average measures are very similar.

While some metrics are very alike (for instance, the average betweenness $\langle b \rangle$), some differences among others are consistent with the fact that the AS and AS+ maps are a coarse-grained representation of the IR map. The IR level map is, for instance, sparser, and its average minimum path distance is larger. The IR map has a small average degree, because routers have a finite capacity and, therefore, can have a limited number of connections. On the contrary, ASs can have in principle any number of connections, since they represent the aggregation of a large number of routers. This implies that AS maps have a greater number of vertices with a high number of connections (hubs), providing the shortcuts needed to produce a small average minimum path distance.

3.3 Scale-free properties

The analysis of the average measures presented in the previous section makes clear that the Internet does not resemble a star-shaped architectures with just a few gigantic hubs and a multitude of singly connected vertices. The same measurements rule out as well the possibility of a random graph structure or a regular grid architecture. These evidences suggest a peculiar topology that will be clearly identified by looking at the detailed distributions. In particular, Faloutsos *et al.* [53] pointed out for the first time that the degree properties of the Internet AS maps are characterized by a probability distribution that a vertex has d edges with the form $p_d \sim d^{-\gamma}$, where $\gamma \simeq 2.1$ is a characteristic exponent. This behavior signals the presence of *scale-free* degree properties; *i.e.* there is no characteristic degree above which the probability is decaying exponentially to zero. In other words, there is a statistically significant probability that a vertex has a very large number of connections compared to the average degree $\langle d \rangle$. In addition, the implicit divergence of $\langle d^2 \rangle$ is signalling the extreme heterogeneity of the degree pattern, since it implies that statistical fluctuations are unbounded. The work of Faloutsos *et al.* was followed by different studies of AS maps [36, 138, 137], AS+ maps [35], and IR maps [60, 29]. Here, we will revisit the analysis of scale-free properties in recent AS, AS+, and IR level maps.

We start by considering the integrated degree probability $P_d = \text{Prob}[d_i > d]$. In the case of a pure power-law probability distribution $p_d \sim d^{-\gamma}$, we expect the functional behavior $P_d = ad^{1-\gamma}$, where a is a normalization constant. In Fig. 3.3 we show the degree distribution for the AS, AS+, and IR maps. For the AS map a clear power law decay with exponent $\gamma = 2.1 \pm 0.1$ is observed, as it has been already reported elsewhere [53, 36, 138, 137]. The reported distribution is also stable in time as found by analyzing different time snapshot of the AS level maps obtained by the NLANR [138, 137]. As noted in Ref. [35], the degree distribution for the AS+ enriched data deviates from a pure power law at intermediate degrees. This anomaly might or might not be related to the biased enrichment of the Internet sampling (see Ref. [35]). While this represents an important point in the detailed description of the degree properties, it is not critical concerning the scale-free nature of the Internet. With respect to the network physical properties, it is just the large degree region that is actually effective. Indeed, recent studies about network resilience to removal of vertices [5, 37] and virus spreading [117] have shown that the relevant parameter is the ratio $\kappa = \langle d^2 \rangle / \langle k \rangle$ between the first two moments of the degree distribution. If $\kappa \gg 1$ then the network manifests some properties that are not observed for networks with exponentially bounded degree distributions. For instance, we can randomly

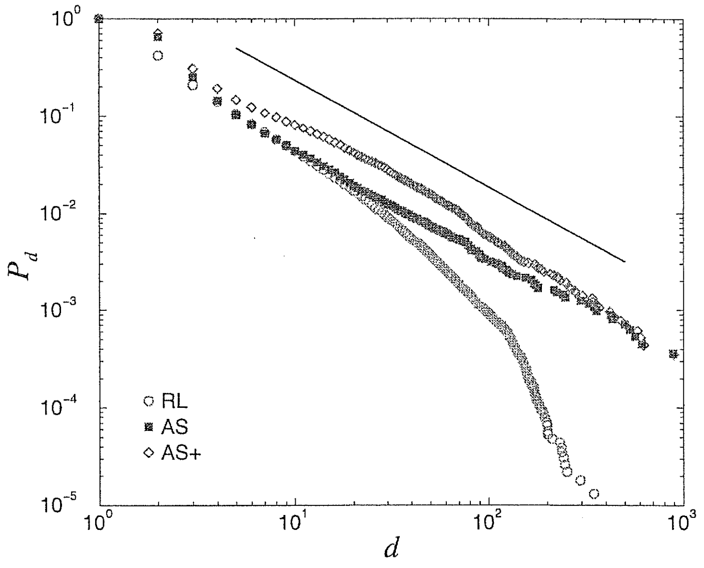


Figure 3.3: Integrated degree distribution P_d = Prob[$d_i > d$] for the AS, AS+, and IR maps. The solid line corresponds to a power law decay $p_d \sim d^{1-\gamma}$ with exponent $\gamma = 2.1$.

remove practically all the vertices in the network and a giant connected component [23] will still exist. In both the AS and AS+ maps, in fact, we observe a wide degree distribution, with the same dependency for very large d. The factor κ is mainly determined by the tail of the distribution, and is very similar for both maps. In particular, we estimate $\kappa = 265$ and $\kappa = 222$ for the AS and AS+ maps, respectively. With such a large values, for all practical purposes (resilience, virus spreading, traffic, etc.) the AS and AS+ maps behave almost identically.

The degree distribution of the IR level map has a power-law behavior that is, however, smoothed by a clear exponential cut-off. The existence of a power-law tendency for small degrees is better seen for the probability distribution p_d = Prob[$d_i = d$], as shown in Fig. 3.4. A power law fit of the form $p_d = a(1-\gamma)d^{-\gamma}$ for $d \leq 300$ yields the exponent $\gamma = 2.1 \pm 0.1$, in perfect agreement with the exponent found for the integrated degree distribution in the AS map. Nevertheless, for $d \gg 50$ the IR map integrated degree distribution follows a faster decay. This picture is consistent with a finite size scaling of the form $p_d = d^{-\gamma} f(d/d_c)$ [45]. Here d_c is a characteristic degree beyond which the distribution decays faster than a power law, and $f(x)$ has the asymptotic behavior $f(x)$ = const. for $x \ll 1$ and $f(x) \ll 1$

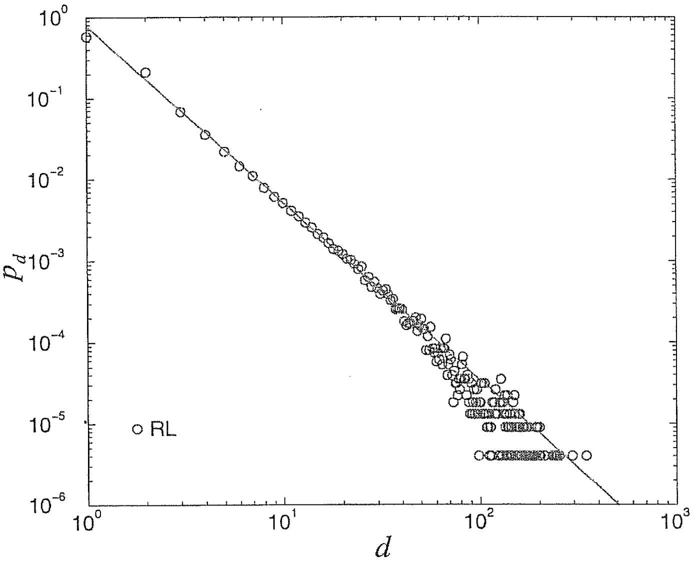


Figure 3.4: Degree distribution $p_d = \mathrm{Prob}[d_i = d]$ for the IR map. The solid line is a power law decay $p_d \sim d^{-\gamma}$ with $\gamma = 2.1$.

for $x \gg 1$. Deviations from the power law behavior at large degrees have been also observed for the larger maps reported in Ref. [29]. In that work, the integrated probability distribution is fitted to the Weibull distribution $P_d = a \exp[-(d/d_c)^\beta]$. While we do not want to enter into the details of the different fitting procedures, we suggest that the more general fitting form $p_d = d^{-\gamma} f(d/d_c)$, in which γ is an independent fitting parameter, is likely a better option.

The presence of truncated power laws must not be considered a surprise, since it finds a natural place in the context of scale-free phenomena. Actually, bounded scale-free distributions (*i.e.* power-law distributions with a cut-off) are implicitly present in every real world system because of finite-size effects or physical constraints. Truncated power laws are observed also in other real networks [6] and different mechanisms have been proposed to explain the cut-off for large degrees. Actually, we can distinguish two different kinds of cut-offs in real networks. The first is an exponential cut-off, $f(x) = \exp(-x)$, which can be explained in terms of a finite degree capacity of the network elements [6] or incomplete information [100]. This is likely what is happening at the IR level, where the finite capacity constraint (maximum number of router interfaces) is, in our opinion, the dominant mechanism affecting the tail of the degree distribution. In this perspective, larger and more recent

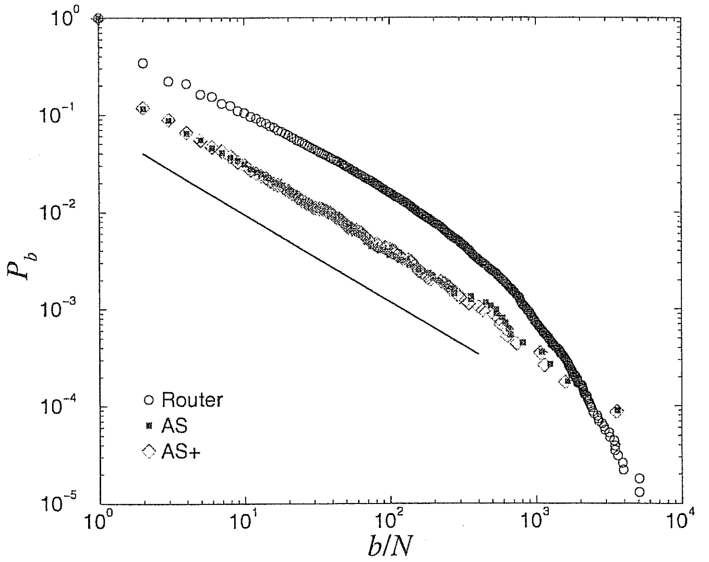


Figure 3.5: Integrated betweenness distribution P_b = Prob$[b_i > b]$ for the AS, AS+, and IR maps. The solid line is a power law decay $P_b \sim b^{1-\gamma_b}$ with $\gamma_b = 1.9$.

samples at the IR level could present a shift in the cut-off due to the improved technical router capabilities and the larger statistical sampling. A second possibility is given by a very steep cut-off such as $f(x) = \theta(1 - x)$, where $\theta(x)$ is the Heaviside step function. This is what happens in growing networks with a finite number of elements. Since SF networks are often dynamically growing networks, this case represents a network which has grown up to a finite number of vertices N. The maximum degree d_c of any vertex is related to the network age. The scale-free behavior is evident up the d_c and then decays as a step function since the network does not possess any vertex with degree d larger than d_c. By inspecting Fig. 3.3, this second possibility appears realized at the AS level. Indeed, the dominant mechanism at this level is the finite size of the network, while degree limits are not present, since each AS is a collection of a large number of routers, and it can handle a very large degree load.

The connection between finite capacity and bounded distributions becomes evident also if we consider the betweenness. This magnitude is a static estimate of the amount of traffic that a vertex supports. Hence, if a router has a bounded capacity, the betweenness distribution should also be bounded at large betweenness. On the contrary, this effect should be absent for the AS maps. The integrated betweenness distribution $P_b = \text{Prob}[b_i > b]$ for the AS, AS+, and IR maps is shown in Fig. 3.5. The AS and AS+ distributions are practically the same and they are well fitted by a power law $P_b \sim b^{1-\gamma_b}$ with an exponent $\gamma_b = 1.9 \pm 0.1$. In the case of the IR map, on the other hand, the betweenness distribution follows a truncated power law, in analogy to what is observed for the degree distribution. The betweenness distribution, therefore, corroborates the equivalence between the AS and AS+ maps, and the existence of truncated power laws for the IR map.

Finally, it is worth to stress that while the power law truncation is an expected feature of finite systems, the scale-free regime is the important signature of an emergent cooperative behavior in the Internet dynamical evolution. This dynamics play therefore a central role in the understanding and modeling of the Internet. In this persepective, the developing of a statistical mechanics approach to complex networks [3] is providing a new dynamical framework where the distinctive statistical regularities of the Internet can be understood in term of the basic processes ruling the appearance or disappearance of vertices and edges.

3.4 Hierarchy and correlations

The topological metrics analyzed so far give us a distinction between the AS and IR maps with respect to the large degree and betweenness properties. The difference becomes, however, more evident if we consider properties related with the existence of hierarchy and correlations. The primary known structural difference in the Internet is the distinction between *stub* and *transit* domains. Vertices in stub domains have edges that go only through the domain itself. Stub domains, on the other hand, are connected via a gateway vertex to transit domains that, on the contrary, are fairly well interconnected via many paths. This hierarchy can be schematically divided into international connections, national backbones, regional networks, and local area networks. Vertices providing access to international connections or national backbones are of course on top level of this hierarchy, since they make possible the communication between regional and local area networks. Moreover, in this way, a small average minimum path length can be achieved with a small average degree. This hierarchical structure will introduce some correlations in the network structure, and it is an important issue to understand how these features manifest at the topological level. In order to exploit the presence of hierarchies in Internet maps we introduce two metrics based on the clustering coefficient and the nearest neighbor average degree [138, 137].

The previously defined clustering coefficient is the average probability

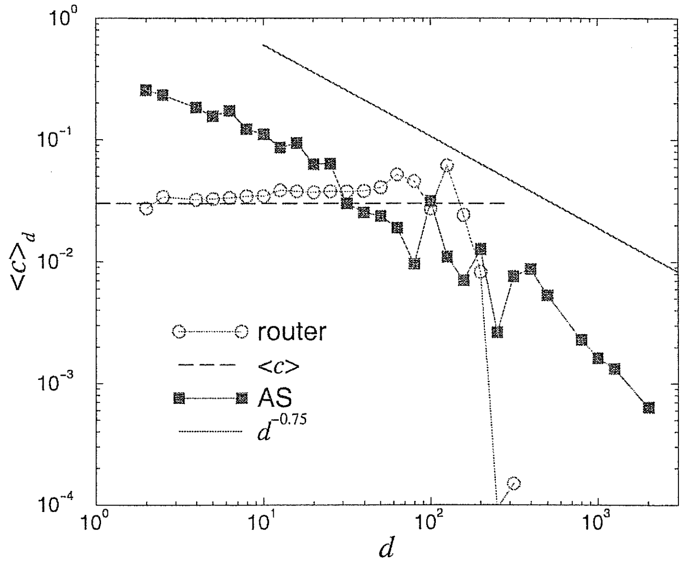


Figure 3.6: Average clustering coefficient as a function of the vertex degree for the AS, AS+, and IR maps. The solid line is given by the power law decay $\langle c \rangle_d \sim d^{-0.75}$. The horizontal dashed line marks the average clustering coefficient $\langle c \rangle = 0.03$ computed for the IR map.

that two neighbors l and m of a vertex i are connected. Let us consider the *adjacency matrix* a_{ij}, that indicates whether there is a connection between the vertices i and j ($a_{ij} = 1$), or the connection is absent ($a_{ij} = 0$). Given the definition of the clustering coefficient, it is easy to see that the number of edges in the subgraph identified by the nearest neighbors of the vertex i can be computed as $e_i = (1/2) \sum_{lm} a_{il} a_{lm} a_{mi}$. Therefore, the clustering coefficient c_i measures the existence of *correlations* in the adjacency matrix, weighted by the corresponding vertex degree. In section 3.2 we have shown that the clustering coefficient for the AS, AS+, and IR maps is four orders of magnitude larger than the one expected for a random graph and, therefore, that they are far from being random. Further information can be extracted if one computes the clustering coefficient as a function of the vertex degree [138, 137]. In Fig. 3.6 we plot the average clustering coefficient $\langle c \rangle_d$ for vertices with degree d. In the case of the AS and AS+ maps this quantity follows a similar trend that can be approximated by a power law decay with an exponent around 0.75. For the IR map, however, except for a sharp drop for large values of d, attributable to low statistics, it is almost constant, and equal to the average clustering coefficient $\langle c \rangle = 0.03$. This implies that, in the AS and AS+ maps, vertices with a small number of connections have larger local clustering coefficients than those with a large degree. This behavior is consistent with the picture described in the previous section of highly clustered regional networks sparsely interconnected by national backbones and international connections. The regional clusters of ASs are probably formed by a large number of vertices with small degree but large clustering coefficients. Moreover, they should also contain vertices with large degrees that are connected with the other regional clusters. These large degree vertices will be on their turn connected to vertices in different clusters which are not interconnected and, therefore, will have a small local clustering coefficient. On the contrary, in the IR level map these correlations are absent. Somehow the domain hierarchy does not produce any signature at the single router scale, where the geographic constraints and degree bounds probably play a more important role.

These observations for the clustering coefficient are supported by another metric related with the correlations between vertex degrees. These correlations are quantified by the probability $p(d' \mid d)$ that, given a vertex with degree d, it is connected to a vertex with degree d'. With the available data, a direct plot of $p(d' \mid d)$ results very noisy and difficult to interpret [59]. Thus in Ref. [138, 137] we suggested to measure instead the nearest neighbors

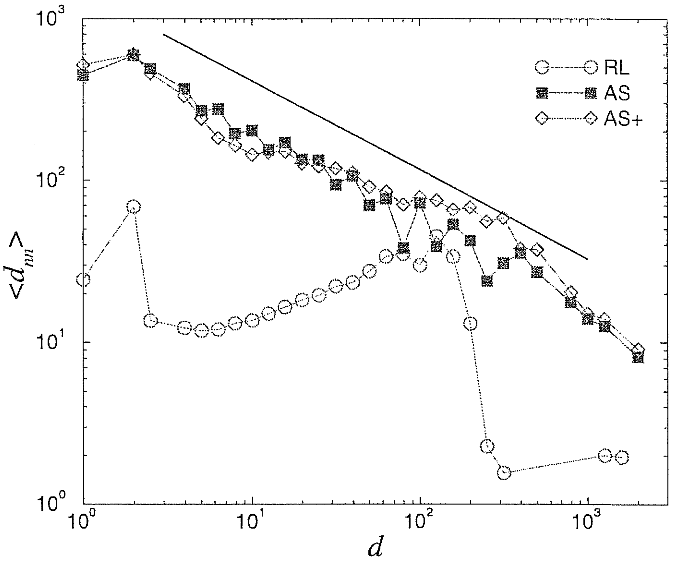


Figure 3.7: Nearest neighbors average degree for the AS, AS+, and IR maps. The solid line is given by the power law decay $\langle d_{nn}\rangle_d \sim d^{-0.55}$. The horizontal dashed line marks the value in the absence of correlations, $\langle d_{nn}\rangle_d^0 = \langle d^2\rangle \,/\, \langle d\rangle = 26.9$, computed for the IR map.

average degree of the vertices of degree d,

$$\langle d_{nn}\rangle_d = \sum_{d'} d'\, p(d' \mid d)$$

and to plot it as a function of the degree d. If there are no degree correlations (*i.e.* for a random graph), then $p^0(d' \mid d) = d' p_{d'} / \langle d\rangle$ and we obtain $\langle d_{nn}\rangle_d^0 = \langle d^2\rangle \,/\, \langle d\rangle$, which is independent of d. The corresponding plots for the AS, AS+, and IR maps are shown in Fig. 3.7. For the AS and AS+ maps we observe a power-law decay for more than two decades, with a characteristic exponent 0.55, clearly indicating the existence of correlations. On the contrary, the IR map displays again an almost constant nearest neighbors average degree, very similar to the expected value for a random graph with the same degree distribution, $\langle d_{nn}\rangle_d^0 \simeq 30$. Again, the sharp drop for large d can be attributed to the low statistics for such large degrees. Therefore, also in this case the two levels of representation show very different features.

It is worth remarking that the present analysis of the hierarchical and correlation properties shows a very good consistency of results in the case of the AS and AS+ maps. This points out a robustness of these features that can thus be considered as general properties at the AS level. On the other hand, the IR map shows a marked difference that must be accounted for when developing topology generators. In other words, Internet protocols working at different representation levels must be thought as working on different topologies. Topology generators as well must include these differences, depending on the level at which we intend to model the Internet topology.

3.5 Other real graphs

In a more general scope we extend the analysis made above for the Internet maps to the study of other real graphs In general, real networks are not uncorrelated and correlations may have different origins. Let us reconsider the example of the Internet. Due to installation costs, the Internet has been designed with a hierarchical structure. This hierarchy can be schematically divided in international connections, national backbones, regional networks, and local area networks. Vertices providing access to international connections or national backbones are of course on top level of this hierarchy, since they make possible the communication between regional and local area networks. Moreover, in this way, a small average minimum path distance can be achieved with a small average degree. This hierarchical structure will introduce some correlations in the network topology. For instance, it is expected that vertices with high degrees are connected to vertices with small degrees.

On the contrary, in social networks well connected people tends to be connected with well connected people [103]. Let us take the example of scientific co-authorship graph. A scientist writing a lot of papers have in general a larger probability to write a paper with another scientist who has also a lot of papers, than with one with a few papers. In fact, if F_i is the number of papers of scientist i and $F_i \ll N$ then the probability that two scientist write a paper together is roughly F_iF_j/N. Now, F_i is in general a monotonic increasing function of the scientist degree d_i (number of collaborators) and, therefore, scientists with a high degree will have a better chance to make a new article together, *i.e.* to be connected.

In Fig. 3.8 we plot $\langle c \rangle_d$ vs d for the different real networks. According to this measure, two different classes of graphs emerge. One in which $\langle c \rangle_d$ does not exhibit a strong dependency with d, except for finite size effects at the largest degrees. This behavior is typical of random graphs, where the probability that two neighbors of a vertex are connected by an edge is a constant, and equal two the probability that any two vertices selected at random are connected. On the contrary, there is another class where $\langle c \rangle_d$ follows an evident decay with increasing the vertex degree d. Thus, in this case, low degree vertices form local sub-graphs that are well connected. At

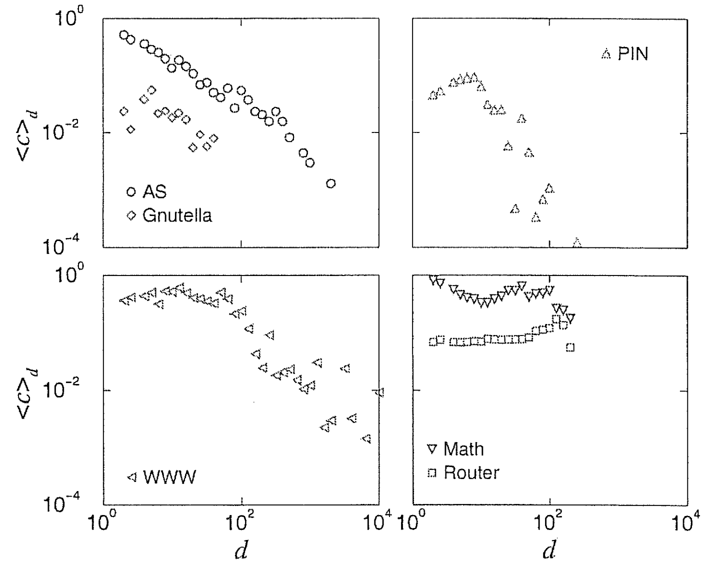


Figure 3.8: Clustering coefficient as a function of the vertex degree for some real graphs. AS and Router are the autonomous system [102] and router [69] level graph representations of the Internet, respectively. WWW a sub-graph of the WWW network, a data set collected by the Notre Dame group on Complex Networks (http://www.nd.edu/~networks). Gnutella is the Gnutella peer to peer network, provided by Clip2 Distributed Search Solutions. PIN if the protein-protein interaction graph of *Saccharomices Cerevisiae* as obtained from two hybrid experiments [70]. Math is the co-authorship graph obtained from all relevant journals in the field of mathematics and published in the period 1991-1998 [13].

the same time they are connected to other parts of the graph by high degree vertices, having a few edges between the subgraphs they connect but giving a small average minimum path distance. This picture makes evident the existence of some hierarchy [115, 138] or modularity [121].

In Fig. 3.9 we plot $\langle d_{nn} \rangle$ vs d for several real networks. Also in this case we found the emergence of two different classes of graphs. In one of them the average nearest neighbor degree exhibits a power law decay with increasing vertex degree. This is a strong evidence of the existence of disassortative (or negative) correlations, where large degree vertices tend to be connected with low degree ones and viseversa. On the other hand, for some of the graphs an increasing tendency is observed denoting the presence of assortative (or positive) correlations, where the edges connect vertices with similar degrees.

