

**A STATISTICAL PHYSICS PERSPECTIVE
OF COMPLEX NETWORKS:
FROM THE ARCHITECTURE OF THE INTERNET AND
THE BRAIN TO THE SPREADING OF AN EPIDEMIC**

THÈSE N° 3270 (2005)

PRÉSENTÉE À LA FACULTÉ SCIENCES DE BASE

Institut de théorie des phénomènes physiques

SECTION DE PHYSIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Thomas PETERMANN

physicien diplômé EPF
de nationalité suisse et originaire d'Emmen (LU)

acceptée sur proposition du jury:

Prof. P. De Los Rios, directeur de thèse
Prof. G. Caldarelli, rapporteur
Prof. W. Gerstner, rapporteur
Prof. M. Tomassini, rapporteur

Lausanne, EPFL
2005

“Die Gesellschaft besteht nicht aus menschlichen Körpern und Gehirnen.
Sie ist ein Netzwerk von Kommunikation.”

Niklas Luhmann (1989)

Acknowledgments

The present work would not have been possible without the help of a large number of people. First of all, I am indebted to Paolo De Los Rios for his constant support during the whole period of this Ph.D. thesis. He provided me with numerous stimulating inputs and his door was always open enabling me to share with him an almost infinite number of discussions, which I will not forget soon. I am grateful to everything he has transmitted to me.

I further would like to thank to Francesco Piazza for the many valuable discussions and that he has devoted much of his time to a careful reading of this thesis. His suggestions and comments considerably improved the manuscript.

I am also grateful to Marc Barthélemy, particularly for the stimulating discussions during his visit. Furthermore, I appreciated the valuable comments and clarifications regarding some aspects of this thesis from Guido Caldarelli, Fredrik Liljeros, Petr Tinyakov and Alessandro Vespignani. A special thank goes to Mark Buchanan who encouraged me to enter into this truly exciting field of research.

Finally, I am most grateful to my parents who always supported me in many ways and who ultimately enabled me to step my feet into the world of networks that surround us.

This work was financially supported through the EC-Fet Open project COSIN IST-2001-33555 and through the OFES-Bern (CH).

A statistical physics perspective of complex networks: from the architecture of the Internet and the brain to the spreading of an epidemic

Statistical physics has revealed itself as the ideal framework to describe large networks appearing in a variety of disciplines such as sociology, communication technology or neuroscience. Despite the diversity of these systems, they appear to exhibit a similar topological complexity such as the presence of small-world or scale-free patterns. The former property refers to a high global and local interconnectedness, whereas the latter means that the frequencies of the number of connections per node, i.e. the degrees, are distributed according to a decaying power law. This ubiquity at the topological level raises several questions. First of all, it should be verified whether the observed topology obtained through the measuring process corresponds to the real one. It is also important to understand the influence of topology on dynamic processes running on a network. Furthermore, we wish to explain how specific factors shape network topology.

By implementing the measuring process as a treelike exploration, we demonstrated for scale-free network models that the exponent of the degree distribution of the explored network is smaller than the original one. This means that the low-degree nodes are underrepresented. Since such an exploration in principle mimics the discovery of the Internet map, the corresponding exponents should not be taken at face value.

As mentioned above, topology plays a crucial role in different dynamic processes taking place on complex networks. An example of paramount importance is the spread of an epidemic. In such a context, it does not come as a surprise that a virus spreads more easily on a network in which global distances are small. This topological property is one of the conditions that allows one to ignore dynamical correlations and to describe the process in the framework of a mean-field approximation. This description, which we derived at different levels, uncovers the role of the degree. However, the influence of the local interconnectedness on the spreading behaviour remains elusive. By systematically exploiting spatial and temporal correlations that govern the spreading dynamics, we further elaborated two methods which quantitatively describe how local substructures influence the spreading behaviour.

In the simplest model for a small-world network, a high global interconnectedness originates from the addition of long-range connections to a regular lattice. In a situation where a cost is associated with the lengths of the links, it is interesting to explore whether the emergence of small-world topology conflicts with a minimisation of the wiring costs. We found that, if the lengths of the additional links are distributed according to a decaying power law, small-world networks can be constructed in a very economical way. As further intriguing consequences, an increase of the exponent of the length distribution optimises the distribution

of flows of traffic over the links while making the networks less vulnerable with respect to random failures of connections.

Overall, this study has led to a series of results related to the topology of complex networks. More precisely, we have investigated how the topology is obtained, what its role in dynamic processes is and what factors shape it.

Une approche des réseaux complexes inspirée par la physique statistique: de l'architecture de l'Internet et du cerveau à la diffusion d'une épidémie

La physique statistique s'est révélée être un cadre idéal afin de décrire de grands réseaux qui apparaissent dans une multitude de disciplines telles que la sociologie, les technologies de communication et les neurosciences. Malgré la diversité de ces systèmes, ils sont presque identiques si l'on regarde à un niveau statistique la manière dont les nœuds sont connectés les uns avec les autres. Cette similitude topologique se manifeste par la propriété 'petit monde' et par l'invariance d'échelle. La première de ces propriétés signifie une haute interconnexion au niveau global et local, tandis que la seconde se réfère au nombre de connexions par nœud, c'est-à-dire au degré. L'invariance d'échelle implique que les degrés sont distribués selon une loi de puissance décroissante. Cette ubiquité topologique soulève plusieurs questions. Tout d'abord, il s'agit de vérifier si la topologie obtenue par un processus de mesure correspond à celle du réseau en question. Il est également important de comprendre l'influence de la topologie sur des processus dynamiques se déroulant sur un réseau. En outre, nous souhaitons connaître comment certains facteurs déterminent la topologie d'un réseau.

En implémentant le processus de mesure mentionné ci-dessus comme une exploration arborescente, nous avons démontré pour des modèles de réseaux invariants d'échelle que l'exposant de la distribution des degrés du réseau exploré est plus petit que celui du réseau de départ. Cela veut dire que les nœuds de degré bas sont sous-représentés. Comme une telle exploration imite la découverte de l'atlas de l'Internet, les exposants correspondants doivent être interprétés avec prudence.

La topologie joue également un rôle crucial dans plusieurs processus dynamiques pour lesquels le réseau représente la "trame". Un exemple d'une importance primordiale est la diffusion d'une épidémie. Dans un tel contexte, il est peu surprenant qu'un virus se propage plus facilement sur un réseau caractérisé par des distances globales courtes. Cette propriété topologique a également pour effet que les corrélations dynamiques sont faibles ce qui permet de décrire le processus dans le cadre de l'approximation du champ moyen. Cette description que nous avons dérivée sur plusieurs niveaux fournit une interprétation du rôle du degré dans la dynamique de diffusion. Pourtant, l'influence de l'interconnexion locale sur le comportement de diffusion reste largement inconnue. Par une investigation systématique des corrélations temporelles et spatiales qui accompagnent la dynamique de l'épidémie, nous avons développé deux méthodes qui décrivent d'une façon quantitative comment des structures d'interconnexion locales déterminent le comportement de diffusion.

Dans le modèle le plus simple d'un réseau petit monde, la haute intercon-

nexion globale est reproduite en ajoutant de longs liens sur un réseau régulier. Dans le cas où un coût est associé aux longueurs des connexions, il est intéressant d'examiner la condition pour qu'une topologie petit monde apparaisse si l'on désire minimiser le coût de câblage. Nous avons démontré que des réseaux petit monde peuvent être créés d'une façon très économique si les longueurs des connexions sont distribuées selon une loi de puissance décroissante. Comme autre conséquence intéressante, nous avons trouvé qu'une augmentation de l'exposant de la distribution des longueurs optimise la répartition des flux de données sur les connexions. En même temps, une telle augmentation rend les réseaux moins vulnérables par rapport à des défaillances aléatoires au niveau des connexions.

Dans l'ensemble, cette étude a mené à une série de résultats concernant la topologie des réseaux complexes. Plus précisément, nous avons examiné comment la topologie est obtenue, quel rôle elle joue dans des processus dynamiques et quels facteurs la déterminent.

Komplexe Netzwerke

mit den Augen des Statistischen Physikers betrachtet: Von der Architektur des Internets und des Gehirns zur Ausbreitung einer Epidemie

Die statistische Physik erwies sich als idealer Rahmen zur Beschreibung grosser Netzwerke, wie sie in verschiedensten Disziplinen - so etwa in der Soziologie, der Kommunikationstechnologie oder den Neurowissenschaften - auftreten. Obwohl diese Systeme von der Natur her sehr verschiedenartig sind, besitzen sie eine ähnliche topologische Komplexität. Diese ist typischerweise charakterisiert durch die Präsenz von 'small-world'-artigen oder skalenfreien Vernetzungsmustern. Während small-world für eine hohe globale und lokale Vernetzung steht, bedeutet Skalenfreiheit, dass die Verteilung der Grade - d.h. die Verteilung der Anzahl Verbindungen pro Knoten - einem abfallenden Potenzgesetz folgt. Diese Allgegenwärtigkeit auf topologischer Ebene wirft verschiedene Fragen auf. Zullererst sollte sichergestellt werden, ob die durch den Messprozess erhaltene der tatsächlichen Topologie entspricht. Weiter ist es äusserst wichtig, die Rolle der Topologie in sich auf Netzwerken abspielenden dynamischen Prozessen zu verstehen. Zudem wollen wir begreifen, wie spezifische Faktoren die Topologie eines Netzwerkes bestimmen.

Indem wir den oben erwähnten Messprozess als baumartige Erkundung implementierten, zeigten wir für skalenfreie Netzwerkmodelle, dass der Exponent der Gradverteilung des erkundeten Netzwerkes kleiner ist als derjenige des ursprünglichen Netzwerkes. Dies weist darauf hin, dass die Knoten mit kleinem Grad unterrepräsentiert sind. Da eine derartige Erkundung an sich der Ermittlung des Atlas des Internets gleichkommt, sollten die entsprechenden Exponenten mit Vorsicht interpretiert werden.

Die Topologie spielt auch eine entscheidende Rolle in verschiedenen dynamischen Prozessen, die sich in einem Netzwerk abspielen können. Ein besonders relevantes Beispiel hierfür ist die Ausbreitung einer Epidemie. In einem derartigen Kontext erstaunt es wenig, dass sich ein Virus in einem Netzwerk mit kurzen globalen Distanzen leicht ausbreitet. Diese topologische Eigenschaft bedingt auch, dass dynamische Korrelationen schwach sind und der Prozess daher im Rahmen der mittleren Feldapproximation beschrieben werden kann. Diese Beschreibung, welche wir auf verschiedenen Ebenen herleiteten, liefert eine Interpretation der Rolle des Grades in der Ausbreitungsdynamik. Auf welche Art die lokale Vernetzung das dynamische Verhalten beeinflusst, bleibt jedoch schwer fassbar. Mittels einer systematischen Untersuchung der zeitlichen und räumlichen Korrelationen, welche die Dynamik der Epidemie begleiten, entwickelten wir zwei Methoden, die quantitativ beschreiben wie lokale Vernetzungsstrukturen das Ausbreitungsverhalten bestimmen.

Im einfachsten Modell für ein small-world Netzwerk wird die hohe globale

Vernetzung durch hinzufügen langer Verbindungen auf ein regelmässiges Gitter erreicht. Falls nun aber die Länge einer Verbindung mit Kosten verknüpft ist, stellt sich die Frage, ob die small-world Eigenschaft auch dann resultiert, wenn zugleich die Vernetzungskosten minimiert werden sollen. Wir wiesen nach, dass small-world Netzwerke auf eine sehr sparsame Art erzeugt werden können, indem die Verbindungslängen gemäss einem abfallenden Potenzgesetz verteilt werden. Zudem fanden wir, dass eine Erhöhung des Exponenten der Längenverteilung die Netzwerke hinsichtlich zufälliger Ausfälle von Verbindungen weniger verwundbar macht und die Verteilung von Datenflüssen durch die Verbindungen optimiert.

Insgesamt führte diese Studie zu einer Reihe von Erkenntnissen betreffend der Topologie von komplexen Netzwerken. So untersuchten wir, wie man die Topologie erhält, welche Rolle sie in dynamischen Prozessen spielt und durch welche Faktoren sie bestimmt wird.

Contents

Acknowledgments	ii
Abstract	v
Version abrégée	vii
Kurzfassung	ix
1 Introduction	1
1.1 Statistical physics and graph theory	1
1.2 Examples of complex networks	3
1.3 Topological measures	7
1.4 The topology of real networks	11
1.5 On the role and the origin of topology	13
1.6 Overview	15
2 Preliminaries	17
2.1 Random Graphs	17
2.2 Random networks with a given $P(k)$	19
2.3 Watts and Strogatz' seminal idea	20
2.4 Growth and preferential attachment	23
2.5 The intrinsic vertex fitness model	26
3 Exploration of scale-free networks	31
3.1 The exploration algorithm	32
3.2 Barabási-Albert networks	33
3.3 Other scale-free networks	35
3.4 Discussion	36
4 Epidemic spreading	39
4.1 The contact process	39
4.2 Mean-field approximation	41
4.2.1 Random bimodal networks	43
4.2.2 Homogeneous networks	45
4.3 Exact formulation	47

4.4	Two-step description	48
4.4.1	Networks of degree four	50
4.4.2	Arbitrary degree	59
4.5	Cluster approximations	60
4.5.1	Revisiting the mean-field and pair approximations	61
4.5.2	Further Systematic Improvement	66
4.6	Discussion	76
5	Spatial small-world networks	79
5.1	Topology	81
5.2	Wiring costs	85
5.3	Distribution of flows	87
5.4	Robustness	89
5.5	Discussion	92
6	General conclusions and outlook	95
A	Full subgraph developments	101
	Bibliography	105
	Curriculum Vitae	113

Chapter 1

Introduction

1.1 Statistical physics and graph theory

Thermodynamics deals with processes associated to heat, such as the melting of ice or the evaporation of water. In the first case, the added heat is used to transform the water initially present as ice into the liquid phase, and in the second, we have a transition from the liquid to the gaseous phase, hence the name *phase transition*. At a microscopic level, water is composed of a huge number of H_2O molecules, and it is the way these are arranged that determines the phase. While their positions are constrained in the solid phase, they move around rather freely in the liquid. Therefore, the existence of different possible phases is not related to the nature of a single H_2O molecule, rather, its roots lie in the collectivity. *Statistical physics* now deduces the macroscopic behaviour from the microscopic laws according to which the H_2O molecules interact and from the statistical distributions of their positions and velocities [1]. Thus it builds the bridge between the molecular world and that visible to the naked eye.

And there was *graph theory* [2] which is the study of abstract mathematical objects composed of a set of vertices that are connected through edges which can be directed or not. The vertices are also referred to as nodes or points; and links, lines or connections are synonyms for edges. Henceforth, these notions are used in an interchangeable way. The applicability of graph theory was already recognised in the 18th century by the Swiss mathematician Leonhard Euler [3]. The puzzle he considered concerned the Prussian town Königsberg, which is divided into 4 parts by the river Pregel (see Fig. 1.1). He wondered if it is possible to take a walk, starting in any part of the town, such that every bridge is crossed exactly once. Euler realised that the shape or size of the town parts do not matter and that this problem is thus best solved by representing the town parts as the nodes and the bridges as the links of a graph. More precisely, the resulting representation is a multigraph since C is connected to both A and B by more than one edge. This representation clearly shows that any town part can be reached by an odd number of bridges, which implies that the desired walk is impossible. At the time,

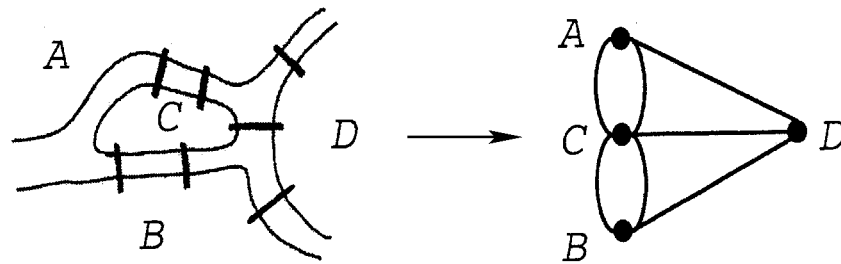


Figure 1.1: A simplified view of the town of Königsberg (left) and its graph representation (right). The nodes A, B, C and D represent the respective town parts the edges are the bridges between them [4].

the study of graph theory remained a branch of discrete mathematics dealing with such problems, mostly because there was only little data about real networks or graphs.

More recently however, connectivity data about numerous real-world networks ranging from the Internet to biological networks has increasingly become available. These networks, which will be described in more detail in the next section, are also very large - sometimes containing millions of nodes. It has thus become more interesting to analyse the topology at a statistical level - from local to global, large-scale properties - rather than to look at specific nodes and investigate to what other nodes they are connected. This statistical characterisation also called for explanations of the observed properties, and this is where the framework of statistical physics comes in: “microscopic” models able to reproduce some large-scale, i.e. “macroscopic”, properties were proposed. As many real networks are growing objects, the dynamic aspect has been a major ingredient in this renewed interest in graph theory. The concept of phase transition also enters as in some network models, a parameter allows to tune between different topological regimes or phases. Moreover, as the spread of viruses often displays a wave-like behaviour, the tools of statistical physics are also useful when it comes to dynamic processes taking place on networks.

In the remaining part of this chapter, the real-world networks of interest will be introduced. We then describe a set of topological measures - at the statistical level - and discuss them for the given examples. The following section highlights the importance of topology and what factors shape it in various contexts, finishing with an overview of the problems investigated in this thesis.

1.2 Examples of complex networks

The networks for which data has increasingly become available in recent years stem from disciplines as diverse as communication technology, sociology and neuroscience. Despite the apparent diversity of these systems, they exhibit a very similar topological complexity at a statistical level, hence the name *complex networks*. While these topological aspects are addressed in the following sections, we here describe these networks, focusing on the examples relevant to our study.

The first example we shall introduce is the Internet [5], this network being relevant to all parts of this thesis. For a majority of people, the word “Internet” means access to an e-mail account and the ability to mine data through one of the most popular public web search engines. The Internet is in fact much more than that, and here we refer to it as a network of computers and other telecommunication devices, i.e. routers, connected through physical links such as cables or wireless connections. The topology of this technological communication network can also be studied at a coarse-grained level, that is at the interdomain (also referred to as autonomous systems) level: each domain - composed of hundreds of routers - is represented by a single node, and an edge is drawn between two domains if there is at least one route that connects them. Fig. 1.2 shows the Internet at the router level (left) and at the autonomous systems level (right). The former was obtained by probing the Internet from a single source whereas so-called skitter traces (a mapping technology involving multiple sources) were

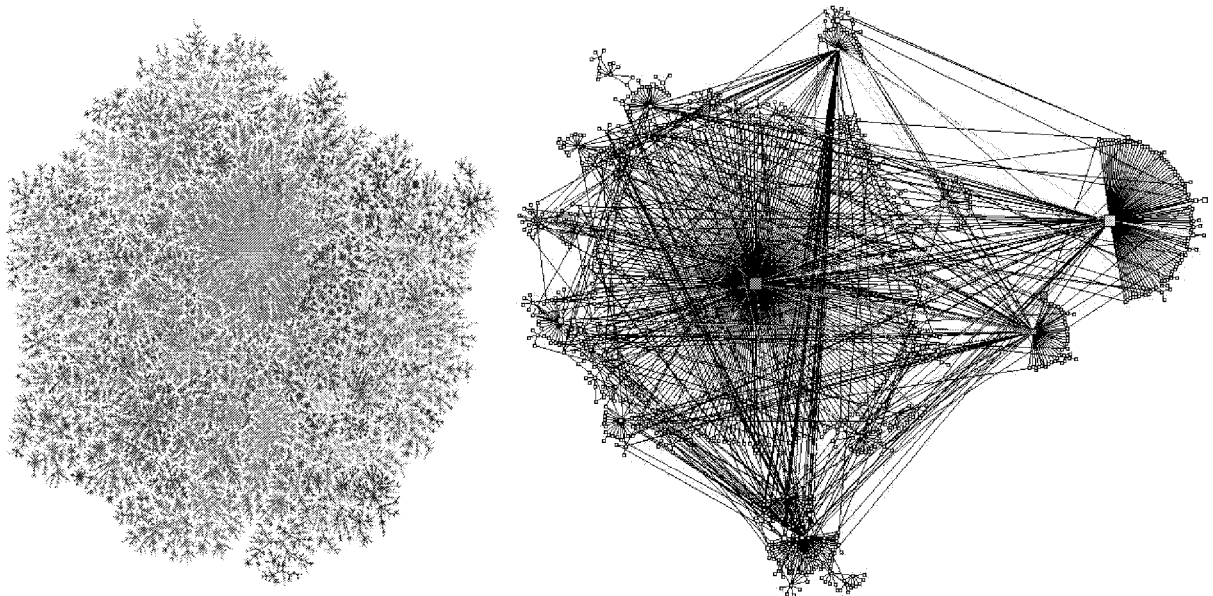


Figure 1.2: Left: Two-dimensional image of a router level Internet map collected by H. Burch and B. Cheswick [6]. Right: Two-dimensional image depicting the Internet’s connectivity at the interdomain level, reconstructed from skitter traces [7].

used to construct the latter. We comment on the statistical topology of the Internet at these two levels in the next section while the reliability of these Internet maps is the content of Chapter 3.

The Internet should not be confused with the World Wide Web. The latter is a directed network whose nodes are the html-documents (web pages) and the edges are the hyperlinks. It can be regarded as a virtual network, the Internet providing the stage. As hyperlinks are visible, it is much easier to analyse its topology, and the size of the World Wide Web for which information is available is much larger than for the Internet. The statistical analysis for the Internet by Govindan and Tangmunarunkit [8] was for example based on 150'000 routers whereas Broder et. al [9] investigated a fraction of the Web containing 2×10^8 web pages. Moreover, the World Wide Web is not a purely communication technological network: as it is used by cyber communities to exchange ideas and by increasingly more people to purchase goods, social and economic factors, among others, shape its topology.

In a technological context, we shall mention two more networks. The first is the power grid in which generators, transformers and substations play the role of the nodes and the edges represent high-voltage transmission lines. The second example is at a much smaller spatial scale, namely electronic circuits. In this case, nodes correspond to electronic components (e.g. logic gates in digital circuits, resistors, capacitors and diodes) and edges are wires in a broad sense. These are two examples where the links correspond to physical wires to which a cost is associated. This cost which is essentially the total length of all the links, appears to be subject to minimisation constraints. In this respect, these cases are relevant to the investigations reported in Chapter 5.

A major part of this thesis concerns the spread of an epidemic for which the Internet - in the case of a computer virus - or a social contact network - for human infectious diseases - represent the underlying arenas. It is therefore worth briefly discussing social networks. While in a social context, the nodes represent individuals, the notion of a link is much more vague. In order to circumvent this ambiguity, one usually focuses on a specific type of social interaction or requires a well-defined condition for two individuals to be connected. Examples of the former strategy include the web of human sexual contacts, scientific - or movie actor collaboration networks while "knowing each other on a first name basis" exemplifies the latter. Sexually transmitted diseases clearly spread on the web of human sexual contacts where any two individuals are connected if they had a sexual contact in a given time window. An example of such a network was obtained from a Swedish survey of sexual behaviour [10]. The other examples of social networks mentioned above are not necessarily playgrounds for viruses, but it is still worth commenting on them. Actors who have played together in the same movie or scientists who have written a paper together clearly know

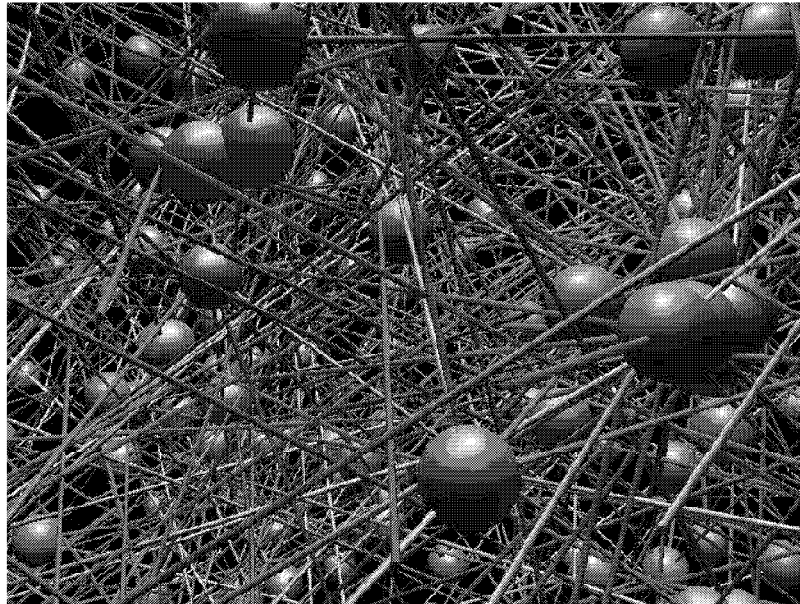


Figure 1.3: Representation of a social group of people in Canberra, Australia, and how they are socially linked to one another. The radii of the spheres (individuals) are proportional to the respective number of connections, by A. Klövdahl [11].

each other well¹, giving rise to interesting social networks. Fig. 1.3 visualises the acquaintances of a large group of people where the “first name” criterion lies at its basis.

The last category of examples we shall introduce are neural networks at several levels. While artificial neural networks have attracted the scientific community for quite a long time, see e.g. [12], connectivity data about real neuronal networks has become available only rather recently, and as we will see, they are characterised by a similar topological complexity as the social and communication technological examples. The building blocks of a neural network are the neurons and the synapses. Every neuron consists of numerous dendrites (input part), a soma (central and processing part) and an axon (output part). The dendrites and axon of a neuron are also referred to as its projections or neurites. In simple terms, the soma releases an action potential or spike through the axon depending on the input signals received through the dendrites. At a synapse, two neurons meet; more precisely, the axon of the presynaptic neuron ends and a dendrite of the postsynaptic neuron begins. Such a neuronal network is mapped to a (directed) graph by representing the neurons as nodes and the synapses as edges. Examples are the neural network of the nematode worm *C. elegans* or *in-vitro* grown neuronal networks (Fig. 1.4). The topology of such networks can also be analysed

¹With the exception of experimental high-energy physics where collaborations sometimes involve hundreds of researchers



Figure 1.4: Electron micrograph of a neuronal network being maintained in culture. The two large blobs are the cell bodies (somas) best conserved, and the flat structures correspond to somas of pyramidal cells. Concerning the projections, the thicker “wires” are axons and the thinner ones correspond to dendrites. Courtesy by J. Berger, B. Götze and M. Kiebler (Max Planck Institute for Developmental Biology, Tübingen, Germany).

at a more coarse-grained level. The mammalian neocortex can for example be regarded as a network of cortical areas linked through numerous fibre bundles. Clearly, these graphs do not tell us anything about brain function. Yet, brain functional networks are obtained through the following mapping: the brain is divided into areas, i.e. voxels of cubic shape, which represent the nodes and any two nodes are connected if the activities of the two corresponding voxels are correlated, that is, if the correlation coefficient associated to the dynamic activity is higher than a certain threshold [13]. The above mentioned anatomical examples are believed to have evolved so as to use a minimum amount of wiring resources [14]. Therefore, they are relevant to the investigations in Chapter 5. The functional example was mentioned for completeness.

1.3 Topological measures

In this section we introduce several quantities which will help us discussing the topology of the above examples thereafter.

The first set of measures we shall introduce concerns the global interconnect- edness of a network, that is how far any two nodes are away from each other. More precisely, by computing the shortest path between all pairs of nodes, one obtains a distribution of their lengths, the length being the number of links con- tained in the shortest path, also referred to as the *distance* [15]. The maximum of this distribution is called *diameter* and the mean value is referred to as the *mean distance*, the latter being used more often in this study. This is a reasonable way to characterise global interconnect- edness only for a connected network. That is, for a graph composed of several isolated clusters, the diameter and the mean distance would be infinite since for all pairs of nodes where one node is in another cluster than the other node, no path exists. A way to overcome this inconvenience is simply to exclude all these pairs. However this does not properly capture this topological property as fragmented networks clearly are poorly interconnected at a global level. A more elaborate strategy is to pass to the *global efficiency* [16] which is the average of the reciprocals of the shortest path lengths

$$E_{\text{glob}} = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}}, \quad (1.1)$$

N being the total number of nodes, and d_{ij} is the distance between node i and j . The contribution of pairs of nodes between which no path exists is thus 0, and despite its simplicity, Eq. (1.1) is a much more sophisticated way to quantify global interconnect- edness as it allows for example to compare fragmented and fully connected graphs in a reasonable way. For directed networks, E_{glob} also characterises reasonably the global interconnect- edness if each pair is considered in both directions.

The *local* interconnect- edness of a network is of equal importance and it can for example be quantified in terms of the *clustering coefficient* [17] which is the probability that two nodes are connected, given that they share a nearest neigh- bour. Focusing on a specific node i which is connected to k_i other nodes, the *local* clustering coefficient is

$$C_i = \frac{E_i}{\binom{k_i}{2}} = \frac{2E_i}{k_i(k_i - 1)}, \quad (1.2)$$

that is the number of edges E_i between the neighbours of node i divided by the value this quantity can maximally take on, ensuring $0 < C_i < 1$. The *total* clustering coefficient (that of the entire network in question) is then obtained through

$$C = \frac{1}{N} \sum_{i=1}^N C_i = \frac{1}{N} \sum_{i=1}^N \frac{2E_i}{k_i(k_i - 1)}. \quad (1.3)$$

The (total) clustering coefficient is related to the density of triangles in that large C indicates a high density whereas small C implies a predominantly treelike topology. In analogy to Eq. (1.1), a local efficiency measure can be obtained by applying this equation to the subgraph composed of the neighbours of node i and their interconnections, leading to a quantity similar to the local clustering coefficient [16]. Certainly, it would be unsound to equate a high clustering with high local interconnectedness. For example, the square lattice is highly interconnected at a local level, but has $C = 0$. Therefore, longer loops need to be considered as well. In Chapter 4, we systematically classify, among other things, loops of length 4. For directed networks, local interconnectedness is usually analysed in terms of densities of subgraphs or *motifs* [18]. In the case of triangles for example, the presence of a direction gives rise to several, topologically distinguishable types of such substructures.

As we will describe in the next section, many real networks are highly interconnected, both locally and globally. We call systems exhibiting these two topological features *small-world networks* [17], although this concept is sometimes used when referring to a small diameter only. In fact, its origin dates back to the 1960's when Milgram used the United States' postal service to probe the structure of the American society, finding that any two people are on average separated by just six acquaintances [19]. Hence, in earlier times (and also nowadays as far as its common use is concerned), a small world predominantly referred to a social context and to global interconnectedness. In order to estimate whether a given network shows small-world features in our sense, its clustering coefficient C and mean distance $\langle d \rangle$ are compared with a random graph having the same number of vertices and edges. Random graphs [20, 21] which will be introduced in detail in the next chapter, are poorly clustered and exhibit small path lengths. A given network thus exhibits small-world patterns if $C \gg C_{\text{rand}}$ and $\langle d \rangle \sim \langle d_{\text{rand}} \rangle$.

There is another important fact when it comes to characterising the topology of a complex network: not all nodes have the same number of edges. The corresponding measure is the *degree distribution* $P(k)$ which gives the probability that a randomly chosen node has *degree* k , that is k edges. According to the degree distribution $P(k)$, real networks fall into three major classes: (i) *scale-free* networks characterised by $P(k) \sim k^{-\gamma}$, with γ -values often between 2 and 3, (ii) *broad-scale* networks with a degree distribution that has a power-law regime followed by a sharp cut-off and (iii) *single-scale* networks whose $P(k)$ has a fast (most likely of an exponential type) decaying tail [22]. In all three cases, the degree distribution is a decreasing function of k , and we will see that the nature of this decay is in many respects crucial. For all these classes, there are only a few high-degree nodes (i.e. *hubs*), however there are many more of them in a scale-free network than in a single-scale graph. In the former example, the hubs sew together the network whereas in the latter, there are too few of them so that they are usually not even

referred to as hubs. We again point out how the situation changes if one considers directed graphs. In that case, there are two degree distributions, namely one for the in-degree and another one for the out-degree, each link contributing to both of them.

The degree distribution is only a first way to characterise the degree-related topology of a complex network. It is for example interesting to ask what degree a node most likely has, given that it is connected to a node of degree k . A way to measure these topological *degree correlations* is to take the average degree of the nearest neighbours of all nodes of degree k , i.e. $k_{\text{nn}}(k)$. When categorising networks according to the degree correlations, three cases are imaginable [23]: (i) $dk_{\text{nn}}(k)/dk > 0$ implies that high-degree nodes are more likely connected to other high-degree nodes, and vertices with few links are connected to other low-degree nodes. Networks of this type are said to exhibit *assortative mixing*. (ii) $dk_{\text{nn}}(k)/dk < 0$ means that it is more probable for hubs to be connected to low-degree nodes, this property being referred to as *disassortative* mixing. Finally (iii) $dk_{\text{nn}}(k)/dk = 0$ indicates that such topological degree correlations are absent. A more refined way to quantify this type of topological correlations is by means of $P(k'|k)$, i.e. the probability that a node has degree k' , given that it is connected to one of degree k . This is elaborated in more detail in Chapter 4 and relates to the average nearest neighbour degree through $k_{\text{nn}} = \int k'P(k'|k)dk'$. Moreover, it is straightforward to generalise these measures to directed networks.

Especially in the context of heterogeneous networks that are highly clustered, it is interesting to see which nodes account for this high density of triangles. It therefore seems useful to plot the clustering coefficient C versus the degree k [24]. If $C(k)$ decreases with k , the low-degree nodes are mostly involved in the formation of triangles giving rise to densely interconnected clusters and the hubs, which contribute to $C(k)$ in a negligible way, sew together the clusters. If this decay takes the form of a power law, the many densely interconnected clusters combine to form larger, but less cohesive groups, which combine again to form even larger and even less interconnected clusters. This self-similarity indicates the presence of a nested modularity or of a *hierarchical* organisation. In the case $C(k) \simeq \text{const.}$, all nodes contribute equally to the density of triangles, and this behaviour thus characterises a *non-hierarchical* network. In principle, $dC(k)/dk > 0$ could also be imagined, but the topological interpretation of this behaviour is elusive.

In certain situations, it is desirable to rank the nodes of a given graph according to their importance, leading to an ordered list starting with the vertices which are most “central” in a sense to be defined. At first sight, one might think that the degree reasonably characterises centrality, but as Fig. 1.5 shows, the node which builds the bridge between the left and the right parts (C) is in some sense more central as its removal would entail a fragmentation of this network into two isolated clusters. On the contrary, if the node with the highest degree

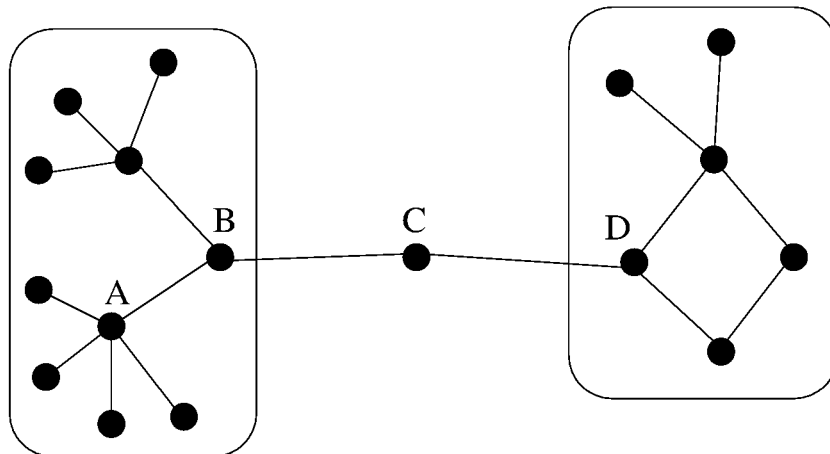


Figure 1.5: A network composed of two clusters joined by vertex C. Depending on how centrality is defined, different nodes play the role of the most central ones.

(A) were cut out, the vertices above B would remain connected to the right part of the graph. Clearly, the degree merely contains local information and the extent to which a vertex is central, is influenced by the global structure of the graph. Possible measures include the *closeness centrality* [25]

$$C_C(i) = \frac{1}{\sum_{j=1}^N d_{ij}},$$

d_{ij} being the distance between node i and j , and *graph centrality* [26]

$$C_G(i) = \frac{1}{\max_{1 \leq j \leq N} d_{ij}}. \quad (1.4)$$

With the second definition, we obtain $C_G(A) = 1/5$, $C_G(B) = C_G(D) = 1/4$ and $C_G(C) = 1/3$, implying that the vertex C is the most central if such a distance criterion is used. When it comes to robustness, i.e. the behaviour of a network with respect to vertex removal, it is important to identify the nodes through which most of the shortest paths go. This goes beyond graph centrality and the corresponding measure is the *betweenness centrality* [27] defined by

$$C_B(i) = \sum_{jk} \frac{n_{jk}(i)}{n_{jk}}, \quad (1.5)$$

n_{jk} being the number of shortest paths going from node j to k , and $n_{jk}(i)$ contains only those going through vertex i . As the name suggests, $C_B(i)$ quantifies to what extent the vertex i lies between others. At a statistical level, it is interesting to study C_B as a function of the degree k , giving again clues concerning the hierarchical organisation of a network. Moreover, the above measures can be adapted such that they apply to edges which will be done in Chapter 5 for Eq. (1.5).

1.4 The topology of real networks

Equipped with a number of topological measures to characterise complex networks, we now discuss in this respect the examples relevant to the present study.

