Universidade de Aveiro Departamento de Física

2008

João Gama Oliveira

Estudo de propriedades dinâmicas de redes complexas

Study of dynamical properties of complex networks

Universidade de Aveiro Departamento de Física
2008

João Gama Oliveira

Estudo de propriedades dinâmicas de redes complexas

Study of dynamical properties of complex networks

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Física, realizada sob a orientação científica do Doutor José Fernando Ferreira Mendes, Professor Catedrático do Departamento de Física da Universidade de Aveiro, e do Doutor Albert-László Barabási, Professor do Departamento de Física da Northeastern University (Boston).

Apoio financeiro da FCT
(SFRH/BD/14168/2003) no âmbito do
III Quadro Comunitário de Apoio.

To my parents

o júri

presidente

Doutor Domingos Moreira Cardoso
Professor Catedrático da Universidade de Aveiro

Doutor José Fernando Ferreira Mendes
Professor Catedrático da Universidade de Aveiro

Doutor Albert-László Barabási
Emil T. Hofman Professor da Universidade de Notre Dame

Doutor Jorge Manuel dos Santos Pacheco
Professor Associado com Agregação da Faculdade de Ciências da Universidade de Lisboa

Doutora Maria Augusta Oliveira Pereira dos Santos
Professora Associada da Faculdade de Ciências da Universidade do Porto

Doutor António Luís Campos Sousa Ferreira
Professor Associado da Universidade de Aveiro

Doutor Marian Boguñá Espinal
Investigador da Facultad de Física da Universidade de Barcelona

Doutor Sergey Dorogovtsev
Investigador Coordenador da Universidade de Aveiro

acknowledgements

This thesis is the result of a PhD program that started in March 2003 at the University of Aveiro with my advisor José Fernando Mendes. From the beginning it was planned that I would spend around six months per year in my co-advisor László Barabási's group at Notre Dame University. I would like to thank both for their support and ideas to follow my PhD program, and for introducing me to many scientists with whom I worked. In particular, Sergey Dorogovtsev in Aveiro and Alexei Vázquez at Notre Dame and Princeton were also great supporters and helped me in many aspects of my PhD.

I would like to thank also the group I worked with in Aveiro: Sasha Goltsev, Sasha Samukhin, Massimo Ostilli, Zhang Peng, Américo Costa, Joana Miguéns, Manuel Barroso, António Luís, Fernão Abreu, Pedro Pombo, Ricardo Dias, Patrícia Silva, Ancai Wu, Xin-Jian Xu, Masayuki Hase, Eshan Khajeh, Farrukh Mukhamedov, Alexander Povolotsky, Bruno Gonçalves and Pavel Krapivsky. I thank the the good moments spent with all of them and the support in some way or another I received from them.

For the same reasons, from the group I worked with at Notre Dame, I would like to thank Suzanne, Alexei, Márcio, Stefan, Eivind, Soon-Hyung, Erzso, Zoli, Gábor, Andi, César, Nick, Marta, Luigi, Sameet, Julian, Pu, Kwang-Il, Juyong, Nicolle and Zoltán Toroczkai.

I'll never forget the help by Janet and László when I was assaulted at Notre Dame (or, better, in South Bend, just outside Notre Dame University), a story that ran through the network of networkers :)

In addition I would like to thank Juan Almendral, Luis López and Miguel Sanjuán for the fruitful periods I spent in Universidad Rey Juan Carlos, Madrid, as well as Roberto Onody and Paulo Alexandre de Castro for the period in São Carlos, University of São Paulo.

From the Department of Physics of Aveiro, I would also like to thank Francisco Reis, Clarisse Soares, Susana Fernandes, Fátima Bola, Regina Silva, Cláudia Santos, Cristina Rei, Emília Fonseca and Fernando Oliveira, for their continued support in many technical issues during the PhD.

Also I thank Nuno, Dina, Sérgio, Elena, Tang, Fu, Emília, João, Mário, António, Andreia and Sónia, with whom I shared office in Aveiro during the first period of the PhD.

Of course I thank all the friends with whom I shared good moments during this time in Aveiro: Vitor, Francesca, Celso, Antónia, Meire, Marcelo, Natália, Fabiane, Mariana, Roger, Manjate, Estevão, Cheo, Cléber, Paulo, Raquel, Patrícia, César, Alexandre, Ricardo, Tang, Massimo, Américo, Peng, Joana, Sara, Lúcia, Nélia, Berta, Pedro, Aneesh; and also in South Bend: Alexei, Márcio, Tatiana, Paulo, Ígor, Hilma, Praveen, César, Andi, Sergio, Tingting, Xiaomao, Lalo, Luigi, Sameet and Julian. Also friends from Porto, where I spent part of the time during the PhD: Arnaldo, Elisabete, Ricardo Cruz, Ricardo Coelho, Paulo, Rui, Ana, André, Patrícia, Susana, Hugo and Miguel.

Last but not least, I thank my close parents to whom this thesis is dedicated.

palavras-chave

Teoria de grafos, redes complexas, física estatística, sistemas dinâmicos, auto-organização crítica, dinâmicas sociais

resumo

Na última década houve grandes desenvolvimentos na área de teoria de grafos e suas aplicações interdisciplinares. Teoria de grafos (ou redes) é um campo de matemática discreta, que, por abstracção dos detalhes de um problema exceptuando a ligação entre os seus elementos, é capaz de uma descrição das suas características estruturais que de outra maneira não seria possível. Muitos sistemas na natureza, e em particular na sociedade, são bem representados por, ou evoluem tendo como base, redes complexas.
Neste trabalho apresentamos alguns avanços para a compreensão das características estruturais genéricas destas redes e sistemas. A tese divide-se em duas partes principais:

Na primeira parte faz-se um estudo da estrutura de redes, começando com uma breve introdução histórica do desenvolvimento da teoria de redes e de conceitos básicos, continuando com um conjunto de exemplos de redes previamente estudadas bem como modelos (Capítulo 1). Seguidamente, apresentamos um estudo teórico de propriedades estruturais como a distância entre vértices e a presença de subgrafos em redes (Capítulo 2). O último capítulo desta primeira parte é dedicado a um estudo detalhado de propriedades estruturais da rede real de colaborações científicas promovida pelo V Programa Quadro da União Europeia, FP5 (Capítulo 3).

Na segunda parte, dividida em três capítulos, processos dinâmicos tendo como base duas redes são investigados: primeiro, a frequência com que os números ocorrem na World-Wide Web (Capítulo 4); segundo, a estatística temporal de actividades humanas, e seus modelos baseados em teoria de filas de espera, que será aqui introduzida (Capítulo 5); e, terceiro, um modelo teórico servindo como base para o estudo de interacções em redes sociais (Capítulo 6).

No Capítulo 7 apresentam-se conclusões gerais, possível trabalho futuro e a lista de publicações resultante do trabalho realizado.

keywords

Graph theory, complex networks, statistical physics, dynamical systems, self-organized criticality, social dynamics

abstract

In the last decade there have been great developments in graph theory, namely in its interdisciplinary applications. Graph (or network) theory is a field of discrete mathematics, which, by abstracting away the details of a problem except the connectivity between its elements, is capable of describing important structural features that would be impossible with all the details retained. Many systems in nature, and in particular in society, are either well represented by, or evolve on the framework of, so called complex networks. Here we present some advances in understanding the generic structural characteristics of these networks and systems. The thesis is divided in two main parts:

In the first part, we present a study of networks' structure, beginning with a brief historical introduction and of basic concepts of network research, continuing with a set of well studied network examples and models (Chapter 1). Next, we present a theoretical investigation of structural properties such as the intervertex distance and the presence of subgraphs in networks (Chapter 2). The last chapter of this first part is devoted to a detailed study of structural properties of the real-world network of scientific collaborations promoted by the European Union's Fifth Framework Programme, FP5 (Chapter 3).

In the second part, divided in three chapters, dynamical processes based on two networks are investigated: First, the frequency with which numbers occur on the World-Wide Web (Chapter 4); second, the statistics of the timing of human activities, and their models based on queueing theory, which will be introduced here (Chapter 5); and third, a theoretical queueing model serving as base for the study of interactions on social networks (Chapter 6).

In Chapter 7 we present general conclusions, outlook future work and the list of publications resulting from the work developed.

Contents

List of Figures

Chapter 1

Introduction

1.1 Brief historical introduction

The study of networks has had a long history in mathematics and the sciences. In 1736 the mathematician Leonard Euler became interested in an enigma called the Königsberg Bridge Problem. The city of Königsberg (today Kaliningrad in Russia) was divided by the river Pregel into four parts as shown in Fig. 1.1. Seven bridges connected the land masses. There was a popular question among the inhabitants of the city: "Is there any single path that crosses all seven bridges exactly once each?". Euler proved the impossibility of such path, making use of a graph representation of the problem (the black dots and lines in Fig. 1.1).

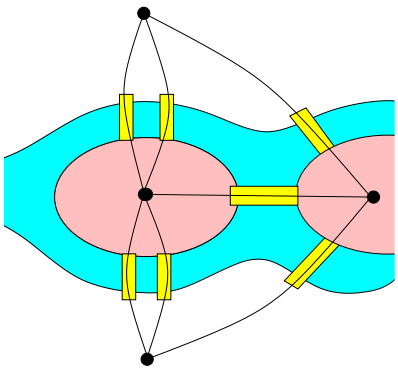

Figure 1.1: Simplified scheme of the 7 bridges (yellow) connecting the land masses in Königsberg. The graph representation consists of the black dots and lines connecting them.

A *graph* (or *network*) is a mathematical object consisting of points, called *vertices* or *nodes*, and lines, called *edges* or *links*, connecting the points — see an example in Fig. 1.2. In this way all the details of the original problem are removed except the connectivity between its elements: There are four vertices representing the four land masses and seven edges joining them (representing the bridges). The bridge problem can then be reformulated in mathematical terms as whether there exists any *Eulerian path* on the graph of the Fig. 1.1, which is precisely a path that traverses each edge exactly once. Euler proved that there is not by observing that, since any such path must both enter and leave every vertex it passes through, except the first and last, there can at most be two vertices with odd degree, where the *degree* of a vertex is the number of edges attached to it. Since all four vertices in the Königsberg graph have odd degree, the bridge problem does not have a solution.

Euler's proof is considered by many to be the first theorem in the field of discrete mathematics known as *Graph Theory*, which has become the main mathematical tool for describing the properties of empirical (real-world) networks. The elements and their connections can be almost anything — people and friendships (Social networks), computers and communication lines (Internet), chemicals and reactions (Biological networks), scientific papers and citations (Information networks), etc. By abstracting away the details of a problem, graph theory is capable of describing important structural features with a clarity that would be impossible with all the details retained. However, despite graph theory is a powerful and general language many authors distinguish it from *Network Theory* (or *Science of Networks*) in three main aspects: (1) by focusing on the properties of real-world networks, network theory is concerned with empirical as well as theoretical questions; (2) it frequently takes the view that networks are not static, but evolve in time according to certain dynamical rules; (3) it aims to understand networks not just as structural objects, but also as the framework on which distributed dynamical systems evolve.

In between the Königsberg bridge problem and the 1990's there were many important developments of graph theory. Of remarkable importance in the 1950's is the work of Solomonoff and Rapoport [1], Gilbert [2], and Erdős and Rényi [3], who began to think of graphs as the medium through which various modes of influence (like information or disease) could propagate. Associated with this trend was the notion that graphs are properly regarded as stochastic objects and therefore that graph properties can be thought in terms of probability distributions. In this way, Solomonoff

and Rapoport first propose a model of a random graph, in the sense that it is composed of a collection of vertices randomly connected by a certain number of edges. A particularly important result obtained was that when the ratio of the number of edges to vertices increases, the graph reaches a point at which it undergoes an abrupt change from a collection of disconnected vertices to a connected state in which the graph contains a *giant component*. More precisely if a graph has N vertices and L edges, the mean degree of a vertex,$\langle k \rangle$, is given by

$$\langle k \rangle = \frac{2L}{N}. \tag{1.1}$$

Then Solomonoff and Rapoport predicted the existence of a phase transition from a fragmented graph (with several small disconnected components) for $\langle k \rangle < 1$ to one dominated by a giant component (whose size tends to infinity as $N \to \infty$) for $\langle k \rangle > 1$. Erdős and Rényi, to whom this result is many times attributed, rediscovered their result independently and gave a major contribution to the development of random graph theory publishing eight papers on random graphs between 1959 and 1968, the most important of which in 1960 [4] dealing with the evolution of some *structural properties* (see Section 1.2) of random graphs as the mean degree is increased. In the mean time, sociologists were starting to apply the ideas of graph theory to social networks, but only in the late 1960's Stanley Milgram, a social psychologist, brought the field into the public consciousness with his famous small-world experiments [5]. In these experiments a target individual and a group of 296 starting volunteers living in the USA were selected, and a document was mailed to each of the starters containing instructions on how to proceed. The participants should try to get the document to the target person by passing it to someone they knew on first name basis and who they believed either would know the target, or might know somebody who did. These acquaintances were then asked to do the same, repeating the process until the document reached the designated target. The number of steps between source and target varied from 2 to 12, with average value 6.2. This small value, when compared to the size of the network, N (in this case the population of the USA), is the origin of the small-world expression. In simple terms, the small-world effect can be understood by realizing that if a person has on average $\langle k \rangle$ acquaintances, then the number of persons contained in a "circle" ℓ steps away from the starting person is approximately $\langle k \rangle^{\ell}$, meaning that to reach the USA population we need only about $\ell = 6$ steps. More precisely we say that a network is a *small world* whenever the average distance between every pair of vertices, $\bar{\ell}$ — where

two nearest neighbor vertices are separated by the unit distance — scales logarithmically with N, *i.e.* $\bar{\ell} \sim \log N$ (see Section 1.2.2). In 1965 Derek Price published an article in the journal Science [6] investigating the network of citations between scientific papers, in which each vertex represents a paper and citations are represented by directed edges from the citing to the cited paper. Price seems to have been the first to observe power-law degree distributions in a network, which are now known to occur in a number of different kinds of networks, often called *scale-free* networks (given that a power-law distribution has no natural scale). A decade later he published another paper [7] proposing a possible mechanism for the observed power laws. Based on previous work by Herbert Simon [8], he proposed that papers that have many citations receive more citations in proportion to the number they already have, and called this process "cumulative advantage", demonstrating that it generates power-law distributions.

In the beginning of the 1980's the mathematician Béla Bollobás proposed the *configuration model* for random graphs with given degree sequence [9] (see Section 1.4.2), which constituted another important development in graph theory, and published a book summarizing the mathematics of random graphs [10]. In 1982, the physicist Rodney Baxter published a book [11] with exact solutions of several statistical physics models, one of which the Ising model on a *Bethe lattice*, a regular[1], deterministic graph (already introduced in 1935 by Hans Bethe [12]) whose properties are close to the configuration model. This may have been the starting point from where the statistical physics community got involved in network theory studies, leading to numerous developments with statistical physics methods being applied to large networks. In 1998 Sidney Redner [13] independently re-obtained Price's power-law degree distribution observations using two large databases of citations of physics papers. In another 1998 article [14], Duncan Watts and Steven Strogatz successfully proposed a model to explain the small-world effect observed earlier by Milgram. In 1999 the cumulative advantage process was rediscovered independently by László Barabási and Réka Albert in what turned out to be the most cited paper of network theory until now [15], introducing the famous BA model and the term *preferential attachment*, concepts later developed by them together with Hawoong Jeong in Ref. [16], and solved by Sergey Dorogovtsev, José Mendes and Alexander Samukhin in Ref. [17], and also independently by Pavel Krapivsky, S. Redner and F. Leyvraz in

[1] In a regular graph all vertices have the same degree.

Ref. [18]. The exact solution was obtained by Bollobás *et. al* in Ref. [19]. The increasing technological capabilities of collecting and processing data, together with these fundamental papers, resulted in a burst of interest in the theory of so called *complex networks*, where the term complex has its origin in the fact that they cannot be modeled by *classical random graphs* (as proposed by Solomonoff and Rapoport, or Erdős and Rényi — Section 1.4.1), in the sense that they are small worlds and have high *clustering coefficient* (*i.e.* high probability that if three vertices are connected by two edges, then the third edge is also present — Section 1.2.3), and/or heterogeneous distribution of degrees of their vertices, often well approximated by power law (Section 1.2.1). Due to the referred increase in data availability, many empirical results (1.3) were obtained for networks like the Internet [20, 21], the World Wide Web [22, 23, 24, 25], e-mail networks [26, 27], social networks [33, 34], biological networks [36, 37, 38, 39, 40]. Also networks more specific to traditional physics have been studied, like networks of free energy minima by Jonathan Doye [28], gradient networks by Zoltán Toroczkai *et. al* [29, 30], or the conformation of polymers by Luís Amaral *et. al* [31], or traditional physics effects on networks, like Bose-Einstein condensation by Ginestra Bianconi and Barabási [32]. A series of review articles by Strogatz [41], Albert and Barabási [42], Dorogovtsev, Alexander Goltsev, and Mendes [43, 44], Mark Newman [45], Tim Evans [46], and Yamir Moreno *et. al* [47], and books by Dorogovtsev and Mendes [48], Pastor-Satorras and Alessandro Vespignani [49], Rick Durret [50], and Guido Caldarelli [51] have been published since then, denoting the rapid evolution the field has been having. General audience books by Watts [52, 53], Bernardo Huberman [54], Barabási [55], and Buchanan [56], among others, were published showing how the subject is interesting to the public in general as well. Also a book consisting of a collection of articles in the field was edited by Newman, Barabási and Watts [57], forming a good summary of its development and state-of-the-art.

Searching for universality both in the structure (or *topology* as frequently termed by physicists) of networks and in the dynamics of their evolution, in addition to uncovering generic properties of real networks, these studies signal the emergence of a new set of modeling tools that considerably enhance our ability to characterize and model complex interactive systems. It is not surprising that physics has been responsible for most of network theory and complex systems studies, since it has been evolving from its traditional areas of research to the study of organization and its emergence in all its forms. It is on this framework that the work presented in this thesis is based.

1.2 Structural properties of networks

We will here introduce and define more precisely some notions of graph theory [58, 59] and structural properties already mentioned in the historical introduction, as well as new ones that will be used in the following chapters, allowing us to characterize and distinguish different kinds of networks. Unlike the graph of Fig. 1.1, which has more than one edge between vertices (and therefore called a *multigraph*) we will consider properties of graphs without multiple edges (see Fig. 1.2). As we will see, however, many of these properties allow us to establish general features and principles of universality between different networks. This may point, in fact, that graph theory, and therefore complex networks, are so general tools that they can possibly be used in almost everything around us, at all scales of the Universe. Despite the structural properties, there are also others, like intrinsic properties of vertices — for example nodes can be colored according to specificities of the case under study. We will not mention the study of these properties here, however in Chapter 3 we will make use of colors to distinguish different characteristics of the real-world network studied there. As already emphasized in the previous Section, we will be mostly interested in the situation of large networks.

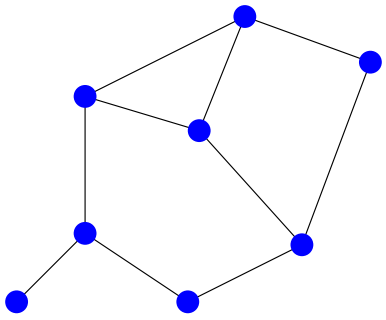

Figure 1.2: Example of a graph without multiple edges.

1.2.1 Adjacency matrix and basic notions

The structure of a graph is completely characterized by a matrix called *adjacency matrix*:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ a_{21} & a_{22} & \dots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \dots & a_{NN} \end{pmatrix} \tag{1.2}$$

where N is the number of vertices in the graph and (for unweighted graphs)

$$a_{ij} = \begin{cases} 1 & \text{if vertex } i \text{ is linked to vertex } j \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{1.3}$$

Every property of the graph can be extracted from its adjacency matrix, since it is fully described by it. For example, the most elementary property of node i, the *degree* k_i, is given by

$$k_i = \sum_{j=1}^{N} a_{ij} \tag{1.4}$$

providing the number of links it has (also called its *connectivity*).

Accordingly the average degree, is given by $\langle k \rangle = \sum_{i,j} a_{ij}/N = 2L/N$ (Eq. 1.1), where L is the number of edges in the graph and $\langle \cdots \rangle$ means average over a particular graph. Yet, the average degree does not probe the degree variations present in the network, which are better characterized by the *degree probability distribution*, P_k, providing the probability that a node has exactly k links. For most networks (called scale-free networks), P_k is a heterogeneous, slowly decaying function, many times well approximated by a power law $P_k \sim k^{-\gamma}$, where γ is the degree exponent, with a cutoff at $k_{cut} \equiv k_{max} \sim N^{1/(\gamma-1)}$ [48, 60]. In scale-free networks the majority of nodes has low degree (of order 1) but a few, called *hubs*, have very high degree (of order of k_{cut}). In most real world networks $N \gg 1$, so that also $k_{cut} \gg 1$, and k can be taken as a real variable, and P_k as a probability density function[2]: $P(k)$. Networks can be *directed*, with links having a specific direction, or *undirected* (when the adjacency matrix is symmetric). The number of in-links of a node in a directed network is its *in-degree*, and the number of links going out is its *out-degree*. Also,

[2]However, note that many times $P(k)$ will be called a probability distribution, as is common in physics.

a network can be *weighted* [61, 62] (with its edges having a specific weight, representing the strength or importance of the link) or *unweighted* (all edges have the same weight, namely 1). In this thesis we consider the simplest undirected, unweighted networks.

1.2.2 Distance measures

Processes taking place along the links of a network, such as package routing on the Internet, traveling via air or contacting a virus from an infected individual are often affected by the length of the paths between two nodes through the network. A *path* between two nodes is defined as a sequence of edges which links them. A graph is *connected* if for any pair of nodes i and j, there is a path from i to j. In unweighted graphs, every edge has weight 1, *i.e.* the distance between two neighboring nodes takes the unit value. In general, there are many paths connecting any two nodes i and j. The number of such paths of length l is given by the (i, j) element of the l-th power of the adjacency matrix (Eq. 1.2).

A useful distance measure is the length of the shortest path, the *geodesic*, ℓ_{ij}, between vertices i and j. The *mean shortest path length*, defined as the average geodesic over all pairs $< ij >$ of vertices,

$$\bar{\ell} = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \ell_{ij} \tag{1.5}$$

is an important structural quantity, characterizing the dimensions of the network. Another distance measure, the *diameter*, gives the maximal separation between a pair of vertices in a network, $\max_{<ij>} \ell_{ij}$.

In Chapters 2 (Section 2.1) and 3 (Section 3.2.2) we will investigate the *k-dependent geodesic*, $\ell(k)$, defined as

$$\ell(k) = \frac{1}{NP(k)} \sum_{\{i:k_i=k\}} \frac{1}{N} \sum_{j=1}^{N} \ell_{ij} \,, \tag{1.6}$$

and giving the average distance of a vertex of degree k to all other vertices.

The distribution of shortest path lengths, $\mathcal{P}(\ell)$ (where we drop the index of ℓ_{ij}), is usually a narrow function, with small average value[3]. This small average value signals a *small-world* network, for which the relative width of the distribution tends to zero as the network size $N \to \infty$ [63, 64, 65].

[3]Note that Eq. 1.5 is equivalent to $\bar{\ell} = \sum_\ell \ell \mathcal{P}(\ell) = \sum_k \ell(k) P(k)$.

Thus, for a large network, almost all pairs of vertices are at distance $\bar{\ell}$, from where, following the reasoning already introduced in Section 1.1, $N \sim \langle k \rangle^{\bar{\ell}}$, and

$$\bar{\ell} \sim \frac{\log N}{\log \langle k \rangle} \,. \tag{1.7}$$

This formula[4] means that $\bar{\ell}$ grows slower than any power of N, so that making an analogy with D-dimensional lattices[5], for which $\bar{\ell} \sim N^{1/D}$, small-world networks are many times said to be infinite-dimensional objects: The number of neighbors a node can have increases with system size.

1. Assign vertex j distance zero, to indicate that it is zero steps away from itself, and set $d \leftarrow 0$.
2. For each vertex l whose assigned distance is d, follow each attached edge to the vertex m at its other end and, if m has not already been assigned a distance, assign it distance $d+1$. Declare l to be a predecessor of m.
3. If m has already been assigned distance $d+1$, then there is no need to do this again, but l is still declared a predecessor of m.
4. Set $d \leftarrow d+1$.
5. Repeat from step 2 until there are no unassigned vertices left.

Now the shortest path (if there is one!) from i to j is the path we get by stepping from i to its predecessor, and then to the predecessor of each successive vertex until j is reached. If a vertex has two or more predecessors, then there are two or more shortest paths, each of which must be followed separately if we wish to know all shortest paths from i to j. In unweighted graphs ℓ_{ij} is the number of predecessors in each shortest path.

Besides the computation of ℓ_{ij}, this algorithm can be applied, with slight modifications, to the computation of betweenness centrality (see Sections 1.2.6 and 3.2.3).

1.2.3 Clustering coefficient

This property measures the extent to which the neighbors of a particular node are connected to each other. Formally, the *local clustering coefficient* [14, 68, 69], C_i, of node i is defined as (see Fig. 1.2.3)

[4]For a more precise derivation see Section 2.1.3

[5]In a lattice all vertices have the same degree, and are arranged in a specified ordered (deterministic) manner.

$$C_i = \frac{2n_i}{k_i(k_i-1)}\,, \tag{1.8}$$

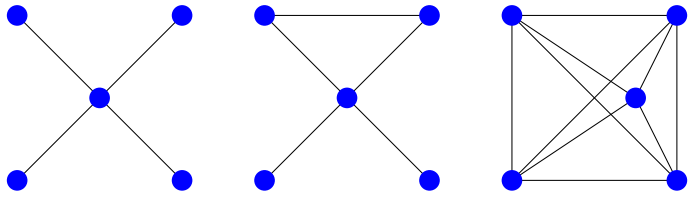

Figure 1.3: Illustration of the definition of local clustering coefficient for three graphs: From left to right C_i for the node in the center is respectively 0, 1/6 and 1.

where n_i denotes the number of links connecting the k_i neighbors of node i to each other[6]. Accordingly, we can define the average clustering coefficient of a network as

$$\langle C\rangle = \frac{1}{N}\sum_{\{i:k_i>1\}} C_i\,, \tag{1.9}$$

where the sum runs over all vertices of degree larger than one. A useful measure is also given by the k-dependent clustering coefficient [70]:

$$C(k) = \frac{1}{NP(k)}\sum_{\{i:k_i=k\}} C_i\,, \tag{1.10}$$

finding that for certain models of scale-free networks $C(k)\sim k^{-1}$ [70, 71], a result corroborated for some empirical networks as well [39]. More generally, in complex networks, usually $C(k)\sim k^{-\alpha}$.

To avoid confusion, it should be noted that another measure of clustering was already in use in the sociology literature before the one defined in Ref. [14], Eq. 1.8, namely the *transitivity* or simply *clustering*. Contrary to the average clustering coefficient (Eq. 1.9), instead of being given by the mean of the ratios, the transitivity T is given by the ratio of the means:

$$T = \frac{\langle 2n_i\rangle}{\langle k_i(k_i-1)\rangle}. \tag{1.11}$$

[6]It is therefore defined only for vertices of degree $k_i > 1$, and denotes the probability that the neighbors of node i are themselves connected.

In the mathematical and physical literature it seems to have been first discussed by Barrat and Weigt [72]. However, there are significant differences in its algorithmic computation time [73], and usually the average clustering coefficient (Eq. 1.9) is used.

As we will see in Chapters 2 and 3, real-world networks are usually highly clustered, when compared to a random graph with the same average degree and number of vertices (see Section 1.4 for the definition of random graph).

1.2.4 Correlations

Correlations in networks may be present in a number of different manners. For example, the clustering coefficient measures a kind of 3-node correlations [68]. Here we will introduce correlations of degrees of nearest neighbor vertices, which describe organizational properties [74, 75, 76] that the degree distribution does not address: Given a degree sequence of all the nodes, do high-degree vertices in a network preferentially associate with other high-degree vertices, or are they mainly connected to low-degree ones? This question has different types of answers depending on the level of detail one wishes to use to address it. The degree correlation coefficient [77] is a number between -1 and 1, representing the Pearson correlation coefficient of the degrees at either ends of an edge (see Section 3.2.5 of Chapter 3). Networks in which hubs are preferentially connected to other hubs are called *assortative*, and have a positive degree correlation coefficient. Social networks tend to be assortative, while most of the networks in biology or communication tend to be *disassortatively* mixed: hubs in these networks preferentially connect to smaller nodes [77, 78].