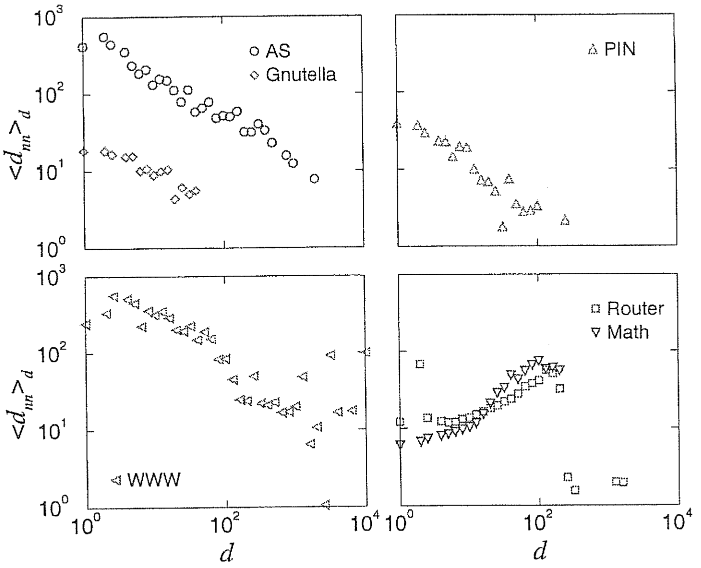


Figure 3.9: Average nearest-neighbors degree as a function of the vertex degree for some real graphs.

Notice, that the subdivision attending either the clustering coefficient or the average nearest-neighbor degree coincide.

3.6 Conclusions

The increasing availability of larger Internet maps and the proliferation of growing networks models with scale-free features have recently stimulated a more detailed statistical analysis aimed at the identification of distinctive metrics and features for the Internet topology. At this respect, in the present work we have presented a detailed statistical analysis of several metrics on Internet maps collected at the router and autonomous system levels. Our analysis confirms the presence of a power-law (scale-free) behavior for the degree distribution, as well as for the betweenness distribution, that can be associated to a measure of the load of the vertices in the maps. The exponential cut-offs observed in the IR maps, associated to the limited capacity of the routers, are absent in the AS level, which conglomerate a large number of routers and are thus able to bear a larger load. The analysis of the clustering coefficient and the nearest neighbors average degree show in a quantitative way the presence of strong correlations in the Internet degree at the AS level, correlations that can be related to the hierarchical distribution of this network. These correlations, on the other hand, seem to be nonexistent at the IR level. The correlation properties clearly indicate the presence of strong differences between the IR and AS levels of representation. Our findings represent a step forward in the characterization of the Internet topology, and will be helpful for scrutinizing more thoroughly the actual validity of the network models proposed so far, and as ingredient in the elaboration of new and more realistic Internet topology generators. A first step in this direction has been already given in the network model proposed in Ref. [59].

The conclusions obtained from the analysis of the Internet maps were extended to the study of other real graphs. It was corroborated that the metrics introduced here can be used to discriminate between different graphs that appear similar with respect to the degree distribution. They have been used for instance in Refs. [121, 89] to get a better understanding of the topology of some real networks.

Chapter 4

Growing networks with local rules

Once we know the existence of power law degree distributions, degree correlations and clustering hierarchy it is interesting to know their origin. It is known that growing network models with global evolution rules exhibit degree correlations. For instance, non-trivial degree correlations has been obtained in the linear preferential attachment models discussed in the second chapter [79] and in a growing network model without any preferential attachment [31]. However, the degree correlations obtained in these global models are not sufficiently strong to account for the features observed in real graphs. In this chapter we study different "microscopic" mechanisms that lead to graphs with the degree correlations and clustering hierarchy as observed in the previous chapter. The term "microscopic" means that we will investigate local evolution rules that involve a vertex and its neighbors. As it will be shown the preferential attachment, the inverse proportionality between the average clustering coefficient and the vertex degree, and degree correlations are common features of graph models build by local rules.

4.1 Random walk on a net

In this section we will study the evolution of a graph where we know about new vertices by simply exploring the graph, with applications to the citation and WWW graphs. We will focus on different "microscopic" mechanisms, where the term "microscopic" means that we will investigate local evolution rules that involve a vertex and its neighbors. A "macroscopic" approach based on effective attachment rates can be found in [80]. There are different ways to obtain information about the documents (articles, web pages)

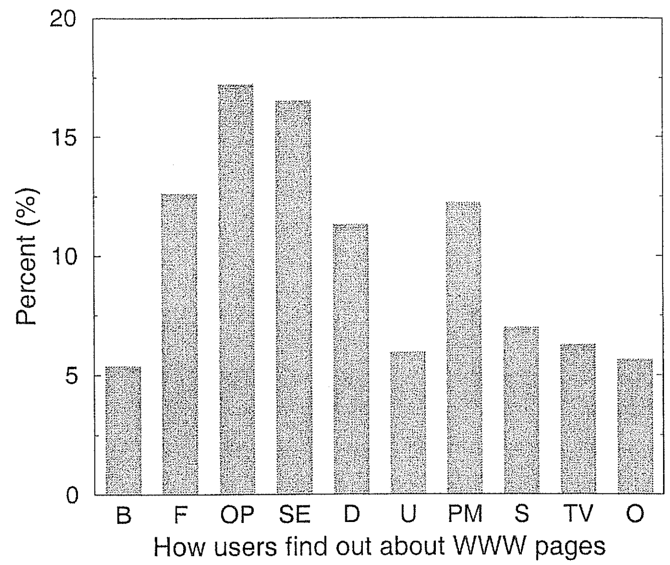


Figure 4.1: How users find about WWW pages according to the 1998 GVU's WWW Survey [128]. The different sources are: books (B), friends (F), other web pages (OP), search engines (SE), (D) directories, printed media (PM), (S) signatures at end of email messages, television advertisements (TV) and other (O). Notice that the higher percent is reached for other web pages (OP).

in these graphs, like looking at directories (citation index, web crawler), commercial spots, pointed by a friend, or following the references (citations, hyper-links) that are contained in the documents that we already know. In the case of the citation graph, we very often found new articles from the citation list of an article that we already know and, later on, we can repeat the process with these new articles. On the other hand, it is known that with a high probability people know about new web pages by surfing on the WWW.

In Fig. 4.1 we can see that the two major contribution to how people find out about new web pages are following the hyper-links of other web pages or using search engines. The first source can be characterized modeling the WWW "surfers" as random walks on the WWW graph. Let us assume that the walk starts from a page selected at random and, on each page, with probability q_e it decides to follow one link on that page or else to jump to another random page. Then, the probability v_i that a page i will be visited is given by

$$v_i = \frac{1-q_e}{N} + q_e \sum_j J_{ij} \frac{v_j}{d_j^{ou}}, \tag{4.1}$$

where J_{ij} is the adjacency matrix. It is quite interesting to notice that this probability of being visited by a random surfer is often used by search engines as a page rank criteria [64], as it is the case of the popular Google [27]. Hence, the two main sources through which new pages are visited are characterized by Eq. (4.1) and, therefore, the main properties of the in-degree distribution of the WWW graph should be computed starting on it. Moreover, when we visit new pages we in general do not create a hyper-link to it. In a first approximation this can be modeled introducing a probability q_v that a visited vertex (page) increases its in-degree by one (a hyper-link is created to it).

In a mean-field approximation one can replace the sum in Eq. (4.1) by Θd_i^{in}, resulting

$$v_i = \frac{1-q_e}{N} + q_e \Theta d_i^{in}. \tag{4.2}$$

where Θ is the average probability that a vertex pointing to vertex i is visited. To compute Θ we should take into account that the probability that a vertex i has an in-edge coming from a vertex with out-degree d^{ou} is $d^{ou} p_{d^{ou}} / \langle d^{ou} \rangle$. This edge will be selected at random among the d^{ou} out-edges and, therefore, with probability $1/d^{ou}$. Thus,

$$\Theta = \sum_{d^{ou}} \frac{d^{ou} p_{d^{ou}}}{\langle d^{ou} \rangle} \frac{1}{d^{ou}} v_{d^{ou}} = \frac{\langle v \rangle}{\langle d^{ou} \rangle}. \tag{4.3}$$

Moreover, when a walk is performed $\langle v \rangle N$ vertices are visited and, therefore, $q_v \langle v \rangle N$ edges are added in average, resulting

$$\langle d^{ou} \rangle = \langle d^{in} \rangle = q_v \langle v \rangle N \frac{\nu_s}{\nu_a}. \tag{4.4}$$

where ν_s and ν_a are the number of surfers and the number of newly added pages per unit time, respectively. Thus, from Eqs. (4.3) and (4.4) we finally obtain

$$\Theta = \frac{\nu_a}{q_v \nu_s N}. \tag{4.5}$$

On the other hand, the probability that the in-degree of a vertex of in-degree $d^{(in)}$ increases by one when a surfer walks on the graph is given by $A(d^{(in)}) = q_v v(d^{(in)})$ and, therefore, from Eqs. (4.2) and (4.5) it follows that

$$A(d^{(in)}) = \frac{1}{N} \left[q_v (1-q_e) + q_e \frac{\nu_a}{\nu_s} d^{(in)} \right]. \tag{4.6}$$

The degree distribution corresponding to this attachment rate can be easily obtained using the rate equation approach [81, 79]. Indeed, the number of vertices $n_{d^{in}}(t)$ with in-degree d^{in} satisfy the rate equations

$$\frac{\partial n_{d^{in}}}{\partial t} = \nu_s A_{d^{in}-1} n_{d^{in}-1} - \nu_s A_{d^{in}} n_{d^{in}} + \nu_a \delta_{d^{in}0}. \tag{4.7}$$

Now we should take into account that the number of vertices on the WWW graph grows exponentially and, in such a case, $\nu_a \propto N$. Moreover, assuming that each surfer has its own (or group of) web page (pages) then the number of surfers is expected to be proportional to the number of web pages, *i.e.* $\nu_s \propto N$. Thus,

$$\frac{\nu_s}{\nu_a} = \alpha, \tag{4.8}$$

where α is a constant. If this condition is satisfied then the in-degree distribution reaches a stationary state and we can write $n_{d^{in}}(t) = Np_{d^{in}}$, where $p_{d^{in}}$ is the stationary probability that a vertex has in-degree d^{in}. Substituting this expression in Eq. (4.7) we obtain

$$p_{d^{in}} = \frac{1}{1+a} \frac{\Gamma[a(\gamma-1)+d]}{\Gamma[a(\gamma-1)]} \frac{\Gamma[(1+a)(\gamma-1)+1]}{\Gamma[(1+a)(\gamma-1)+d+1]} \tag{4.9}$$

where

$$\gamma = 1 + \frac{1}{q_e}, \qquad a = \alpha q_v (1-q_e) \tag{4.10}$$

with the asymptotic behavior for large in-degree

$$p_{d^{in}} \sim (d^{in})^{-\gamma}. \tag{4.11}$$

Hence, the random walk model on a directed graph leads to a power law in-degree distribution, with an exponent $\gamma \geq 2$. Notice that the power law exponent does not depend on q_v and, therefore, we expect that generalizations of the rule of creating an edge to a visited vertex will not change this exponent. For instance, one can divide the vertices in classes in such a way that the edges can be only created among vertices of the same class, and the resulting power law exponent should be the same. Moreover, the power law exponent does not depend on α.

We can go beyond the in-degree distribution and compute the clustering coefficient as a function of the total degree $d = d^{in} + d^{ou}$ of a vertex. For this purpose we consider the graph as undirected and compute the number e_i of edges among the neighbors of a vertex i. Since the only dynamics in this model is given by the random walk it results that

$$\frac{\partial e_i}{\partial t} = q_v \left(q_e \Theta d_i^{in} + q_e v_i \right). \tag{4.12}$$

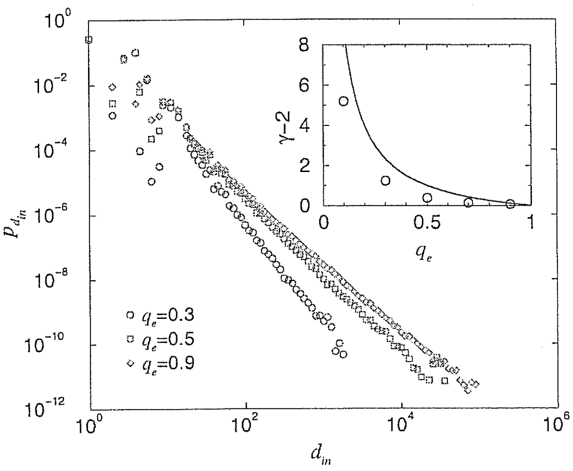


Figure 4.2: In-degree distribution of the random walk model for different values of the probability to continue the walk q_e and for graph size $N = 10^6$. In all cases we take average over 100 realizations. The inset shows the exponent γ obtained from the fit to the power law $p_{d^{in}} \sim (d^{in})^{-\gamma}$ (circles) together with the analytical prediction (continuous line).

The first term in the right hand side is the probability that a vertex with an out-edge to i is visited and the second the probability that vertex i is visited and the walk follows one of it out-edges to visit an out-neighbor vertex. In all cases the visited vertex is selected with probability q_v. Using Eqs. (4.2), (4.5) and taking into account that $\partial_t d_i^{in} = A(d_i^{in})$ we can rewrite (4.12) as

$$\frac{\partial e_i}{\partial t} \approx (1 + q_e)\frac{\partial d_i^{in}}{\partial t}, \tag{4.13}$$

Integrating this equation with the boundary condition $e(d^{in} = 0) = 0$ we obtain the clustering coefficient.

$$\langle c \rangle_d = \frac{2e(d)}{d(d-1)} = \frac{2(1+q_e)}{d} + \frac{2(1+q_e)(1-d^{ou})}{d(d-1)}, \tag{4.14}$$

Thus, for large d the clustering coefficient scale as

$$\langle c \rangle_d \approx \frac{2(1+q_e)}{d}. \tag{4.15}$$

Random walk model

We now study a particular random walk model by means of numerical simulations and compare its properties with the analytical results obtained above.

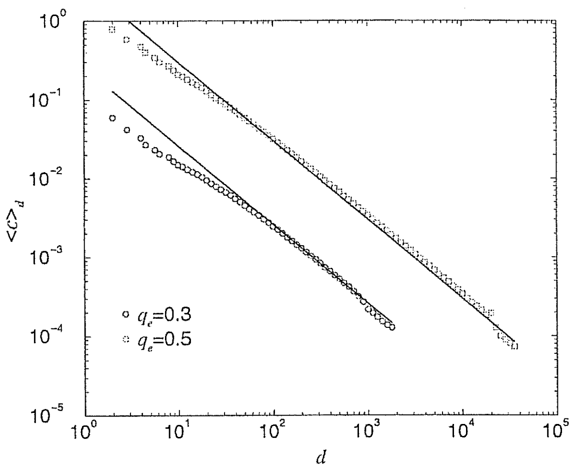


Figure 4.3: Clustering coefficient as a function of vertex degree of the random walk model, for different values of the probability to continue the walk q_e and for graph size $N = 10^6$. In all cases we take average over 100 realizations. The solid lines correspond with the power law decay $C(d) = 2(1 + q_e)/d$.

We have made some simplifications in order to reduce the number of parameters and investigate the influence of the most important parameter q_e. The model is defined as follows: *Initial condition*: starting with one vertex and an empty set of edges, iteratively perform the following rules,

- *Adding*: A new vertex is created with an edge pointing to one of the existing vertices, which is selected at random.

- *Walking*: if an edge is created to a vertex in the network then with probability q_e an edge is also created to one of its nearest neighbors. When no edge is created go to the *adding* rule.

The first simplification is that there is only one "surfer" in the network, *i.e.* $\nu_s = 1$. Second, each time the "surfer" decides not to follow one of the edges of the visited vertex it stops its search, and a new vertex starts a new search from a vertex selected at random. In other words the jump to a random vertex is coupled with the addition of new vertices resulting $\nu_a = 1 - q_e$. Finally, each time a vertex is visited an edge is created to it, thus $q_v = 1$. Hence, the in-degree distribution is given by Eq. (4.9) with

$$\gamma = 1 + \frac{1}{q_e}, \qquad a = 1. \tag{4.16}$$

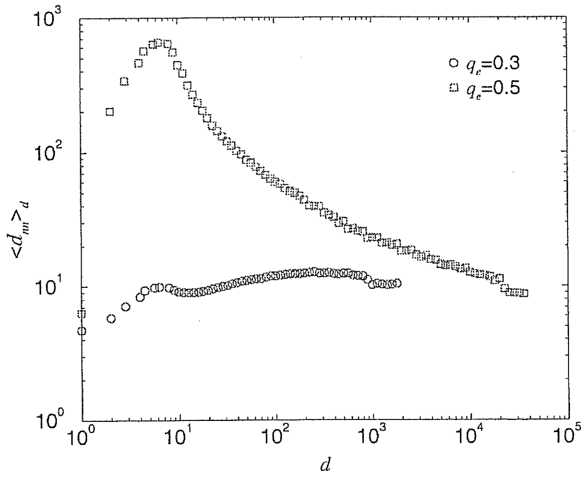


Figure 4.4: Average neighbor degree as a function of vertex degree of the random walk model, for different values of the probability to continue the walk q_e and for graph size $N = 10^6$. In all cases we take average over 100 realizations.

We have made numerical simulations of this random walk model up to graph sizes $N = 10^6$ making average over 100 realizations. In Fig. 4.2 we show a log-log plot of the in-degree distribution for different values of q_e. The power law decay for large degrees is evident. The exponent γ obtained from the fit to the numerical data is shown in the inset, together with the predicted dependency in Eq. (4.16). The analytical values overestimate the power law exponent but the qualitative picture is the same. For $q_e \to 0$ the power law exponent is so large that the degree distribution cannot be distinguished from an exponential. On the contrary, for $q \to 1$ it approaches is minimum value $\gamma = 2$. We attribute the quantitative disagreement to the mean-field approximation performed in the step from Eq. (4.1) to (4.2). On the other hand, the behavior of the average clustering coefficient with respect to the vertex degree is shown in Fig. 4.3. In this case the analytical asymptotic behavior in Eq. (4.15) is in very good agreement with the numerical data.

We were not able to obtain a prediction for the scaling of the average neighbor degree with the vertex degree. In this case our analysis relies on numerical simulations. In Fig. (4.4) we plot $\langle d_{nn} \rangle$ vs. d for two values of q_e. For $q_e = 0.3$ and for small values of q_e the average neighbor degree does not exhibit a strong dependency with d and, therefore, the graphs appear as uncorrelated. On the contrary, for $q_e = 0.5$ and in general for larger values of q_e it shows a peak around $d = 10$ and then decays with increasing degree.

This decay becomes even faster with increasing q_e. We have not found an explanation for this qualitative change of behavior yet. It is worth noticing that the experimental data for the WWW yield $\gamma \approx 2.1$, that can be obtained in our model for some $q_e > 0.5$, resulting in negative correlations as in the real data (see Fig. 3.9).

Recursive search model

In the random walk one follows only one edge of the visited vertices. However, one may consider an exhaustive search following all the edges recursively. The main idea of a recursive search is thus to be connected to one vertex of the network and any time we get in contact with a new vertex we follow all its edges, exploring in this way a larger part of the network. This can be modeled modifying the walking rule as follows,

- *Walking*: if an edge is created to a vertex in the network then with probability q_e an edge is also created to each of its nearest neighbors. When no edge is created go to the *adding* rule.

As for the previous model we have $\nu_s = 1$, $\nu_a = 1 - q_e$ but $A(d^{in})$ is not given by Eq. (4.6).

$q_e = 0$: In this case only the *adding* rule is performed, hence $A(d^{in}) = 1/N$ independent of d^{in}. The fact that $A(d^{in})$ scales as N^{-1} carries as a consequence that $n_{d^{in}}(N) = Np_{d^{in}}$ is the stationary solution of Eq. (4.7), where $p_{d^{in}}$ is the stationary probability to find a vertex with in-degree d^{in}. Substituting this expression in Eq. (4.7) one obtains

$$p_{d^{in}} = 2^{-(d^{in}+1)}. \tag{4.17}$$

$q_e = 1$: Also for this limiting case the in-degree distribution can be computed exactly. Let us determine $A(d^{in})$ using the following fact. Any vertex i with in-degree d_i^{in} has d_i^{in} vertices with an edge to it, which will be denoted by x_j $(j = 1, 2, \ldots, d_i^{in})$. At the same time each of these x_j vertices may have other vertices with an edge to it. The following result holds: any vertex with an edge to any of the vertices x_j has also an edge to i. The proof is straightforward, if when a vertex is added it creates an edge to any of the vertices x_j then with probability $q_e = 1$ it creates an edge to all the nearest neighbors of x_j, among which vertex i is contained; end of the proof. Hence, the probability that when a vertex is added it creates an edge to vertex i is just the probability $(1 + d_i^{in})/N$ that the first edge is connected to i or to any of the d_i^{in} vertices with an edge to i, *i.e.* $A(d^{in}) = (1 + d^{in})/N$. As for

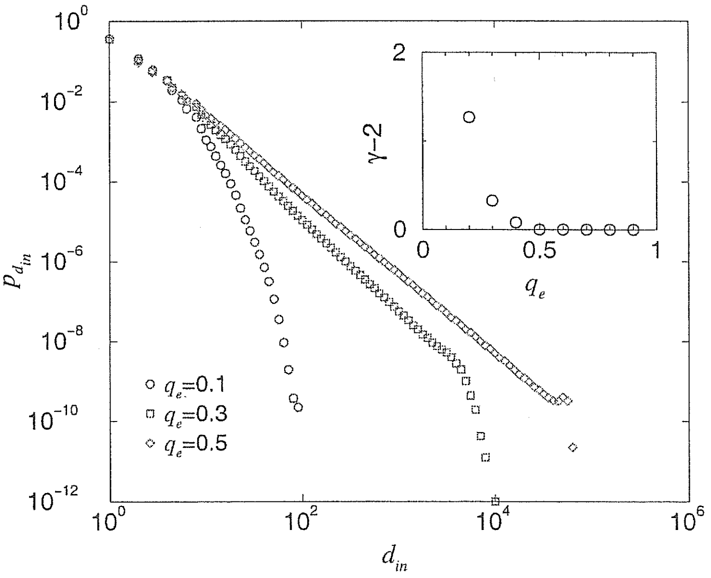


Figure 4.5: Log-log plot of the in-degree distribution of the recursive search model for different values of q_e. The inset shows the exponent γ obtained from the power law fit $p_{d^{in}} \sim (d_{in})^{-\gamma}$ to the numerical data.

$q_e = 0$ $A(d^{in})$ scales as $1/N$ and, therefore, the stationary solution is of the form $n_{d^{in}}(N) = Np_{d^{in}}$. Then from Eq. (4.7) it follows that

$$p_{d^{in}} = \frac{1}{(d^{in}+1)(d^{in}+2)}. \tag{4.18}$$

Notice that also in this case, although it is not implicitly assumed, there is a preferential attachment.

The limiting cases $q_e = 0$ and $q_e = 1$ are described by in-degree distributions which are qualitative different. For $q_e = 0$ the distribution is exponential with a finite average in-degree. On the contrary, for $q_e = 1$, the distribution follows a power law decay $p_{d^{in}} \sim {d^{in}}^{-\gamma}$ for large d^{in}, with $\gamma = 2$. This power law decay goes up to the largest possible degree $d^{in} \sim N^{1/(\gamma-1)} \sim N$ while $p_{d^{in}} = 0$ for $d^{in} \geq N$. Hence, for $q_e = 1$ and large N the average in-degree scale as

$$\langle d^{in}\rangle(N) = \langle d^{ou}\rangle(N) = a + \ln N, \tag{4.19}$$

where a is independent of N and clearly $\langle d^{in}\rangle$ diverges in the thermodynamic (large network sizes) limit. In a mean-field approximation one can neglect the existence of loops in the network and, in such a case, the "walking" rule will take place on a tree. Each vertex on the tree will have in average $\langle d^{ou}\rangle(N)$ sons, which is just the average out-degree after N vertices have been added. Moreover, if a vertex is visited then each of its sons will be visited with probability q_e. Hence, when the vertex $N+1$ is added, its average out-degree $\langle d^{ou}\rangle(N+1)$ will be given by the average number of vertices visited during the walk, *i.e.*

$$\langle d^{ou}\rangle(N+1) = 1+q_e\langle d^{ou}\rangle(N)+[q_e\langle d^{ou}\rangle(N)]^2+\ldots = \frac{1}{1-q_e\langle d^{ou}\rangle(N)}. \quad (4.20)$$

If there is a stationary state then $\langle d^{ou}\rangle(N+1) = \langle d^{ou}\rangle(N) = \langle d^{ou}\rangle$. In this case Eq. (4.20) yields two solutions. One of them diverges when $q_e \to 0$, which is not admissible since $\langle d^{ou}\rangle = 1$ for $q_e = 0$. The other solution is

$$\langle d^{ou}\rangle = \langle d^{in}\rangle = \frac{1-\sqrt{1-4q_e}}{2q_e}. \quad (4.21)$$

This solution is valid for $q_e \leq q_c = 1/4$ and, therefore, the average out degree does not converge to an stationary value when $q_e > q_c$. In this last region the average out degree will increases logarithmically with N, as in the extreme case $q_e = 1$ (see Eq. (4.19)). Now, since $\langle d^{in}\rangle = \langle d^{ou}\rangle$ and $\langle d^{in}\rangle$ approaches a stationary state for any $\gamma > 2$ and diverges otherwise, we expect that the in-degree distribution has a power law exponent $\gamma > 2$ for $q_e < q_c$ and $\gamma \leq 2$ for $q_e > q_c$. Moreover, taking into account that the fastest divergence is obtained for $q_e = 1$, where $\gamma = 2$, we conclude that for $q_e > q_c$ the power law exponent is constant and equal to $\gamma = 2$.