We again start with the Internet due to its particular relevance for this thesis. At the router level, it was observed to be highly interconnected globally while being characterised by the degree distribution $P(k) \sim k^{-\gamma}$. Faloutsos *et al.* studied a subset composed of 3888 nodes, finding an average degree $\langle k \rangle = 2.57$, a mean distance $\langle d \rangle = 12.15$, that of a corresponding random graph being of the same order of magnitude ($\langle d_{\text{rand}} \rangle = 8.75$), and $\gamma = 2.48$ for the exponent of the degree distribution [28]. Based on a larger subset ($N = 150000$), $\langle k \rangle = 2.66$, $\langle d \rangle = 11 \sim \langle d_{\text{rand}} \rangle = 12.8$ and the degree distribution shown in Fig. 1.6a characterised by $\gamma = 2.4$ resulted [8]. In fact, the router map also exhibits a much larger clustering coefficient than that of a corresponding random graph [29], making it a network with small-world features. Beyond this simple characterisation, the absence of both degree correlations and hierarchy was noted [24, 29]. It was further found that the betweenness distribution follows a truncated power law reflecting that the degree of any router is subject to an upper bound [29]. At the interdomain (or autonomous systems) level, the small-world property appears to be conserved [29, 30, 31], and the degrees are distributed according to $P(k) \sim k^{-\gamma}$ with $\gamma \simeq 2.2$ [28, 30] as shown in Fig. 1.6b. In contrast to the router level, the Internet displays degree correlations, namely disassortative mixing, and a hierarchical structure at

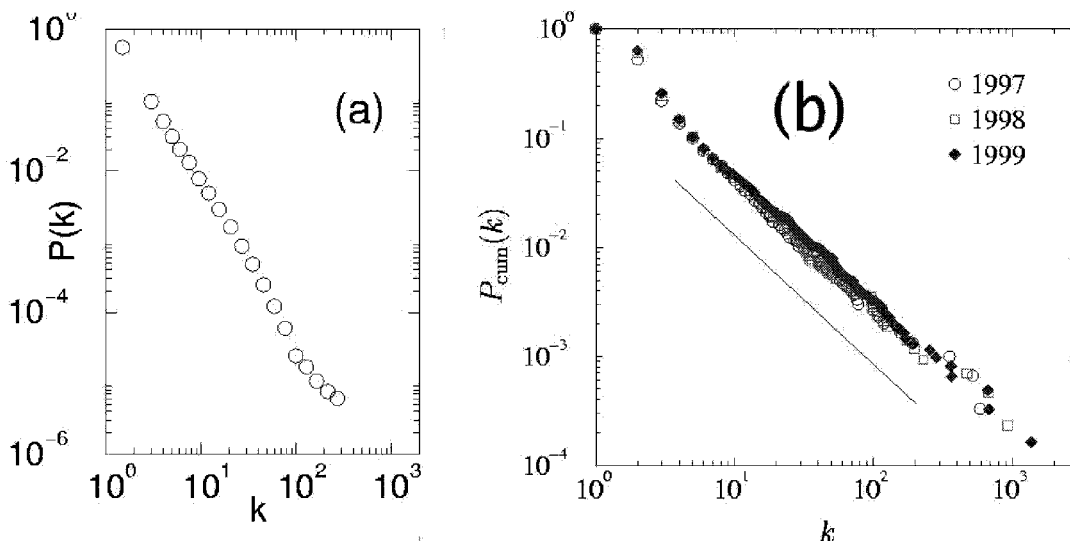


Figure 1.6: (a) Degree distribution of the Internet at the router level [33]. (b) The cumulated degree distribution for the 1997, 1998 and 1999 snapshots of the Internet at the autonomous system level. The power-law behaviour is characterised by a slope -1.2 , which yields an exponent $\gamma = 2.2$ [30].

the autonomous system level [23, 24, 31, 32]. It was further observed that the vertex betweenness centrality is proportional to the vertex degree for the network at the interdomain level.

The interconnectedness of the power grid of the Western United States, whose graph is composed of 4941 nodes and 6596 edges, is characterised by $\langle d \rangle = 18.7 \sim \langle d_{\text{rand}} \rangle = 12.4$ and $C = 0.080 \gg C_{\text{rand}} = 0.005$ [17], i.e. it falls into the class of small-world networks. Its degree distribution exhibits an exponential tail [33] which was also observed for the electric power grid of Southern California [22]. Furthermore, all nodes contribute equally to the clustering coefficient [24], indicating the lack of a hierarchical structure.

The topology of both analogical and digital circuits is very similar to that of the power grid, despite the spatial extension orders of magnitudes smaller. That is, integrated circuits show small-world patterns along with a single-scale degree distribution, i.e. $P(k)$ is characterised by a sharp cut-off [34].

Although rather different in their nature, social networks are characterised by a similar topological complexity as the above technological examples. Real data is available for the movie actor and scientific collaboration networks spanning several disciplines. The former is characterised by a mean distance $\langle d \rangle \simeq \langle d_{\text{rand}} \rangle$, a clustering coefficient $C \simeq 100 \cdot C_{\text{rand}}$ [17], and its degree distribution is described by $P(k) \sim k^{-\gamma}$ with $\gamma = 2.3 \pm 0.1$ for large k [35], it thus is a scale-free small-world network. Moreover, assortative mixing patterns were observed in this network [23]. Interestingly, scientific collaboration networks have an almost identical topology - at the statistical level - i.e. they are also assortatively mixed scale-free small-world networks [36, 37, 38]. In fact, the assortative mixing property, i.e. people working with many others tend to collaborate with other “hubs”, appears to be a characteristic property of such social networks [23]. When it comes to sexual networks, real data is much harder to obtain, making it difficult to compute network properties beyond the degree distribution, such as clustering coefficients or path lengths. The Swedish survey on sexual behaviour mentioned in the previous section led to the result that the frequency distribution of the total number of partners is described by $P(k) \sim k^{-\gamma}$, for large k , with $\gamma_{\text{female}} = 3.5 \pm 0.2$ and $\gamma_{\text{male}} = 3.3 \pm 0.2$ respectively [39]. However, it is believed that the mean distance of sexual networks is rather short as contacts over long distances are quite common, see e.g. [40]. At the same time, clustering clearly is zero in a heterosexual network. Moreover, the density of loops of length 4 is low since people watch each other and prior partnerships, making it for example unlikely that a woman W_1 who has had a sexual relationship with a man M_A will start one with a man M_B after W_2 has been a sexual partner of M_A and M_B [41]. For these reasons, sexual networks are not small worlds in our sense.

Finally, the topologies of the neural networks introduced above are as follows. Both the neural network of the worm *C. elegans* and *in-vitro* grown neuronal

networks fall into the class of single-scale small-world networks [17, 22, 42]. At a more coarse-grained level, several mammalian cortices (namely the macaque visual cortex, the macaque cortex and the cat cortex) were analysed, finding that the cortical areas are highly interconnected, both at a local and global level [43]. As there is currently very little empirical information on the number of connections per unit at the level of small cortical populations, it remains elusive whether the cortex is a single-scale or a scale-free network. As addressed in the next section, scale-free networks are particularly robust with respect to random removal of nodes; and by adopting such an indirect reasoning, Martin *et al.* conjectured the cortex to be a scale-free network [44]. More information about brain connectivity can be obtained at the functional level, i.e. through the dynamic correlations between brain areas as explained above. The resulting graphs were found to belong to the class of scale-free small-world networks [45]. Furthermore, the positive slopes of the curves corresponding to $C(k)$ and $k_{\text{nn}}(k)$ for these networks question the frequently assumed hierarchical structure of the brain.

1.5 On the role and the origin of topology

We have now given empirical evidence that networked systems from a wide range of disciplines are characterised by a similar topological complexity. This is interesting and immediately raises several questions. Are small-world or scale-free patterns crucial for the functioning of certain systems such as the brain, or do they lie at the roots when it comes to understanding processes such as the spread of diseases? Conversely, one can wonder whether the observed topology of these systems is the result of some common organising principles. Let us comment on these quests with a few examples.

Until a few years ago, it was a long-standing problem why computer viruses are so persistent [46]. The problem resided in the inaccurate modelling of the spreading process: very often, the topology of the connection patterns was not taken into account at all - or improperly, for example by using a square lattice. The topological aspect was simply beyond question. Yet, since we are now aware that the Internet - the place where computer viruses spread - is a scale-free network, this ingredient can be taken into account, finding that there will be a finite number of infected computers for any positive infection rate [47]. In other words, the threshold of the infection rate below which the epidemic dies out, is zero. This behaviour is due to the hubs whose presence causes the entire network to be invaded by a virus. The lesson topology teaches us does not stop at this point; but rather, it also gives clues regarding the immunisation of the network. That is, only if high-degree nodes are more likely to be immunised, the epidemic threshold becomes non-zero [48]. This example shows that topology is in many respects a crucial factor.

It is also the scale-free architecture of the Internet which makes it resilient to random failures, yet very fragile with respect to intentional attacks [49, 50, 51, 52]. In simple terms, if a randomly chosen small fraction of nodes is cut out, these nodes most likely have only a few connections, resulting in no major change in the network structure. On the other hand, if high-degree nodes are removed with preference, the network fragments into several isolated parts. For a single-scale network where most nodes have the same degree and $P(k) \sim \exp(-k/k_0)$ for large k , an intentional attack is less dramatic than for a scale-free network whereas it turns out to be more vulnerable with respect to random failures. This can be understood by applying the same reasoning as above.

The importance of topology was also noted in a neural context. Regarding associative memory tasks, a fully connected network composed of N neurons stores the information in the $N^2/2$ bonds. Although such a connectivity displays a good performance, computer time and memory generally grow with N^2 , and it is unlikely that biological evolution has led to such a topology. In fact, a much sparser connectivity, i.e. a scale-free network, also leads to a good associative memory where computer memory and time are proportional to N [53]. Another example concerns the response of a neural network to sensory input. When an odour is presented to an insect, the approximately 800 neurons of the olfactory antennal lobe develop coherent oscillations after a very short period of time [54]. A fast system response is a characteristic feature of a system with a small diameter whereas coherent oscillations are typically exhibited by locally highly interconnected topologies, such as regular lattices. A small-world connectivity combines these two properties, thus the topology again plays a crucial role [55].

The above examples illustrate the crucial role played by network topology in several contexts. But as the reasoning related to the robustness of scale-free networks equally applies to biological networks [e.g. for a metabolic network in which metabolites (chemical substances) are connected through chemical reactions], this example could suggest that natural evolution has applied a selective pressure to the development of an optimal topology, given the environmental constraints. In other situations, an optimisation of topology can result from design or self-organisation. Other factors that shape the topology of complex networks are spatial constraints. That is, in the case where (i) the positions of the nodes can no longer be ignored and (ii) a cost is associated to the lengths of the links, the fact that the network is embedded in a geographical space constrains the resulting topology [56].

1.6 Overview

In the previous section, we generally exposed the type of questions arising in the field of complex networks. If we again take the problem concerning the walk over the Königsberg bridges, we see that the relevant questions can be divided into the following parts. First, the original system has to be transformed into a graphical representation (town parts become nodes and bridges turn into edges). It follows the analysis of the topology of the resulting graph (all nodes have an odd number of edges) which permits to draw conclusions regarding a dynamic process taking place on the graph (impossibility of an Euler walk, i.e. a walk so as every bridge is crossed exactly once). In addition, it is in many cases important to understand the origin of the observed graph topology (why were the bridges built that way?) [4]. We shall now give an overview of the content of this thesis, and we will see that every selected chapter falls into one of the above mentioned parts of a typical graph-theoretical problem.

In previous sections we have introduced topological measures at a statistical level and analysed several real-world examples in terms of them. Although the described systems were very diverse in nature, they are characterised by a similar topological complexity (e.g. by small-world or scale-free patterns). In Chapter 2, we introduce several models which are able to reproduce selected statistical properties of complex networks. We will see that the concept of growth or the interpolation between randomness and regularity are major ingredients in these models. Some models can be regarded as explanations of the observed properties as they demonstrate how simple organising principles give rise to complex large-scale topologies. In most cases however, these “organising principles” are merely effective mechanisms of no particular profundity which allow to reproduce certain topological features. The advantages of these models are that they depend only on a few parameters. It therefore becomes more convenient to work with these models rather than with the raw datasets, especially when it comes to studying the influence of a selected topological property - this property being well controlled by an appropriate model - e.g. in network performance. Hence, this chapter serves as a basis for subsequent investigations.

While transforming the town of Königsberg into a graphical representation is relatively straightforward, this becomes much more difficult for large networks. In the case of the Internet (at the router level), this task is complicated further as the topology has to be probed by sending data packets from a selected number of routers to given targets. In Chapter 3, we investigate for scale-free network models whether an exploration of this type leads to reliable connectivity data and discuss the implications of our findings.

When it comes to dynamic processes taking place on graphs, the spread of an epidemic is a much more intricate process than an Euler walk. We already noted above that a scale-free architecture fundamentally influences the spread-

ing behaviour of an epidemic. This result was obtained by modelling the system at the mean-field level [47]. In Chapter 4, we investigate the assumptions which underly the mean-field description and derive different levels of this type of approximation. Another challenging quest regards how the spreading behaviour is influenced by the presence of (short) loops in the underlying network. Clearly, the local topology partly determines the number of paths along which a virus can propagate and has thus an effect on the spreading behaviour. Different strategies in order to gain insights about the role of the loop structure have been proposed. An interpretation of how clustering, i.e. a high density of triangles, influences the stationary spreading behaviour was obtained by mapping the epidemic process onto bond percolation [57]. Another approach is to abandon the mean-field level and take into account spatial correlations which govern the epidemic dynamics. Matsuda et al. first used the ordinary pair approximation in order to study a population dynamical problem [58]. This approximation, as its name anticipates, accounts for pair correlations and lies at the basis of improved pair models [59, 60] which uncover the role of the loop structure in a rather indirect way: clustering enters by making a number of assumptions about the open (\sphericalangle) and closed (\triangle) triple correlations. In the remaining part of Chapter 4, we present two methods that take into account the presence of loops. The first systematically exploits temporal correlations while a systematic description of spatial correlations lies at the basis of the second.

We also looked into a problem corresponding to “why were the bridges built that way?” While it does not come as a big surprise that the small-world property makes it easy for a virus to spread, it is a challenging problem to explain whether the very same topological property emerges when it comes to the economical construction of a network with physical links whose costs grow with their lengths. This is an interesting question because a high global interconnectedness essentially relies on long edges connecting far away nodes, as it will be explained in the next chapter. In most small-world models, the lengths of these long links are distributed uniformly. Yet, power-law decaying distributions were measured for systems created through self-organization, design and evolution and for which the wiring costs are a key factor in their formation, namely for the Internet [61], integrated circuits [62] and the human cortex [63]. Some modelling effort taking into account the constraint of wiring minimisation has been made for systems where the connection lengths are [64] or are not distributed according to a power law [65, 66, 67, 68], and such length distributions emerge quite naturally when wiring costs along with shortest path lengths are minimised [69]. In Chapter 5, we re-analyse the small-world phenomenon from a wiring cost perspective and discuss the distribution of traffic over the links and the robustness for networks whose connection lengths are distributed according to a decaying power law.

The major conclusions are drawn in Chapter 6 along with a number of suggestions for further investigations.

Chapter 2

Preliminaries

The increased availability of network-topological data has fostered the elaboration of models, their purpose being severalfold. On the one hand, it is important to have simple “recipes” that allow to reproduce the measured properties of these complex interwoven systems. On the other hand, when it comes to the study of dynamic processes taking place in a network, it is often preferable to work with simple models depending on a few parameters than with the raw connectivity data. In this chapter, the network models used throughout this thesis are briefly sketched, focusing on the ingredients and the properties relevant to our study.

2.1 Random Graphs

Up to almost a decade ago, when there was only little quantitative knowledge about real networks, these structures were believed to be essentially random and therefore modeled as so-called *random graphs* [2, 20, 21]. Such a graph is constructed as follows: Starting with N isolated nodes (e.g. laid out on a circle), any pair is connected with probability p . Variation of the parameter p thus allows to tune the connectedness of this type of graph (Fig. 2.1).

Let us now discuss the statistical properties of this model. The average degree is easily obtained since the $E = pN(N - 1)/2$ edges are shared by the N nodes, thus for large N , we have

$$\langle k \rangle = \frac{2E}{N} \simeq pN.$$

The simplicity of this model also allows for an analytical derivation of the degree distribution. The probability that k edges are attached to an arbitrarily chosen node reads

$$P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k},$$

where the two appearing exponents sum up to $N - 1$ accounting for the absence of self-loops. In the limit $N \gg 1$ and $p \ll 1$, such that $pN = \langle k \rangle$, this distribution becomes

$$P(k) = \frac{\langle k \rangle^k}{k!} e^{-\langle k \rangle},$$

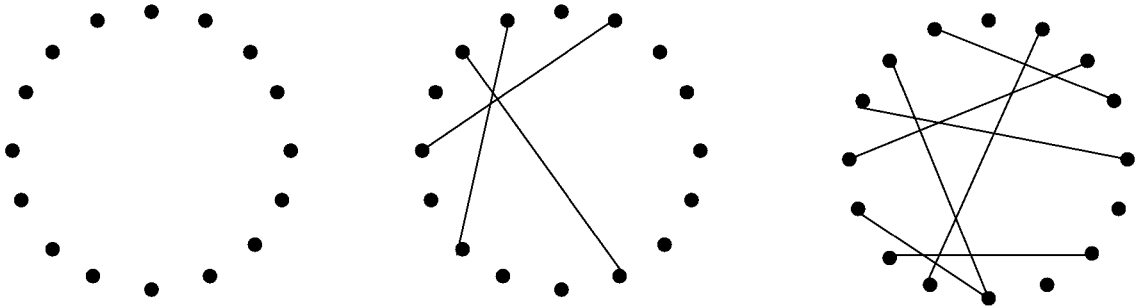


Figure 2.1: Schematic representation of the random graph model: starting from a set of disconnected vertices (left), each pair of vertices is connected with a probability p . Middle and right: two realisations with a small p and a larger value of this parameter (right).

thus strongly decaying as $k \rightarrow \infty$. We further mentioned in the previous chapter that real networks exhibit a strong degree of local interconnectedness. In a random graph, the presence of an edge between nodes A and B obviously does not depend on whether these two vertices are connected via a node C. As a consequence, the clustering coefficient is

$$C = p = \frac{\langle k \rangle}{N}.$$

If random graphs characterised by the same mean degree and different sizes are compared, we see that C decreases with N . These networks therefore mostly exhibit a treelike topology although long loops are present. In fact, the latter statement can be understood through the following reasoning. In analogy to the clustering coefficient, the density of loops of length l (denoted by Q_l) corresponds to the probability of finding a chain of $l - 2$ connected edges given that the two end vertices are connected via another node A. As long as $l \ll N$, we have

$$Q_l = \binom{N-3}{l-3} p^{l-2} \simeq N^{l-3} p^{l-2} = \frac{\langle k \rangle^{l-2}}{N}, \quad (2.1)$$

the binomial factor accounting for the number of possible sets of length $l - 3$ corresponding to the inner vertices of the chain. For $\langle k \rangle > 1$, we see that the density of loops increases with their length. Finally, let us look at the global interconnectedness, in terms of the average distance between any two nodes. If p is chosen too small, the random graph consists of several isolated clusters, and only these pairs of nodes between which a path exists are considered in this case. For p sufficiently large and by assuming a treelike topology, we can argue that the number of nodes N at a distance $\langle d \rangle$ from an arbitrarily chosen one obeys $N \sim \langle k \rangle^{\langle d \rangle}$, implying

$$\langle d \rangle \sim \frac{\ln N}{\ln \langle k \rangle}.$$

This scaling is much slower than that of a D -dimensional regular lattice where $\langle d \rangle \sim N^{1/D}$ and expresses the small-world phenomenon.

In summary, random graphs are characterised by a Poisson degree distribution and tend not to be clustered - two properties rarely seen in real networks. On the other hand, their random nature reproduces a high degree of global interconnectedness. This insight is used in the Sec. 2.3.

2.2 Random networks with a given $P(k)$

In some situations, it is desirable to compare a complex network characterised by a certain degree distribution $P(k)$ with its random counterpart, that is, with a network entirely random except that the degrees of its nodes are distributed according to the distribution $P(k)$ [70]. Such a graph of size N is constructed as follows [71, 72]:

1. To each of the N vertices, attach k ends of edges, k being a random integer drawn according to the distribution $P(k)$.
2. Connect the free ends of links randomly.

Step 2 is often performed such that multiple connections and self-loops are avoided. This constraint can give rise to correlations such that the resulting network no longer falls into the desired class. However for large N this effect can be neglected. Fig. 2.2 illustrates this procedure for a bimodal network characterised by the degree distribution $P(k) = (\delta_{k2} + \delta_{k3})/2$.

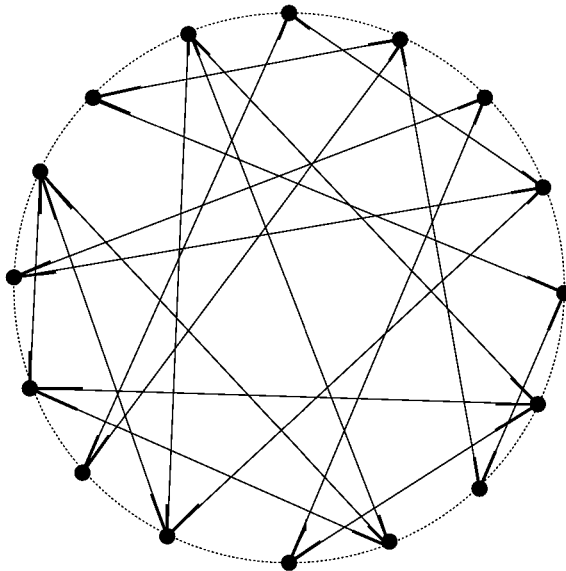


Figure 2.2: Construction of a bimodal random network of size $N = 16$ where half the nodes have degree 2 and the other half degree 3. After the nodes are assigned the aforementioned numbers of ends of edges, they are connected at random, not allowing for multiple connections and self-loops.

The statistical properties not related to the degrees are essentially inherited from the random graph model: the random networks considered in this section are poorly clustered, and their mean distance increases only logarithmically with the system size.

2.3 Watts and Strogatz' seminal idea

As already pointed out earlier, real-world networks are usually highly interconnected both locally *and* globally. High local interconnectedness is a typical property of a regular lattice while the global analogue regards random graphs. Watts and Strogatz combined these two insights and proposed a model that interpolates between a regular lattice and a random graph, thus capturing both a rich

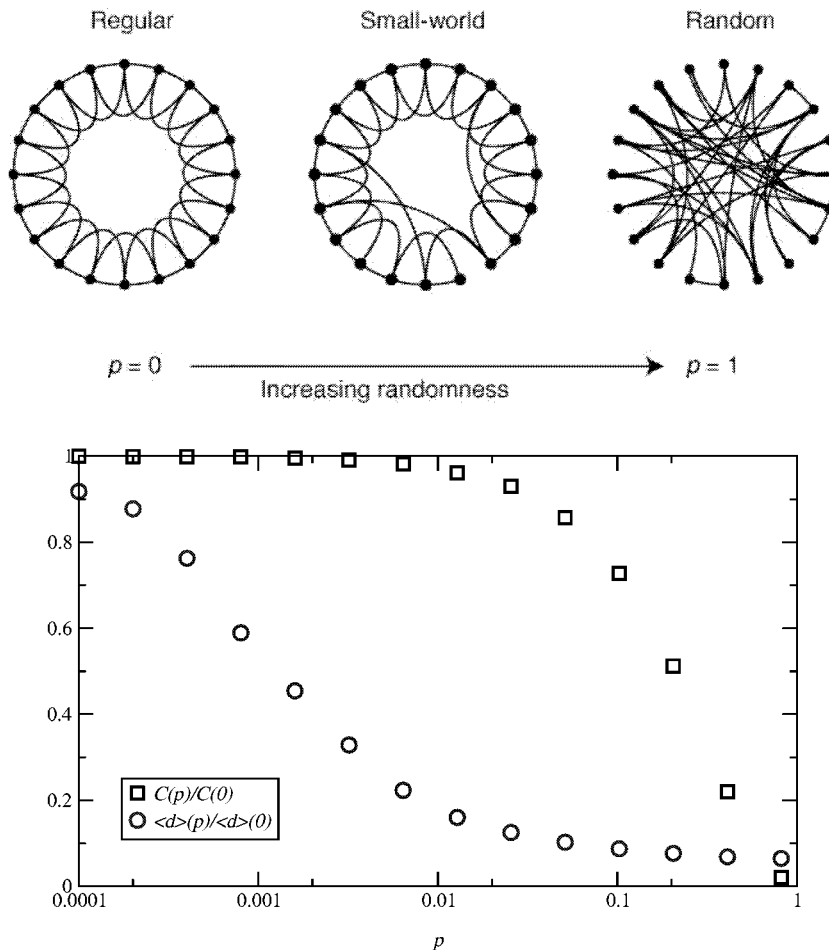


Figure 2.3: Upper panel: the number of re-wirings parametrised by the probability p allows to interpolate between a regular lattice and a random graph [17]. Lower panel: the degree of local and global interconnectedness expressed by the mean distance and the clustering coefficient as a function of the rewiring probability p . The figure shows the nonlinear effect on $\langle d \rangle$ and that C is significantly reduced only for large values of p .

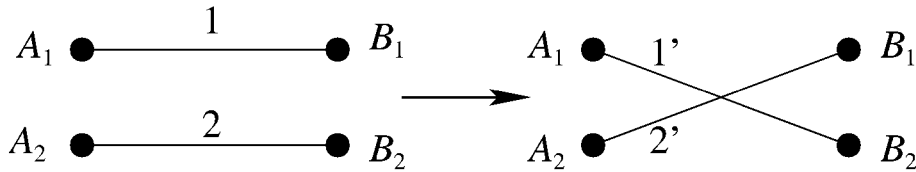


Figure 2.4: Rewiring procedure not affecting the degree distribution: the end vertices of two arbitrarily chosen links are exchanged.

loop structure and small distances between any pair of nodes [17]. In the original formulation, the interpolation is done as follows: starting from a regular lattice (e.g. a ring where nodes two units apart are also directly connected), every edge is rewired with a probability p , that is, the existing link connecting nodes A and B is removed and replaced by a new one between nodes A and C, C being chosen randomly (upper panel of Fig. 2.3). This corresponds to establishing a long-range connection or *shortcut* between nodes A and C. Interestingly, even for low values of the parameter p , the mean distance is already reduced dramatically with respect to the original lattice while the clustering coefficient is almost unaltered, as the lower panel of Fig. 2.3 shows. It is therefore the simplest imaginable model that effectively reproduces the two properties in question.

The emergence of the small-world property depends neither on the chosen regular lattice (corresponding to $p = 0$) nor on the details of the rewiring procedure. Clearly, the mechanism described above affects the degree distribution. If this is an undesirable side-effect, one can perform the rewiring as illustrated in Fig. 2.4 [73]:

- Choose randomly two links (link 1 connecting node A_1 with B_1 and connection 2 linking vertex A_2 with B_2) that do not share a common node.
- Remove these two links and establish two new connections between A_1 and B_2 as well as A_2 and B_1 .

This ensures the degrees of all nodes to be unchanged. This mechanism is particularly useful when it comes to constructing “homogeneous” small-world networks which differ in the precise loop structure. Here the concept of homogeneity is not used in the strict sense that identical connectivity patterns are “seen” from every node, but rather that the degrees are distributed according to the distribution $P(k) = \delta_{kK}$, K being the constant degree. In the case $K = 4$, the application of this rewiring procedure to a square lattice (Fig. 2.5a), to a Kagomé lattice (Fig. 2.5b) or to a topology shown on the left in Fig. 2.3 leads to small-world networks with only loops of length 4, only triangles, or a combination of them - if loops consisting of more than 4 edges are not considered. In Sec. 4.4, we use exactly these types of disordered networks in order to illustrate a method which gives a quantitative interpretation of the loop structure in a spreading phenomenon.

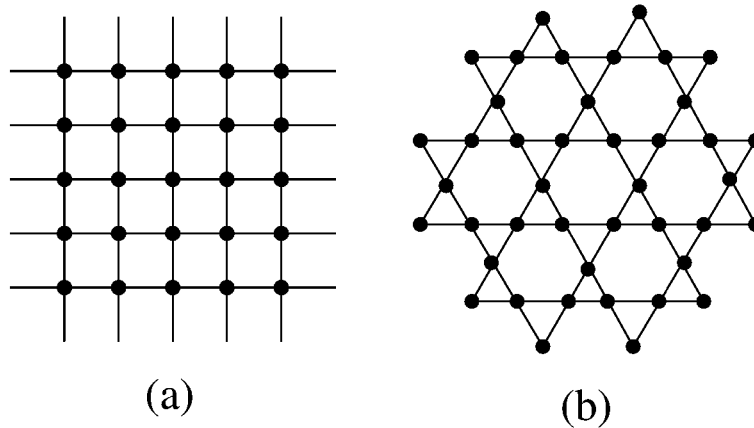


Figure 2.5: Other regular lattices which may be used as point of departure to build small-world networks. In a corresponding context such as the spreading of an epidemic, the square lattice (a) allows to study the effect of loops of length 4 whereas the Kagomé lattice (b) - usually used in condensed matter physics as this geometry represents one of the most frustrated anti-ferromagnetic systems - has only triangles.

The above described rewiring procedures could lead to a fragmentation of the network into several clusters which makes an analytical investigation of this model difficult [74]. Instead, the addition of long-range connections (e.g. emanation of a shortcut at every node with probability p) without removing existing edges essentially leads to the same behaviour of the mean distance, and the network remains fully connected even for $p = 1$.

The precise statistical-physical nature of the onset of small-world behaviour has been a matter of debate. While Watts and Strogatz only discussed this question for a fixed system size N , this onset was later conjectured to be a crossover phenomenon [75]: in 1 dimension and for $N \ll N^* \sim p^{-1}$, the scaling of the mean distance is that of a regular lattice ($\langle d \rangle \sim N$) and also referred to as a large world, whereas for $N \gg N^*$, we have the behaviour observed in random graphs ($\langle d \rangle \sim \ln N$, small world). Based on more rigorous arguments and for general dimension D , the present system was shown to exhibit a critical point at $p = 0$ [74]. It has to be pointed out further that the various rewiring procedures described above all give rise to uniform shortcut-length distributions. In Chapter 5 we discuss, among other things, the nature of this transition for systems where this distribution is a decaying power-law.

In summary, the interpolation between a regular lattice and a random graph reproduces well the high degree of local and global interconnectedness seen in real-world networks. When it comes to the degree distribution, the graphs constructed with this model fall into the category of single-scale networks [22], that is, $P(k)$ is essentially Poissonian for the first rewiring mechanism and if the shortcuts are only added; and $P(k)$ equals that of the initial network for the rewiring

procedure illustrated in Fig. 2.4. This behaviour differs from a number of real-world networks, i.e. those characterised by a scale-free degree distribution. In the two remaining sections of this chapter, two prototype models reproducing such degree distributions are introduced.

2.4 Growth and preferential attachment

In the models discussed so far, the total number of nodes has always been a fixed quantity. This is in contrast to many real-world examples, such as the World Wide Web: a large number of websites are being created every day, connecting themselves to existing sites through hyperlinks. Thereby not all sites already present in the Web are equally likely to acquire a new incoming connection. Yet, if the number of hyperlinks pointing to a given website is used as an indicator of its popularity, a newly created website might preferentially establish a link to a popular one rather than to a site to which only few others point - the “rich get richer” phenomenon. Barabási and Albert used these two ingredients in order to formulate a model which indeed reproduces a power-law degree distribution as is shown below [35]: starting with a set of m_0 vertices and E_0 edges between them, a new node with $m \leq m_0$ edges is added at every time step (growth). The nodes to which the new edges attach are thereby chosen with probability

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}, \quad (2.2)$$

k_j corresponding to the degree of vertex j . This is the rule of preferential attachment. Several approaches were proposed in order to solve this model, namely continuum theory [35, 76], a master-equation approach [77] and a rate-equation approach [78], here we describe the first. The properties of the network after a very long period of growth do not depend on the details of the initial core, the latter are therefore ignored. At time t , the network then consists of $N = t/\Delta t$ nodes, and the preferential attachment rule (2.2) translates into

$$k_i(t + \Delta t) = k_i(t) + m \frac{k_i(t)}{\sum_{j=1}^N k_j(t)} = k_i(t) + m \frac{k_i(t)}{2mt/\Delta t} \quad (2.3)$$

since the sum of all degrees is twice the number of edges, that is $2mt/\Delta t$ if the initial core is ignored. In the continuum limit ($\Delta t \rightarrow 0$), Eq. (2.3) becomes

$$\frac{dk_i(t)}{dt} = \frac{k_i(t)}{2t}, \quad (2.4)$$

and with the initial condition $k_i(\tau_i) = m$, τ_i corresponding to the “birth time” of node i , it has the solution

$$k_i(t) = m \left(\frac{t}{\tau_i} \right)^{1/2}. \quad (2.5)$$

Eq. (2.5) indicates that, in a doubly logarithmic representation, the degrees of all

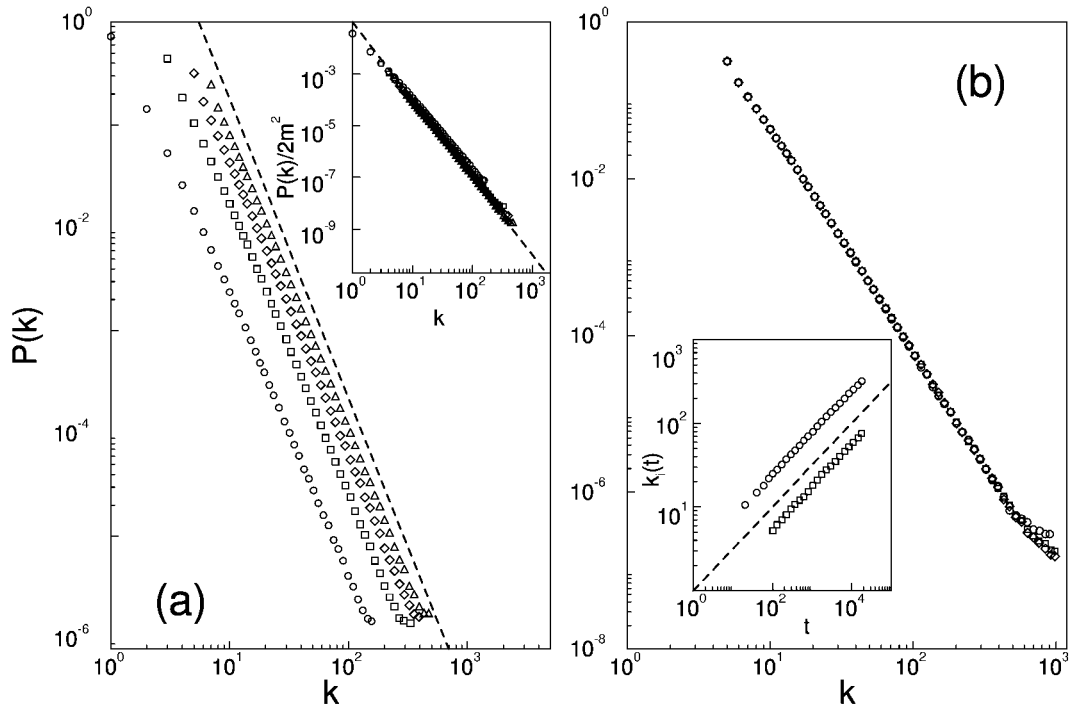


Figure 2.6: Numerical simulations of network evolution according to the Barabási-Albert model ($\Delta t = 1$): (a) Degree distribution, with $N = m_0 + t = 300000$ and several values of $m = m_0$ (\circ : $m = m_0 = 1$, \square : $m = m_0 = 3$, \diamond : $m = m_0 = 5$ and \triangle : $m = m_0 = 7$). The slope of the dashed line is $\gamma = 2.9$, providing the best fit to the data. The inset shows the rescaled distribution $P(k)/2m^2$ [motivated by Eq. (2.8)] for the same values of m , the slope of the dashed line being $\gamma = 3$. (b) $P(k)$ for $m_0 = m = 5$ and various system sizes (\circ : $N = 100000$, \square : $N = 150000$ and \diamond : $N = 200000$). The inset shows the time evolution for the degree of two vertices, added to the system at $t_1 = 5$ and $t_2 = 95$. Here $m = m_0 = 5$ and the dashed line has slope 0.5, as predicted by Eq. (2.5) [76].

nodes grow as straight lines with slope $1/2$, the only difference being the intercept which corresponds to the respective time of birth (inset of Fig. 2.6b). Eq. (2.5) further tells us that nodes can be classified according to their degree or age since these two quantities stand in a monotonic relation to each other. More precisely, the probability that a node has a degree $k_i(t)$ smaller than k , $P[k_i(t) < k]$, is given by

$$P[k_i(t) < k] = P\left(\tau_i > \frac{m^2 t}{k^2}\right). \quad (2.6)$$

Since the nodes enter into the network at a constant rate, the birth times τ_i are uniformly distributed, that is,

$$P(\tau'_i < \tau_i) \simeq \frac{\tau_i}{t}, \quad (2.7)$$

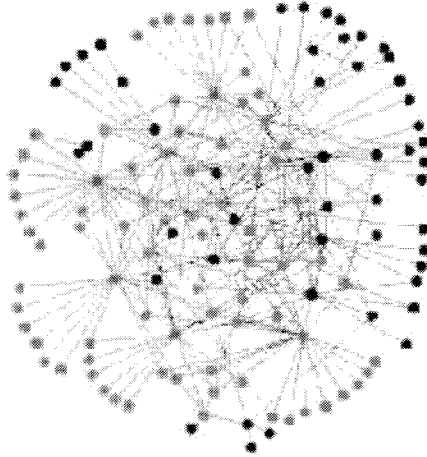


Figure 2.7: A scale-free network of 130 nodes, grown according to the preferential attachment model. The five biggest nodes are shown in red, and they are in contact with 60% of the nodes (green) [79].

the latter approximation holding in the long-term limit. Introducing Eq. (2.7) into Eq. (2.6) leads to

$$P\left(\tau_i > \frac{m^2 t}{k^2}\right) = 1 - \frac{m^2}{k^2}$$

and for the degree distribution, we finally obtain

$$P(k) = \frac{\partial P[k_i(t) < k]}{\partial k} = 2m^2 k^{-\gamma}, \quad (2.8)$$

that is a power law with exponent $\gamma = 3$ independent from m . Fig. 2.6 shows that this result is in agreement with corresponding numerical simulations, and Fig. 2.7 visualises a realisation of this model. In the following chapter we will use the above reasoning when looking at the exploration of such a network during the growth process.

Other values of the exponent γ can be obtained if the attachment kernel (2.2) is generalised. If the attachment rate (2.2) remains asymptotically linear in k , degree distributions $P(k) \sim k^{-\gamma}$ with $2 < \gamma < \infty$ result, the precise value of γ depending on the preasymptotic behaviour [78, 80]. Sub-linear attachment kernels ($\Pi(k_i) \sim k_i^\alpha$, $\alpha < 1$) lead to stretched exponential distributions while super-linear forms ($\Pi(k_i) \sim k_i^\alpha$, $\alpha > 1$) have as outcome a gelation process where there is a node connected to almost every other node of the graph.

The degree distribution is only a first way to characterise the degree-related topology. It is also important to see if high-degree nodes are more likely connected to other high-degree nodes, to those with few connections or if there is no preference. These degree correlations can for example be expressed by means of the average nearest neighbour degree $\overline{k_{mn}}$ as introduced in Chapter 1 and in the case of the model by Barabási and Albert, this quantity does not depend on

k , expressing the absence of such degree correlations [5]. This fact will also be exploited in the next chapter.

Other properties of this model are more difficult to compute analytically. The average distance was shown to scale with the system size as [81]

$$\langle d \rangle \sim \frac{\ln N}{\ln \ln N},$$

and numerical simulations indicate that the clustering coefficient depends on N as [33]

$$C \sim N^{-0.75}.$$

This scaling is slower than that predicted by the random graph model, but it is still too fast to explain the behaviour of real graphs.