More detailed representations of degree correlations are given by the mean degree of the nearest neighbors (nn) of a vertex as a function of its degree [21] given by

$$\langle k\rangle_{nn}(k) = \frac{1}{kNP(k)} \sum_{\{i:k_i=k\}} \sum_{j=1}^{k} k_{nn,j}\,, \tag{1.12}$$

where $k_{nn,j}$ is the degree of the j-th nearest neighbor of vertex i and the first sum runs over all vertices of degree k. Also to measure correlations two-dimensional histograms of the degrees of the vertices at the ends of an edge, *i.e.* the joint degree-degree distribution $P(k,k')$. For uncorrelated networks $\langle k\rangle_{nn}(k)$ is independent of k and $P(k,k')$ factorizes to the product of the degree distributions $P(k)\,P(k')$.

We will use these measures in the study of a real network in Chapter 3, Section 3.2.5.

1.2.5 Subgraphs (cycles, trees)

Subgraphs are subsets of connected vertices in a graph, and provide important information about the structure of many real networks (Section 1.3) [79]. For example, in cellular regulatory networks feed-forward loops[7] play a key role in processing regulatory information [80, 81], while in protein interaction networks highly connected subgraphs represent evolutionary conserved groups of proteins [82]. In a similar way, *cycles*, a special class of subgraphs, offer evidence for autonomous behavior in ecosystems [83], cyclical exchanges give stability to social structures [84], and cycles contribute to reader orientation in hypertext [85]. Finally, understanding the nature and frequency of cycles is important for uncovering the equilibrium properties of various network models [86]. Another class of subgraphs are *trees*, *i.e.* subgraphs without cycles (see Fig. 2.1b-e in Chapter 2). Trees are important because many times it is possible to assume that a graph is *tree-like* (*i.e.* has very few loops), an approximation that greatly simplifies calculations.

1.2.6 Centrality measures

Centrality measures allow us to probe the influence of a vertex in the network as a whole. The simplest centrality measure of a vertex is its degree, giving us the number of connections to other vertices. A more significant centrality measure is the *betweenness centrality* [87], which measures the extent to which a vertex m lies on the paths between other vertices. It is defined as

$$\sigma_m = \frac{1}{(N-1)(N-2)} \sum_{\{i,j:i\neq j\neq m\}} \frac{B_{(i,m,j)}}{B_{(i,j)}}, \tag{1.13}$$

where $B_{(i,j)}$ is the number of shortest paths between nodes i and j, $B_{(i,m,j)}$ is the number of such shortest paths passing through vertex m, and the sum is taken over all pairs of vertices i and j which do not include m. Here we introduce the pre-factor $1/[(N-1)(N-2)]$ (where N is the total number of vertices) in order to account for normalization, so that $0 \leq \sigma_m \leq 1$, useful for the calculations of Section 3.2.3.

[7] Cycles (or loops) are sequences of distinct connected vertices, except the first and last which are the same.

The *centrality index* of a vertex v (used in Section 2.1.5) is defined as [88]

$$c_v = \frac{(N-1)}{\sum_u \ell_{vu}}, \tag{1.14}$$

where ℓ_{vu} is the length of the shortest path between vertices u and v, and the sum is over all vertices u of the graph for which there is a path to v.

1.3 Networks in the real world

Networks are ubiquitous in our world[8], and many times it is not easy to classify them into a single category, since they are themselves interconnected. This is one of the reasons why network theory is an interdisciplinary research field, and maybe also one of the reasons why most networks share the same generic properties. For example, the World Wide Web can be seen as a communication network, or information network, or technological network, or even as a social network (as illustrated, for example, by personal blogs, or social networking websites). The increasing availability of electronic databases has already established a wide list of complex networks, serving empirical network studies, which are to be modeled by theoretical research. Here we list and classify a few examples of networks which can be found around us. Many authors separate the networks in more classes (distinguishing between communication or information networks, for example) or in a different manner. However, in many situations this is merely conventional due to the reason referred above. Having this in mind, we thus list in this section examples of different networks in three main classes: Social Networks; Communication, Information and Technological Networks; and Biological Networks.

1.3.1 Social Networks

- Friendship Networks

  Friendship networks have been studied for a long time in Sociology [89, 90, 91], given their relevance to understand many social phenomena. Friendships are represented by links between vertices representing people. These studies usually collect more information about each individual (rather than just his or her connections), allowing, for example, to see how society

[8] Indeed, every system involving interactions between its elements has an underlying network, vertices representing the elements and edges representing interactions.

organizes itself according to gender, race, personal interests, etc. The data collection for these networks has been usually made by recurring to questionnaires, which limits the studies to small nets. However, recently, a new generation of websites[9] permitting users to create and share content (as well as having fun!) by establishing connections between them [92], may potentially be a way to turn much easier the analysis of these networks by applying automatic methods, provided the privacy of each user is kept safe.

- Scientific Collaboration Networks

  Collaboration networks are represented by graphs whose vertices represent (for example) scientists who worked together, coauthoring at least one publication represented by one edge [93, 94]. On a coarser level a collaboration network can represent collaborations between scientific institutions, such as universities and/or industry related entities [95]. In this thesis we study two collaboration nets, one of each type as mentioned: The coauthorship network of mathematical publications in Section 2.2 of Chapter 2 [96, 97] and, in Chapter 3, the network of collaborations arising from the Fifth Framework Programme, an initiative which sets out the priorities for the European Union's research and technological development, promoting collaborations between scientific institutions and industry related entities [98, 99].

- Movie Actor Networks

  The Internet Movie Database[10] is the source of one of the largest social networks open to study. Based on all movies since the 1880's, the network has over 400,000 actors as its nodes and movies that represent the links between any of them [14, 100]. The degree distribution of the actor network has a power-law tail [101], and its clustering coefficient is much larger than that of a random network of similar size.

- The Network of Human Sexual Contacts

  Sexually transmitted diseases like AIDS spread on the subset of the social network described by sexual relationships. Although precise data about the links of this network is quite hard to collect, a few investigations have given us insights about its topology. Liljeros et al. [102]

[9] For example LinkedIn (`www.linkedin.com`) or Facebook (`www.facebook.com`).

[10] URL: `www.imdb.com`

have estimated the degree distribution of the sex web using a survey about the number of sexual partners of 2810 Swedish individuals. Their investigation shows that the distribution of the number of sexual contacts of both men and women follow power laws. This finding has a strong impact on epidemiological studies aimed at eradicating diseases spreading on sexual contact networks, as scale free networks with degree exponents under 3 were found to allow diseases with arbitrarily low virulence to stay endemic and to show no improvement upon random immunization of their nodes [26, 103].

1.3.2 Communication, Information and Technological Networks

- The Internet

  The Internet, a network of physical cables between computers, routers and other telecommunication devices, is one of the favorite models of network studies [48, 49]. Its structure, defined at two different levels of detail, is continuously mapped, and the huge number of nodes and links provide good statistical grounds for the measurement of many network features. At the most basic level the vertices are routers, while edges are the physical connections between them. The Autonomous System (AS) level is a coarse-grained view of the Internet, where each autonomous Internet domain (defined by local data routing, such as the whole network domain of the University of Aveiro) is represented by a single vertex. Maps at both levels have been publicly available since 1999 [20, 104, 105, 106, 107], when Faloutsos *et al.* [20] measured the degree distribution at both levels and concluded that both follow power laws. Further studies of these networks showed that they also display small world behavior ($\bar{\ell}$ around 9 for the router, 3 for the AS level, Internet), along with high clustering coefficients (see Sections 1.2.2 and 1.2.3) [21, 108].

- The World-Wide Web (WWW)

  The World-Wide Web (WWW) [48, 54], often incorrectly referred to as the "Internet", is a huge network of Web pages linked by directed URL hyperlinks [25, 109, 110]. It is the largest available network[11], with a number of web pages on the order of 10^{10}, yet it is also very typical

[11] For a daily estimation of its size see `http://www.worldwidewebsize.com/`

in many of its properties: high clustering and small world behavior with an average path length estimated to be around 16 [23, 25, 111] (meaning that *on average* 16 clicks are enough to go from one Web page to another). Moreover, both distributions of the ingoing and the outgoing links are power laws with scaling over more than five orders of magnitude [23, 24, 112, 113, 111]. In a coarse-grained network representation of the World-Wide Web, each web domain (or website[12]) like the whole `www.ua.pt` page system is represented as a node, while any hyperlink from a document in this domain to another domain defines an edge between them. This bird-eye view of the WWW also gives us a scale free network, and an even smaller cyber-world: the average path length of this graph is 3.1 [111].

In Chapter 4 we look at the Web from a different perspective. Using it as a database, we study the frequency of numerals in its documents, finding that it is much richer and complex than would be predicted by classical Benford's law [114] which states a logarithmic decay for the frequency of the first digit in numbers occurring in databases.

- E-mail networks

  The strucuture of e-mail networks, with electronic addresses as nodes and e-mails as the links, has been investigated based on data stored in server log files [27, 115, 116]. The importance of this communication network comes from its ability to spread viruses [117], a process similar to natural virus spreading along social interactions [26]. Thus, the finding that e-mail networks have scale free degree distribution explains the surprising prevalence of old viruses[13], in spite of easy-access anti-virus software [103, 118, 119].

- Articles citation networks

  Citation networks reflect the way research articles of different scientific areas build on previous knowledge. They can be constructed using online databases of scientific papers; links of these networks are the references between them [6, 13, 120]. These references are directed links, and studies of their topology indicate that the in-degree distribution of these networks follow power laws [6, 13], while the out-degree distribution has a well-defined maximum and an

[12] Not to be confused with *Web page*: a website is formed by a set of web pages.

[13] For a list of known viruses see for example `http://www.wildlist.org/`

exponential tail [121]. These citation networks are particular cases of *citation graphs* for which new connections emerge only between a new vertex and already existing ones.

- Power grids

  Power grids are networks of generators, transformers and substations linked by high-voltage transmission lines spanning a whole country or region, distributing electric current. Statistical studies on the power grid covering western states in the USA indicate that they are small world networks with relatively high average clustering coefficient and an exponential degree distribution [14, 52, 101]. Recent interest in vulnerabilities of the power grid has been triggered by extensive electricity blackouts which affected large regions of the eastern United States [122, 123, 124].

- Telephone network

  Defined as a network whose vertices represent telephone numbers and the directed edges calls from one number to another, the phone-call (directed) network connecting people who had long-distance conversation via AT&T (in the course of one day in the USA), was mapped out by Aiello et al. [125, 126] and was found to have a power law degree distribution both for incoming and outgoing calls. Mobile phone calls network has also been analyzed as a weighted, undirected network [127], again have heterogeneous degree distribution, as well as non-trivial clustering and correlations.

- Language networks

  Words in a human language can be linked in several ways. Defined as graphs of words linked if they appear no more than two words apart with a frequency higher than a chosen threshold, co-occurrence networks based on the British National Corpus[14] were found to have a degree distribution with two distinct regimes of power law scaling [128]. Word co-occurrence networks hint at methods used by people to organize concepts while choosing them for communication [128, 129, 130, 131, 132]. Perhaps not surprisingly, this abstraction of human language into a network also has a degree distribution with power law tail, along with a very

[14] URL: `http://info.ox.ac.uk/bnc/`

high clustering coefficient. In Section 2.2 we use a semantic network of English synonyms database to corroborate the theoretically obtained results of that section.

1.3.3 Biological Networks

- Metabolic Pathways, Protein Interaction Networks, Genetic Regulatory Networks

  Many biological systems can be usefully represented as networks. Perhaps the classic example of a biological network is the network of metabolic pathways, which is a representation of metabolic substrates and products with directed edges joining them if a known metabolic chemical reaction exists that acts on a given substrate and produces a given product. Molecular biologists study huge maps of metabolic pathways. Studies of the statistical properties of metabolic networks can be found, for example, in Refs. [36, 37, 39, 40, 133, 134].

  A separate network is the network of mechanistic physical interactions between proteins[15], which is usually referred to as a protein interaction network (or 'interactome'). These networks have been studied by a number of authors [38, 82, 136, 137, 138].

  Another important class of biological network is the genetic regulatory network. The expression of a gene, *i.e.*, the production by transcription and translation of the protein for which the gene codes, can be controlled by the presence of other proteins, both activators and inhibitors, so that the genome itself forms a switching network with vertices representing the proteins and directed edges representing dependence of protein production on the proteins at other vertices [139, 140, 141].

- Neural networks

  The worm *C. elegans* is the only organism with a completely mapped neural network. It has 282 neurons and close to 2000 connections (synapses or gap junctions) [142]. This small but dense network has an exponential degree distribution and quite high clustering coefficient [14, 101]. Functional magnetic resonance imaging techniques can be used to measure the activity of regions of the human brain. Correlations between these regions can define a functional network

[15] Not to be confused with the protein folding process, whose network representation has also been recently given in Refs. [31, 135]

of brain sites connected by common patterns of activity. These networks are dynamic, and the details of their structure is interesting for functional studies of the brain. Nonetheless, their large-scale organization is scale free, with high clustering coefficients [143].

- Food webs and networks of ecosystems

  Food webs are networks of species linked by predator-prey interactions. These networks have been mapped out in a few habitats by ecologists who use them to investigate interactions between different species [144]. A few independent studies on food webs of different sizes have shown that they are highly clustered, and the average path length between species is below 3 [145, 146, 147]. The nature of their degree distribution is unclear, mostly due to the small size of these systems.

- Disease networks

  Disease networks, where diseases and genes are linked in a bipartite [16] network if the disease is caused by mutations in the gene, have been studied [148, 149, 150] revealing the existence of distinct disease-specific functional modules associated with characteristic genes and therefore with proteins, which may allow the production of more specific drugs for diseases [151].

1.4 Network models

In the past few years a series of network models have been developed to explain nontrivial generic properties of real-world networks, such as the small world property, scale free degree distribution or high clustering. In this section we review the most influential models.

1.4.1 Classical random graph

The simplest random networks are so-called classical random graphs (CRG's) [1, 2, 3, 4]. In simple terms, these are maximally random networks under the constraint that the mean degree of their vertices, $\langle k \rangle$, is fixed. The number of vertices N is also fixed in these uncorrelated graphs. There

[16] A bipartite network is one formed by two distinct classes of nodes with links existing only between nodes of distinct classes.

are two main versions of CRG's: The Erdős and Rényi model [4] is a statistical ensemble of all possible graphs of precisely N vertices and precisely L edges, where each member of the ensemble has equal probability of realization; on the other hand, in the Gilbert model [2], each pair of N vertices is connected with some probability p. This produces a statistical ensemble of all possible graphs of N vertices. The members of this ensemble are weighted with some statistical weights. In the thermodynamic limit (infinitely large networks), these two versions are equivalent, and $\langle k\rangle = 2L/N = p(N-1)$. The degree distribution of a CRG has a Poisson form:

$$P(k) = \frac{\mathrm{e}^{-\langle k\rangle}\langle k\rangle^k}{k!}. \tag{1.15}$$

Here $\langle k\rangle$ is fixed as $N \to \infty$. This is an extremely rapidly decreasing distribution (faster than exponential) after the peak close to its natural scale $\langle k\rangle$. All moments converge.

Given that all pairs of vertices are connected with the same probability p, the clustering coefficient of a CRG is $\langle C\rangle = p$.

The limit with fixed $\langle k\rangle$ as $N \to \infty$ (*i.e.* $p \to 0$ when $N \to \infty$) corresponds to a *sparse graph* for which the mean number of connections of a vertex is much less than the number of connections of a vertex in a fully connected graph (also called *complete graph*). This limit is the most interesting given that it is when the network's *giant connected component* — a subgraph of mutually reachable vertices whose size is a non-vanishing fraction of N (when $N \to \infty$) — is formed. Otherwise the network is only a set of separated trees (Section 1.2.5). It turns out that in CRG's, the giant connected component exists if the mean number of connections of a vertex exceeds one, $\langle k\rangle > 1$. At $\langle k\rangle = 1 \ll N$ there is a phase transition where the giant connected component is born[17].

The giant connected component is also typically present in complex networks, whose main difference to CRG's is the presence of high clustering coefficient (in contrast to the vanishing $\langle C\rangle = p$) and the broad, slowly decaying, degree distribution (in contrast to Eq. 1.15). Common to both complex networks and CRG's is the fact that both show the small-world effect (Section 1.2.2).

[17] This phase transition is the equivalent to the one observed in percolation theory in the infinite-dimensional limit [42, 44].

1.4.2 Configuration model

The *configuration model* (introduced by Bollobás [9]) is the first natural generalization of classical random graphs. In very simple terms, the configuration model is a maximally random graph with a given degree distribution $P(k)$. This complex random equilibrium network (actually an ensemble of networks) is uncorrelated. The configuration model produces tree-like graphs.

More precisely, the configuration model generalizes the classical random graph to a graph with generic degree distribution by drawing a degree sequence k_i $(i = 1, ..., N)$ from the desired distribution $P(k)$. A well known algorithm to generate a graph according to this model was given by Molloy and Reed [152, 153].

1.4.3 Small-world network model

Watts and Strogatz proposed a specific class of complex networks [14], which display the small-world effect, and named them small-world networks[18]. These are lattices with high clustering (e.g., a trigonal lattice), where randomly chosen vertices are connected by long-range shortcuts. Actually, a small-world network is a superposition of a lattice and a classical random graph. Due to the strong clustering of the lattice, a small-world network has high clustering. Due to the compactness of the classical random graph, a small-world network is compact.

1.4.4 Preferential attachment model

The most popular self-organization mechanism of networks is preferential attachment (or preferential linking): vertices of high degree attract new connections with higher probability. More precisely, the probability that a new edge becomes attached to a vertex with k connections is proportional to a 'preference' function of k, $f(k)$ [17]. The resulting structure of the growing net is determined by the form of this function.

Scale-free degree distributions may emerge only if the preference function is linear, that is

$$f(k) = \frac{(k + A)}{(\langle k \rangle + A)}\,, \tag{1.16}$$

[18] Not to be confused with a small world, which is a network displaying the small-world effect (see Section 1.2.2).

where A is a constant. This seems to be a widespread situation in real networks. This form of preference produces γ exponents between 2 and ∞.

Models of evolving systems based on this concept were proposed by Yule [154] and Simon [8]. To growing networks, this idea was applied by Price [7] — a linear preference function — and by Barabási and Albert [15] — a proportional preference function. Specifically, in the BA model the probability that a new vertex i becomes attached to a vertex j already in the network is given by

$$\Pi_{i\to j} = \frac{k_j}{\sum_l k_l}\,, \tag{1.17}$$

and corresponds to the case when $A = 0$ in the preference function Eq. 1.16. In the BA model, the growing network is a citation graph: At each time step, a new vertex is added to the network and becomes attached to m vertices, according to Eq. 1.17. A simple and fast algorithm to generate a BA network consists of the following steps:

1. Start with m_0 completely connected vertices.

2. Initialize a linear array where each vertex i of the network is present k_i times. [At this step $k_i \equiv m_0 - 1, \forall i$, and the array size is $m_0 \times (m_0 - 1)$].

3. At each step add a vertex to the network, and randomly choose m elements of the array of the previous step, to which the new vertex will connect. (To avoid multiple connections, if the same vertex is chosen more than once, then choose another random element until there is no repetition.)

4. Update the array by adding to it m new entries corresponding to the new vertex, and another m entries each corresponding to each selected vertex in the previous step.

5. Repeat from step 3 until the desired network size N is reached.

In Chapter 2 we will use a more general algorithm in order to generate networks according to $f(k)$ in Eq. 1.16 with $A > 0$. The difference from the previous algorithm is that with probability $m_0/(m_0 + A)$ the connection is chosen preferentially (according to Step 3), otherwise, with complementary probability, the connection is chosen randomly between the existing vertices.

1.4.5 Deterministic models

In Chapter 2 (see Figure 2.1) we use a set of deterministic growing graphs [70, 155] to study structural properties of networks. These graphs are built by using a set of rules up to a certain number of vertices. They correctly reproduce many features of real networks, allowing exact analytic calculations, and can be used as tools to recursively guess new ones.

Chapter 2

Structure of complex networks

The first step toward a complete characterization of complex networks consists in a reliable description of their structural properties, which play a relevant role in the functionality of real networks as well as in the dynamical patterns of processes taking place on them. In this chapter we present a theoretical study of two of the fundamental properties discussed in Section 1.2 of Chapter 1. In Section 2.1, we study the average shortest path length (Eq. 1.5) as a function of degree, $\ell(k)$ for several types of networks. In Section 2.2 we investigate the abundance of subgraphs and cycles in networks with both well defined degree distribution, $P(k)$, and k-dependent clustering coefficient, $C(k)$ (Eq. 1.10).

2.1 k-dependent geodesic in complex networks, $\ell(k)$

In this section we study the mean length $\ell(k)$ of the shortest paths between a vertex of degree k and the rest of the vertices (see Section 1.2.2) in growing networks, where correlations are non-negligible. In a number of deterministic scale-free networks we observe a power-law correction to a logarithmic dependence, $\ell(k) = A\ln[N/k^{(\gamma-1)/2}] - Bk^{\gamma-1}/N + \ldots$ in a wide range of network sizes. Here N is the number of vertices in the network, γ is the degree distribution exponent, and the coefficients A and B depend on a network. We compare this law with a corresponding $\ell(k)$ dependence obtained for random scale-free networks growing through the preferential attachment mechanism. In stochastic and deterministic growing trees with an exponential degree distribution,

we observe a linear dependence on degree, $\ell(k) \cong A \ln N - Bk$. We compare our findings for growing networks with those for uncorrelated graphs.

2.1.1 Introduction

The mean intervertex distances in networks were extensively studied both in the framework of empirical research [23] and analytically [100, 65, 156, 157]. The typical size dependence of the mean intervertex separation is logarithmic, $\bar{\ell}(N) \propto \ln N$. However, the mean intervertex distance is an integrated, coarse characteristic. One may be interested in a more delicate issue—the position of an individual vertex in a network. Recently Holyst *et al.* [158], have considered the question: how far are vertices of specific degrees from each other? They have shown that in uncorrelated networks, the mean length of the shortest path between vertices of degrees k and k' is $\ell(k,k') \cong D + A \ln N - A \ln(kk')$, where D is independent of N, k, and k', and the coefficient A depends only of the mean branching ratio of the network. Note the coincidence of the coefficients of $\ln N$ and $\ln(k,k')$ in this result. The authors of Ref. [158], also calculated $\ell(k,k')$ of networks with nonzero clustering though without degree-degree correlations. In this case, they have arrived at the same expression as above but with coefficients of $\ln N$ and $\ln(k,k')$ additionally depending on the clustering. Here we present our observations for another (though related) characteristic—the mean length of the shortest paths from a vertex of a given degree k to the remaining vertices of the network, $\ell(k)$. This quantity is related to $\ell(k,k')$ in the following way:

$$\ell(k) = \sum_{k'} P(k')\ell(k,k')\,, \tag{2.1}$$

and so

$$\bar{\ell} = \sum_{k} P(k)\ell(k) = \sum_{k,k'} P(k)P(k')\ell(k,k')\,. \tag{2.2}$$

In simple terms, we reveal the smallness of a network from the point of view of its vertex of a given degree. Our objects of interest are growing (and so inevitably correlated) networks.

In Section 2.1.2 we list our main observations, so that readers not interested in details may restrict themselves to this section. Section 2.1.3 contains the discussion of the $\ell(k)$ dependence in uncorrelated networks for the sake of comparison. In Section 2.1.4 we explain in detail how the results were obtained and describe particular cases. In Section 2.1.5 we make a few remarks on the

degree-dependent intervertex separation in various networks and discuss relations of this quantity to centrality measures (see Section 1.2.6) used in sociology [88, 160].

2.1.2 Main observations

For the purpose of the analytical description of $\ell(k)$ we use simple deterministic graphs. Deterministic small worlds were considered in a number of recent papers [70, 155, 161, 162, 163, 164, 165, 166, 167, 168, 169] and have turned out to be a useful tool. (We called these networks *pseudofractals*. Indeed, at first sight, they look as fractals. However, they are infinite dimensional objects, so that they are not fractals.) These graphs correctly reproduce practically all known network characteristics. We use a set of deterministic scale-free models with various values of the degree distribution exponent γ, $P(k) \propto k^{-\gamma}$ (see Fig. 2.1). We consider deterministic graphs with γ in the range between 2 and ∞, where a graph with $\gamma = \infty$ has an exponentially decreasing (discrete) spectrum of degrees.

In the studied scale-free deterministic graphs, in a wide range of the graph sizes, the mean separation of a vertex of degree k from the remaining vertices of the network is found to follow the dependence:

$$\ell(k) = A \ln\left[\frac{N}{k^{(\gamma-1)/2}}\right] - B\,\frac{k^{\gamma-1}}{N} + \dots . \tag{2.3}$$

The constants A and B (as well as the sign of B) depend on a particular network.

In stochastic growing scale-free networks, we observe a dependence $\ell(k, N)$ shown in Figure 2.2. This figure demonstrates the results of the simulations of networks growing by the preferential attachment mechanism with a linear preference function [17]. While the dependence on $\ln N$ is linear practically in the entire range of observation, $\ell(k)$ vs. $\ln k$ is of a more complex form (see Fig. 2.2). The derivative $d\ell(k)/d\ln k$ is non-zero at $k = 1$ and at large degrees, $\ell(k)$ is fitted by a linear function of $\ln k$ with a larger slope. One should note that in all growing networks considered in this section, new connections cannot emerge between already existing vertices. These networks are often called "citation graphs".

In the specific point $\gamma = 3$, correlations between the degrees of the nearest neighbors in these graphs are anomalously low. In this situation, the main contribution to $\ell(k)$ reduces to $\ell(k) \propto \ln(N/k)$, which coincides with the result for equilibrium uncorrelated networks (see the next section).

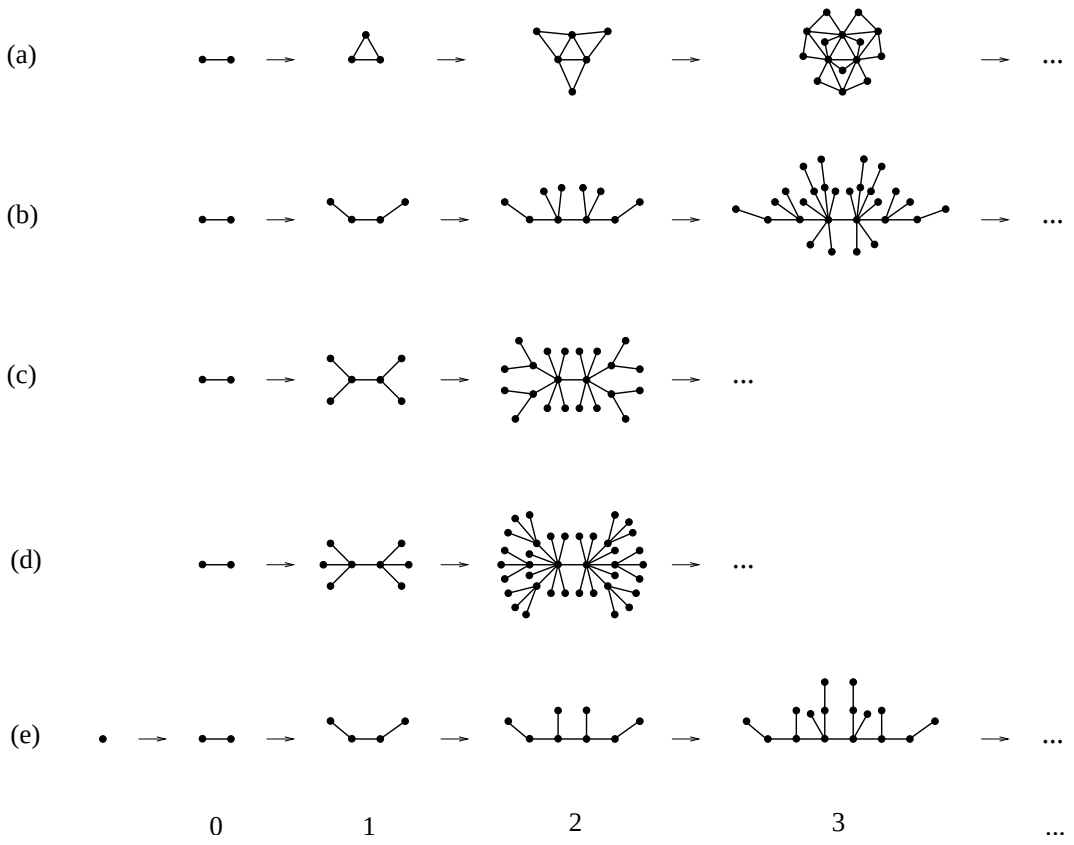

Figure 2.1: The set of deterministic graphs that is used in this section. (a) A scale-free graph with the exponent of the degree distribution $\gamma = 1 + \ln 3/\ln 2 = 2.585\ldots$ [43, 70]. At each step, each edge of the graph transforms into a triangle. (b) A scale-free tree graph with $\gamma = 1 + \ln 3/\ln 2 = 2.585\ldots$ [161]. At each step, a pair of new vertices is attached to the ends of each edge of the graph. (c) A scale-free tree graph with $\gamma = 3$. At each step, a pair of new vertices is attached to the ends of each edge plus a new vertex is attached to each vertex of the graph. (d) A scale-free tree graph with $\gamma = 1 + \ln 5/\ln 2 = 3.322\ldots$. At each step, a pair of new vertices is attached to the ends of each edge plus two new vertices are attached to each vertex of the graph. (e) A deterministic tree graph with an exponentially decreasing spectrum of degrees [161]. At each step, a new vertex is attached to each vertex of the graph. In all these graphs, a mean intervertex distance grows with the number N of vertices as $\ln N$.