To investigate the behavior for $0 < q_e < 1$ and the existence of a non trivial threshold q_c as predicted by the mean-field approach, we have made numerical simulations of the recursive search model for different values of q_e up to graph sizes $N = 10^5$. For each value of q_e the in-degree distribution was averaged over 100 runs of the algorithm. The resulting in-degree distribution is shown in Fig. (4.5). For $q_e = 1$ the decay for large in-degrees is very fast, and can be fitted by a power law decay with a very large exponent or equivalently by an exponential decay. On the contrary, for larger q_e the exponent becomes smaller and the power law behavior becomes more evident. Finally, for $q_e \geq q_c = 0.5 \pm 0.1$ the exponent becomes independent of q_e and equals $\gamma = 2$, in agreement with the mean-field prediction. However, the numerical threshold is two times the value obtained from Eq. (4.21).

In ordinary critical phenomena the absence of any typical length scale takes place at the critical point, which is observed at a precise value of the order parameter. For the present model, however, the absence of a characteristic in-degree is not only manifested at a precise value of q_e but in the whole interval $q_c \leq q_e \leq 1$. These features are very similar to those observed in some sandpile models [129, 139], the paradigm of self-organized critical systems [7, 8]. As in these models[141, 142], there is a time scale separation between the addition of new vertices and their "walk" through the network. In the thermodynamic limit ($N \to \infty$) the phase diagram of the model is divided in a sub-critical ($0 \leq q_e < q_c$) and a critical region ($q_c \leq q_e \leq 1.$), where the power law exponent does not depend on the control parameter. Hence, the results presented here suggest that for $q_c \leq q_e \leq 1$ the present model is in a self-organized critical state.

4.2 Connecting nearest-neighbors

In social graphs it is more probable that two vertices with a common neighbor get connected than two vertices chosen at random [104]. Clearly this property leads to large clustering coefficients since it increases the number of connections between the neighbors of a vertex, as it has been already observed in a model proposed by Davidsen, Ebel and Bornholdt [39]. The basic assumption of their model is that the evolution of social connections is mainly determined by the creation of new relations between pairs of individuals with a common friend.

In this context we introduce the concept of potential edge, as the potential connection between two disconnected vertices with common neighbors. Moreover, the graph dynamics will be defined by the transition rates between the three possible states of a pair of vertices: disconnected (s), connected by a potential edge (p) or by an edge (e). Let d_i^* be the number of potential edges incident to vertex i, potential degree to abbreviate. We can write the rate equations for the evolution of the number of vertices with degree d and potential degree d^* however we have found problems in solving them. Instead we will use the continuum approach [11, 46]. In this case we neglect fluctuations and write mean-field equations for the evolution of d_i and d_i^*,

$$\begin{aligned} \frac{\partial d_i}{\partial N} &= \nu_{s\to e}(N - d_i - d_i^*) + \nu_{p\to e} d_i^* - (\nu_{e\to s} + \nu_{e\to p}) d_i, \\ \frac{\partial d_i^*}{\partial N} &= \nu_{s\to p}(N - d_i - d_i^*) + \nu_{e\to p} d_i - (\nu_{p\to s} + \nu_{p\to e}) d_i^*, \end{aligned} \tag{4.22}$$

where $\nu_{x\to y}$ is the transition rate from state x to state y per unit of N, and $N - d_i - d_i^*$ is just the number of remaining nodes, that are not connected by a potential edge nor by an edge to node i. The creation (deletion) of a potential edge incident to a vertex is associated with the creation (deletion) of an edge incident to one of its neighbors and, therefore,

$$\begin{aligned} \nu_{s\to p} &= \nu_{s\to e} d_i, \\ \nu_{p\to s} &= \nu_{e\to s} d_i. \end{aligned} \tag{4.23}$$

In the following we will neglect any process where an edge is deleted, *i.e.*

$$\nu_{e\to s} = \nu_{e\to p} = 0. \tag{4.24}$$

This assumption may seem to crude for social networks where it is known that social relations can be lost but it is realistic in many cases. For instance, in the network of scientific collaborations two scientists are said to be connected if they have co-authored a paper. It is clear that this connection cannot be lost in time because the fact that they have written a paper together cannot be changed. In general, if the connection between two vertices is given by the occurrence of certain event (co-authoring a paper, being in the cast of a the same film, having a sexual relation) in the past history then this connection cannot be lost and, therefore, our approximation holds. Another crucial assumption is related to the fact that the transition from potential edge to an edge has a higher probability of occurrence than the transition from disconnected to an edge. In fact, the connection of two disconnected vertices without a common neighbor is a process that models the creation of a social relation between two social entities chosen at random. We thus assume

$$\nu_{s\to e} = \frac{\mu_0}{N^2}. \tag{4.25}$$

On the other hand, the creation of an edge between two vertices with a common neighbor, that is with a potential edge between them, models the creation of a social relation between two "friends" of a social entity. In this case we assume

$$\nu_{p\to e} = \frac{\mu_1}{N}. \tag{4.26}$$

Under these approximations the system of equations (4.22) is reduced to

$$\begin{aligned} N\frac{\partial d_i}{\partial N} &= \mu_0 + \mu_1 d_i^*, \\ N\frac{\partial d_i^*}{\partial N} &= \mu_0 d_i - \mu_1 d_i^*, \end{aligned} \tag{4.27}$$

This system of differential equations is linear and, therefore, can be easily integrated resulting that, for $N \gg N_i$

$$d_i(N) = d_0 \left(\frac{N}{N_i}\right)^{\beta}, \quad d_i^*(N) = d_0^* \left(\frac{N}{N_i}\right)^{\beta}, \tag{4.28}$$

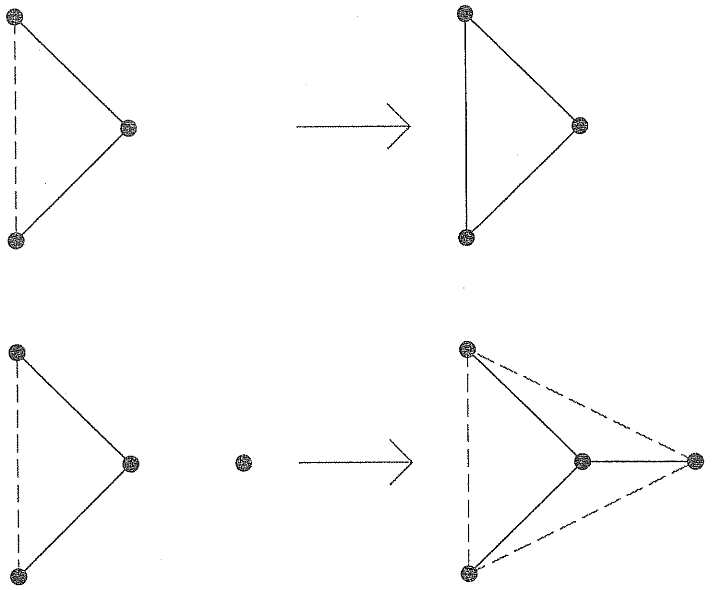


Figure 4.6: Schematic representation of the two evolution rules of the connecting nearest-neighbors model. Top: with probability u a potential edge (dashed line) becomes an edge (continuum lines). Bottom: with probability $1-u$ a new vertex is added to the graph (disconnected vertex in the left), then it is connected with an edge to a vertex selected at random and by potential edges to its neighbors (right).

where N_i is the size of the graph when vertex i was added to it and

$$\beta = \frac{\mu}{2}\left(-1+\sqrt{1+4\frac{\mu_0}{\mu_1}}\right) . \tag{4.29}$$

Now, if the vertices are added at a constant rate then

$$\begin{aligned} P(d_i > d) &= P\left[d_0\left(\frac{N}{N_i}\right)^{\beta} > d\right] \\ &= \int_0^N \frac{dN_i}{N}\Theta\left[d_0\left(\frac{N}{N_i}\right)^{\beta} - d\right] , \end{aligned} \tag{4.30}$$

and, therefore,

$$p_d = \frac{\partial P(d_i > d)}{\partial d} \sim d^{-\gamma}, \tag{4.31}$$

with

$$\gamma = 1 + \frac{1}{\beta}. \tag{4.32}$$

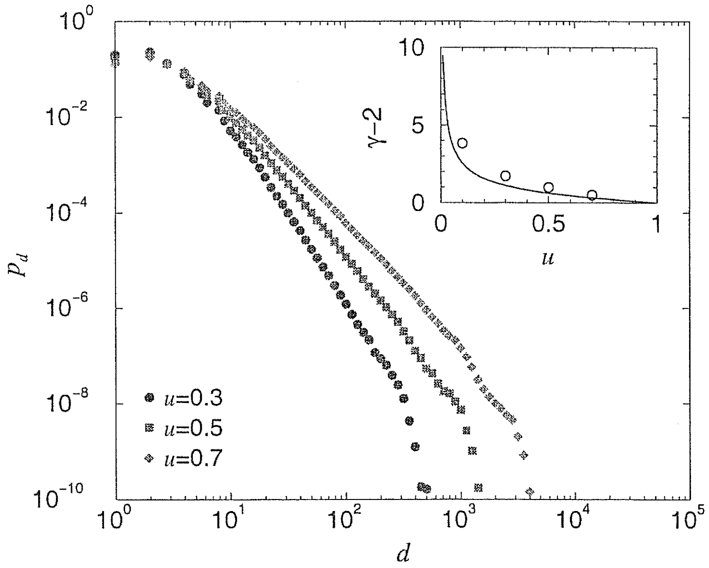


Figure 4.7: Degree distribution of the connecting nearest neighbors model for different values of the addition rate u, graph size $N = 10^6$ and average over 100 realizations. The inset shows the exponent γ obtained from the fit to the power law $p_d = ad^{-\gamma}$ (circles) together with the analytical prediction (continuous line).

Notice that the main ingredient leading to this power law behavior is given by Eq. (4.23). On the contrary, if $\nu_{s\to p}$ would be independent of the vertex degree an exponential decay would be obtained.

We can also compute the clustering coefficient as a function of the vertex degree. The main contribution to the evolution of e_i, the number of edges among the neighbors of vertex i, is given by the transition *potential edge* $\longrightarrow$ *edge*. In fact, if the potential edge connecting a vertex i to another vertex j, with common neighbor k, becomes an edge then vertex i gain one neighbor (vertex j) and a new edge among its neighbors (that connecting j and k). Neglecting other contributions we have

$$\frac{\partial e_i}{\partial N} = \nu_{p\to e} d_i^* = \mu_1 \frac{d_i^*}{N}. \tag{4.33}$$

Integrating this equation using Eq. (4.28) it results that

$$\langle c \rangle_d = \frac{2e(d)}{d(d-1)} \approx \frac{2\mu_1}{d}. \tag{4.34}$$

Thus, once again we obtain the inverse proportionality between $\langle c \rangle_d$ and

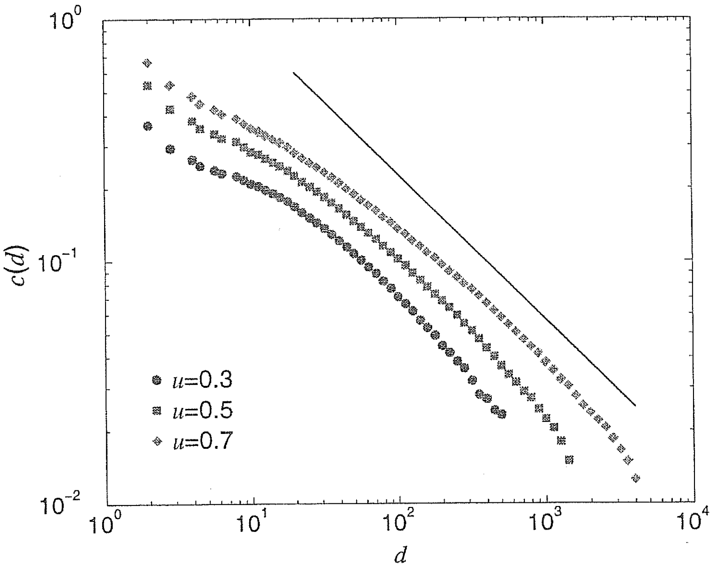


Figure 4.8: Clustering coefficient as a function of vertex degree of the connecting nearest neighbors model for different values of the addition rate u, graph size $N = 10^6$ and average over 100 realizations. The solid line is a power law decay with exponent 0.6.

vertex degree d, in this case due to the conversion of potential edges between vertices with a common neighbor into edges.

Connecting nearest-neighbors model

To check these results we have made numerical simulations of a variant of the model proposed by Davidsen, Ebel and Bornholdt [39], defined as follows: Starting with a single vertex and an empty set of edges iteratively perform the following rules,

- With probability $1 - u$ introduce a new vertex in the graph, create an edge from the new vertex to a node j selected at random, (implying the creation of a potential edge between the new vertex and all the neighbors of j).
- With probability u convert one potential edge selected at random into an edge.

A schematic representation of this rules is shown in Fig. 4.6. Actually, in the Davidsen, Ebel and Bornholdt [39] model the number of vertices is fixed and each time a new vertex is added one vertex is removed from the graph. We consider the growing variant because in this case is easier to determine some properties analytically. For very large N we expect that both variants have the same qualitative behavior.

These evolution rules fit into the equations written above after setting

$$\mu_0 = 1, \qquad \mu_1 = \frac{u}{1-u}. \tag{4.35}$$

Thus, from Eqs. (4.29) and(4.32) it follows that

$$\gamma(u) = 1 + \frac{2(1-u)}{u}\left(-1+\sqrt{1+4\frac{1-u}{u}}\right)^{-1}, \tag{4.36}$$

with the limiting cases

$$\gamma(0) = \infty, \qquad \gamma(1) = 2. \tag{4.37}$$

Thus, the power law exponent γ takes its minimum value when $u \to 1$ corresponding to a low rate of addition of vertices and it grows with decreasing u corresponding to higher rates of vertex addition. In Fig. 4.7 we plot the degree distribution as obtained from numerical simulations. For intermediate degrees it exhibits a power law decay $p_d \sim d^{-\gamma}$. The value of γ obtained from the fit to the numerical data is shown in the inset, together with the analytic curve given by Eq. (4.36). In the region $u \to 0$ where the graph is quite sparse ($\langle d \rangle \sim 1$) there is a good quantitative agreement between theory and simulations. However, when u gets closer to 1 deviations are observed. In this last region the exponent $\gamma \approx 2$, yielding large fluctuations in the vertex degrees. These fluctuations were nevertheless neglected in the continuum approach.

In Fig. 4.8 we plot the clustering coefficient as a function of the vertex degree. It follows a power law decay for large degrees but with an exponent smaller than 1. Thus, also in this case we found a disagreement between the continuum approach predictions and numerical simulations. On the other hand, the average neighbor degree as a function of the vertex degree is shown in Fig. 4.9. It increases with increasing d, *i.e.* the graphs generated using this model exhibit positive degree correlations. This result is in very good agreement with the observations made for social graphs that are also characterized by positive degree correlations. Hence, the connecting nearest-neighbors mechanism generates many of the topological properties of social networks.

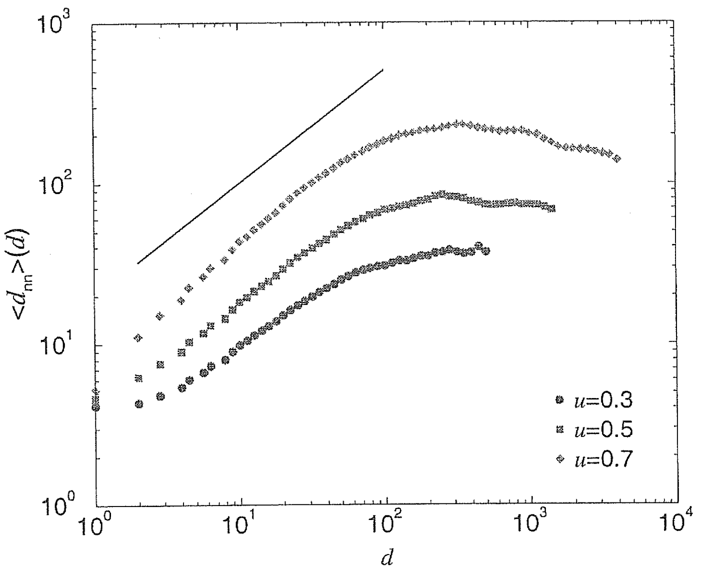


Figure 4.9: Average degree among the neighbors of a vertex with degree d of the connecting nearest neighbors model for different values of the addition rate u, graph size $N = 10^6$ and average over 100 realizations. The solid line is a power law grow with exponent 0.6.

4.3 Duplication-divergence

The evolution of some real graphs is given by a replication or partial replication of its local structure. An example is the genome that evolves, among other mechanisms, through single gene or full genome duplications [113] and mutations that lead to the differenciation of the duplicate genes. The evolution of the genome can be translated into the evolution of the protein-protein interaction network where each vertex represents the protein expressed by a gene, as we will see in the sixth chapter. After gene duplication both the expressed proteins will have the same interactions. This corresponds to the addition of a new vertex in the network with edges pointing to the neighbors of its ancestor. In addition positive and negative mutations can be modeled by the creation and lost, respectively, of the edges leading to the divergence of the duplicates. Another example is the WWW where new web pages may be created making a copy or a partial copy of the hyperlinks present in other web pages [78]. In this case the duplication models the copy process and the divergence the deletion or addition of hyperlinks in the duplicated pages.

In a first approximation we will assume that the processes of duplication and divergence are not coupled but take place independently one of the other.

Moreover, we will also assume that the creation and deletion of edges take place at random and that they are independent of the degree of the vertices at the edges ends, or any other topological property. Under these approximations, the evolution of the degree of a vertex (the number of interacting partners) is given by

$$\frac{\partial d_i}{\partial N} = \frac{\nu_d}{N} d_i + \nu_c (N - d_i) - \nu_l d_i, \tag{4.38}$$

where ν_d, ν_c, and ν_l are the rates per unit of vertex added of duplications, edge creation and edge lost, respectively. By definition, each duplication implies the addition of a new node and, therefore,

$$\nu_d = 1. \tag{4.39}$$

We will further assume that

$$\nu_c = \frac{\mu_0}{N}, \qquad \nu_l = \frac{\mu_1}{N} \tag{4.40}$$

otherwise the stationary graph will be empty or fully connected, both being unreal. Then, substituting Eqs. (4.39) and (4.40) into Eq. (4.38) we obtain

$$N \frac{\partial d_i}{\partial N} = \mu_0 + (1 - \mu_1) d_i. \tag{4.41}$$

The integration of this equation yields

$$d_i(N) = \left(d_i(N_i) + \frac{\mu_0}{1-\mu} \right) \left(\frac{N}{N_i} \right)^{\beta} - \frac{\mu_0}{1-\mu}, \tag{4.42}$$

where N_i and $d_i(N_i)$ are the graph size and degree of vertex i when vertex i was added to the graph, and

$$\beta = 1 - \mu_1. \tag{4.43}$$

Here we have implicitly assumed that

$$\mu_1 < 1, \tag{4.44}$$

otherwise the stationary state will be an empty graph.

From Eq. (4.42) it follows that

$$P(d_i > d) = P\left[\left(d_i(N_i) + \frac{\mu_0}{1-\mu} \right) \left(\frac{N}{N_i} \right)^{\beta} - \frac{\mu_0}{1-\mu} > d \right]. \tag{4.45}$$

This probability should be computed taking into account that both N_i and $d_i(N_i)$ are random variables. If the duplications take place at a constant rate then the probability density of N_i is given by $p(N_i) = 1/N$ independent of N_i. Moreover, the probability that a node has degree $d_i(N_i)$ when it is introduced is just the probability that its ancestor has this degree. If the graph is in a stationary state then $P[d_i(N_i) = d] = p_d$, is just the degree distribution. Hence

$$P(d_i > d) = \sum_{d'} p_{d'} \int_1^N \frac{dN_i}{N} \Theta \left[\left(d' + \frac{\mu_0}{1-\mu} \right) \left(\frac{N}{N_i} \right)^\beta - \frac{\mu_0}{1-\mu} > d \right]. \tag{4.46}$$

For $N \gg 1$ we finally obtain

$$p_d = \frac{\partial P(d_i > d)}{\partial d} \sim \left(\frac{\mu_0}{1-\mu} + d \right)^{-\gamma}, \tag{4.47}$$

with

$$\gamma = 1 + \frac{1}{1-\mu_1}. \tag{4.48}$$

The origin of this power law degree distribution is determined by the second term in the right hand side of Eq. (4.41), associated with the vertex duplications and subsequent edge lost. These are local mechanisms and, as in the models describe before, they lead to an effective preferential attachment manifested as a power law degree distribution.

The next step is thus to investigate if the duplication-divergence model satisfies the inverse proportionality between the average clustering coefficient and vertex degree. If the creation of new interactions takes place at random, *i.e.* they appear between randomly chosen vertices, then the average clustering coefficient will be negligible for large graph sizes N. There is however one source of new interactions giving an appreciable contribution. In the duplication process, if the ancestor is a self-interacting protein then the ancestor and the duplicate may have an interaction among them [144]. Let us assume that this happens with a probability q_v. Thus, if a neighbor of a vertex i is duplicated it will gain a new neighbor (the copy) and with probability q_v an edge between its neighbors (that between the copy and its ancestor) and therefore

$$\frac{\partial e_i}{\partial t} \approx q_v \frac{\partial d_i}{\partial t}. \tag{4.49}$$

where we have neglected any other process leading to new interactions and the edge lost. The integration of this equation yields

$$\langle c \rangle_d = \frac{2e(d)}{d(d-1)} \approx \frac{2q_v}{d}. \tag{4.50}$$

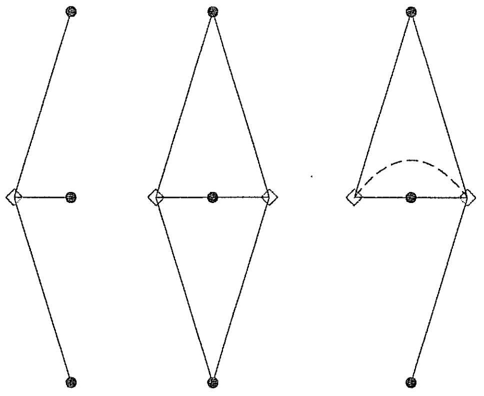


Figure 4.10: Schematic representation of the coupled duplication-divergence model evolution rules. Left and middle: A vertex (◇) is being duplicated. Right: The divergence of the duplicates is manifested as a coupled lost of interactions, where the coupling is given by the restriction that for each neighbor (●) at least one of the duplicates should preserve an edge to it. Moreover, due to the existence of self-interactions, a new edge can be created between the duplicates (dashed line).

Hence, under these assumptions we obtain the inverse proportionality behavior. The inclusion of the edge lost may change this result. We do not have any analytical prove but since this process contributes to the lost of triangles and it has a higher impact in high degree vertices then we expect that $\langle c \rangle_d$ decays faster than d^{-1}.