To summarise, the above described dynamic perspective along with the idea of preferential attachment provide us with a model able to reproduce a power-law degree distribution. In some situations, e.g. in the case of protein interaction networks, it was argued that such an attachment rule results from the laws governing their evolution [82, 83]. These networks grow by copying (replicating) existing nodes (proteins) borrowing some of their links and adding some new others. But the preferential attachment rule requires that the newly entering node knows the degrees of all the nodes already present in the system. In several contexts, this is certainly not a reasonable assumption. In the next section, we discuss how scale-free networks may emerge when no growth is involved, but in a situation where links between nodes are established according to a deeper organising principle.

2.5 The intrinsic vertex fitness model

Caldarelli *et al.* explored an alternative mechanism in a situation where the total number of nodes N is fixed, giving rise to scale-free networks [84]. The underlying idea is that the nodes are not merely indistinguishable elements, but to every node a real variable quantifying its fitness is assigned. In a social context, this intrinsic property measures the authoritativeness or the social success and the values are drawn from a given probability distribution $\rho(x)$. Pairs of nodes are then connected with probability $f(x_i, x_j)$, f being a symmetric function of its arguments, and x_i and x_j correspond to the respective fitness values. This is an elementary implementation of a link formation process giving rise to a mutual benefit between the two elements involved.

In the following, we describe how the degree distribution can be computed from the above ingredients and then give examples of combinations of ρ and f that reproduce scale-free networks.

If the fitness values are assumed to lie within the interval $0 < x < \infty$, i.e. $\int_0^\infty \rho(x) dx = 1$, the mean degree of a node with fitness x is

$$k(x) = N \int_0^\infty f(x, y) \rho(y) dy = NF(x). \quad (2.9)$$

In analogy to the previous section, the probability that a node with fitness x has a degree less than k reads

$$P[k(x) < k] = P\left[x < F^{-1}\left(\frac{k}{N}\right)\right] = \int_0^{F^{-1}\left(\frac{k}{N}\right)} \rho(x) dx.$$

One therefore obtains for the degree distribution

$$P(k) = \frac{\partial P[k(x) < k]}{\partial k} = \rho\left[F^{-1}\left(\frac{k}{N}\right)\right] \frac{d}{dk} F^{-1}\left(\frac{k}{N}\right), \quad (2.10)$$

which holds exactly in the $N \rightarrow \infty$ limit [85].

Let us now consider the link formation function $f(x_i, x_j) = (x_i x_j) / x_M^2$ where x_M is the largest value of x in the network, the denominator ensuring $f(x_i, x_j) \leq 1$. Eq. (2.9) then leads to

$$F(x) = \frac{x}{x_M^2} \int_0^\infty y \rho(y) dy = \frac{x \langle x \rangle}{x_M^2}$$

and with Eq. (2.10), the degree distribution becomes

$$P(k) = \frac{x_M^2}{N \langle x \rangle} \rho\left(\frac{x_M^2}{N \langle x \rangle} k\right). \quad (2.11)$$

The choice of this functional form of f thus allows the degrees to be distributed essentially in the same way as the fitnesses. In particular, scale-free networks subject to $P(k) \sim k^{-\gamma}$ can be obtained by distributing the fitnesses according to $\rho(x) \sim x^{-\gamma}$. This choice can be justified by arguing that power laws appear rather generically in many contexts when one ranks, for example, people according to their incomes or cities according their populations, etc. This is the so-called Zipf law which establishes that the rank $R(x) \sim x^{-\alpha}$ in a quite universal fashion [86]. Although the mechanism “power law in - power law out” is hardly surprising, it still provides a new path to construct scale-free networks and exploits the widespread occurrence of Zipf’s law in society. In order to explore this model further, we shall choose an exponential distribution for the fitnesses, i.e. $\rho(x) = e^{-x}$, and links between vertices are established with probability $f(x_i, x_j) = \theta[x_i + x_j - z(N)]$ where $\theta(x)$ is the usual Heaviside step function. In other words, the sum of the two fitnesses in question has to be larger than the value $z(N)$ for a connection to be formed. This is inspired by the formation of protein interaction networks. The free energies (fitnesses) obey a Boltzmann distribution (i.e. an exponential distribution) and any two proteins bind, thus forming a protein complex and corresponding to a link, if the sum of the free energies in question is larger than a given threshold. As above, by applying Eq. (2.9) to this combination of ρ and f , the mean degree of a node with fitness x is obtained. The degree distribution is then computed according to

$$P(k) = \frac{1}{N} \int_0^\infty e^{-x} \delta[k - k(x)] dx$$

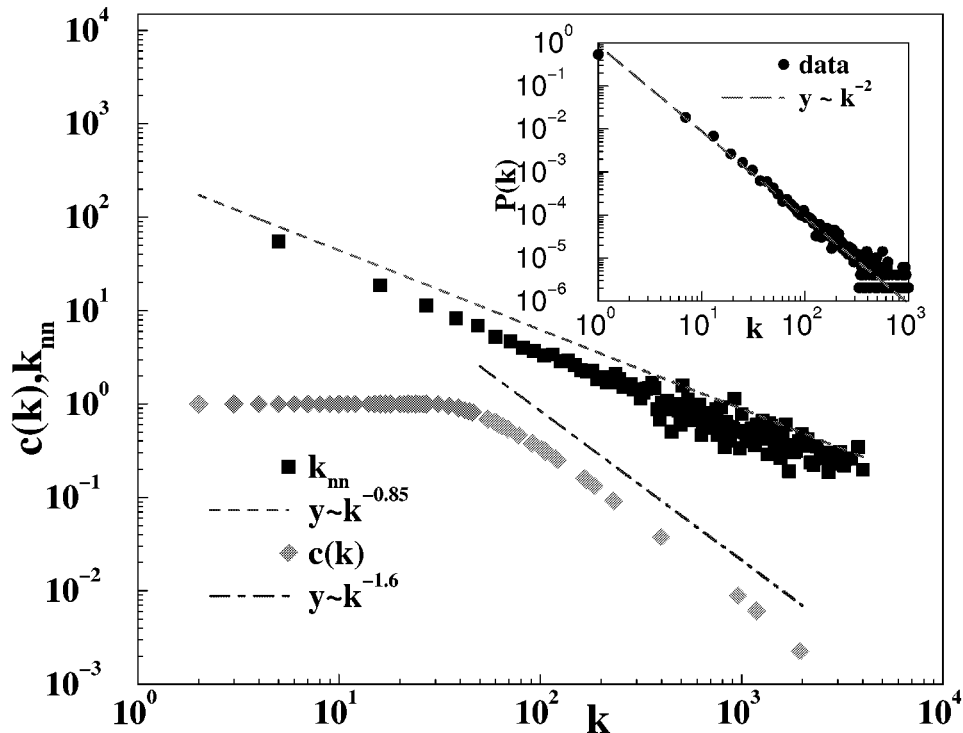


Figure 2.8: Average neighbour degree $\overline{k_{nn}}$ and clustering coefficient as functions of k as well as degree distribution (inset) for networks generated with the threshold rule and an exponential fitness distribution [84].

rather than by using Eq. (2.10), the latter applying to smooth functions only. For this choice of ρ and f , scale-free networks indeed emerge, more precisely

$$P(k) = k^{-2} e^{-z(N)} \delta(k - N). \quad (2.12)$$

The δ -function with the exponential amplitude arises from the fact that nodes with fitness $x_i > z(N)$ are connected to all other nodes, but this is irrelevant for large enough z and N . The inset of Fig. 2.8 numerically confirms Eq. (2.12), and the main figure further shows that this model yields non-trivial behaviours of the degree-dependent clustering coefficient and average nearest-neighbour degree - unlike the preferential attachment model.

The just described choice allows to generate scale-free networks obeying $P(k) \sim k^{-2}$ only. More generally, it was found that, given a fitness distribution density $\rho(x)$, there always exists a symmetric linking probability function $f(x, y)$ such that the resulting random network is scale-free with a given real exponent [87]. In the following chapter, scale-free networks generated according to the fitness model are explored.

In summary, the “good get richer” mechanism described in this section also gives rise to scale-free networks. The model involves a fitness probability distribution and a linking probability function whose appropriate choices permit the decay exponent to take on a given real value. As far as topological proper-

ties beyond the degree distribution are concerned, a non-trivial behaviour of the degree-dependent clustering coefficient, among other things, results. With respect to the model proposed by Barabási and Albert, it successfully reproduces some properties observed in real-world scale-free networks.

Chapter 3

Exploration of scale-free networks

In the elaboration of the network models described in the previous chapter which aim to reproduce some of the measured properties, little attention was paid to the quality of the connectivity data. At an initial stage, it was thus hardly doubted that the measured topology could differ from the real one. In the case of the Internet, the degree distribution is $P(k) \sim k^{-\gamma}$, with $\gamma = 2.15 \pm 0.05$ at the domain level [28] and $\gamma \simeq 2.45$ at the router level [8, 28]. These values clearly lie below the prediction of the original version of the preferential attachment model ($\gamma = 3$). The preferential attachment rule was experimentally tested, finding that the attachment rate is linear for nodes of degree $k \gtrsim 10$ [88] and if the attachment probability is assumed to be linear in k only asymptotically, exponents in the range $2 < \gamma < 3$ indeed result [80] - as already mentioned in Sec. 2.4.

Another explanation of the discrepancy between the measured γ -values and the prediction of the Barabási-Albert model might be that the measuring process “distorts” the real topology of the Internet. It is therefore useful to describe here how one obtains the map of the Internet from which the statistical properties are derived. The most popular method is a tree-like exploration implemented through the recursive use of the `traceroute` command: `traceroute` finds a path (usually a short, but not necessarily the shortest one) from the node where the command is executed to another given node. By repeating the procedure asking `traceroute` to find paths to all other possible nodes (addressed by their IP number), the outcome is a representation of the Internet showing a rather small number of loops. This is due to the fact that `traceroute` usually uses the same paths: if a node D can be reached from A through both B and C , `traceroute` mostly detects only one of them. Yet, more than one path is uncovered if traffic over an already discovered one is so high that it becomes more convenient to switch to a different path. Data collection with this technique yielded the γ -values reported above.

In this chapter, we first describe how we mimicked this tree-like exploration and then present numerical results for the Barabási-Albert and the fitness model introduced in the previous chapter. For the former, the outcome is supported by an analytical argument. The main finding is that such an exploration indeed re-

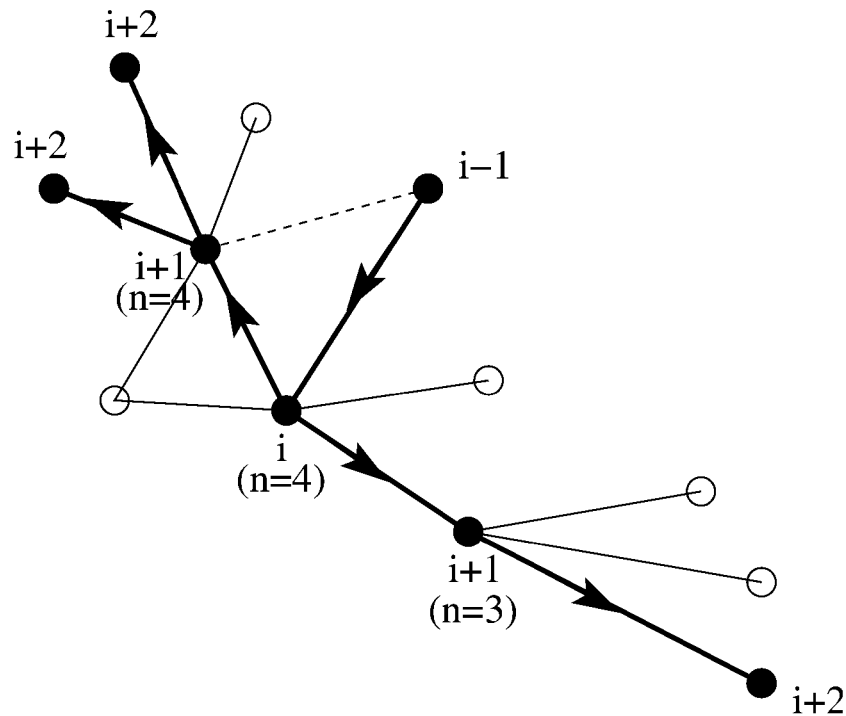


Figure 3.1: Three sequential steps ($i - 1 \rightarrow i$, $i \rightarrow i + 1$, $i + 1 \rightarrow i + 2$) of the exploration algorithm. Arriving at a specific node, the network is further explored by following $[p \cdot n]$ of the n possible paths leading to vertices that have not been visited yet. “Seen” vertices are filled circles whereas empty circles correspond to unexplored nodes. The dashed line symbolizes the exclusion from the exploration as it leads to an already visited node, enforcing the “measured” network to be treelike although cycles can be present in the original graph.

duces the exponent of the degree distribution. Finally, we discuss the implications and mention the research activity that this work has stimulated.

3.1 The exploration algorithm

The mapping of the Internet by means of the `traceroute` tool was modelled by exploring the scale-free network in question according to the following algorithm (Fig. 3.1):

1. Starting at one of the nodes with the highest connectivity k_{max} , one chooses a number of links emanating from that node, namely $[p \cdot k_{max}]$ at random where $[x]$ denotes the integer part of x . The parameter p , $0 < p \leq 1$, allows to tune the exploration.
2. The just chosen links lead to an equal number of vertices. At each of these vertices, count the number of nearest neighbours that have not been visited

yet, and let us denote it by n . From the n possible paths, randomly chose $[p \cdot n]$ of them in order to continue the exploration.

3. Repeat step 2 until no more nodes can be visited.

Before proceeding with the results, a few remarks are in order: (i) By exploring a network according to the above rules, loops are fully lost whereas with the traceroute technique, they are detected, though rarely. (ii) If one started at a node with a few connections only, then there are two possibilities. Either none of the explored links leads to a hub, thus missing the part of the network which accounts for the scale-free topology. If, on the other hand, a hub is reached, then essentially the same network is “seen” as if one started at that hub. We therefore chose the algorithm as it is described above. (iii) In the analytical treatment below, links leading to nodes which have not been visited yet are explored with probability p . But we checked that this formulation is equivalent to the above in that both versions produce the same numerical results.

If the above algorithm is applied to a network, a graph with less nodes and less links results, since at every node only a fraction p of the available links is followed. The intuitive result would be that every node sees just a fraction of its edges, so that all degrees should be reduced by a factor p , the original degree distribution $P(k)$ becoming $\tilde{P}(k) = P(p \cdot k)$. Yet, as is shown in the next section for scale-free networks, the effect of this probabilistic pruning is much more dramatic.

3.2 Barabási-Albert networks

The simulations of the exploration of networks grown according to Barabási-Albert model (Fig. 3.2) show that the measured exponent γ_m differs from the value of the original network ($\gamma = 3$). This represents a situation where the newly entering node establishes just $m = 1$ new connection to an already existing node with probability (2.2), and the exploration is parametrised by $p = 0.5$, giving rise to $\gamma_m \simeq 2.5$. We can wonder if this is a finite-size effect and the correct exponent is recovered for very large networks or whether the exploration induces this change.

A simple analytical argument can be formulated using the lack of degree correlations in the preferential attachment model. Recalling that the degree of node i evolves in time according to

$$\frac{dk_i(t)}{dt} = \frac{k_i(t)}{2t} \quad (3.1)$$

implies that the ages of neighbouring nodes are uncorrelated. This allows us to look at the exploration during the growth process. That is, we label the initial node as reachable; with probability p , a new node is said to be reachable if it connects to at least one reachable node. In a specific realisation at any time t , the

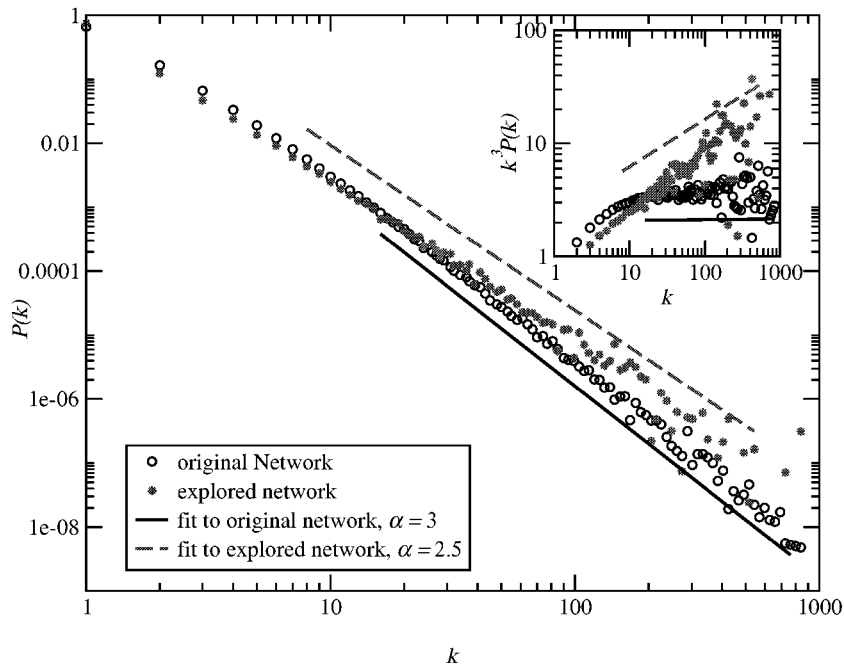


Figure 3.2: Degree distribution of a Barabási-Albert network grown with $m = 1$. Circles: original network; stars: explored network with $p = 0.5$. The best fit to the original network is with $\gamma \simeq 3$, and to the explored network with $\gamma \simeq 2.5$. Inset: rescaled degree distribution $k^3 P(k)$, such that the data for the original network are constant, and the residual power-law behaviour of the explored network is more evident. Here each of the 10 realisations of networks of size $N = 10^6$ was explored 10 times.

rate of change of the number of reachable nodes $dN(t)/dt$ is 0 if the new node will be unreachable and 1 otherwise (here $dt = 1$ without loss of generality). By averaging over many realisations, this rate can be interpreted as the probability that the new node is reachable, and it obeys the iterative equation

$$\frac{dN(t)}{dt} = 1 - \left[1 - p \int_0^t \frac{dN(t')}{dt'} q_t(t') dt' \right]^m \quad (3.2)$$

where $q_t(t') dt'$ is the probability that the new node attaches to one who entered the network during the time interval $[t', t' + dt']$. Through the solution of Eq. (3.1), that is $k_i(t) = m(t/\tau_i)^{1/2}$, τ_i being the birth time of node i , the rule of preferential attachment translates into $q_t(t') = a(t')^{-1/2}$ and with the normalisation $\int_0^t q_t(t') dt' = 1$, we obtain $q_t(t') = 1/(2\sqrt{tt'})$. In other words, the factor $q_t(t')$ gives the older nodes, i.e. those with a higher degree, a larger weight. The integral in Eq. (3.2) then corresponds to the probability that the targeted node is reachable, and the full right hand side implements the condition that at least one of the m targets will be part of the explored network. Since $N(t)$ can grow at most linearly, we make the assumption that $dN/dt \sim t^\alpha$ with α expected to be negative. With $q_t(t')$ as derived above, the integral in Eq. (3.2) becomes

$t^\alpha/(2\alpha + 1)$. In the limit $t^\alpha p/(2\alpha + 1) \ll 1$, the right hand side of Eq. (3.2) can be linearised, leading to one term proportional to t^α . A comparison with the left hand side results in

$$\alpha = \frac{mp - 1}{2}. \quad (3.3)$$

Therefore, as long as $mp < 1$, the temporal density of reachable nodes decreases in time. In order to derive the degree distribution, we here use a shorter approach than in Sec. 2.4. The equivalence of degree and age is reflected in the corresponding density distributions as $P_m(k)dk = \rho(\tau)d\tau$, implying that the degree distribution of the explored network is obtained through

$$P_m(k) = \rho(\tau) \frac{dk}{d\tau}. \quad (3.4)$$

From the solution of Eq. (3.1), that is $k \sim \tau^{-1/2}$, we have $\tau \sim k^{-2}$ and as a consequence $d\tau/dk \sim k^{-3}$. The growth rate of the explored network therefore becomes $\rho(\tau) = dN(\tau)/d\tau \sim \tau^\alpha \sim k^{-2\alpha}$. Eq. (3.4) thus reads $P_m(k) \sim k^{-\gamma_m}$ where $\gamma_m = 2\alpha + 3$ and with Eq. (3.3), we obtain

$$\gamma_m = mp + 2. \quad (3.5)$$

For the case depicted in Fig. 3.2 ($m = 1, p = 0.5$), we find $\gamma_m \simeq 2.5$ which is in very good agreement with the numerical result. We therefore expect that, as long as $mp < 1$, the exponent characterising the explored network differs from that corresponding to the original graph.

3.3 Other scale-free networks

In order to check whether the distortion of the degree distribution is intimately related to the growth and preferential attachment, we applied our exploration algorithm to scale-free networks generated with the fitness model introduced in Sec. 2.5. Let us briefly recall its ingredients: to every node a random variable x , drawn from a density distribution $\rho(x)$, is assigned and any pair of nodes (x, y) is connected with probability $f(x, y)$, f being symmetric in its arguments. The choice $f(x, y) = xy$ permits to map the fitness distribution $\rho(x)$ onto the degree distribution. For example, distributing the fitnesses according to $\rho(x) \sim x^{-3}$ generates networks characterised by $P(k) \sim k^{-3}$. But this model also allows for the generation of scale-free networks in a less trivial way: the combination of $\rho(x) \sim e^{-x}$ with a threshold-type connection probability $f(x, y) = \theta(x + y - c)$ results in graphs whose degrees are distributed as $P(k) \sim k^{-2}$. More generally, given the distribution of the fitnesses, it is always possible to find a symmetric linking probability function such that scale-free networks with a given exponent emerge.

Fig. 3.3 shows that the “power law in - power law out” as well as the threshold-type network also exhibit the distortion effect observed for the Barabási-Albert

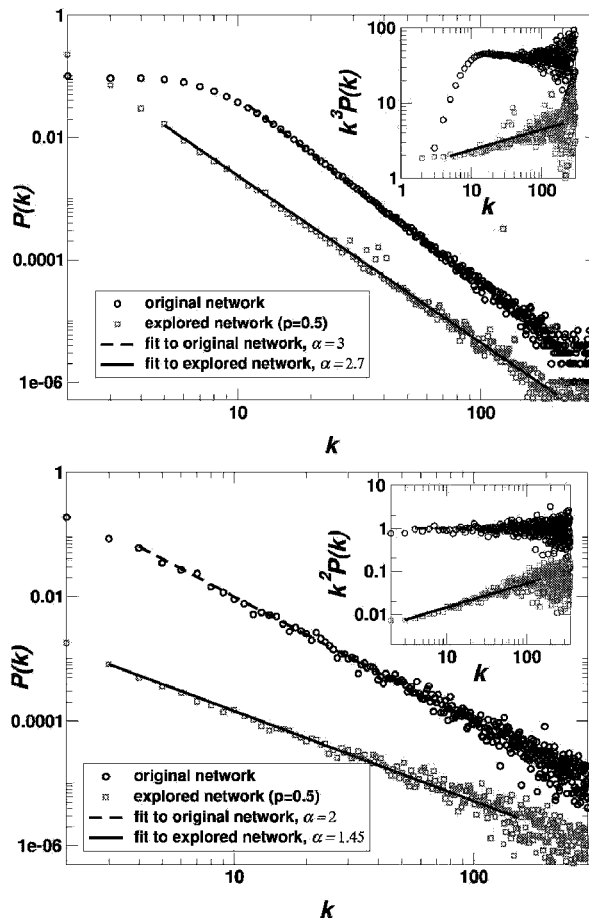


Figure 3.3: Degree distributions of the original and explored networks for the fitness model. Upper panel: $\rho(x) \sim x^{-3}$ and $f(x, y) = xy$. Lower panel: Combination of exponentially distributed fitnesses and threshold-type linking probability function. The distortion effect is best seen in the insets where the values are rescaled such that the data corresponding to the original networks are flat. These results were obtained for 10 different realisations of networks composed of 10^5 nodes, each exploring 10 times.

model. In the former, $\gamma = 3$ leads to $\gamma_m \simeq 2.7$ and for the latter, the explorer “sees” $\gamma_m \simeq 1.5$ with respect to the original value $\gamma = 2$. We do not have an analytical understanding for the reduction of the degree-distribution exponent for this model. As a matter of fact, the presence of degree correlations and the lack of an explicit time evolution hinder the formulation of an equation similar to (3.2).

3.4 Discussion

We have analysed whether the tree-like exploration of scale-free networks can give rise to systematic errors reflected in a change of the degree-distribution exponent. Interestingly, in all cases the measured exponent γ_m is smaller than γ ,

an indication that the exploration process penalises nodes with small degree with respect to nodes with large degree. This is reasonable, since a node with few connections has fewer paths reaching it (and some bottlenecks, since all these paths have ultimately to flow through its few connections) than a high degree node. The ultimate result is therefore that high degree nodes are fairly well represented in the final distribution, whereas the number of nodes with few connections is underestimated¹. This intuitive picture rationalises our numerical and analytical finding that the measured exponent is smaller than the real one.

Obviously, the exploration of scale-free networks can be extended for example by using multiple sources instead of a single one [90]. From every source, the `traceroute` command is then executed to every other node, leading to a more complete representation of the underlying network. Clauset and Moore found that in order to obtain a good estimate of γ , it is necessary to use a number of sources which grows linearly with the average degree of the underlying graph [91]. Moreover, it is even possible that a scale-free topology is “seen” when exploring a random graph [92]. This effect is very pronounced when using a small number of sources and it is related to the fact that the degrees of nodes far from the source are strongly underestimated. This numerical observation was put on a firmer ground, finding that an explored random graph is characterised by $P(k) \sim k^{-1}$ for k below the average degree [93].

Overall, the conclusion of this investigation remains that the data obtained by the recursive execution of the `traceroute` command should not be taken at face value.

¹The sensitivity of low-degree nodes further leads to significantly inaccurate estimates of γ for graphs having many more edges than nodes [89]

Chapter 4

Epidemic spreading

The content of the previous chapter can be interpreted as dealing with a process on a complex network aiming to measure its large-scale topology. In the following, we shall investigate the spreading of an epidemic which can be regarded as a *dynamic* process taking place on a network: a computer virus spreads on the Internet, and HIV (as a biological example) propagates on top of the web of human sexual contacts. As already mentioned in Chapter 1, we focus on the elaboration of methods which lead to an understanding of distinct topological properties - mostly at a statistical level - in the spreading behaviour.

We first describe in detail how the contact process is modelled, along with the chosen connectivity patterns. A set of mean-field-type approximations involving several assumptions is then developed, leading to non-trivial correlations in the time evolution of the fraction of infected nodes with a given degree. Describing the dynamics of the system at an exact level, we first rigorously derive the major mean-field equation and then use this description in order to develop two methods which give an understanding of the local topology in the spreading behaviour. The first consists in taking into account temporal correlations and the second is a systematic exploration of spatial correlations based on the choice of a cluster whose size determines the range up to which such correlations are taken into account. This latter method is - in principle - similar to the cluster variation method of statistical physics [94]. Finally, these methods are critically discussed, especially in view that they can be improved systematically.

4.1 The contact process

The dynamic laws that describe the spreading of an infectious disease are determined by the contact structure which underlies the population. We therefore model the epidemic as a dynamic process on top of a given network that does not change in time. The nodes of the network represent individuals, and the links correspond to relationships between individuals along which an infective agent can propagate.

Since the aim of this chapter is the investigation of the role of specific topological properties, especially the local interconnectedness, we adopt a rather simple epidemiological model where the individuals can be only in two possible states, namely infected (I) or susceptible (S). Because the nodes repeatedly run through the cycle susceptible \rightarrow infected \rightarrow susceptible, it is called SIS model. In the physics community, it has recently been formulated as follows [47]: A node susceptible to the disease gets infected with probability $\nu\Delta t$ if it is connected to at least one infected nearest neighbour. On the other hand, infected nodes recover spontaneously with probability $\delta\Delta t$. Under certain circumstances, this version of the SIS model is formally advantageous with respect to its conventional formulation, where infected nodes can infect neighbouring susceptible vertices with probability $\nu\Delta t$ [95]. In the latter case, susceptible nodes become infected with probability $1 - (1 - \nu\Delta t)^{k_{\text{inf}}}$, k_{inf} being the number of infected nearest neighbours. The former version shall be referred to as the simplified SIS model. By rescaling the time unit, we can reduce the number of parameters to one: the time evolution is determined by the effective spreading rate $\lambda \equiv \nu/\delta$, and the recovery rate is set to 1. The quantitative details of the behaviour of the system still depend on the choice of Δt . In particular, the method described in Sec. 4.4 gives a quantitative interpretation of the effect of short loops by performing two time-steps exactly. This effect is of higher order in Δt , such that their influence is not seen in the continuous-time limit ($\Delta t \rightarrow 0$). As long as $\Delta t > 0$, we set this quantity to 1 without lack of generality and formulate it for the simplified SIS model. The discrete version of the simplified SIS model is also adopted in the two following sections where the bases for the two-step description are elaborated. The methodology discussed in Sec. 4.5 explores spatial correlations and can be formulated both in discrete and continuous time. We describe it for the continuous-time version of the conventional SIS model.

The other model constituent concerns the underlying network. We shall not attempt to examine the combined effect of the degree distribution, degree correlations and the loop structure. But since the various sections of this chapter aim to come to an understanding of these different topological properties in epidemic spreading, appropriate network models are used. In Sec. 4.2, a rigorous mean-field approximation is derived, that is a description which ignores spatiotemporal correlations. As will be shown, the only topological property entering at this level is the number of nearest neighbours of any node. We illustrate this approximation for homogeneous and random bimodal networks, the latter being introduced in Sec. 2.2 and representing a case with a simple degree-correlation structure. The methods beyond the mean-field level described in Secs. 4.4 and 4.5 unravel the role of short loops in the spreading process. The former applies to homogeneous networks as well as to its disordered versions explained in Sec. 2.3 while the latter provides a way to investigate spreading phenomena in the extremal cases only, i.e. either for an entirely regular or random homogeneous network. A random homogeneous network is thereby constructed analogous to a random bimodal network,

that is according to the algorithm given in Sec. 2.2.

4.2 Mean-field approximation

A first approach to describe the spreading dynamics consists in ignoring spatial correlations. In other words, the probability of finding a pair of two infected neighbouring sites A and B is assumed to be the probability that site A is infected times the probability that B is infected. At this level of approximation, the system is described by assigning to all the N nodes of the network a probability P_i^t that node i is infected at time t . As a consequence, its probability to be susceptible is $1 - P_i^t$ since there are only two possible states. The simplified discrete-time SIS model then translates into

$$P_i^{t+1} = \lambda(1 - P_i^t) \left[1 - \prod_{j \text{nn} i} (1 - P_j^t) \right] \quad (4.1)$$

where the product in Eq. (4.1) runs over all the nearest neighbours j of node i , and the factor in the square bracket gives the probability that at least one neighbour of node i is infected at time t . As infected nodes recover spontaneously with probability 1, there is no term corresponding to this transition. Eq. (4.1) shall be referred to as the site approximation. In order to obtain an equation which permits to study the spreading phenomenon on a topology of which we merely know some statistical properties rather than its full connectivity patterns, we transform Eq. (4.1) into the degree-space. By applying $1/N_k \cdot \sum_{i=1}^N \delta_{k_i k} [\cdot]$, $N_k = \sum_{i=1}^N \delta_{k_i k}$ being the number of nodes of degree k , the site approximation reads

$$\rho_k^{t+1} = \lambda k \sum_{k'} P(k'|k) (\rho_{k'|k}^t - \rho_{k,k'}^t). \quad (4.2)$$

This equation describes how the fraction of infected nodes of degree k at time $t + 1$

$$\rho_k^{t+1} = \frac{\sum_{i=1}^N \delta_{k_i k} P_i^{t+1}}{N_k}$$

is determined by the topological degree correlation factor $P(k'|k)$, by the probability that a node of degree k' is infected at time t , given that it is connected to a node of degree k

$$\rho_{k'|k}^t = \frac{\sum_{i=1}^N \delta_{k_i k} \sum_{j \text{nn} i} \delta_{k_j k'} P_j^t}{N_{kk'}}$$

and by the probability to find a connected pair of infected nodes, one having degree k and the other degree k'

$$\rho_{k,k'}^t = \frac{\sum_{i=1}^N \delta_{k_i k} P_i^t \sum_{j \text{nn} i} \delta_{k_j k'} P_j^t}{N_{kk'}}.$$

Thus the latter quantity is not simply $\rho_k \rho_{k'}$, but rather a heterogeneous topology induces correlations although pair correlations are being ignored. The only

approximation involved in Eq. (4.2) is $\prod_{j \text{nni}} (1 - P_j^t) \simeq 1 - \sum_{j \text{nni}} P_j^t$ holding in the vicinity of the epidemic threshold λ_c below which all the P_i 's vanish. The quantity $P(k'|k)$ more precisely is the probability that an arbitrarily chosen node has degree k' given that it is connected to one of degree k . From the total number of links connecting two nodes one of which having degree k and the other k' , i.e. $N_{kk'} = \sum_{i=1}^N \delta_{k_i k} \sum_{j \text{nni}} \delta_{k_j k'} = N_{k'k}$ ¹, this degree-correlation factor is obtained through

$$P(k'|k) = \frac{N_{kk'}}{\sum_l N_{kl}}$$

and satisfies the normalisation $\sum_{k'} P(k'|k) = 1$. This quantity further fulfills the detailed balance condition [96]

$$kP(k'|k)P(k) = k'P(k|k')P(k') = \langle k \rangle P(k, k')$$

where $P(k) = N_k/N$ is the degree distribution, $\langle k \rangle$ the mean degree and

$$(2 - \delta_{kk'})P(k, k') = (2 - \delta_{kk'}) \frac{N_{kk'}}{\sum_{l,l'} N_{ll'}} = (2 - \delta_{kk'}) \frac{N_{kk'}}{\langle k \rangle N}$$

is the joint probability that two nodes of degrees k and k' are connected. We shall now perform a series of approximations in order to simplify Eq. (4.2). Assuming

$$\rho_{k,k'}^t = \rho_{k'|k}^t \rho_k^t, \quad (4.3)$$

the temporal evolution of ρ_k is dictated by

$$\rho_k^{t+1} = \lambda k (1 - \rho_k^t) \sum_{k'} P(k'|k) \rho_{k'|k}^t \rho_k^t \quad (4.4)$$

and with

$$\rho_{k'|k}^t = \rho_{k'}^t, \quad (4.5)$$

it reduces to

$$\rho_k^{t+1} = \lambda k (1 - \rho_k^t) \sum_{k'} P(k'|k) \rho_{k'}^t. \quad (4.6)$$

The effect of the assumptions (4.3) and (4.5) is studied below for random bimodal networks. While the Eqs. (4.2) and (4.4) involve different densities or fields, i.e. ρ_k , $\rho_{k'|k}$ and $\rho_{k,k'}$, Eq. (4.6) expresses the time evolution of ρ_k fully by itself. Clearly, the field is still of a vectorial nature, that is, there is not a single equation for ρ , i.e. $\sum_k P(k) \rho_k$, but rather, the various ρ_k 's are coupled into the system of Eqs. (4.6). This system shall therefore be referred to as the degree-dependent mean-field equation. From the stationary-state solution of Eq. (4.6), one obtains the epidemic threshold [96]²

$$\lambda_c = 1/\Lambda \quad (4.7)$$

¹In the case $k = k'$, N_{kk} gives twice the number of links connecting two nodes both of which having degree k .

²In this reference, Eq. (4.6) is derived in a more heuristic way.

where Λ is the largest eigenvalue of the connectivity matrix defined by $C_{kk'} = kP(k'|k)$. As could already have been anticipated from Eqs. (4.1) and (4.2), for example, topological properties beyond the degree such as the presence of loops, do not enter within the site approximation. In order to come to such an understanding, we will have to abandon the assumption described at the beginning of this section.

In the case of an uncorrelated network, which obeys $P(k'|k) = k'P(k')/\langle k \rangle$, Eq. (4.7) becomes [96]

$$\lambda_c = \frac{\langle k \rangle}{\langle k^2 \rangle}. \quad (4.8)$$

This means that if the average degree is finite, the larger the degree fluctuations the smaller the epidemic threshold. For a scale-free network with $P(k) \sim k^{-\gamma}$, with $2 < \gamma < 3$, $\langle k^2 \rangle \rightarrow \infty$ and λ_c becomes even 0, hence for any infection probability $\lambda > 0$, a finite number of number of individuals is infected in the stationary state [47]. Since the Internet belongs to this class of networks, this result solved the long-standing problem of the long persistence of computer viruses [46]. For such a degree distribution, the presence of degree correlations does not modify the result $\lambda_c = 0$ [97].

4.2.1 Random bimodal networks

We shall now illustrate to what extent the various levels of site approximations introduced above are able to reproduce the observed behaviour for the example of a SIS-type process on a random network where half the nodes have degree 3 and the other half has 6 nearest neighbours, in short $P(k) = (\delta_{k3} + \delta_{k6})/2$.

Fig. 4.1 reports the *prevalence*, that is the average number of infected individuals, in the stationary state versus the effective spreading probability at the level of simulation, if the state of every node is modelled by a probability [Eq. (4.1)] and if the evolution were subject to the degree-dependent mean-field equation (4.6). Both site approximations underestimate the observed threshold, that is, the numerical steady-state solution of Eq. (4.1) suggests $\lambda_c \simeq 0.195$ and Eq. (4.8) - the analytical solution of Eq. (4.6) for an uncorrelated network - results in $\lambda_c = 0.2$. This indicates that this shift is rooted in the approximations (4.3) and (4.5). Fig. 4.2 shows the quantities $\rho_{k'|k}$, ρ_k and $\rho_{k,k'}$ where $k, k' \in \{3, 6\}$ as measured from the numerical solution of Eq. (4.1). Panel (a) illustrates that the approximation (4.5) does not hold exactly; and when it comes to $\rho_{k,k}$, the approximation $\rho_{k,k} = \rho_{k|k}\rho_k$ overestimates the original values for $k = 3$ [panel (b)] and corrects the curve downwards for $k = 6$ [panel (d)]. The second approximation ($\rho_{k|k} = \rho_k$) shifts the curves back in the opposite direction. In the case $k \neq k'$ [panel (c)], assuming $\rho_{k,k'} = \rho_{k'|k}\rho_k$ lowers the corresponding slope whereas $\rho_{k'|k} = \rho_{k'}$ does not bring about a further correction.