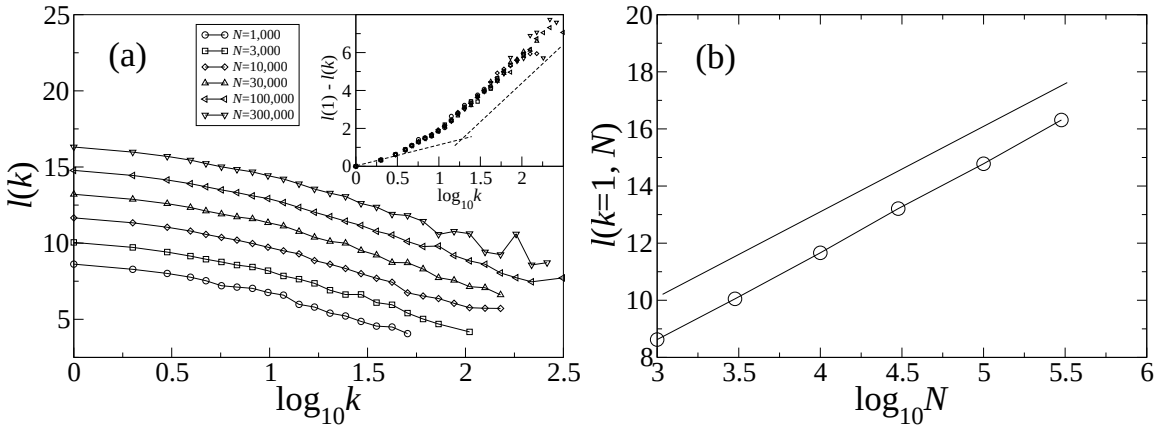

Figure 2.2: Degree-dependent mean intervertex separation in a random scale-free network (tree) growing through the mechanism of preferential attachment. At each time step a new vertex is added. It becomes attached to a vertex selected with probability proportional to the sum of the degree of this vertex and a constant A — "additional attractiveness" [17] (see Section 1.4.4). Here we use $A = 1$. (a) $\ell(k)$ vs. $\log_{10} k$ for networks of $N = 1000$, 3000, 10 000, 30 000, 100 000, and 300 000, vertices. Each of the first four curves were obtained after 50 runs, while for the networks of 100 000 and 300 000 vertices, 20 and 5 runs were used, correspondingly. Binning was made at large degrees, which allowed us to reduce noise. The inset demonstrates that in this network, the difference $\ell(k = 1) - \ell(k)$ does not depend on the size N. In the inset, for the sake of clearness we do not show lines connecting points. The dashed lines highlight two limiting behaviors. As k approaches its minimal value $k = 1$, $\ell(k = 1) - \ell(k) \approx 1.0 \log_{10} k \approx 0.43 \ln k$ for all studied network sizes, while at large degrees, $\ell(k = 1) - \ell(k) \approx \text{const} + 4.1 \log_{10} k \approx \text{const} + 1.8 \ln k$. (b) The dependence of $\ell(k = 1)$ on $\log_{10} N$. For comparison, a line with a slope 3 is shown.

Formula (2.3) fails at $\gamma \to \infty$. E.g., it cannot be applied for networks with an exponential degree distribution. In growing trees with this distribution, we observe the dependence:

$$\ell(k) \cong A \ln N - Bk \,, \tag{2.4}$$

where the constants A and B depend on a network. In particular, we found that this law is exact in deterministic graphs (trees) with an exponential degree distribution [e.g., graph (e) in Fig. 2.1] at least up to very large sizes. Moreover, we observed the same dependence in a simulated stochastically growing tree with random attachment. In this tree (with an exponential degree distribution), at each time step, a new vertex is attached to a randomly selected vertex of the net. The result of the simulation of this network is shown in Fig. 2.3(a). In both the networks—graph (e) in Fig. 2.1 and the corresponding stochastic net with random attachment—the slope of the degree dependence turned out to be $-1/2$. More generally, if in a growing tree of this kind, at each step, n new vertices become attached to a vertex, the slope of the degree dependence equals $-1/(n+1)$ [see Fig. 2.3(b)].

All networks that we studied, had the generic property:

$$\max_k \ell(k) \approx 2 \min_k \ell(k) \,, \tag{2.5}$$

in the large network limit. As is natural, the maximum value of $\ell(k)$ is attained at the minimal degree of a vertex in a network, and vice-versa, the minimum value of $\ell(k)$ is attained at the maximum degree.

2.1.3 $\ell(k)$ of an uncorrelated network

The configuration model [170, 171, 172, 173, 174] is a standard model of an uncorrelated (equilibrium) random network (Section 1.4.2). The mean intervertex distance $\bar{\ell}$ in these networks is estimated in the following way, Ref. [100] (see also Refs. [65, 157]). The mean number of m-th nearest neighbors of a vertex is

$$z_m = z_1 (z_2/z_1)^{m-1} \,, \tag{2.6}$$

where $z_1 = \langle k \rangle$ is the mean number of the nearest neighbors of a vertex, *i.e.* the mean degree. $z_2 = \langle k^2 \rangle - \langle k \rangle$ is the mean number of the second nearest neighbors of a vertex. z_2/z_1 is the branching coefficient of the network. By using formula (2.6), one can get $\bar{\ell}$: $z_{\bar{\ell}} \sim N$, so $\bar{\ell}(N) \approx \ln N / \ln(z_2/z_1)$.

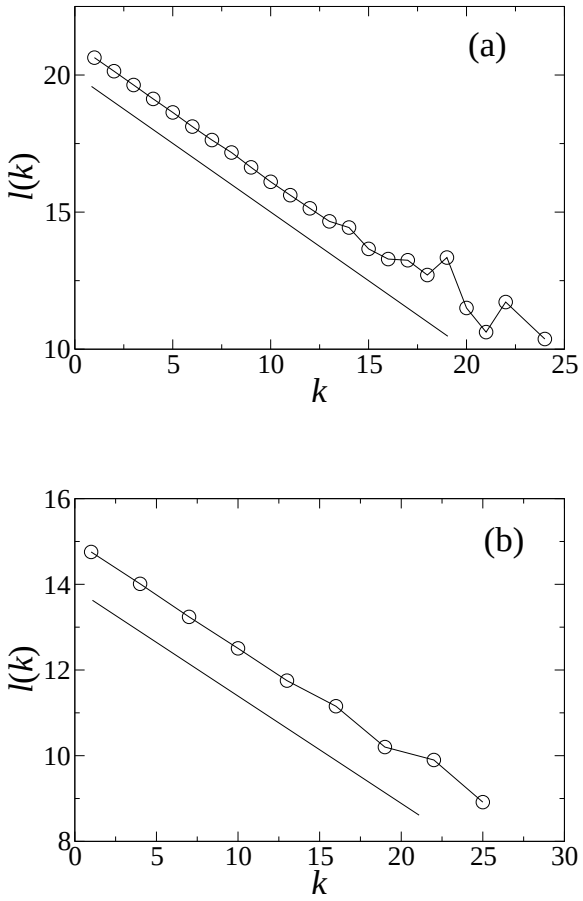

Figure 2.3: Degree-dependent mean intervertex separation in stochastic networks (trees) growing under the mechanism of random attachment. These networks have exponential degree distributions. (a) At each time step, a vertex is attached to a randomly chosen vertex of the network. The dependence is the result of the simulation of the network of 10^5 vertices, 50 runs. For comparison, a line with a slope $-1/2$ is shown. (b) At each time step, 3 vertices are attached to a randomly chosen vertex of the network. The dependence is presented for the network of 9998 vertices, 50 runs. The initial configuration consists of two vertices connected by an edge. For comparison, a line with a slope $-1/4$ is shown. Note that in these plots $\max \ell(k) \approx 2 \min \ell(k)$. In other words, in these networks, there are no vertices of degree greater than $k_{\max}$: $\ell(k_{\max}) = \max \ell(k)/2$. Note fluctuations in the range of the highest degrees.

Similarly, for the mean number of m-th nearest neighbors of a vertex of degree k, we have

$$z_m(k) = k(z_2/z_1)^{m-1}\,. \tag{2.7}$$

So, the estimate is $k(z_2/z_1)^{\ell(k)-1} \sim N$ and thus

$$\ell(k) \approx \frac{\ln(N/k)}{\ln(z_2/z_1)}\,. \tag{2.8}$$

Here we neglected an additional constant independent of N and k which would be excess precision.

Relation (2.7) is evident. It also may be obtained strictly by using the Z-transformation technique [100, 175]:

$$z_m(k) = \left[x\frac{d}{dx}\phi_1^k(\phi_1(\ldots\phi_1(x)))\right]_{x=1}\,. \tag{2.9}$$

$\phi_1(x) = \phi(x)/z_1$ is the Z-transformation of the distribution of the number of edges of an end vertex of an edge with excluded edge itself. $\phi(x)$ is the Z-transformation of the degree distribution of the network: $\phi(x) \equiv \sum_k P(k)x^k$ (see Ref. [100]). Formula (2.9) is a direct consequence of the following features of the configuration model: (i) the network has a locally tree-like structure, (ii) vertices of the network are statistically equivalent, (iii) correlations between degrees of nearest neighbor vertices are absent. Relation (2.9) together with $\phi_1(1) = \phi(1) = 1$ readily leads to relation (2.8).

Note that expression (2.8) also follows from the mentioned result of Holyst *et al.*, Ref. [158], that is $\ell(k,k') \approx \ln[N/(kk')]/\ln(z_2/z_1)$ for the configuration model. Substituting this result into formula (2.1) and ignoring terms independent of N and k immediately gives expression (2.8). In its turn, substituting expression (2.8) into formula (1) leads to the standard formula for the configuration model: $\bar{\ell} \approx \ln N/\ln(z_2/z_1)$.

One point should be emphasized. In the configuration model, the logarithmic size dependence of the (degree-independent) mean intervertex distance $\bar{\ell}(N) \sim \ln N$ is valid only for degree distributions with a finite second moment $\langle k^2\rangle$. If $\langle k^2\rangle$ diverges as $N \to \infty$, $\bar{\ell}(N)$ grows slower than $\ln N$. One can see that the result (2.8) may be generalized to any given form $\bar{\ell}(N)$ of the size-dependence of the mean intervertex distance. In this general case, the degree-dependent separation is expressed in terms of the function $\bar{\ell}(N)$, namely, putting in evidence also its N-dependence, $\ell(k,N) \sim \bar{\ell}(N/k)$.

2.1.4 Derivations

In this section we study a degree-dependent intervertex separation in the deterministic graphs of Fig. 2.1. Graphs (a) – (d) have a discrete spectrum of vertex degrees with a power-law envelope. Graph (e) has a discrete spectrum of vertex degrees with an exponential envelope. We also list some basic characteristics of these graphs. We stress that the main structural characteristics (clustering, degree–degree correlations [176, 21, 177, 38, 178, 77], etc.) of these deterministic networks are quite close to those of their stochastic analogs (see [70]).

(A) Graph (a) in Fig. 2.1.—This graph was proposed in Ref. [43] and extensively studied in Ref. [70]. The growth starts from a single edge ($t = 0$). At each time step, each edge of the graph transforms into a triangle. Actually, we have a deterministic version of a stochastic growing network with attachment of a new vertex to a randomly chosen edge, see Ref. [179]. The number of vertices of the graph is $N_t = 1 + (3^t + 1)/2$. ($t = 0, 1, 2, \ldots$ is the number of the generation.) In the large network limit, the mean degree of the graph is $\langle k\rangle \to 4$.

Degrees of the vertices in the graph take values $k(s) = 2^s$, $s = 1, 2, \ldots, t$. The spectrum of degrees has a power-law envelope. This spectrum corresponds to a continuum scale-free spectrum $P(k) \propto k^{-\gamma}$ with exponent $\gamma = 1 + \ln 3/\ln 2 = 2.585\ldots$. Note that this network has numerous triangles, which suggests high clustering. In more detail, by definition, the average clustering coefficient of a vertex of degree k is (see also the equivalent Eq. 1.10)

$$C(k) = \left\langle \frac{c(k)}{k(k-1)/2} \right\rangle_k = \frac{\langle c(k)\rangle_k}{k(k-1)/2}\,. \tag{2.10}$$

Here, $c(k)$ is the number of triangles attached to a vertex of degree k, and $\langle\cdots\rangle_k$ means the averaging over all vertices of degree k. One can see that in this graph (as well as in its stochastic version)

$$C(k) = \frac{2}{k}\,. \tag{2.11}$$

[Indeed, by construction, the number of triangles attached to a vertex of degree k in the graph is $k - 1$. So, $C(k) = (k-1)/[k(k-1)/2] = 2/k$.] This gives, for the mean clustering,

$$\langle C\rangle = \sum_k P(k)C(k) = \frac{4}{5}\,, \tag{2.12}$$

while the standard clustering coefficient (transitivity), i.e., the density of loops of length 3 in a

network,

$$T = \frac{\sum_k P(k)C(k)k(k-1)}{\sum_k P(k)k(k-1)}\,, \tag{2.13}$$

approaches zero in the infinite network limit, $T = 0$. Note the difference between the finite mean clustering of the network and its zero clustering coefficient.

In principle, one may derive an exact analytical expression for the degree-dependent separation by using recursion relations and the Z-transformation technique. However, these calculations turn out to be cumbersome. Instead, here we only check that some analytical formula for $\ell(k)$ is valid in a sufficiently large number of generations of a deterministic graph, up to, say, $t \sim 10$ or 12. So, we confirm a guessed expression in networks of sizes up to $N \sim 10^5$. In fact, we implement the following approach:

(i) Find the mean separation values $\ell_t(s)$ for all kinds of vertices in each of several first generations of the deterministic graph [t is the number of generation, and $k = 2^s$, $s = 1, 2, \ldots, t$];

(ii) by using this array of numbers, guess the form of $\ell_t(s)$;

(iii) check this result by computing directly $\ell_t(s)$ for several extra generations of the graph.

There are few computations in stage (i): we have to find only t values of $\ell_t(s)$ in a t generation of a graph. For sufficiently small networks, these values can be found even without a computer. Step (ii) also turns out to be rather easy since we already know the structure of the analytical expressions for a mean intervertex distance in these networks (see Ref. [70]). Step (iii) may be performed by using a computer to count paths. This approach is based on our experience with problems on these graphs and was checked in Ref. [70] for related quantities. Our guess actually exploits underlined recursion relations without revealing them. Nonetheless, we can only claim that the analytical expressions, obtained in this way, are valid at the studied generations of our deterministic graphs. In principle, there exists a (small) chance that at some higher generation (or generations), these formulas fail. Thus, the results of this section should be considered only as observations of $\ell(k)$ for a set of networks of a modest size.

In this way, we get

$$\ell_t(s) = \frac{1}{2(N_t - 1)}[2(2t - s + 5)3^{t-2} - 3^{s-1} + 1]\,. \tag{2.14}$$

This formula is valid for $t \geq 1$. We checked it up to $t = 12$, which corresponds to $N_t = 265\,722$. We also checked that this formula leads to the known exact formula for the mean intervertex distance $\bar{\ell}$ for any t and so that N [70]. An asymptotic form of this expression is

$$\ell(k, N) = \frac{4}{9\ln 3}\ln N - \frac{2}{9\ln 2}\ln k - \frac{k^{\gamma-1}}{6N} + \frac{4}{9}\frac{\ln 2}{\ln 3} + \frac{10}{9} + \dots \tag{2.15}$$

at large N, where N is the total number of vertices in the graph. This leads to formula (2.3).

One can see that the minimum value of $\ell(k)$ is $\ell_{\min} = \ell(k = 2^t) \cong 2t/9$, where $t \cong \ln N/\ln 3$. On the other hand, its maximum value is $\ell_{\max} = \ell(k = 2) \cong 4t/9$. So, we arrive at relation (2.5): $\ell_{\max} = 2\ell_{\min}$.

(B) Graph (b) in Fig. 2.1.—This graph was proposed in Ref. [161]. At each time step, each edge of the graph transforms in the following way: each end vertex of the edge gets a new vertex attached [see Fig. 2.1, graph (b), instant 0 → instant 1]. This graph is very similar to graph (a). In particular, the exponent of its degree distribution is the same, $\gamma = 1 + \ln 3/\ln 2 = 2.585\dots$. The difference is that the graph is a tree, so the mean degree $\langle k\rangle \to 2$ as $N \to \infty$.

The total number of vertices in the graph is $N_t = 3^t + 1$. The vertices have degrees $k(s) = 2^s$, where $s = 0, 1, 2, \dots, t$. In the same way as for graph (a), we find the expression

$$\ell_t(s) = \frac{1}{2(N_t - 1)}[(4t - 2s + 9)3^{t-1} - 3^s]\,, \tag{2.16}$$

which is observed starting with $t = 0$. This leads to the asymptotic relation

$$\ell(k, N) = \frac{2}{3\ln 3}\ln N - \frac{1}{3\ln 2}\ln k - \frac{k^{\gamma-1}}{2N} + \frac{3}{2} + \dots\,, \tag{2.17}$$

that is, to formula (2.3).

The minimum value of $\ell(k)$ is $\ell_{\min} = \ell(k = 2^t) \cong t/3$, where $t \cong \ln N/\ln 3$. The maximum value is $\ell_{\max} = \ell(k = 1) \cong 2t/3$, i.e., again, we arrive at relation (2.5).

(C) Graph (c) in Fig. 2.1.—At each step, (i) a new vertex becomes attached to each end vertex of each edge of this graph and, simultaneously, (ii) a new vertex becomes attached to each vertex of the graph. This produces a growing deterministic scale-free tree with exponent $\gamma = 3$, which is a deterministic analog of the Barabási-Albert model [15, 16] (for exact solution of the stochastic model, see Refs. [17, 176, 18]).

The number of vertices in the graph is $N_t = 1 + (4^{t+1} - 1)/3$. Their degrees take values $k(s) = 2^s - 1$, $s = 1, 2, 3, ..., t + 1$. The observed degree-dependent separation is

$$\ell_t(s \geq 2) = \frac{1}{9(N_t - 1)}[2(6t - 3s + 10)4^t - 4^s - 1]\,. \tag{2.18}$$

Asymptotically, this is

$$\ell(k, N) = \frac{1}{\ln 4} \ln N - \frac{1}{2 \ln 2} \ln k - \frac{k^{\gamma-1}}{9N} + \frac{\ln 3}{2 \ln 2} + \frac{2}{3} + \dots \tag{2.19}$$

for $k, N \gg 1$ (note that the maximum degree of a vertex in this graph is $k_{\max} \sim N^{1/2}$). This leads to expression (2.3) with $\gamma = 3$, which coincides with result (2.8) for uncorrelated networks. This is an understandable coincidence. Indeed, correlations between degrees of the nearest neighbor vertices in this deterministic graph, as well as in the Barabási-Albert model are anomalously week. So, the result must be close to that for an uncorrelated network.

The minimum value of $\ell(k)$ in this graph is $\ell_{\min} = \ell(k = 2^{t+1} - 1) \cong t/2$, where $t \sim \ln N / \ln 4$. The maximum value is $\ell_{\max} = \ell(k = 1) \cong t$, so that relation (2.5) is fulfilled.

(D) Graph (d) in Fig. 2.1.—At each step, (i) a pair of new vertices is attached to ends of each edge of the graph plus (ii) two new vertices are attached to each vertex of the graph. This results in the value of the γ exponent greater than 3, $\gamma = 1 + \ln 5 / \ln 2 = 3.322\dots$.

The number of vertices in the graph is $N_t = (3 \cdot 5^t + 1)/2$. Degrees of the vertices are $k(s) = 3 \cdot 2^{s-1} - 2$, $s = 1, 2, 3, ..., t + 1$. The observed expression for the degree-dependent separation is

$$\ell_t(s) = \frac{1}{8(N_t - 1)}[(72t - 36s + 71 + 5^{3-s})5^{t-1} + 2\,5^{s-1} - 6]\,. \tag{2.20}$$

The corresponding asymptotic expression is of the following form:

$$\ell(k, N) = \frac{6 \ln N}{5 \ln 5} - \frac{3 \ln k}{5 \ln 2} - \frac{5^{-\ln 3 / \ln 2}}{4N} k^{\gamma-1} + 1.232 + \dots\,, \tag{2.21}$$

where the contribution $1.232\dots = [6 \ln(2/3)]/(5 \ln 5) + (3 \ln 3)/(5 \ln 2) + 7/12$. Again, now with the graph where $\gamma > 3$, we arrive at formula (2.3).

In this graph, we have $\ell_{\min} = \ell(k = 3 \cdot 2^t - 2) \cong 3t/5$ and $\ell_{\max} = \ell(k = 1) \cong 6t/5$, where $t \cong \ln N / \ln 5$.

The important feature of the expressions for $\ell(k, N)$ in deterministic scale-free networks with $\gamma \neq 3$ were non-equal coefficients of $\ln N$ and $\ln k$. For comparison we have measured $\ell(k, N)$ in

a random growing scale-free network growing through the mechanism of preferential attachment with a linear preference function [17]. At each time step, a new vertex emerges and becomes attached to a vertex chosen with probability proportional to the sum of its degree and a constant A. Exponent $\gamma = 3 + A$. We use $A = 1$, so that $\gamma = 4$. The resulting degree-dependent separations are shown in Fig. 2.2(a) for networks of up to 300 000 vertices. One can see in the inset that in these random networks, the difference $\ell(k = 1, N) - \ell(k, N)$ is independent of N in contrast to the deterministic graphs (a)—(d). Furthermore, $[\ell(k = 1, N) - \ell(k, N)]/\log_{10} k \approx 1.0$ as $\log_{10} k$ approaches zero [i.e., $d\ell(k, N)/d\ln k \approx -0.43$]. However, at large k, we find a linear dependence on $\log_{10} k$ with a larger slope, namely 4.1 [i.e., $d\ell(k, N)/d\ln k \approx -1.8$]. In its turn, $\ell(k = 1, N)$ is well fitted by a linear dependence on $\log_{10} N$ with a slope approximately 3.1, see Fig. 2.2(b) [i.e., $d\ell(k = 1, N)/d\ln N \approx 1.35$]. The difference in these slopes — 4.1 and 3.1 — is in sharp contrast to uncorrelated networks. The ratio of these slopes, 1.3 is close to what we had for deterministic graphs according to Eq. (2.3) with $\gamma = 4$ substituted, namely, $(\gamma - 1)/2 = 1.5$. Moreover, Fig. 2.2(a) shows that for each network size, $\ell_{\max} \approx 2\ell_{\min}$, as was observed in deterministic graphs.

One should note that the contribution $\sim k^{\gamma-1}/N$ to $\ell(k, N)$ for the deterministic graphs, is noticeable only in a narrow neighborhood of $k_{\max}$, if results are presented in the form $\ell(k, N)$ vs. $\ln k$. On the other hand, the linear dependence $\ell(k, N)$ on $\ln k$ is realized in a much wider range of $\ln k$. In Eq. (2.15)—graph (a), it is valid for all degrees up to nearly $k_{\max}$, and in Eqs. (2.17), (2.19), and (2.21)—graphs (b), (c), and (d), respectively, this law is observable for $k \gg 1$. It is in this region that we compared the rations of the coefficients of $\ln k$ and $\ln N$ in deterministic and stochastic growing scale-free networks.

(E) Graph (e) in Fig. 2.1.—At each time step, a new vertex becomes attached to each vertex of the graph. The growth starts with a single vertex ($t = -1$). The total number of vertices in the graph is $N_t = 2^{t+1}$. The degree distribution is exponential. One can check that the number of vertices of degree k at time t is $N_t(k \le t) = 2^{t+1-k}$, $N_t(k = t + 1) = 2$ (t is assumed to be greater than -1).

By using the above described procedure, we find the exact expression:

$$\ell_t(k) = \frac{2^t}{2^{t+1} - 1}\,(2t + 2 - k)\,. \tag{2.22}$$

This formula shows that the linear dependence on degree is valid for any k. For the large graphs we

have

$$\ell(k,N) \cong \frac{\ln N}{\ln 2} - \frac{k}{2}\,, \tag{2.23}$$

which confirms formula (2.4).

In this graph, $\ell_{\min} \cong \ln N/(2\ln 2) \cong \ell_{\max}/2$ which coincides with relation (2.5).

Graph (e) has a close stochastic analog—a tree, where at each step, a new vertex is attached to a randomly chosen vertex. It is easy to obtain the asymptotic expression for the mean shortest path length $\bar{\ell}(N)$ in this network. Let us consider even more general model. Let at each time step, n new vertices be attached to a randomly selected vertex. Then the total number of vertices N grows as $N_t \cong nt$. For the total length of the shortest paths between vertices in the network at time $t+1$ one can right:

$$\frac{N_{t+1}(N_{t+1}-1)}{2}\,\bar{\ell}(t+1) = \frac{N_t(N_t-1)}{2}\,\bar{\ell}(t)$$
$$+\frac{1}{N_t}\,N_t\left(1\cdot n + 2\,\frac{n(n-1)}{2} + n(N_t-1)[\bar{\ell}(t)+1]\right). \tag{2.24}$$

The first term on the right-hand side of this equation is the total length of the shortest paths in the network at time t. The second term is the increase of this total length due to the attachment of n new vertices to a randomly chosen vertex. The factor $1/N_t$ is due to the random choice. The term $1\cdot n$ is the sum of the paths connecting the new vertices to their "host". The term $2\cdot n(n-1)/2$ is the total length of the paths between the new vertices. The last term in the large parentheses is the sum of the lengths of the paths connecting the n new vertices and the $N_t - 1$ old vertices distinct from the vertex receiving new connections. In the large network limit, Eq. (2.24) is readily reduced to the following one:

$$\frac{N^2}{2}n\frac{d\bar{\ell}}{dN} = -\frac{n(n+1)}{2}\bar{\ell} + nN \cong nN, \tag{2.25}$$

and so we have

$$\bar{\ell} \cong 2\ln N, \tag{2.26}$$

independent of n.

The calculation of $\ell(k)$ is a more difficult problem. So, for comparison, we present here only the result of the simulation of this stochastic network. Figure 2.3(a) demonstrates that the dependence

$\ell(k)$ in the stochastically growing network is a linear function with the same slope $-1/2$ as in the deterministic small world (e) in Fig. 2.1.

We also considered more general deterministic graphs of this type, where n new vertices become attached to each vertex of a network at each time step. The resulting dependence $\ell(k)$ is a linear function but with slope $-1/(n+1)$. Figure 2.3(b) shows that $\ell(k)$ of the corresponding stochastically growing networks has the same form. We also checked that $\ell(k=1,N) \approx 2\ln N$, as in expression (2.26) for $\bar{\ell}(N)$.

2.1.5 Discussion and summary

Several points should be emphasized:

(i) One can estimate a typical value of the correction term in formula (2.3). At the maximum degree $k_{\max} \sim N^{1/(\gamma-1)}$, this term is of the order of $k_{\max}^{\gamma-1}/N \sim$ const. This should be compared to $\ln[k_{\max}^{1/(\gamma-1)} \sim \ln N]$.

(ii) One should indicate that law (2.4), i.e., a linear dependence $\ell(k)$, was obtained only for growing trees with an exponential degree distribution. In non-tree growing networks with random attachment (at each time step, a new vertex becomes attached to several randomly chosen vertices), we observed a non-linear dependence.

(iii) The relative width of the distribution of the intervertex distance in infinite small worlds approaches zero [63, 64, 65]. In other words, vertices of an infinite small world are almost surely mutually equidistant (Section 1.2.2). This circumstance does not allow one to measure $\ell(k)$ in an infinite network with the small-world effect, for which $\ell(k) \equiv \bar{\ell}$. However, even in very large real-world networks (e.g., in the Internet [177]), the distribution of the intervertex distance is still broad enough. So, in real networks, $\ell(k)$ is a measurable characteristic, as we will see in Chapter 2 for a real world network of scientific collaborations.