Coupled duplication-divergence model

In practice the processes of duplication and divergence cannot be decoupled. The protein-protein interaction network has a functional role in the organism and, therefore, the lost of certain interactions can result in the death of the corresponding organism. According to the classical model [113] after duplication the duplicate genes have fully overlapping functions. Later on, one of the copies may either become nonfunctional due to degenerative mutations or it can acquire a novel beneficial function and become preserved by natural selection. In a more recent framework [54, 87] it is proposed that both duplicate genes are subject to degenerative mutations loosing some functions but jointly retaining the full set of functions present in the ancestral gene. To investigate the influence of the coupling between duplication and divergence we introduce the following model: At each time step a vertex is added according to the following rules

- *Duplication*: a vertex i is selected at random. A new vertex $i\prime$ with an edge to all the neighbors of i is created. With probability q_v an edge between i and $i\prime$ is established (self-interacting proteins).

- *Divergence*: for each of the vertices j connected to i and $i\prime$ we choose randomly one of the two edges (i, j) or $(i\prime, j)$ and remove it with probability $1 - q_e$.

A schematic representation of this rules is shown in Fig. 4.10. For practical purposes the algorithm starts with two connected vertices and repeat the duplication-divergence rules N times. Since genome evolution analysis [144, 68] supports the idea that the divergence of duplicate genes takes place shortly after the duplication, we can assume that the divergence process always occurs before any new duplication takes place; i.e., we are in presence of a time scale separation between duplication and mutation rates. This allows us to consider the number of vertices in the network, N, as a measure of time (in arbitrary units). It is worth remarking that the algorithm does not include the creations of new edges, *i.e.* the developing of new interactions between gene products, other than those due to self-interactions. This process has been argued to have a probability relatively smaller than the divergence one[144]. However, we have tested that the introduction in the coupled duplication-divergence algorithm of a probability to develop new random connections does not change the network topology substantially.

In order to provide a general analytical understanding of the model, we use a mean-field approach for the moments distribution behavior. Let $\langle d\rangle (N)$ be the average degree of the network with N vertices. After a duplication event $N \to N + 1$ we have that the average degree is given by

$$\langle d\rangle (N+1) = \frac{N \langle d\rangle (N) + 2q_v + (2q_e - 1) \langle d\rangle (N)}{N+1}. \tag{4.51}$$

On average, there will be a gain proportional to $2q_v$ because of the interaction between duplicates, to $2\langle d\rangle (N)$ because of duplication, and a loss proportional to $2(1 - q_e) \langle d\rangle (N)$ due to the divergence process. For large N, taking the continuum limit, we obtain a differential equation for $\langle d\rangle$. For $q > 1/2$, $\langle d\rangle$ grows with N but saturates to the stationary value $\langle d\rangle = 2q_v/(1 - 2q_e) + \mathcal{O}(N^{2q_e-1})$, On the contrary, for $q_e > 1/2$, $\langle d\rangle$ grows with N as N^{2q_e-1}. At $q_e = q_1 = 1/2$ there is a dramatic change of behavior in the large scale degree properties. Analogous equations can be written for higher order moments $\langle d^l\rangle$ using the rate equations

$$\frac{\partial n_d}{\partial N} = A_{d-1}n_{d-1} - A_d n_d - \frac{n_d}{N} + 2q_v G_{d-1} + 2(1 - q_v)G_d, \tag{4.52}$$

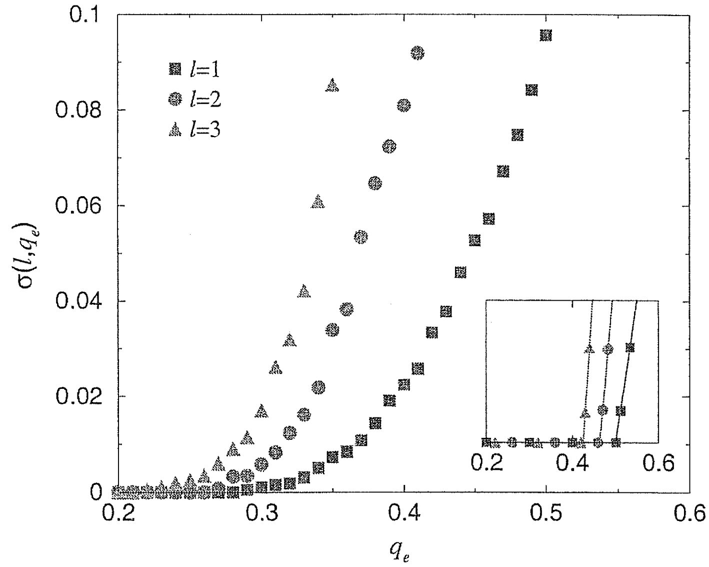


Figure 4.11: The exponent $\sigma_l(q_e)$ as a function of q_e for different values of l. The symbols were obtained from numerical simulations of the model. The moments $\langle d^l \rangle$ were computed as a function of N in networks with size ranging from $N = 10^3$ to $N = 10^6$. The exponents $\sigma_n(q)$ are obtained from the power law fit of the plot $\langle d^l \rangle$ vs. N. In the inset we show the corresponding mean-field behavior, as obtained from Eq. (4.56), which are in qualitative agreement with the numerical results.

where

$$A(d^{in}) = \frac{1}{N}(q_v + q_e d)\,, \tag{4.53}$$

$$G_d = \sum_{d' \geq d} \binom{d'}{d} \frac{n_{d'}}{N} \left(\frac{q_e}{2}\right)^d \left(1 - \frac{q_e}{2}\right)^{d'-d}. \tag{4.54}$$

The first two terms in the right hand side of Eq. (4.52) result from the duplication of a neighbor of a vertex (with probability $q_e d/N$) and the duplication of a vertex with the creation of an edge between the duplicates (with probability q_v/N), yielding the attachment rate in Eq. (4.53). Moreover, the last three terms are given by the divergence of the duplicates, where with probability n_d/N a vertex with degree d is replaced by two duplicates (factor two in the last two terms). Thus, the coupling of the duplication and divergence mixes the equations for different n_d. We cannot give an exact derivation of n_d but we can compute the moments of the degree distribution

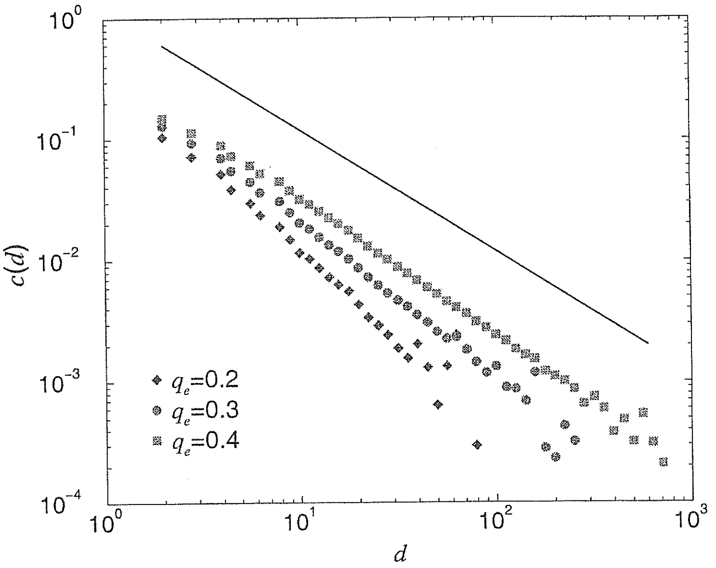


Figure 4.12: Clustering coefficient as a function of vertex degree of the coupled duplication-divergence model for different values of q_e, graph size $N = 10^6$ and average over 100 realizations. The solid line is a power law decay with exponent 1.

[77]. Multiplying Eq. (4.52) by d^l and summing over d we obtain

$$M_l = \sum_d p_d d^l \sim N^{\sigma_l(q_e)}, \tag{4.55}$$

where

$$\sigma_l(q_e) = lq_e + 2\left[\left(\frac{1+q_l}{2}\right)^l - 1\right], \tag{4.56}$$

provided $\sigma_l(q_e) > 0$. If $\sigma_l(q_e) < 0$ the corresponding moment approaches a stationary value for large N. Then, for all l we find a value q_l at which the moments cross from a divergent behavior to a finite value for $N \to \infty$. In particular for $l = 1$ we have $q_1 = 1/2$ (as obtained above) and for $l = 2$ we obtain $q_2 = 2\sqrt{3} - 3 \approx 0.46$. Moreover, the nonlinear behavior with l is indicative of a multi-fractal degree distribution.

In order to support the analytical calculations, we performed numerical simulations of the coupled duplication-divergence model with graph size ranging from $N = 10^3$ to 10^6. In Fig. 4.11 we report the generalized exponents $\sigma_l(q_e)$ as a function of the divergence parameter q. As predicted by the analytical calculations, $\sigma_l = 0$ at a critical value q_l. The general phase diagrams obtained is in good qualitative, but not quantitative, agreement with the mean-field predictions and the multi-fractal picture. Noticeably, multi-fractal features are present also in a recently introduced model of growing networks [47] where, in analogy with the duplication process, newly added vertices inherit the network degree properties from parent vertices. Multi-fractality, thus, appears related to local inheritance mechanisms. Multi-fractal distributions have a rich scaling structure where the scale-free behavior is characterized by a continuum of exponents. This behavior is, however, opposite to usual exponentially bounded distributions. Even if the evolution rules of the coupled duplication-divergence model are local they introduce an effective linear preferential attachment. However, because the edge deletion of duplicate vertices introduce additional heterogeneity in the problem we obtain a multi-fractal behavior.

The coupling between duplication and divergence is however less relevant to determine the scaling of the average clustering coefficient with vertex degree. In fact, for the coupled duplication-divergence model Eq. (4.49) also applies, obtaining the inverse proportionality in Eq. (4.50). In Fig. 4.12 we plot $\langle c \rangle_d$ vs. d for different values of q_e, manifesting a power law decay but with an exponent larger than 1. With decreasing q_e (increasing the lost of edges) the power law decay deviates more and more from the predicted behavior $\langle c \rangle_d \sim d^{-1}$. This picture corroborates our hypothesis that if the edge lost is sufficiently large then a faster decay should be observed. On the other hand, the average neighbor degree as a function of the vertex degree for different values of q_e is despited in Fig. 4.13. The existence of negative degree correlations are manifested by a power law decay $\langle d_{nn} \rangle \sim d^{-0.1}$.

4.4 Conclusions

After analyzing these models we can conclude that growing networks based on local evolution rules exhibit an effective linear preferential attachment. It is true that when we take a vertex at random the selection does not imply any degree preference, other than the one imposed by the degree distribution. However, if we take a neighbor of that vertex then some preference is induced. In fact the probability that vertex i is a neighbor of the randomly selected vertex is simply

$$\frac{d_i}{\sum_j d_j} \tag{4.57}$$

which is exactly the linear preferential attachment considered in the Barabási *et al* model [12]. Therefore, the connection to a neighbor of a vertex selected at random leads an effective linear preferential attachment.

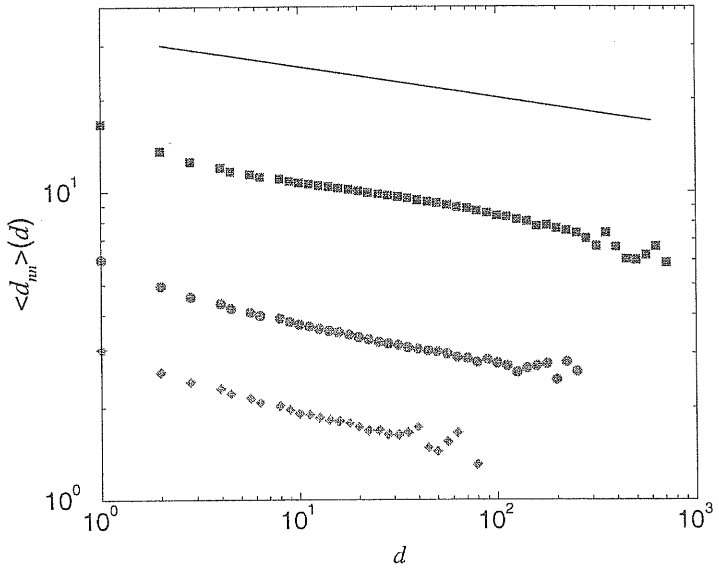


Figure 4.13: Average degree among the neighbors of a vertex with degree d of the coupled duplication-divergence model for different values of q_e, graph size $N = 10^6$ and average over 100 realizations. The solid line is a power law decay with exponent 0.1.

Another important consequence of the local models considered above is the inverse proportionality between the average clustering coefficient and the vertex degree, or more general $\langle c \rangle \sim d^{-\beta}$. This result is determined by the fact that when a new edge is created to a vertex then with a certain probability an edge will also be created to one or more of its neighbors. Thus, locality is again a crucial point. On the other hand, even if we were not able to find an analytical explanation, these local models are also characterized by degree correlations among connected vertices.

Hence, the growing models with local rules exhibit some of the common features of real graphs, the effective preferential attachment [73, 13, 104, 115, 138], an average clustering coefficient that decreases with increasing vertex degree [138, 121], and degree correlations [115, 138, 103].

Chapter 5

Statistical mechanics on graphs with degree correlations

In the previous chapter it was shown that many real networks exhibit correlations. In particular, we observed the existence of two types of correlations: degree-degree correlations and clustering hierarchy. A second step will then be to study the influence of correlations and compare the results with those obtained for uncorrelated graphs. Some works in that direction are starting to emerge. Newman [103] have studied the problem of percolation on graphs with an arbitrary degree distribution and degree correlations. Moreover, Boguña and Pastor Satorras [101] have analyzed the problem of epidemic spreading. In both cases they observed that the precise value of the threshold depends on the magnitude of the degree correlations.

In this chapter we develop a statistical mechanics approach on top of random graphs with an arbitrary degree distribution and arbitrary degree correlations. Using the Bethe-Peierls [19, 118] and the replica formalism [96] for dilute systems we compute the free energy and related thermodynamic properties. We discuss the problem of percolation and its generalizations to site and bond percolation making emphasis on the influence of degree correlations and the general conditions for the existence of a giant component [136]. Latter we focus our attention in the NP-hard optimization problem of vertex covering [55]. We have chosen the vertex covering problem because in this way we can also investigate the influence of the degree distribution and degree correlations on the computational complexity [55, 114]. We show that such correlations may lead to a qualitatively different solution structure as compared to uncorrelated networks, resulting in a higher complexity of the network in a computational sense. We also investigate the influence of correlations on the performance of different heuristic algorithms to find the minimal vertex cover, obtaining that simple heuristic algorithms fail to find a minimal vertex cover in the highly correlated case, whereas uncorrelated networks seem to be simple from the point of view of combinatorial optimization. Moreover, a real application, to the vertex cover of the Internet and other real graphs, is also discussed. Part of these results have been published in [140].

5.1 Statistical mechanics on graphs

Consider the set of undirected graphs with N vertices and arbitrary degree distribution p_d. Following a randomly chosen edge, we will find a vertex of degree d with probability

$$q_d = \frac{dp_d}{\langle d \rangle}. \tag{5.1}$$

We further assume correlations between adjacent vertices: The probability that a randomly chosen edge connects two vertices of degrees d, d' is given by $(2 - \delta_{d,d'})e_{dd'}$. The conditional probability that a vertex of degree d is reached following any edge coming from a vertex of degree d', is

$$p(d'|d) = e_{dd'}/q_d \ , \tag{5.2}$$

thus explicitly depends on both d and d'. Consistency with the degree distribution requires $\sum_{d'=0}^{\infty} e_{dd'} = q_d$, and $e_{dd'}$ has to be symmetric. For uncorrelated graphs $e_{dd'} = q_d q_{d'}$ factorizes resulting $p(d'|d) = q_{d'}$ that is independent of d.

Let us now consider a general statistical-mechanics model with discrete degrees of freedom $x_i = 0, 1$ defined on vertices $i = 1, ..., N$, and interactions $J_{ij} = 0, 1$ defined on edges, with the partition function

$$Z = \sum_{\{x_i=0,1\}} e^{\mu \sum_i x_i} \prod_{i<j|J_{ij}=1} \chi(x_i, x_j), \tag{5.3}$$

where μ is the chemical potential. J is the adjacency matrix with entries $J_{ij} = 1$ if vertices i and j are adjacent (connected), and $J_{ij} = 0$ else. $\chi(x_i, x_j)$ is an arbitrary interaction term and its precise form will depend on the model under consideration. The only disorder present in Eq. (5.3) is given by the edges J_{ij}. Generalizations to disordered interactions, as present e.g. in spin-glasses, random local fields or non-binary discrete variables are straightforward. The free energy will then given by

$$-F = \overline{\ln Z}, \tag{5.4}$$

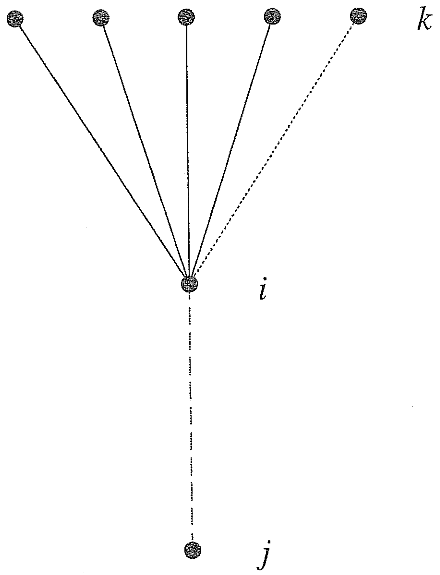


Figure 5.1: Schematic representation of the Bethe-Peierls approach. We focus on the subtree rooted in i with deleted edge (i,j) and iteratively compute the partition function, expressed in terms of that of the subtree rooted in j with deleted edge (j,k).

where the over-bar denotes the disorder average over interactions J_{ij},

$$\overline{A} = \frac{1}{\mathcal{N}} \int \prod_{i<j} dJ_{ij} P(J_{ij}) \prod_{i=1}^{N} \delta\left(\sum_j J_{ij} - d_i\right) A, \tag{5.5}$$

where

$$\mathcal{N} = \prod_{i<j} dJ_{ij} P(J_{ij}) \prod_{i=1}^{N} \delta\left(\sum_j J_{ij} - d_i\right), \tag{5.6}$$

is a normalization factor and $P(J_{ij})$ is the probability distribution of the interaction matrix elements. Notice that the delta functions enforce the constraints on the degree distribution while the degree correlations are taken into account in the definition of $P(J_{ij})$,

$$P\left(J_{ij}\right) = \left(1 - \frac{\langle d\rangle}{N} \frac{e_{d_i d_j}}{q_{d_i} q_{d_j}}\right) \delta\left(J_{ij}\right) + \frac{\langle d\rangle}{N} \frac{e_{d_i d_j}}{q_{d_i} q_{d_j}} \delta\left(J_{ij} - 1\right), \tag{5.7}$$

where $\langle d\rangle$ is the average degree.

5.1.1 Bethe-Peierls iterative approach

Since the graphs are locally tree-like, the model can be solved by the iterative Bethe-Peierls scheme which becomes exact if only one pure state is present. The free energy can be expressed in terms of simple effective-field distributions acting on vertices of given degree. In the case of multiple pure states this has to be generalized to the cavity approach, see e.g. [94] for the example of a spin-glass on a Bethe lattice with constant vertex degree. Alternatively, one can apply the replica approach. The simple Bethe-Peierls solution corresponds to the assumption of replica symmetry (RS), whereas the full cavity approach is able to handle also the case of replica symmetry breaking (RSB).

Take any edge (i,j), i.e. $J_{ij}=1$. Let us introduce $Z_x^{(i|j)}$ as the partition function of the subtree rooted in i, with deleted edge (i,j), and with x_i fixed to the value x. This partition function can be calculated iteratively (see Fig. 5.1),

$$\begin{aligned} Z_0^{(i|j)} &= \prod_{k\neq j|\ J_{ik}=1} \left(\chi(0,0)Z_0^{(k|i)} + \chi(0,1)Z_1^{(k|i)}\right) \\ Z_1^{(i|j)} &= e^{\mu} \prod_{k\neq j|\ J_{ik}=1} \left(\chi(1,0)Z_0^{(k|i)} + \chi(1,1)Z_1^{(k|i)}\right) \end{aligned} \tag{5.8}$$

The effective fields

$$h_{(i|j)} = \ln\left(\frac{Z_1^{(i|j)}}{Z_0^{(i|j)}}\right) , \tag{5.9}$$

are thus determined by the iterative description (dividing the second by the first equation in (5.8))

$$h_{(i|j)} = \mu + \sum_{k\neq j|\ J_{ik}=1} u(h_{(k|i)}) , \tag{5.10}$$

where $u(h_{(k|i)})$ is the effective field induced by x_k on site i, and is given by

$$u(h_{(k|i)}) = \ln\left(\frac{\chi(1,0)+\chi(1,1)e^{h_{(k|i)}}}{\chi(0,0)+\chi(0,1)e^{h_{(k|i)}}}\right) . \tag{5.11}$$

Let us now assume, that the model has only one pure state, which corresponds to the assumption of RS. In this case, the iterative procedure given by Eq. (5.10) converges to well-defined distributions $P_d(h)$ and $Q_d(u)$ of effective fields $h_{(i|j)}$ and $u(h_{(i|j)})$ restricted to vertices of excess degree d. They are determined by the self-consistency equation

$$P_d(h) = \int \prod_{t=1}^{d-1} du_t Q_d(u_t)\delta\left(h-\mu-\sum_t u_t\right), \tag{5.12}$$

$$Q_d(u) = \sum_{d_1} p(d_1|d) \int \prod_{t=1}^{d_1-1} du_t Q_{d_1}(u_t)\delta\left[u - f_\mu\left(\mu+\sum_t u_t\right)\right], \tag{5.13}$$

where

$$f_\mu(h) = \frac{1}{2}\ln\left[\frac{e^{\mu h}\chi(1,1)+\chi(1,0)}{e^{\mu h}\chi(0,1)+\chi(0,0)}.\right]. \tag{5.14}$$

Notice that u is the contribution to the local field given by one neighbor while h given by all $d-1$ neighbors.

5.1.2 Replica approach

The Bethe-Peierls solution corresponds to the assumption of replica symmetry. Hence, the iterative equations for the local fields can be also obtained by a direct calculation of the partition function and the free energy using the replica formalism [96]. The main difficulty in computing the free energy is caused by the average over the disorder in Eq. (5.4). The replica trick overcome this exploiting the identity

$$\overline{\ln Z} = \lim_{n\to 0}\frac{\overline{Z^n}-1}{n}. \tag{5.15}$$

Now, for integer n the n^{th} moment of the partition function in Eq. (5.3) can be rewritten as

$$\overline{Z^n} = \sum_{\{x_i=0,1\}} \overline{e^{\mu\sum_{ia}x_i^a}\prod_{a,i<j|J_{ij}=1}\chi(x_i^a,x_j^a)}, \tag{5.16}$$

where $a = 1,...,n$ is the copy, or replica, index. To compute the average over the disorder we write the degree constraints in the integral form

$$\delta\left(\sum_j J_{ij}-d_i\right) = \int\frac{d\psi_i}{2\pi}e^{i(\sum_j J_{ij}-d_i)\psi_i}. \tag{5.17}$$

For sparse graphs ($\langle d\rangle \ll N$) from Eqs. (5.5), (5.7) and (5.16), it follows that

$$\begin{aligned}\overline{Z^n} &= \frac{1}{\mathcal{N}}\sum_{\{x_i^a=0,1\}}\int\prod_i\left(\frac{d\psi_i}{2\pi}e^{-id_i\psi_i}\right) \\ &\times \exp\left(-\frac{\langle d\rangle N}{2}+\frac{\langle d\rangle}{2N}\sum_{ij}e^{i(\psi_i+\psi_j)+\mu H_o\sum_{i,a}x_i^a}\prod_a\chi(x_i^a,x_j^a)\right)\end{aligned} \tag{5.18}$$

In analogy with a previou calculation for uncorrelated graphs [84] we introduce the functional order parameter

$$\rho_d(\vec{\sigma}) = \frac{1}{N}\sum_i \delta(\vec{\sigma}-\vec{x}_i)\delta_{dd_i}e^{i\psi_i}, \tag{5.19}$$

and its complex conjugate $\hat{\rho}_d(\vec{\sigma})$. Then, tracing over the spins x_i^a and integrating out the ψ_i variables [83] one is left with the following expression for the n^{th} moment of the partition function

$$\overline{z^n} = \int \prod_{d,\vec{\sigma}} d\rho_d(\vec{\sigma})d\hat{\rho}_d(\vec{\sigma})e^{-nNF}, \tag{5.20}$$

with the free energy

$$\begin{aligned} -nF &= -\langle d\rangle \sum_{d\vec{\sigma}} \rho_d(\vec{\sigma})\hat{\rho}_d(\vec{\sigma}) + \sum_d p_d \ln \sum_{\vec{\sigma}} e^{\mu\sum_a \sigma_a}\left(\hat{\rho}_d(\vec{\sigma})\right)^d \\ &+ \frac{\langle d\rangle}{2}\left[1+\sum_{\vec{\sigma}_1\vec{\sigma}_2 d_1 d_2} \frac{e_{d_1d_2}}{q_{d_1}q_{d_2}}\rho_{d_1}(\vec{\sigma}_1)\rho_{d_2}(\vec{\sigma}_2)\prod_a \chi\left(\sigma_1^a,\sigma_2^a\right)\right]. \end{aligned} \tag{5.21}$$

The main contribution to the free energy in the thermodynamic limit is evaluated *via* the following functional saddle point equations:

$$\rho_d(\vec{\sigma}) = q_d \frac{e^{\mu\sum_a \sigma^a}\left(\hat{\rho}_d(\vec{\sigma})\right)^{d-1}}{\sum_{\vec{\sigma}_1} e^{\mu\sum_a \sigma_1^a}\left(\hat{\rho}_d(\vec{\sigma}_1)\right)^d}, \tag{5.22}$$

$$\hat{\rho}_d(\vec{\sigma}) = \sum_{\vec{\sigma}_1 d_1} \frac{e_{dd_1}}{q_d q_{d_1}}\rho_{d_1}(\vec{\sigma}_1)\prod_a \chi(\sigma^a,\sigma_1^a). \tag{5.23}$$

In the limit $n \to 0$ we can check that the order parameters are normalized.