In order to find a threshold value which lies closer to the prediction by the site approximation (4.1), we shall solve Eq. (4.4). Inspired from Fig. 4.2a, we make

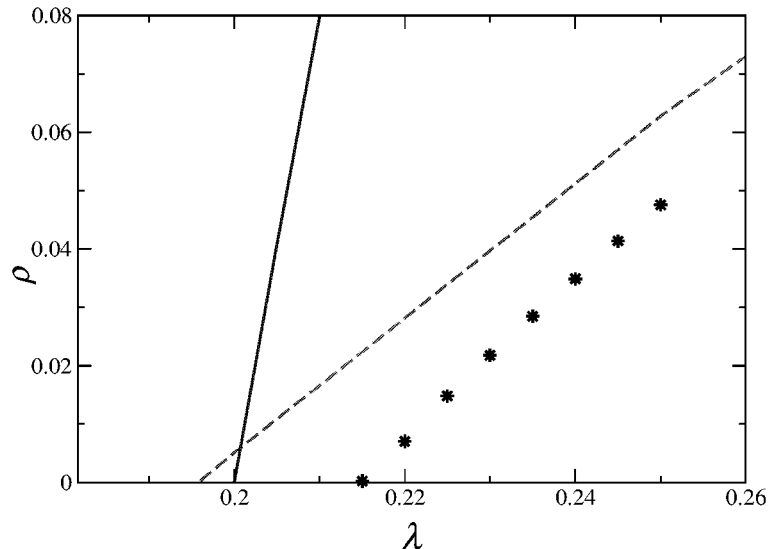


Figure 4.1: Fraction of infected individuals in the stationary state for a SIS-type process on a random bimodal network. Both the site approximation (dashed line) and the degree-dependent mean-field equation (solid line) underestimate the epidemic threshold obtained by the simulation result (stars). For the latter, averages over 10 realisations of networks of size $N = 10^4$ were taken where the system was relaxed into equilibrium from 10 different starting configurations for each realisation.

the ansatz $\rho_{k'|k} = (1 + c_{k'k})\rho_{k'}$ with $c_{33} = -0.18$, $c_{36} = 0.07$, $c_{63} = -0.08$ and $c_{66} = 0.05$ where the numerical values of c_{kl} are obtained from the slopes of $\rho_{k|l}$. Since $P(3|3) = P(3|6) = 1/3$ and $P(6|3) = P(6|6) = 2/3$, the stationary state of Eq. (4.4) reads

$$\begin{aligned}\rho_3 &= \lambda(1 - \rho_3)[(1 + c_{33})\rho_3 + 2(1 + c_{63})\rho_6] \\ \rho_6 &= 2\lambda(1 - \rho_6)[(1 + c_{36})\rho_3 + 2(1 + c_{66})\rho_6].\end{aligned}$$

In the vicinity of the epidemic threshold, $\rho_3, \rho_6 \ll 1$ and the above pair of equations can be linearised and written in the following matrix form

$$\begin{pmatrix} \lambda(1 + c_{33}) - 1 & 2\lambda(1 + c_{63}) \\ 2\lambda(1 + c_{36}) & 4\lambda(1 + c_{66}) - 1 \end{pmatrix} \begin{pmatrix} \rho_3 \\ \rho_6 \end{pmatrix} = 0.$$

Setting the determinant to 0 leads to a quadratic equation for λ , the solution of interest being $\lambda_c \simeq 0.195$ which is in good agreement with the numerical finding.

It is further worth noting that the difference of the threshold values as obtained from Eqs. (4.1) and (4.7) is larger for networks with richer degree distributions, i.e. if more degrees are present in the underlying network, as long as $\langle k^2 \rangle$ is not extremely large.

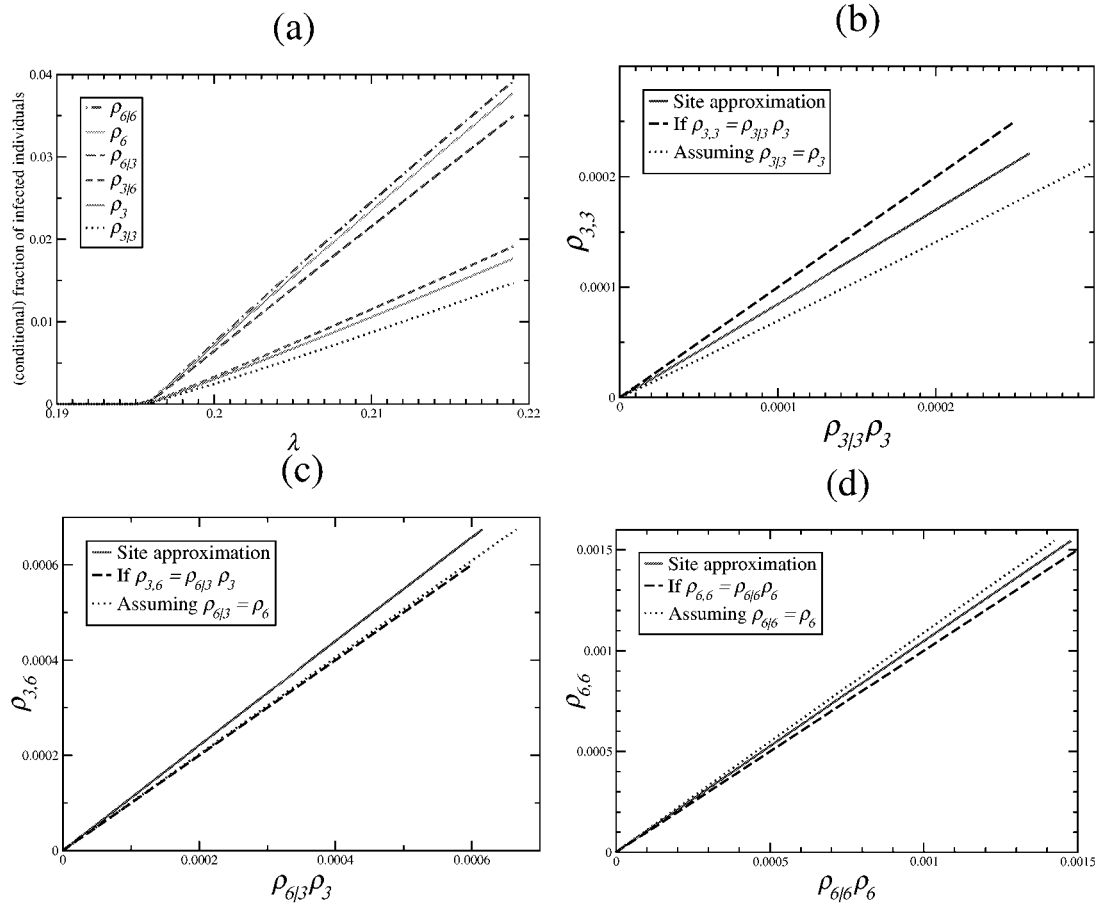


Figure 4.2: Validation of the approximations (4.3) and (4.5). Panel (a) reports the (conditional) fraction of infected individuals in the vicinity of the epidemic threshold, and the panels (b)-(d) illustrate how $\rho_{k,k'}$ changes upon performing the approximations mentioned above. The curves corresponding to the site approximations (solid lines) result from averaging the stationary state solutions of Eq. (4.1) over all nodes and over 10 realisations of networks of size $N = 10^4$.

4.2.2 Homogeneous networks

In a network where every node has K nearest neighbours, i.e. in a homogeneous network, the above considerations simplify considerably. Although Eq. (4.1) in its form still holds, we can take into account the homogeneity by saying that an arbitrarily chosen node is infected at time t with probability P^t , thus omitting the index i . This equation then reads

$$P^{t+1} = \lambda(1 - P^t)[1 - (1 - P^t)^K] \quad (4.9)$$

which is a mean-field approximation in its original sense since it involves only one field, namely P^t . From its linearised version

$$P^{t+1} = \lambda K(1 - P^t)P^t, \quad (4.10)$$

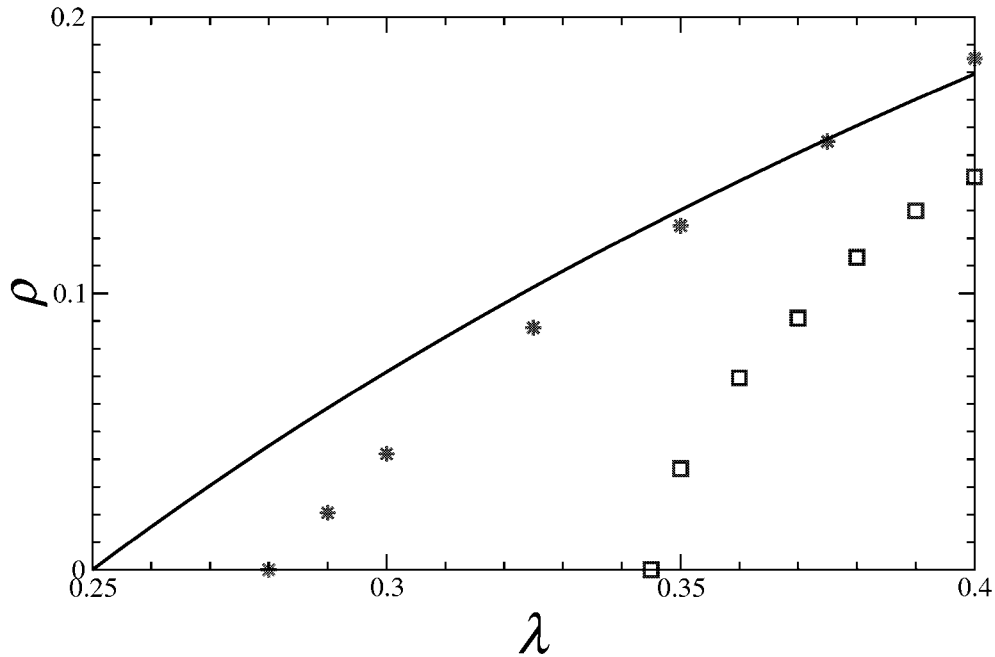


Figure 4.3: Epidemic spreading on homogeneous networks of degree 4. The average number of infected individuals ρ as a function of the effective spreading probability λ in the steady state is shown. The simulation result for the square lattice (squares) was obtained by relaxing the system (of size $N = 10^4$) into equilibrium for 10 different initial configurations. In the case of the random network (stars), we further averaged over 10 realizations of networks consisting of $N = 10^5$ nodes. This figure shows that the simulations exhibit higher epidemic thresholds with respect to the mean-field approximation.

we easily extract the threshold value

$$\lambda_c = \frac{1}{K}. \quad (4.11)$$

This is in agreement with Eq. (4.8) since $\langle k \rangle = K$ and $\langle k^2 \rangle = K^2$. It has to be stressed further that P^t can directly be interpreted as a density of infected individuals ρ^t ($= \rho_K^t = \rho_{K|K}^t$), due to the underlying topological homogeneity. Along these lines, Eq. (4.10) is the homogeneous analogue of each of the Eqs. (4.2), (4.4) and (4.6). Eq. (4.9) therefore also results in λ_c given by Eq. (4.11), i.e. the threshold value is not shifted as observed for inhomogeneous topologies.

Fig. 4.3 contrasts the mean-field prediction with simulation results for two types of networks of constant degree $K = 4$, namely a random homogeneous network and a square lattice. The random network exhibits a smaller epidemic threshold and it lies closer to the prediction $\lambda_c = 1/4$. Both of these observations are rather intuitive: (i) In a random network, global distances are small, making it easy for a virus to spread, thus even for a small effective spreading probability, a non-zero fraction of individuals is infected in the stationary state. (ii) The mean-

field approximation is also known as the homogeneous mixing hypothesis meaning that every individual is equally likely to transmit the virus to any other individual. This is just another way of saying that spatial correlations are not taken into account and also corresponds to ignore any underlying spatial structure. Since the random network comes closer to this assumption than the square lattice, the corresponding simulation result is in better agreement with that predicted by Eq. (4.11).

4.3 Exact formulation

In this section, we introduce the formalism that will serve as point of departure for deriving the methods described in Secs. 4.4 and 4.5. For completeness, it is also shown how the mean-field approximation (4.9), which was heuristically obtained in the previous section, is recovered.

On an exact level, we shall describe the epidemic dynamics by assigning a probability $\mathcal{P}_t(\mathbf{x})$ to each configuration \mathbf{x} at every instant of time t . The vector \mathbf{x} contains the states x_i of all the nodes i of the network, x_i being either 0 (susceptible) or 1 (infected). The system probabilities satisfy at any time t

$$\sum_{\mathbf{x}} \mathcal{P}_t(\mathbf{x}) = 1.$$

and evolve in time according to

$$\mathcal{P}_{t+1}(\mathbf{x}) = \sum_{\mathbf{y}} \mathcal{W}_{\mathbf{y} \rightarrow \mathbf{x}} \mathcal{P}_t(\mathbf{y}). \quad (4.12)$$

The transition matrix of the system $\mathcal{W}_{\mathbf{y} \rightarrow \mathbf{x}}$ is obtained from the matrix $W_{y_l \rightarrow x_l}^l$ which shall denote the probability that the state of the arbitrary site l changes from y_l to x_l , through

$$\mathcal{W}_{\mathbf{y} \rightarrow \mathbf{x}} = \prod_{l=1}^N W_{y_l \rightarrow x_l}^l,$$

N being the total number of nodes in the network. The matrix elements representing the probabilities for the possible events at the site l are given by the simplified version of the SIS model, namely

$$\begin{aligned} W_{1 \rightarrow 0}^l &= 1 & W_{0 \rightarrow 1}^l &= \lambda \left[1 - \prod_{j \text{ nnl}} (1 - y_j) \right] \\ W_{1 \rightarrow 1}^l &= 0 & W_{0 \rightarrow 0}^l &= 1 - \lambda \left[1 - \prod_{j \text{ nnl}} (1 - y_j) \right], \end{aligned}$$

or in a more compact form

$$W_{y_l \rightarrow x_l}^l = 1 - x_l + \lambda(2x_l - 1)(1 - y_l) \left[1 - \prod_{j \text{ nnl}} (1 - y_j) \right].$$

The products in the above expressions have to be taken over all the nearest neighbours j of node l . The factor $1 - \prod_{j \text{nn}l} (1 - y_j)$ is 1 if at least one $y_j = 1$ and 0 otherwise.

Before deriving our methods which give insight into the role of the loop structure, we show how the mean-field approximation (4.9) is retrieved through this formalism. At that level, the sites are considered independently from each other, and we write for the system probability

$$\mathcal{P}_t(\mathbf{x}) = \prod_{l=1}^N P_t(x_l), \quad (4.13)$$

i.e. the system is described by the single variable $P_t(1)$ [the probability of being susceptible is $P_t(0) = 1 - P_t(1)$]. Here the index i is again omitted since the underlying network is supposed to be homogeneous. Its dynamics is obtained from Eq. (4.12) by summing it over all possible configurations \mathbf{x} , x_0 held fixed

$$\sum_{\{x_j\}_{j \neq 0}} \mathcal{P}_{t+1}(\mathbf{x}) = \sum_{\mathbf{y}} \mathcal{P}_t(\mathbf{y}) \underbrace{\sum_{\{x_j\}_{j \neq 0}} \mathcal{W}_{\mathbf{y} \rightarrow \mathbf{x}}}_{W_{y_0 \rightarrow x_0}^0}, \quad (4.14)$$

where the node 0 can be chosen in an arbitrary way. The left hand side of the above equation corresponds to the probability that node 0 is in state x_0 at time $t + 1$. With the ansatz (4.13), the time evolution is

$$P_{t+1}(1) = \lambda P_t(0) [1 - P_t(0)^K].$$

which is equivalent to Eq. (4.9) since $P_t(1) = P_t$ and $P_t(0) = 1 - P_t$.

4.4 Two-step description

A strategy that serves to incorporate *local ordering* properties - i.e. the loop structure - is to take into account temporal correlations. Thus, departing from Eq. (4.12) and performing two time-steps exactly, we expect that the way the second neighbours are arranged, enters very naturally into the description. For example, the cases where two nearest neighbours of an arbitrary node are also directly connected (presence of a triangle), where they are linked via a second neighbour (giving rise to a loop of length 4) or where the only path goes through the original node (treelike structure), lead to different results. We now derive the general equation, special cases are then looked at within the following subsections.

As outlined above, we iterate Eq. (4.12) once

$$\mathcal{P}_{t+1}(\mathbf{x}) = \sum_{\mathbf{y}} \left[\mathcal{W}_{\mathbf{y} \rightarrow \mathbf{x}} \sum_{\mathbf{z}} \mathcal{W}_{\mathbf{z} \rightarrow \mathbf{y}} \mathcal{P}_{t-1}(\mathbf{z}) \right].$$

We now again pass to a site approximation. By summing the above equation over all possible configurations \mathbf{x} , x_0 held fixed, we get

$$P_{t+1}(x_0) = \sum_{\mathbf{z}} \left[\mathcal{P}_{t-1}(\mathbf{z}) \sum_{\{y_l\}_{l=0,1,\dots,K}} (W_{y_0 \rightarrow x_0} \prod_{j=0}^K W_{z_j \rightarrow y_j}) \right], \quad (4.15)$$

0 again being an arbitrarily chosen node. The system probability $\mathcal{P}_{t-1}(\mathbf{z})$ is given by Eq. (4.13), and the nodes $1, 2, \dots, K$ denote the nearest neighbours of the arbitrarily chosen node 0. If the set comprising the node 0, its nearest (nodes $1, 2, \dots, K$) and second neighbours is denoted by \mathcal{N}^2 , Eq. (4.15) can be written as

$$P_{t+1}(x_0) = \underbrace{\sum_{\{z_r\}_{r \notin \mathcal{N}^2}} \left[\prod_{s \notin \mathcal{N}^2} P_{t-1}(z_s) \right]}_1 \sum_{\{z_u\}_{u \in \mathcal{N}^2}} \left[\prod_{v \in \mathcal{N}^2} P_{t-1}(z_v) \sum_{\{y_l\}_{l=0,1,\dots,K}} (W_{y_0 \rightarrow x_0} \prod_{j=0}^K W_{z_j \rightarrow y_j}) \right]$$

since in the W -factors, only z -states associated to nodes belonging to \mathcal{N}^2 appear. A tour de force calculation then leads to

$$\begin{aligned} P_{t+1}(1) = & \lambda \left[\lambda \sum_{\alpha_1=1}^K \langle f_{\alpha_1} \rangle_{t-1} - \lambda^2 \left(\sum_{\alpha_1=1}^K \sum_{\alpha_2=\alpha_1+1}^K \langle f_{\alpha_1} f_{\alpha_2} \rangle_{t-1} + \sum_{\alpha_1=1}^K \langle f_0 f_{\alpha_1} \rangle_{t-1} \right) \right. \\ & + \lambda^3 \left(\sum_{\alpha_1=1}^K \sum_{\alpha_2=\alpha_1+1}^K \sum_{\alpha_3=\alpha_2+1}^K \langle f_{\alpha_1} f_{\alpha_2} f_{\alpha_3} \rangle_{t-1} + \sum_{\alpha_1=1}^K \sum_{\alpha_2=\alpha_1+1}^K \langle f_0 f_{\alpha_1} f_{\alpha_2} \rangle_{t-1} \right) + \dots \\ & \dots - (-\lambda)^K \left(\sum_{\alpha_1=1}^K \sum_{\alpha_2=\alpha_1+1}^K \dots \sum_{\alpha_K=\alpha_{K-1}+1}^K \langle f_{\alpha_1} f_{\alpha_2} \dots f_{\alpha_K} \rangle_{t-1} \right. \\ & \left. \left. + \sum_{\alpha_1=1}^K \sum_{\alpha_2=\alpha_1+1}^K \dots \sum_{\alpha_{K-1}=\alpha_{K-2}+1}^K \langle f_0 f_{\alpha_1} f_{\alpha_2} \dots f_{\alpha_{K-1}} \rangle_{t-1} \right) \right]. \quad (4.16) \end{aligned}$$

Thereby we introduced the variable

$$\begin{aligned} f_{\alpha} & \equiv (1 - z_{\alpha}) \left[1 - \prod_{\sigma \text{ nna}} (1 - z_{\sigma}) \right] \\ & = \begin{cases} 1 & \text{if } z_{\alpha} = 0 \text{ and at least one } z_{\sigma} = 1, \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

and the expectation value of a function of $\{z_k\}_{k \in \mathcal{N}^2}$,

$$\left\langle g(\{z_k\}_{k \in \mathcal{N}^2}) \right\rangle_t \equiv \sum_{\{z_k\}_{k \in \mathcal{N}^2}} \left[\prod_{l \in \mathcal{N}^2} P_t(z_l) g(\{z_m\}_{m \in \mathcal{N}^2}) \right] \quad (4.17)$$

for notational convenience. In the following, an expectation value of a product of n f -factors will be referred to as a term of n -th order although it is proportional to $\lambda \cdot \lambda^n$. As is illustrated in detail in the following subsection, every term of Eq. (4.16) corresponds to a subgraph of the graph composed of the nodes \mathcal{N}^2

and whose links are according to the network under investigation. It can already be anticipated that the first term accounts for the degree (distribution) only, whereas the contributions of higher order will give insight about the role of the loop structure.

4.4.1 Networks of degree four

In this subsection, we elaborate the implications of our two-step description for topologies where every node has 4 nearest neighbours, the analysis being restricted to the stationary state. In the left panel of Fig. 4.4, we show the simulation

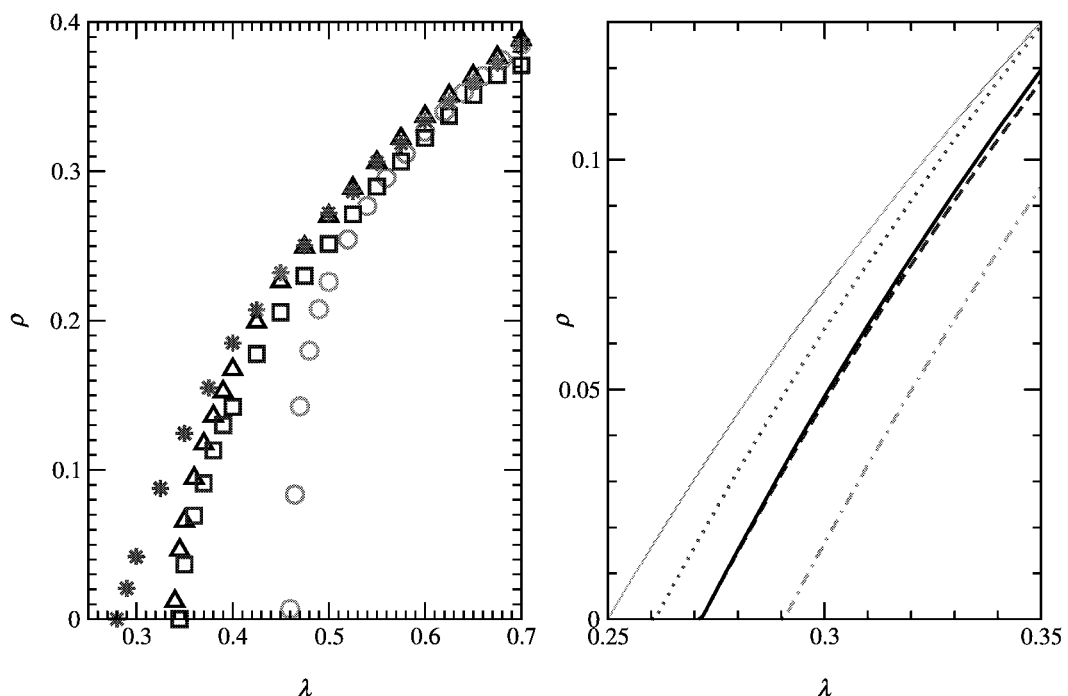


Figure 4.4: Equilibrium prevalences for the epidemic process on top of different networks characterized by $P(k) = \delta_{k4}$. Left: Simulation results for the homogeneous random graph (\star), the Kagomé lattice (\triangle), the square lattice (\square) and the ring-type network (\circ). These results were obtained by averaging over 10 different initial configurations for networks composed of $N = 10^4$ nodes. For the Molloy-Reed graph, we additionally averaged over 10 realisations of networks. Right: The ordinary mean-field approximation (thin solid) ignores (local) ordering properties and yields $\lambda_c = 0.25$ for any homogeneous network of degree 4. For the Molloy-Reed network (dotted), the Kagomé lattice (solid), the square lattice (dashed) and the ring (dotted-dashed), different steady-state prevalences are obtained at the two-step level.

results for the random homogeneous network as well as the regular graphs introduced in section 2.3. The ring-type network exhibits the largest epidemic threshold. For both the square and Kagomé lattices, the critical value is $\lambda_c \simeq 0.34$. We therefore anticipate that either four plaquettes or two triangles (per node) lead to the same effect in the regime of low prevalences. Finally, the lowest epidemic threshold is found if the population is arranged on a homogeneous random network (of degree 4). The last result is very intuitive since in such a graph, global distances are small, making it more easy for a virus to spread. Therefore, even if the effective spreading rate is rather low, a finite fraction of the population will be infected in the stationary state, hence the small value for the location of the onset of the epidemic. In summary, these results indicate that the poorer the loop structure, the lower the corresponding epidemic threshold.

In the right part of Fig. 4.4, the one-step (ordinary mean-field) and two-step site approximations are reported. The former corresponds to the steady-state solution of Eq. (4.9) for which $\rho = 0$ at $\lambda_c = 1/4$ according to Eq. (4.11). All the networks in question are therefore treated identically, the loop structure being ignored at this level of description. Yet, the two-step solutions [Eq. (4.15)] are diverse for the different graphs. Going from right to left, the curves correspond to the ring, the square lattice, the Kagomé lattice and the Molloy-Reed network, that is they appear in the same sequence as at the level of simulation. Furthermore, the curves corresponding to the Kagomé and square lattice also meet the x -axis at the same value of λ . It has to be noted that the two-step estimates for the threshold values are still considerably inaccurate especially for the ring and lattices, but this just highlights the presence of higher-order spatiotemporal correlations. However, the important point is that the degeneracy associated to the one-step description disappears at the two-step level.

On the basis of Eq. (4.16), we shall now analytically study the effect of local ordering properties upon the epidemic spreading, leading to a quantitative understanding of the threshold value.

Random network

We shall now evaluate all the terms of Eq. (4.16) for a locally treelike topology. Fig. 4.5 shows the subgraphs representing the terms in Eq. (4.16), in increasing order. Thereby the correspondence is as follows: Given the term $\langle f_\alpha f_\beta \rangle$, the nodes α and β are represented by filled circles whereas their nearest neighbours are drawn by empty circles. The links which enter at the level of the subgraph in question, are represented by solid lines whereas the ignored ones are dashed. If we denote the second neighbours of the central vertex 0 by $l1, l2, l3$ for $l = 1, 2, 3, 4$ and follow Eq. (4.17), the first order contribution for $\alpha_1 = 1$ (subgraph in Fig.

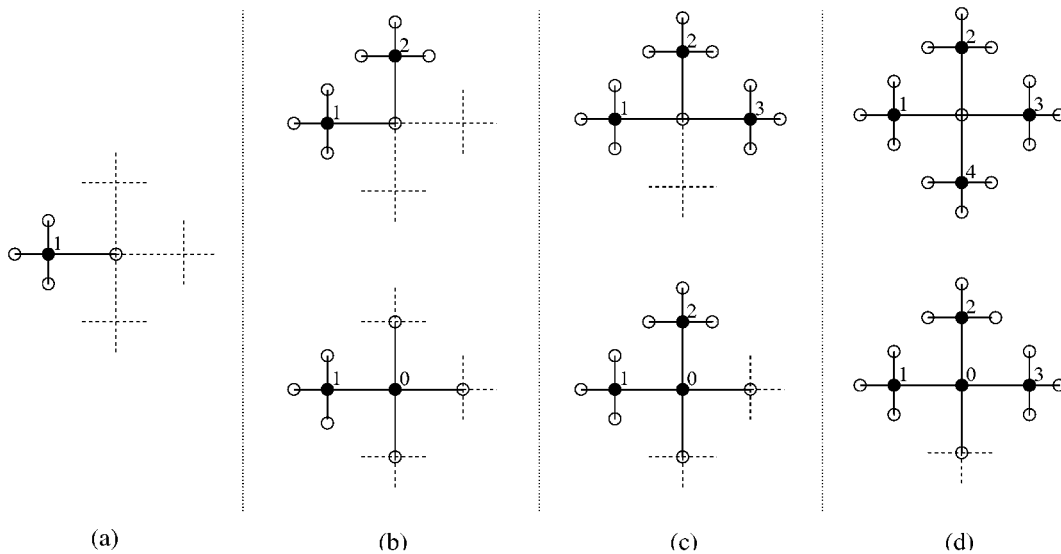


Figure 4.5: Pictorial representation of the terms of Eq. (4.16) for a treelike topology. The order of a specific subgraph is given by the number of filled circles, (a) corresponding to $\langle f_1 \rangle$, (b) to $\langle f_1 f_2 \rangle$ and $\langle f_0 f_1 \rangle$, and so forth. In this vein, the indices of the f -factors appearing in the expectation values correspond to filled circles whereas empty circles represent their nearest neighbours. The dashed lines are the links of the complete graph which are not contained in a specific subgraph.

4.5a) is

$$\begin{aligned} \langle f_1 \rangle &= \sum_{z_1} \sum_{z_0} \sum_{z_{11} \dots z_{13}} \left\{ P(z_1) P(z_0) P(z_{11}) P(z_{12}) P(z_{13}) \right. \\ &\quad \left. \times \underbrace{(1 - z_1) [1 - (1 - z_0)(1 - z_{11})(1 - z_{12})(1 - z_{13})]}_{f_1} \right\} \\ &= 4P + \mathcal{O}(P^2) \end{aligned}$$

where the sum over the z -variables to which no circles are associated, has been carried out trivially. Furthermore, we again have set $P \equiv P(1)$ in the third line (this is also done below), and the time index was omitted since we are only interested in the steady state. This term appears with multiplicity 4 (due to $\sum_{\alpha_1=1}^K$), giving the contribution $16P$ to first order in P .

Fig. 4.5b shows the subgraphs representing $\langle f_1 f_2 \rangle$ (upper part) and $\langle f_0 f_1 \rangle$ (lower part). Their contributions are

$$\begin{aligned} \langle f_1 f_2 \rangle &= \sum_{z_0} \sum_{z_1} \sum_{z_2} \sum_{z_{11} \dots z_{13}} \sum_{z_{21} \dots z_{23}} \left\{ P(z_0) P(z_1) P(z_2) \right. \\ &\quad \left. \times P(z_{11}) P(z_{12}) P(z_{13}) P(z_{21}) P(z_{22}) P(z_{23}) f_1 f_2 \right\} \\ &= P + \mathcal{O}(P^2), \end{aligned}$$

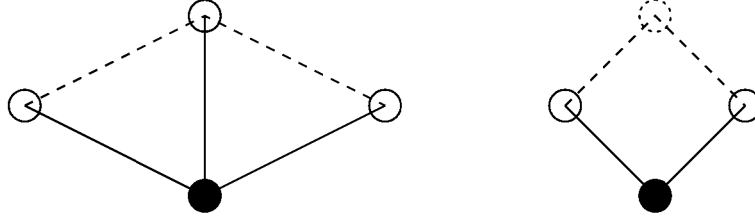


Figure 4.6: Loops of length 4 in a network. Primary quadrilaterals (left) are two adjacent triangles and therefore involve only nearest neighbours (solid empty circles) of the central node (filled circle). A secondary quadrilateral (right) involves a second neighbour (dotted empty circle) as well.

occurring $\binom{4}{2} = 6$ times and

$$\begin{aligned} \langle f_0 f_1 \rangle &= \sum_{z_0} \sum_{z_1} \sum_{z_2 \dots z_4} \sum_{z_{11} \dots z_{13}} \left\{ P(z_0) P(z_1) P(z_2) P(z_3) P(z_4) \right. \\ &\quad \left. \times P(z_{11}) P(z_{12}) P(z_{13}) f_0 f_1 \right\} \\ &= 9P^2 + \mathcal{O}(P^3), \end{aligned}$$

thus not giving a contribution to first order in P . As a consequence, the second-order term (that is the one proportional to $\lambda \cdot \lambda^2$) is $6P$.

As the procedure should now be clear, we only give the results for the remaining orders. The upper subgraph of Fig. 4.5c represents the term $\langle f_1 f_2 f_3 \rangle$. Its contribution is $P + \mathcal{O}(P^2)$. The lower subgraph corresponds to $\langle f_0 f_1 f_2 \rangle$, yielding $18P^3 + \mathcal{O}(P^4)$. As the former term has multiplicity $\binom{4}{3} = 4$, the total third-order contribution is $4P$.

As far as the fourth order is concerned (Fig. 4.5d), the subgraph involving node 0 as filled circle neither gives a contribution whereas the term $\langle f_1 f_2 f_3 f_4 \rangle$ having multiplicity 1 also gives $P + \mathcal{O}(P^2)$, thus totally yielding $\lambda \cdot \lambda^4 P$.

Collecting these results, we obtain the following condition that determines the epidemic threshold for a treelike topology

$$1 = \lambda(16\lambda - 6\lambda^2 + 4\lambda^3 - \lambda^4), \quad (4.18)$$

which is satisfied by $\lambda_c \simeq 0.2609$. This is the value that we found numerically (second curve from the left in the right panel of Fig. 4.4).

Graphs with loops

It is easy to imagine that the preceding analysis yields different results when triangles and loops of length 4 are present. Caldarelli *et al.* [32] have classified loops of length 4 in a complex network into *primary* and *secondary quadrilaterals* (Fig. 4.6). In the former case, the external vertices, which the loop is composed of, are all nearest neighbours whereas secondary quadrilaterals are plaquettes, the external nodes being two nearest and one second neighbour. With these topological

	E	Q_1	Q_2
square lattice	0	0	4
Kagomé lattice	2	0	0
ring	3	2	2

Table 4.1: Loop properties for our regular non-treelike networks.

measures, the loop structure of a strictly homogeneous graph can quantitatively be characterised as follows: By choosing an arbitrary node, the number of edges between its nearest neighbours is denoted by E . Q_1 and Q_2 shall refer to the number of primary and secondary quadrilaterals. For the networks in question, we report the corresponding values in Table 4.1.

Let us now look at the subgraph development for the square lattice whereby we focus on the important changes with respect to the treelike case. The full development is given in the appendix (Table A.1). We have already noticed that the first-order term is fully determined by the degree distribution, therefore the $\lambda \cdot \lambda$ coefficient is 16, as in the treelike case. At order 2, the term $\langle f_1 f_2 \rangle$ (upper subgraph of Fig. 4.5b) splits into two subgraphs in the presence of plaquettes (Fig. 4.7). The right subgraph is the same as in the treelike case, yet the left yields a contribution $2P + \mathcal{O}(P^2)$. Their multiplicities are 4 (left) and 2 (right) summing up to $\binom{4}{2} = 6$. The resulting $\lambda \cdot \lambda^2$ coefficient is therefore -10. Although different subgraphs enter into the development also at the orders ≥ 3 , the coefficients appearing in the equation determining the epidemic threshold do not change.

The second-order subgraphs for the Kagomé lattice are depicted in Fig. 4.8. Both contributions are $P + \mathcal{O}(P^2)$. The one involving a triangle appears 6 times whereas the right subgraph has multiplicity 4. We therefore obtain the same second-order coefficient as for the square lattice. An analysis for the higher-order subgraphs yields no difference with respect to the square lattice. These two cases

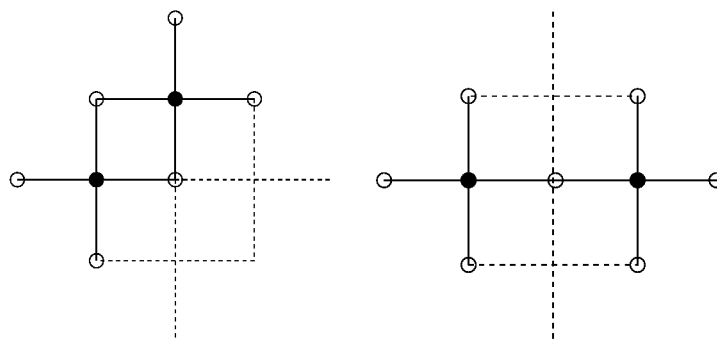


Figure 4.7: Second-order subgraphs not involving the central node as filled circle for the square lattice. To first order in P , the left subgraph yields the contribution $2P$, the right one P . For further explanations, see Fig. 4.5.

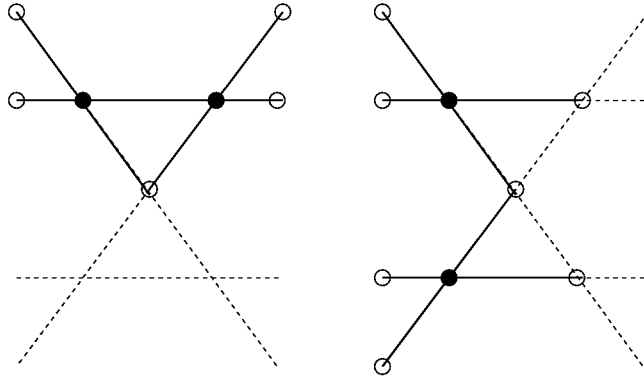


Figure 4.8: Subgraphs of second order for the Kagomé lattice, both contribute with $P + \mathcal{O}(P^2)$. See Fig. 4.5 for details regarding line- and circle styles.

are therefore equivalent at the two-step level for $P \ll 1$. In Table 4.2, we summarise the coefficients for these two lattices as well as for the ring, the full elaborations being given in the appendix (Tables A.2 and A.3).

Of course the idea is now to extend Eq. (4.18) such that it holds for all the investigated graphs. Our findings suggest that the local ordering properties enter in the following way into the equation determining the epidemic threshold:

$$1 = \lambda \left[16\lambda - (6 + 2E + Q_1 + Q_2)\lambda^2 + (4 + Q_1)\lambda^3 - \lambda^4 \right]. \quad (4.19)$$

However this is not the full story. What about loops of length 5? Let us argue why they do not enter in the framework of a two-step description. Although there exist such loops involving only first and second neighbours (Figs. 4.9a and b), the loop may be closed only between two second neighbours (Fig. 4.9c).

Obviously such a connection is ignored at the two-step level. In the language of graph theory [15] (p. 154), the latter case corresponds to a *fundamental* loop whereas the former examples can be reduced to loops of length 3 and 4. At the 2-step level, only loops up to length 4 enter into the description for the following reason: if the central node is infected at time t , it can causally affect only vertices two links away, corresponding to a chain of 4 links. Obviously, it matters whether the first and the last node of this chain are identical. In this case, we have a loop of length 4. Otherwise it cannot be distinguished whether the topology is fully

$\lambda \cdot \lambda^n$ - coeff.	$n = 1$	$n = 2$	$n = 3$	$n = 4$
square lattice	16	-10	4	-1
Kagomé lattice	16	-10	4	-1
ring	16	-16	6	-1

Table 4.2: Coefficients of the two-step threshold equation for our networks having in common $P(k) = \delta_{k4}$, but differing in the loop pattern.