(iv) The degree-dependent mean intervertex distance may be considered as a measure of "centrality" of a given degree vertex in a network. How does this relate to other centrality characteristics [160], first of all to the centrality index of a vertex [88]? Recall from Section 1.2.6 that the centrality index of a vertex v is defined as $c_v = (N-1)/\sum_u \ell_{vu}$, where ℓ_{vu} is the length of the shortest path between vertices u and v, N is the number of vertices in the graph, and the sum is over all vertices

of the graph. (The centrality index is often given without the $N-1$ factor.) One may see that the mean centrality index $c(k)$ of a vertex of degree k is related (but not equal) to $1/\ell(k)$. Nevertheless, there is a special case—graphs where every vertex of a given degree k has the same value of the sum of intervertex distances between this and the rest of the vertices. So, this value is exactly $(N-1)\ell(k)$, and consequently $c(k) = 1/\ell(k)$. This situation is realized in our deterministic graphs. Thus, in the deterministic graphs, we actually found the inverse centrality index, but in random networks, $c(k)$ and $\ell(k)$ are different characteristics.

In conclusion, we have studied the mean length of the shortest paths between a vertex of degree k and the other vertices in growing networks with power-law and exponential degree distributions. In the investigated deterministic and random networks, we have observed dependences $\ell(k)$ which strongly differ from those for uncorrelated networks. Our results characterize the compactness of a network from the point of view of a vertex with a given number of connections.

2.2 Evolution of subgraphs and cycles in complex networks

Subgraphs and cycles are often used to characterize the local properties of complex networks (see Section 1.2.5). Here we show that the subgraph structure of real-world networks (see also Section 1.3) is highly time dependent: as the network grows, the density of some subgraphs remains unchanged (which we called Type II), while the density of others (Type I) increase at a rate that is determined by the network's degree distribution and clustering properties. This inhomogeneous evolution process, supported by direct measurements on several real networks and on the deterministic model of Fig. 2.1a, leads to systematic shifts in the overall subgraph spectrum and to an inevitable overrepresentation of some subgraphs and cycles.

2.2.1 Introduction

Motivated by practical and theoretical questions, recently a series of statistical tools have been introduced to evaluate the abundance of subgraphs [80, 81, 82, 79] and cycles [180, 181, 165, 183], offering a better description of a network's local structure. Yet, most of these methods were designed to capture the subgraph structure of a specific snapshot of a network, characterizing static graphs. Most

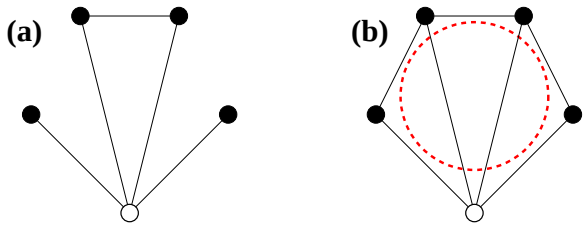

Figure 2.4: Examples of subgraphs and cycles with a central vertex. The subgraph shown in (a) has $n = 5$ vertices and $n - 1 + t = 5$ edges, where $t = 1$ represents the number of edges connecting the neighbors of the central vertex (empty circle) together. In (b) we show a subgraph with $t = 3$ edges among the neighbors, such that the central vertex and its neighbors form a cycle of length $h = 5$, highlighted by the dotted circle

real networks, however, are the result of a growth process, and keep evolving in time [42, 43]. While growth often leaves some of the network's global features unchanged, it alters its local, subgraph based structure, potentially modifying everything from subgraph densities to cycle abundance. Yet, the currently available statistical methods cannot anticipate or describe such potential changes.

In this section we show that during growth the subgraph structure of complex networks undergoes a systematic reorganization. We find that the evolution of the relative subgraph and cycle abundance can be predicted from the degree distribution $P(k)$ and the degree-dependent average clustering coefficient $C(k)$. The results indicate that the subgraph composition of complex networks changes in a very inhomogeneous manner: while the density of many subgraphs is independent of the network size, they coexist with a class of subgraphs whose density increases at a subgraph dependent rate as the network expands. Therefore in the thermodynamic limit a few subgraphs will be highly overrepresented [80, 81], a prediction that is supported by direct measurements on a number of real networks for which time resolved network topologies are available. This finding questions our ability to characterize networks based on the subgraph abundance obtained from a single topological snapshot. We show that a combined understanding of network evolution and subgraph abundance offers a more complete picture.

2.2.2 Subgraphs

We consider subgraphs with n vertices and $n-1+t$ edges, whose central vertex has links to $n-1$ neighbors, which in turn have t links among themselves (Fig. 2.4a). The total number of n-node subgraphs that can pass by a node with degree k is $\binom{k}{n-1}$. Each of these n-node subgraphs can have at most $n_{\rm p} = (n-1)(n-2)/2$ edges between the $n-1$ neighbors of the central node. The probability that there is an edge between two neighbors of a degree k vertex is given by the clustering coefficient $C(k)$ (Section 1.2.3). Therefore, the probability to obtain t connected pairs and $n_{\rm p}-t$ disconnected pairs is given by the binomial distribution of n_p trials with probability $C(k)$. The expected number of (n,t) subgraphs in the network is obtained after averaging over the degree distribution, resulting in

$$N_{nt} = g_{nt} N \sum_{k=1}^{k_{\rm max}} P(k) \binom{k}{n-1} \binom{n_{\rm p}}{t} C(k)^t [1-C(k)]^{n_{\rm p}-t} \ , \tag{2.27}$$

where $k_{\rm max}$ is the maximum degree and the geometric factor g_{nt} takes into account that the same subgraph can have more than one central vertex. For instance, a triangle will be counted three times since each vertex is connected to the others, therefore $g_{31} = 1/3$. For networks where $P(k) \sim k^{-\gamma}$ and $C(k) \sim k^{-\alpha}$, where γ and α are the degree distribution and clustering hierarchy exponents, in the thermodynamic limit $k_{max} \gg 1$, Eq. (2.27) predicts the existence of two subgraph classes [79]

$$\frac{N_{nt}}{N} \sim \begin{cases} C_0^t k_{\rm max}^{n-\gamma-\alpha t} \ , & n-\gamma-\alpha t > 0 \ , \quad \text{Type I} \ , \\ C_0^t \ , & n-\gamma-\alpha t < 0 \ , \quad \text{Type II} \ . \end{cases} \tag{2.28}$$

Therefore, for the Type I subgraphs the N_{nt}/N density increases with increasing network size, and N_{nt}/N is independent of N for Type II subgraphs. In the following we provide direct evidence for the two subgraph types in three real networks for which varying network sizes are available: coauthorship network of mathematical publications [96], the autonomous system representation of the Internet [21, 177], and the semantic web of English synonyms [184]. In each of these networks the maximum degree increases as $k_{max} \sim N^{\delta}$. We estimated δ from the scaling of the degree distribution moments with the graph size, $\langle k^n \rangle \sim N^{\delta(n+1-\gamma)}$, with $n = 2, 3, 4$. Furthermore, we find that C_0 from $C(k) = C_0 k^{-\alpha}$ also depends on the network size as $C_0 \sim N^{\theta}$, where θ can be estimated using

Network	γ	α	δ	θ	ζ_3	ζ_4	ζ_5
Co-authorship	2.4	0.0	0.6	0.00	0.6	1.6	2.6
Internet	2.2	0.75	1.0	0.20	0.3	0.7	1.2
Language	2.7	1.0	0.40	0.68	0.7	1.4	2.0
Model	2.6	1	0.63	0	0	0	0

Table 2.1: Characteristic exponents of the investigated real networks and the deterministic model of Fig. 2.1a. The exponents are defined through the scaling of the degree distribution $P(k) \sim k^{-\gamma}$, the clustering coefficient $C(k) = C_0 k^{-\alpha}$, with $C_0 \sim N^{\theta}$, the largest degree $k_{max} \sim N^{\delta}$, and the number of h-cycles $N_h/N \sim N^{\zeta_h}$.

$C_0 = \sum_{k\geq 2} C(k)/\sum_{k\geq 2} k^{-\alpha}$, giving a better estimate than a direct fit of $C(k)$. The exponents characterizing each network are summarized in Table. 2.1.

In Fig. 2.5 we show the density of all five vertex subgraphs ($n = 5$) as a function of t. For the Internet and Language networks C_0 increases with N, therefore the subgraph's density increases with the network size for all subgraphs. This consequence of the non-stationarity of the clustering coefficient is subtracted by normalizing N_{nt} by C_0^t. For the co-authorship graph with $\alpha = 0$ (Table 2.1), only Type I subgraphs are observed, as predicted by (2.28). In contrast, for the Internet and semantic networks $\alpha > 0$, therefore the overrepresented Type I phase is expected to end approximately at the phase boundary predicted by (2.28). Indeed, left to the arrow denoting the $n - \gamma - \alpha t$ phase boundary we continue to observe a systematic increase in N_{5t}/NC_0^t, as expected for Type I subgraphs. In contrast, beyond the phase boundary the subgraph densities obtained for different network sizes are independent of N, collapsing into a single curve.

We compared also our predictions with direct counts in the growing deterministic network model [70] of Fig. 2.1a, characterized by a degree exponent $\gamma = 1 + \ln 3/\ln 2 \approx 2.6$ and a degree dependent clustering coefficient $C(k) = C_0 k^{-\alpha}$, with $C_0 = 2$ and $\alpha = 1$. In Fig. 2.5d we show the number of ($n = 5$,t) subgraphs for different values of t and graph sizes. The arrow indicating the predicted phase transition point $n - \gamma - \alpha t = 0$ clearly separates the Type I from the Type II subgraphs, a numerical finding that is supported by exact calculations as well. Note that only one Type II $n = 5$ subgraph is present in the deterministic network, due to its particular evolution rule.

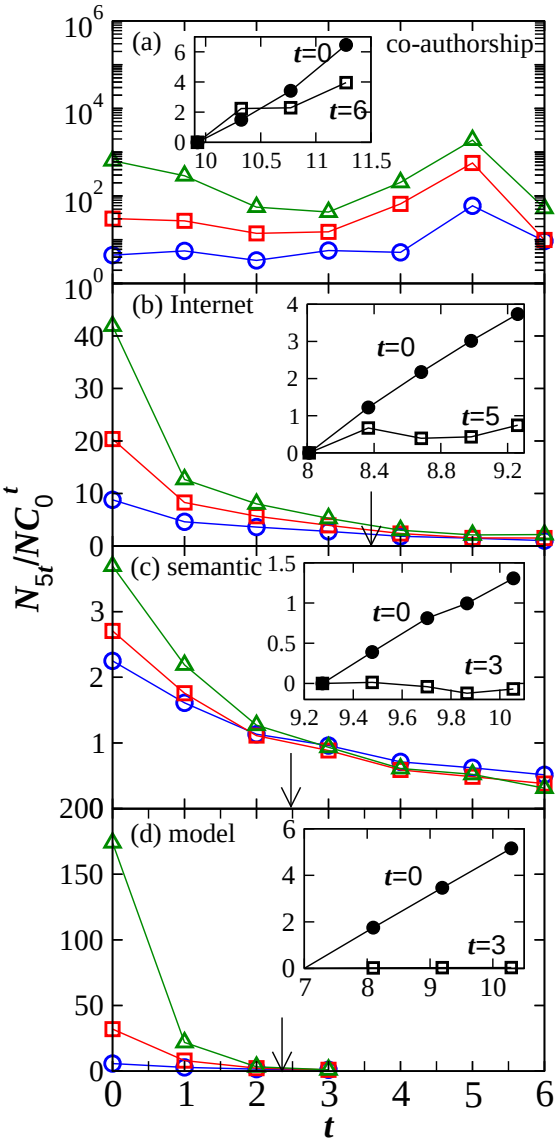

Figure 2.5: Number of ($n = 5$,t) subgraphs for the co-authorship (a), Internet (b), semantic (c) networks and the deterministic model (d) as a function of t. Different symbols correspond to different snapshots of the networks evolution, from early stage (circles) to intermediate (squares) and current (*i.e.* largest) (triangles). N_{nt} depends strongly on t (spanning several orders of magnitude) making difficult to observe the N dependence. Thus we normalized all the quantities (N_{5t}, C_0 and N) to the first year available. The arrows correspond to the phase boundary $5 - \gamma - \alpha t = 0$, with Type I and II subgraphs to the left and right of the arrow, respectively. In the insets showing the system size dependence we plot $\log N_{5t}$ vs $\log N$ for different values of t.

2.2.3 Cycles

The formalism developed above can be generalized to predict cycle abundance as well. Consider the set of centrally connected cycles shown in Fig. 2.4b. If the central vertex has degree k, we can form $\binom{k}{h-1}$ different groups of h vertices, $h-1$ selected from its k neighbors and the central vertex. Each ordering of the $h-1$ selected neighbors corresponds to a different cycle, therefore we multiply with half of the number of their permutations $(h-1)!$ (assuming that 123 is the same as 321). Finally, to obtain the number of h-cycles we multiply the result with the probability of having $h-2$ edges between consecutive neighbors, $C(k)^{h-2}$, and sum over the degree distribution $P(k)$, finding

$$\frac{N_h}{N} = g_h \sum_{k=h-1}^{k_{max}} P(k) \frac{(h-1)!}{2} \binom{k}{h-1} C(k)^{h-2} \,, \tag{2.29}$$

where g_h is again a geometric factor correcting multiple counting of the same cycle. Note that (2.29) represents a lower bound for the total number of h-cycles, which also include cycles without a central vertex. Depending on the values of h, γ and α the sum in (2.29) may converge or diverge in the limit $k_{max} \to \infty$. When it converges, the density of h-cycles is independent of N (Type II), otherwise it grows with N (Type I). Since in preferential attachment models without clustering the density of h-cycles decreases with increasing N [185], we conclude that clustering is the essential feature that gives rise to the observed high h-cycle number in such real networks like the Internet [180]. To further characterize the cycle spectrum, we need distinguish two different cases, $0 < \alpha < 1$ and $\alpha \geq 1$.

$0 < \alpha < 1$: In the $k_{max} \to \infty$ limit the cycle density follows

$$\frac{N_h}{N} \sim \begin{cases} C_0^{h-2} \,, & h < h_c \,, \\ C_0^{h-2} k_{max}^{(1-\alpha)(h-h_c)} \,, & h > h_c \,, \end{cases} \tag{2.30}$$

where $h_c = (\gamma - 2\alpha)/(1-\alpha)$. Therefore, large cycles ($h > h_c$) are abundant, their density growing with the network size N. As $\alpha \to 1$ the threshold $h_c \to \infty$, therefore the range of h for which the density is size-independent expands significantly.

Direct calculations using (2.29) show that N_h exhibits a maximum at some intermediate value of h (see Fig. 2.6a), already reported for the deterministic model [165, 182]. The maximum represents a finite size effect, as the characteristic cycle length h^*, corresponding to the maximum of N_h, scales

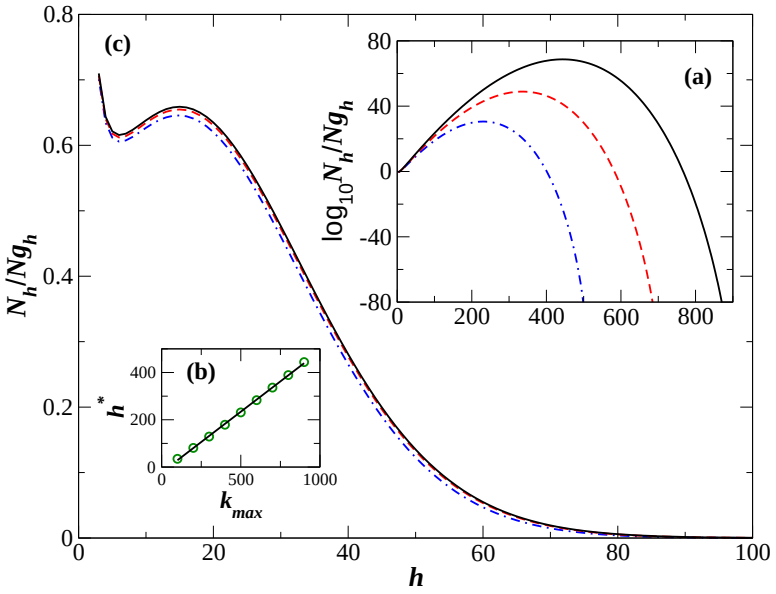

Figure 2.6: Number of h-cycles as computed from (2.29), using $\gamma = 2.5$, (a) $C_0 = 1$ and $\alpha = 0.9$, (b) h value at which N_h has a maximum as a function of k_{max}, (b) $C_0 = 2$ and $\alpha = 1.1$, and $k_{max} = 500$ (dashed-dotted), 700 (dashed) and 900 (solid).

as $h^* \sim k_{max}$ (Fig. 2.6b). Yet, next we show that this behavior is not generic, but depends on the value of α.

$\alpha \geq 1$: For all $\gamma > 2$ only Type II cycles are expected ($N_h/N \sim C_0^{h-2}$), as suggested by the divergence of h_c in the $\alpha \to 1$ limit. If $C_0 > 1$ the number of h-cycles continues to exhibit a maximum and the characteristic cycle length h^* scales as $h^* \sim k_{max}$. If $C_0 < 1$, however, the number of h-cycles decreases with h, although a small local minima is seen for small cycles. More important, in this case N_h/N is independent of the network size (see Fig. 2.6c), in contrast with the size dependence observed earlier (Fig. 2.6a and [165]). Thus, for networks with $\alpha > 1$ or $\alpha = 1$ and $C_0 < 1$ the cycle spectrum is stationary, independent of the stage of the growth process in which we inspect the network.

Our predictions for the cycle abundance are based on centrally connected cycles, in which a central vertex is connected to all vertices of the cycle (Fig. 2.4b). In the following we show that our predictions capture the scaling of all h-cycles as well, not only those that are centrally connected. For this in Fig. 2.7 we plot the number of $h = 3, 4, 5$ cycles (*i.e.* all cycles as well as those that

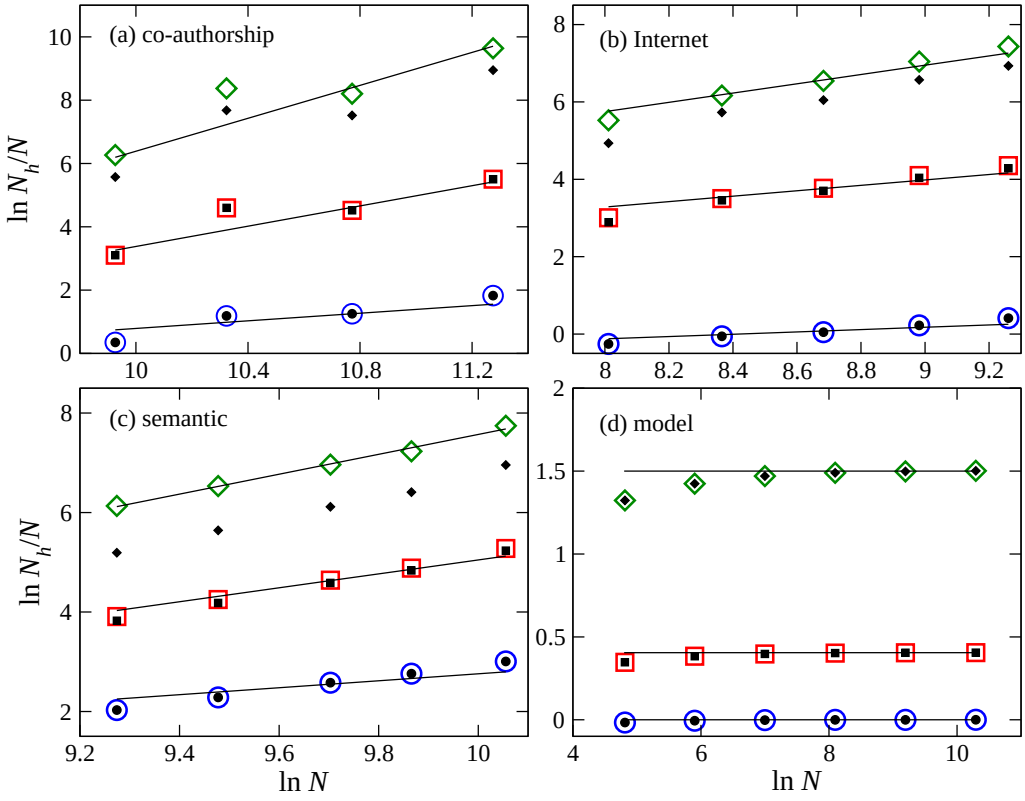

Figure 2.7: Density of all (open symbols) and centrally connected (filled symbols) cycles with $h = 3$ (circles), 4 (squares) and 5 (diamond) cycles as a function of the graph size. The continuous lines correspond with our predictions (Tab. 2.1).

are centrally connected) as a function of the graph size for the studied real and model networks, together with our predictions (continuous line). First we note that in many cases ($h = 3$ and 4) the full cycle density and the density of the centrally connected cycles overlap. In the few cases ($h = 5$) where there are systematic differences between the two densities the N-dependence of the two quantities is the same, indicating that our calculations correctly predict the scaling of all cycles.

For the co-authorship and Internet graphs $\alpha < 1$ and $h_c < 3$, therefore the $h = 3, 4, 5$ cycles are predicted to be in the Type I regime ($h > h_c$). In this case $N_h/N \sim N^{\zeta_h}$, where $\zeta_h = \theta(h-2) + \delta(1-\alpha)(h-h_c)$. For the language graph $\alpha = 1$, therefore $\zeta_h = \theta(h-2)$. For the deterministic model a direct count of the h-cycles reveals that they are of Type II, *i.e.* their density is independent of N [165], in agreement with our predictions for $\alpha \geq 1$. These predictions are shown as continuous lines in Fig. 2.7, indicating a good agreement with the real measurements.

2.2.4 Conclusion

Our results offer evidence of a quite complex subgraph dynamics. As the network grows, the density of the Type II subgraphs remains unchanged, being independent of the system size. In contrast, the density of the Type I subgraphs increases in an inhomogeneous way. Indeed, each (n,t) subgraph has its own growth exponent ζ_{nt}, which means that their density increases in a differentiated manner: the density of some Type I subgraphs will grow faster than the density of the other Type I subgraphs. Thus, inspecting the system at several time intervals one expects significant shifts in subgraphs densities. As a group, with increasing network size the Type I subgraphs will significantly outnumber the constant density Type II subgraphs. Therefore the inspection of the subgraph density at a given moment will offer us valuable, but limited information about the overall local structure of a complex network. Note that nearest neighbor degree correlations, described by $P(k, k')$, were neglected. However, the $P(k)$ and the $C(k)$ functions already allow us to predict with high precision the future shifts in subgraph densities, indicating that a precise knowledge of the global network characteristics can help us to fully understand the local structure of the network at any moment. These results will eventually lead us to reevaluate a number of concepts, ranging from the potential characterization of complex networks based on their subgraph spectrum to our understanding of the impact of subgraphs on processes taking place on complex networks [29, 186].

Chapter 3

University and industry interplay FP5 network

3.1 Introduction

Understanding the relationship between research and industry is essential to improve the quality of life in any society. Ranging from faster application of new discoveries to knowing whether or where investment should be applied, this flow of knowledge between research and industry has long been of general interest. Yet, knowledge is a special resource whose study demands new techniques. The traditional approach to resources is based on scarcity since they are usually finite, but knowledge cannot be seen this way because it grows, and the more it is used the more it spreads [187]. In addition, existing studies on the research and industry interplay [95, 188, 189] have neglected its network character. Our approach consists in analyzing this issue from a complex network viewpoint [42, 48, 45]. In this approach, the interaction between research and industry is best described as a network whose vertices represent either companies or institutions devoted to research, and each edge represents collaboration between any two of them. Hence, we can quantitatively study how research and industry influence each other, by recurring to data describing a real system.

Here, we focus our attention in the Framework Programme (FP), a mechanism aiming to improve the transference of knowledge in the European Union (EU) by setting out its priorities for research

and technological development. The data to generate the corresponding FP network were gathered from the CORDIS website[1] by a Perl[2] script. Since, at the time the data was collected, the 6th programme was under execution and the 7th was being planned, we focused our study in the 5th Framework Programme (FP5)—covering the period from 1998 to 2002—in order to analyze a completely finished programme. Despite the presence of more than 25,000 participants, they can be split in two major groups: Companies and Universities. The first is made of over 16,700 companies and other industry related participants who expect their investments in R+D+I to be profitable. The second group can be regarded as the opposite, more than 8,500 participants involved in some type of research for whom results do not necessarily return immediate income (see Appendix A). Exploring the relationship between these two groups not only provides a good example of the interplay between structure and information flow, but also offers a glimpse on how research links with innovation and if the distance between basic research, applications and products reduces [190].

It is worth remarking that we are mainly interested in the capacity of the FP5 to create and transfer information and nothing can be said about this issue inside each node. Notice that some participants are large institutions or companies with complex organization charts, which may have several projects whose coordination cannot be guaranteed in general. However, our main concern is how to set the means to integrate research, development and innovation efficiently, not if these means are successfully used.

3.2 Analysis of the data

To characterize the FP5, in this section we compute five important features in any network: degree distribution, shortest path length distribution, betweenness centrality, clustering coefficient and the degree-degree correlation. The description of these properties is given in Section 1.2. More details about the network dataset can be found in the Appendix A.

[1] Community Research and Development Information Service: `http://cordis.europa.eu`

[2] Open Source programming language: `http://www.perl.org/`

3.2.1 Degree distribution

The question wether empirical distributions are or are not power laws is still object of study [191, 192] despite the many situations of scientific interest where they occur and of their significant implications on the phenomena under study. Many times it is safer to report heterogeneous distributions (or heavy tailed distributions) instead of reporting power laws. Here we will refer to the observed distributions as power laws, even though sometimes this may be questionable. In this way, we find that the probability that a University collaborates with k other Universities (i.e., the degree distribution of the Universities) decays as a power law, $P(k) \sim k^{-\gamma_U}$ with $\gamma_U = 1.76$. Similarly, Companies follow a power law with $\gamma_C = 2.76$. The two distributions can be seen in Fig. 3.1, where a log-log scale is used in the plot, providing evidence for the scale-free topology [15] of both networks. The degree distribution of the whole FP5 network is also well approximated by a power law with exponent γ close to 2.1.

Note that the degree distribution of Universities is described by a power law with $\gamma_U < 2$, implying that their mean degree grows in time. Indeed the first moment (i.e. mean degree in this case) of a distribution with a power–law tail diverges when its exponent is less than 2. This result suggests that Universities form an accelerated growing network [43, 193], where the total number of edges grows faster than a linear function of the total number of vertices and, consequently, it is verified that $1 < \gamma < 2$.

To elucidate this issue, we computed the average degree $\langle k \rangle$ during several years to check its tendency. Though we only have the data corresponding to 4 years (table 3.1), they are enough to confirm the existence of an accelerated growth since the average degree is not constant (46% increase for the network of Universities in the four year period). But if the collaborations grow faster than proportional to the number of participants, it is because they do not emerge by the mere increase of participants. Not only new participants contribute to increase the number of collaborations, but also the old ones, meaning that some form of synergy exists encouraging the creation of new collaborations between Universities.

On the other hand, the average degree of Companies also grows (though significantly slower) during the four year span of the dataset (table 3.1). However, the fact that $\gamma_C > 2$ suggests that this increase should be transient. Therefore, although the creation of collaborations is encouraged

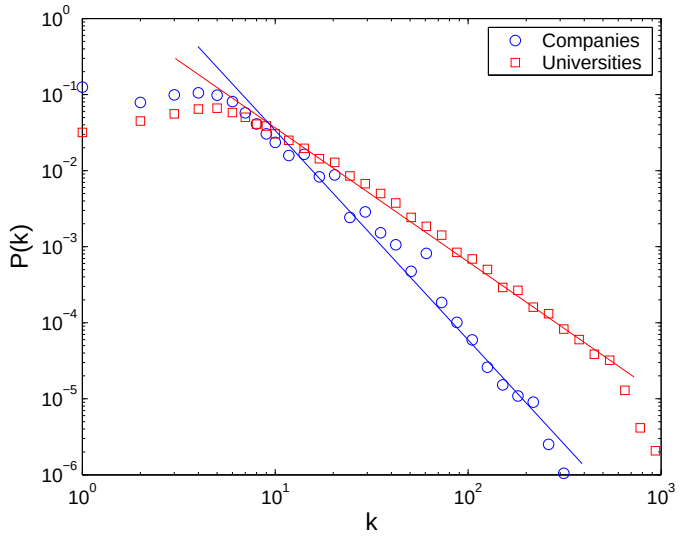

Figure 3.1: This graph depicts with red squares the probability that a University collaborates with k other Universities, that is, its degree distribution. The degree distribution of Companies is shown with blue circles. Data were log-binned. We find that both distributions follow a power law tail, $P(k) \sim k^{-\gamma}$, thus having a scale-free topology, with vertices connecting each other in a heterogeneous manner: Most vertices have few connections, but some have a very large degree. The best fit for the straight region of the curves gives $\gamma_U = 1.76 \pm 0.01$ with a correlation coefficient $R = 0.998$ for Universities, and $\gamma_C = 2.76 \pm 0.03$ with $R = 0.991$ for Companies. However, the fact that Universities show $\gamma_U < 2$ whereas Companies have $\gamma_C > 2$ implies that the mean degree of Universities grows in time but not the mean degree of Companies. This result suggests that some form of synergy encourages the creation of new collaborations mainly between Universities, while the network of Companies is less dynamic in this respect.