Replica symmetric ansatz

Under the assumption of replica symmetry the local fields are given by

$$\rho_d(\vec{\sigma}) = q_d \int dh P_d(h)\frac{e^{h\sum_{a=1}^n \sigma_a}}{(1+e^h)^n}, \tag{5.24}$$

$$\hat{\rho}_d(\vec{\sigma}) = \int du Q_d(u)\frac{e^{u\sum_{a=1}^n \sigma_a}}{(1+e^u)^n}. \tag{5.25}$$

Substituting these expressions in Eqs. (5.22) and (5.23) we obtain the following set of coupled equations

$$P_d(h) = \int \prod_{t=1}^{d-1} du_t Q_d(u_t)\delta\left(h-\mu-\sum_t u_t\right), \tag{5.26}$$

$$Q_d(u) = \sum_{d_1} p(d_1|d) \int \prod_{t=1}^{d_1-1} du_t Q_{d_1}(u_t) \delta \left[u - f_\mu \left(\mu + \sum_t u_t \right) \right], \qquad (5.27)$$

where

$$f_\mu(h) = \frac{1}{2} \ln \left[\frac{e^{\mu h} \chi(1,1) + \chi(1,0)}{e^{\mu h} \chi(0,1) + \chi(0,0)} \right]. \qquad (5.28)$$

$P_d(h)$ is the average probability distribution of effective fields and $Q_d(u)$ is that of cavity fields, acting on sites with degree d. These equations coincide with those obtained using the Bethe-Peierls scheme (see Eqs. (5.12) and (5.13)). Moreover, we would like to stress that the strong inhomogeneities present in the graph are correctly taken into account and handled via the computation of the whole probability distributions. Finally, substituting Eqs. (5.24) and (5.25) into Eq. (5.21) we obtain the following expression for the free energy

$$\begin{aligned} F &= \sum_d (d-1) p_d \int \prod_{t=1}^{d} du_t Q_d(u_t) \ln \left(1 + e^{\mu + \sum_t u_t}\right) \\ &- \frac{\langle d \rangle}{2} \sum_{d_1 d_2} e_{d_1 d_2} \int dh_1 dh_2 P_{d_1}(h_1) p_{d_2}(h_2) \\ &\times \ln \left[e^{\mu(h_1+h_2)} \chi(1,1) + e^{\mu h_1} \chi(1,0) + e^{\mu h_2} \chi(0,1) + \chi(0,0) \right] \qquad (5.29) \end{aligned}$$

Then, using Eqs. (5.26)-(5.29) we can compute the different thermodynamic quantities.

5.2 Percolation

The problem of percolation is equivalent to the Ising model at zero temperature and zero magnetic field, where the size of the giant component is just the magnetization [52]. We can exploit this analogy to study the percolating properties of random graphs with arbitrary degree correlations. In the infinite temperature Ising model if two vertices interact they have to be in the same state and, therefore, percolation theory corresponds with

$$\chi_{perc}(x_i, x_j) = \delta_{x_i, x_j}, \qquad (5.30)$$

and $\mu = 0$. Thus, substituting $\chi(x_i, x_j)$ and μ by these expressions in Eq. (5.27) it results that

$$Q_d(u) = \sum_{d_1} p(d_1|d) \int \prod_{t=1}^{d_1-1} du_t Q_{d_1}(u_t) \delta \left(u - \sum_t u_t \right). \qquad (5.31)$$

Let $\pi_d = 1 - Q_d(0)$ be the probability that an edge incident to a vertex of degree d carries a constraint, *i.e.* the neighbor is in the state 1. According to Eq. (5.31) $u = 0$ corresponds with $\sum_t u_t = 0$ resulting

$$1 - \pi_d = \sum_{d_1} p(d_1|d)\,(1 - \pi_{d_1})^{d_1 - 1}\,. \tag{5.32}$$

Moreover, if we assume that the magnetization of the largest cluster is positive then the local fields acting on a vertex can be either 0 or strictly positive. If it is strictly positive then the vertex is in the largest cluster and, therefore, the fraction of vertices S in the giant component is given by

$$S = 1 - \sum_d p_d(1 - \pi_d)^d. \tag{5.33}$$

These results have been already obtained by Newman [110] using the generating function formalism.

5.2.1 Dilute percolation

Let us generalize this result to the site percolation problem. In this case a fraction f of the vertices is removed from the graph and the new giant component is computed. Since the vertex removal is independent of the vertex degree this is equivalent to replace the original degree distribution and correlations by: (i) the probability that a vertex selected at random has degree d and it has not been removed, and (ii) the probability that if we select a vertex at random and follow one of its edges we end in a vertex with degree d' that has not been removed, *i.e.*

$$p_d \to (1-f)p_d, \qquad p(d'|d) \to (1-f)p(d'|d), \tag{5.34}$$

Substituting Eqs. (5.34) in Eqs. (5.33) and (5.32) we get

$$S = 1 - f - (1-f)\sum_d p_d u_d^d, \tag{5.35}$$

$$u_d = f + (1-f)\sum_{d'} p(d'|d) u_{d'}^{d'-1}, \tag{5.36}$$

where the term $-f$ (f) in Eq. (5.35) (Eq. (5.36)) gives the probability of hitting a removed vertex. One solution to these equations is $u_d = 1$ yielding $S = 0$. This solution is valid whenever the equation for the u_d is stable under successive approximations. That is, if we start with $u_d(t) = 1 - \rho_d(t)$ and compute the successive approximation $\rho_d(t+1)$ then we should obtain that $\rho_d(t) \to 0$ in the limit $t \to \infty$. For $\rho_d(t) \ll 1$ the last equation is approximated by the linear map

$$\rho_d(t+1) = \sum_{d'} L_{dd'} \rho_{d'}(t), \tag{5.37}$$

with

$$L_{dd'} = (1-f)C_{dd'}, \qquad C_{dd'} = (d'-1)p(d'|d). \tag{5.38}$$

The stability of the solution $u_d = 1$ is then related to the largest eigenvalue of $L_{dd'}$. If it is smaller (larger) than 1 the solution is stable (unstable). Since $L_{dd'}$ is linear in f the stability condition can be written as

$$f > f_c, \qquad (1-f_c)\Lambda_{max} = 1, \tag{5.39}$$

where Λ_{max} is the largest eigenvalue of $C_{dd'}$ provided that $\Lambda_{max} > 1$. If $\Lambda_{max} < 1$ the graph does not have a giant component even for $f = 0$. Moreover, since $C_{dd'}$ is a positive matrix then Λ_{max} has the lower and upper bounds $\min_d \sum_{d'} C_{dd'}$ and $\max_d \sum_{d'} C_{dd'}$, yielding

$$\min_d \langle d \rangle_{nn}(d) \leq 1 + \Lambda_{max} \leq \max_d \langle d \rangle_{nn}(d). \tag{5.40}$$

where

$$\langle d \rangle_{nn}(d) = \sum_{d'} p(d'|d)d', \tag{5.41}$$

is the average degree among the neighbors of a vertex with degree d [115]. Eq. (5.40) can be used to determine, based on a simple topological measure, whether or not a given graph is robust under vertex removal.

On the other hand, in the bond percolation problem a fraction f of the edges is removed from the graph and the new giant component is computed. Since the edge removal is made at random this is equivalent to keep the original degree distribution and replace the degree correlations by the probability that if we select a vertex at random and follow one of its edges we end in a vertex with degree d' that has not been removed, *i.e.*

$$p_d \to p_d, \qquad p(d'|d) \to (1-f)p(d'|d). \tag{5.42}$$

Substitution of Eq. (5.42) in Eqs. (5.33) and (5.32) yields

$$S = 1 - \sum_d p_d u_d^d, \tag{5.43}$$

$$u_d = f + (1-f)\sum_{d'} p(d'|d)u_{d'}^{d'-1}. \tag{5.44}$$

Note that the only difference between the site and bond percolation problems (see Eqs. (5.35) and (5.36)) is the equation for the giant component while that for u_d is identical. Hence, Eqs. (5.39) and (5.40) are also valid for the bond percolation problem.

In what follows we consider some particular graphs in order to analyze the effects of correlations. Depending on the monotony of $\langle d \rangle_{nn}(d)$ the degree correlations can be classified in: uncorrelated if it is independent of d, assortative or positive if it increases with increasing d and disassortative or negative if it decreases with decreasing d. A similar definition has been introduced in Ref. [103] using a correlation coefficient.

5.2.2 Graphs with random mixing

For random graphs with no constraint other than the one imposed by the degree distribution we have $p(d',d) = q_{d'}$. In this case the lower and upper bounds in Eq. (5.40) are equal giving for the largest eigenvalue

$$\Lambda_{max}^{unco} = \frac{\langle d^2 \rangle}{\langle d \rangle} - 2. \tag{5.45}$$

Alternatively one can compute Λ directly from the eigenvalue problem of $C_{dd'}$. Then from Eq. (5.39) we obtain $1 - f_c = 1/(\langle d^2 \rangle / \langle d \rangle - 2)$ [37]. Hence, if the second moment $\langle d^2 \rangle$ diverges the threshold equals 1, *i.e.* the network is robust under random vertex or edge removal. Furthermore, consider the case in which the degree correlations can be decomposed into two components

$$p(d'|d) = \alpha q_{d'} + (1-\alpha)\delta p(d'|d) \tag{5.46}$$

with $0 < \alpha < 1$ and $\delta p(d'|d) > 0$ for all (d, d'). Varying the parameter α one interpolates between the uncorrelated graphs ($\alpha = 1$) and a graph with arbitrary degree correlations given by $\delta p(d'|d)$. In this case from Eq. (5.40) we obtain $\Lambda_{max} \geq \alpha \Lambda_{max}^{unco}$ and, therefore, if the network is robust for the uncorrelated case it will also be robust for any $\alpha > 0$. This immediately implies that any graph with a divergent second moment and a finite amount of random mixing of the edges does not have a percolation threshold.

5.2.3 Graphs with assortative mixing

Assortative correlations allow us to show that the divergence of the second moment is not a necessary condition for the absence of the threshold. Let us consider a network with degree correlations

$$p(d'|d) = \alpha \delta_{dd'} + (1-\alpha)\delta p(d'|d), \tag{5.47}$$

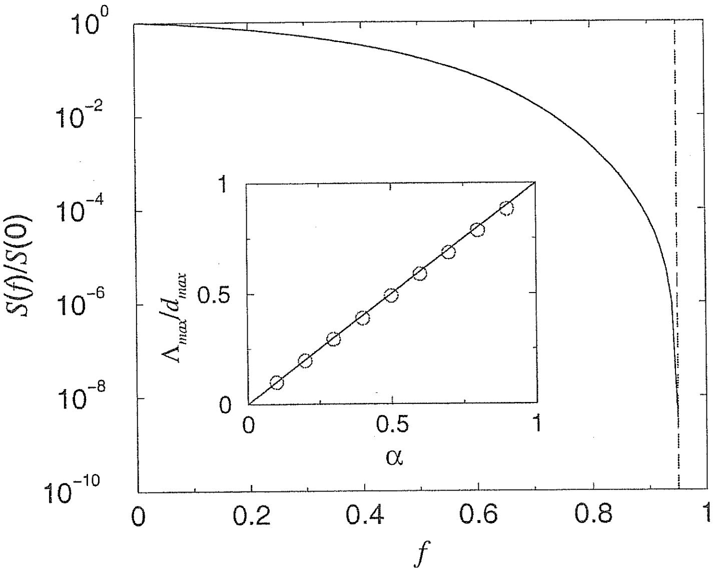


Figure 5.2: Size of the giant component for a graph with $p_d = cd^{-3.5}$ $(2 \leq d \leq 100)$ and degree correlations $p(d'|d) = \alpha\delta_{dd'} + (1-\alpha)q_{d'}$ with α, as computed from Eq. (5.35). The dashed line marks the percolation threshold obtained using perturbation theory (Eq. (5.49)). The inset shows the largest eigenvalue relative to d_{max} as a function of α. The points where computed numerically and the line is the perturbation theory dependency $\Lambda_{max}/d_{max} = \alpha$.

with $0 < \alpha < 1$ and $\delta p(d'|d) > 0$ for all (d, d'). $\alpha = 1$ corresponds to a fully assortative graph made up of sub-graphs with fixed degree. In this case $C_{dd'} = d'\delta_{dd'}$ (see Eq. 5.38) is already diagonal. The largest eigenvalue is $\Lambda_{max} = d_{max}$, where d_{max} is the largest degree. If d_{max} diverges for $N \to \infty$ then $f_c = 1$. For the more general case $0 < \alpha < 1$ we compute the largest eigenvalue using perturbation theory [127] around $\alpha = 1$, obtaining

$$\Lambda_{max}(\alpha) = \alpha d_{max} + (1-\alpha)C_{d_{max}d_{max}}. \tag{5.48}$$

This result is valid whenever $(1-\alpha)C_{d_{max}d_{max}} \ll \alpha d_{max}$. In general $C_{d_{max}d_{max}}$ decreases with increasing d_{max}, resulting

$$\Lambda_{max}(\alpha) \approx \alpha d_{max}, \tag{5.49}$$

for $d_{max} \gg 1/\alpha$. Hence, for any $\alpha > 0$ and any unbounded degree distribution we have $f_c = 1$, *i.e.* there is no percolation threshold. In Fig. 5.2 we show the validity of the perturbation theory for a particular perturbation $\delta p(d'|d)$.

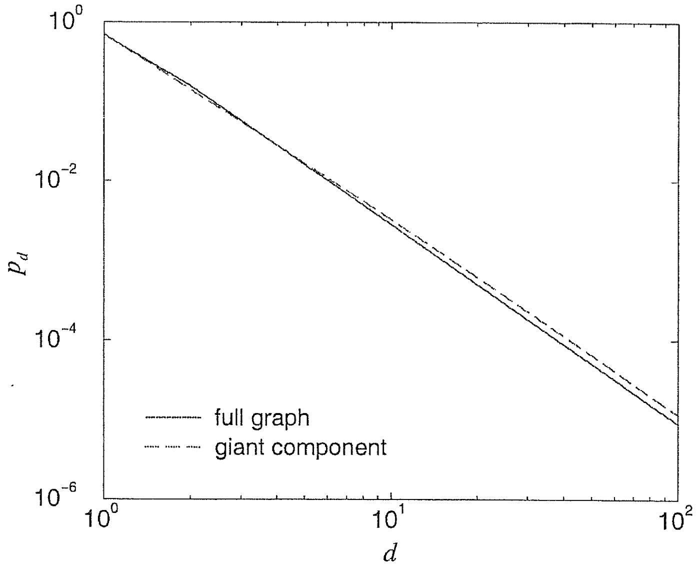


Figure 5.3: Comparison between the degree distribution in the full graph and in the giant component in disassortative graphs, with degree correlations and degree distribution in Eq. (5.53), with $\gamma = 2.5$ and $\alpha = 0.5$.

Thus, as in the fully assortative case, if $\alpha > 0$ and d_{max} diverges $f_c = 1$. Therefore, we can conclude that the divergence of the second moment is not a necessary condition.

5.2.4 Graphs with disassortative mixing

Is the divergence of the second moment a sufficient condition for $f_c = 1$? The answer is no as shown by the following example of a disassortative graph. Consider a graph made by a collection of stars, with a central vertex connected to vertices of degree 1, interconnected among them. Take a vertex with degree $d > 1$ (the center of a star) and an edge incident to it. Then with probability g_d a vertex at the other end is chosen at random among all vertices with degree $d' > 1$, otherwise it is connected to a vertex with $d' = 1$ chosen at random, *i.e.*

$$\begin{aligned} p(d'|d) &= \frac{(1-g_{d'})d'p_{d'}}{\sum_s(1-g_s)sp_s}\Theta(d'-1)\delta_{d,1} \\ &+ (1-g_d)\delta_{d',1}\Theta(d-1) \\ &+ g_d\frac{g_{d'}d'p_{d'}}{\sum_s g_s s p_s}\Theta(d'-1)\Theta(d-1). \end{aligned} \tag{5.50}$$

where $\Theta(x)$ is the unitary step function ($\Theta(x) = 0$ for $x \leq 0$ and $\Theta(x) = 1$ for $x > 0$). The first term in the right hand side is the contribution from the connection of vertices with degree 1 to the centers of the stars. The second accounts for the connection of the stars to their leafs. The last one represents the interconnection among the stars.

The fraction of vertices with degree 1 is obtained self-consistently from the condition $p_1 = \sum_{d>1}(1 - g_d)dp_d$. Moreover, the average degree of the neighbors of a vertex with $d > 1$ is given by

$$\langle d \rangle_{nn} = 1 + g_d \left(\frac{\sum_{d'>1} g_{d'} d'^2 p_{d'}}{\sum_s g_s s p_s} - 1 \right), \tag{5.51}$$

and, therefore, these graphs are disassortative for any monotonic decreasing function g_d. To analyze the percolation properties of this graph we computed exactly the largest eigenvalue of $C_{dd'} = (d' - 1)p(d'|d)$, resulting

$$\Lambda_{max} = \frac{\sum_d g_d^2 (d-1) dp_d}{\sum_s g_s s p_s}. \tag{5.52}$$

Hence, the conditions for the existence of a giant component ($\Lambda_{max} > 1$) or resilience to damage ($\Lambda_{max} = \infty$) are modulated by g_d and, therefore, the disassortative correlations given by g_d have a great impact on the percolation properties. For instance, let us consider

$$g_d = d^{-\alpha}, \quad p_d = cd^{-\gamma}, \tag{5.53}$$

with $\gamma < 3$ ($\langle d^2 \rangle = \infty$). From Eq. (5.51) it follows that $\langle d \rangle_{nn} - 1 \sim d^{-\alpha}$ so that with increasing α the graph gets more and more disassortative. Moreover, Λ_{max} diverges for $\alpha < \alpha_c$ where

$$\alpha_c = (3 - \gamma)/2, \tag{5.54}$$

and it is finite otherwise. Thus, for small values of α the graph is robust but for $\alpha > \alpha_c$ it becomes fragile. It is worth noticing that the value of α above which the giant component disappears ($\Lambda_{max} < 1$) is larger than α_c. Besides, for large degrees, the degree distribution of the vertices in the giant component is still a power law, and it decays slower than that of the whole graph, as shown in Fig. 5.3. Thus, disassortative correlations compete against the formation of the giant component and, the divergence of $\langle d^2 \rangle$ is not a sufficient condition to get a robust graph with $f_c = 1$.

5.2.5 Epidemic spreading

The connection between percolation theory and models of epidemic spreading is well known [61]. Two general classes of epidemiological models can be related to percolation problems, the Susceptible-Infected-Removed (SIR) and the Susceptible-Infected-Susceptible (SIS) classes. The SIR model assumes that individuals can exist in three classes and that once they get infected they can not catch the infection again. This model can be mapped into a bond percolation problem taking f as the probability that the disease will be transmitted from one vertex to another and the size of the giant component as the size of the outbreak. Hence, all the conclusions drawn above for the bond percolation problem can be translated to the language of epidemic spreading for the SIR model on top of graphs with degree correlations, extending in this way a previous study by Newman for uncorrelated graphs [109].

On the other hand, the SIS model allows individuals to move through the cycle of infection so that the prevalence (number of infected individuals) attains a stationary value. The SIS model on top of graphs with degree correlations has been recently analyzed by Boguña and Pastor-Satorras [101] as a function of the effective spreading rate λ. They obtained the epidemic threshold (the value of λ above which the solution with zero prevalence is unstable) $\lambda = 1/\Lambda'_{max}$, where Λ'_{max} is the largest eigenvalue of the matrix

$$C'_{dd'} = dp(d'|d). \tag{5.55}$$

This approach is quite similar to the one presented here for site percolation with the remark that $C'_{dd'}$ is different (see Eq. (5.38)). In fact, if y_d is an eigenvector of $C'_{dd'} = dp(d'|d)$ corresponding to the eigenvalue Λ' then y_d/d is an eigenvector of $C''_{dd'} = d'p(d'|d)$ corresponding to the same eigenvalue. This last matrix is that of Eq. (5.38), but replacing d' by $d'-1$. However, this subtle difference makes the SIS and dilute percolation different. We have computed the largest eigenvalue of $C''_{dd'}$ for the disassortative graph considered above(Eq. (5.50)). Taking the limit $\langle d^2 \rangle \gg 1$ one gets

$$\Lambda'_{max} \approx \frac{\sum_d (1-g_d) d^2 p_d}{\sum_s (1-g_s) s p_s}, \tag{5.56}$$

where g_d is again a decreasing function of d. In this case, independent of the form of g_d, the divergence of the second moment of the degree distribution implies the divergence of Λ'_{max}. Moreover, the same conclusion is obtained if g_d is an increasing function of d. The conditions for the existence of a finite prevalence in the SIS model have been recently addressed in [21], where the divergence of the second moment has been shown to be a sufficient condition for the absence of the phase transition in the SIS model. Nevertheless, we have shown that this conclusion does not hold for dilute percolation. This essential difference is rooted in the existence of an additional dimension in the SIS model, given by the time evolution of the density of infected sites.

5.3 Vertex covering

Another application of the formalism developed above is given by the vertex cover problem. It belongs to the basic NP-hard optimization problems [55] and, therefore, it is expected to require a solution time which is growing exponentially with the graph size. Let us be more precise. Given a graph with vertices $i \in \{1, ..., N\}$ and edges $\{(i,j) | 1 \leq i < j \leq N, J_{ij} = 1\}$, a *vertex cover* V is a subset of vertices, $V \subset \{1, ..., N\}$, such that at least one end-vertex of every edge is contained in V. So no edge (i,j) is allowed to exist with $i \notin V$ and $j \notin V$. Of course, the set of all vertices forms a trivial vertex cover. The hard optimization problem consists in finding the *minimal vertex cover.*

Using the the hard-sphere lattice gas representation introduced by Hartmann and Weight [149]: $x_i = 1$ if $i \notin V$, and $x_i = 0$ if $i \in V$, the vertex cover condition can be rewritten as

$$\prod_{i<j|J_{ij}=1} (1 - x_i x_j) = 1, \tag{5.57}$$

Hence, in our formalism, the minimal vertex cover will correspond with

$$\chi_{vc}(x_i, x_j) = 1 - x_j x_j. \tag{5.58}$$

Moreover, the chemical potential μ can be used to fix the cardinality (the fraction of covered vertices) of the vertex cover, minimal ones are obtained in the limit $\mu \to \infty$.