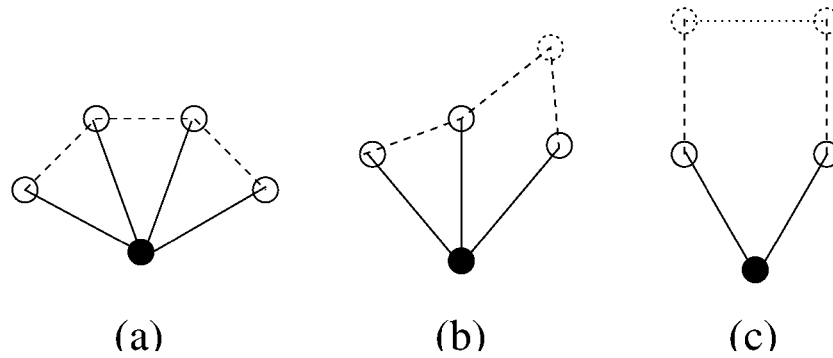


Figure 4.9: Loops of length 5 involving different hierarchies of nearest neighbours. Connections emanating from the central node (filled circle) are drawn as a solid line, links going from nearest neighbours (empty solid circles) to other nearest neighbours or second neighbours (empty dotted circles) are dashed, and second neighbours are connected by a dotted line. The *pentalateral* (a) involves nearest neighbours only, in (b) the loop traverses a second neighbour and in the one in (c) lacking in internal connections, a link between two second neighbours serves to close it.

treelike or if loops of length greater than 4 are present. Along these lines, it has to be expected that loops up to length $2n$ enter within an n -time-step description. In contrast, the presence of higher-order quadrilaterals modifies the coefficients of Eq. (4.19). Fig. 4.10a shows what we shall call a *tertiary* quadrilateral: the three nearest neighbours of the central node are all connected to another common node. Evidently, the presence of a tertiary quadrilateral implies $Q_2 = \binom{3}{2} = 3$ secondary quadrilaterals. In a *fourth-order* quadrilateral (Fig. 4.10b), 4 nodes share two common vertices as nearest neighbours, implying the presence of $Q_3 = \binom{4}{3} = 4$ tertiary quadrilaterals and $Q_2 = \binom{4}{2} = 6$ secondary quadrilaterals. In a network of degree $K > 4$, quadrilaterals up to order K can in principle be found.

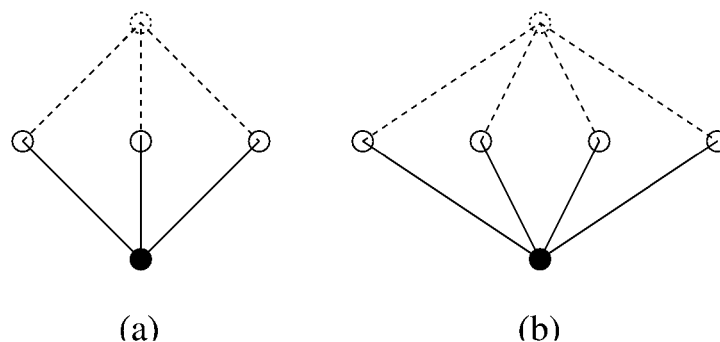


Figure 4.10: Quadrilaterals of orders 3 and 4. The n (a: $n = 3$, b: $n = 4$) nearest neighbours (empty solid circles) of the central node (filled circle) share another vertex (dotted empty circle) as nearest neighbour.

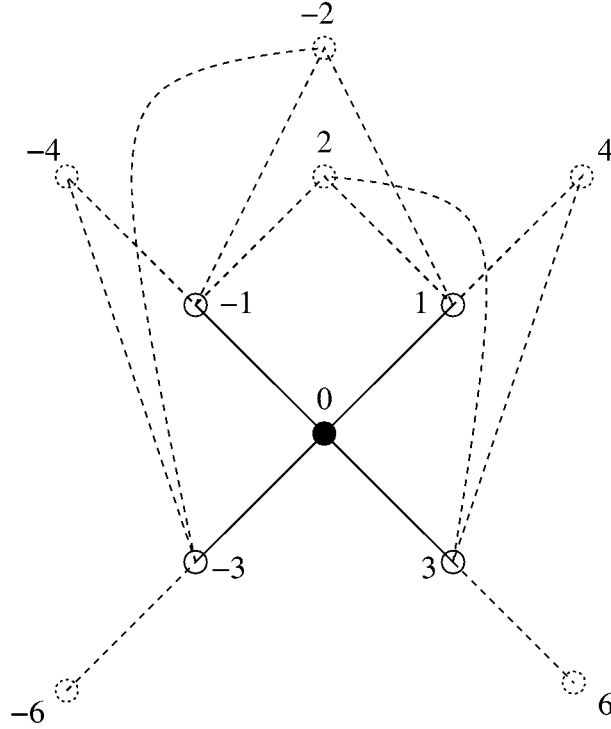


Figure 4.11: This figure visualises how the nearest (empty solid circles, vertices $\{-3,-1,1,3\}$) and second neighbours (empty dotted circles) of an arbitrarily chosen node (filled circle, node 0) are arranged in a one-dimensional lattice with additional connections between sites 3 units apart. The sets $\{0,\{-1,1,3\},2\}$ as well as $\{0,\{-3,-1,1\},-2\}$ are forming tertiary quadrilaterals.

A one-dimensional lattice with additional connections between the nodes i and $i+3$ for all i (instead of $i+2$ as in the ring investigated up to now) possesses the neighbourhood structure shown in Fig. 4.11, i.e. it is characterized by $E = Q_1 = 0, Q_2 = 8, Q_3 = 2$ and $Q_4 = 0$. By applying our formalism to this case and to a network that has fourth-order quadrilaterals, Eq. (4.19) generalises to

$$1 = \lambda \left[16\lambda - (6 + 2E + Q_1 + Q_2)\lambda^2 + (4 + Q_1 + Q_3)\lambda^3 - (1 + Q_4)\lambda^4 \right], \quad (4.20)$$

the coefficients of order 3 and 4 being modified only.

Introducing disorder

The networks considered up to now lack in the small world property characterising social networks on which the epidemic process is occurring. By starting with a ring-like network where nodes two units apart are also directly connected and repeating the second rewiring procedure described in section 2.3 a certain number of times, we obtain graphs of fixed degree $K = 4$ that are simultaneously

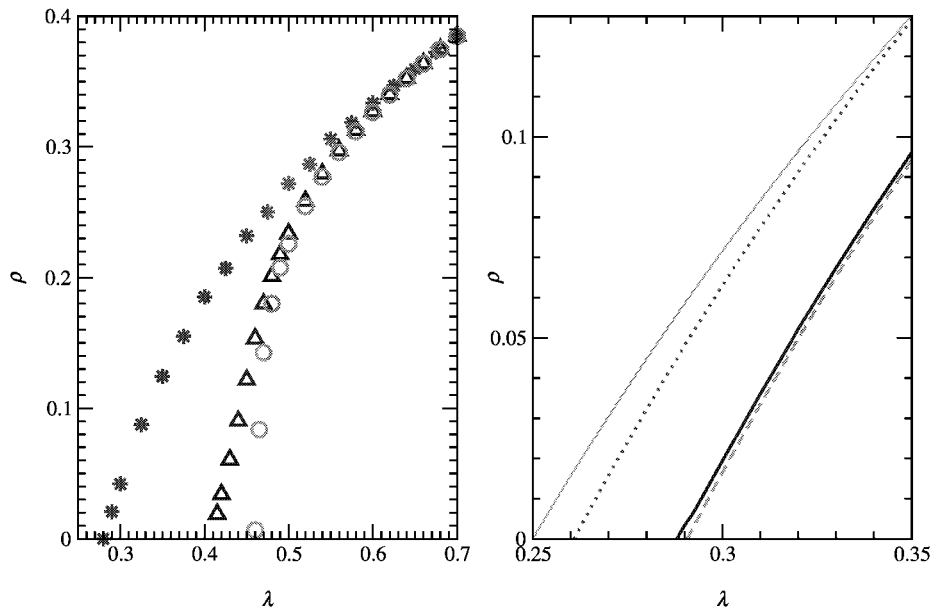


Figure 4.12: Steady-state prevalence for the ring-type network having undergone different degrees of randomisation ($N = 10^4$, 10 different initial configurations, 10 realisations for the cases involving randomness). Left: Simulation result for the original network (\circ), its partially rewired version (\triangle) and the entirely random network (\star). Right: The conventional mean-field approximation (thin solid) treats all the networks in question identically whereas at the two-step level, the Molloy Reed network (dotted), the partially randomized ring (solid) and the fully ordered ring (dashed) appear in the same sequence as in the left panel.

highly clustered, and in which the average distance between pairs of nodes is small [17].

The left part of Fig. 4.12 reports the simulation result for the equilibrium prevalence of the epidemic process on the disordered ring. Systems of size $N = 10^4$ were used, and the rewiring procedure was repeated $n = 100$ times. For completeness, the two limiting cases (fully ordered ring and random network) are also depicted. This panel shows the considerable effect on the steady-state spreading behaviour of the rather small number of re-wirings.

The right panel of this figure depicts the ordinary mean-field approximation (predicting $\lambda_c = 0.25$ for all cases) and the numerical solutions of the two-step description (4.15). It has to be noted that for the partially rewired ring lacking in strict homogeneity, Eq. (4.15) was solved at every node, therefore involving the set $P_i(x_i)$, $i = 1, 2, \dots, N$, the resulting prevalence being given by $1/N \sum_{i=1}^N P_i(1)$. Again at the double-step level, the networks in question are treated differently, as it could already be observed in Fig. 4.4. Here the curve corresponding to

the partially randomised ring lies closer to the original network in proportion to the simulation result. This is due to the small world property which can have a considerable effect on the location of the onset of the epidemic. Obviously, the simulation result uncovers the real effect of this global topological property whereas at the two-step level, it is the slightly poorer loop structure that accounts for the corresponding shift in the epidemic threshold.

Of course the quantities E, Q_1, Q_2, Q_3 and Q_4 are no longer reasonable for a partially randomised network due to the lack of strict homogeneity, but rather its local ordering properties can be quantified by averaging these values over all the nodes of the network. The emerging topological parameters \bar{E} and \bar{Q}_i ($i = 1, 2, 3, 4$) are essentially the *clustering-* (up to the factor $K(K-1)/2 = 6$) and *grid-coefficients*, i.e. the densities of triangles and loops of length 4 [17, 32]. We may therefore replace E and the number of quadrilaterals (of the different orders) in Eq. (4.20) by its mean-values, yielding the following estimate for the epidemic threshold condition

$$1 = \lambda \left[16\lambda - (6 + 2\bar{E} + \bar{Q}_1 + \bar{Q}_2)\lambda^2 + (4 + \bar{Q}_1 + \bar{Q}_3)\lambda^3 - (1 + \bar{Q}_4)\lambda^4 \right]. \quad (4.21)$$

For our partially randomised ring, we have $\bar{E} = 2.883, \bar{Q}_1 = 1.886, \bar{Q}_2 = 1.958$ and $\bar{Q}_3 = \bar{Q}_4 = 0$, leading to $\lambda_c \simeq 0.2892$. This value corresponds approximately to where the corresponding curve in the right part of Fig. 4.12 meets the x -axis.

4.4.2 Arbitrary degree

The implications of our two-step description have been illustrated for homogeneous networks of degree 4 in the previous subsection. This was a convenient choice as there exists a number of familiar simple graphs obeying $P(k) = \delta_{k4}$, differently ordered. Of course our formalism enables us to generalise the obtained threshold condition (4.20) to an arbitrary degree K , which is the subject of this subsection.

Let us again look at a fully treelike network, using Eqs. (4.16) and (4.18) as guidelines. The λ^2 -coefficient 16 incorporating the degree distribution is simply $4 \cdot 4$ since $\langle f_\alpha \rangle = 4P + \mathcal{O}(P^2)$ and α runs from 1 to 4. The remaining coefficients -6, 4 and -1 correspond to the binomial coefficients $-\binom{4}{2}, \binom{4}{3}$ and $-\binom{4}{4}$. Indeed the threshold equation for a treelike network of degree K derived by Eq. (4.16) is

$$1 = \lambda \left[\lambda K^2 - \sum_{\kappa=2}^K \lambda^\kappa \binom{K}{\kappa} \right]. \quad (4.22)$$

Repeating the graph developments for homogeneous networks characterised by different values of K and varying loop structures reveals that the very same

correction terms enter into Eq. (4.22), yielding

$$\begin{aligned}
1 = \lambda & \left\{ \Theta(K-1)K^2\lambda \right. \\
& - \Theta(K-2) \left[\binom{K}{2} + 2E + Q_1 + Q_2 \right] \lambda^2 \\
& + \Theta(K-3) \left[\binom{K}{3} + Q_1 + Q_3 \right] \lambda^3 \\
& \left. - \sum_{\kappa=4}^K \Theta(K-\kappa) \left[\binom{K}{\kappa} + Q_\kappa \right] (-\lambda)^\kappa \right\}
\end{aligned} \tag{4.23}$$

where E is again the number of connections between the nearest neighbours of an arbitrarily chosen node, Q_n denotes the number of quadrilaterals of order n and $\Theta(x)$ is the step-function defined by

$$\Theta(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

The equation that describes the role of local ordering in a disordered network subject to $P(k) = \delta_{k,K}$ is again obtained simply by replacing the corresponding quantities by their mean-values ($E \rightarrow \bar{E}$, $Q_i \rightarrow \bar{Q}_i$) in Eq. (4.23), providing an improved estimate for the epidemic threshold.

4.5 Cluster approximations

Performing two steps exactly in the epidemic dynamics - as it was done in the previous section - is only one way to learn about the role of topological properties beyond the degree in the spreading behaviour. Another approach consists in taking into account spatial correlations, that is within this section, we shall abandon the assumption that the states of two connected nodes evolve in time independently from one another. The method is illustrated for the conventional SIS model, and we give the results for the continuous-time limit. The fact that much attention has been given to this limit of this model mainly led to this choice, thus allowing for a comparison with existing approaches. While the effective spreading *probability* λ ($= \lambda\Delta t$ since $\Delta t = 1$) was the central parameter in the above discrete-time considerations, this role is now played by the effective spreading *rate* λ which corresponds to a probability per unit time: infected nodes infect neighbouring susceptible nodes with probability $\lambda\Delta t$ ($\Delta t \rightarrow 0$). By modelling the spreading dynamics in terms of rates, the mean-field approximation can be derived in a somewhat different way than that described in Sec. 4.2. With this rate-based heuristic approach, one can obtain a pair approximation, that is a description which takes into account pair correlations. Up to this level of correlations, this is indeed a reasonable strategy. But if one wants to keep track of higher-order correlations (e.g. the density of plaquettes of four infected nodes

in the case of the square lattice), a more general starting point reveals itself as advantageous.

4.5.1 Revisiting the mean-field and pair approximations

In this subsection, we first sketch the rate-based heuristic approach and then derive the mean-field and pair approximations by using an exact description, which was introduced in Sec. 4.3, formulated for the conventional SIS model. The various higher-order approximations are elaborated in the following subsection.

Conventional approach

The rate of change of an average quantity f (such as the fraction of sites in a particular state) is described as

$$\dot{f} = \sum_{x \in X} \sum_{e_x \in E_x} r(e_x)(f_{e_x} - f), \quad (4.24)$$

where X is the set of all sites, and E_x represents the set of all events that can occur at x . A particular event e_x changes the average from f to f_{e_x} and occurs at rate $r(e_x)$ [59].

At the mean-field level, the dynamics is described in terms of the density of infected individuals ρ_1 , and the fraction of susceptible nodes obeys $\rho_0 = 1 - \rho_1$. If we interpret ρ_1 as the probability that an arbitrarily chosen node is infected, ρ_1 can be altered either through recovery or infection. A recovery changes ρ_1 to 0 and occurs at rate 1 while an infection changes ρ_1 to 1 and occurs at rate $\lambda K \rho_1$ since any *infected* node can infect neighbouring nodes at a rate λ . As a consequence, Eq. (4.24) reads

$$\dot{\rho}_1 = 1(0 - \rho_1) + \lambda K \rho_1(1 - \rho_1) = -\rho_1 + \lambda K \rho_0 \rho_1. \quad (4.25)$$

At this level of description, pair correlations are thus fully ignored.

In the framework of the standard pair approximation [58], the dynamics is described in terms of the doublet densities ρ_{xy} ($x, y \in \{0, 1\}$), this quantity corresponds to the probability that a randomly chosen pair is in configuration (xy) . They are related to the global densities ρ_x and local densities (conditional probabilities) $\rho_{x|y}$ by: $\rho_{xy} = \rho_{yx} = \rho_x \rho_{y|x} = \rho_y \rho_{x|y}$. The global and local densities satisfy

$$\sum_{x=0}^1 \rho_x = 1 \quad \text{and} \quad \sum_{x=0}^1 \rho_{x|y} = 1 \quad \text{for any } y \in \{0, 1\}$$

Eq. (4.24) tells that the density of infected individuals and the doublet density ρ_{11} evolve in time according to

$$\begin{aligned} \dot{\rho}_1 &= -\rho_1 + \lambda K \rho_{0|1} \rho_1 \\ \dot{\rho}_{11} &= -2\rho_{11} + 2\lambda \rho_{10} + 2\lambda(K-1)\rho_{1|01}\rho_{10}. \end{aligned} \quad (4.26)$$

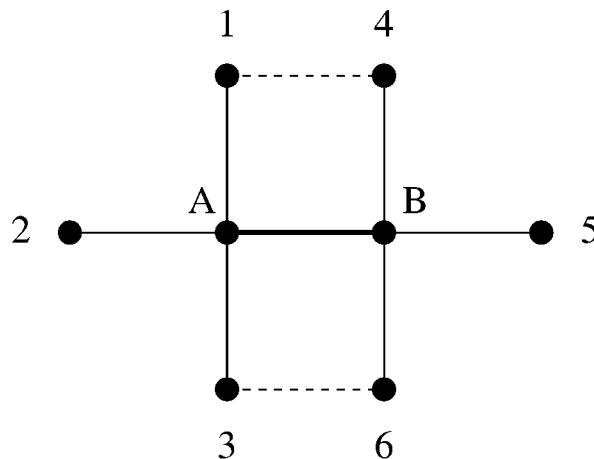


Figure 4.13: An arbitrarily chosen link and its nearest neighbourhood within a homogeneous network characterized by the degree distribution $P(k) = \delta_{k4}$. The dashed lines indicate the connections which are present in the case of a square lattice.

The first of Eqs. (4.26) can also be regarded as the result of substituting ρ_0 by $\rho_{0|1}$ in Eq. (4.25), i.e. the susceptible node that becomes infected has to be a nearest neighbour of the vertex which will transmit the infective agent. The second of Eqs. (4.26) consists of a recovery term and two transmission terms. More precisely, the first term describes the destruction of (11)-pairs which are being transformed into either (10)- or (01)-pairs. Both transitions occur at rate 1 (i.e. the recovery rate), hence the factor 2. The two remaining terms correspond to the creation of (11)-pairs. The factor 2 in these terms is needed because we do not assume any asymmetry between sites, which means $\rho_{10} = \rho_{01}$. A (11)-pair can be created from a (10)-pair either if the infective agent proceeds along the connection within that pair (second term) or if the susceptible node is infected by one of the other $K - 1$ nearest neighbours of it (third term, see also Fig. 4.13). This path involves the conditional probability $\rho_{1|01}$ [i.e. the probability of finding an infected node adjacent to a (01)-pair] which is approximated by $\rho_{1|0}$ as in the ordinary pair approximation, only nearest-neighbour correlations are taken into account. In order to solve the Eqs. (4.26), the system has to be closed. The set $\rho_1, \rho_{1|1}$ is a suitable choice, but ρ_{11}, ρ_{10} works equally well.

Fig. 4.14 contrasts the solutions of Eqs. (4.25) and (4.26) with the simulations for two different homogeneous networks of degree $K = 4$, i.e. a square lattice and a random network where 4 links are attached to every node³. The pair approximation provides a rather good description of the equilibrium dynamics on top of a

³This figure is the continuous-time analogue of Fig. 4.3 for the conventional SIS model, except that the corresponding pair approximation was not shown in the earlier figure. The quantitative differences can be seen by comparing the threshold values implied by the simulation results: the thresholds of the rates are higher than those of the probabilities.

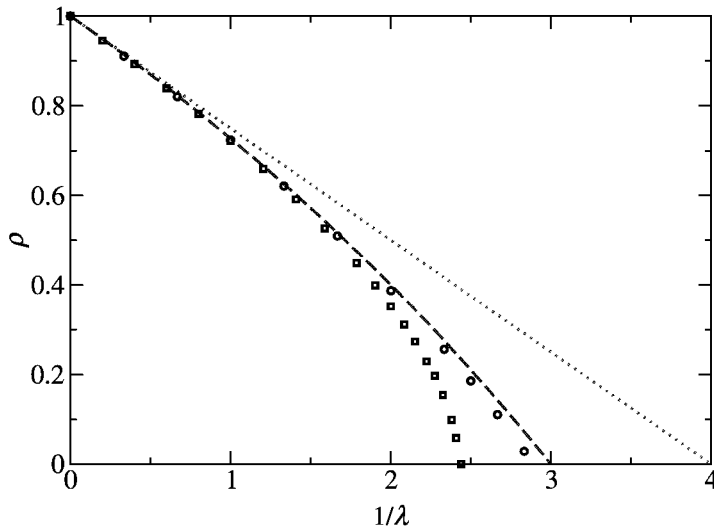


Figure 4.14: Epidemic spreading on homogeneous networks of degree 4. The average number of infected individuals ρ (prevalence) as a function of the inverse effective spreading rate $1/\lambda$ in the steady state is shown. The simulation result for the square lattice (squares) was obtained by relaxing the system (of size $N = 10^4$) into equilibrium for 10 different initial configurations, henceforth this shall be referred to as the number of iterations. In the case of the random network (circles), we further averaged over 10 realisations of networks consisting of $N = 10^5$ nodes. The adopted time-step was $\Delta t = 0.01$ for both examples. This figure shows that the simulations exhibit higher epidemic thresholds with respect to the approximations. The mean-field description (dotted line) yields $\lambda_c = 1/4$ whereas the pair approximation (dashed line) leads to $\lambda_c = 1/3$ for the epidemic threshold. The latter is also in better agreement with the simulation results for $1/\lambda \rightarrow 0$.

random homogeneous network, whereas the deviation from the simulation result is remarkable if the population is arranged on a square lattice whose topology is characterized by the presence of many loops of short length.

We shall now adapt the formalism introduced in Sec. 4.3 to the conventional SIS model. While this description enabled us to develop our two-step approximation, in this section it will serve as a starting point in order to investigate the role of spatial correlations beyond the pair level.

The conventional SIS model at an exact level

With respect to the formulation developed in Sec. 4.3, it is merely the factor $W_{y_l \rightarrow x_l}^l$ which changes since we are no longer using the simplified SIS model. We therefore again describe the dynamics in terms of $\mathcal{P}_t(\mathbf{x})$ which is the probability

that the system is in configuration \mathbf{x} at time t and evolves in time according to

$$\mathcal{P}_{t+\Delta t}(\mathbf{x}) = \sum_{\mathbf{y}} \mathcal{W}_{\mathbf{y} \rightarrow \mathbf{x}} \mathcal{P}_t(\mathbf{y}). \quad (4.27)$$

This is a more general formulation than Eq. (4.12) which allows for the derivation of the continuous-time limit $\Delta t \rightarrow 0$ at a later stage. The transition matrix appearing in Eq. (4.27) can be written as

$$\mathcal{W}_{\mathbf{y} \rightarrow \mathbf{x}} = \prod_{l=1}^N W_{y_l \rightarrow x_l}^l, \quad (4.28)$$

where the local transition matrix $W_{y_l \rightarrow x_l}^l$ is given by the conventional SIS model. Recalling that infected nodes recover spontaneously with rate 1 and infect neighbouring susceptible nodes with rate λ , we have

$$\begin{aligned} W_{1 \rightarrow 0}^l &= \Delta t & W_{0 \rightarrow 0}^l &= \prod_{j \text{ nnl}} (1 - y_j \lambda \Delta t) \\ W_{1 \rightarrow 1}^l &= 1 - \Delta t & W_{0 \rightarrow 1}^l &= 1 - \prod_{j \text{ nnl}} (1 - y_j \lambda \Delta t), \end{aligned}$$

where the products have to be taken over the nearest neighbours of site l . By using the binary variable x_l in addition to y_l , the above expressions are summarised as

$$W_{y_l \rightarrow x_l}^l = x_l + (1 - 2x_l) \left[y_l \Delta t + (1 - y_l) \prod_{j \text{ nnl}} (1 - y_j \lambda \Delta t) \right]. \quad (4.29)$$

The above set of equations will serve as starting point for various approximations, be it in discrete or continuous time. In the latter case, only the terms up to order 1 in Δt have to be taken into account, but this limit shall be carried out later on. As most of the existing methods are formulated in continuous time, we will elaborate the approximations for this case in order to allow for a comparison.

Derivation of the mean-field and pair approximations

In the following, it is shown how the approximations (4.25) and (4.26) are recovered from the exact description (4.27).

By repeating the procedure described in the second part of Sec. 4.3 with $W_{y_l \rightarrow x_l}^l$ given by Eq. (4.29), one finds for the probability that an arbitrarily chosen site is infected at time $t + \Delta t$

$$P_{t+\Delta t}(1) = 1 - \Delta t P_t(1) - P_t(0) [1 - \lambda \Delta t P_t(1)]^K \quad (4.30)$$

whose continuous-time limit ($\Delta t \rightarrow 0$) is

$$\dot{P}(1) = -P(1) + \lambda K P(0) P(1),$$

which is easily identified with Eq. (4.25) since $P(1) = \rho_1$ and $P(0) = \rho_0$.

Let us now see how the pair approximation is obtained by using our formalism. For this purpose, we sum Eq. (4.27) over all possible configurations, x_A and x_B held fixed, where A and B are the two sites of an arbitrarily chosen pair

$$\sum_{\{x_i\}_{i \notin \{A,B\}}} \mathcal{P}_{t+\Delta t}(\mathbf{x}) = \sum_{\mathbf{y}} \mathcal{P}_t(\mathbf{y}) \underbrace{\sum_{\{x_i\}_{i \notin \{A,B\}}} \mathcal{W}_{\mathbf{y} \rightarrow \mathbf{x}}}_{W_{y_A \rightarrow x_A}^A W_{y_B \rightarrow x_B}^B}. \quad (4.31)$$

The left hand side of the above equation corresponds to the probability that the pair AB is in state $(x_A x_B)$ at time $t + \Delta t$, which shall be denoted by $P_{t+\Delta t}(x_A x_B)$. By adopting the enumeration introduced in Fig. 4.13, we obtain from Eq. (4.29) for the transition probability $(y_A y_B) \rightarrow (x_A x_B)$

$$\begin{aligned} W_{y_A \rightarrow x_A}^A W_{y_B \rightarrow x_B}^B &= \tau_A \tau_B + \Delta t (1 - 2x_A) [y_A - \lambda(1 - y_A)(y_B + y_1 + y_2 + y_3)] \tau_B \\ &\quad + \Delta t (1 - 2x_B) [y_B - \lambda(1 - y_B)(y_A + y_4 + y_5 + y_6)] \tau_A \end{aligned} \quad (4.32)$$

where the linearisation in Δt has been carried out at this point due to technical convenience and

$$\tau_i = \tau_i(x_i, y_i) \equiv x_i + (1 - 2x_i)(1 - y_i), \quad (4.33)$$

an abbreviation which will also be used below. Furthermore the expression (4.32) only involves state variables y_i where i is either A, B or one of its nearest neighbours. The sum over the remaining y_j is therefore carried out trivially. Taking into account correlations up to range 2 only, we write for the probability that the pair AB and its nearest neighbours are in given states

$$\begin{aligned} P_t \begin{pmatrix} y_1 & y_4 \\ y_2 & y_A & y_B & y_5 \\ y_3 & y_6 \end{pmatrix} &= P_t(y_A y_B) \cdot P_t(y_1|y_A) P_t(y_2|y_A) P_t(y_3|y_A) \\ &\quad \times P_t(y_4|y_B) P_t(y_5|y_B) P_t(y_6|y_B). \end{aligned} \quad (4.34)$$

The conditional probabilities in the above ansatz are expressed as

$$P(y_i|y_A) = \frac{P(y_i y_A)}{P(y_A)},$$

where $P(y_A) = \sum_{x=0}^1 P(x y_A)$. Using this ansatz and performing the remaining summations, the continuous-time limit of Eq. (4.31) leads to the system (for general K)

$$\begin{aligned} \dot{P}(00) &= 2P(10) \left[1 - \lambda(K-1) \frac{P(00)}{P(0)} \right] \\ \dot{P}(10) &= P(11) - P(10) + \lambda P(10) \left[2(K-1) \frac{P(00)}{P(0)} - K \right] \\ \dot{P}(11) &= -2P(11) - 2\lambda P(10) \left[(K-1) \frac{P(00)}{P(0)} - K \right]. \end{aligned} \quad (4.35)$$

By identifying the pair probabilities $P(xy)$ with the doublet densities ρ_{xy} and since $\rho_{00}/\rho_0 = 1 - \rho_{10}/\rho_0$, the system of Eqs. (4.35) corresponds to Eqs. (4.26).

In summary, in the conventional derivation of the mean-field and pair approximations based on Eq. (4.24) - described at the beginning of this subsection - the rate of change of an average density is directly expressed by all the different events that can alter its value in a rather heuristic way [Eqs. (4.25) and (4.26)]. As we showed, this derivation of the approximations becomes an automatic procedure involving

- an initial summation of the system probability $\mathcal{P}_{t+\Delta t}(\mathbf{x})$ over almost all possible states in order to obtain $P_{t+\Delta t}(x)$ or $P_{t+\Delta t}(x_A x_B)$ [Eqs. (4.14) and (4.31)],
- an ansatz corresponding to the approximation [Eqs. (4.13) and (4.34)],
- and the continuous-time limit.

However, the last step is not really imperative. Our methodology works equally well in discrete time. If Δt is set to 1, $\lambda \Delta t = \lambda$ then corresponds to a probability rather than to a rate and higher-order terms in λ appear in the equations⁴. As an example, the discrete-time evolution at the mean-field level is governed by Eq. (4.30). Obviously, the results quantitatively differ from the continuous-time limit as was already pointed out in the context of Figs. 4.14 and 4.3. The full advantage of this formalisation will become clear in the next subsection.

It is also important to note that topological properties beyond the degree distribution do not enter at the level of the standard pair approximation. In the case of the square lattice, the nodes 1 and 4 as well as 3 and 6 are connected (in Fig. 4.13) whereas these links are missing in its random counterpart. Various improvements upon the ordinary pair approximation have been proposed. Instead of deriving the higher-order correlations from the dynamics of the system, these pair models consist in making a number of biologically motivated assumptions involving parameters that characterize the topology of the underlying network [59, 60, 99]. We shall compare our approach with these improved pair models in the next section.

4.5.2 Further Systematic Improvement

The difference between the simulation results and the pair approximation in Fig. 4.14 is rooted in the negligence of correlations of range greater than 2. Especially slightly above the onset of the epidemic, where a small fraction of the nodes is infected, the pairs of sites should not be considered independently, and higher-order dynamical correlations have to be taken into account. In other words, the state x_i of node i at time $t + \Delta t$ is determined by all the states of its

⁴This approach was pursued for the simplified SIS model in the previous section.

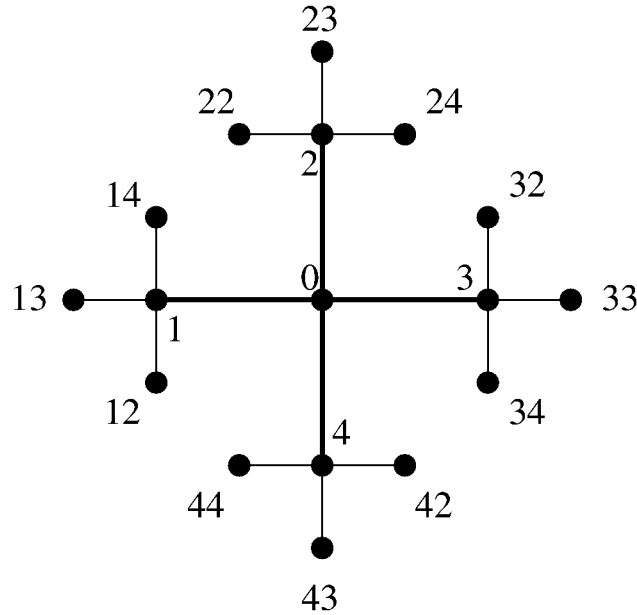


Figure 4.15: An arbitrary node (denoted by 0) with its corresponding star-like fundamental cluster within a homogeneous random network of degree $K = 4$.

nearest neighbours, i.e. it is not the case that the states of the various nearest neighbours at time t contribute independently from each other to the state x_i at time $t + \Delta t$.

We therefore want to incorporate the longer correlation range by extending the fundamental cluster (site, pair) to a star or square, respecting the underlying network's topology. Different spatial patterns are thus embedded very naturally by our method. The equations, to which the dynamics of the higher-order correlations are subject to, are derived in a very straightforward way by our formalism. The binary nature of these equations allows for a very efficient solution by the computer. On the other hand, the equations can be simplified further by taking into account the underlying symmetries. This procedure will be illustrated for the triangular and square lattices. Performing this extension, we find an improved description of the steady state as well as the dynamics.

Alternatively, it is possible to derive the dynamics of triple correlations by using Eq. (4.24) [98]. Although this approach has the advantage that no specific cluster must be chosen, it is a rather difficult undertaking.

Homogeneous Random Network

As the local topology is fully treelike, we shall use a star as our fundamental element. In contrast to regular lattices, this extension is a unique choice. Fig. 4.15 shows an arbitrarily chosen node in a homogeneous random network and two hierarchies of its nearest neighbours, also introducing the notation which is adopted below.

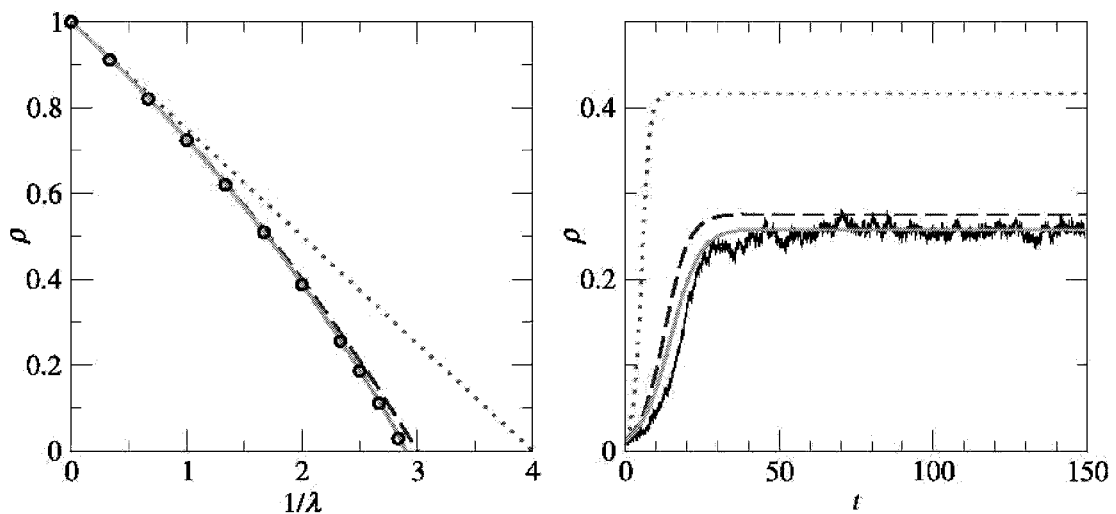


Figure 4.16: Spreading behaviour for a population arranged on a random network obeying $P(k) = \delta_{k4}$. The left panel shows the simulation result (circles) and different levels of approximations for the equilibrium dynamics. The former was obtained by performing 10 iterations on 10 different realizations of networks of size $N = 10^5$, the transition probabilities being determined by $\Delta t = 0.01$. The star approximation (solid line) is in excellent agreement with the simulation result, yielding also an accurate description of the critical region. The mean-field (dotted line) and pair approximation (dashed line) have been plotted again for comparison. The right part shows the invasion of an infective agent (infection rate $\lambda = 3/7$) on the same type of topology. At the mean-field level (dotted line), the initial prevalence of 0.01 increases to its equilibrium value during 10 time units only. The pair approximation (dashed line) provides a further improvement, and the star approximation (solid line) is in remarkable agreement with the stochastic simulation whose parameters N and Δt correspond to those already mentioned above.

The probability that at time t , node 0 is in state x_0 and its nearest neighbours 1, 2, 3, 4 are in the states $\{x_1, x_2, x_3, x_4\}$ is denoted by

$$P_t \begin{pmatrix} x_2 \\ x_1 \ x_0 \ x_3 \\ x_4 \end{pmatrix}$$

and obtained by summing $\mathcal{P}_t(\mathbf{x})$ over all possible configurations, $\{x_0, x_1, x_2, x_3, x_4\}$ held fixed. The probability that the nearest and second-nearest neighbours of node 0 are in given states, is given by the ansatz

$$P_t(\{y_j\}_{j \in \mathcal{N}}) = P_t \begin{pmatrix} y_2 \\ y_1 \ y_0 \ y_3 \\ y_4 \end{pmatrix} \prod_{l=1}^4 P_t(y_{l2}y_{l3}y_{l4}|y_l y_0), \quad (4.36)$$

where \mathcal{N} represents the set of nodes depicted in Fig. 4.15 and the conditional probabilities are

$$P_t(y_{l2}y_{l3}y_{l4}|y_l y_0) = \frac{P_t \begin{pmatrix} y_{l4} \\ y_{l3} & y_l & y_0 \\ y_{l2} \end{pmatrix}}{P_t(y_l y_0)}.$$

The pair probability appearing in the above expression is extracted from the corresponding star probabilities by

$$P_t(y_l y_0) = \sum_{y_{l2}=0}^1 \sum_{y_{l3}=0}^1 \sum_{y_{l4}=0}^1 P_t \begin{pmatrix} y_{l4} \\ y_{l3} & y_l & y_0 \\ y_{l2} \end{pmatrix}.$$

With these ingredients, Eq. (4.27) leads to the following continuous-time evolution of the star probabilities

$$\dot{P} \begin{pmatrix} x_2 \\ x_1 & x_0 & x_3 \\ x_4 \end{pmatrix} = \sum_{\{y_j\}_{j \in \mathcal{N}}} \left(P(\{y_j\}_{j \in \mathcal{N}}) \prod_{i=0}^4 \left[(1-2x_i)[y_i - \lambda(1-y_i) \sum_{j \text{ nni}} y_j] \prod_{k \neq i} \tau_k \right] \right) \quad (4.37)$$

with $P(\{y_j\}_{j \in \mathcal{N}})$ given by Eq. (4.36) and τ_k according to Eq. (4.33). The binary character of this system of $2^5 = 32$ equations permits a very efficient numerical implementation. On the other hand, if one takes into account the symmetries of the problem, the degrees of freedom can be reduced to 10, this procedure will be shown for the regular lattices. The left panel of Fig. 4.16 shows the striking agreement of the star approximation with the simulation result, all along from a high effective spreading rate to its threshold value, for the equilibrium situation. Its right part opposes the various approximations to the stochastic simulation for the case of the invasion of an infective agent, the initial prevalence being set to 0.01. Whereas the steady state is reached rather quickly in the mean-field description, the slope of the star approximation is in remarkable agreement with the simulation. As correlations of a greater range are taken into account, it can also be observed that the system equilibrates more gradually, that is $\dot{\rho}(t \simeq 30)$ for the star approximation is considerably smaller than $\dot{\rho}(t \simeq 10)$ for the curve corresponding to the mean-field description.