(since when the FP5 was finished the mean number of collaborations had risen from 10 to 26 and some participants had surpassed 2,500 collaborations) these results reveal that the synergy is more pronounced between Universities. In this sense, the FP5 is less effective in improving the network of Companies and Universities seem to take more advantage of this opportunity to create new collaborations.

Also noticeable in table 3.1 is the fact that the number of Companies increases faster than the number of Universities (72% and 64% increase respectively in the four year period), indicating another difference in the evolution of both networks.

	N	$\langle k \rangle$	$\langle C \rangle$
Year	Univ–Comp	Univ–Comp	Univ–Comp
1999	3075–4658	17.2–6.2	0.65–0.58
2000	5377–9359	21.9–6.8	0.66–0.53
2001	7355–13905	27.7–7.9	0.67–0.53
2002	8522–16765	31.9–8.2	0.68–0.59

Table 3.1: Evolution of Universities and Companies during the FP5. Here we show the total number of vertices N, the average degree $\langle k \rangle$ and the average clustering coefficient $\langle C \rangle$ during the four years that the FP5 lasted.

3.2.2 Shortest paths

The distance between vertices is the number of edges in the shortest path which links them (Section 1.2.2). Defining the set of participants which can be linked through a path as a connected component, we find that the largest connected component of Universities spans 93.7% of the network (7,987 vertices) while for Companies it is made of 10,801 nodes (64.4%). Hence, while almost all Universities are linked in only one component, Companies are more fragmented and one third of them fall in other smaller components (actually, the second biggest component contains only 48 participants). This result shows that Universities are important to compact the network since the largest connected component of the complete network (U+C) comprises 88.7% of the Companies and 96.0% of the Universities (i.e. 23,055 vertices in total). In addition, the largest distance in the

network of Universities is 7 and the average distance is $\bar{\ell} = 3.34$ whereas, in the case of Companies, the farthest pair is separated by 14 edges and the average distance is[3] $\bar{\ell} = 5.67$. This can be seen in Fig. 3.2 where we plot the geodesic distribution, $\mathcal{P}(\ell)$ versus ℓ. Hence, also here Universities are essential for Companies since the largest distance in the entire network is only 8 and the average distance is $\bar{\ell} = 3.14$, which implies that, on average, there are only two intermediaries between two participants.

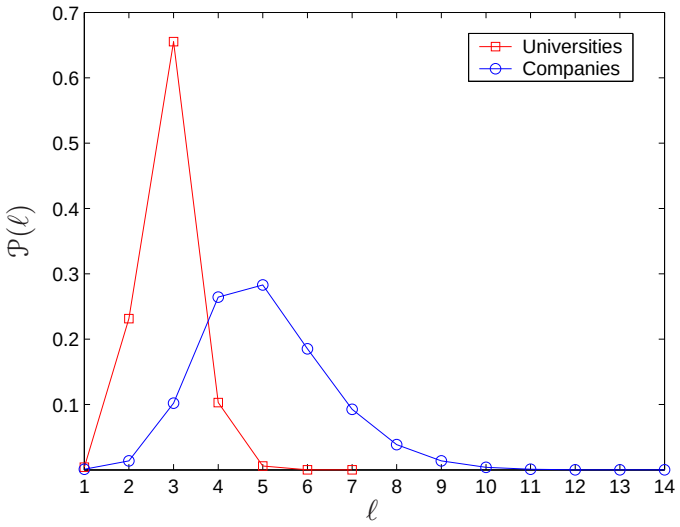

Figure 3.2: The distribution of shortest paths in the largest connected component of Universities (red squares) and Companies (blue circles) displays the presence of the small-world effect. The mean value is $\bar{\ell} = 3.34$ for Universities and $\bar{\ell} = 5.67$ for Companies. Moreover, while the farthest pair of Companies has 13 intermediaries, for Universities the maximum separation is 7 edges. Therefore, Universities are important for Companies since, when they cooperate, in the whole FP5 network the largest distance reduces to 8 and the average distance to 3.14.

The average distance is a coarse characteristic though. As a finer measure, it is possible to compute the average distance of a vertex of degree k to all other vertices in the largest component [194]. In Fig. 3.3 we plot $\ell(k)$ for both networks on a log-linear scale.

Therefore, albeit both networks display the so-called small-world effect [41], there are important

[3]Both average distances are approximately the value obtained for a random graph [10] with the same number of nodes and average degree. For Universities is $\bar{\ell} \approx \log N / \log\langle k \rangle = 2.61$ and for Companies is $\bar{\ell} = 4.62$.

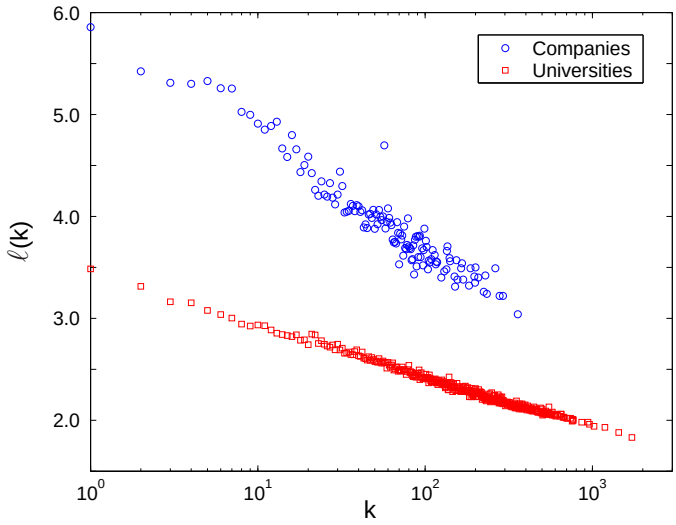

Figure 3.3: The average distance of a participant with k partners to all other participants in the largest connected component is depicted. Universities are the red squares and Companies are the blue circles. It can be seen the logarithmic dependence since it is verified that $\ell(k) \sim -\beta \log k$ where $\beta_U = 0.503 \pm 0.003$ with $R = 0.994$ for Universities and $\beta_C = 1.13 \pm 0.03$ with $R = 0.958$ for Companies. The decay is faster (*i.e.* $\beta_C > \beta_U$) in the net with the larger value of exponent γ (see Fig. 3.1), providing empirical evidence for the network models of Section 2.1. Note that the lowest degree vertices in the network of Universities show a distance to other vertices comparable to the one of the highest degree vertices in the network of Companies. Also note that in both networks $\max \ell(k) \approx 2 \min \ell(k)$ as had been previously observed in Section 2.1, Ref. [194].

differences. The presence of Universities eases the flow of information since they are much closer to each other than Companies. This could be expected since the main purpose of a company is to satisfy its shareholders, which does not include the spread of information from which competitors can take advantage. But, interestingly, the consequences of this fact go beyond. When Universities are excluded from the projects, Companies become isolated despite Universities are only one third of the participants. Companies tend to form clusters, turning difficult (if not impossible) the communication between them and, consequently, little can be developed or innovated since other results are not available to work with. Thus the natural tendency of Companies to protect their findings would finish killing R+D+I. The presence of Universities contributes to moderate this.

3.2.3 Betweenness centrality

To further investigate the interplay between the two kinds of participants, we can also measure the betweenness centrality (1.2.6 Eq. 1.13) in the FP5.Since its computation for the whole FP5 is an extremely time-consuming task, we focus our study on one of its subprograms: 'Promotion of innovation and encouragement of small and medium sized enterprises participation' (SME), which is formed by 195 research institutions and 212 Companies (see Appendix). Given our ability to split the SME into Universities and Companies, several different situations are considered. The average betweenness of the SME, taken over all its vertices, turns out to be $\langle\sigma\rangle = 5.19 \cdot 10^{-3}$. Considering only those vertices m which are Universities, we find that their average betweenness among all other vertices in the SME is $\langle\sigma_U\rangle = 6.76 \cdot 10^{-3}$. Likewise, we obtain $\langle\sigma_C\rangle = 3.74 \cdot 10^{-3}$ for Companies.

Now, if we only take into account those shortest paths whose endpoints are Companies, the betweenness measures the role Universities play in linking Companies: $\langle\sigma_{CUC}\rangle = 5.44 \cdot 10^{-3}$; on the other hand, when the endpoints are Universities, the average betweenness of Companies is $\langle\sigma_{UCU}\rangle = 2.34 \cdot 10^{-3}$. Thus, we see that the role Universities play between Companies is more than twice the one played by Companies between Universities. Moreover, given that $\langle\sigma_U\rangle > \langle\sigma\rangle > \langle\sigma_C\rangle$, we observe again the central function played by research institutions in the FP5 network.

3.2.4 Clustering coefficient

The *clustering coefficient* of a vertex i is defined as $C_i = 2n_i/[k_i(k_i - 1)]$, where n_i is the number of edges connecting its k_i nearest neighbors (Section 1.2.3). It equals 1 for a participant at the center of a completely connected cluster, and 0 for a node whose neighbors are not linked at all. Taking the average of the clustering coefficient, we obtain $\langle C \rangle = 0.68$ for Universities and $\langle C \rangle = 0.59$ for Companies, which are much higher than the average clustering coefficient of a random graph [10] with the same number of nodes and average degree (namely, $\langle C \rangle = \langle k \rangle / N$). Moreover, $\langle C \rangle$ is independent of the number N of participants in both cases (see table 3.1), in contrast with the prediction of a scale-free model [15] where $\langle C \rangle \sim N^{-0.75}$ [42, 64]. This high and size-independent average clustering coefficient evidences the organization of Universities and Companies in modules.

However, when we measure the clustering coefficient of a node with k links, $C(k)$, for both networks (Fig. 3.4), we find that it decays as a power law for large k. We therefore infer that the two nets have hierarchical modularity, which is characterized by the scaling law $C(k) \sim k^{-\alpha}$, in contrast to some scale-free or modular networks where the clustering coefficient is degree-independent [39].

This result suggests that Universities and Companies have an inherent self-similar structure [195], being made of many highly connected small modules, which integrate into larger modules, which in turn group into even larger modules (Fig. 3.5A). Actually, we observe that 4,333 Universities (50.8%) and 10,564 Companies (63.5%) have $C_i = 1$, indicating the presence of many totally connected groups. This is due to the fact that most of these entities participate in only one project, having as neighbors other vertices, which in turn are all connected between them by virtue of the participation in the project. Furthermore, given that this result suggests weak geographical constraints [196], we searched for communities in them [197] and found precisely that they were not based on nationality (Fig. 3.5B), whence, the FP is successfully applying a policy which avoids its segregation by nationality.

3.2.5 Degree-degree correlations

An interesting question is which vertices pair up with which others. It may happen that vertices connect randomly, no matter how different they are. Usually, however, there is a selective linking, i.e. there is some feature which makes more (or less) likely the connection (see Section 1.2.4).

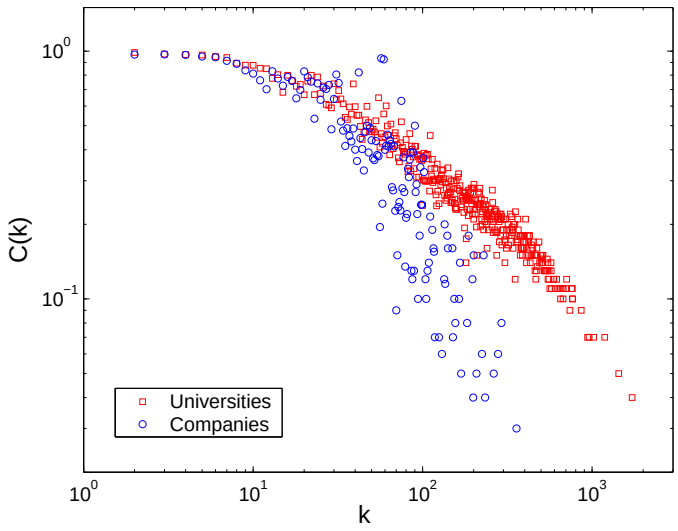

Figure 3.4: In this graph the clustering coefficient as a function of k is shown. After the initial plateau, where $C(k)$ is approximately constant, it approximately decays as a power law, $C(k) \sim k^{-\alpha}$, where $\alpha_U = 0.54 \pm 0.01$ with $R = 0.97$ for Universities (red squares) and $\alpha_C = 1.05 \pm 0.06$ with $R = 0.86$ for Companies (blue circles). We therefore conclude that both networks have hierarchical modularity since scale-free and modular networks are degree-independent, whereas hierarchical modularity is characterized by the power-law decay $C(k) \sim k^{-\alpha}$.

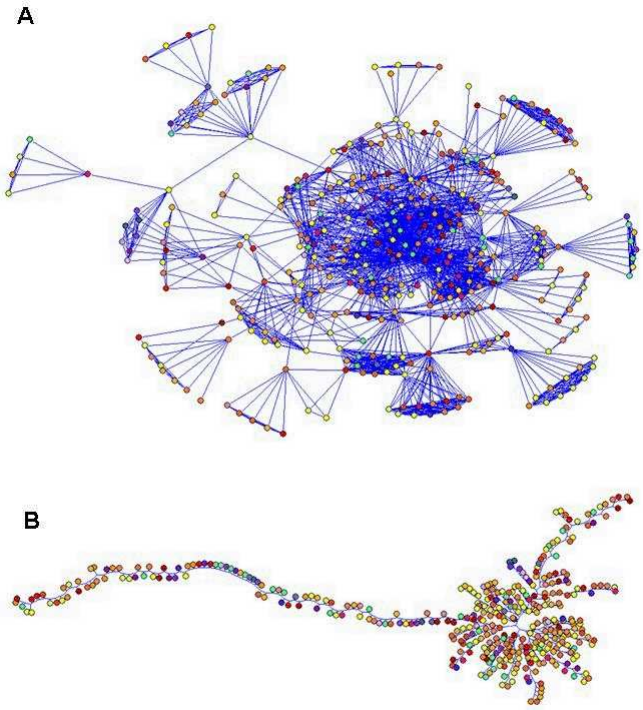

Figure 3.5: The existence of hierarchical modularity in the networks of Universities and Companies suggests that they have a self-similar structure. Since projects in the FP are classified in 8 subprograms depending on their objectives, we choose, for clarity, to illustrate in **A** this self-similar structure with the smallest one: 'Promotion of innovation and encouragement of small and medium sized enterprises participation' (SME)—see Appendix A. Also, to verify if there is a bias by nationality in the collaborations, we searched for communities reflecting groups of participants collaborating strongly among them. In the networks of Universities, Companies and both together (even when they are analyzed by subprogram) the result was similar to **B,** corresponding to the SME subprogram. If we color the nodes according to their nationalities and arrange them in space with a free for noncommercial use, standard algorithm software [198], we find that they are all mixed.

A first approach to elucidate this issue is by means of the *joint degree-degree distribution* $P(k, k')$, which gives us the probability of finding an edge connecting vertices of degree k and k'. We see that for Companies the distribution has sharp peaks for $k = k'$ (Fig. 3.6A). This network thus seems to display assortative mixing, i.e. if one chooses at random a vertex of degree k then, with considerable probability, it will be connected to vertices of degree k. In other words, Companies with similar degree tend to collaborate more frequently than Companies with different degrees.

Notice that the fact (mentioned in the previous section) that many entities participate in only one project may, by itself, explain these peaks: If the X participants of a certain project have no other projects each of them has degree $X - 1$ and each of their neighbors has degree $X - 1$, giving rise to an assortative trend. On the other hand one can also argue that, when a Company has high degree it is due to being involved in many projects. It is then reasonable to assume that nodes with high degree represent large institutions, given that only these can deal with many projects at the same time. That being the case, the observed assortativity means that the spread of information between Companies depends on the institution's size. On the contrary, for Universities $P(k, k')$ is scattered throughout the plane $k - k'$ (Fig. 3.6B). While there are still peaks along the line $k = k'$, the presence of many others for $k \neq k'$ is clear, suggesting that Universities are less selective in what regards the size of their partners.

It is important to remark, however, that the joint degree-degree distribution requires many observations in order to obtain good statistics. For example, if we focus our analysis in the range $[0, 200]$, we need about 200×200 points, otherwise fluctuations are important and the plot is far from smooth [199]. To avoid this problem, one uses the average degree of the nearest neighbors of a vertex of degree k, $\langle k \rangle_{nn}(k)$, which is a coarser but less fluctuating measure. To compute it, we find all participants with k links and take the average degree of all their neighbors. The results are shown in Fig. 3.7, and confirm those obtained through the joint degree-degree distributions. To emphasize the presence of the cut-off due to the finite size of the network, the points obtained from less than 10 observations are plotted as crosses (Universities in red and Companies in blue) and the rest of the points as squares (Universities) or circles (Companies). Considering then only the circles and the squares, we confirm that collaborations between Companies are size-dependent (positive slope) whereas those between Universities are not (no slope).

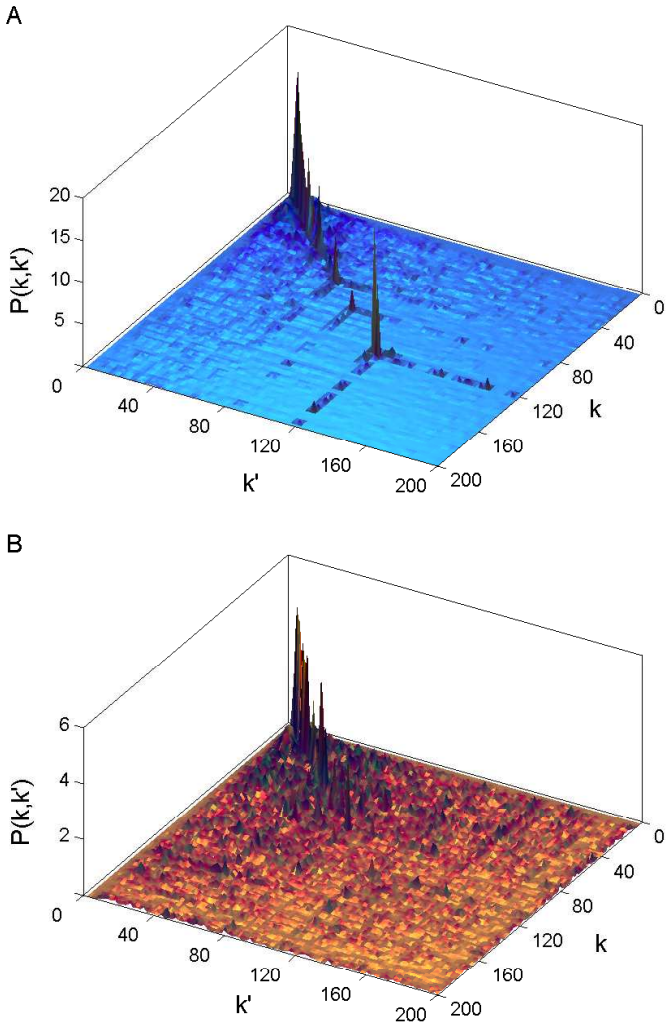

Figure 3.6: Determination of the mixing through the joint degree-degree distribution. The X and Y axes represent the degrees k and k' and the Z axis gives the corresponding joint degree-degree probability in per mill. The range is limited from 0 to 200 to illustrate a clearer picture. The joint degree-degree distribution of Companies (**A**) peaks on the line $k = k'$ which implies that the mixing is assortative. Since the number of links held by a participant is related to its size, we infer that Companies with similar sizes tend to collaborate more frequently than Companies with different sizes. The joint degree-degree distribution of Universities (**B**) is distributed throughout the X-Y plane which suggests that Universities do not have assortative mixing and thus choose their collaborators in a less selective manner.

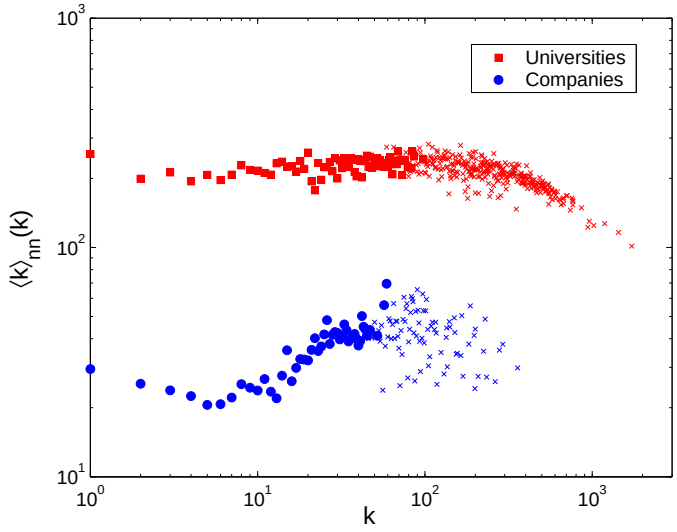

Figure 3.7: In this plot the average degree of the nearest neighbors of a vertex with k links, $\langle k \rangle_{nn}(k)$, is shown. To mark the proximity to the cut-off, the points obtained from less than 10 observations are plotted as crosses (Universities in red and Companies in blue) and the remaining points as squares (Universities) or circles (Companies). In this manner, it can be seen that these points are biased downwards due to the finite size of the network. Then, once focusing our attention on the circles and the squares, we find that Companies have assortative mixing, while Universities link between them regardless their degrees.

It is also interesting to analyze how Universities and Companies link each other, which can be done as follows. We search for all Companies with k links and then compute the average degree of all their neighboring Universities. Note that the former degrees are always calculated in the corresponding network, thus a Company with degree k has k neighbor Companies, though it may have more links (to Universities) in the complete FP5 network. Analogously, we can find all Universities with k links to average the degrees of all neighbor Companies. The results are depicted in Fig. 3.8 where, as before, it is used a $\log - \log$ scale. Again, we plot as squares (Universities) or circles (Companies) the points obtained from more than 10 observations to identify the region where the tendency is well defined. We find that, while Companies link to Universities independently of their sizes, Universities with high degree tend to collaborate with large Companies.

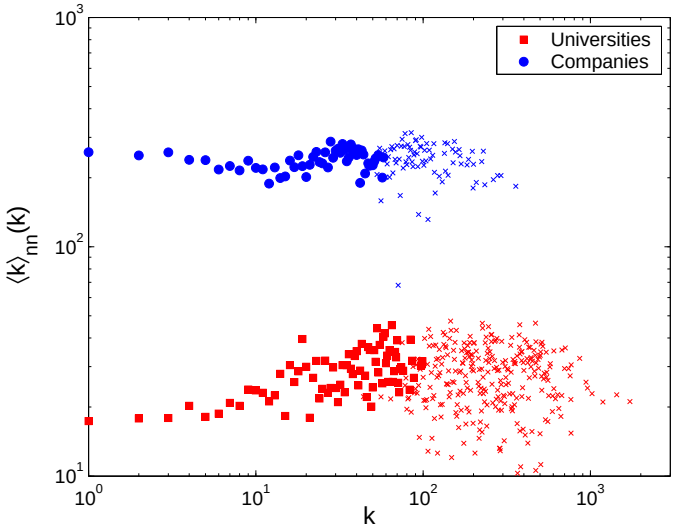

Figure 3.8: Here we plot the average degree of the nearest Companies of a University with k links to other Universities (red squares) and the average degree of the nearest Universities of a Company with k links to other Companies (blue circles). As before, if we only consider the circles and the squares, we find that Companies link to Universities independently of their degrees while Universities with high degree collaborate mainly with Companies which have also high degree.

Finally, another way to quantify the mixing in the FP5 is by means of the *assortativity coefficient* [77], which is just the Pearson correlation coefficient of the degrees of connected vertices. In this case, we obtain what type of mixing takes place in the network by means of a single number

instead of a distribution. If e_{jk} is the probability that a randomly chosen edge has vertices with degree j and k at either end, the assortativity coefficient takes the following form:

$$r = \frac{\sum_{jk} q_j q_k (e_{jk} - q_j q_k)}{\sum_k k^2 q_k - \left(\sum_k k q_k\right)^2}$$

where $q_k = \sum_j e_{jk}$ and $q_j = \sum_k e_{jk}$. This coefficient verifies that $-1 \leq r \leq 1$, being positive when the network is assortative and negative when it is disassortative. We find $r_C = 0.13$ for the network of Companies and $r_U = 0.06$ for Universities, corroborating an assortative trend usual in social networks [78].

Therefore, Companies and Universities differ in the way they establish collaborations. Companies are organized hierarchically, where positions in that hierarchy are related to the size: The assortative trend in the network of Companies suggests that large corporations are reluctant to choose as partners small companies. Between Universities, however, size is not important and it is common to find a large institution collaborating with a small one. But if we analyze which partners Universities choose among Companies, we check that large institutions in Universities prefer working with large Companies. On the contrary, Companies select their collaborators between Universities regardless of their sizes. We can then conclude that large Companies play indeed a leading role in the FP5 while Universities play the role of bridges between participants which are separated in the hierarchical structure of Companies.

3.3 Discussion

We have presented here a study of the interplay between research and industry in the scope of the Fifth Framework Programme. Using network theory methods, we perform several measures that allow us to quantify the features of this relationship and assess their potential improvements. Naturally, the FP5 network does not include all interactions between university and industry (such as the recruitment of graduates by companies, the transfer of knowledge through scientific and technical literature or industry conferences). Furthermore, as already mentioned in Section 3.1, it also neglects the fact that internal connections in an institution (e.g. between different departments) may be absent, which would mean that a node in the studied network would split into disconnected nodes. While these issues may significantly influence the flow of information in the network, addressing all

of them requires information that is beyond reach for most researchers at this point. The presented analysis thus represents a starting point for a quantitative understanding of the university-industry interplay network. It is possible, however, to foresee advances in these directions, given the increasing availability of information on how institutions self-organize.

The results point to the central function played by Universities in the FP5 network in reducing the distance between research and applications. Indeed, we show that Universities play a crucial role in connecting the network of Companies, which would otherwise be separated in many small clusters. While the network of Universities is well integrated and established in accordance to what is observed for other social networks, the same doesn't seem to apply for the Companies network, mainly due to its relatively small largest connected component. Competition is probably the origin of this effect, which is moderated by the presence of Universities. It seems reasonable, then, to conclude that special attention should be devoted to company-company collaborations. Supporting this, is also the fact that new collaborations arise at a higher rate between Universities.

Our observations suggest in addition that Companies and Universities establish collaborations differently: While Companies seem to exhibit a hierarchical structure in terms of their size, Universities are less selective in their collaborations. We also observed that both networks display hierarchical modularity and that communities in the FP5 network are not nation-based. The FP appears then to mix all nationalities of the European Union, thus reaching one of its main goals: Promote the transfer of knowledge throughout Europe.

Chapter 4

Frequency of numbers on the World Wide Web

4.1 Introduction

The distribution of numbers in human documents is determined by a variety of diverse natural and human factors, whose relative significance can be evaluated by studying the numbers' frequency of occurrence. Although it has been studied since the 1880's, this subject remains poorly understood. Here, we obtain the detailed statistics of numbers in the World Wide Web, finding that their distribution is a heavy-tailed dependence which splits in a set of power-law ones. In particular, we find that the frequency of numbers associated to western calendar years shows an uneven behavior: 2004 represents a 'singular critical' point, appearing with a strikingly high frequency; as we move away from it, the decreasing frequency allows us to compare the amounts of existing information on the past and on the future. Moreover, while powers of ten occur extremely often, allowing us to obtain statistics up to the huge 10^{127}, 'non-round' numbers occur in a much more limited range, the variations of their frequencies being dramatically different from standard statistical fluctuations. These findings provide a view of the array of numbers used by humans as a highly non-equilibrium and inhomogeneous system, and shed a new light on an issue that, once fully investigated, could lead to a better understanding of many sociological and psychological phenomena.

4.2 Motivation

Already in the early 1880's, Newcomb [200] noticed a specific uneven distribution of the first digits of numbers, which is now known as Benford's law [114]. The observed form of this distribution indicates the wide, skewed shape of the frequency of occurrence of numbers in nature [201, 202, 203] — as an illustration, note that in these first two sentences the numerals 114, 200, 201, 202, 203 and 1880 all occur twice. Benford's law is directly derived by assuming that a number occurs with a frequency inversely proportional to it, meaning that the frequencies of numbers in the intervals $(1, 10)$, $(10, 100)$, $(100, 1000)$, etc. are equal. Yet, this assumption lacks convincing quantitative support and understanding, in part due to scanty data available. In our days, this problem can be tackled by resorting (with the help of search engines) to the enormous database constituted by the World Wide Web.