In this case from Eq. (5.27) we obtain

$$\begin{aligned} Q_d(u) &= \sum_{d_1} p(d_1|d) \int \prod_{t=1}^{d_1-1} du_t Q_{d_1}(u_t) \\ &\times \delta\left[u + \left(1 + \sum_t u_t\right) \Theta\left(1 + \sum_t u_t\right)\right]. \end{aligned} \tag{5.59}$$

where $\Theta(x)$ is the unitary step function, $\Theta(x) = 0$ for $x \leq 0$ and $\Theta(x) = 1$ for $x > 0$. From Eq. (5.59) it follows that the cavity fields can only take the values $u = -1$ and $u = 0$. Let $\pi_d = Q_d(-1)$ be the probability that an edge incident to a vertex of degree d carries a constraint, *i.e.* that is not yet covered by the neighbor vertex. Then, form Eq. (5.59) we obtain the following self-consistent equation

$$\pi_d = \sum_{d_1} p(d_1|d)\,(1-\pi_{d_1})^{d_1-1}\,. \tag{5.60}$$

Moreover, taking into account that the minimal size of the vertex cover is the average fraction of vertices with $x_i = 0$ we obtain

$$x_c = 1 - \sum_d p_d(1-\pi_d)^{d-1}\left(1 - \frac{d-2}{2}\pi_d\right). \tag{5.61}$$

The expressions obtained above are related to the validity of the replica symmetry, i.e. to the existence of a single connected cluster of minimal vertex covers in configuration space. As observed in [150, 149], the replica symmetry is related to the *local* stability of this solution. In presence of replica symmetry breaking, Eq. (5.60) has no stable solution. Since it has to be solved by numerical iteration in the general case, an instability prevents the program from convergence and thus provides a precise tool to detect replica symmetry breaking without any replica symmetry breaking calculation. In fact, if we substitute the approximate solution $\pi_d(t)$ in the right hand side of Eq. (5.60) then we will obtain the successive approximation $\pi_d(t+1)$. Let us write $\pi_d(t) = \pi_d^{(0)} + \rho_d(t)$, where $\pi_d^{(0)}$ is one root of Eq. (5.60) and $\rho_d(t) \ll \pi_d^{(0)}$ a perturbation around it. Substituting this expression into Eq. (5.60) and neglecting quadratic terms in $\rho_d(t)$ we obtain

$$\rho_d(t+1) = -\sum_{d_1} L_{dd_1}\rho_{d_1}(t), \tag{5.62}$$

where

$$L_{dd_1} = (d_1 - 1)p(d_1|d)(1-\pi_d^{(0)})^{d_1-1}. \tag{5.63}$$

Hence, the stability of the root $\pi_d^{(0)}$ is related to the largest eigenvalue Λ_{max} of L_{dd_1}. In this case the solution is asymptotically stable for

$$\Lambda_{max} < 1. \tag{5.64}$$

5.3.1 Uncorrelated graphs

For uncorrelated graphs $p(d_1|d) = q_{d_1}$ and the expressions obtained above become simpler. In this case $\pi_d = \pi$ is independent of d resulting

$$\pi = G_1(1-\pi), \qquad x_c = 1 - G_0(1-\pi) - \frac{\langle d\rangle}{2}\pi^2, \tag{5.65}$$

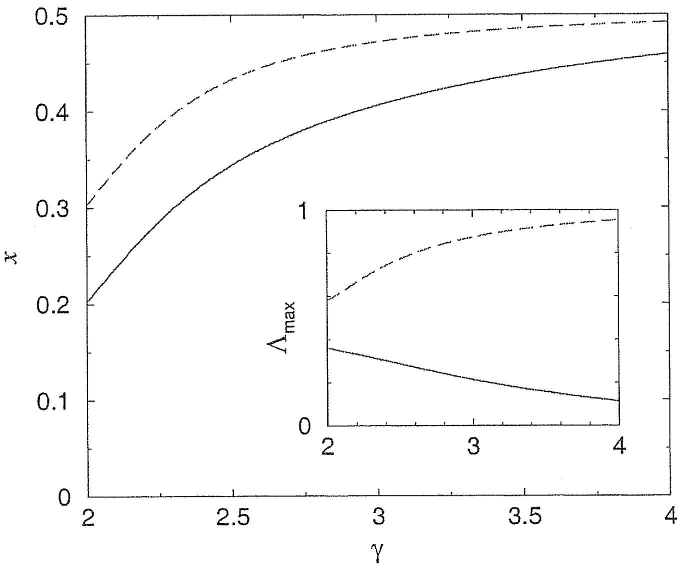


Figure 5.4: Minimal vertex cover for uncorrelated graphs with power law degree distribution $p_d \sim d^{-\gamma}$ with $d_{min} \leq d \leq d_{max}$, for $d_{min} = 1$ (continuous line) $d_{min} = 2$ (dashed line) and $d_{max} = 10^5$. In the inset we plot the largest eigenvalue as a function of γ for the same parameters.

where

$$G_0(x) = \sum_d p_d x^d, \qquad G_1(x) = \sum_d q_d x^{d-1}, \tag{5.66}$$

are the generating functions of p_d and q_d, respectively. Moreover, from Eq. (5.63) it follows

$$L_{dd_1} = (d_1 - 1) q_{d_1} (1-\pi)^{d_1 - 1}. \tag{5.67}$$

Notice that L_{dd_1} is independent of d and the largest eigenvalue can then be computed easily resulting

$$\Lambda_{max} = G_1'(1-\pi). \tag{5.68}$$

In particular let us consider uncorrelated networks with power law degree distribution

$$p_d = \frac{d^{-\gamma}}{\sum_{d'} d'^{-\gamma}}, \qquad d_{min} \leq d \leq d_{max}. \tag{5.69}$$

where d_{min} and d_{max} are the minimum and maximum degree, respectively. In Fig. 5.4 we plot x_c as computed from Eq. (5.65) as a function of γ for $d_{min} = 1$, 2 and $d_{max} = 10^5$. With increasing d_{min} the size of the minimal vertex cover increases but the qualitative picture is the same: x_c decreases with decreasing γ, that is with increasing the probability to find high degree vertices. Moreover, in the inset we plot Λ_{max} as a function of γ. For this magnitude we observe a different behavior dependent on the minimum degree. For $d_{min} = 1$ we have that Λ_{max} decreases with increasing γ while for $d_{min} = 2$ the opposite takes place. In any case $\lambda_{max} < 1$ and, therefore, the replica symmetric solution is stable.

5.3.2 Assortative graphs

To see the influence of correlations, we concentrate on networks having equal degree distributions but different correlation properties. We restrict our attention to scale-free graphs with the degree distribution in Eq. (5.69). For vertex cover, interesting effects are expected to appear for positive correlations, or assortative networks. We therefore consider

$$p(d_1|d) = r\delta_{dd_1} + (1-r)q_{d_1}, \tag{5.70}$$

where $0 \leq r \leq 1$. Varying the parameter r we can interpolate between uncorrelated graphs $r = 0$ and fully assorted graphs $r = 1$. In fact, one can easily check that the parameter r is exactly the correlation coefficient introduced in Ref. [103].

In Fig. 5.5 we show the resulting size of the minimal vertex covers for different values of γ as a function of r. The replica symmetric solution, as obtained from Eq. (5.61) breaks at a certain value of r that depends on γ. There, the solution-space structure changes drastically, from being unstructured, or replica symmetric, in the low correlated case to being clustered, or replica symmetry breaking, for sufficiently high correlations. To check the stability of the solution we have computed the largest eigenvalue Λ_{max} as a function of r. The results are also shown in Fig. 5.5. The replica symmetric solution breaks when $\Lambda_{max} = 1$.

5.3.3 Heuristic algorithms

Once we know the minimal vertex cover size we can test the performance of heuristic algorithms to determine it. In particular we will consider: generalized leaf-removal, high degree vertex removal and a local method. The first two are global algorithms in the sense that we need to know the whole network topology while in the third one we will need only information about nearest neighbors.

The leaf-removal algorithm was proposed by Bauer and Golinelli [17, 16]. It is based on the fact that if we have an edge connecting a vertex A with

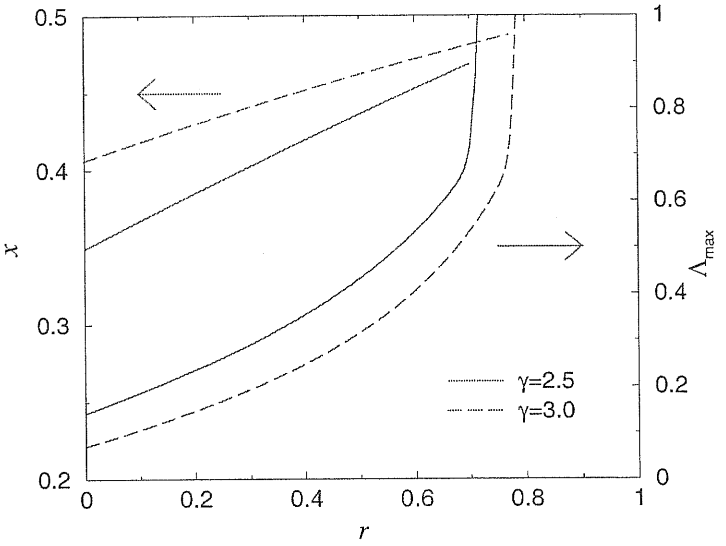


Figure 5.5: Left: Minimal vertex cover size for a network with degree distribution $p_d \sim d^{-\gamma}$ and degree correlations given by Eq. (5.70). The lines correspond to the analytical solution for $\gamma = 2.5$ (continuous line) and $\gamma = 3.0$ (dashed line). The curves stop at the point where replica symmetric solution breaks. Right: Largest eigenvalue Λ_{max} as a function of r for the same values of γ.

degree 1 and a vertex B with degree $d \geq 1$ then covering vertex B is better or equally better than covering vertex A. Thus, one iteratively select a vertex with degree one, cover its neighbor and removes both vertices together with their incident edges. Moreover, all vertices with current degree zero are also removed from the graph. The procedure stops when there are no more leafs. If the remaining graph is empty then all edges will be covered and the fraction of vertices covered is the minimal vertex cover size. On the contrary, if some loops remain then we need a generalization of the algorithm to cover the remaining edges. Let S_{core} be the fraction of vertices in the graph after the leaf-removal algorithm stops. Bauer and Golinelli [16] have shown that for random graphs with $\langle d \rangle < e$ the core size goes to zero when $N \to \infty$. On the contrary for $\langle d \rangle > e$ S_{core} is finite. Later, Weigt [148] pointed out that this transition for $S_{core} = 0$ to $S_{core} > 0$ corresponds with the replica symmetry breaking.

To determine the vertex cover size in the region $S_{core} > 0$ Weigt [148] introduced a generalization of the leaf-removal algorithm that gives an upper

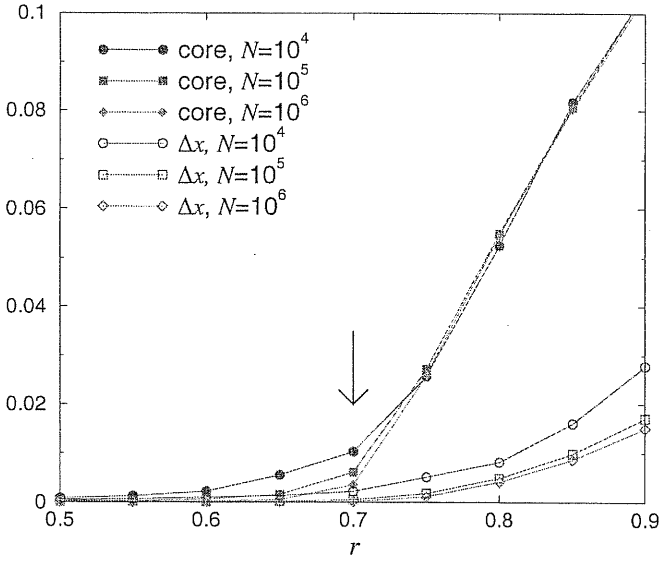


Figure 5.6: The core size S_{core} and error Δx for graphs with degree distribution and correlations given by Eq. (5.69) and (5.70) as a function of r for $\gamma = 2.5$. The networks were generated using a modification of the Molloy and Reed algorithm [98] using $d_{min=1}$ and d_{max} limited by the graph size N. The arrow marks the value of r above which the replica symmetric solution become unstable.

bound for the minimal size. It is defined as follows.

- *Generalized leaf-removal*: Select the vertex of minimal current degree from the graph and cover all its neighbors. Then remove the considered vertices together with their incident edges. This step is iterated until the full graph is removed.

If, for some graph, this algorithm stops without having ever chosen vertices of current degree $d \geq 2$, then the algorithm is identical to leaf-removal and, therefore, the constructed vertex cover is minimal. Overestimations may appear if the algorithm is forced to select also vertices of higher degree $d \geq 2$, where the error can be at most $d - 1$. Thus, summing $(d - 1)(1 - \delta_{d,0})/N$ over all iteration steps, we get an upper bound Δx on the total error made in estimating x_c using the above heuristic algorithm [140]. If Δx goes to zero in the large-N limit, the algorithm has consequently constructed an almost minimal vertex cover.

The replica symmetry breaking was also found above for random graphs with degree correlations given by Eq. (5.70). Hence, we expect that the leaf-removal algorithm is able to find the minimal vertex cover size for low degree correlations but it can fail for strongly correlated graphs. To check this hypothesis we have generated random networks, with the degree distribution in Eq. (5.69) and d_{max} limited by the graph size and computed the vertex cover size x, core size S_{core} and error Δx using the generalized leaf-removal algorithm. The results for $\gamma = 2.5$ are shown in Fig. 5.6. Below the replica symmetry breaking point both S_{core} and Δx goes to zero with increasing graph size. Thus, as for random graphs, in the replica symmetric region the leaf-removal algorithm gives the minimal vertex cover size. On the contrary, above the replica symmetry breaking point both S_{core} and Δx becomes finite and therefore we cannot guarantee that the generalized leaf-removal algorithm yields the minimal vertex cover size.

In power law networks the existence of an appreciable probability to find high degree vertices suggest us that covering the high degree vertices can be also satisfactory. Thus, we propose the following heuristic algorithm for leaf removal.

- *Max-d-removal*: Select a vertex of current maximal degree, cover it and remove all its incident edges. Continue until the full graph is removed.

In general one can select the vertex to be covered with a probability

$$w(d_i)/\sum_j w(d_j)$$

as proposed by Weigt [148] but we have found that choosing a vertex with maximal degree is the must efficient way. As it is shown in Fig. 5.7 the max-d-removal algorithm is nearly as good as the generalized leaf-removal.

In general the structure of real networks is only partially known and therefore global algorithms like the generalized leaf-removal or the max-d-removal become useless. In this case we need local algorithms that use information about the structure of the network in their neighbor to decide, locally, a partially efficient vertex cover. For instance, vertex covers have found applications in monitoring the Internet traffic [26] and in denial of service attack prevention [120]. However, the structure of the router level graph representation of the Internet is far for being complete. To construct the vertex cover in this case we propose the following procedure.

- *Local algorithm*: For each edge cover the ending vertex with the largest degree, if they are equal chose one of them at random.

Thus, each vertex with degree d will be covered if at least one of its neighbors

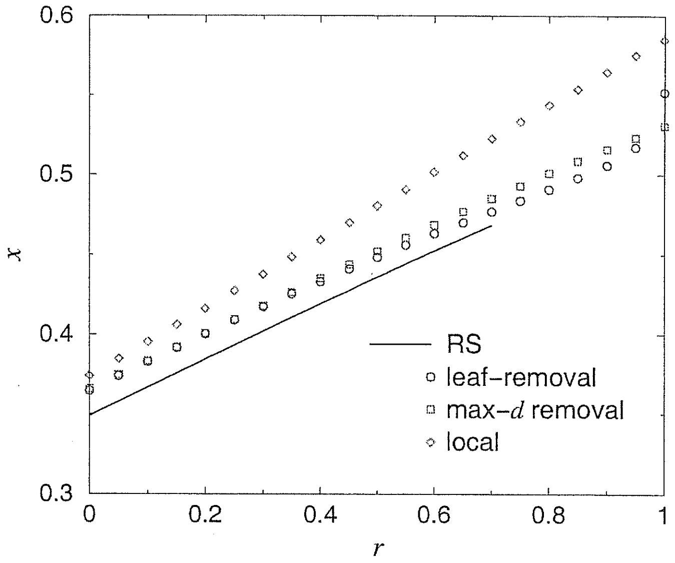


Figure 5.7: Vertex cover size obtained from different heuristic algorithms, for graphs with degree distribution and correlations given by Eq. (5.69) and (5.70) as a function of r for $\gamma = 2.5$. The networks were generated using a modification of the Molloy and Reed algorithm [98] using $d_{min=1}$ and d_{max} limited by the graph size N. The replica symmetric solution for $d_{max} \gg 1$ is also shown for comparison.

has degree $d_1 < d$, *i.e.*

$$x = 1 - \sum_d p_d \left[\sum_{d_1 > d} p(d_1|d) + \frac{1}{2} p(d|d) \right]^d \tag{5.71}$$

where the term $\frac{1}{2}p(d|d)$ takes into account that in case the degrees are equal the selection is made at random. Notice that Eq. (5.71) is valid for any graph because the decision to cover each edge depends only on the degree of the ending vertices.

Surprisingly, this simple algorithm have a good performance as its is shown in Fig. 5.7. For a power law degree distribution with $\gamma = 2.5$ and the assortative degree correlations considered above it overestimate the vertex cover size in less than 10% and similar results are obtained for other values of γ. Moreover, it performs better the slower is the degree of correlations. This last observation is of great importance because we have shown that many real networks exhibit negative correlations and, therefore, the performance of the local algorithm will in these cases be even better. In fact, using this local

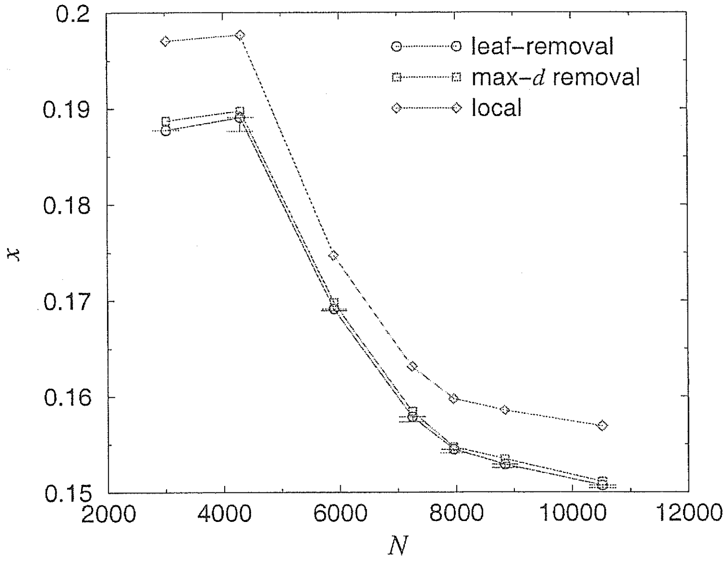


Figure 5.8: Vertex cover size of the AS graph representation of the Internet as a function of the number of vertices N using different heuristic algorithms. The size of the error bars is given by Δx as defined in the text.

Graph	N	gen. leaf-removal	max-d removal	local algorithm	Δx
AS	10515	0.1507	0.1511	0.1569	0.0003
Router	228298	0.286	0.289	0.313	0.003
Math	78837	0.51	0.51	0.53	0.1

Table 5.1: Vertex cover size for different real graphs (see Tab. 2.1) and obtained using different heuristic algorithms

algorithm, we has computed the vertex cover of the AS graph representation of the Internet obtaining that in this case it overestimates the minimal vertex cover size in less that 5%, as it is shown n Fig. 5.8. However, for real graphs with assortative correlations, like the mathematical co-authorship and the Internet router level graphs, the error becomes larger. The vertex cover size for these graphs is shown in Tab. 5.1, as computed from the three heuristic algorithms discussed above. As it can be seen, for the co-autorship graph (Math) the error Δx is of the order of the vertex cover size.

5.4 Conclusions

The existence of a finite amount of random mixing of the connections between vertices is sufficient to make the graph robust under vertex or edge removal provided $\langle d^2 \rangle \to \infty$. Assortative correlations makes the situation even better, they can lead to a graph robust to random damage even with a finite second moment of the degree distribution. However, the solution of optimization problems become harder. On the contrary, disassortative correlations compete again the formation of the giant component and can make a graph fragile even with a divergent second moment. Moreover, the solution of optimization problems becomes easier.

Chapter 6

Assigning proteins functions from protein-protein interaction data

The determination of proteins' activity and functionality in cells is a costly task requiring an extensive biochemical analysis. The inevitably faster techniques for genes and protein sequencing has largely increased the number of known protein sequences but the determination of their structures and functions proceeds at a much slower pace. Furthermore, in the last few years, the possibility to study organisms on a genome or proteome wide-scale has radically changed the approach to the problem at hand. The sequencing of entire genomes and the possibility to access gene's co-expression patterns have moved the attention from the study of single proteins or small complexes to that of the entire proteome (see [67] for an historical perspective). In this context, the search for reliable methods for proteins' function assignment is of uttermost importance. These methods ought to be useful both to complement experiment of functional genomics and to correct possible errors due to the experimental method itself.

Previous approaches to deduce the unknown function of a class of proteins have exploited sequence similarities with proteins of known function or clustering of co-regulated genes [154, 62]. It has been noted [76], however, that the correlation between sequence and function is far less evident than that between sequence and structure. A different perspective has been proposed by Pellegrini *et al.* [119]. They explored the hypothesis that functionally linked proteins (those participating in the same metabolic pathway, for example) have evolved in a correlated way and , therefore, very likely have the same phylogenetic profile. Uetz [132] and Ito [70] adopted a two-hybrid assay technique to map the pairwise interactions among proteins of

Saccharomyces Cerevisiae and used such map to classify proteins of unknown function depending on the known functions of the partners directly interacting with them. This relies on the assumption that if two proteins interact, then they are most likely participating to some common activity and therefore they share at least one functional class. Similar methods exploiting this idea have been presented in [125, 65]. More recently, two different groups [56, 66] have envisaged a strategy to identify complexes composed of three or more proteins. A function was assigned to all the proteins which participate in a complex with at least a component of known function. On the other hand, two-hybrid experiments are known to be prone to false positives and it is hard to establish to which extent protein-protein interaction networks can be considered complete and error free. It would be therefore highly desirable to dispose of a general approach which allows not only to make predictions for otherwise uncharacterized proteins, but that, at the same time, addresses the reliability/robustness of the prediction itself upon eventual errors or incomplete informations.

Here we propose a general and flexible approach to make function prediction once the map of physical interactions between proteins and the functional classification of them is only partially given. The strategy is based on the assumption that "interaction implies a common function", and optimizes the number of shared functionalities among *all* interacting proteins, with the constraints imposed by the whole network of interactions and by the subset of proteins with already assigned functional classes. With this method the whole network of interactions through its complex topology is taken into account and it is worth noting that the study of the network topology itself offers unexpected insights on the functionality of the relative organism as a whole [143, 72].

6.1 Protein interaction network

Protein-protein interactions are intrinsic to virtually every cellular process ranging from DNA replication, transcription, splicing and translation, to secretion, cell cycle control, intermediary metabolism, formation of cellular macro-structures and enzymatic complexes. The formation of large cellular structures such as the cytoskeleton, the nuclear scaffold, and the mitotic spindle result from complex interactions between proteins. Relatively smaller structures such as nuclear pores, centrosomes and kinetochores are beginning to be characterized and, in each case, protein-protein interactions seem to play a crucial role.

Apart from the evident structural requirements provided by a plethora of

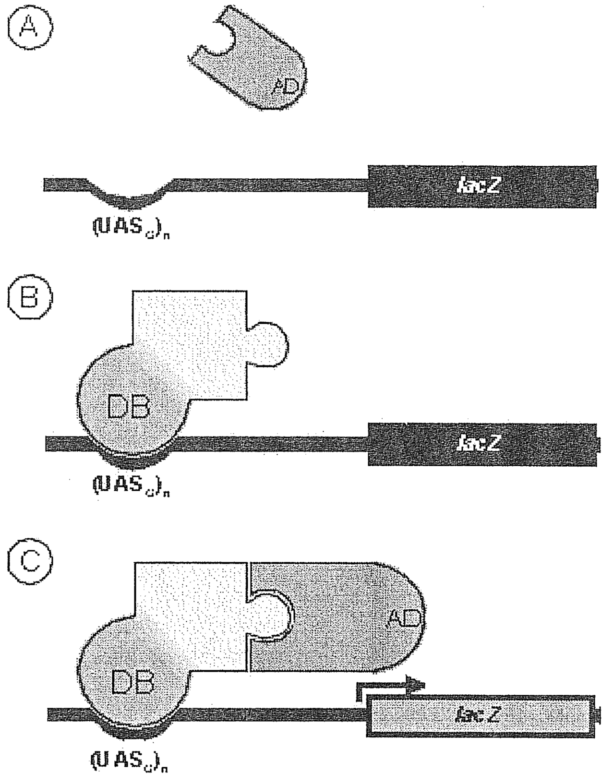


Figure 6.1: Principle of the two-hybrid experiment. (A), (B) Two prove proteins, one containing the DNA-binding domain (DB) and one that contains an activation domain (AD), are co-transfected into an appropriate host strain. If the proteins interact the DB and AD are brought into proximity and can activate the transcription of reporter genes (here *LacZ*). Reprinted from [133].

protein-protein interactions, there are a large number of transient protein-protein interactions that control and regulate a large number of cellular processes. All modifications of proteins involve such transient protein-protein interactions. Indeed kinases, phosphatases, glycosyl transferases, acyl transferases and proteases interact only transiently, i.e. for a limited period of time, with their protein substrates. Such protein-modifying enzymes encompass a large number of fundamental processes such as cell growth, the cell cycle, metabolic pathways and signal transduction. Surprisingly, very large protein complexes also mediate many of these enzymatic activities.