Square Lattice

The presence of loops characterising the square lattice strongly affects the epidemic dynamics as we have already seen above. In order to arrive at a level of description beyond the pair approximation, we shall use the square as our fundamental cluster. This seems to be a natural choice, although it is not unique as discussed below. In analogy to the above considerations for the homogeneous random network, the probability that the corners of the square $ABCD$ are in the

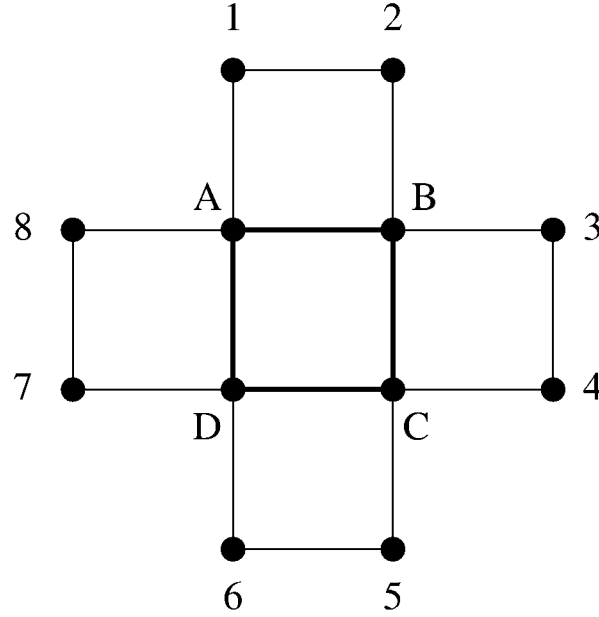


Figure 4.17: An arbitrarily chosen square within a 2-dimensional lattice and the denotation of the nearest neighbours of its corners. The former serves as the fundamental element within the square approximation.

states $\{x_A, x_B, x_C, x_D\}$ at time t is deduced from the system probability by

$$P_t \begin{pmatrix} x_A & x_B \\ x_D & x_C \end{pmatrix} = \sum_{\{x_i\}_{i \notin \{A, B, C, D\}}} \mathcal{P}_t(\mathbf{x}).$$

If the nearest neighbours of the vertices A, B, C and D are enumerated according to Fig. 4.17, we write for the probability that the nodes comprised within these 5 squares (i.e. the nearest neighbours of the central plaquette) are in given states

$$P_t(\{y_i\}_{i \in \{A, B, C, D, 1, 2, \dots, 8\}}) = P_t \begin{pmatrix} y_A & y_B \\ y_D & y_C \end{pmatrix} \cdot P_t(y_1 y_2 | y_A y_B) P_t(y_3 y_4 | y_B y_D) \quad (4.38) \\ \times P_t(y_5 y_6 | y_C y_D) P_t(y_7 y_8 | y_A y_C)$$

with

$$P_t(y_1 y_2 | y_A y_B) = \frac{P_t \begin{pmatrix} y_1 & y_2 \\ y_A & y_B \end{pmatrix}}{P_t(y_A y_B)}$$

involving the pair probability

$$P_t(y_A y_B) = \sum_{x_1=0}^1 \sum_{x_2=0}^1 P_t \begin{pmatrix} x_1 & x_2 \\ y_A & y_B \end{pmatrix}$$

and analogously for the other factors appearing in (4.38). At this point, we could again write down an equation of the type (4.37), but we shall explicitly make use

of the symmetries of the problem in order to reduce the computational load. The $2^4 = 16$ plaquette probabilities are subject to

$$\begin{aligned}
P_t \begin{pmatrix} 00 \\ 00 \end{pmatrix} &\equiv q_{0,t} \\
P_t \begin{pmatrix} 10 \\ 00 \end{pmatrix} &= P_t \begin{pmatrix} 01 \\ 00 \end{pmatrix} = P_t \begin{pmatrix} 00 \\ 10 \end{pmatrix} = P_t \begin{pmatrix} 00 \\ 01 \end{pmatrix} \equiv q_{1,t} \\
P_t \begin{pmatrix} 11 \\ 00 \end{pmatrix} &= P_t \begin{pmatrix} 01 \\ 01 \end{pmatrix} = P_t \begin{pmatrix} 00 \\ 11 \end{pmatrix} = P_t \begin{pmatrix} 10 \\ 10 \end{pmatrix} \equiv q_{2,t}^A \\
P_t \begin{pmatrix} 10 \\ 01 \end{pmatrix} &= P_t \begin{pmatrix} 01 \\ 10 \end{pmatrix} \equiv q_{2,t}^C \\
P_t \begin{pmatrix} 11 \\ 10 \end{pmatrix} &= P_t \begin{pmatrix} 11 \\ 01 \end{pmatrix} = P_t \begin{pmatrix} 10 \\ 11 \end{pmatrix} = P_t \begin{pmatrix} 01 \\ 11 \end{pmatrix} \equiv q_{3,t} \\
P_t \begin{pmatrix} 11 \\ 11 \end{pmatrix} &\equiv q_{4,t}.
\end{aligned}$$

The exact description (4.27) leads to the following continuous-time dynamics for these quantities

$$\begin{aligned}
\dot{q}_0 &= 4q_1 - 8\lambda T_1 q_0 \\
\dot{q}_1 &= -q_1 + 2q_2^A + q_2^C + \lambda[-2q_1(1 + 2T_1 + T_2) + 2T_1 q_0] \\
\dot{q}_2^A &= -2q_2^A + 2q_3 + \lambda[-4q_2^A + 2T_1(q_0 + 3q_1) + 2T_2(q_1 - q_2^A)] \\
\dot{q}_2^C &= -2q_2^C + 2q_3 + \lambda(-4q_2^C + 4T_1 q_1 - 4T_2 q_2^C) \\
\dot{q}_3 &= -3q_3 + q_4 + \lambda[2q_1 + 4q_2^A - 4q_3 - 2T_1(q_0 + 2q_1) + 2T_2(q_1 + 2q_2^A + 2q_2^C)] \\
\dot{q}_4 &= -4q_4 + \lambda[8q_2^C + 16q_3 - 8T_2(q_1 + q_2^A + q_2^C)]
\end{aligned} \tag{4.39}$$

where

$$T_1 = \frac{t_1^A}{p_0^A} \quad \text{and} \quad T_2 = \frac{t_2^C}{p_1^A}$$

involving the following triplet- and pair probabilities given by the square probabilities through

$$\begin{aligned}
t_1^A &= P \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = q_1 + q_2^A, & t_2^C &= P \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} = q_2^C + q_3, \\
p_0^A &= P(00) = q_0 + 2q_1 + q_2^A & \text{and} & p_1^A &= P(10) = q_1 + q_2^A + q_2^C + q_3.
\end{aligned}$$

Since

$$q_0 + 4q_1 + 4q_2^A + 2q_2^C + 4q_3 + q_4 = 1,$$

the square approximation in the form (4.39) represents a dynamical system of 5 degrees of freedom, in contrast to 16 if the symmetries were not exploited.

The left part of Fig. 4.18 shows the systematic improvement brought about by the square- and the bi-square approximations in dynamic equilibrium. The latter is a description whose fundamental cluster is composed of two squares. Its

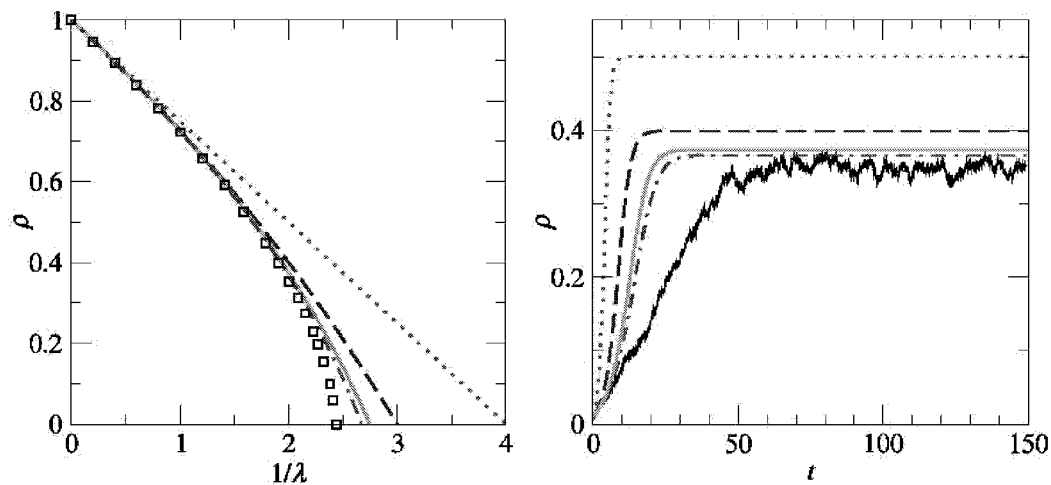


Figure 4.18: Prevalences for the epidemic process on a square-lattice structured population. The left side illustrates the result for the equilibrium state. The mean-field (dotted line) and pair approximations (dashed line) are levels of description at which topological properties beyond the degree distribution do not enter. The approximations involving the square (solid line) and a rectangle composed of two squares (dashed-dotted line) as fundamental units systematically approach the steady-state behaviour, as predicted by the simulations (squares, for its details see Fig. 4.14). The right panel reports on the dynamics for $\lambda = 0.5$. By taking into account correlations of a greater range, the slope during the transient time decreases as a comparison of the mean-field (dotted line), pair (dashed line), square (solid line) and bi-square approximations (dashed-dotted line) shows. The difference between the simulation result (for $N = 10^4$, $\Delta t = 0.01$) and the bi-square approximation remains significant during the invasion period due to the considerable effect of random events at overall low prevalence.

prediction of the epidemic threshold ($\lambda_c \simeq 0.38$) is still lower than the simulation result ($\lambda_c \simeq 0.41$): this highlights the crucial role of the higher-order spatial correlations in lattice structured populations. However, the square approximation describes the spreading behaviour very accurately for $1/\lambda \lesssim 2$. The right panel of Fig. 4.18 represents the improvements upon the dynamics. Note that from a certain characteristic time the simulation lags behind all the approximations as a direct consequence of the stochasticity which is particularly important at low prevalences. However this characteristic time is shifted to the right as higher-order correlations of a greater range are taken into account.

An improvement upon the standard pair approximation can also be obtained as follows [59]. Instead of deriving the square probabilities from the dynamics of

the system, one can write them as

$$P \begin{pmatrix} x_i & x_a \\ x_j & x_b \end{pmatrix} = P(x_i)P(x_a)P(x_b)P(x_j)C_{ia}C_{ab}C_{bj}C_{ji}T_{\square iabj}$$

involving the relative pair and square correlation factors C_{xy} and $T_{\square iabj}$. For a straight triple, it is supposed

$$P(x_i x_a x_b) = P(x_i)P(x_a)P(x_b)C_{ia}C_{ab}T_{\triangle iab}.$$

By setting the relative correlation factors $T_{\square iabj}$ and $T_{\triangle iab}$ to 1 and using the fact that on the square lattice 1/3 of the triples are straight and 2/3 form part of a square, one obtains an improvement for $P(x_i|x_a x_b)$. In other words, the conditional probability $P(x_i|x_a x_b)$ is not simply set to $P(x_i|x_a)$ as it is done in the ordinary pair approximation, but rather the loop structure is incorporated while still using pairs as building blocks.

Triangular Lattice

In its ordinary formulation, the fact that two sites can have neighbours in common, is simply ignored by the pair approximation. By means of the triangular lattice, we show how the method described in this section has to be applied, i.e. what the next level of description beyond the pair approximation is.

The clue is to use the triangle as the basic element. In analogy to the previous cases, the probability that the vertices of a triangle ABC are in the states $\{x_A, x_B, x_C\}$ at time t is obtained through

$$P_t \begin{pmatrix} x_A x_B \\ x_C \end{pmatrix} = \sum_{\{x_i\}_{i \notin \{A,B,C\}}} \mathcal{P}_t(\mathbf{x}).$$

Fig. 4.19 shows the neighbourhood of an arbitrarily chosen triangle within this lattice. For the probability that the vertices depicted in this figure are in given states, we suppose

$$P_t(\{y_i\}_{i \in \{A,B,C,1,2,\dots,9\}}) = P_t \begin{pmatrix} y_A y_B \\ y_C \end{pmatrix} \cdot P_t(y_1|y_A y_B)P_t(y_4|y_B y_C)P_t(y_7|y_C y_A) \\ \times P_t(y_8 y_9|y_A)P_t(y_2 y_3|y_B)P_t(y_5 y_6|y_C).$$

The conditional probabilities appearing in the above expression can be written as fractions involving site- and pair probabilities. The latter are deduced from the triangle probabilities in analogy to previous explanations. As the triangle

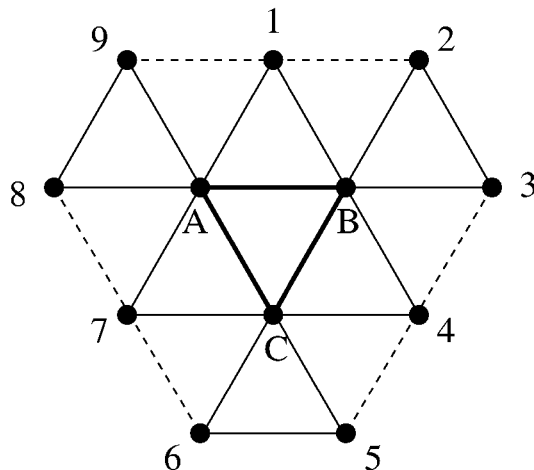


Figure 4.19: An arbitrarily chosen triangle and its nearest neighbourhood. The dashed lines indicate that the corresponding links are ignored.

correlations are subject to the symmetries

$$\begin{aligned}
 P_t \begin{pmatrix} 00 \\ 0 \end{pmatrix} &\equiv t_{0,t} \\
 P_t \begin{pmatrix} 10 \\ 0 \end{pmatrix} &= P_t \begin{pmatrix} 01 \\ 0 \end{pmatrix} = P_t \begin{pmatrix} 00 \\ 1 \end{pmatrix} \equiv t_{1,t} \\
 P_t \begin{pmatrix} 11 \\ 0 \end{pmatrix} &= P_t \begin{pmatrix} 10 \\ 1 \end{pmatrix} = P_t \begin{pmatrix} 01 \\ 1 \end{pmatrix} \equiv t_{2,t} \\
 P_t \begin{pmatrix} 11 \\ 1 \end{pmatrix} &\equiv t_{3,t},
 \end{aligned}$$

a further simplification can be performed, and finally the continuous-time triangle dynamics is governed by the equations

$$\begin{aligned}
 \dot{t}_0 &= 3[t_1 - 2\lambda(A_1 + A_2)] \\
 \dot{t}_1 &= -t_1 + 2t_2 + 2\lambda(-2t_1 + 3A_1 + 2A_2 - 2A_3 - A_4) \\
 \dot{t}_2 &= -2t_2 + t_3 + 2\lambda(t_1 - 3t_2 - 3A_1 - A_2 + 4A_3 + 2A_4) \\
 \dot{t}_3 &= 3[-t_3 + 2\lambda(t_1 + 3t_2 + A_1 - 2A_3 - A_4)].
 \end{aligned} \tag{4.40}$$

where

$$A_1 = \frac{p_1 t_0}{s_0}, \quad A_2 = \frac{t_0 t_1}{p_0}, \quad A_3 = \frac{p_0 p_1}{s_0} \quad \text{and} \quad A_4 = \frac{t_1 t_2}{p_1}$$

depending on the pair probabilities $p_1 = P(10) = t_1 + t_2$, $p_0 = P(00) = t_0 + t_1$ and the site probability $s_0 = P(0) = t_0 + 2t_1 + t_2$. Because of the constraint

$$t_0 + 3t_1 + 3t_2 + t_3 = 1,$$

we have three degrees of freedom in the triangle approximation (4.40).

As far as the equilibrium prediction is concerned, the triangle approximation provides a very good description for $1/\lambda < 3$ (Fig. 4.20, left panel). The difference

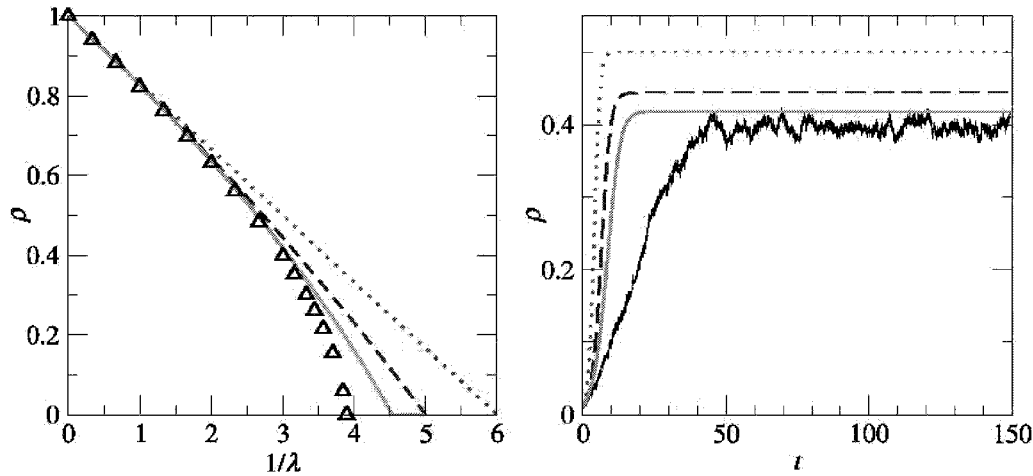


Figure 4.20: Results for the SIS model on a triangular lattice. Left: Steady-state prevalence as predicted by various approximations and through computer simulations (triangles) whose parameters are $N = 10^4$, $\Delta t = 0.01$ and 10 iterations. The mean-field description (dotted line) yields an epidemic threshold $\lambda_c = 1/6$. With respect to the pair approximation (dashed line), the description based on the triangle (solid line) provides a better approximation of the simulation result. Right: Invasion dynamics for an effective spreading rate $\lambda = 1/3$. The upper two curves show the mean-field (dotted line) and pair dynamics (dashed line). The improvement brought about by the triangle approximation (solid line) still lags behind the simulation result due to the same reason as in the case of the square lattice. The latter was obtained for $N = 10^4$ and $\Delta t = 0.01$ as well.

between its threshold prediction ($1/\lambda_c \simeq 4.5$) and the simulation result ($1/\lambda_c \simeq 3.9$) is of the same order of magnitude as the plaquette approximation in the case of the square lattice. Concerning the dynamics (Fig. 4.20, right panel), we also observe a lag between the simulation and the approximations, and the slope during the transient time is slightly improved as one goes from the pair to the triangle approximation.

The strategy outlined at the end of the last subsection can also be applied to the triangular lattice [59]. In addition to the open triplet probability, the triangle probability is written as

$$P\left(\begin{array}{c} x_i \\ x_a x_b \end{array}\right) = P(x_i)P(x_a)P(x_b)C_{ia}C_{ab}T_{\Delta iab}.$$

One then obtains an analogous correction for $P(x_i|x_a x_b)$ involving a parameter θ denoting the fraction of triplets in closed form which is $2/5$ in the triangular lattice. Interestingly, the simplest elaboration of this approach ($\tau_{\Delta iab} = \tau_{\angle iab} \equiv \tau_{iab}$) reproduces the invasive period reasonably accurate if θ is chosen larger than

its correct value ($\theta \simeq 0.6$). Keeling *et al.* [60] and Rand [99] also developed improved pair models based on this approach.

4.6 Discussion

The spreading of an epidemic, be it the propagation of an infectious disease or a computer virus, was modeled as a dynamic process on a networked structure. We used two versions of the simple SIS model where nodes representing individuals or computers are either susceptible or infected. In the first, infected nodes recover with probability Δt , susceptible nodes become infected with probability $\lambda \Delta t$ if they are connected to at least one infected nearest neighbour, and Δt was set to 1. In the second, more conventional version, the recovery of infected nodes occurs with rate 1 while they infect neighbouring susceptible sites at a rate λ . In the methods discussed below, we also describe to which of these models they apply. This choice is motivated below. As far as the connectivity patterns representing the contact structure are concerned, we chose simple model networks each of which enabling us to tune specific topological properties and thus allowing to study its effect on the spreading behaviour. In particular, random bimodal networks were used to learn about the role of degree correlations, and networks with a fixed degree (regular lattices, homogeneous random networks and adequate small-world networks) led to new insights about the role of short loops.

With the hypothesis that the state of any node is not felt by its nearest neighbours, that is by ignoring spatial correlations, we rigorously derived a sequence of mean-field equations involving various further approximations. The main conclusion gained at this level of description is that the fraction of infected pairs, one node having degree k the other k' , is not simply the fraction of infected nodes of degree k times that of degree k' . In other words, the heterogeneous topology induces correlations although the probability of finding an infected pair of nodes is assumed to be the product of the two site probabilities in question. These mean-field considerations presented in Sec. 4.2 were performed for the discrete-time version of the simplified SIS model although they equally apply to the other formulation of the SIS model mentioned above.

The remaining part of this chapter was devoted to the role of short loops in the spreading process, and since we did not aim to come to an understanding of the combined effect of the loop structure and degree-related topological properties, homogeneous networks were used. Describing the epidemic dynamics of the entire population as a Markovian process, we derived a two-step description that takes into account temporal correlations. This approach revealed to be very prolific if one wants to unravel the role of loops of short length in the contact network regarding epidemic spreading. Indeed it leads to a subgraph development where the complete graph involves the connectivity patterns of two hierarchies of nearest neighbours (of an arbitrarily chosen node). Within this novel approach serving to

tract a probabilistic system, the local topology, be it treelike or be loops of length 3 or 4 present, therefore enters very naturally. The analytically obtained condition for the location of the onset of the epidemic then serves as a guiding equation elucidating the role of clustering and grid-like ordering in epidemic spreading.

In principle it is possible to apply our two-step description to more complex networks where different degrees are present, uncovering the effect of the degree-dependent densities of triangles and quadrilaterals on the critical value. Likewise, loops of length $\leq 2n$ are expected to enter within an n -time-step description, also providing a natural classification of them. However the major insight gained by the strategy of exploring temporal correlations is best illustrated as it was done in Sec. 4.4.

It is worth noting that this method only applies to a discrete-time formulation. This is easily seen by looking at the threshold condition, e.g. Eq. (4.21) or (4.23): λ in fact stands for $\lambda\Delta t$ and by performing the continuous-time limit ($\Delta t \rightarrow 0$) from a two-step equation, only terms proportional to Δt^r with $r \leq 2$ are kept. Loops would thus have a vanishing effect in this limit. However, describing the epidemic dynamics in discrete time is not an unreasonable assumption: in order to measure the transmission probability, an interval of time needs to be fixed and this probability is then derived from the number of occurred transmissions between two individuals, repeating the experiment many times. The two-step approach could also be applied to the discrete-time version of the conventional SIS model. However, the coefficients in the threshold condition are then expected to depend on the number of triangles and quadrilaterals in a more complex manner, that is the simplified SIS model is particularly illustrative in the context of this method.

In the remaining section, we explored spatial correlations in order to come to an understanding of the role of the local topology in the spreading phenomenon. The method we proposed in Sec. 4.5 consists in choosing a fundamental cluster composed of a certain number of nodes as well as links connecting them. A definite probability is assigned to each possible configuration of the basic element. The size of the fundamental cluster represents the range up to which spatial correlations are exactly taken into account. At a level beyond the pair approximation, the choice of the basic element is guided by the underlying network's topology. In the case of the square lattice, clusters composed of at least one plaquette serve as the fundamental element; for random networks the local treelike structure is incorporated by using the star as the basic cluster. Spatial patterns beyond the degree distribution are therefore embedded in a very natural way by this method. By adopting the same point of departure as for the two-step approach, that is describing the epidemic dynamics of the entire population as a discrete-time Markovian process, the appearing probabilities (probability that a cluster and its nearest neighbourhood is in a given configuration) are expressed in terms of the fundamental cluster probabilities. The continuous-time dynamics emerges as a limiting case ($\Delta t \rightarrow 0$).

With respect to the ordinary (rather heuristic) derivation of the mean-field

and pair approximation, these descriptions are derived with the help of our formalism by using the site or the pair respectively as fundamental clusters in a very automatic way. Independently of the specific choice of the cluster, the binary character of the resulting equations allows for a very efficient solution by the computer. Likewise, a further simplification can be reached if the symmetries which the fundamental cluster probabilities are subject to, are taken into account. As soon as correlations of range greater than 2 are not ignored, our method yields improved estimates for the location of the onset of the epidemic. In the case of the random network, the star approximation already leads to an excellent description of the steady state and the transient dynamics. In the regular counterpart, many squares have to be included within the corresponding fundamental unit in order to attain the same level of accuracy. This is due to the presence of stronger correlations caused by the high local ordering. The method was also illustrated for a triangular lattice and contrasted to approaches that make a certain number of assumptions about the higher-order correlations which lead to improved pair models.

We have focused on homogeneous networks, since in this case a fundamental cluster is identified most easily. The homogeneity lies indeed at the basis of our cluster approximations since it must be possible to express the probability appearing on the right of our master equations [see e.g. Eq. (4.37)] entirely in terms of the fundamental cluster probabilities, such as in Eqs. (4.36) and (4.38). In principle, our method can be extended to slightly heterogeneous systems, for example to a random network where two different degrees are present. The dynamics is then described in terms of two different star-like clusters (according to the occurring degrees), this hybridisation involving the constraint that the pair probabilities derived from the two clusters must coincide.

However the novelty of this method lies in the formalism which essentially consists in a more general starting point and its associated systematic improvability rather than the specific results for the selected epidemiological model and geometrical examples.

In summary, the mean-field description led to a non-trivial relationship between degree correlations and the time evolution of the fraction of infected pairs connecting nodes with given degrees; and the methods presented in Secs. 4.4 and 4.5 provide ways to study the role of the local topology in a dynamic probabilistic system. The two-step approach has the advantage that it unravels the role of short loops for networks with an arbitrary degree of disorder, but this effect is only seen if the dynamics is modelled in discrete time. On the other hand, cluster approximations apply to both discrete- and continuous-time dynamics, but they rely on the fact that the process takes place either on a fully regular or entirely random topology.

Chapter 5

Spatial small-world networks

The problems investigated up to now were related to networks which did not “live” in a (geographical) space, i.e. a space in which a metric is defined. All that mattered, for example in the previous chapter, was: who is connected to whom, i.e. which individual/computer can possibly transmit the virus to whom. Thus the positions where the nodes were drawn were of a minor importance. In that sense, all the networks depicted in Fig. 5.1 are identical. That is, if we label the nodes and establish a matrix A with $A_{ij} = 1$ if i is connected to j and $A_{ij} = 0$ otherwise (i.e. the adjacency matrix), there exists a way of labeling the nodes for the Figs. 5.1a-f such that the resulting matrix A is always the same. Clearly, in the case of Watts and Strogatz’ model, we started with a lattice that “lives” in a D -dimensional space, but the reason for that was merely to borrow from it a high local interconnectedness - that is a property of regular lattices. In other words, the Euclidean distance between nodes has not been a quantity of interest.

The following networks indeed do not “live” in a geographic space: the World Wide Web, social networks to some extent¹ and any type of virtual network, e.g. the protein folding network where nodes are protein configurations and a (directed) link between any two nodes is established if there exists a corresponding transition between the two nodes [100] - or generally, any network where the links have no real or physical correspondence. Yet, networks like the Internet, the human brain at the anatomical level or integrated electronic circuits clearly “live” in a Euclidean space, and the locations of the nodes are essential when it comes to describing these systems. In fact, it is believed that the brain has evolved so as to minimise the wiring costs related to the lengths of the axons and dendrites [101, 102, 103], computer chips were designed such that a low amount of wire is used and also for computer networks such as the Internet, long physical² connections are more costly. More importantly, these spatial constraints have shaped these networks, and it is essential to understand how topology is affected by these constraints.

¹Indirectly, geography matters in a social context in that people usually have most of their friends in the same neighbourhood, village or town and some who live elsewhere.

²For obvious reasons, satellite or wireless connections are not subject to that constraint.

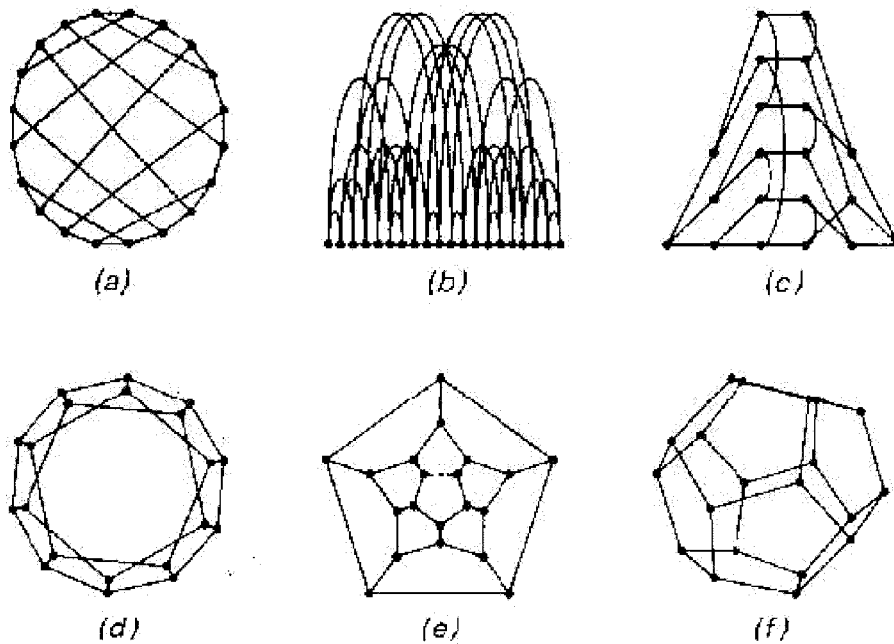


Figure 5.1: Six different ways of laying out the same network. (a) nodes arranged around a circle; (b) nodes arranged along a line; (c) nodes arranged across the page according to distance from a particular node; (d) 2D layout with network and spatial distances as close as possible; (e) planar layout; (f) 3D layout [104].

When we think of network topology, one point immediately comes to mind: is it possible to construct small-world networks in which the wiring costs are somehow minimised? At first sight, we hardly think that this question could be answered in an affirmative way since the small-world phenomenon is rooted in the presence of shortcuts, i.e. long-range connections, which are costly. However, the length distribution of the shortcuts $q(l)$ appears to be crucial. While the majority of the modelling effort has been done for uniform distributions, decaying power laws were reported for the above systems [61, 62, 63]. A length distribution $q(l) \sim l^{-\alpha}$ implies the absence of a characteristic connection length, i.e. the presence of multiple length scales. It was even conjectured that the fundamental mechanism behind the small-world phenomenon is neither disorder nor randomness as described in Sec. 2.3, but rather the scale-free length distribution [105]. Moreover, the navigability in a small world with such a connection-length distribution depends on the corresponding decay exponent [106], and the nature of random walks over the network is also affected [107].

In this chapter, we re-analyse the small-world phenomenon for systems where $q(l) \sim l^{-\alpha}$ and investigate how the topology relates to the wiring costs. We complement the picture in that we also study the distribution of flows over the links and the robustness related to random failures *and* overload of connections, finishing with concluding remarks.

5.1 Topology

As already mentioned in Sec. 2.3, the behaviour of the mean distance characterising the small-world (SW) phenomenon essentially depends neither on the chosen regular lattice (as point of departure) nor on the details of the rewiring procedure, that is, it does not matter whether local connections are replaced by shortcuts or if the latter are simply added. Here, we supplement a D -dimensional lattice ($D = 1, 2$) of linear size L - thus consisting of $N = L^D$ nodes - subject to periodic boundary conditions, with pN additional connections whose lengths are distributed according to the probability density $q(l) \sim l^{-\alpha}$. Due to the boundary conditions, it is most convenient to allow for lengths $2 \leq l \leq L/2$ which also determines the normalisation of $q(l)$. More precisely, any such additional connection is established by first choosing its length according to $q(l)$; it is then put on the lattice by randomly choosing one endpoint and the other at distance l (with no preference to a particular direction), such that no pair of sites is connected by more than one additional connection. It thus depends on the value of l that is being drawn whether the corresponding additional link is a real shortcut implying that it connects far away nodes.

For small values of p , this formulation of the SW model corresponds to the case where at every site, a link is added with probability p - the other endpoint being chosen as above, but it has the advantage that p can be greater than 1 which allows us to look beyond the simple probabilistic version. Clearly, a certain amount of real shortcuts - that is, long additional connections - is required for SW topology to emerge. A first picture is obtained by plotting the mean distance versus p for different values of α , namely 0, 0.5, 1, ..., 3 and a fixed system size (Fig. 5.2), the lowest curve ($\alpha = 0$) corresponding to $\langle d \rangle(p)$ in Fig. 2.3. At a fixed value of p , $\langle d \rangle$ increases with α . Conversely, in order to have a given mean distance, the value of p that must be chosen increases with α . By looking at the region $p > 2$, the points corresponding to $\alpha \leq 1.5$ lie much below the three uppermost curves. This may be the signature of a topological transition. However, in order to investigate its precise nature, the N -dependence of $\langle d \rangle$ needs to be considered as well, as we pointed out in Sec. 2.3, and this is done in the following. The above reasoning suggests that SW features are present if the length-distribution exponent α is smaller than a critical value α_c . In order to find this critical exponent, we look at the probability that an arbitrarily chosen additional link is a real shortcut, that is,

$$P_c(L) = \int_{(1-c)L/2}^{L/2} q(l)dl, \quad (5.1)$$

c being small but finite, and require the presence of the order of one such connection:

$$P_c(L)p^*(L)L^D \simeq 1 \quad (5.2)$$

where $p^*(L)L^D$ is the desired number of additional connections. The interpretation of $p^*(L)$ is as follows: for a given L (and α), the presence of many more

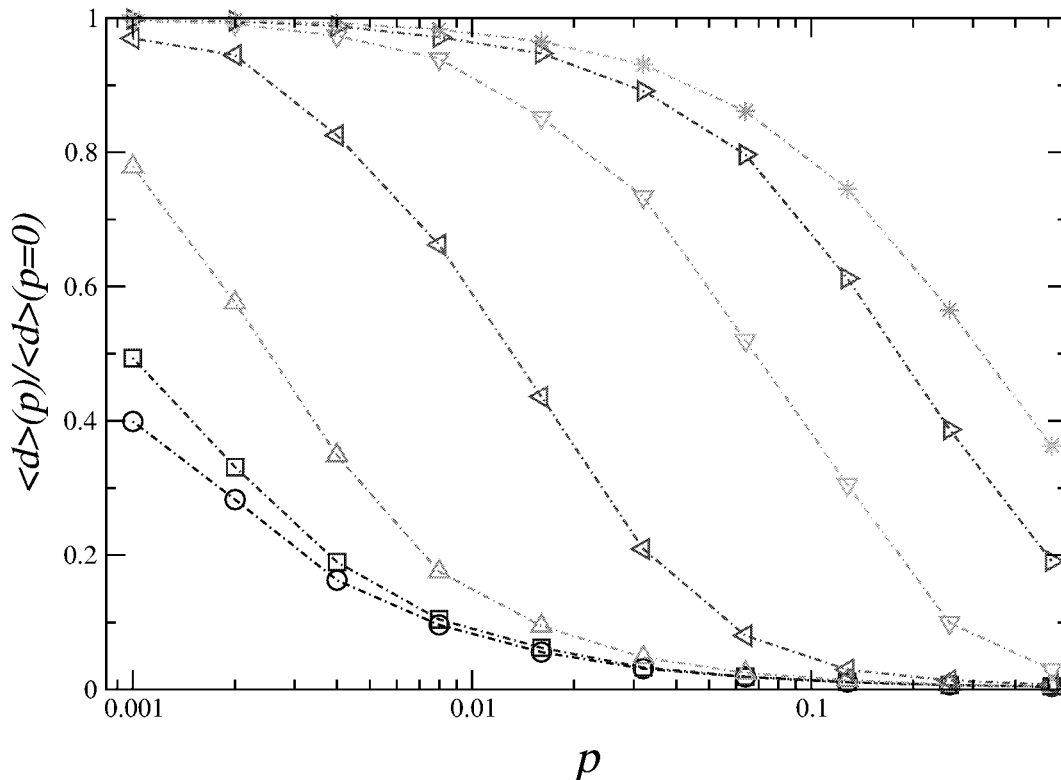


Figure 5.2: Relative mean distance versus p for 1-dimensional lattices ($N = 10^4$). The curves correspond to the following length-distribution exponents: $\alpha = 0$ (\circ), $\alpha = 0.5$ (\square), $\alpha = 1$ (\triangle pointing upwards), $\alpha = 1.5$ (\triangle pointing leftwards), $\alpha = 2$ (\triangle pointing downwards), $\alpha = 2.5$ (\triangle pointing rightwards) and $\alpha = 3$ (\star). These results were obtained by averaging over 10 realisations of networks.

additional links than the desired number [$pL^D \gg p^*(L)L^D$] implies SW topology, that is the mean distance scales as in a random network ($\langle d \rangle \sim \ln N = D \ln L$). On the contrary, if $pL^D \ll p^*(L)L^D$, one essentially observes a regular lattice characterised by $\langle d \rangle \sim L$ also referred to as a *large world* (LW). Evaluating Eq. (5.1) gives

$$P_c(L) \sim \begin{cases} 1 - (1 - c)^{1-\alpha} & \text{if } \alpha < 1, \\ -\ln(1 - c) / \ln L & \text{if } \alpha = 1, \\ L^{1-\alpha} [(1 - c)^{1-\alpha} - 1] & \text{if } \alpha > 1. \end{cases} \quad (5.3)$$

and with Eq. (5.2), we find for the critical fraction of additional links

$$p^*(L) \sim \begin{cases} L^{-D} & \text{if } \alpha < 1, \\ \ln(L) / L^D & \text{if } \alpha = 1, \\ L^{\alpha-D-1} & \text{if } \alpha > 1. \end{cases} \quad (5.4)$$

where the c -dependence has been ignored since we are only interested in the scaling with L . Eq. (5.4) can also be interpreted differently if p rather than L is

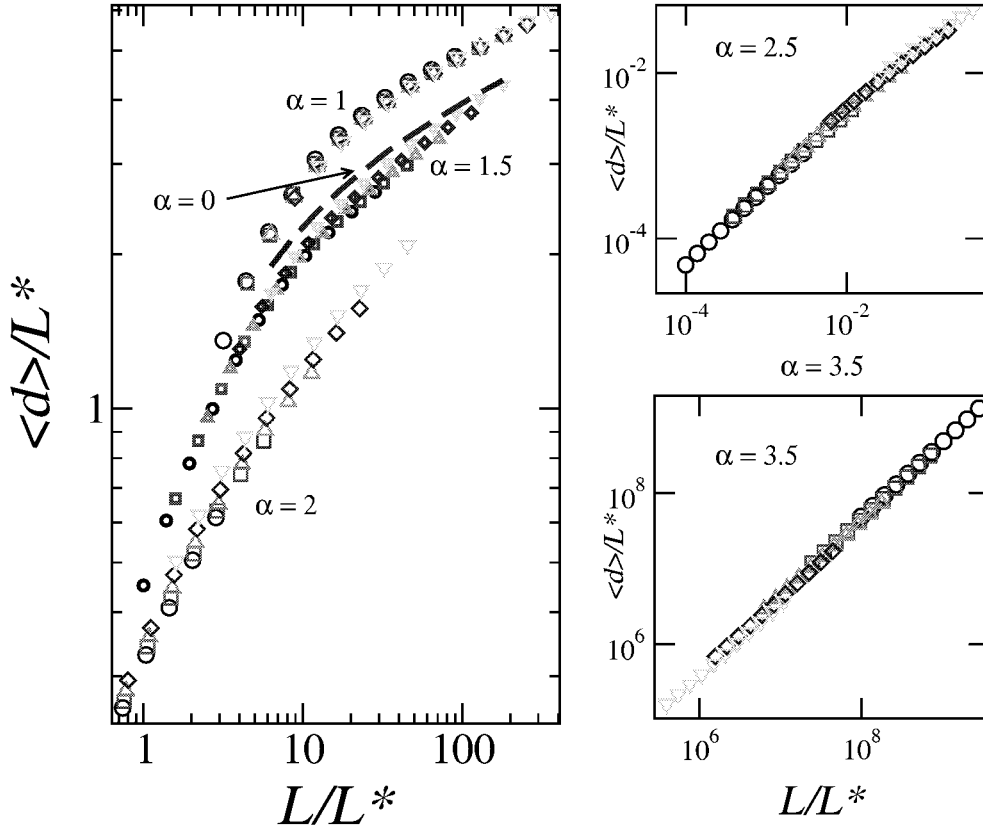


Figure 5.3: Mean distance versus linear system size, both of these quantities being rescaled by $L^*(p)$, for $p = 0.001$ (\circ), $p = 0.002$ (\square), $p = 0.004$ (\triangle), $p = 0.008$ (\diamond) and $p = 0.016$ (∇). The exponent α ranges from 0 to 3.5 as indicated. The data collapses confirm Eqs. (5.4) and (5.6) also for $\alpha > \alpha_c$ (lower right panel). These results were obtained by averaging over 100 realisations of networks (of different sizes) and in each case, the mean distance was computed by constructing the minimum spanning tree from a number of randomly chosen “root” nodes, this number being chosen such that an accurate estimate could be obtained.

fixed: there exists a critical length $L^*(p)$ such that $L \gg L^*(p)$ corresponds to SW topology and $L \ll L^*(p)$ characterises a large world. This is easily seen by raising these inequalities to the power of D and multiplying them with p since then, again numbers of additional connections are contrasted and the above picture holds. Mathematically, we replace p^* by p and L by L^* , leading to

$$L^*(p) \sim \begin{cases} p^{-1/D} & \text{if } \alpha \leq 1, \\ p^{\frac{1}{\alpha-D-1}} & \text{if } \alpha > 1. \end{cases} \quad (5.5)$$

with some sort of logarithmic correction of minor importance for $\alpha = 1$. For the special case $\alpha = 0$, this result was previously derived by other means [74, 108, 109]. We further see that $L^*(p) \rightarrow \infty$ as $\alpha \rightarrow (D + 1)^-$. This suggests $\alpha_c = D + 1$ and implies that SW topology only emerges for $\alpha < \alpha_c$. Before contrasting this result

with the findings of other authors, we shall provide a unified SW-LW picture and verify it numerically. The mean distance can be expressed as

$$\langle d \rangle = L^* \mathcal{F}_\alpha \left(\frac{L}{L^*} \right), \quad (5.6)$$

the scaling function obeying [74, 75, 108]

$$\mathcal{F}_\alpha(x) \sim \begin{cases} x & \text{if } x \ll 1, \\ \ln x & \text{if } x \gg 1 \end{cases} \quad (5.7)$$

where the second line of Eq. (5.7) may also read $[\ln(x)]^{s(\alpha)}$, $s(\alpha) > 0$. Fig. 5.3 shows the rescaled mean distances as a function of the rescaled linear system size for different values of α and $p = 0.001, 0.002, \dots, 0.016$ in each set for the case $D = 2$. The observed data collapses for all the length-distribution exponents confirm Eq. (5.5) obtained by our simple argument as well as Eq. (5.6). We numerically verified Eq. (5.7), especially in the limit $L/L^* \ll 1$, the logarithmic tail of \mathcal{F}_α further being exhibited best for small α .