One should note that the profoundly wide form of the distribution of numbers in human documents is determined by two sets of factors. The first includes general natural reasons of which the most important is the multi-scale organization of our World. The second are 'human factors' including the current technological level of the society, the structure of languages, adopted numeral and calendar systems, history, cultural traditions and religions, human psychology, and many others. By analyzing the occurrence frequency of numbers we can estimate the relative significance and role of these factors.

4.3 Frequency of Numbers on the Web

The frequency of occurrence of numbers in the World Wide Web (or simply Web) necessarily reflects the distribution of numbers in all human documents, allowing us to effectively study their statistics by using search engines, which supply the approximate number of web pages (or web documents) containing the Arabic numeral that we are looking for. In this respect, the Web provides us with huge statistics. Yet, the frequencies of occurrence of distinct kinds of numbers are very different [204]: for example, one can see that 777 and 1000 occur much more frequently than their neighbors (Table 4.1). Here we report on the markedly distinct statistics of different types of natural numbers

Table 4.1: **Typical numbers with high frequencies of occurrence**

Example	Description
1000	powers of 10
2460, 2465	'round' numbers: multiples of 10 and 5
777,171717	numbers easy to remember or symmetric
$512 = 2^9$	powers of 2
666,777	numbers with strong associations
78701	popular zip codes
866, 877	toll free telephone numbers
1812	important historical dates
747, 8086	serial numbers of popular products
314159	beginning parts of mathematical constants

(or, rather, positive integers) in the Web documents, collected[1] through the currently most popular search engine, Google [205]. We consider separately (i) powers of 10 and (ii) non-round integers, and find that in both of these cases, the number $N(n)$ of pages containing an integer n decays as a power law, $N(n) \sim n^{-\beta}$, over many orders of magnitude. The observed values of the β exponent strongly differ for the different types of numbers, (i) and (ii), and also differ from 1, thus contradicting the above mentioned assumption of inverse proportionality for their frequency of occurrence.

Note that, previously, scale-free (*i.e.* power-law) distributions were observed for processes in the WWW [22, 24] and its structural characteristics [23, 15]. However, and in contrast to these studies, we use the WWW as a database for measuring one of the basic distributions in nature. In order to explain the observed distributions, we treat the global array of numbers as a non-equilibrium, evolving system with a specific influx of numbers, and, as a reflection of this non-equilibrium nature, we find a 'critical behavior' of $N(n)$ in the neighborhood of $n = 2004$ (the current year at the time the measurements were made): near this point, the frequency of WWW documents follows a power law, $N(n) \sim (2005 - n)^{-\alpha}$.

[1]The data was collected by using a Linux shell script together with the open source text web browser Lynx available at `http://lynx.isc.org/`.

Finally, we show that the statistics of variations of the frequencies of WWW pages which contain close numbers of the same kind, dramatically disagrees with the standard distribution of statistical fluctuations. We observe, namely, that the amplitude of these variations, $\delta N(n)$, is much greater than what would be expected for standard statistical fluctuations. Consequently, the frequencies of pages containing different numbers fluctuate not independently, these fluctuations being a reflection of those of the influx of numbers.

4.4 Current-year Singularity

In the second week of December 2004, we obtained the frequency of WWW documents corresponding to positive integers n in the range between 1 and 100,000 (Fig. 4.1a). This plot contains a set of regularly distributed peaks, which indicate that different types of numbers occur with very unlike frequencies. For example, the number of documents containing round (ending with 0) numbers is much higher than that for non-round numbers. Furthermore, the special number 2004 occurs with a remarkably high frequency: 3,030,000,000 pages. For comparison, among 8,058,044,651 WWW pages covered by the used search engine, a single character string a occurs in about 8,000,000,000 pages, while the numbers 0, 1 and 1000 occur in 2,180,000,000, 4,710,000,000 and 154,000,000 pages, respectively. The high, asymmetric peak of $N(n)$ around $n = 2004$ (Fig. 4.1b) is naturally identified as the contribution of documents containing numbers associated to years; below $n = 2005$, this peak can be fitted by a power law, following $N(n) \sim (2005 - n)^{-\alpha}$, where $\alpha = 1.2 \pm 0.1$ (inset of Fig. 4.1b). Therefore, in the vicinity of 2004, $N(n)$ increases with n much faster than the total number of pages in the WWW grows with time, which indicates that there are many pages with numbers associated to years that disappear from the WWW (or at least, are updated) after a while. Indeed, our observations prove that the amount of pages holding a number $n < t$ (where t is time measured in years) in the region of the 'critical singularity' decreases with t approximately following $N(n, t) \sim (t - n)^{-\alpha}$.

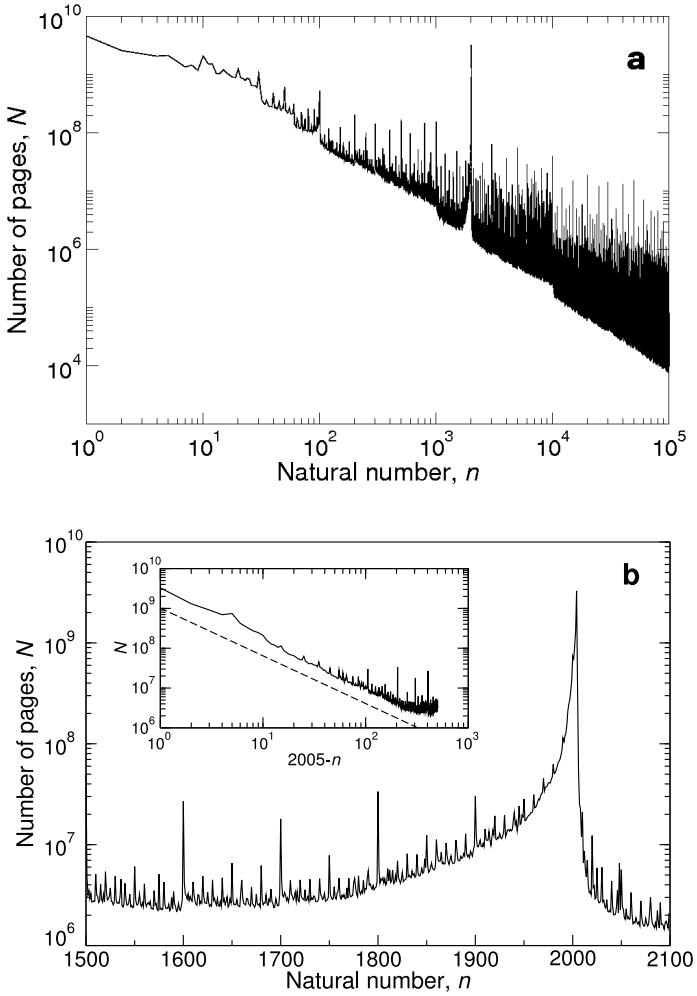

Figure 4.1: **a,** The frequencies $N(n)$ of WWW pages containing numbers n up to 100,000 on a log-log plot. Note the peak at $n = 2004$. **b,** The part of the distribution around $n = 2004$ shown in more detail on a log-linear plot. The asymmetric form of the peak gives an idea about the relation between the stored volumes of information on the past and on the future: the former is much more referred to than the latter. In the inset, the low-n part of this peak is plotted versus the difference $2005 - n$ on a log-log plot ($1500 < n < 2005$). A power-law behavior is observed practically in the entire range where the contribution of numbers associated to years is main. The slope of the dashed line is -1.2. It was not possible to find a reliable fit to the dependence for $n \geq 2005$. These plots also demonstrate a hierarchy of peaks for documents holding numbers of different kinds.

4.5 Power-law Distributions

We find that the frequency of occurrence of natural numbers, considered without separating them into distinct classes (Fig. 4.1a), is a slowly decreasing dependence. Nevertheless, it can hardly be fitted by any power law because it is, in fact, the result of the superposition of distributions of distinct kinds of numbers, which, in turn, are power laws having different exponents. In order to proceed, we then compare the statistics of the WWW documents which hold two 'extreme' types of numbers: (i) powers of 10, which should occur with the highest frequencies due to the common decimal numeral system, and, contrastingly, (ii) non-round numbers (i.e. those with a non-zero digit in the end) which are, on average, the most indistinctive ones, therefore occurring with the lowest frequencies. It is worth remarking that, even though the non-round include many peculiar numbers, such as 777 for example, we find that their contribution does not change the statistics noticeably.

The strikingly high frequency of occurrence of powers of 10 in the WWW allows us to obtain the statistics for numbers up to 10^{127} (Fig. 4.2a), a range that is restricted by the limited size of strings being accepted by the used search engine (128 characters)[2]. Two distinct regions are seen in the distribution. The region of relatively 'small' numbers, up to 10^{11} (Fig. 4.2b), is of a power-law form, $N(n) \sim n^{-\beta}$, where $\beta = 0.50 \pm 0.02$, hence close to the law $N(n) \sim 1/\sqrt{n}$; note that this exponent is much smaller than 1 and far smaller than the values of the exponents of typical Zipf's law distributions [15, 206], these being mostly in the range between 2 and 3. For comparison, the occurrence frequencies of a character string $baaa\ldots a$ of varying length were also measured, a quite different, far from straight line, dependence having been observed (Fig. 4.2c). For n larger than 10^{11}, we observe an extremely slow decrease of the frequency of occurrence of pages containing powers of 10 (Fig. 2a). It is worth noting that the crossover between these two regimes turns out to be rather close to the maximum 32 digit binary number, which is about 0.4×10^{10}.

For properly measuring the occurrence frequency of non-round numbers, we use a set of intervals selected in their wide range, each of which having a width of 50 numbers, so that the relative variation of the frequency of WWW pages inside a specific interval is sufficiently small. In addition, these

[2]Search engines find the number of pages containing a given positive integer in the WWW and not the total number of times this integer occurs in the Web. This difference is not essential in our study, since we are mostly interested in the tail of the distribution. Indeed, the probability that a large number occurs several times in the same page is low.

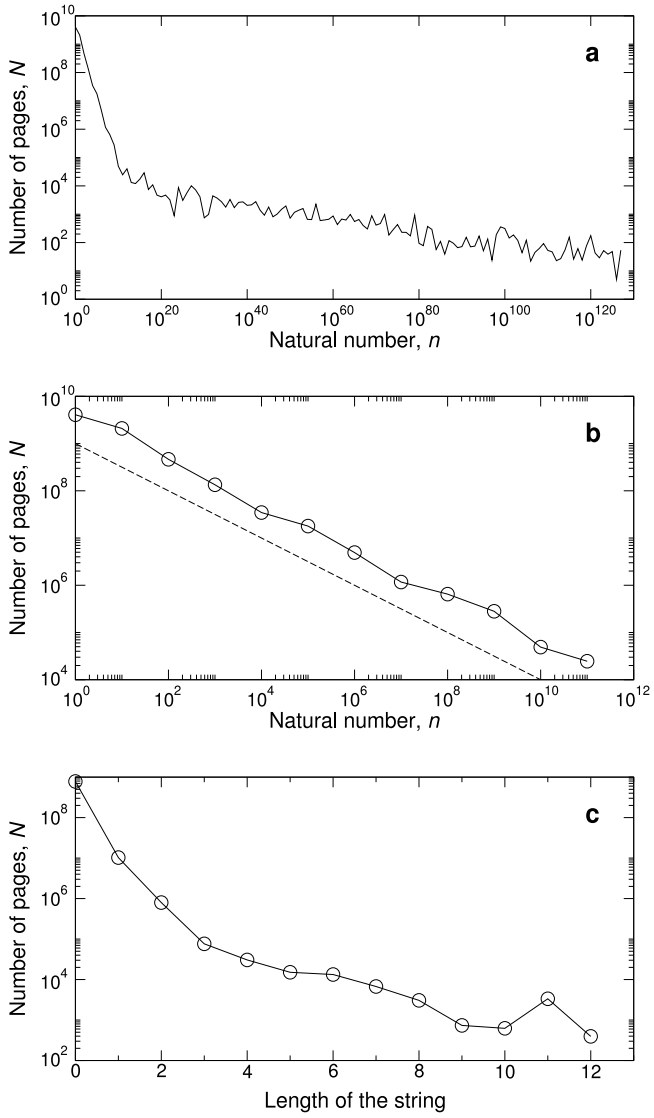

Figure 4.2: The frequencies of Web pages containing powers of 10. **a,** The full log-log plot up to the maximal searchable 10^{127}. **b,** The power-law-like part of the distribution. The slope of the dashed line is -0.5. We emphasize that the power-law dependence is observed over 11 orders of magnitude, which is a uniquely wide range. **c,** For comparison, the number of WWW documents containing a character string $baaa\ldots a$ of varying length on a log-linear plot (the length of the string is the equivalent to the exponent in the power of 10). Note the difference from **b**.

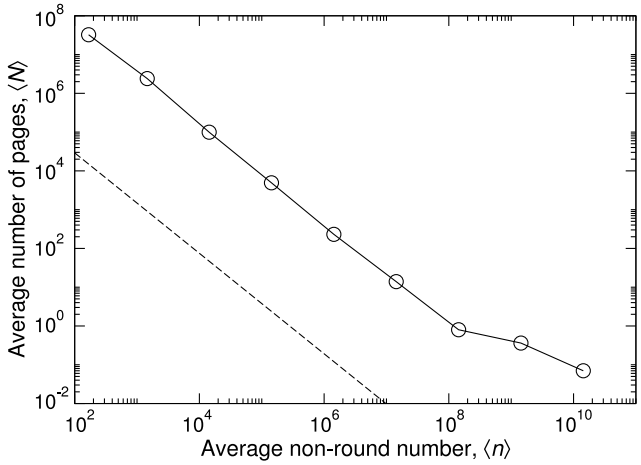

Figure 4.3: Log-log plot of the frequencies of WWW pages holding non-round numbers. The circles show the average amounts of pages with non-round numbers taken from relatively narrow intervals (50 numbers). Each interval is centered at the $\langle n \rangle$ coordinate of a circle. The dashed line has slope -1.3. Note that the power-law behavior is observed over 6 orders of magnitude. Non-round numbers occur much less frequently than powers of 10, which explains the essentially narrower range of numbers in this plot than in Fig. 4.2a. For instance, presently, and as far as search engines report, there are no WWW documents with the number 12345789014.

intervals are chosen far from the powers of 10, whose close neighborhood includes numbers, such as, for instance, 1009, that occur more often and whose distribution does not follow a clear power law. Within each of these intervals, we take the average values of n and $N(n)$, and denote them by $\langle n \rangle$ and $\langle N \rangle$, respectively; the resulting dependence (Fig. 4.3) has a prominent power-law region with exponent $\beta = 1.3 \pm 0.05$, which strongly differs from that ascertained for powers of 10. As numbers grow, the ratio of the amount of WWW documents with powers of 10 to that with non-round numbers increases, following the $n^{0.8}$ dependence.

A few mechanisms generating power-law distributions [206] are known [8, 154, 207, 208, 209]. Most of these mechanisms explain power laws as a result of a specific self-organization of a non-equilibrium system, and we treat our observations in the spirit of these approaches. Evidently, the array of numbers in human documents is an evolving system, and the stochastic growth of this

array is due to a permanent influx of numbers, added with new documents. The added numbers (among which may also occur new distinct ones, that were not employed previously) are chosen from a distribution which is determined by the one for the existing numbers. Here we do not discuss a specific model exploiting this mechanism and generating the observed complex distributions[3], but instead, we explain the reason for the unusual small values of exponents which we observed — $\beta = 0.5$ and 1.3 (Figs. 4.2b and 4.3), while typical Zipf's law exponents are 2 and greater. At least, Zipf's law exponents must take values greater than 1. At first sight, this difference seems surprising, since the mechanisms of the power laws are quite similar. But, importantly, these two sets of exponents are defined for different distributions. In our non-traditional case, the observed power law describes the behavior of the frequency of WWW pages with a given natural number n, namely $N(n) \sim n^{-\beta}$. In contrast, typical Zipf's law exponent γ occurs in a power law for a quite different quantity: in our terms, this quantity is the amount, $m(N)$, of distinct numbers, where each of them occurs in every of N Web pages. So, we have the relation $m(N) \sim N^{-\gamma}$. One can show that the exponents β and γ satisfy a simple relation, $\beta = 1/(\gamma - 1)$ [48]. As a result, if the γ exponent is greater than 2, which is typical for simple linear growth processes, the β exponent is smaller than 1, as in Fig. 4.2b. On the other hand, nonlinear growth may produce exponents γ below 2, which gives β greater than 1, as in Fig. 4.3.

4.6 Fluctuations of the Number of WWW Pages

The distributions reported here demonstrate that the frequencies of WWW pages holding numbers even of the same kind (for example, non-round numbers) strongly fluctuate from number to number. For documents containing non-round integers, we obtain the dependence of the fluctuations' amplitude (i.e. dispersion), $\sqrt{\langle (N - \langle N \rangle)^2 \rangle} = \sqrt{\langle N^2 \rangle - \langle N \rangle^2}$, on the average frequency, $\langle N \rangle$, of these documents (Fig. 4.4). For calculating these dispersions and mean values, we used the same intervals as in Fig. 4.3. The resulting dependence turns out to be proportional, $\sqrt{\langle N^2 \rangle - \langle N \rangle^2} \approx 0.1 \langle N \rangle$, over a broad region of values $\langle N \rangle$, which crucially differs from the square root behavior of standard

[3]Without knowing the details of the evolution of the global array of numbers, one can only propose a class of evolutionary stochastic models with unknown parameters. So, we cannot calculate the observed values of the exponents but can explain the range of these values.

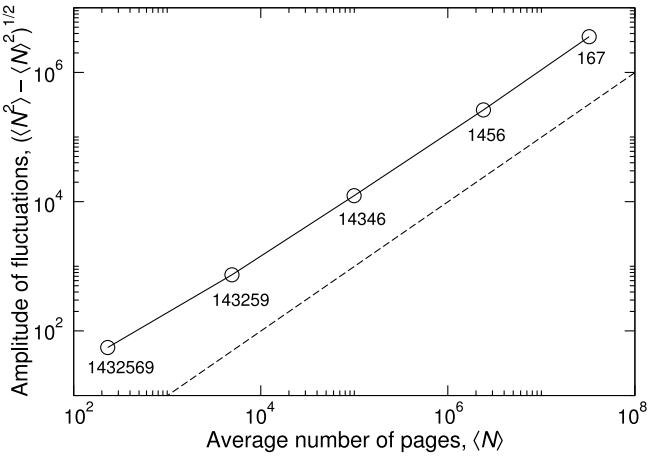

Figure 4.4: Amplitude of the fluctuations, $\sqrt{\langle N^2\rangle - \langle N\rangle^2}$, of the frequencies of pages containing non-round numbers versus their mean values, $\langle N\rangle$, on a log-log plot. The data (circles) were obtained resorting to the same intervals as in Fig. 4.3. Next to each circle the average (non-round) number, $\langle n\rangle$, for the corresponding interval is indicated. The dashed line has slope 1. One can see that $\sqrt{\langle N^2\rangle - \langle N\rangle^2} \approx 0.1\langle N\rangle$ for $\langle N\rangle > 10^3$.

statistical fluctuations [210]. The usual reason for such a strong difference is that the fluctuations of the quantities under study are not statistically independent [211, 212]. In this respect, there is only one factor in the evolution of the array of numbers which can break the statistical independence of fluctuations, namely, the variation of the influx of numbers. So, the observed proportional law proves that the variations of the occurrence frequencies of numbers are an outcome of the fluctuations of their global influx in the WWW.

4.7 Discussion and Conclusions

These observations suggest a new view of the array of integers in the WWW (and in nature) as a complex, evolving, inhomogeneous system. The statistics of numbers turns out to be far more rich and complex than one might expect based on classical Benford's law. Moreover, our findings provide a tool for extracting meaningful information from statistical data on the frequency of occurrence of

numbers. As an illustration, consider the two integers, 666 and 777, with clear associations. We find that these numbers occur in the WWW with frequencies of 11,800,000 and 13,600,000 pages, respectively, which are 1.25 and 1.65 times higher than, on average, the occurrence frequencies of their non-round neighbors. These deviations are to a great extent higher than what one would anticipate from the relative amplitude of fluctuations, 0.1. Therefore, we can reasonably compare the amounts of pages containing 666 and 777 obtained after subtracting the numbers of pages holding the neighbors of these two integers. These subtractions give 2,400,000 and 5,400,000 pages for 666 and 777, respectively. It is the difference (or, rather, the relative difference) between the two last amounts that should be used as a starting point for a subsequent comparative analysis. The proposed approach is very suggestive. Indeed, by analyzing the frequencies of occurrence of specific 'popular' numbers with clear interpretations one could evaluate the relative significance of the corresponding underlining factors of this popularity.

Many more questions lie ahead: How do the occurrence frequencies of specific numbers vary in time? How do different numbers correlate and co-occur in WWW documents? It is well known that humans can easily memorize only up to rather limited sequences of digits [213, 214], which are, therefore, many times replaced by words (like, for instance, the IP addresses of computers). Then, how does the statistics of numbers relate to the organization of human memory and to semantics? Our findings quantitatively show the key role of the common decimal numeral system — a direct consequence of the number of fingers. How do other numeral systems (the binary system, for example) influence the general statistics of numbers?

The global array of numbers is surmised to be a "numeric snapshot of the collective consciousness" [204]. So, the study of their statistics could lead to a better understanding of a wide circle of sociological and psychological phenomena. The distribution of numbers in human documents contains a wealth of diverse information in an integrated form. The detailed analysis of the general statistics of numbers in the WWW could allow the effective extraction and evaluation of this hidden information.

Chapter 5

Timing of human dynamics

5.1 Introduction

Humans participate on a daily basis in a large number of distinct activities, from electronic communication, such as sending emails or browsing the web, to initiating financial transactions or engaging in entertainment and sports. Given the number of factors that determine the timing of each action, ranging from work and sleep patterns to resource availability, it appears impossible to seek regularities in the apparently random human activity patterns, apart from the obvious daily and seasonal periodicities. Therefore, in contrast with the accurate predictive tools common in physical sciences, forecasting human and social patterns remains a difficult and often elusive goal. Yet, the need to understand the timing of human actions is increasingly important. Indeed, uncovering the laws governing human dynamics in a quantitative manner is of major scientific interest, requiring us to address the factors that determine the timing of human actions. But these questions are driven by applications as well: most human actions have a strong impact on resource allocation, from phone line availability and bandwidth allocation in the case of Internet or Web use, all the way to the design of physical space for retail or service oriented institutions. Despite these fundamental and practical driving forces, our understanding of the timing of human initiated actions is rather limited at present [215].

The interest in addressing the timing of events in human dynamics is not new: it has a long history in the mathematical literature, leading to the development of some of the key concepts

in probability theory [60], and has reemerged at the beginning of the 20$^{\text{th}}$ century as the design problems surrounding the phone system required a quantitative understanding of the call patterns of individuals. But most current models of human activity assume that human actions are performed at constant rate, meaning that a user has a fixed probability to engage in a specific action within a given time interval. These models approximate the timing of human actions with a Poisson process, in which the time interval between two consecutive actions by the same individual, called the waiting or inter-event time, follows an exponential distribution [216] (Eq. 5.3 in the next Section). Poisson processes are the base of the celebrated Erlang formula [217],

$$E(q,c) = \frac{q^c}{c!}\left(\sum_{i=0}^{c}\frac{q^i}{i!}\right)^{-1}, \tag{5.1}$$

predicting the number of phone lines, c, required in an institution, and where E is the fraction of callers that find all lines full and q is the number of calls starting per unit time (*i.e.* the Poisson process rate, see Figs. 5.1a-c in the next Section). Also, they represent the basic approximation in the design of most currently used Internet protocols and routers [218]. Yet, the availability of large datasets recording selected human activity patterns increasingly question the validity of the Poisson approximation. Indeed, an increasing number of recent measurements indicate that the timing of many human actions systematically deviate from the Poisson prediction, the waiting or inter-event times being better approximated by a heavy tailed or Pareto distribution [219, 220, 221, 222]. The difference between a Poisson and a heavy tailed behavior is striking: the exponential decay of a Poisson distribution implies that the consecutive events follow each other at relatively regular time intervals and forbids very long waiting times. In contrast, the slowly decaying heavy tailed processes allow for very long periods of inactivity that separate bursts of intensive activity.

It has been recently proposed by Barabási that the bursty nature of human dynamics is a consequence of a queuing process driven by human decision making [219]: whenever an individual is presented with multiple tasks and chooses among them based on some perceived priority parameter, the waiting time of the various tasks will be Pareto distributed. In contrast, first-come-first-serve and random task execution, common in most service oriented or computer driven environments, lead to a uniform Poisson-like dynamics. Yet, this work has generated just as many questions as it resolved. What are the different classes of processes that are relevant for human dynamics? What

determines the scaling exponents? Do we have discrete universality classes (and if so how many) as in critical phenomena [223], or the exponents characterizing the heavy tails can take up arbitrary values, as it is the case in network theory [42, 43, 49]? Is human dynamics always heavy tailed?

In this chapter we aim to address some of these questions by studying the different universality classes that can appear as a result of the queuing of human activities. We first review, in Section 5.2, the frequently used Poisson approximation, which predicts an exponential distribution of interevent times. In Section 5.3 we present evidence that the interevent time probability density function (pdf) $P(\tau)$ of many human activities is characterized by the power law tail

$$P(\tau) \sim \tau^{-\alpha} \ . \tag{5.2}$$

In Section 5.4 we discuss the general characteristics of the queueing models for how humans time their various activities. In Sections 5.5-5.6 we study two classes of queuing models designed to capture human activity patterns. We find that restrictions on the queue length play an important role in determining the scaling of the queuing process, allowing us to document the existence of two distinct universality classes, one characterized by $\alpha = 3/2$ (Section 5.5) and the other by $\alpha = 1$ (Section 5.6). In Section 5.7 we discuss the relationship between interevent and waiting times. Finally, in Section 5.8 we discuss the applicability of these models to explain the empirical data, as well as outline future challenges in modeling human dynamics.

5.2 Poisson processes

Consider an activity performed with some regularity, such as sending emails, placing phone calls, visiting a library, or browsing the web. We can keep track of this activity by recording the timing of each event, for example the time each email is sent by an individual. The time between two consecutive events we call the *interevent time* for the monitored activity and will be denoted by τ. Given that the interevent time can be explicitly measured for selected activities, it serves as a test of our ability to understand and model human dynamics: proper models should be able to capture its statistical properties.

The most primitive model of human activity would assume that human actions are fundamentally periodic, with a period determined by the daily sleep patterns. Yet, while certain periodicity is

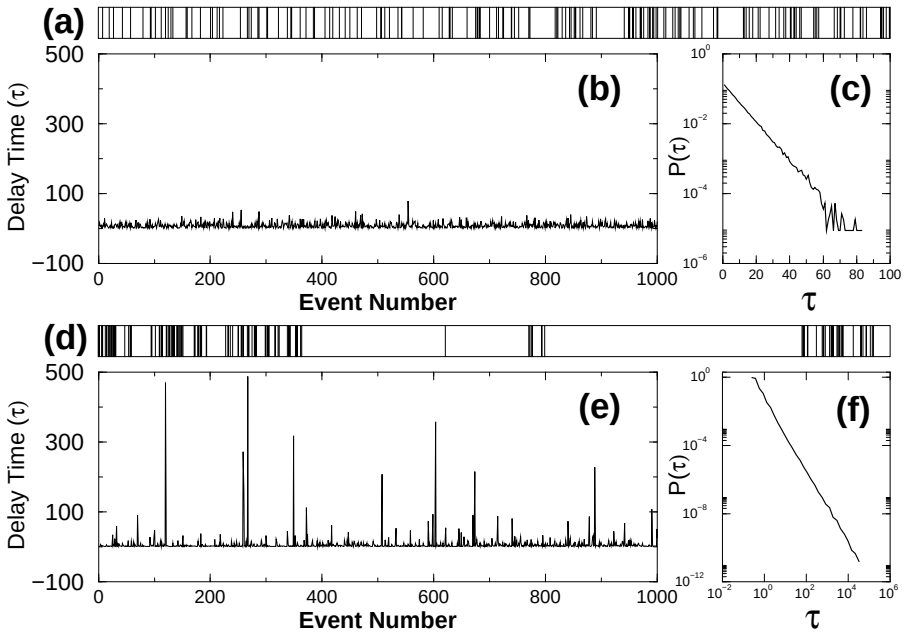

Figure 5.1: The difference between the activity patterns predicted by a Poisson process (top) and the heavy tailed distributions observed in human dynamics (bottom). **(a)** Succession of events predicted by a Poisson process, which assumes that in any moment events take place with probability q. The horizontal axis denotes time, each vertical line corresponding to an individual event. Note that the interevent times are comparable to each other, long delays being virtually absent. **(b)** The absence of long delays is visible on the plot showing the delay times τ for 1,000 consecutive events, the size of each vertical line corresponding to the gaps seen in (a). **(c)** The probability of finding exactly n events within a fixed time interval is $\mathcal{P}(n;q) = e^{-qt}(qt)^n/n!$, which predicts that for a Poisson process the inter-event time distribution follows $P(\tau) = qe^{-q\tau}$, shown on a log-linear plot in (c) for the events displayed in (a, b). **(d)** The succession of events for a heavy tailed distribution. **(e)** The waiting time τ of 1,000 consecutive events, where the mean event time was chosen to coincide with the mean event time of the Poisson process shown in (a-c). Note the large spikes in the plot, corresponding to very long delay times. (b) and (e) have the same vertical scale, allowing to compare the regularity of a Poisson process with the bursty nature of the heavy tailed process. **(f)** Delay time distribution $P(\tau) \simeq \tau^{-2}$ for the heavy tailed process shown in (d,e), appearing as a straight line with slope -2 on a log-log plot.

certainly present, the timing of most human actions are highly stochastic. Indeed, periodic models are hopeless in capturing the time we check out a book from the library, beyond telling us that it should be within the library's operation hours. The first and still most widely used stochastic model of human activity assumes that the tasks are executed independently from each other at a constant rate λ, so that the time resolved activity of an individual is well approximated by a Poisson process [216]. In this case the probability density function (pdf) of the recorded interevent times has the exponential form

$$P(\tau) = \lambda e^{-\lambda\tau} \ . \tag{5.3}$$

In practice this means that the predicted activity pattern, while stochastic, will display some regularity in time, events following each other on average at $\tau \approx \langle\tau\rangle = 1/\lambda$ intervals. Indeed, given that for a Poisson process $\sigma = \sqrt{\langle\tau^2\rangle - \langle\tau\rangle^2} = \langle\tau\rangle$ is finite, very long waiting times (*i.e.* large temporal gaps in the sequence of events) are exponentially rare. This is illustrated in Fig. 5.1a, where we show a sequence of events generated by a Poisson process, appearing uniformly distributed in time (but not periodic).