In general, assemblies of proteins have been analyzed using two complementary approaches: the biochemical and the genetic. Apart from these

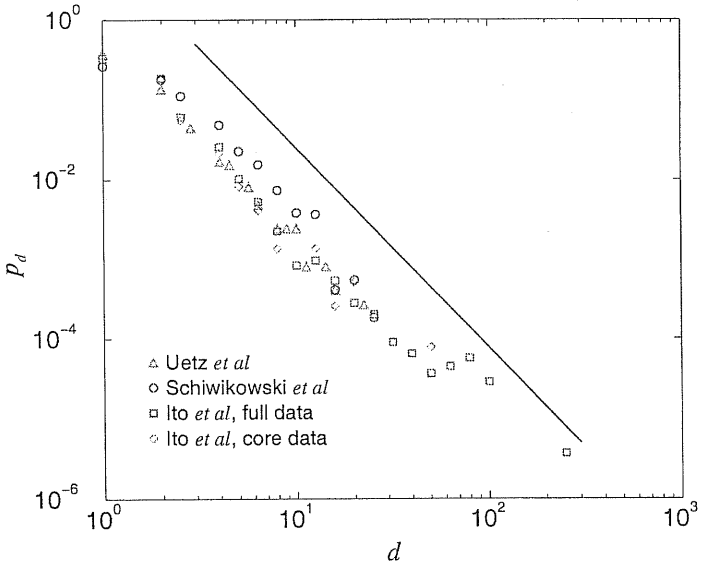


Figure 6.2: Degree distribution of the protein interaction graph. The solid line is a power law decay $p_d \sim d^{-\gamma}$ with $\gamma = 2.5$.

two methods, a new technology has been developed during the past decade [88]. This technique, entitled "two-hybrid" or "interaction trap", enables not only the identification of interacting partners but also the characterization of known interaction couples and even embodies the technological means to manipulate protein-protein interactions. The two-hybrid technique exploits the fact that the DNA-binding domain of the transcription factor GAL4 is incapable of activating transcription unless physically, but not necessary covalently associated with an activating domain (see Fig 6.1).

We have analyzed the graph representing the protein interaction network (PIN) of *Saccharomices Cerevisiae* [125, 70] as obtained from different two-hybrid experiment sets. The first data set reported by Uetz *et al* [132] contains 957 identified interactions (edges) among 1004 proteins. The second data set reported by Schwikowski *et al* [125] is composed 2238 identified interactions between 1825 proteins. On the other hand, Ito *et al* [70] have reported two data sets. A full data set with all the 4549 interactions they detected among 3278 proteins and a core data set with less uncertainty containing 841 interactions among 997 proteins. The degree distribution of these data sets is shown in Fig. 6.2. As it can be seen, there is an appreciable overlap between the degree distribution obtained form the three different sources. Moreover, they can be fitted by a power law decay $p_d \sim d^{-\gamma}$ yielding a power

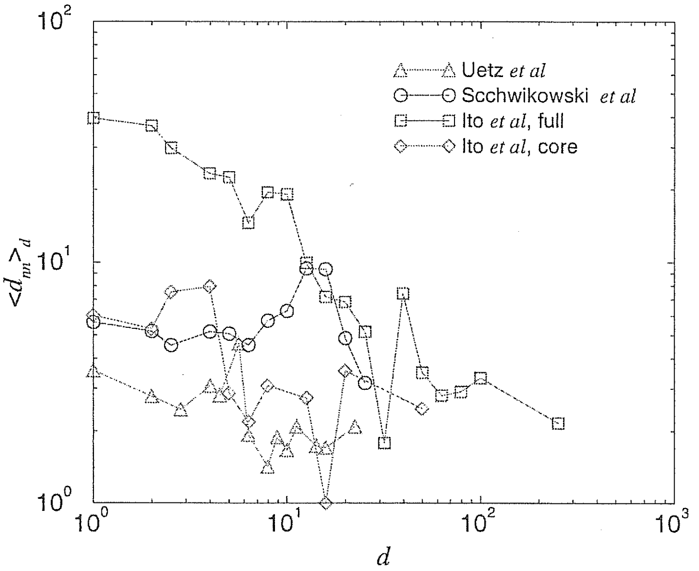


Figure 6.3: Average nearest-neighbors degree as a function of the vertex degree for the protein interaction graph.

law exponent $\gamma \approx 2.5$ in agreement with previous reports [144, 72].

Nevertheless, the plot of the average degree among nearest-neighbors in Fig. 6.3 reveals some visible differences. The full data set of Ito *et al*, the largest among them, exhibits a stronger decreasing dependency with increasing vertex degree than the rest. Hence, we do not know up to what extent this data is reliable and, therefore, in the following we will use the second largest data set, that reported by Schwikowski *et al* [125].

6.1.1 Comparison with the coupled duplication divergence model

Proteins are divided in families according to their sequence and functional similarities [63, 131]. The existence of these families can be explained using the evolutive hypothesis that all proteins in a family evolved from a common ancestor [113]. This evolution is thought to take place through gene or entire genome duplications, gaining redundant genes. After the duplication redundant genes diverge evolving to perform different biological functions. According to the classical model [113] after duplication the duplicate genes have fully overlapping functions. Later on, one of the copies may either become nonfunctional due to degenerative mutations or it can acquire a novel beneficial function and become preserved by natural selection. In a more recent framework [54] it is proposed that both duplicate genes are subject to degenerative mutations loosing some functions but jointly retaining the full set of functions present in the ancestral gene.

The evolution of the genome can be translated into the evolution of the protein-protein interaction network where each vertex represents the protein expressed by a gene. After gene duplication both the expressed proteins will have the same interactions. This corresponds to the addition of a new vertex in the network with edges pointing to the neighbors of its ancestor. In addition positive and negative mutations can be modeled by the creation and lost, respectively, of the edges leading to the divergence of the duplicates. These are the main ingredients of the coupled duplication divergence model discussed in the fourth chapter and defined as follows. At each time step a vertex is added according to the following rules

- *Duplication*: a vertex i is selected at random. A new vertex $i\prime$ with an edge to all the neighbors of i is created. With probability q_v an edge between i and $i\prime$ is established (self-interacting proteins).
- *Divergence*: for each of the vertices j connected to i and $i\prime$ we choose randomly one of the two edges (i, j) or $(i\prime, j)$ and remove it with probability $1 - q_e$.

The peculiarities of the coupled duplication-divergence model manifest quantitatively in different features characterizing the topology of the protein interaction network. Among these the tendency to generate biconnected triplets and quadruples of vertices. These are sets of vertices connected by a simple cycle of edges, thus forming a triangle or a square. In the coupled duplication divergence model triangle formation is a pronounced effect since with probability $q_v q_e$ the duplicating genes and any neighbor of the parent gene will form a new triangle. Analogously, duplicating genes and any couple of neighbors of the parent gene will form a new square with probability q_e^2.

An indication of triangles formation in networks is given by the clustering coefficient $C_\triangle = 3N_\triangle/N_\wedge$ [110] where $N_\triangle$ is the number of biconnected triplets (triangles) and $N_\wedge$ is the total number of simply connected triplets. Similarly it is possible to define the square coefficient $C_\square = 4N_\square/N_\sqcap$, with $N_\square$ the number of squares in the network and $N_\sqcap$ the number of simply connected quadruples. By measuring these quantities in the yeast *Saccharomices Cerevisiae* protein interaction network [125], we obtain $C_\triangle = 0.23$ and $C_\square = 0.11$. These values are one order of magnitude larger than those obtained for a scale-free random graph and with other growing network models, for which it has been shown that the clustering coefficient is algebraically

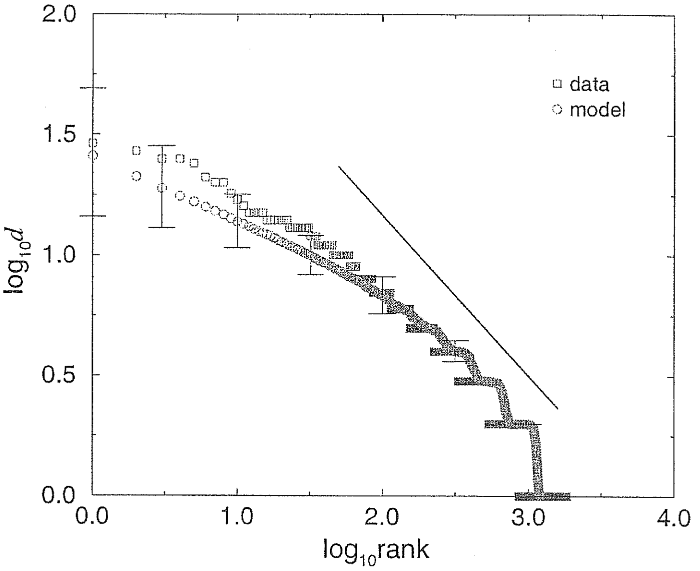


Figure 6.4: Zipf plot of vertex degrees for the *Saccharomices Cerevisiae* protein interaction network [125] and the coupled duplication-divergence model with $q_v = 0.1$, $q_e = 0.3$ with $N = 1825$. The vertex degree d is plotted as a function of its rank in decreasing order of d. Error bars represent standard deviation on a single network realization. The straight line is a power law with exponent $1/(1-\gamma)$ with $\gamma = 2.5$, which will correspond to a power law connectivity distribution $p(k) \sim k^{-\gamma}$.

decaying with the network size [3]. On the contrary, the coupled duplication-divergence model shows clustering coefficients saturating at a finite value, and it is possible to tune the parameters q_v and q_e in order to recover the real data estimates, keeping the average degree as that of the PIN $\langle d\rangle \approx 2.4$. A reasonable agreement with the values obtained for the real PIN is found when $q_v \simeq 0.1$ and $q_e \simeq 0.3$, which yield networks with $C_\Delta = 0.10(5)$ and $C_\Box = 0.10(2)$. The value of q_v obtained in this way is close to the fraction of self-interacting proteins reported for *S. Cerevisiae* (0.04) [132]. Thus, considering that q_v is an effective parameter that takes into account self-interactions but that may also include other effects, the agreement is very good.

Noticeably, for these values of the parameters the coupled duplication-divergence model generates networks where other quantities are in good agreement with those obtained from experimental data. A pictorial representation of this agreement is provided in Fig. 6.4, where we compare the Zipf plot of the degree obtained from 10^3 realizations of the coupled duplication-

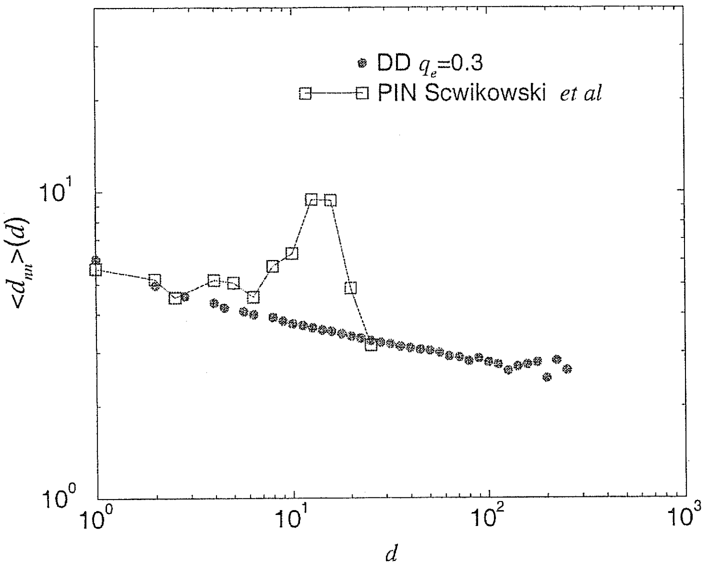


Figure 6.5: Comparison of the average nearest-neighbors degree as a function of the vertex degree for the protein interaction graph and the coupled duplication-divergence model.

divergence model with optimized q_v and q_e and that of the yeast *Saccharomices Cerevisiae* protein interaction network. The generated networks are composed by $N = 1825$ vertices as for the real data. The agreement is very good, considering the relatively large statistical fluctuations we have for this network size. Error bars on the numerical data refer to statistical fluctuations on single realizations. It is worth noticing, that despite the evident multi-fractal nature of the coupled duplication-divergence model, for a single realization of size consistent with that of the protein interaction network, the intermediate d behavior can be approximated by an effective algebraic decay with exponent $1/(\gamma - 1)$ with $\gamma \simeq 2.5$ as found in Ref.[72]. However, the plot in Fig. 6.4 shows a curvature that deviates form the algebraic behavior, evidencing the multi-fractal nature of the degree distribution. Moreover, using the same parameters we also found a reasonable agreement with the degree correlations, as it is shown in Fig. 6.5 for the average nearest-neighbor degree as a function of the vertex degree. The different for the largest degrees is probably due to finite size effects.

The existence of a multi-fractal behavior does not change, at least qualitatively, the main results obtained in the previous chapter. To show that we examine the behavior of the coupled duplication divergence model under ran-

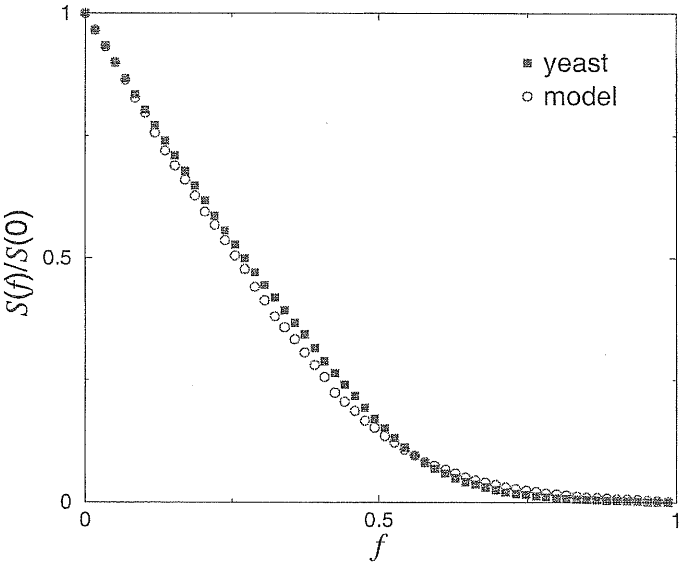


Figure 6.6: Relative size of the giant component after a fraction f of the vertices is removed for the coupled duplication divergence model and the yeast protein interaction network.

dom vertex removal and compare it with those obtained for the yeast protein interaction network. Resilience to damage is indeed considered an extremely relevant property for a network. From an applicative point of view it gives a measure of how robust a network is against disruptive modifications and how far one can go in altering it without destroying its connectivity and therefore functionality [74, 72]. In the random deletion process (site percolation problem) a fraction f of vertices and their incident edges is removed. Fig. 6.6 shows the relative size of the giant component versus the fraction of removed vertices for the yeast protein interaction network and graphs generated using the coupled duplication-divergence model, using the model parameters obtained above. As in pure power law graphs, the tolerance to damage is determined by the scale-free nature of its multi-fractal distribution and the obtained curves are in very good agreement with the corresponding ones for the yeast data.

6.2 Global optimization method

Once we know the topology of the graph substrate let us introduce our protein function assignment method. Suppose that a subset of the proteins are

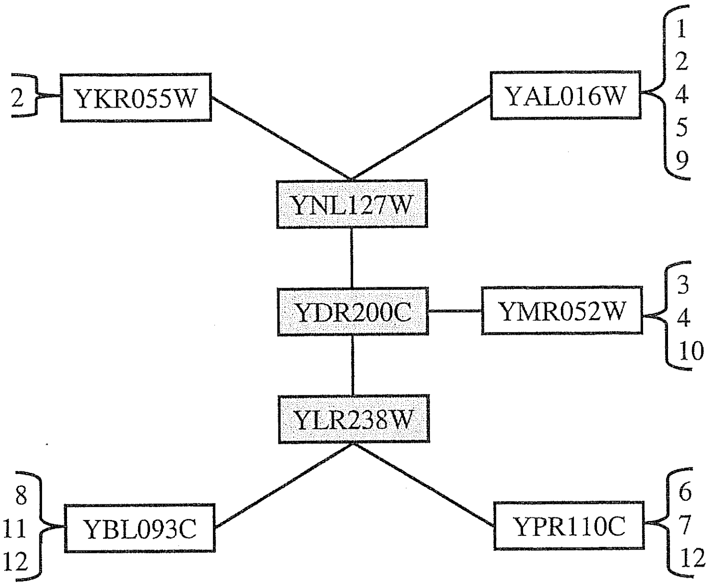


Figure 6.7: Schematic representation of the protein function assignment method. Sub-graph of the protein-interaction network of the yeast *Saccharomices Cerevisiae*. Proteins in gray boxes are unclassified (unknown function) while the others are classified proteins with the functions within the brackets and labelled according to the following criteria: 1- cell growth; 2- budding, cell polarity and filament formation; 3- pheromone response, mating-type determination, sex-specific proteins; 4- cell cycle check point proteins; 5- cytokinesis; 6- rRNA synthesis; 7- tRNA synthesis; 8- transcriptional control; 9- other transcription activities; 10- other pheromone response activities; 11- stress response; and 12- nuclear organization. Given one of these proteins of unknown function if we take as a prediction the function that appears more in the neighbor proteins of known function then we obtain the following classification (from top to bottom) YNL127W (2), YDR200C (3,4,10) and YLR238W (12). Our method, however, considers also the interactions among unclassified proteins. If we iterate once more the "majority rule" by taking into account the interactions between the three unclassified proteins, we obtain the following classification YNL127W (2,4), YDR200C (3,4,10) and YLR238W (12). In this way we find another possible function for YNL127W. This is actually the spirit of our method, we take advantage of the prediction we are making for proteins of unknown function and apply a global optimization method.

known to belong to one or more functional classes. A function σ_i, chosen among all F possible functions is assigned to each unclassified protein i. F is the total number of functions we consider and the finer the definition of function the greater is F. Assigning to an unclassified protein the most common function(s) present among the classified interacting proteins ("majority rule" assignment) is rather straightforward and it is the method adopted in [125, 65]. In most cases [93], however, very few unclassified proteins have more than one interacting protein with known function. In addition, in these few lucky cases, the interacting proteins with known functions do not usually have shared functionalities (see Fig.(6.7)). In this perspective, the use of the "majority rule" assignment of the functionality is rather unsatisfactory since *all links between proteins with unknown function are completely neglected.* This implies a very partial use of the knowledge of the protein-protein interaction network. Most important, the final configuration of assigned functions to unclassified proteins ought to be consistent with the rules used to determine the functions themselves. In other words to an unclassified protein with one or more unclassified partner(s) must be assigned functions consistently with the functions assigned to the unclassified partners themselves. This points out to a method in which unknown proteins influence the "majority rule" assignment in a self-consistent way once their functions have been selected.

Our functional prediction strategy is based on a global optimization principle: a score or energy is associated to any given assignment (configuration) of functions for the whole set of unclassified proteins. The score is lower in configurations that maximize the overlap of functions in interacting proteins. The new ingredient is that the contribution to the total score of a given functional assignment of an unclassified protein is computed as the number of classified and unclassified neighbor proteins with that function.

To each unclassified protein i a function σ_i is assigned, chosen among the F possible ones in order to globally minimize the following score function:

$$E = -\sum_{i<j=1}^{n} J_{ij}\delta_{\sigma_i,\sigma_j} - \sum_{i=1}^{n} h_i(\sigma_i). \tag{6.1}$$

where $J_{i,j}$ is the adjacency matrix of the interaction network for the unclassified proteins ($J_{i,j}$ is equal to 1 if protein i and j interact and are unclassified, 0 otherwise), $\delta_{i,j}$ is the discrete delta function and $h_i(\sigma_i)$ is proportional to the number of classified partners of protein i with function σ_i. The "majority rule" of [132, 70] corresponds to minimize solely the second term on the *r.h.s.* of equation (6.1). The above can be achieved with local methods (*i.e.* considering successively and independently each protein). In our method, the contribution to the total score of assigning a protein i to functional class σ_i depends also on the assignment made to all other proteins, resulting in a much harder computational task. The advantage is that the underlying requirements that "interaction requires a common function" is applied also to interactions between previously unclassified proteins, that are completely ignored with the "majority rule".

Hence, the determination of the functions of all unclassified proteins in the network becomes a global optimization problem and can not be done on the basis of the local environment only. Finding the optimal function assignment corresponds to determine the minimal energy (ground state) of what in statistical mechanics is known as the Pott's model [152] with a random field, the latter represented by the proteins with known function. In general the resulting computational problem is frustrated. It is, in fact, generally impossible to satisfy all the constraints imposed by classified proteins on their interacting unclassified partners. This lead to a multiplicity of optimal solutions which contains the minimal amount of frustration. The existence of multiple solutions allows the objective assignment of multiple functions to most unclassified proteins (see Fig. 6.7). Depending on the complexity of the underlying graph and on the boundary conditions, the energy minimization represents a hard computational task.

To overcome the computational difficulties and find the configuration or configurations that minimize E we perform a simulated annealing [123] introducing an effective temperature T. We start with an initial random configuration $\{\sigma_i\}$. Then, at each Monte-Carlo step, we select one protein at random and change its state from σ_i to σ_i', where σ_i' is selected at random among the possible states of protein i with the constraint $\sigma_i' \neq \sigma_i$. We then compute the energy difference $\Delta E = E' - E$ between these two configurations. If $\Delta E \leq 0$ we accept the new configuration. If $\Delta E > 0$ we accept the new configuration with probability $\exp(-\Delta E/T)$. Otherwise we return to the original configuration. This Monte-Carlo step is repeated until E reaches an stationary value and, when this happens, T is decreased by a tiny amount. These two process, equilibration at a given T and decrease of T is repeated until the protein states remain unchanged for a sufficiently long time. At the end the protein states are our predicted functional classification. Since the minimum energy solution is not unique the simulating annealing process was repeated several times starting from different initial configurations. Finally we compute the fraction of times p_{is} the protein i was observed in the final state s, which give us an estimate of the probability that protein i belongs to the functional classification s.

In instances of this type simulated annealing technique is an appropriate tool, allowing to obtain the optimal solutions. Indeed, the optimization

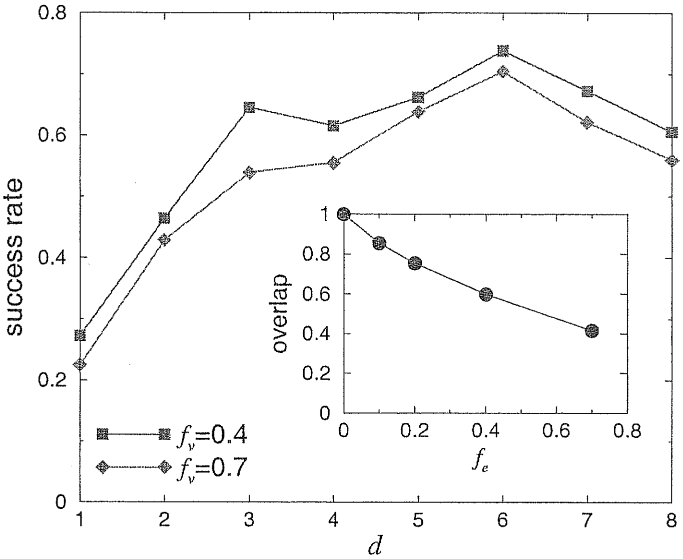


Figure 6.8: Prediction accuracy of the protein function assignment method. The success rate of our method after a fraction f_v of classified proteins have been set unclassified. Each point represents the probability that the functional classification of these proteins, as predicted by our method, coincides with their real classification as a function of the number of their interacting partners d and for two values of f_v. The inset shows the overlap ($\Theta_i(f_e)$ average over all unclassified proteins) between the functional predictions obtained using two different interaction networks, the original and one with a fraction f_e of its links rewired, as a function of f_e.

procedure is repeated several times to account for the non uniqueness of the optimal configurations and a prediction for the functional classification is finally made by taking those functions that, for each unclassified protein, occurred more often in the whole set of simulated annealing processes .