Some attention was previously given to this type of problem. Sen et al.'s results suggest $\alpha_c = D + 1$ in agreement with our finding although they oddly defined the SW behaviour via the scaling of the clustering coefficient rather than by means of the scaling of distances [110]. Our argument can also be formulated for a model where a link is added at any site \mathbf{x} with probability p , the other endpoint (\mathbf{y}) being chosen with probability $\sim |\mathbf{x} - \mathbf{y}|^{-\alpha}$ [111], to which we will refer as the variation of our model. This indeed differs from adding pN links whose lengths are distributed according to $\sim l^{-\alpha}$ for any value of p because of the following reason. In the version treated in this section, for any link to be added, we first choose its length according to the probability density distribution

$$q(l) = \frac{l^{-\alpha}}{\int_2^{L/2} \tilde{l}^{-\alpha} d\tilde{l}} \quad ,$$

then randomly choose one endpoint and the other at the drawn distance l with no preference of a specific direction. Hence the length distribution of the additional links does not depend on the dimensionality of the lattice as the lengths are chosen first. Yet, in the version where a link is added at any site with probability p , the other endpoint being chosen with probability $\sim |\mathbf{x} - \mathbf{y}|^{-\alpha}$, the resulting link lengths are influenced by the dimension of the lattice in that the emerging distribution is

$$q_s(l) = \frac{l^{D-1-\alpha}}{\int_2^{L/2} \tilde{l}^{D-1-\alpha} d\tilde{l}} \quad ,$$

in essence simply because the length is chosen after the site in question “sees” the dimensionality of the lattice, which changes the normalisation accordingly. For this version, $q(l)$ in Eq. (5.1) has to be replaced by $\int d\Omega l^{D-1} q(l)$, and the subsequent steps result in $\alpha_c = 2D$. In Ref. [111] this critical exponent is derived

through the following rescaling argument. The lattice is divided into blocks of linear size b such that $1 \ll b \ll L$, one block thus containing b^D nodes. Any two blocks are connected if there is at least one additional link from one node in the first block to another in the second. From the number of additional links at this coarse-grained level, a corresponding \tilde{p} can be extracted, and the above conjecture was found through $\tilde{p} = f(b)p$. The variation of the model is equivalent to the case where a connection is added between any pair of sites \mathbf{x} and \mathbf{y} with a probability proportional to $|\mathbf{x} - \mathbf{y}|^{-\alpha}$ [112]. In this reference, $\alpha_c = 2D$ was found with mathematical rigour. Let us finally note that in one dimension and for small p , the two models do coincide. This is in agreement with the fact that in that case, the two predictions for α_c are equal ($\alpha_c = D + 1 = 2D = 2$ if $D = 1$).

5.2 Wiring costs

The moments $\langle l \rangle$ and $\langle l^2 \rangle$ play a crucial role as far as the wiring costs of the networks are concerned. The total wiring cost $C_W = pL^D \langle l \rangle$ is also an important quantity, its minimisation governing, for example, the evolution of cortical networks [101]. We find for the first two moments the scaling relations summarised in Tab. 5.2, the expressions for integer α being modified by logarithmic corrections. In 2 dimensions, SW topology can be realized even if $\langle l \rangle = \text{const}$ (that is, for $2 < \alpha < 3 = \alpha_c$) whereas this is not the case in 1 dimension where $\langle l \rangle$ becomes finite in the $L \rightarrow \infty$ limit only above $\alpha_c = 2$. Moreover, if $D = 3$, it is even possible to have $\langle l \rangle = \mathcal{O}(1) = \langle l^2 \rangle$ while still being in the SW regime for $3 < \alpha < 4 = \alpha_c$. An appropriate choice of the parameters D and α is thus the key to constructing networks which are both efficient (SW topology) and economical (low wiring costs).

It is furthermore interesting to have a closer look at the relationship between the wiring costs and the topology. As α varies, one can ask what mean distance results, given a total amount of wiring length (i.e. the total cost) to be used to establish the additional connections. Fig. 5.4a reports these dependencies for $\alpha = 0, 1, 1.5$ and 1.75 (going from the uppermost to the lowest set) for 1 dimensional topologies of 10^4 sites. The largest value of $\langle d \rangle$ (the leftmost circle) corresponds to the length scale $L^* < 10^3 \ll 10^4 = L$, thus all the points in the figure represent the system in the SW regime. It can clearly be seen that the mean distance

	$0 \leq \alpha < 1$	$1 < \alpha < 2$	$2 < \alpha < 3$	$\alpha > 3$
$\langle l \rangle$	L	$L^{2-\alpha}$	const	const
$\langle l^2 \rangle$	L^2	$L^{3-\alpha}$	$L^{3-\alpha}$	const

Table 5.1: Behaviour of the moments of the connection-length distribution as a function of the linear system size L .

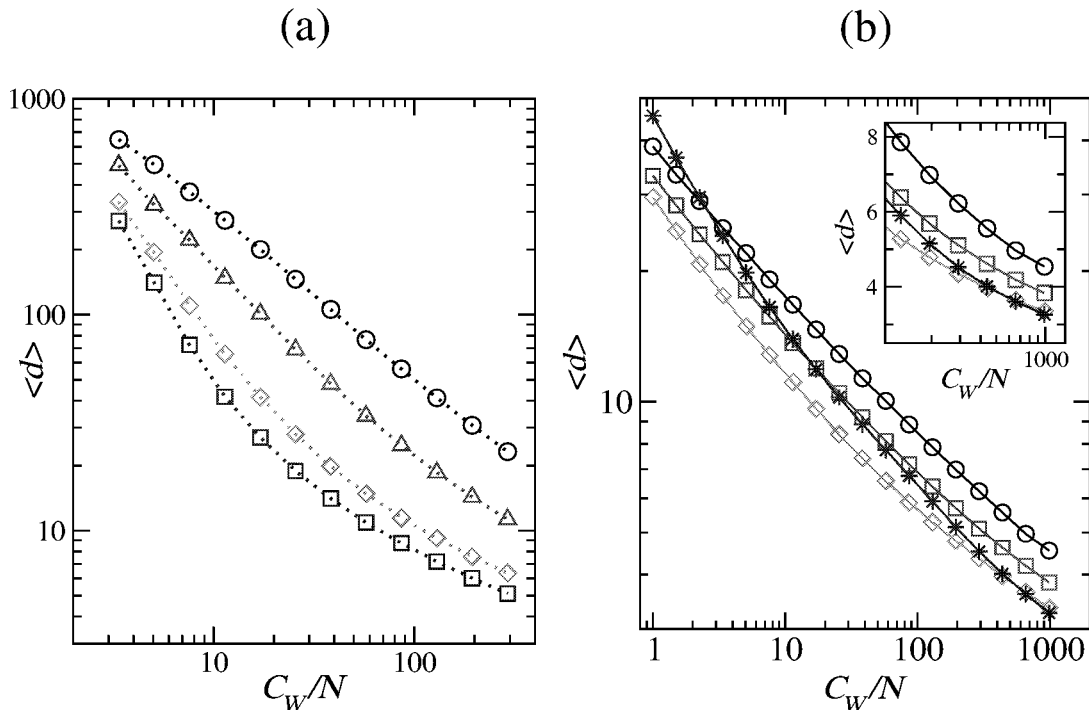


Figure 5.4: (a) Mean distance as a function of the total wiring costs (divided by the number of sites) for 1 dimensional topologies ($N = 10^4$). The curves (\circ : $\alpha = 0$, \triangle : $\alpha = 1$, \diamond : $\alpha = 1.5$ and \square : $\alpha = 1.75$) show that the mean distance decreases with α for a fixed value of C_W/N . (b) Analogous results for $D = 2$ [$N = 500 \times 500$ and $\alpha = 0$ (\circ), $\alpha = 1$ (\square), $\alpha = 2$ (\diamond) and $\alpha = 3$ (\star)], the inset enlarges the rightmost part of the curves. All the points shown here result from averaging over 100 realisations of networks.

decreases with α at fixed wiring costs C_W/N , i.e. the larger α the smaller the world. Let us now see how this behaviour can be understood in terms of the analytical elaborations from the previous section. When Eq. (5.6) is expressed in terms of $x = C_W/N = p\langle l \rangle$, L^* being taken from Eq. (5.5), we obtain

$$\frac{\langle d \rangle}{L} \sim \begin{cases} \frac{1}{x} \mathcal{F}_0(4x) & \text{if } \alpha = 0 \\ \frac{1}{\ln(L/2)} x^{-1} \mathcal{F}_1[4x \ln(L/2)] & \text{if } \alpha = 1 \\ x^{\frac{1}{\alpha-2}} \mathcal{F}_\alpha \left[\text{const}(\alpha) x^{\frac{1}{2-\alpha}} \right] & \text{if } 1 < \alpha < 2. \end{cases} \quad (5.8)$$

Since $\mathcal{F}_\alpha(x) \sim \ln x$ in the SW regime, this effectively raises the exponents of x appearing to the left of \mathcal{F}_α in Eq. (5.8) slightly and thus qualitatively explains the numerical results shown in Fig. 5.4a, at least for $x = C_W/N \lesssim 20$. Moreover, we made similar observations in 2 dimensions (Fig. 5.4b).

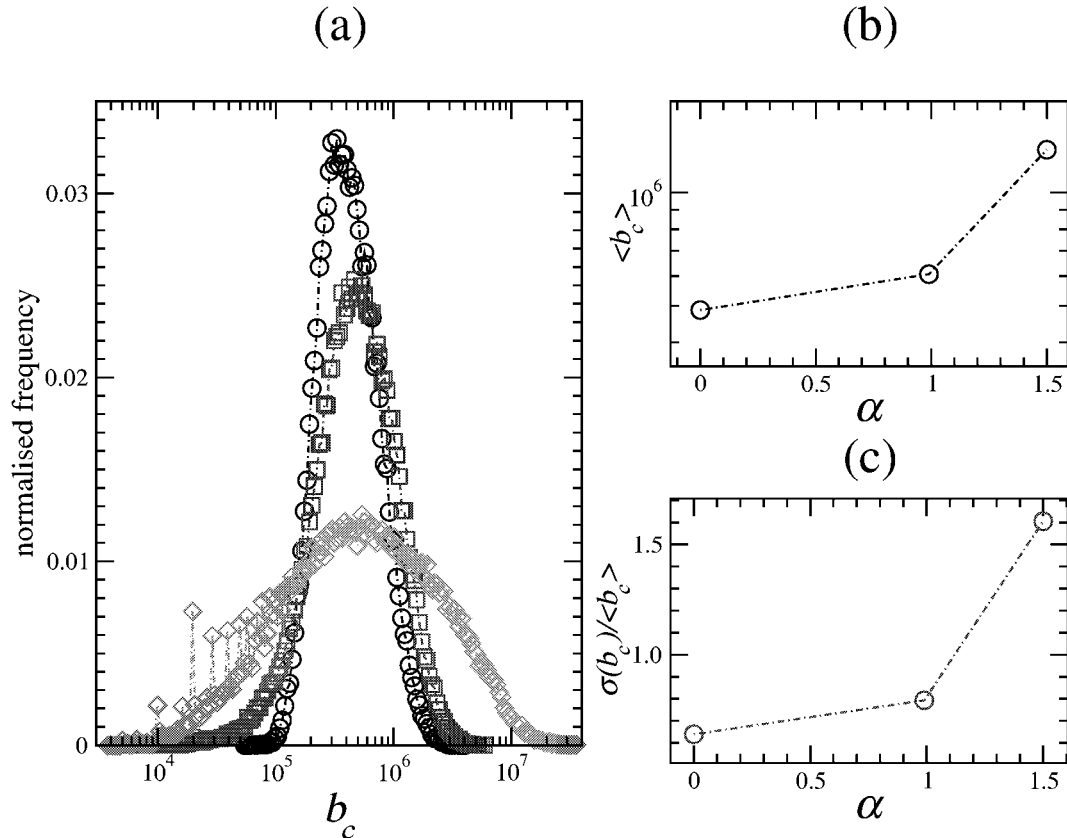


Figure 5.5: Distribution of flows over the additional links for 1-dimensional lattices ($p = 0.1$, $N = 10^4$, 100 realisations). (a): full distributions (\circ : $\alpha = 0$, \square : $\alpha = 1$ and \diamond : $\alpha = 1.5$). (b): mean values of the betweenness centralities and (c): relative variances.

5.3 Distribution of flows

Up to this point, we have seen that a positive value of α in the connection-length distribution leads to favourable properties: for example, given a certain amount of wiring length, choosing a higher α (but still below the value $\alpha_c = D+1$) leads to a smaller world. In order to obtain a more complete picture, we also studied the implications of our model regarding the distribution of flows over the additional links. The corresponding quantity is the link betweenness centrality which was introduced in Chapter 1. Let us briefly recall this measure. If every node sends one packet (of information) to every other node, there are

$$b_c(s) = \sum_{A,B} \frac{n_{AB}(s)}{n_{AB}} \quad (5.9)$$

packets flowing through link s , where n_{AB} is the number of shortest paths between nodes A and B , and $n_{AB}(s)$ counts only those going through connection s , the

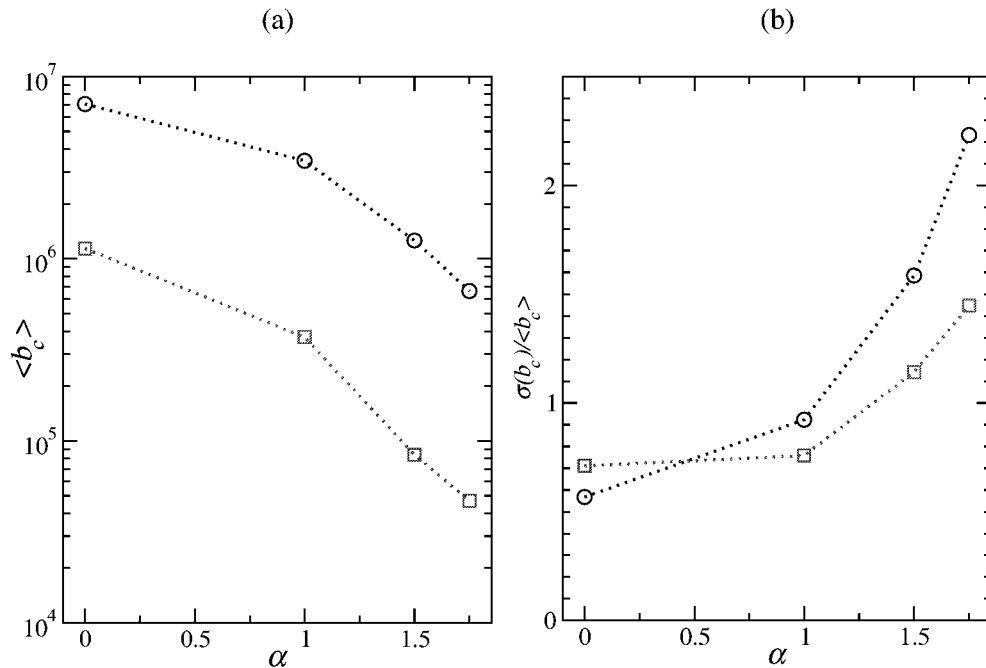


Figure 5.6: Betweennesses for 1D networks ($N = 10^4$) where the wiring costs associated to the additional links are fixed. First (a) and second (b) moments (\circ : $C_W/N = 10$, \square : $C_W/N = 100$). These results were obtained by averaging over 100 realisations of networks.

sum running over all pairs of nodes³. Fig. 5.5 reports the full distribution of the betweennesses (a), the corresponding mean values (b) and fluctuations (c) for 1D networks with a fixed number of additional links ($p = 0.1$, $N = 10^4$) and $\alpha = 0, 1, 1.5$. These computations were performed with the help of an efficient algorithm [38, 113]. The observation that $\langle b_c \rangle$ increases with α can be understood through the following argument [114]: $\langle b_c \rangle$ is obtained from Eq. (5.9) by summing over all links and dividing the result by E , the total number of them, that is,

$$\langle b_c \rangle = \frac{1}{E} \sum_{A,B} \frac{1}{n_{AB}} \underbrace{\sum_s n_{AB}(s)}_{n_{AB} d_{AB}} = \frac{1}{E} \sum_{A,B} d_{AB} = \frac{N^2}{E} \langle d \rangle \quad (5.10)$$

where d_{AB} is the distance between nodes A and B, i.e. the length in terms of the number of links of the corresponding shortest path. Hence, if the links of the underlying lattice are ignored, we indeed expect that $\langle b_c \rangle$ increases with α since (i) N and E are independent of α and (ii) $\langle d \rangle$ is an increasing function of α (Fig. 5.2). Regarding the fluctuations, one can argue similarly. The second

³It is assumed that every packet is routed through the shortest path.

moment of the betweenness centrality, i.e.

$$\langle b_c^2 \rangle = \frac{1}{E} \sum_s \sum_{A,B} \frac{n_{AB}(s)}{n_{AB}} \sum_{C,D} \frac{n_{CD}(s)}{n_{CD}} = \frac{1}{E} \sum_{A,B} \frac{1}{n_{AB}} \sum_{C,D} \frac{1}{n_{CD}} \sum_s n_{AB}(s)n_{CD}(s) \quad (5.11)$$

is essentially determined by the extent to which shortest paths overlap since for a given link s , the factor $n_{AB}(s)n_{CD}(s)$ is non-zero only if the link s is part of a shortest path from A to B and of another one from C to D. Evidently, the higher α the smaller the fraction of real shortcuts, i.e. those carrying the traffic, and the very same links also overlap more and more. This qualitatively explains the behaviour of $\sigma(b_c)/\langle b_c \rangle = \sqrt{\langle b_c^2 \rangle / \langle b_c \rangle^2 - 1}$ observed in Fig. 5.5.

It is also interesting to study the betweenness distribution when the wiring resources (to build the network) are fixed. In Fig. 5.6, we report the numerical results for $C_W/N = 10$ (lower sets) and $C_W/N = 100$ (upper sets), the remaining parameters being $D = 1$ and $N = 10^4$. Panel (a) illustrates that $\langle b_c \rangle$ now decreases with α . This is in agreement with Eq. (5.10), i.e. $\langle b_c \rangle \sim \langle d \rangle / E$, since both $\langle d \rangle$ (see Fig. 5.4) and $1/E \sim (pN)^{-1} = \langle l \rangle / C_W$ are decreasing functions of α . Regarding the variances, panel (b) shows that

$$\left. \frac{\sigma(b_c)}{\langle b_c \rangle} \right|_{C_W/N=10} > \left. \frac{\sigma(b_c)}{\langle b_c \rangle} \right|_{C_W/N=100}$$

for $\alpha \gtrsim 0.5$. This indicates that increasing the total costs, thus providing more additional links, reduces the degree to which shortest paths overlap - if we adopt the reasoning given for $p = \text{const}$ - which is reasonable.

5.4 Robustness

In the previous sections of this chapter, we have investigated properties of spatial small-world networks (with connection lengths that decay as a power law) whose connectivity patterns were not changed, once the additional links had been added. Yet, it is interesting to study the behaviour of the networks with respect to link deletion. The associated concept, that is, robustness, can for example be defined as the extent to which the mean distance increases when a definite fraction of additional links is failing. Two different types of failure can be distinguished. On the one hand, it is possible that certain connections are malfunctioning for whatever reason, resulting in *random failures*. The other scenario is directly related to the traffic on the network: as the links may only be able to carry flow amounts which lie below a certain threshold, *overload* of the additional connections occurs which is the second type we shall investigate. We further merely compare networks with equal (initial) wiring costs, that is the trivial case $p = \text{const}$ is not studied in this section.

As far as random failures are concerned, Fig. 5.7 illustrates to what extent the mean distances of 1D networks increase for $C_W/N = 10$ (left) and $C_W/N = 100$

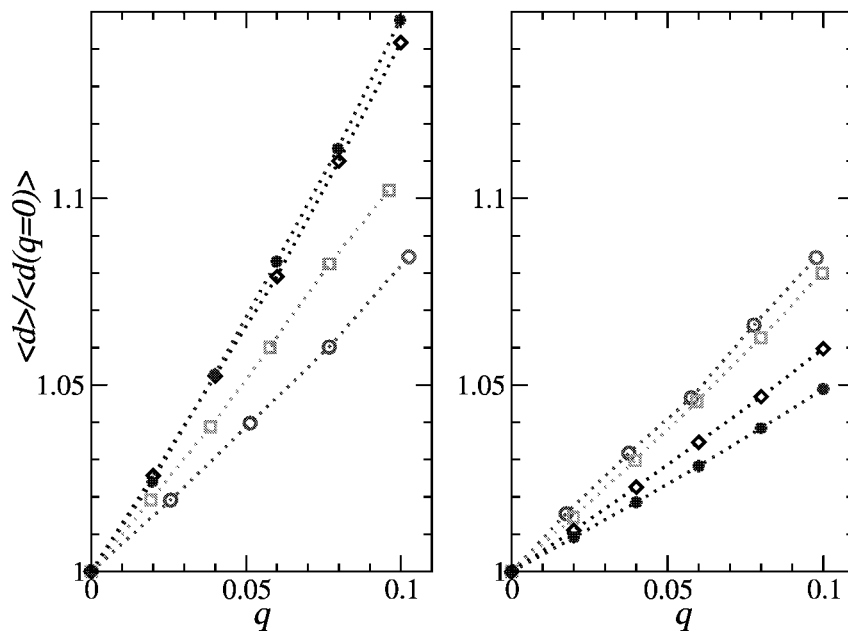


Figure 5.7: Robustness with respect to random failures. These figures show to what extent the mean distance increases when a fraction q of the additional connections is removed. The initial 1D networks have a fixed cost: $C_W/N = 10$ (left) and $C_W/N = 100$ (right), and α takes on the values 0 (\circ), 1 (\square), 1.5 (\diamond) and 1.75 (\star). These results were obtained by averaging over 1000 realisations of networks consisting of $L = 10^4$ nodes.

(right), α taking the values 0, 1, 1.5 and 1.75 if up to 10% of the additional links are deleted. We find that for low costs ($C_W/N = 10$), $\alpha = 0$ corresponds to the most robust system, becoming more fragile as α increases. Yet, at high wiring costs ($C_W/N = 100$), the WS-type network ($\alpha = 0$) is most vulnerable, and the robustness related to random failures undergoes an inversion between these two cost values, simply reflecting the non-trivial behavior of $\langle d \rangle(p)$. We observed an analogous effect in 2 dimensions. Changing the network structure through such random failures obviously entails a redistribution of the flows. Fig. 5.8 shows that $\langle b_c \rangle$ increases with the failure fraction q , which does not come as a surprise.

In the case of overload, we found the behaviour to be independent of the wiring costs of the initial network. The vulnerability always increases with α , that is, the relative increase of the mean distance (caused by deleting a certain fraction of the most charged links) is lowest for $\alpha = 0$. Fig. 5.9 reports this behaviour for 1D networks, the lowest set corresponding to $\alpha = 0$ and the uppermost to $\alpha = 1.75$ ($C_W/N = 10$). This finding is in agreement with the arguments given above in the

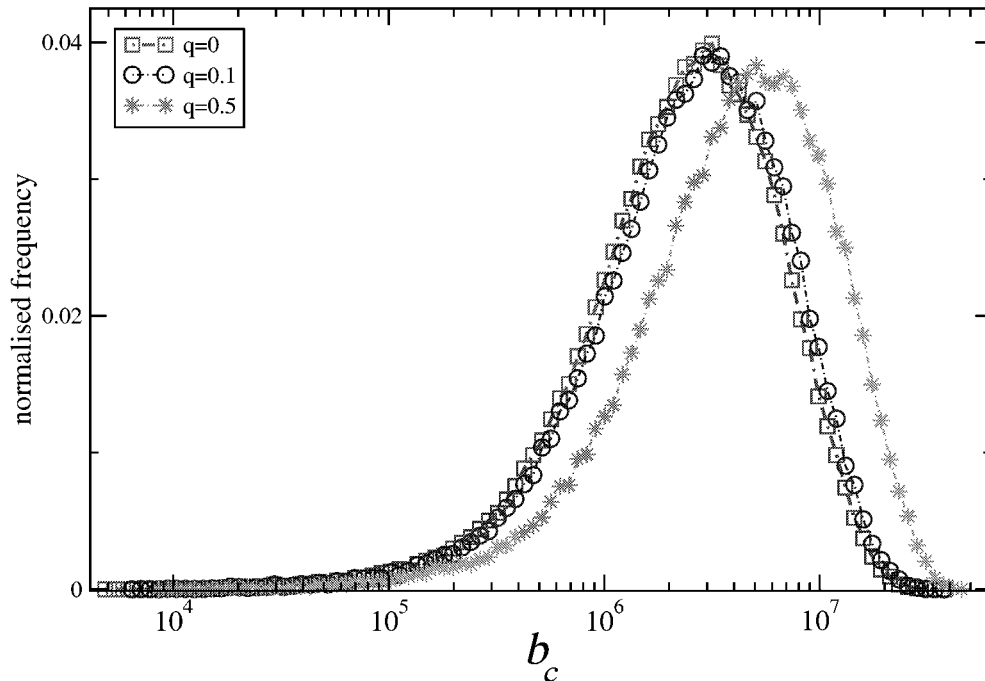


Figure 5.8: Link betweenness centrality distribution for several (random) removal fractions q . The distribution associated to the initial network (characterised by $C_W/N = 10$, corresponding to $q = 0$ and \square) is shifted rightwards as q increases to 0.1 (\circ) and 0.5 (\star). These results were obtained by averaging over 1000 realisations of networks of size ($N = 10^4$).

previous section. The higher α the smaller the fraction of real shortcuts carrying the majority of the traffic and making $\langle d \rangle$ small. As a consequence, $\langle d \rangle$ increases the faster upon their removal the larger α . Furthermore, this result does not depend on the details of the overloading process: whether a given fraction of the most charged links is removed simultaneously or the failure is accomplished in a cascade-like fashion, the dependency of the robustness from α remains unaltered. The bold symbols in the inset of Fig. 5.9 show the cascade-like case, illustrating that $\langle d \rangle / \langle d \rangle(q = 0)$ increases with α as well. In 2 dimensions, we obtained analogous results (Table 5.2).

α	0	1	2	3
$\langle d \rangle / \langle d \rangle(q = 0)$	1.05 ± 0.01	1.08 ± 0.01	1.24 ± 0.02	1.37 ± 0.04

Table 5.2: Overload-related robustness (simultaneous removal of 10% of the most charged connections). The values shown here were obtained by simulating the process for 100 different realisations of networks consisting of $N = 100 \times 100$ sites with initial wiring costs $C_W/N = 10$. The emerging picture is that networks with low α are most robust.

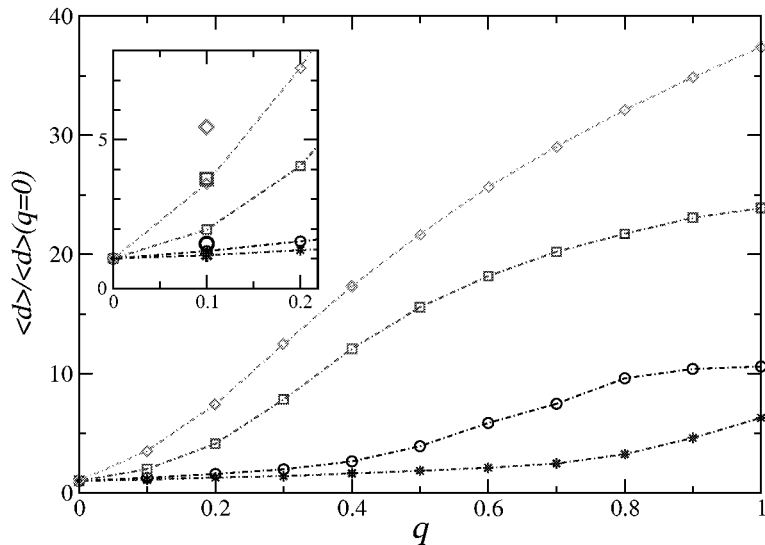


Figure 5.9: Robustness with respect to overload. This figure shows the extent to which the mean distance increases upon removing a fraction q of the most charged, in terms of flow, links for 1D networks ($N = 10^4$, $C_W/N = 10$) for $\alpha = 0$ (\star), $\alpha = 1$ (\circ), $\alpha = 1.5$ (\square) and $\alpha = 1.75$ (\diamond). The inset enlarges the region $0 \leq q \leq 0.2$ and the relative increase of $\langle d \rangle$ upon a cascade-like removal of links is also shown (for $q = 0.1$, bold symbols). Here we averaged over 10 realisations of failures and networks.

5.5 Discussion

This chapter was devoted to our speculation that there is more about the small-world phenomenon as was known until very recently, especially in the case where the links have a physical correspondence, e.g. computer connections, axonal or dendritic wire (brain) or metallic wire (integrated circuits). In Watts and Strogatz' model where the resulting shortcut-length distribution is uniform, building a small world is a costly affair. In the context of flows over the network, most of the traffic is carried by the few shortcuts (which are $\mathcal{O}(L)$ long). This makes the system very vulnerable with respect to their deletion.

Since it is theoretically interesting and because power-law decaying connection-length distributions have been reported in various contexts, we performed a detailed analysis of appropriate model networks. The model consisted in supplementing a D -dimensional lattice with additional links whose lengths are distributed according to $\sim l^{-\alpha}$. Instead of the lattice, one could also start out with a random distribution of nodes in D -dimensional space with only nearest neighbour connections, the nearest neighbours being defined via Voronoi tessellations [115].

By means of a simple analytical argument and through extensive numerical simulations, we showed that small-world topology can be realised for $\alpha < \alpha_c =$

$D+1$. By studying the scaling of the moments of the length distribution, we found that our small-world networks can be constructed in a very economical way if D and α are chosen appropriately. In 3 dimensions, it is for example possible to build a small world while $\langle l \rangle = \mathcal{O}(1) = \langle l^2 \rangle$. If networks differing in α with equal wiring costs are compared, it turns out that the larger α , the smaller the world (as long as $\alpha < \alpha_c$). Networks with a positive α also show favourable attributes when it comes to the distribution of flows over the links. Finally, above a certain value of the wiring costs C_W/N , high α networks are more robust than low α ones with respect to random failure of links while for the case of overload - be it the *simultaneous* removal of the most charged connections or a *cascade-like* overload - the vulnerability increases with α . These findings complement the observation that power-law decaying connection-length distributions emerge quite naturally when wiring costs along with shortest paths are minimised [69].

Chapter 6

General conclusions and outlook

In this interdisciplinary study, we have employed the framework of statistical physics in order to address several questions regarding the topology of large and complex networked systems. The selected problems fall into the categories (a) ‘how to obtain the topology’, (b) ‘topology as cause’ and (c) ‘topology as effect’, these keywords providing an overall characterisation of the field of complex networks. More precisely, it is not a trivial task to get the graphical representation for many classes of networks (a); it is important to unravel the role of topology as far as dynamic processes taking place on networks are concerned (b), as well as to understand what factors shape the topology of complex networks (c). In the following, we briefly summarise the results of this study, draw our conclusions and give suggestions for future investigations.

Obtaining the graphical representation (a) is particularly challenging in the case of the Internet (at the router level). Unlike the World Wide Web where links are visible in the form of hyperlinks, the Internet’s topology must be queried indirectly, e.g. by sending data packets from an arbitrarily chosen router to a set of targets. We examined whether the map obtained through such an exploration reflects the real topology or if this discovery method introduces systematic errors resulting in a distortion of the topology. In order to approach this problem, we mimicked the probing process by applying a treelike exploration method to networks whose topology we know, making it possible to compare the original connectivity patterns with the observed ones. At a qualitative level, the overall topology is fairly well captured, that is homogeneity or heavy-tailed degree distributions are usually conserved by the mapping process. Dramatic effects, such as the appearance of power-law degree distributions for underlying homogeneous graphs, are found only in very peculiar cases. At a more quantitative level, we numerically found for scale-free model networks, namely for the preferential attachment model by Barabási and Albert and for the intrinsic vertex fitness model, that the explorer “sees” a smaller exponent than the original one. For the first of these models, we were able to support our finding with a simple analytical argument. The interpretation of a reduced exponent is that low-degree vertices

are underrepresented in the explored network. This is reasonable because these vertices have fewer paths reaching them as the betweenness centrality generally increases with the degree in the case of scale-free networks. However, a deeper and model-independent understanding of this effect is still missing. This is of major importance and should, in principle, allow for a prediction of the real exponent, given the measured one. Clearly, the degree distribution is only one topological property of a complex network. Regarding the interconnectedness, a treelike exploration gives a good estimate for the diameter along with an underrepresentation of short loops. It would thus be interesting to study to what extent $C(k)$ is distorted, i.e. whether the mapping process conserves the hierarchical structure. It would also be an interesting problem to investigate if degree correlations are subject to such biases.

While our simulations were based on single source explorations, other authors pointed out how the mapping improves if more than one source is used. All the mentioned efforts and suggestions may ultimately lead to more efficient mapping strategies and to a set of rules describing how to extrapolate for the real topology from the measured one.

The second and major part of this thesis was devoted to dynamic processes taking place on complex networks (b). More precisely, we studied the spreading of an epidemic and examined how the topology influences the spreading behaviour. The importance of topology in this context has been noted only rather recently: it had been a long-standing problem why computer viruses are so persistent until the scale-free topology of the Internet was shown to lie at the root of the absence of any finite epidemic threshold. Interestingly, this result can be obtained through the simplest mean-field approximation.