The Poisson process was originally introduced by Poisson in his major work applying probability concepts to the administration of justice [224]. Today it is widely used to quantify the consequences of human actions, such as modeling traffic flow patterns or accident frequencies [216], and is commercially used in call center staffing [225], inventory control [226], or to estimate the number of congestion caused blocked calls in mobile communications [218]. It has been established as a basic model of human activity patterns at a time when data collection capabilities on human behavior were rather limited. In the past few years, however, thanks to detailed computer based data collection methods, there is increasing evidence that the Poisson approximation fails to capture the timing of many human actions.

5.3 Empirical results

Evidence that non-Poisson activity patterns characterize human activity has first emerged in computer communications, where the timing of many human driven events is automatically recorded.

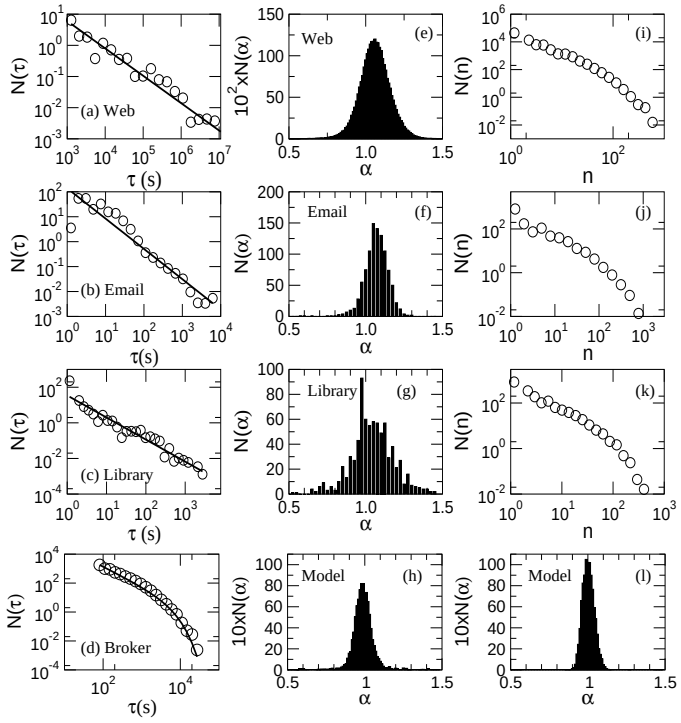

Figure 5.2: (a) The interevent time distribution between (a) two consecutive visits of a website by a single user; (b) two consecutive library loans made by a single individual; (c) two consecutive emails sent out by a user. For (a-c) we show as a straight line the $\alpha = 1$ scaling. (d) The interevent time distribution between two consecutive transactions made by a stock broker. The distribution follows a power-law with the exponential cut-off $P(\tau) \sim \tau^{-1.3} \exp(-\tau/\tau_0)$. (e-g) The distribution of the exponents (α) characterizing the interevent time distribution of users browsing the website (e), individual loans from the library (f) and the emails sent by different individuals (g). The exponent α was determined only for users whose total activity levels exceeded certain thresholds, the values used being 15 web visits (e), 15 emails (f) and 10 books (g). (h,l) We numerically generate for 10,000 individuals interevent time distributions following a power-law with exponent $\alpha = 1$. The distribution of the measured exponents follows a normal distribution similar to the distribution observed in (e-g). If we double the time window of the simulation (h) the deviation around the average becomes much smaller (l). (i-k) The distribution of the *number of events* in the studied systems: number of HTML hits for each user (i), the number of books checked out by each user (j) and the number of emails sent by different individuals (k), indicating that the overall activity patterns of individuals is also heavy tailed.

For example, measurements capturing the distribution of the time differences between consecutive instant messages sent by individuals during online chats [227] have found evidence of heavy tailed statistics. Professional tasks, such as the timing of job submissions on a supercomputer [228], directory listings and file transfers (FTP requests) initiated by individual users [229] were also reported to display non-Poisson features. Similar patterns emerge in economic transactions [230, 231], in the number of hourly trades in a given security [232] or the time interval distribution between individual trades in currency futures [233]. Finally, heavy tailed distributions characterize entertainment related events, such as the time intervals between consecutive online games played by users [234]. Note, however, that while these datasets provide clear evidence for non-Poisson human activity patterns, most of them do not resolve individual human behavior, but capture only the aggregated behavior of a large number of users. For example, the dataset recording the timing of job submissions looks at the timing of *all jobs* submitted to a computer, by any user. Thus for these measurements the interevent time does not characterize a *single* user but rather a *population* of users. Given the extensive evidence that the activity distribution of the individuals in a population is heavy tailed, these measurements have difficulty in capturing the origin of the observed heavy tailed patterns. For example, while most people send only a few emails per day, a few send a very large number on a daily basis [116, 27].

If the activity pattern of a large number of users is simultaneously captured, it is not clear where the observed heavy tails come from: are they rooted in the activity of a single individual, or rather in the heavy tailed distribution of user activities? Therefore, when it comes to our quest to understand human dynamics, datasets that capture the long term activity pattern of a *single* individual are of particular value such as the timing of printing jobs submitted by users [235] or the activity patterns of individual email users [116]. These measurements offer direct evidence that the heavy tailed activity patterns emerge at the level of a *single* individual, and are not a consequence of the heterogeneous distribution of user activity. Despite this evidence, a number of questions remain unresolved: Is there a single scaling exponent characterizing all users, or rather each user has its own exponent? What is the range of these exponents? Next we aim to address these questions through the study of six datasets, each capturing individual human activity patterns of different nature. First we describe the datasets and the collection methods, followed by a quantitative characterization of

the observed human activity patterns.

Web browsing: Automatically assigned cookies allow us to reconstruct the browsing history of approximately 250,000 unique visitors of the largest Hungarian news and entertainment website (origo.hu), which provides online news and magazines, community pages, software downloads, free email and search engine, capturing 40% of all internal Web traffic in Hungary [222, 236]. The site receives 6,500,000 HTML hits on a typical workday. We used the log files of the site to collect the visitation pattern of each visitor between 11/08/02 and 12/08/02, recording with second resolution the timing of each download by each visitor [222]. The interevent time, τ, was defined as the time interval between consecutive page downloads (clicks) by the same visitor.

Email activity patterns: This dataset contains the email exchange between individuals in a university environment, capturing the sender, recipient and the time of each email sent during a three and six month period by 3,188 [116] and 9,665 [27] users, respectively. We focused here on the data collected by Eckmann [116], which records 129,135 emails with second resolution. The interevent time corresponds to the time between two consecutive emails sent by the same user.

Library loans: The data contains the time with second resolution at which books or periodicals were checked out from the library by the faculty at the University of Notre Dame during a three year period. The number of unique individuals in the dataset is 2,247, together participating in a total of 48,409 transactions. The interevent time corresponds to the time difference between consecutive books or periodicals checked out by the same patron.

Trade transactions: A dataset recording all transactions (buy/sell) initiated by a stock broker at a Central European bank between 6/1999 and 5/2003 helps us quantify the professional activity of a single individual, giving a glimpse on the human activity patterns driving economic phenomena. In a typical day the first transactions start at 7AM and end at 7PM and the average number of transactions initiated by the dealer in one day is around 10, resulting in a total of 54,374 transactions. The interevent time represents the time between two consecutive transactions by the broker. The gap between the last transaction at the end of one day and the first transaction at the beginning of the next trading day was ignored.

The correspondence patterns of Einstein, Darwin and Freud: We start from a record containing the sender, recipient and the date of each letter [237, 238, 239] sent or received by the three scientists

during their lifetime. The databases used in our study were provided by the Darwin Correspondence Project[1], the Einstein Papers Project[2] and the Freud Museum of London[3]. Each dataset contains the information about each sent/received letter in the following format: SENDER, RECIPIENT, DATE, where either the sender or the recipient is Einstein, Darwin or Freud. The Darwin dataset contained a record of a total of 7,591 letters sent and 6,530 letters received by Darwin (a total of 14,121 letters). Similarly, the Einstein database contained 14,512 letters sent and 16,289 letters received (total of 30,801). For Freud we have 3,183 (2,675) sent (received) leters. Note that 1,541 letters in the Darwin database and 1,861 letters in the Einstein database were not dated or were assigned only potential time intervals spanning days or months. We discarded these letters from the dataset. Furthermore, the dataset is naturally incomplete, as not all letters written or received by these scientists were preserved. Yet, assuming that letters are lost at a uniform rate, they should not affect our main findings. For these three datasets we do not focus on the interevent times, but rather the *response* or *waiting times* τ_w. The waiting time, τ_w, represents the time interval between the date of a letter received from a given person, and the date of the next letter from Darwin, Einstein or Freud to him or her, *i.e.* the time the letter waited on their desk before a response was sent. To analyze Einstein, Darwin, and Freud's response time we have followed the following procedure: if individual A sent a letter to Einstein on DATE1, we search for the next letter from Einstein to individual A, sent on DATE2, the response time representing the time difference $\tau_w = \text{DATE2} - \text{DATE1}$, expressed in days. If there are multiple letters from Einstein to the recipient, we always consider the first letter as the response, and discard the later ones. Missing letters could increase the response time, the magnitude of this effect depending on the overall frequency of communication between the respective correspondence partners. Yet, if the response time would follow a distribution with an exponential tail, then randomly distributed missing letters would not generate a power law waiting time: they would only shift the exponential waiting times to longer average values. Thus the observed power law cannot be attributed to data incompleteness.

In the following we will break our discussion in three subsections, each focusing on a specific class of behavior observed in the studied individual activity patterns.

[1] http://www.lib.cam.ac.uk/Departments/Darwin/

[2] http://www.einstein.caltech.edu/

[3] http://www.freud.org.uk

5.3.1 The $\alpha = 1$ universality class: Web browsing, email, and library datasets

In Fig. 5.2a-c we show the interevent time distribution between consecutive events for a single individual for the first four studied databases: Web browsing, email, and library visitation. For these datasets we find that the interevent time distribution has a power-law tail

$$P(\tau) \sim \tau^{-\alpha} \tag{5.4}$$

with exponent $\alpha \approx 1$, independent of the nature of the activity. Given that for these activity patterns we collected data for thousands of users, we need to calculate the distribution of the exponent α determined separatelly for each user whose activity level exceeds a certain threshold (*i.e.* avoiding users that have too few events to allow a meaningful determination of $P(\tau)$). As Fig. 5.2e-g shows, we find that the distribution of the exponents is peaked around $\alpha = 1$.

The scattering around $\alpha = 1$ in the measured exponents could have two different origins. First, it is possible that each user is characterized by a different scaling exponent α. Second, each user could have the same exponent $\alpha = 1$, but given the fact that the available dataset captures only a finite time interval from one month to several months, with at best a few thousand events in this interval, there are uncertainties in our ability to determine numerically the exponent α. To demonstrate that such data incompleteness could indeed explain the observed scattering, in Figs. 5.2h and 5.2l we show the result of a numerical experiment, in which we generated 10,000 time series, corresponding to 10,000 independent users, the interevent time of the events for each user being taken from the same distribution $P(\tau) \sim \tau^{-1}$. The total length in time of each time series was chosen to be $1,000,000$. We then used the automatic fitting algorithm employed earlier to measure the exponents in Figs. 5.2e-g to determine numerically the exponent α for each user. In principle for each user we should observe the same exponent $\alpha = 1$, given that the datasets were generated in an identical fashion. In practice, however, due to the finite length of the data, each numerically determined exponent is slightly different, resulting in the histogram shown in Fig. 5.2h. As the figure shows, even in this well controlled situation we observe a scattering in the measured exponents, obtaining a distribution similar to the one seen in Figs. 5.2e-g. The longer the time series, the sharper the distribution is (Fig. 5.2l), given that the exponent α can be determined more accurately.

The distributions obtained for the three studied datasets are not as well controlled as the one used in our simulation: while the length of the observation period is the same for each user, the activity level of the users differs widely. Indeed, as we show in Fig. 5.2i-k, the activity distribution of the different users, representing the number of events recorded for each user, also spans several orders of magnitude, following a fat tailed distribution. Thus the degree of scattering of the measured exponent α is expected to be more significant than seen in Fig. 5.2h and l, since we can determine the exponent accurately only for very active users, for which we have a significant number of datapoints. Therefore, the obtained results are consistent with the hypothesis that each user is characterized by a scaling exponent in the vicinity of $\alpha = 1$, the difference in the numerically measured exponent values being likely rooted in the finite number of events we record for each user in the datasets. This conclusion will be corroborated by our modeling efforts, that indicate that the exponents characterizing human behavior take up discrete values, one of which providing the empirically observed $\alpha = 1$.

As we will see in the following sections, an important measure of the human activity patterns is the waiting time, τ_w, representing the amount of time a task waits on an individual's priority list before being executed. For the email dataset, given that we know when a user receives an email from another user and the time he sends the next email back to her, we can determine the email's waiting or response time. Therefore, we define the waiting time as the difference between the time user A receives an email from user B, and the time A sends an email to user B. In looking at this quantity we should be aware of the fact that not all emails A sends to B are direct responses to emails received from B, thus there are some false positives in the data that could be filtered out only by reading the text of each email (which is not possible in the available datasets).

5.3.2 The $\alpha = 3/2$ universality class: The correspondence of Einstein, Darwin and Freud

In the case of the correspondence patterns of Einstein, Darwin and Freud we will focus on the response time of the authors, partly because we will see later that this has the most importance from the modeling perspective. As shown in Fig. 5.3, the probability that a letter will be replied to in τ_w days is well approximated by a power law (Eq. 5.4) with $\alpha = 3/2$, the scaling spanning

four orders of magnitude, from days to years. Note that this exponent is significantly different from $\alpha = 1$ observed in the earlier datasets, and we will show later that modeling efforts indeed establish $\alpha = 3/2$ as a scaling exponent characterizing human dynamics.

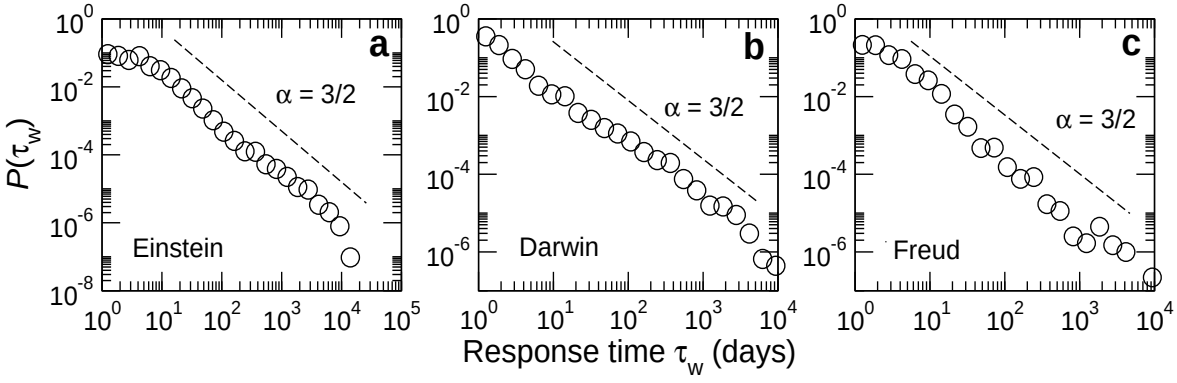

Figure 5.3: Distribution of the response times for the letters replied to by Einstein, Darwin and Freud, as indicated on each plot. Note that the distributions are well approximated with a power law tail with exponent $\alpha = 3/2$. While for Darwin and Einstein the datasets provide very good statistics (the power law regime spanning 4 orders of magnitude), the plot corresponding to Freud's responses is not so impressive, yet still being well approximated by the power law distribution. Note that while in most cases the identified reply is indeed a response to a received letter, there are exceptions as well: many of the very delayed replies represent the renewal of a long lost relationship.

The dataset allows us to determine the interevent times as well, representing the time interval between two consecutive letters sent by Einstein, Darwin or Freud to any recipient. We find that the interevent time distribution is also heavy tailed, albeit the quality of scaling is not as impressive as we observe for the response time distribution. This is due to the fact that we do not know the precise time when the letter is written (in contrast with the email, which is known with second resolution), but only the day on which it was mailed. Given that both Einstein and Darwin wrote at least one letter most days, this means that long interevent times are rarely observed. Furthermore, owing to the long observational period (over 70 years), the overall activity pattern of the two scientists has changed significantly, going from a few letters per year to as many 400-800 letters/year during the later, more famous phase of their professional life. Thus the interevent time, while it appears to follow a power law distribution, it is by no means stationary. On the contrary, the observed response

time distribution is stationary.

5.3.3 The stock broker activity pattern

For the stock broker we again focus on the interevent time distribution, finding that the best fit follows $P(\tau) \sim \tau^{-\alpha} \exp(-\tau/\tau_0)$ with $\alpha = 1.3$ and $\tau_0 = 76$ min (see Fig. 1d). This value is between $\alpha = 1$ observed for the users in the first three other datasets and $\alpha = 3/2$ observed for the correspondence patterns. Yet, given the scattering of the measured exponents, it is difficult to determine if this represents a standard statistical deviation from $\alpha = 1$ or $\alpha = 3/2$, the two values expected by the modeling efforts (see Sections 5.5 and 5.6), or it stands as evidence for a new universality class. At this point we believe that the former case is valid, something that can be decided only once data for more users will become available[4]. The exponential cutoff is not inconsistent with the modelling efforts either: as we will show in Appendix B.2, such cutoffs are expected to accompany all human activity patterns with $\alpha < 2$.

5.3.4 Qualitative differences between heavy tailed and Poisson activity patterns

The heavy tailed nature of the observed interevent time distribution has clear visual signatures. Indeed, it implies that an individual's activity pattern has a bursty character: short time intervals with intensive activity (bursts) are separated by long periods of no activity (Figs. 5.1d-f). Therefore, in contrast with the relatively uniform activity pattern predicted by the Poisson process, for a heavy tailed process very dense successions of events (bursts) are separated by very long gaps, predicted by the slowly decaying tail of the power-law distribution. This bursty activity pattern agrees with our experience of an individual's normal email usage pattern: during a single session we typically send several emails in quick succession, followed by long periods of no email activity, when we focus on other activities.

[4]However, see the next Chapter for a possible explanation.

5.4 Capturing human dynamics: queuing models

The empirical evidence discussed in the previous Section raises several important questions: Why does the Poisson process fail to capture the temporal features of human activity? What is the origin of the observed heavy tailed activity patterns in human dynamics? To address these questions we need to inspect closely the processes that contribute to the timing of the events in which an individual participates.

Most of the time humans face simultaneously several work, entertainment, and family related responsibilities. Indeed, at any moment an individual could choose to participate in one of several tasks, ranging from shopping to sending emails, making phone calls, attending meetings or talks, going to a theater, getting tickets for a sports event, and so on. To keep track of the various responsibilities ahead of them, individuals maintain a *to do* or *priority* list, recording the upcoming tasks. While this list is occasionally written or electronically recorded, in many cases it is simply kept in memory. A priority list is a dynamic entity, since tasks are removed from it after they are executed and new tasks are added continuously. The tasks on the list compete with each other for the individual's time and attention. Therefore, task management by humans is best described as a *queuing process* [240, 241], where the queue represents the tasks on the priority list, the server is the individual which executes them and maintains the list, and some selection protocol governs the order in which the tasks are executed. To define the relevant queuing model we must clarify some key features of the underlying queuing process, ranging from the arrival and service processes to the nature of the task selection protocol, and the restrictions on the queue length [240]. In the following we discuss each of these ingredients separately, placing special emphasis on their relevance to human dynamics.

Server: The server refers to the individual (or agent) that maintains the queue and executes the tasks. In queuing theory we can have one or several servers in parallel (like checkout counters in a supermarket). Human dynamics is a *single server* process, capturing the fact that an individual is solely responsible for executing the tasks on his/her priority list[5].

Task Arrival Pattern: The arrival process specifies the statistics of the arrival of new tasks to the queue. In queuing theory it is often assumed that the arrival is a Poisson process, meaning that

[5]However interactions between individuals may influence the execution of tasks, see next Chapter.

new tasks arrive at a constant rate λ to the queue, randomly and independently from each other. We will use this approximation for human queues as well, assuming that tasks land at random times on the priority list. If the arrival process is not captured by a Poisson distribution, it can be modeled as a renewal process with a general distribution of interarrival times [240]. For example, our measurements indicate that the arrival time of emails follows a heavy tailed distribution, thus a detailed modeling of email based queues must take this into account. We must also keep in mind that the arrival rate of the tasks to the list is filtered by the individual, who decides which tasks to accept and place on the priority list and which to reject. In principle the rejection of a task is also a decision process that can be modeled as a high priority short lived task.

Service process: The service process specifies the time it takes for a single task to be executed, such as the time necessary to write an email, explore a web page or read a book. In queuing theory the service process is often modeled as a Poisson process, which means that the distribution of the time devoted to the individual tasks has the exponential form (5.3). However, in some applications the service time may follow some general distribution. For example, the size distribution of files transmitted by email is known to be fat tailed [242, 243], suggesting that the time necessary to review (read) them could also follow a fat tailed distribution. In queuing theory it is often assumed that the service time is independent of the task arrival process or the number of tasks on the priority list. While we adopt this assumption here as well, we must also keep in mind that the service time can decrease if too many tasks are in the queue, as humans may devote less time to individual tasks when they have many things to do.

Selection protocol or queue discipline: The selection protocol specifies the manner in which the tasks in the queue are selected for execution. Most human initiated events require an individual to weigh and prioritize different activities. For example, at the end of each activity an individual needs to decide what to do next: send an email, do some shopping or place a phone call, allocating time and resources for the chosen activity. Normally individuals assign to each task a priority parameter, which allows them to compare the importance of the different tasks on the list. The time a task waits before it is executed depends on the method the agent uses to choose the task to be executed next. In this respect three selection protocols are particularly relevant for human dynamics:

(*i*) The simplest is the first-in-first-out (FIFO) protocol, executing the tasks in the order they

were added to the list. This is common in service oriented processes, like the first-come-first-serve execution of orders in a restaurant or getting help from directory assistance and consumer support.

(*ii*) The second possibility is to execute the tasks in a random order, irrespective of their priority or time spent on the list. This is common, for example, in educational settings, when students are called on randomly, and in some packet routing protocols.

(*iii*) In most human initiated activities task selection is not random, but the individual tends to execute always the highest priority item on his/her list. The resulting execution dynamics is quite different from (*i*) and (*ii*): high priority tasks will be executed soon after their addition to the list, while low priority items will have to wait until all higher priority tasks are cleared, forcing them to stay longer on the list. In the following we show that this selection mechanism, practiced by humans on a daily basis, is the likely source of the fat tails observed in human initiated processes.

Queue Length or System Capacity: In most queuing models the queue has an infinite capacity and the queue length can change dynamically, depending on the arrival and the execution rate of the individual tasks. In some queuing processes there is a physical limitation on the queue length. For example, the buffers of Internet routers have finite capacity, so that packets arriving while the buffer is full are systematically dropped. In human activity one could argue that, given the possibility to maintain the priority list in a written or electronic form, the length of the list has no limitations. Yet, if confronted with too many responsibilities, humans will start dropping some tasks and not accept others. Furthermore, while keeping track of a long priority list is not a problem for an electronic organizer, it is well established that the immediate memory of humans has finite capacity of about seven tasks [213, 244]. In other words, the number of priorities we can easily remember, and therefore the length of our priority list, is bounded. These considerations force us to inspect closely the difference between finite and an unbounded priority lists, and the potential consequences of the queue length on the the waiting time distribution.

In this paper we follow the hypothesis that the empirically observed heavy tailed distributions originate in the queuing process of the tasks maintained by humans, and seek appropriate models to explain and quantify this phenomenon. Particularly valuable are queuing models that do not contain power law distributions as inputs, and yet generate a heavy tailed output. In the following we will focus on priority queues, reflecting the fact that humans most likely choose the tasks based

on their priority for execution.

In the empirical datasets discussed in Section 5.3 we focused on both the interevent time and the waiting time distribution of the tasks in which humans participate. In the following two Sections we focus on the *waiting time* of a task on the priority list rather than the interevent times. In this context he waiting time, τ_w, represents the time difference between the arrival of a task to the priority list and its execution, thus it is the sum of the time a task waits on the list and the time devoted to executing it. In Section 5.7 we will return to the relationship between the empirically observed interevent times and the waiting times predicted by the discussed models.

5.5 Variable queue length models: $\alpha = 3/2$ universality class

Our first goal is to explore the behavior of priority queues in which there are no restrictions on the queue length. Therefore, in these models an individual's priority list could contain arbitrary number of tasks. As we will show below, such models offer a good approximation to the surface mail correspondence patterns, such as that observed in the case of Einstein, Darwin and Freud (see Section 5.3.2). Therefore, we will construct the models with direct reference to the the datasets discussed in Section 5.3. We assume that letters arrive at rate λ following a Poisson process with exponential arrival time distribution. Replacing letters with tasks, however, provides us a more general model, in principle applicable to any human activity. The responses are written at rate μ, reflecting the overall time a person devotes to his correspondence. Each letter is assigned a discrete priority parameter $x = 1, 2, \ldots, r$ upon arrival, such that always the highest priority unanswered letter (task) will be always chosen for a reply. The lowest priority task will have to wait the longest before execution, and therefore it dominates the waiting time probability density for large waiting times. This model was introduced in 1954 by Cobham [245] to describe some manufacturing processes. Most of the analytical work in queuing theory has concentrated on the waiting time of the lowest priority task, finding that the waiting time distribution follows [246]

$$P(\tau_w) \sim A\tau_w^{-3/2} \exp\left(-\frac{\tau_w}{\tau_0}\right) , \tag{5.5}$$

where A and τ_0 are functions of the model parameters, the characteristic waiting time τ_0 being given by

$$\tau_0 = \frac{1}{\mu\left(1-\sqrt{\rho}\right)^2} \ , \tag{5.6}$$

where $\rho = \lambda/\mu$ is the traffic intensity. Therefore, the waiting time distribution is characterized by a power law decay with exponent $\alpha = 3/2$, combined with an exponential cutoff.

The model can be extended to the case where the priorities are not discrete, but take up continuous values $0 \leq x < \infty$ from an arbitrary $\eta(x)$ distribution. The Laplace transform of the waiting time distribution for this case has been calculated in Ref. [240], but the resulting equation is difficult to invert, forcing us to study the model numerically (Fig. 5.4). The natural control parameter is $\rho = \lambda/\mu$, allowing us to distinguish three qualitatively different regimes:

Subcritical regime, $\rho < 1$: Given that the arrival rate of the tasks is smaller than the execution rate, the queue will be often empty. This significantly limits the waiting time, most tasks being executed soon after their arrival. The simulations indicate that the waiting time distribution exhibits an asymptotic scaling behavior consistent with Eq. 5.5 (Fig. 5.4). While in the $\rho \to 0$ limit we observe mainly the exponential decay, as ρ approaches 1 a power law regime with exponent $\alpha = 3/2$ emerges, combined with the exponential cutoff.