The dataset used in this work is relatively small and, therefore, simulating annealing techniche should in principle give the exact ground states. In is interesting to know however what will happens when the method is applied to other organisms with a larger protein-protein interaction graph. We do not have any definitive answer but we can make some partial conclusions taking into account our analysis for the vertex covering problem. In particular, we found that in uncorrelated or negatively correlated graphs with power law degree distributions we can solve hard optimization problems using polynomial time algorithms with an error of the order N^{-1}. Fortunately, the protein-protein interaction network, and more generally biological networks

[103], exhibit negative correlations. Hence, we expect that optimization problems on top of them can be solved in polynomial time exploiting the graph topology.

6.3 Prediction accuracy

We have applied our functional-prediction method to the yeast *Saccharomices Cerevisiae* protein-protein interaction network. The interaction data was obtained from Ref. [125], it contains $N = 1826$ proteins with $E = 2238$ identified interactions, and some of its topological properties were discussed above. The functional classification was obtained from the MIPS database [1]. The MIPS finer classification scheme contains $F = 424$ functional categories, plus two categories for proteins with no assigned function: "CLASSIFICATION NOT YET CLEAR-CUT" and "UNCLASSIFIED PROTEINS". The data contains $n = 441$ proteins in these two last categories [125]. We used our global optimization method obtaining functional assignments for all the proteins listed within these two categories.

6.3.1 Protein function uncertainty

In order to asses the reliability and robustness of our method we have done several tests. As a first test we consider a single protein in the classified proteins set as unclassified. We then assign a function to this protein and define a successful prediction if the most probable function obtained using our method is in the list of the functions actually performed by that protein. By repeating this experiment on several different classified proteins, we measured the success rate in recovering the actual protein functions. This procedure can be repeated by progressively loosing information by ignoring the functional classification of a fraction f_v of classified proteins and adding them to the set of unclassified proteins. In this way, we can get a quantitative estimate of the reliability of our predictions as a function of the amount of available information on the network. Fig. 6.8 shows the percentage of successful predictions as a function of the degree of the proteins for different values of f_e. Some interesting conclusions can be drawn from this test. First we note that the prediction quality for poorly connected vertices, degree 1 and 2, decreases to just 30%. On the contrary, in the case of unclassified proteins with degree larger than 2, even with the loss of a substantial part of the information (up to $f_v = 0.4$) a correct prediction can be made between 60%-70% of the cases, quite independently of the degree of the protein involved in the prediction. This implies that the availability of further information regarding the number of classified proteins would not radically change the rate of prediction success on the proteins with a higher number of interacting partners.

6.3.2 Protein interaction uncertainty

A second test concerns the presence of errors in the protein network. It is known that protein-protein interactions obtained from two hybrid experiments contain an amount of false positives and negatives. The effect of this uncertainty can be modeled by re-wiring a certain fraction f_e of protein interactions. More precisely, with a probability f_e, each reported interaction is ignored and a new interaction is randomly drawn between two proteins that do not interact according to the available data. In this way we obtain a new network with a certain degree of similarity, depending upon f_e, with the original one. We implement our method on this modified network and determine new predictions for the unclassified proteins. A quantitative comparison with the predictions made using the original network is provided by the overlap function defined as follows. Let $p_{is}(f_e)$ the probability, as obtained from our method implemented on the network with a fraction f_e of re-wirings, that the unclassified protein i belongs to the functional classification s. The case $p_{is}(0)$ then corresponds to the functional classification obtained using the original network. The overlap between the protein function prediction in the two networks can be expressed as

$$\Theta_i(f_e) = \sum_{s=1}^{F} \sqrt{p_{is}(0)p_{is}(f_e)}, \tag{6.2}$$

that equals 1 when $p_{is}(f_e) = p_{is}(0)$ for all s. We computed the average of $\Theta_i(f_e)$ restricted to unclassified proteins with d interacting partners, obtaining that it does not vary too much with the node degree. In the inset of Fig. 6.8 we plot the average of $\Theta_i(f_e)$ over all unclassified proteins as a function of f_e. For 10% of errors ($f_e = 0.1$), the overlap is around 0.85%. This shows that a few misplaced interactions due to experimental erroneous results do not preclude an effective evaluation of the proteins' functions. Of course larger levels of errors lower the overlap, signaling that the two networks provide rather different configurations of functional assignment.

6.4 Conclusions

The method we propose appears as a more general tool for the assignment of protein function pointing out that protein-protein interaction data can be an effective framework for deducing the function of unclassified proteins. The method allows also to determine multiple functionalities and takes into account self-consistently the effect of unclassified proteins in the final assignment configuration. Finally, the validity tests performed show that the method tolerates the inherent imperfection and the not complete nature of the protein networks. Since data of protein-protein interactions are accumulating rapidly, we are confident that our method might provide a relevant contribution in obtaining valuable informations from the global-web like view of the protein interactions.

Chapter 7

General conclusions

In this thesis the existence, origin and consequences of degree correlations and hierarchy in complex networks have been investigated. Different metrics were proposed to characterize these correlations and hierarchy. Using them we conclude that, in addition to power law degree distributions, degree-degree correlations and clustering hierarchies are common features of many real networks. Moreover, they can be used to discriminate between different networks that that appear similar simply on the basis of their degree distribution.

As a further step in the understanding of complex networks, one hypothesis for the origin of these degree correlations and clustering hierarchy was analyzed. It was shown that preferential attachment, degree correlations and clustering hierarchy appear naturally in growing graph models with local rules, offering an explanation for the ubiquity of these properties in real graphs.

The properties of models defined on top of graphs with degree correlations were also studied. A Bethe-Peierls and a replica symmetric schemes were developed to compute different magnitudes characterizing these models. A particular attention has been devoted to the problems of dilute percolation and vertex covering. It was obtained that assortative correlations lead to graphs that are more robust under random vertex or edge removal but, at the same time, the problems of combinatorial optimization on top of these graphs become harder. On the contrary, dissasortative correlations make the graph more fragile and facilitates the solution of combinatorial optimization problems.

All these results were applied to the study of the protein-protein interaction network and the protein function prediction. A statistical mechanics model was developed to make protein function predictions using protein interaction data, resulting in a global optimization problem. The reliability of this method was analyzed through its application to the yeast, obtaining satisfactory results. In a more general scope, taking into account that biological networks exhibit disassortative correlations, the results described above indicate that the use of global optimization methods to determine some of their properties is quite efficient, with a better performance than heuristic methods based in simple local schemes.

Bibliography

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

[2] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. In *Proc. 32nd ACM Symp. Theor. Comp.*, pages 171–180, 2000.

[3] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–95, 2001.

[4] R. Albert, H. Jeong, and A.-L. Barabási. The diameter of the world-wide web. *Nature*, 401:130–131, 1999.

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

[6] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small-world networks. *Proc. Nat. Acad. Sci. U.S.A.*, 97:11149–1152, 2000.

[7] P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality: An explanation of the 1/f noise. *Phys. Rev. Lett.*, 59:381–384, 1987.

[8] P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality. *Phys. Rev. A*, 38:364–374, 1988.

[9] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–511, 1999.

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

[11] A.-L. Barabási, R. Albert, and H. Jeong. Scale-free characteristics of random networks: The topology of the world wide web. *Physica A*, 281:69–77, 2000.

[12] A.-L. Barabási, R. Albert, H. Jeong, and G. Bianconi. Power-law distribution of the world wide web. *Science*, 287:2115a, 2000.

[13] A.-L. Barabási, H. Jeong, Z.Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A*, 311:590–614, 2002.

[14] A. Barrat and M. Weigt. On the properties of small-world network models. *Eur. Phys. J. B*, 13, 2000.

[15] M. Barthélémy and L. A. N. Amaral. Small-world networks: Evidence for a crossover picture. *Phys. Rev. Lett.*, 82:3180–3183, 1999.

[16] M. Bauer and O. Golinelli. Core percolation: a new geometric phase transition in random graphs. *Eur. Phys. J. B*, 24:339–352, 2001.

[17] M. Bauer and O. Golinelli. Exactly solvable model with two conductor-insulator transitions driven by impurities. *Phys. Rev. Lett.*, 86:2621–2624, 2001.

[18] J. Berg and M Lassig. Correlated random networks. *cond-mat/0205589.*

[19] H. A. Bethe. *Proc. Roy. Soc. (London) A*, 150:552–575, 1935.

[20] G. Bianconi and A.-L. Barabási. Competition and multiscaling in evolving networks. *Europhys. Lett*, 54:436–442, 2000.

[21] M. Boguña, R. Pastor-Satorras, and A. Vespignani. Absence of epidemic threshold in scale-free networks with connectivity correlations. *cond-mat/0208163.*

[22] B. Bollobás. *Discrete Math.*, 53:1, 1981.

[23] B. Bollobás. *Random Graphs.* London: Academic Press, 1985.

[24] B. Bollobás. Graph theory and networks generation of scale-free graphs. In *School on statistical physics, probability theory and computational complexity*, 2002.

[25] B. Bollobás and O. Riordan. The diameter o a scale-free random graph. *preprint*, 2001.

[26] Y. Breitbart, C.-Y. Chan, M. Garofalakis, R. Rastogi, and A. Silverschatz. In *Proceedings of IEEE INFOCOM, Alaska*, 2001.

[27] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, pages 107–117, 1998.

[28] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer Networks*, 33:309–320, 2000.

[29] A. Broido and K. C. Claffy. In *Proceedings of SPIE International symposium on Convergence of IT and Communication. Denver, 2001.*

[30] G. Caldarelli, R. Marchetti, and L. Pietronero. The fractal properties of internet. *Europhys. Lett.*, 52:386–392, 2000.

[31] D. S. Callaway, J.E. Hopcroft, J. M. Kleinberg, M. E. J. Newman, and S. H. Strogatz. Are randomly grown graphs really random? *Phys. Rev. E*, 64:041902–041909, 2001.

[32] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Network robustness and fragility: Percolation on random graphs. *Phys. Rev. Lett.*, 85:5468–5471, 2000.

[33] J. Camacho, R. Guimerá, and L. A. N. Amaral. Robust patterns in food web structure. *Phys. Rev. Lett.*, 88:228102–228105, 2002.

[34] R. F. Cancho, C. Janssen, and Ricard V. Sol. Topology of technology graphs: Small world patterns in electronic circuits. *Phys. Rev. E*, 64:046119–0461123, 2001.

[35] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. J. Shenker, and W. Willinger. In *Proceedings of IEEE Infocom 2002, New York.*

[36] H. Chou. A note on power-laws of internet topology. *cs.NI/0012019.*

[37] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin. Resilience of the internet to random breakdowns. *Phys. Rev. Lett.*, 85:4626–4628, 2000.

[38] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin. Breakdown of the internet under intentional attack. *Phys. Rev. Lett.*, 86:3682–3685, 2001.

[39] J. Davidsen, H. Ebel, and S. Bornholdt. Emergence of a small world from local interactions: Modeling acquaintance networks. *Phys. Rev. Lett.*, 88:128701–128704, 2001.

[40] M. Argollo de Menezes, C. F. Moukarzel, and T. J. P. Penna. First-order transition in small-world networks. *Europhys. Lett.*, 50:574–579, 2000.

[41] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. Ising model on networks with an arbitrary distribution of connections. *Phys. Rev. E*, 66:016104–016108, 2002.

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

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

[44] S. N. Dorogovtsev and J. F. F. Mendes. Scaling properties of scale-free evolving networks: Continuous approach. *Phys. Rev. E*, 63:056125–056144, 2001.

[45] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Adv. Phys.*, 51:1079–1187, 2002.

[46] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Structure of growing networks with preferential linking. *Phys. Rev. Lett.*, 85:4633–4636, 2000.

[47] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Multifractal properties of growing networks. *Europhys. Lett.*, 57:334–340, 2002.

[48] H. Ebel, L.-I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *cond-mat/0201476.*

[49] Electric Engineering and Computer Science Department, University of Michigan, http://topology.eecs.umich.edu/. *Topology project.*

[50] P. Erdos and A. Rényi. On random graphs. *Publications Mathematicaa*, 6:290, 1959.

[51] P. Erdos and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5:17–60, 1960.

[52] J. W. Essam. Percolation theory. *Rep. Prog. Phys.*, 43:833–912, 1980.

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

[54] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. l. Yan, and J. Postlethwait. Preservation of duplicate genes by complementary, degenerative mutation. *Genetics*, 151:1531–1545, 1999.

[55] M. Garey and D. Johnson. *Computers and Intractability: A Guide to the theory of NP-completeness.* Freeman, San Francisco, 1979.

[56] A. Gavin et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147, 2000.

[57] E. M. Gin, M. Girvan, and M. E. J. Newman. The structure of growing social networks. *Phys. Rev. E*, 64:046132–046139, 2001.

[58] K.-I. Goh, B. Kahng, and D. Kim. Universal behavior of load distribution in scale-free networks. *Phys. Rev. Lett.*, 87:278701–278704, 2001.

[59] K.-I. Goh, B. Kahng, and D. Kim. Fluctuation-driven dynamics of the internet topology. *Phys. Rev. Lett.*, 88:108701–108704, 2002.

[60] R. Govindan and H. Tangmunarunkit. Heuristics for internet map discovery. In *Proceedings of the 2000 IEEE INFOCOM Conference, Tel Aviv, Israel, March*, pages 1371–1380, 2000.

[61] P. Grassberger. On the critical behavior of the general epidemic process and dynamical percolation. *Math. Biosci.*, 63:157–172, 1983.

[62] H. C. Harrington, Rosenow, C., and J. Retief. Monitoring gene expression using dna microarrays. *Curr. Opin. Microbiol.*, 3:285–291, 2000.

[63] S. Henikoff, E. A. Greene, S. Pietrokovski, P. Bork, T. K. Attwoodand, and L. Hood. Gene families: the taxonomy of protein paralogs and chimeras. *Science*, 278:609–614, 1997.

[64] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and Marc Najork. Measuring index quality using random walks on the web. In *Proceedings of the 8th International World Wide Web Conference, Toronto, Canada*, pages 213–22, 1999.

[65] H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, , and T. Tagaki. Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast*, 18:523–531, 2000.

[66] Y. Ho et al. Systematic identification of protein complexes in saccharomycescerevisiae by mass spectrometry. *Nature*, 415:180–183, 2002.

[67] T. C. Hodgman. A historical perspective on gene/protein functional assignment. *Bionformatics*, 16:10–15, 2000.

[68] M. A. Huynen and P. Bork. *Proc. Natl. Acad. Sci. USA*, 95:5849, 1998.

[69] Information Sciences Institute, http://www.isi.edu/div7/scan/. *Mapping the Internet within the SCAN project.*

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

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

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

[73] H. Jeong, Z. Néda, and A.-L. Barabási. Measuring preferential attachment for evolving networks. *cond-mat/0104131.*

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

[75] C. Jin, Q. Chen, and S. Jamin. *INET: Internet topology generators.* Tech. Rep. CSE-TR-433-00, EECS Dept., University of Michigan, 2000.

[76] P. D. Karp. What we do not know about sequence analysis and sequence databases. *Bioinformatics*, 14:753–754, 1998.

[77] J. Kim, P. L. Krapivsky, B. Kahng, and S. Redner. Evolving protein interaction networks. *cond-mat/0203167.*

[78] J. Kleinberg, R. Kumar, P. Raphavan, S. Rajagopalan, and A. Tomkins. The web as a graph: measurements, models, and methods. *Proceedings of the 5th International Conference on Combinatorics and Computing, Tokyo, Japan, 26-28 July*, pages 1–17, 1999.

[79] P. L. Krapivsky and S. Redner. Organization of growing random networks. *Phys. Rev. E*, 63:066123–066136, 2001.

[80] P. L. Krapivsky and S. Redner. A statistical physics perspective on web growth. *Computer Networks*, 39:261–276, 2002.

[81] P. L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Phys. Rev. Lett.*, 85:4629–4633, 2000.

[82] J. Laherrere and D. Sornette. Stretched exponential distributions in nature and economy: fat tails with characteristic scales. *Eur. Phys. J. B*, 2:525–539, 1998.

[83] M. Leone. *tatistical Mechanics of spin systems on diluted random structures.* PhD thesis, International School for Advanced Studies, 2002.

[84] M. Leone, A. Vázquez, A. Vespignani, and R. Zecchina. Ferromagnetic ordering in graphs with arbitrary degee distribution. *Eur. Phys. J. B*, 28:191–197, 2002.

[85] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Berg. The web of human sexual contacts. *Nature*, 411:907–908, 2001.

[86] Lucent Bell Labs, http://www.cs.bell-labs.com/who/ches/map/. *Internet mapping project.*

[87] M. Lynch and A. Force. *Genetics*, 154:459, 1999.

[88] Phizicky E. M. and Fields S. Protein-protein interactions: methods for detection and analysis. *Microbiol. Rev.*, 59:94–123, 1995.

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

[90] R. M. May and A. L. Lloyd. Infection dynamics on scale-free networks. *Phys. Rev. E*, 64:066112–066115, 2001.

[91] A. Medina and I. Matta. *BRITE: a flexiblegenerator of Internet topologies.* Tech. Rep. BU-CS-TR-2000-005, Boston University, 2000.

[92] A. Medina, I. Matta, and J. Byers. On the origin of power laws in internet topologies. *Comp. Com. Rev.*, 30:18–, 2000.

[93] M. L. Meyer, , and P. Hieter. Protein networks-built by association. *Nature Biotech.*, 18:1242–1243, 2000.

[94] M. Mézard and G. Parisi. The bethe lattice spin glass revisited. *Eur. Phys. J. B*, 20:217–233, 2001.

[95] M. Mézard and G. Parisi. The cavity method at zero temperature. *cond-mat/0207121*, 2002.

[96] M. Mézard, G. Parisi, and M. A. Virasoro. *Spin glass theory and Beyond.* World Scientific, Singapore, 1987.

[97] S. Milgram. The small world problem. *Psychology today*, 2:60, 1967.

[98] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 6:161–179, 1995.

[99] Y. Moreno, R. Pastor-Satorras, and A. Vespignani. Epidemic outbreaks in complex heterogeneous networks. *Eur. Phys. J. B*, 26:521–529, 2002.

[100] S. Mossa, M. Barthélémy, H. E. Stanley, and L. A. Nunes Amaral. Truncation of power law behavior in scale-free network models due to information filtering. *Phys. Rev. Lett.*, 88:138701–138704, 2002.

[101] M. Bogu na and R. Pastor-Satorras. Epidemic spreading in correlated complex networks. *cond-mat/0205621.*

[102] National Science Foundation, http://moat.nlanr.net/. *The National Laboratory for Applied Network Research (NLANR).*

[103] M. E. J. Newman. Assortative mixing in networks. *cond-mat/0205405.*

[104] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64:R025102–025106, 2001.

[105] M. E. J. Newman. Scientific collaboration networks: I. network construction and fundamental results. *Phys. Rev. E*, 64:016131–016139, 2001.

[106] M. E. J. Newman. Scientific collaboration networks: Ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64:016132–016139, 2001.

[107] M. E. J. Newman. The structure of scientific collaboration networks. *Proc. Nat. Acad. Sci. U.S.A.*, 98:404–409, 2001.

[108] M. E. J. Newman. Spread and epidemic disease on networks. *Phys. Rev. E*, 66:016128–016138, 2002.

[109] M. E. J. Newman. The structure and function of networks. *to appear in Comp. Phys. Comm.*, 2002.

[110] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distribution and their applications. *Phys. Rev. E*, 64:026118–026135, 2001.

[111] M. E. J. Newman and D. J. Watts. Renormalization group analysis of the small-world network model. *Phys. Lett. A*, 263:341–346, 1999.

[112] M. E. J. Newman and D. J. Watts. Scaling and percolation in the small-world network model. *Phys. Rev. E*, 60:7332–7342, 1999.

[113] S. Ohono. *Evolution by gene duplication.* Springer-Verlag, Berlin, 1970.

[114] C. H. Papadimitriou and K. Steiglitz. *Combinatorial optimization: algorithms and complexity.* Prentice Hall, New Jersey, 1982.

[115] R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Dynamical and correlation properties of the internet. *Phys. Rev. Lett.*, 87:258701–258704, 2001.

[116] R. Pastor-Satorras and A. Vespignani. Epidemic dynamics and endemic states in complex networks. *Phys. Rev. E*, 63:066117–066125, 2001.

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

[118] R. Peierls. *Proc. Camb. Phyl. Soc.*, page 477, 1936.

[119] M. Pellegrini, E. Marcotte, M. J. Thompsom, D. Eisemberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA*, 96:4285–4288, 1999.

[120] A. R. Puniyani and R. M. Lukose. Growing random networks under constraints. *cond-mat/0107391.*

[121] Z. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *cond-mat/0206130.*

[122] S. Redner. How popular is your paper? an empirical study of citation distribution. *Eur. Phys. J. B*, 4:131–134, 1998.

[123] Kirkpatrick S., Gelatt C. D., and Vecchi M. P. Optimization by simulated annealing. *Science*, 220:621–680, 1983.

[124] San Diego Supercomputer Center, http://www.caida.org/home/. *The Cooperative Association for Internet Data Analysis (CAIDA).*

[125] B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. *Nature Biotech.*, 18:1257–1261, 2000.

[126] R. V. Solé and J. M. Montoya. Complexity and fragility in ecological networks. *Proc. R. Soc. London B*, 268:2039–2045, 2001.

[127] G. W. Stewart and J.G. Sun. *Matrix perturbation theory.* Academic Press, San Diego, 1990.

[128] GVU's WWW User Survey. How users find out about www pages. Technical report, Georgia Tech Research Corporation, 1998.

[129] B. Tádic and D. Dhar. Emergent spatial structures in critical sandpiles. *Phys. Rev. Lett.*, 79:1519–1522, 1997.

[130] H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker, and W. Willinger. *Comput. Commun. Rev.*, 32:76, 2002.

[131] R. L. Tatusov, E. V. Koonin, and D. J. Lipman. A genomic perspective on protein families. *Science*, 278:631–637, 1997.

[132] P. Uetz et al. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403:623–627, 2000.

[133] W. Van Criekinge and R. Beyaert. Yeast two-hybrid: State of the art. *Biological procedures Online*, 2:1–38, 1999.

[134] A. Vázquez. Statistics of citation networks. *cond-mat/0105031.*

[135] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani. Modeling of protein interaction networks. *cond-mat/0108043.*

[136] A. Vázquez and Y. Moreno. Resilience to damage of networks with degree correlations. *cond-mat/.*

[137] A. Vázquez, R. Pastor-Satorras, and A. Vespignani. Internet topology at the router and autonomous system level. *cond-mat/020608.*

[138] A. Vázquez, R. Pastor-Satorras, and A. Vespignani. Large-scale topological and dynamical properties of internet. *Phys. Rev. E*, 65:066130–066141, 2002.

[139] A. Vázquez and O. Sotolongo-Costa. Self-organized criticality and directed percolation. *J. Phys. A: Math. Gen.*, 32:2633–2644, 1999.

[140] A. Vázquez and M. Weigt. Computational complexity arising from degree correlations in networks. *cond-mat/0207035.*

[141] A. Vespignani and S. Zapperi. Order parameter and scaling fields in self-organized criticality. *Phys. Rev. Lett.*, 78:4793–4796, 1997.

[142] A. Vespignani and S. Zapperi. How self-organized criticality works: A unified mean-field picture. *Phys. Rev. E*, 57:6345–6362, 1998.

[143] A. Wagner. Robutness again mutations in genetic networks of yeast. *Nature Gent.*, 24:355–361, 2000.

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

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

[146] D. J. Watts. *Small Worlds: The Dynamics of Networks between Order and Randomness.* Princeton University Press, Princeton, New Jersey, 1999.

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

[148] M. Weigt. Dynamics of heuristic optimization algorithms on random graphs. *to appear in Eur. Phys. J. B*, 2002.

[149] M. Weigt and A. K. Hartmann. The number of guards needed by a museum: A phase transition in vertex covering of random graphs. *Phys. Rev. Lett.*, 84:6118–6121, 2000.

[150] M. Weigt and A. K. Hartmann. Statistical mechanics perspective on the phase transition in vertex covering finite-connectivity random graphs. *Phys. Rev E*, 63:056127–05645, 2001.

[151] W. Willinger, R. Govindan, S. Jamin, V. Paxson, and S. Shenker. Scaling phenomena in the internet: Critically examining criticality. *Proc. Natl. Acad. Sci. USA*, 99:2573–2580., 2002.

[152] F. Y. Wu. The potts model. *Rev. Mod. Phys.*, 54:235–268, 1982.

[153] S.-H. Yook, H. Jeong, and A.-L. Barabási. Modeling the internet's large-scale topology. *cond-mat/0107417*, 2001.

[154] M. Q. Zhang. Promoter analysis of co-regulated genes in the yeast genome. *Comput. Chem.*, 23:233–250, 1999.