We have shown that ignoring “real space” pair correlations, i.e. assuming that the states of connected infected nodes are uncorrelated, does not imply the absence of pair correlations in “degree space”. In other words, the fraction of infected nodes of a given degree depends on the degrees of the nearest neighbours. Further approximations then led to different levels of mean-field descriptions, the final one corresponding to the simplest type mentioned in the previous paragraph. These descriptions give a precise quantitative interpretation of the degree distribution and the degree correlations in the dynamic behaviour.

While the role of the degree-related topology in the spreading behaviour can be uncovered rather straightforwardly, it is much harder to unravel the corresponding role of loops. Clearly, such an understanding cannot be gained at the mean-field level where the number of nearest neighbours of a given node is the only topological property entering into the approximation. By describing the epidemic dynamics as a Markovian process at an exact level, we derived two methods which yield a quantitative interpretation of how local interconnectedness determines the dynamic behaviour, namely a two-time-step description and cluster approximations.

The former method was elaborated for homogeneous networks in which every node “sees” identical connectivity patterns, and it consists in performing two time steps exactly. This led to a subgraph development allowing to unravel how the way the next-nearest neighbours (of an arbitrarily chosen node) are arranged influences the spreading behaviour. The resulting equation describes how clustering and grid-like ordering determines the onset of an epidemic, not only for strictly homogeneous networks, but also for disordered ones in which every node still has the same degree. The method could in principle be extended to heterogeneous networks, i.e. graphs in which not every node has the same degree. This would lead to an understanding of how the degree distribution, the degree correlations and the degree-dependent loop densities *together* determine the spreading behaviour. The other line of extension is to carry out more than two time steps exactly. In a 3-step description, the simulation result is approximated even better, also providing a natural classification of loops up to length six. More generally, if n steps are performed exactly, the accuracy improves systematically. In addition, the accompanying classification of loops up to length $2n$ might be interesting in its own right, given the current high attention paid to the study of network motifs and subgraphs.

In the second method, one has to choose a fundamental cluster and the exact description then leads to a set of equations for the probabilities that the cluster is in a given configuration. As the choice of this cluster reflects the underlying network topology, this method is more adequate for homogeneous networks. We observed that in order to obtain a given accuracy, the cluster size must be chosen much larger for regular lattices than for random networks. This can be understood if we recall that the local treelike topology effectively makes random networks infinite-dimensional structures. As the accuracy of the mean-field approximation increases with the dimension, i.e. the higher the dimension the weaker the dynamical correlations, the above observation does not come as a surprise. However, further work is needed in order to identify a controlling parameter that determines the required cluster size for a given level of accuracy. In analogy to the different levels of mean-field descriptions, the pair approximation could also be transformed into degree space. In such a description, the (degree-dependent) clustering coefficient could be taken into account by interpreting it as the proportion of open to closed triples. This hints at another approach uncovering the combined effect of the degree-related topology and the loop structure. Since it is currently not known whether the epidemic threshold vanishes for highly clustered scale-free networks, the elaboration of computational strategies in this direction is of the utmost importance.

In principle, one could also imagine n -step cluster approximations, that is exploiting temporal and spatial correlations simultaneously. In the framework of a “two-step pair approximation”, the probability that the nodes up to two links away from an arbitrarily chosen pair are in given states would then have to be expressed in terms of intricately overlapping pairs.

Finally, we wish to stress that our methods are independent of the choice of the epidemiological model. Focusing on the role of topology, the SIS model was an appropriate choice so as to keep the complexity due to the contact process at a minimum. However, if we allowed the nodes to be in three or four possible states, our methodologies could also be formulated for the susceptible \rightarrow infected \rightarrow recovered (SIR) model or for the susceptible \rightarrow exposed \rightarrow infected \rightarrow recovered (SEIR) model. In fact, our methods could also be employed in order to describe a dynamic process other than the spreading of an epidemic since they are systematic descriptions of a system that evolves according to the rules of a probabilistic cellular automaton. One example would be the dynamics of a neuronal network: a neuron can be susceptible to synaptic inputs, fire an action potential thus “infecting” other nodes, and display a refractory period after action potential firing, corresponding to a third state.

Overall, our methods are rigorous approximations for dynamic probabilistic systems of which we merely know the time evolution. In this context, the probability of finding the system in a given configuration at a given time (even in equilibrium) is only implicitly determined through its (Markovian) time evolution. Due to their generality, we believe that our methods will prove useful in different contexts.

In the case of a lethal infectious disease, nodes disappear in the course of time. This means that a network on which a dynamic process takes place can be altered by that process. Therefore, while we clarified the role of topology as far as the spread of an epidemic is concerned (b), it should be beared in mind that a dynamic process of this type can also shape the topology of a network (c). In this thesis, we considered another factor that determines to some extent the topology of complex networks, namely spatial constraints. For a network “living” in a Euclidean space, the positions of the nodes matter and the links have a physical length to which a wiring cost can be associated. By adopting a wiring cost perspective, we investigated under what conditions it is possible to realise small-world networks. At first sight, wiring minimisation seems to conflict with the emergence of a high global interconnectedness since that topological property is due to the presence of long-range connections which are costly. Yet, by supplementing D -dimensional lattices with additional links whose lengths l are distributed according to a decaying power law, i.e. $l^{-\alpha}$, it resulted that small-world networks can be constructed in a very economical way if D and α are chosen appropriately. When it comes to flows over the links, we found that their distribution becomes more optimal as α increases. Concerning random failures of links, we obtained the non-trivial picture that the small-world network by Watts and Strogatz ($\alpha = 0$) is most vulnerable if a large amount of wire is available. In the case of overload, on the other hand, the length distribution alone fully determines the robustness, that is, networks characterised by a high value of α are most vulnerable. As length distributions of the type investigated here have

been observed in a number of real-world networks, such as integrated circuits, the Internet or the human cortex, we believe these results will have intriguing implications in their modelisation.

Besides wiring minimisation, there are other factors which shape the topology of real, spatial small-world networks. The wiring patterns of a neural network are for example subject to the minimisation of conduction delays related to the transmission of signals along axons and dendrites. Furthermore, since ‘growth’ is a characteristic ingredient partly determining the architecture of the Internet and neural networks, it would also be interesting to examine a model involving growth along with simple rules for the establishment of connections with regard to link length distributions.

In summary, this study was devoted to the topology of complex networks: how we obtain it (a), what its role in dynamic processes is (b) and how specific factors shape it (c). In particular, we examined the treelike exploration of scale-free networks and discussed the implications regarding the topology of the Internet. As far as probabilistic processes on networks are concerned, we developed different levels of mean-field approximations and two systematic methods which unravel the role of short loops in the dynamic behaviour. Finally, we demonstrated that many properties of small-world networks are favoured when the shortcut length distribution is a decaying power law rather than a uniform distribution.

In a broader sense, it would for example be interesting to examine problems at the interface of (b) and (c). That is, the dynamics of the network and that of the epidemic process are usually coupled in a real situation, leading to a deeper understanding of the interplay between topology and dynamics. Furthermore, we only dealt with binary networks in which nodes are either connected or not. It might be worthwhile to generalise the ideas presented in this thesis to weighted networks, especially the methods serving to describe dynamic probabilistic processes on networks. In an epidemiological context, this would mean that the likelihood for the virus to be transmitted along a given (social) link depends on its weight or, as far as the Internet is concerned, a virus is more likely to propagate along a connection with a large flow of data. In addition to the concrete open questions given in the above paragraphs, these are all further interesting future directions for which the insights of this thesis provide a valuable basis.

The results of this Ph.D. thesis were published in journals read by the scientific community interested in complex networks and mathematical epidemiology, and part of them were also presented to the neuroscience community.

Appendix A

Full subgraph developments

On the following pages, we show the complete subgraph developments which lead to the coefficients determining the epidemic threshold at the two-step level described in Sec. 4.4 for the square lattice (Tab. A.1), the Kagomé lattice (Tab. A.2) and the ring-type network (Tab. A.3). Every subgraph corresponds to a term in Eq. (4.16), its contribution is obtained by the procedure illustrated in Sec. 4.4.1. The $\lambda \cdot \lambda^n$ -coefficient of the threshold equation is obtained by summing all the $\mathcal{O}(P)$ contributions (taking into account the multiplicities) of the n -th-order subgraphs.

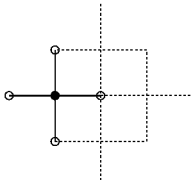
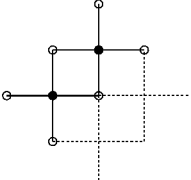
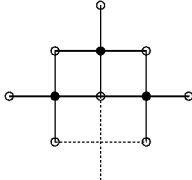
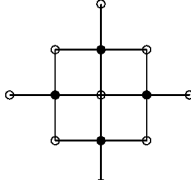
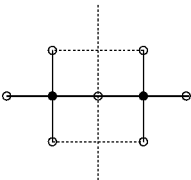
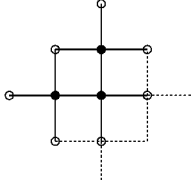
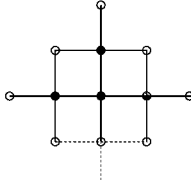
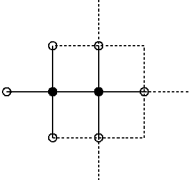
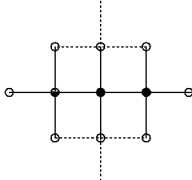
order (n)	1	2	3	4
subgraph				
contribution	$4P + \mathcal{O}(P^2)$	$2P + \mathcal{O}(P^2)$	$P + \mathcal{O}(P^2)$	$P + \mathcal{O}(P^2)$
multiplicity	4	4	4	1
subgraph				
contribution		$P + \mathcal{O}(P^2)$	$2P^2 + \mathcal{O}(P^3)$	$5P^3 + \mathcal{O}(P^4)$
multiplicity		2	4	4
subgraph				
contribution		$9P^2 + \mathcal{O}(P^3)$	$18P^3 + \mathcal{O}(P^4)$	
multiplicity		4	2	
$\lambda \cdot \lambda^n$ -coeff.	16	-10	4	-1

Table A.1: Subgraph development for the square lattice. All the terms in Eq. (4.16) are symbolized by a specific subgraph, its order being given by the number of filled circles. The $\lambda \cdot \lambda^2$ -coefficient -10, as an example, is obtained by summing the various $\mathcal{O}(P)$ contributions, that is $2 \cdot 4 + 1 \cdot 2 + 0 \cdot 4 = 10$, and the negative sign comes from Eq. (4.16). The $\lambda \cdot \lambda^3$ - and $\lambda \cdot \lambda^4$ -coefficients are unaltered with respect to the treelike topology although other subgraphs enter into the development.

order (n)	1	2	3	4
subgraph				
contribution	$4P + \mathcal{O}(P^2)$	$P + \mathcal{O}(P^2)$	$P + \mathcal{O}(P^2)$	$P + \mathcal{O}(P^2)$
multiplicity	4	6	4	1
subgraph				
contribution		$P + \mathcal{O}(P^2)$	$5P^2 + \mathcal{O}(P^3)$	$8P^4 + \mathcal{O}(P^5)$
multiplicity		4	4	4
subgraph				
contribution			$8P^3 + \mathcal{O}(P^4)$	
multiplicity			2	
$\lambda \cdot \lambda^n$ -coeff.	16	-10	4	-1

Table A.2: The subgraphs of all the orders for the Kagomé lattice. See Tab. A.1 for how the coefficients are obtained and as far as further details are concerned.

order (n)	1	2	3	4
subgraph				
contribution	$4P + \mathcal{O}(P^2)$	$2P + \mathcal{O}(P^2)$	$P + \mathcal{O}(P^2)$	$P + \mathcal{O}(P^2)$
multiplicity	4	4	4	1
subgraph				
contribution		$2P + \mathcal{O}(P^2)$	$P + \mathcal{O}(P^2)$	$P^2 + \mathcal{O}(P^3)$
multiplicity		2	2	2
subgraph				
contribution		$P + \mathcal{O}(P^2)$	$3P^2 + \mathcal{O}(P^3)$	$2P^2 + \mathcal{O}(P^3)$
multiplicity		1	3	2
subgraph				
contribution		$P + \mathcal{O}(P^2)$	$5P^2 + \mathcal{O}(P^3)$	
multiplicity		3	1	
$\lambda \cdot \lambda^n$ -coeff.	16	-16	6	-1

Table A.3: The full subgraph development for the ring-type network. See Tab. A.1 for the derivation of the $\lambda \cdot \lambda^n$ -coefficients. With respect to the two lattices treated above, the $\lambda \cdot \lambda^2$ - and $\lambda \cdot \lambda^3$ -coefficients are -16 and 6.

Bibliography

- [1] K. Huang. *Statistical mechanics* (John Wiley and Sons, New York, 1987).
- [2] B. Bollobás. *Random graphs* (Academic Press, London, 1985).
- [3] L. Euler. *Solutio problematis ad geometriam situs pertinentis*. Commentarii academiae scientiarum Petropolitanae **8**, 128-140 (1741).
- [4] A. Vázquez. *Degree correlations and clustering hierarchy in networks: Measures, origin and consequences*. (Ph.D. dissertation, ISAS Trieste, 2002).
- [5] R. Pastor-Satorras and A. Vespignani. *Evolution and structure of the Internet: A statistical physics approach* (Cambridge University Press, Cambridge, 2004).
- [6] <http://www.lumeta.com>
- [7] K. Claffy, T. Monk and D. McRobb. *Internet tomography*. Nature, Web matters, January 7th 1999.
- [8] R. Govindan and H. Tangmunarunkit. *Heuristics for Internet map discovery*. Proc. IEEE INFOCOM **3**, 1371-1380 (2000).
- [9] A. Broder, R. Kumar, G. Maghoul, P. Raghavan, S. Rajalopagan, R. Stata, A. Tomkins and J. Wiener. *Graph structure in the web*. Comput. Netw. **33**, 309-320 (2000).
- [10] B. Levin (Ed.) *Sex i Sverige: Om Sexuallivet i Sverige (Sex in Sweden: On the sexual life in Sweden)* (Natl. Inst. Pub. Health, Stockholm, 1996).
- [11] A. Klovdahl. *View-net: A new tool for network analysis*. Social Networks **8**, 313-342 (1986), see also <http://www.nd.edu/~networks/gallery.htm>
- [12] R. Rojas. *Neural networks: A systematic introduction*. (Springer, Berlin, 1996).
- [13] S. Dodel, J.M. Herrmann and T. Geisel. *Functional connectivity by cross-correlation clustering*. Neurocomp. **44**, 1065-1070 (2002).

- [14] V.A. Klyachko and C.F. Stevens. *Connectivity optimisation and the positioning of cortical areas*. Proc. Natl. Acad. Sci. **100**, 7937-7941 (2003).
- [15] J. Gross and J. Yellen. *Graph Theory and its Applications* (CRC Press, 1999).
- [16] V. Latora and M. Marchiori. *Efficient behaviour of small-world networks*. Phys. Rev. Lett. **87**, 198701 (2001).
- [17] D.J. Watts and S.H. Strogatz. *Collective dynamics of 'small-world' networks*. Nature **393**, 440-442 (1998).
- [18] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon. *Network motifs: Simple building blocks of complex networks*. Science **298**, 824-827 (2002).
- [19] S. Milgram. *The small-world problem*. Psychology Today **1**, 60-67 (1967).
- [20] P. Erdős and A. Rényi. *On random graphs*. Publ. Math. (Debrecen) **6**, 290-297 (1959).
- [21] P. Erdős and A. Rényi. *On the evolution of random graphs*. Publ. Math. Inst. Hung. Acad. Sci. **5A**, 17-61 (1960).
- [22] L.A.N. Amaral, A. Scala, M. Barthélemy and H.E. Stanley. *Classes of small-world networks*. Proc. Natl. Acad. Sci. **97**, 11149-11152 (2000).
- [23] M.E.J. Newman. *Assortative mixing in networks*. Phys. Rev. Lett. **89**, 208701 (2002).
- [24] E. Ravasz and A.-L. Barabási. *Hierarchical organisation in complex networks*. Phys. Rev. E **67**, 026112 (2003).
- [25] G. Sabidussi. *The centrality index of a graph*. Psychometrika **31**, 581-602 (1966).
- [26] P. Hage and F. Harary. *Eccentricity and centrality in networks*. Social Networks **17**, 57-63 (1995).
- [27] L.C. Freeman. *A set of measures of centrality based on betweenness*. Sociometry **40**, 35-41 (1977).
- [28] M. Faloutsos, P. Faloutsos and C. Faloutsos. *On power-law relationships of the Internet topology*. Comput. Commun. Rev. **29**, 251-262 (1999).
- [29] A. Vázquez, R. Pastor-Satorras and A. Vespignani. *Topology, hierarchy and correlations in Internet graphs*. Lect. Notes Phys. **650**, 425-440 (2004).
- [30] R. Pastor-Satorras, A. Vázquez and A. Vespignani. *Dynamical and correlation properties of the Internet*. Phys. Rev. Lett. **87**, 258701 (2001).

- [31] A. Vázquez, R. Pastor-Satorras and A. Vespignani. *Large-scale topological and dynamical properties of the Internet*. Phys. Rev. E **65**, 066130 (2002).
- [32] G. Caldarelli, R. Pastor-Satorras and A. Vespignani. *Structure of cycles and local ordering in complex networks*. Eur. Phys. J. B **38**, 183-186 (2004).
- [33] R. Albert and A.-L. Barabási. *Statistical mechanics of complex networks*. Rev. Mod. Phys. **74**, 47-97 (2002).
- [34] R. Ferrer i Cancho, C. Janssen and R.V. Solé. *Topology of technology graphs: Small-world patterns in electronic circuits*. Phys. Rev. E **64**, 046119 (2001).
- [35] A.-L. Barabási and R. Albert. *Emergence of scaling in random networks*. Science **82**, 509-512 (1999).
- [36] M.E.J. Newman. *The structure of scientific collaboration networks*. Proc. Natl. Acad. Sci. **98**, 404-409 (2001).
- [37] M.E.J. Newman. *Scientific collaboration networks. I. Network construction and fundamental results*. Phys. Rev. E **64**, 016131 (2001).
- [38] M.E.J. Newman. *Scientific collaboration networks. II. Shortest paths, weighted networks and centrality*. Phys. Rev. E **64**, 016132 (2001).
- [39] F. Liljeros, C.R. Edling, L.A.N. Amaral, H.E. Stanley and Y. Åberg. *The web of human sexual contacts*. Nature (London) **411**, 907 (2001).
- [40] L. Falk, M. Lindberg, M. Jurstrand, A. Bäckman, P. Olcén and H. Fredlund. *Genotyping of Chlamydia trachomatis would improve contact tracing*. Sex. Transm. Diseases **30**, 205-210 (2003).
- [41] P.S. Bearman, J. Moody and K. Stovel. *Chains of affection: The structure of adolescent romantic and sexual networks*. Am. J. Soc. **110**, 44-91 (2004).
- [42] O. Shefi, I. Golding, R. Segev, E. Ben-Jacob and A. Ayali. *Morphological characterisation on in-vitro neuronal networks*. Phys. Rev. E **66**, 021905 (2002).
- [43] O. Sporns and J.D. Zwi. *The small world of the cerebral cortex*. Neuroinf. **2**, 145-162 (2004).
- [44] R. Martin, M. Kaiser, P. Andras and M. Young. *Is the brain a scale-free network?* Soc. Neurosci. Abstr. **27**, 814-816 (2001).
- [45] V.M. Eguíluz, D.R. Chialvo, G. Cecchi, M. Baliki and A.V. Apkarian. *Scale-free brain functional networks*. Phys. Rev. Lett. **94**, 018102 (2005).
- [46] S.R. White. *Open problems in computer virus research*. Virus bulletin conference (<http://www.research.ibm.com/antivirus/SciPapers.htm>) 1998.

- [47] R. Pastor-Satorras and A. Vespignani. *Epidemic spreading in scale-free networks*. Phys. Rev. Lett. **86**, 3200-3203 (2001).
- [48] A.-L. Barabási, Z. Dezsö, E. Ravasz, S.-H. Yook and Z. Oltvai. *Scale-free and hierarchical structures in complex networks*. AIP Conf. Proc. **661**, 1-16 (2003).
- [49] R. Albert, H. Jeong and A.-L. Barabási. *Error and attack tolerance of complex networks*. Nature **406**, 378-382,542 (2000).
- [50] 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).
- [51] R. Cohen, K. Erez, D. ben-Avraham and S. Havlin. *Resilience of the Internet to random breakdowns*. Phys. Rev. Lett. **85**, 4626-4628 (2001).
- [52] R. Cohen, K. Erez, D. ben-Avraham and S. Havlin. *Breakdown of the Internet under intentional attack*. Phys. Rev. Lett. **86**, 3682-3685 (2001).
- [53] D. Stauffer, A. Aharony, L. da Fontoura Costa and J. Adler. *Efficient Hopfield pattern recognition on a scale-free neural network*. Eur. Phys. J. B **32**, 395-399 (2003).
- [54] K. MacLeod, A. Bäcker and G. Laurent. *Who reads temporal information contained across synchronised and oscillatory spike trains?* Nature **395**, 693-698 (1998).
- [55] L.F. Lago-Fernández, R. Huerta, F. Corbacho and J.A. Sigüenza. *Fast response and temporal coherent oscillations in small-world networks*. Phys. Rev. Lett. **84**, 2758-2761 (2000).
- [56] S.P. Gorman and R. Kulkarni. *Spatial small worlds: New geographic patterns for an information economy*. Environment and Planning B: Planning and Design **31**, 273-296 (2003).
- [57] M.E.J. Newman. *Properties of highly clustered networks*. Phys. Rev. E **68**, 026121 (2003).
- [58] H. Matsuda, N. Ogita, A. Sasaki and K. Sato. *Statistical mechanics of population - the lattice Lotka-Volterra model*. Prog. Theor. Phys. **88**, 1035-1049 (1992).
- [59] M. Van Baalen. *Pair approximations for different spatial geometries* in: Diekmann, U., Law, R., Metz, J.A.J. (Eds.) *The geometry of ecological interactions: Simplifying spatial complexity*, Cambridge University Press, Cambridge, UK, pp. 359-387, 2000.

- [60] M.J. Keeling, D.A. Rand and A.J. Morris. *Correlation models for childhood epidemics*. Proc. R. Soc. Lond. B **264**, 1149-1156 (1997).
- [61] S.-H. Yook, H. Jeong and A.-L. Barabási. *Modelling the Internet's large-scale topology*. Proc. Natl. Acad. Sci. **99**, 13382-13386 (2002).
- [62] Z.-H. Payman, J.A. Davis and J.D. Meindl. *Prediction of net-length distribution for global interconnects in a heterogeneous system-on-a-chip*. IEEE Trans. on VLSI systems **8**, 649-659 (2000).
- [63] A. Schüz and V. Braitenberg. *The human cortical white matter: Quantitative aspects of cortico-cortical long-range connectivity*. in *Cortical areas: Unity and diversity*, edited by A. Schüz and R. Miller (Taylor & Francis, London and New York, 2002), pp. 377 - 385.
- [64] J. Karbowski. *Optimal wiring principle and plateaus in the degree of separation for cortical neurons*. Phys. Rev. Lett. **86**, 3674-3677 (2001).
- [65] M. Barthélemy. *Crossover from scale-free to spatial networks*. Europhys. Lett. **63**, 915-921 (2003).
- [66] M. Kaiser and C.C. Hilgetag. *Modelling the development of cortical systems networks*. Neurocomp. **58-60**, 297-302 (2004).
- [67] M. Kaiser and C.C. Hilgetag. *Spatial growth of real-world networks*. Phys. Rev. E **69**, 036103 (2004).
- [68] M.T. Gastner and M.E.J. Newman. *The spatial structure of networks*. e-print cond-mat/0407680 (2004).
- [69] N. Mathias and V. Gopal. *Small worlds: How and why*. Phys. Rev. E **63**, 021117 (2001).
- [70] M.E.J. Newman, S.H. Strogatz and D.J. Watts. *Random graphs with arbitrary degree distributions and their applications*. Phys. Rev. E **64**, 026118 (2001).
- [71] M. Molloy and B. Reed. *A critical point for random graphs with a given degree sequence*. Random Struct. Algorithms **6**, 161-179 (1995).
- [72] M. Molloy and B. Reed. *The size of the giant component of a random graph with a given degree sequence*. Combinatorics, Probab. Comput. **7**, 295-305 (1998).
- [73] S. Maslov and K. Sneppen. *Specificity and stability in topology of protein networks*. Science **296**, 910-913 (2002).
- [74] M.E.J. Newman and D.J. Watts. *Renormalization group analysis of the small-world network model*. Phys. Lett. A **263**, 341-346 (1999).

- [75] M. Barthélemy and L.A.N. Amaral. *Small-world networks: Evidence for a crossover picture*. Phys. Rev. Lett. **82**, 3180-3183 (1999); M. Barthélemy and L.A.N. Amaral. *Erratum: Small-world networks: Evidence for a crossover picture*. Phys. Rev. Lett. **82**, 5180 (1999).
- [76] A.-L. Barabási, R. Albert and H. Jeong. *Mean-field theory for random scale-free networks*. Physica A **272**, 173-187 (1999).
- [77] 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).
- [78] P.L. Krapivsky, S. Redner and F. Leyvraz. *Connectivity of growing random networks*. Phys. Rev. Lett. **85**, 4629-4632 (2000).
- [79] A.-L. Barabási. *Linked: The new science of networks - how everything is connected to everything else and what it means for science, business and everyday life* (Hardcover: Perseus, Cambridge, MA, 2002; Paperback: Plume, New York, NY, 2003).
- [80] P.L. Krapivsky and S. Redner. *Organization of growing random networks*. Phys. Rev. E **63**, 066123 (2001).
- [81] B. Bollobás and O. Riordan. *The diameter of a scale-free random graph*. Combinatorica **24**, 5-34 (2004).
- [82] A. Vázquez, A. Flammini, A. Maritan and A. Vespignani. *Modeling of protein interaction networks*. ComPlexUs **1**, 38-44 (2003).
- [83] R.V. Solé, R. Pastor-Satorras, E.D. Smith and T. Kepler. *A model of large-scale proteome evolution*. Adv. Compl. Systems **5**, 43-54 (2002).
- [84] G. Caldarelli, A. Capocci, P. De Los Rios and M.A. Muñoz. *Scale-free networks from varying vertex intrinsic fitness*. Phys. Rev. Lett. **89**, 258702 (2002).
- [85] P.L. Krapivsky and S. Redner. *Finiteness and fluctuations in growing networks*. J. Phys. A **35**, 9517-9534 (2002).
- [86] M. Marsili and Y.-C. Zhang. *Interacting individuals leading to Zipf's law*. Phys. Rev. Lett. **80**, 2741-2744 (1998).
- [87] V.D.P. Servedio, G. Caldarelli and P. Buttà. *Vertex intrinsic fitness: How to produce arbitrary scale-free networks*. Phys. Rev. E **70**, 056126 (2004).
- [88] H. Jeong, Z. Neda and A.-L. Barabási. *Measuring preferential attachment in evolving networks*. Europhys. Lett. **61**, 567-572 (2003).
- [89] A. Clauset and C. Moore. *Why mapping the Internet is hard*. e-print cond-mat/0407339 (2004).

- [90] L. Dall'Asta, I. Alvarez-Hamelin, A. Barrat, A. Vázquez and A. Vespignani. *A statistical approach to the traceroute-like exploration of networks: Theory and simulations*. e-print cond-mat/0406404 (2004); L. Dall'Asta, I. Alvarez-Hamelin, A. Barrat, A. Vázquez and A. Vespignani. *Exploring networks with traceroute-like probes: Theory and simulations*. e-print cs.NI/0412007 (2004).
- [91] A. Clauset and C. Moore. *Accuracy and scaling phenomena in Internet mapping*. Phys. Rev. Lett. **94**, 018701 (2005).
- [92] A. Lakhina, J.W. Byers, M. Crovella and P. Xie. *Sampling biases in IP topology measurements*. Proc. IEEE INFOCOM **1**, 332-341 (2003).
- [93] A. Clauset and C. Moore. *Traceroute sampling makes random graphs appear to have power law degree distributions*. e-print cond-mat/0312674 (2003).
- [94] R. Kikuchi. *A theory of cooperative phenomena*. Phys. Rev. **81**, 988-1003 (1951).
- [95] O. Diekmann and J.A.P. Heesterbeek. *Mathematical epidemiology of infectious diseases* (Wiley, Chichester, 2000).
- [96] M. Boguñá and R. Pastor-Satorras. *Epidemic spreading in correlated complex networks*. Phys. Rev. E **66**, 047104 (2002).
- [97] M. Boguñá, R. Pastor-Satorras and A. Vespignani. *Absence of epidemic threshold in scale-free networks with degree correlations*. Phys. Rev. Lett. **90**, 028701 (2003).
- [98] A.J. Morris. *Representing spatial interactions in simple ecological models*. Ph.D. dissertation, University of Warwick, Coventry, UK (1997).
- [99] D.A. Rand. *Correlation equations and pair approximations for spatial ecologies*. in: McGlade, J. (Ed.) *Advanced ecological theory: Principles and applications*. Blackwell Science, Oxford, UK, pp. 100-142 (1999).
- [100] F. Rao and A. Caffish. *The protein folding network*. J. Mol. Biol. **342**, 299-306 (2004).
- [101] S.B. Laughlin and T.J. Sejnowski. *Communication in neuronal networks*. Science **301**, 1870-1874 (2003).
- [102] D.B. Chklovskii, T. Schikorski and C.F. Stevens. *Wiring optimisation in cortical circuits*. Neuron **34**, 341-347 (2002).
- [103] V.A. Klyachko and C.F. Stevens. *Connectivity optimisation and the positioning of cortical areas*. Proc. Natl. Sci. U.S.A. **100**, 7937-7941 (2003).

- [104] S. Wolfram. *A new kind of science* (Wolfram Media, Champaign IL, 2002), p. 477.
- [105] R. Kasturirangan. *Multiple scales in small-world graphs*. e-print cond-mat/9904055 (1999).
- [106] J.M. Kleinberg. *Navigation in a small world - It is easier to find short chains in some networks than others*. Nature **406**, 845-845 (2000).
- [107] S. Jespersen and A. Blumen. *Small-world networks: Links with long-tailed distributions*. Phys. Rev. E **62**, 6270-6274 (2000).
- [108] M.E.J. Newman and D.J. Watts. *Scaling and percolation in the small-world network model*. Phys. Rev. E **60**, 7332-7342 (1999).
- [109] M. Argollo de Menezes, C.F. Moukarzel and T.J.P. Penna. *Geometric phase-transition on systems with sparse long-range connections*. Physica A **295**, 132-139 (2001).
- [110] P. Sen, K. Banerjee and T. Biswas. *Phase transitions in a network with range-dependent connection probability*. Phys. Rev. E **66**, 037102 (2002).
- [111] C.F. Moukarzel and M. Argollo de Menezes. *Shortest paths on systems with power-law distributed long-range connections*. Phys. Rev. E **65**, 056709 (2002).
- [112] M. Biskup. *On the scaling of the chemical distance in long-range percolation models*. Ann. Probab. **32**, 2938-2977 (2004).
- [113] U. Brandes. *A faster algorithm for betweenness centrality*. J. Math. Soc. **25**, 163 (2001).
- [114] S. Dimitrov, E. Chow and M. Barthélemy. *in preparation* (2005).
- [115] G. Voronoi. *Nouvelles applications des paramètres continus à la théorie des formes quadratiques*. J. Reine Angew. Math. **134**, 198-287 (1908).

Thomas Petermann, *cand. Ph.D.*
Laboratory of Statistical Biophysics, ITP-SB
Ecole Polytechnique Fédérale de Lausanne - EPFL
CH-1015 Lausanne, Switzerland
Tel. : +41 21 693 05 20
E-Mail : Thomas.Petermann@alumni.ethz.ch
URL : <http://marie.epfl.ch/tpeterma>

Born: May 20th 1975
single
Swiss citizen, from Buchrain and Emmen
Military : subofficer within the *National Emergency Operations Centre*, federal council's headquarters

PRESENT POSITION

Research and Teaching Assistant at EPFL's Laboratory of Statistical Biophysics (until Oct. 2003: Institute of Theoretical Physics at the University of Lausanne), leading to the Ph.D. degree (supervised by Prof. P. De Los Rios) May 2002 - present

- Modelling complex networks ranging from the Internet to the brain
- Development of two systematic methods for the description of the spreading dynamics of an epidemic
- Presentation of the results at various international conferences and interdisciplinary schools as well as in international scientific journals (details: see below)
- Supervision of a semester project of a 4th year physics student

EDUCATION

- Physics studies at the Swiss Federal Institute of Technology, Zurich (ETHZ) 1996 - 2001
With exchange year at EPFL (1998/99). Specialisation in computational and statistical physics.
Diploma thesis (at the Institute of Polymer Physics with Prof. H.C. Öttinger): *Selected aspects of the thermodynamics of small systems* (40 pages)
- Grammar school, Reussbühl near Lucerne (maturité type C). 1988 - 1995

EXPERIENCE

Applied Research

- Practical training at ABB Corporate Research Ltd., Baden-Dättwil. Calculation of AC-losses of superconducting slabs and tapes with FEMLAB (report of 11 pages). 2000 (3 months)

Fundamental Research

- Scientific Collaborator in the *Polymer Physics Group* of ETHZ, with Prof. H.C. Öttinger: Completion of Diploma Thesis project. 2001 (1 month)
- Practical training in Nuclear Physics at the University of Manchester, Great Britain (through the International Association for the Exchange of Students for Technical Experience). Participation in the performance of an experiment at CNRS Strasbourg, France; analysis of nuclear reaction data; report of 11 pages. 1999 (2 months)
- Participation at the national contest *Schweizer Jugend forscht (Swiss Young Scientist)* with the project *Berechnung von Planeten- und Satellitenbahnen (Computing the trajectories of planets and satellites)*, predicate: very good, special appreciation: Swiss Astronomical Society. 1994/95

Development

- General practical training at the Diagnostics Division of Roche Molecular Systems Inc., Rotkreuz. Working on a wide range of projects within the development laboratory of the PCR (Genetics) section. 1995 and 1997 (4 months)

Teaching

- Mathematics teacher at the upper grammar school Reussbühl near Lucerne. 2001 (2 weeks)
- Physics exercise lessons for a group of Material Science students, ETHZ. 2000/01 (1 semester)
- Individual lessons in Mathematics, Physics and Chemistry for grammar school students. 1997/98

TALKS

- IBM T.J. Watson Research Center, Group of Biometaphorical Computing and Computational Neuroscience, Yorktown Heights NY, USA, 3/4/2005 (invited).
- Cold Spring Harbor Laboratory, Group of Dmitri Chklovskii (Theoretical neuroscience, principles of brain design), Cold Spring Harbor NY, USA, 3/2/2005 (invited).
- National Institutes of Health (NIH), Laboratory of Systems Neuroscience (NIMH), Bethesda MD, USA, 2/28/2005 (invited).
- European Molecular Biology Laboratory (EMBL), Gene Expression Unit, Heidelberg, Germany, 10/15/2004 (invited).
- *Science of Complex Networks: From Biology to the Internet and WWW*, Aveiro, Portugal, 8/29 – 9/2/2004 (contributed).
- *Journée scientifique des doctorants ITP (Scientific Afternoon of the ITP's Ph.D. students)*, Lausanne, 12/5/2003 (contributed).
- *Living matter: a new challenge to physicists?* (1st transalpine seminar of physics), Les Diablerets, Switzerland, 3/2 – 3/8/2003 (contributed).
- *2nd external seminar of the Institute of Theoretical Physics of the University of Lausanne*, Grimentz, Switzerland, 2/9 – 2/11/2003 (contributed).

POSTERS

- *Lectures on Complex Systems*, Florence, Italy, 10/6 – 10/8/2004.
- *Conference on growing networks and graphs in statistical physics, finance, biology and social systems (COSIN midterm meeting)*, Rome, Italy, 9/1 – 9/5/2003.

WORKSHOPS AND SCHOOLS

- *Conference on Complex Networks: Evolution and Structural Properties (COSIN final meeting)*, Salou (Tarragona), Spain, 3/14 – 3/18/2005.
- *Complex Systems Summer School* organised by the Santa Fe Institute, Santa Fe NM, USA, 6/7 – 7/2/2004. Invited Participant.
- *1st International Workshop on Biologically Inspired Approaches to Advanced Information Technology (BIO ADIT)*, Lausanne, 1/29 – 1/30/2004.
- *Modelling Complex Systems* (7th Granada Seminar on Computational and Statistical Physics), Granada, Spain, 9/2 – 9/7/2002.
- *COSIN Kickoff meeting*, Rome, Italy, 3/11 – 3/14/2002.

POSTGRADUATE COURSES

- *Analysis and modelling using Mathematica*, given by P. Stadelmann, winter term 2002/2003, 14 weeks, 2 hours per week.
- *Quantum field theory: an introduction to the theory of the fundamental interactions*, given by J.-P. Derendinger, summer term 2003, 14 weeks, 4 hours per week.
- *Physics of proteins*, given by Ch. Tang, summer term 2003, twice three hours.
- *Statistical mechanics of complex networks*, given by G. Caldarelli, winter term 2003/2004, 4 weeks, 3 hours per week.

PUBLICATIONS AND SUBMITTED PREPRINTS

- T. Petermann and P. De Los Rios. *Spatial small-world networks: a wiring-cost perspective*, e-print cond-mat/0501429, submitted to *Physical Review Letters* (2005).
- J. Corbo and T. Petermann. *Selfish Peering and Routing*, Proceedings of the Complex Systems Summer School 2004 (organised by the Santa Fe Institute), e-print cs.GT/0410069.
- T. Petermann and P. De Los Rios. *Role of clustering and gridlike ordering in epidemic spreading*, *Physical Review E* **69**, 066116: 1 – 14 (2004), also included in the *Virtual Journal of Biological Physics Research* **7**.
- T. Petermann and P. De Los Rios. *Cluster approximations for epidemic processes: a systematic description of correlations beyond the pair level*, *Journal of Theoretical Biology* **229**, 1 – 11 (2004).
- T. Petermann and P. De Los Rios. *Exploration of scale-free networks: do we measure the real exponents?* *European Physical Journal B* **38**, 201 – 204 (2004).

REFEREEING ACTIVITIES

Physical Review E, Physical Review Letters

LANGUAGES

German: Mother tongue, English: Fluent, French: Fluent, Italian: Fluent

COMPUTER SKILLS

- Languages: C++, Mathematica, Matlab, Turbo Pascal.
- Operating Systems: Dos, Windows 2000, Unix/Linux.
- Software: Mathematica, Maple, Femlab.

HOBBIES AND INTERESTS

- Sports: Rowing (participation in races), Swimming, Beach Volleyball & Cross Country Skiing.
- Music: Playing the Violin (education from 1983 to 1994).
- Others: Cooking, Travelling, Culture in general, Landscape Photography.

Lausanne, 5/10/2005