Critical regime, $\rho = 1$: When the arrival and the response rate of the letters are equal, according to Eqs. 5.5 and 5.6 we should observe a power law waiting time distribution with $\alpha = 3/2$ (Fig. 5.4). This regime would imply that, for example, Darwin responds to all letters he receives, which is not the case, given that their response rate is 0.32 (Darwin), 0.24 (Einstein) and 0.31 (Freud) [220]. In this case it is easy to show that the queue length performs a one-dimensional random walk bounded at $l = 0$. These fluctuations in the queue length will limit the waiting time distribution, as the tasks will wait at most as long as it takes for the queue length to return to $l = 0$. Therefore, the waiting time distribution will have as upper bound the return time distribution of a one-dimensional random walk. It is known, however, that the return time distribution of a random walker follows $P(t) \sim t^{-3/2}$ [247, 248], which is the origin of the 3/2 exponent in Eq. 5.4. This argument indicates that Eq. 5.5 is related to the fluctuations in the length of the priority list.

Supercritical regime, $\rho > 1$: Given that in this regime the arrival rate exceeds the response rate,

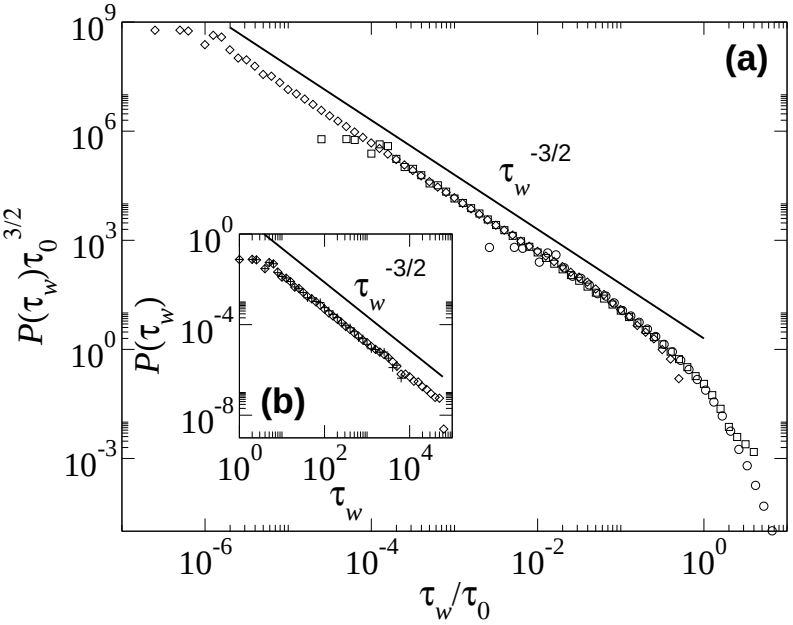

Figure 5.4: Waiting time distribution for tasks in the queueing model discussed in Section 5.5 with continuous priorities. The numerical simulations were performed as follows: At each step we generate an arrival τ_a and service time τ_s from an exponential distribution with rate λ and μ, respectively. If $\tau_a < \tau_s$ or there are no tasks in the queue then we add a new task to the queue, with a priority $x \in [0, 1]$ from uniform distribution, and update the time $t \to t + \tau_a$. Otherwise, we remove from the queue the task with the largest priority and update the time $t \to t + \tau_s$. The waiting time distribution is plotted for three $\rho = \lambda/\mu$ values: $\rho = 0.9$ (circles), $\rho = 0.99$ (squares) and $\rho = 0.999$ (diamonds). The data has been rescaled to emphasize the scaling behavior $P(\tau_w) = \tau_w^{-3/2} f(\tau_w/\tau_0)$, where $\tau_0 \sim (1 - \sqrt{\rho})^{-2}$. In the inset we plot the distribution of waiting times for $\rho = 1.1$, after collecting up to 10^4 (plus) and 10^5 (diamonds) *executed* tasks, showing that the distribution of waiting times has a power law tail even for $\rho > 1$ (supercritical regime). Note, however, that in this regime a high fraction of tasks are never executed, staying forever on the priority list whose length increases linearly with time, a fact that is manifested by a shift to the right of the cutoff of the waiting time distribution.

the average queue length grows linearly as $\langle l(t)\rangle = (\lambda - \mu)t$. Therefore, a $1 - 1/\rho$ fraction of the letters is never responded to, waiting indefinitely in the queue. Given Darwin, Einstein and Freud's small response rate, this regime captures best their correspondence pattern. We can measure the waiting time for each letter that is responded to. In Fig. 5.4 we show the waiting time probability density obtained from numerical simulations, indicating that it follows a power law with exponent $\alpha = 3/2$. Thus the supercritical regime follows the same scaling behavior as the critical regime, but only for the letters that are responded to. The rest of the letters wait indefinitely in the list ($\tau_w = \infty$).

A power law distribution emerges only in $\rho = 1$ and $\rho > 1$ regimes. The $\rho = 1$ regime requires a careful tuning of the human execution rate, so that the execution and the arrival rates are exactly the same. In contrast, for $\rho > 1$ no tuning is necessary, but the number of tasks on the list increases linearly with time, thus many tasks are never executed. This limit is probably the most realistic for human dynamics: we often take on tasks that we never execute, and technically stay on our priority list forever. As we discussed above, this is the case for Einstein, Darwin and Freud, who answer only a fraction of their letters. However, we must not overlook the second important feature of the discussed model: the only exponent it can predict is $\alpha = 3/2$, rooted in the fluctuations of the queue length. While this fully agrees with the correspondence patterns of Einstein, Darwin and Freud, it is significantly higher than the values observed in the empirical data discussed in Section 5.3.1 on web browsing, email communications or library visits, which we found to be scattered around $\alpha = 1$.

5.6 Fixed queue length models: $\alpha = 1$ universality class

According to the model discussed in the previous Section an individual must have the capacity to keep track of tens or hundreds of tasks at the same time. This may be appropriate for surface mail, where the letters pile on our desk until replied to. In contrast, there is extensive evidence from the psychology literature that the number of tasks humans can easily keep in their short term memory is bounded [213]. This leads us to inspect a model in which the length of the priority list remains unchanged [219], a new task being added only when an old task is removed from the list (executed).

We assume that an individual mantains a priority list with L tasks, each task being assigned a priority parameter $x_i,\ i = 1, ..., L$, chosen from an $\eta(x)$ distribution. At each time step with

probability p the individual selects the highest priority task and executes it, removing it from the list. At that moment a new task is added to the list, its priority x_i is again chosen from $\eta(x)$, thus the length L of the list remains unchanged. With probability $1 - p$ the individual executes a randomly selected task, independent of its priority. The $p \to 1$ limit of the model describes the deterministic highest-priority-first protocol, when always the highest priority task is chosen for execution, while $p \to 0$ corresponds to the random choice protocol, introduced to mimic the fact that humans occasionally select some low priority items for execution, before all higher priority items are executed. In the model time is discrete, each task execution corresponding to one unit of time. Implicit in this assumption is the approximation that the service time distribution follows a delta function, *i.e.*, each task takes one unit time to execute.

To understand the dynamics of the model we first study it via numerical simulations with priorities chosen from a uniform distribution $x_i \in [0, 1]$. The simulations show that in the $p \to 1$ limit the probability that a task spends τ_w time on the list has a power law tail with exponent $\alpha = 1$ (Fig. 5.5a). In the $p \to 0$ limit $P(\tau_w)$ follows an exponential distribution (Fig. 5.5a), as expected for the random selection protocol. As the typical length of the priority list differs from individual to individual, it is important for the tail of $P(\tau_w)$ to be independent of L. Numerical simulations indicate that this is indeed the case: changes in L do not affect the scaling of $P(\tau_w)$ [219]. The fact that the scaling holds for $L = 2$ as well indicates that it is not necessary to have a long priority list: even if an individual balances only two tasks at the same time, a bursty heavy tailed interevent dynamics will emerge. Next we focus on the $L = 2$ case, for which the model can be solved exactly, providing important insights into its scaling behavior that can be generalized for arbitrary L values as well.

5.6.1 Exact solution for $L = 2$

For $L = 2$ the waiting time distribution was exactly determined by A. Vázquez [221] (see Appendix B.1), obtaining

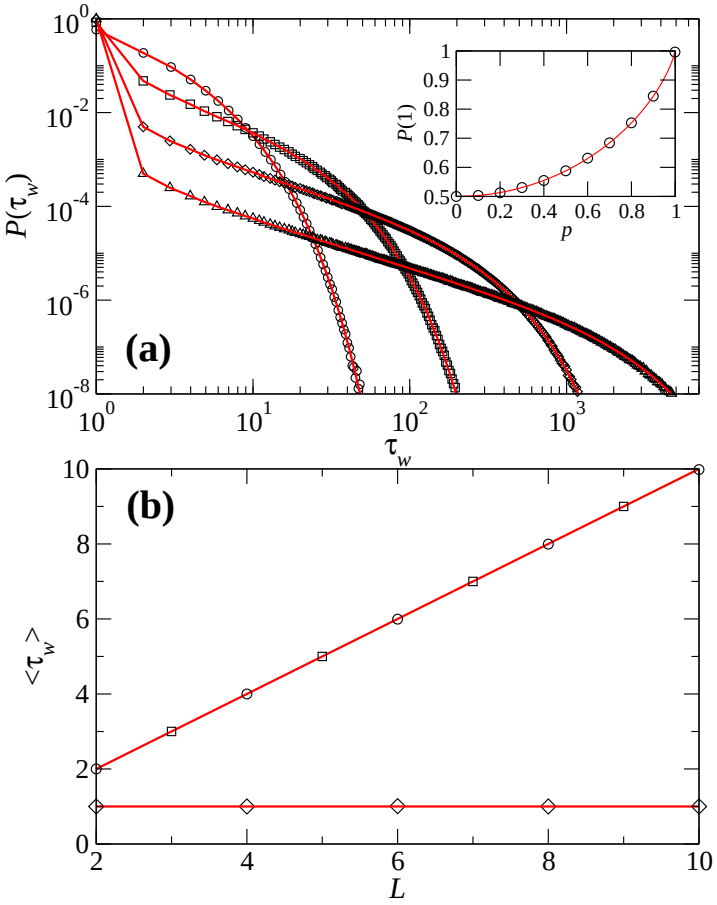

Figure 5.5: (a) Waiting time probability distribution function for the model discussed in Section 5.6 for $L = 2$ and a uniform new task priority distribution function, $\eta(x) = 1$, in $0 \leq x \leq 1$, as obtained from Eq. 5.7 (lines) and numerical simulations (symbols), for $p = 0.5$ (circles), $p = 0.9$ (squares), $p = 0.99$ (diamonds) and $p = 0.999$ (triangles). The inset shows the fraction of tasks with waiting time $\tau = 1$, as obtained from (5.7) (lines) and numerical simulations (symbols). (b) Average waiting time of executed tasks vs the list size as obtained from Eq. B.9 (lines) and numerical simulations (symbols), for $p = 0.0$ (squares), $p = 0.999$ (circles) and $p = 1$ (diamonds).

$$P(\tau_w) = \begin{cases} 1 - \frac{1-p^2}{4p} \ln \frac{1+p}{1-p} \ , & \tau_w = 1 \\ \\ \frac{1-p^2}{4p(\tau_w - 1)} \left[\left(\frac{1+p}{2} \right)^{\tau_w - 1} - \left(\frac{1-p}{2} \right)^{\tau_w - 1} \right] \ , & \tau_w > 1 \end{cases} \tag{5.7}$$

independent of $\eta(x)$ from which the task priorities are selected. In the limit $p \to 0$ from Eq. 5.7 follows that

$$\lim_{p \to 0} P(\tau_w) = \left(\frac{1}{2} \right)^{-\tau_w} \ , \tag{5.8}$$

i.e. $P(\tau_w)$ decays exponentially, in agreement with the numerical results (Fig. 5.5a). This limit corresponds to the random selection protocol, where a task is selected with probability 1/2 in each step. In the $p \to 1$ limit we obtain

$$\lim_{p \to 1} P(\tau_w) = \begin{cases} 1 + \mathcal{O}\left(\frac{1-p}{2} \ln(1-p) \right) \ , & \tau_w = 1 \\ \\ \mathcal{O}\left(\frac{1-p}{2} \right) \frac{1}{\tau_w - 1} \ , & \tau_w > 1 \ . \end{cases} \tag{5.9}$$

In this case almost all tasks have a waiting time $\tau_w = 1$, being executed as soon as they were added to the priority list. The waiting time of tasks that are not selected in the first step follows a power law distribution, decaying with $\alpha = 1$. This behavior is illustrated in Fig. 5.5a by a direct plot of $P(\tau_w)$ in Eq. 5.7 for a uniform distribution $\eta(x)$ in $0 \leq x \leq 1$. For $p < 1$ the $P(\tau_w)$ distribution has an exponential cutoff, which can be derived from Eq. 5.7 after taking the $\tau_w \to \infty$ limit with p fixed, resulting in

$$P(\tau_w) \sim \frac{1-p^2}{4} \frac{1}{\tau_w} \exp\left(-\frac{\tau_w}{\tau_0} \right) \ , \tag{5.10}$$

where

$$\tau_0 = \left(\ln \frac{2}{1+p} \right)^{-1} \ . \tag{5.11}$$

When $p \to 1$ we obtain that $\tau_0 \to \infty$ and, therefore, the exponential cutoff is shifted to higher τ_w values, while the power law behavior $P(\tau_w) \sim 1/\tau_w$ becomes more prominent. The $P(\tau_w)$ curve

systematically shifts, however, to lower values for $\tau_w > 1$, indicating that the power law applies to a vanishing task fraction (see Fig. 5.5a and Eq. 5.10). In turn, $P(1) \to 1$ when $p \to 1$, corroborated by the direct plot of $P(1)$ as a function of p (see inset of Fig. 5.5a).

5.6.2 Numerical results for $L > 2$

Based on the results discussed above, the overall behavior of the model with a uniform priority distribution can be summarized as follows. For $p = 1$, corresponding to the case when *always* the highest priority task is removed, the model does not have a stationary state. Indeed, each time the highest priority task is executed, there is a task with smaller priority x_m left on the list. With probability $1-x_m$ the newly added task will have a priority x'_m larger than x_m, and will be executed immediately. With probability x_m, however, the new task will have a smaller priority, in which case the older task will be executed, and the new task will become the 'resident' one, with a smaller priority $x'_m < x_m$. For a long period all new tasks will be executed right away, until an another task arrives with probability x''_m that again pushes the non-executed priority to a smaller value $x''_m < x'_m$. Thus with time the priority of the lowest priority task will converge to zero, $x_m(t) \to 0$, and thus with a probability converging to one the new task will be immediately executed. This convergence of x_m to zero implies that for $p = 1$ the model does not have a stationary state. A stationary state develops, however, for any $p < 1$, as in this case there is always a finite chance that the lowest priority tasks will also be executed, thus the value of x_m will be reset, and will converge to some $x_m(p) > 0$. This qualitative description applies for arbitrary $L > 2$ values.

To quantify this qualitative picture we studied numerically the $L > 2$ case assuming that $\eta(x)$ is uniformly distributed in the $0 \leq x \leq 1$ interval. To investigate how fast the system approaches the stationary state we compute the average priority of the lowest priority task in the queue, $\langle x_{\min}(t)\rangle$ (see Fig. 5.6a,b) since it represents a lower bound for the average of any other priorities on the list. We find that for any L values $\langle x_{\min}(t)\rangle$ decreases exponentially up to a time scale t_0, when it reaches a stationary value $\langle x_{\min}(\infty)\rangle$. The numerical simulations indicate that

$$t_0 \sim \frac{1}{1-p} \ , \tag{5.12}$$

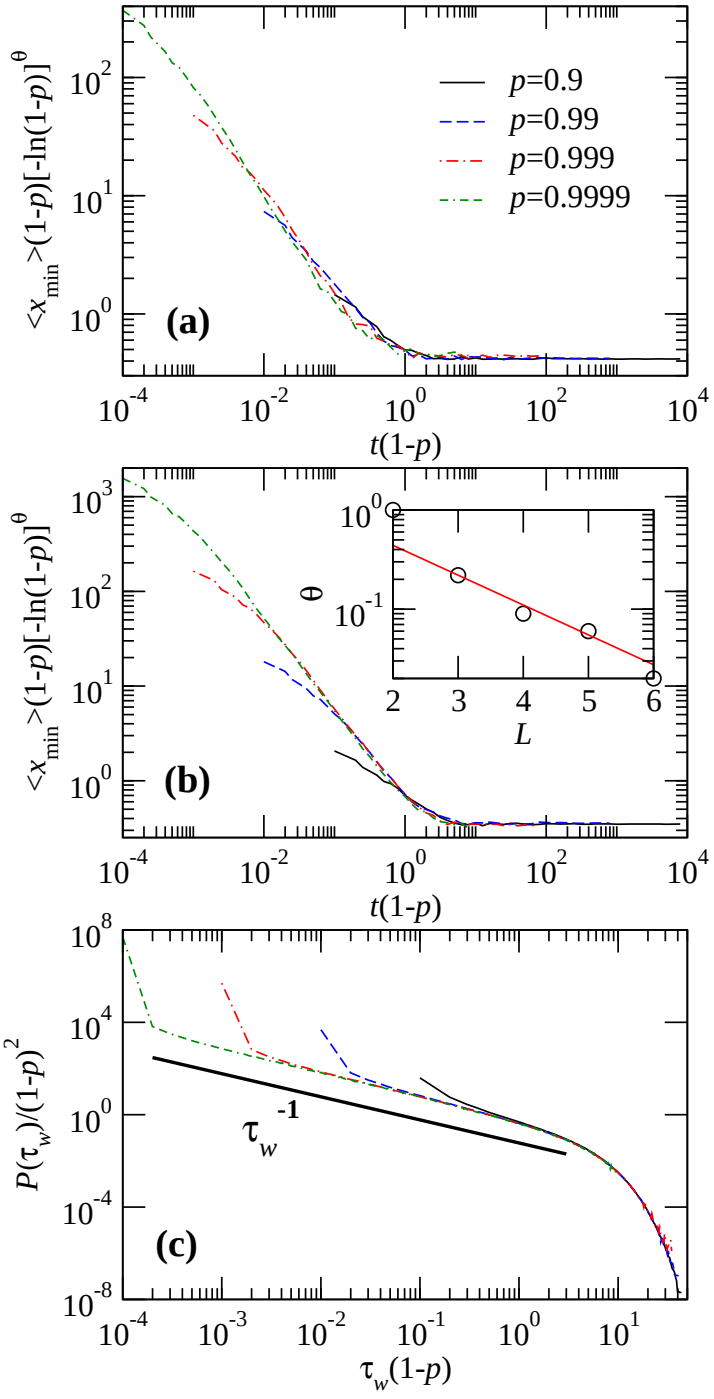

Figure 5.6: Rescaled plot of the average priority of the lowest task priority in the list for $L = 2$ (a) and $L = 3$ (b) and different values of p (see legend). The inset in (b) shows the exponent θ_L for different L (points), indicating that $\theta_L = \theta_3/2^{L-3}$ for $L > 2$ (continuous line). (c) Rescaled plot of the waiting time distribution for $L = 3$. Similar plots are obtained for larger vales of L (data not shown).

$$\langle x_{\min}(\infty)\rangle \sim (1-p)[-\ln(1-p)]^{\theta_L} \ . \tag{5.13}$$

For $L = 2$ we can calculate $\langle x_{\min}(\infty)\rangle$ exactly [221], obtaining

$$\begin{aligned} \langle x_{\min}(\infty)\rangle &= \frac{1-p}{2p}\left(\frac{1+p}{2p}\ln\frac{1+p}{1-p} - 1\right) \\ &\approx \frac{1-p}{2}[-\ln(1-p)] \ , \end{aligned} \tag{5.14}$$

and therefore $\theta_2 = 1$. For $L > 2$ we determined θ_L from the best data collapse, obtaining the values shown in the inset of Fig. 5.6b, indicating that

$$\theta_L = \frac{\theta_3}{2^{L-3}} \ ,$$

where $\theta_3 = 0.22$ is the value of θ_L for $L = 3$. These results support our qualitative discussion, indicating that for all $L \geq 2$ and $0 \leq p < 1$ values the system reaches a stationary state.

Finally we measured the waiting time distribution after the system has reached the stationary state. The results for $L = 3$ are shown in Fig. 5.6c, and similar results were obtained for other $L > 2$ values. The data collapse of the numerically obtained $P(\tau)$ indicates that

$$P(\tau) \sim (1-p)^2 \frac{1}{\tau}\exp\left(-\frac{\tau}{\tau_0}\right) \ , \tag{5.15}$$

when $L > 2$ and $\tau \gg 1$, where

$$\tau_0 \sim \frac{1}{1-p} \tag{5.16}$$

in the $p \to 1$ limit. The simulations indicate that the model's behavior for $L > 2$ is qualitatively similar to the behavior derived exactly for $L = 2$, but different scaling parameters characterize the scaling functions. For any $L \geq 2$, however, the waiting times scale as $P(\tau_w) \sim \tau_w^{-1}$, *i.e.* we have $\alpha = 1$.

5.6.3 Comparison with the empirical data

As the results in the previous Sections show, the model proposed to account for the $\alpha = 1$ universality class has some apparent problems. Indeed, for truly deterministic execution ($p = 1$) the model does

not have a stationary state. The problem was solved by introducing a random task execution ($p < 1$), which leads to stationarity. In this case, however, a p dependent fraction of tasks are executed immediately, and only the rest of the long lived tasks follow a power law. As p converges to zero, the fraction of tasks executed immediately diverges, developing a significant gap between the power law regime, and the tasks displaying $\tau = 1$ waiting time. Is this behavior realistic, or represents an artifact of the model? A first comparison with the empirical data would suggest that this is indeed an artifact, as measurements shown in Fig. 5.2 do not provide evidence of a large number of tasks that are immediately executed. However, when inspecting the measured results we should keep in mind that they represent the intervent times, and not the waiting times (see Ref. [215] for more details related to the email dataset on this issue).

5.7 Relationship between waiting and interevent times

As we discussed above, the empirical measurements provide either the interevent time distribution $P(\tau)$ (Sections 5.3.1 and 5.3.3) or the waiting time distribution $P(\tau_w)$ (Section 5.3.2) of the measured human activity patterns. In contrast the model predicts only the waiting time τ_w of a task on an individual's priority list. What is the relationship between the observed interevent times and the predicted waiting times? The basic assumption of this chapter is that the waiting times the various tasks experience on an individual's priority list are responsible for the heavy tailed distributions seen in the interevent times as well. The purpose of this section is to discuss the relationship between the two quantities.

The model predictions, that the waiting time distribution of the tasks follows a power law, are directly supported by one dataset in each universality class: the email data and the correspondence data. As discussed in Section 5.3, we have measured the waiting time distribution for both datasets, finding that the distribution of the response times indeed follows a power law with exponent $\alpha = 1$ (email) and $\alpha = 3/2$ (correspondence mail) as predicted by the models. Therefore, the direct measurement of the waiting times are likely rooted in the fat tailed response time distribution. For the other three datasets, however, such as web browsing, library visits and stock purchases, we cannot determine the waiting time of the individual events, as we do not know when a given task is added to the individual's priority list.

To explore the broader relationship between the waiting times and the interevent times we must remember that, while during the measurements we are focusing on a specific task (like email), the models assume the knowledge of *all* tasks that an individual is involved in. Thus the empirical measurements offer only a selected subset of an individual's activity pattern. To see the relationship between τ and τ_w next we discuss two different approaches.

Queueing of different task categories: The first approach acknowledges the fact that tasks are grouped in different categories of priorities: we often do not keep in mind specific emails to be answered, but rather remember that we need to check our email and answer whatever needs attention. Given this, one possible modification of the discussed models would assume that the tasks we monitor correspond to specific activity categories, and when we are done with one of them, we do not remove it from the list, but we just add it back with some changed priority. That is, checking our email does not mean that we deleted email activity from our priority list, but only that next has some different priority. If we monitor only one kind of activity, then a proper model would be the following: we have L tasks, each assigned a given priority. After a task is executed, it will be *reinserted* in the queue with a new priority chosen from the same distribution $\eta(x)$. If we now monitor the time at which the different tasks exit the list, we will find that the interevent times for the *monitored* tasks correspond exactly to the waiting time of that task on the list. Note that this conceptual model would work even if the tasks are not immediately reinserted, but after some delay τ_d. Indeed in this case the interevent time will be $\tau = \tau_w + \tau_d$, and as long as the distribution from which τ_d is selected from is bounded, the tail of the interevent time distribution will be dominated by the waiting time statistics.

Interaction between individuals: The timing of specific emails also depends on the interaction between the individuals that are involved in an email based communication. Indeed, if user A gets an email from user B, she will put the email into her priority list, and answer when she gets to it. Thus the timing of the response depends on two parameters: the receipt time of the email, and the waiting time on the priority list. Consider two email users, A and B, that are involved in an email based conversation. We assume that A sends an email to B as a response to an email B sent to A, and vice-versa. Thus, the interevent time between two consecutive emails sent by user A to user B is given by $\tau = \tau_w^A + \tau_w^B$, where τ_w^A is the waiting time the email experienced on user A's

queue, and τ_w^B is the waiting time of the response of user B to A's email. If both users prioritize their tasks, then they both display the same waiting time distribution, *i.e.* $P(\tau_w^A) \sim (\tau_w^A)^{-\alpha}$ and $P(\tau_w^B) \sim (\tau_w^B)^{-\alpha}$. In this case the interevent time distribution $P(\tau)$, which is observed empirically if we study only the activity pattern of user A, follows also $P(\tau) \sim \tau^{-\alpha}$. Thus the fact that users communicate with each other turns the waiting time into observable interevent times.

In summary, the discussed mechanisms indicate that the waiting time distribution of the tasks could in fact drive the interevent time distribution, and that the waiting time and the interevent time distributions should decay with the same scaling exponent. In reality, of course, the interplay between the two quantities can be more complex than discussed here, and perhaps even better mapping between the two measures could be found for selected activities. But these two mechanisms indicate that if the waiting time distribution is heavy tailed, we would expect that the interevent time distribution would also be heavy tailed.

5.8 Discussion

In the following we will discuss the main results obtained in this chapter. A more complete discussion (including model limitations, task optimization and correlations) can be found in Ref. [215].

Universality classes: As summarized in the introduction, the main goal of this chapter was to discuss the potential origin of the heavy tailed distributed interevent times observed in human dynamics. To start we provided evidence that in five distinct processes, each on a different human activity, the interevent time distribution for individual users follows a power law. Our fundamental hypothesis is that the observed interevent time distributions are rooted in the mechanisms that humans use to decide when to execute the tasks on their priority list. To support this hypothesis we studied a family of queuing models, assuming that each task to be executed by an individual waits some time on the individual's priority list and we showed that queuing can indeed generate power law waiting time distributions. We find that a model that allows the queue length to fluctuate leads to $\alpha = 3/2$, while a model for which the queue length is fixed displays $\alpha = 1$. These results indicate that human dynamics is described by at least two universality classes, characterized by empirically distinguishable exponents. Note that while we have classified the models based on the limitations on the queue length, we cannot exclude the existence of models with fixed queue length that scale

with $\alpha = 3/2$, or models with fluctuating length that display scaling with $\alpha = 1$, or some other exponents (see next Chapter).

In comparing these results with the empirical data, we find that email and phone communication, web surfing and library visitation belong to the $\alpha = 1$ universality class. The correspondence patterns of Einstein, Darwin and Freud offer convincing evidence for the relevance of the $\alpha = 3/2$ exponent, and the related universality class, for human dynamics. In contrast the fourth process, capturing a stock broker's activity, shows $\alpha = 1.3$. Given, however, that we have data only for a single user, this value is in principle consistent with the scattering of the exponents from user to user, thus we cannot take it as evidence for a new universality class. One issue still remains without a satisfactory answer: why does email and surface mail (Einstein, Darwin and Freud datasets) belong to different universality classes? We can comprehend why should the mail correspondence belong to the 3/2 class: letters likely pile on the correspondent's desk until they are answered, the desk serving as an external memory, thus we do not need to remember them. But the same argument could be used to explain the scaling of email communications as well, given that unanswered emails will stay in our mailbox until we delete them (which is one kind of task execution). Therefore one could argue that email based communication should also belong to the 3/2 universality class, in contrast with the empirical evidence, that clearly shows $\alpha = 1$ [219, 116].

In addition we argued that in a series of processes the waiting time distribution determines the interevent time distribution as well (see Section 5.7). This argument closes the loop of the chapter's logic, establishing the relevance of the discussed queueing models to the datasets for which only interevent times could be measured. We do not feel, however, that this argument is complete, and probably future work will strengthen this link. In this respect two directions are particularly promising. First, designing queueing models that can directly predict the observed interevent times as well would be a major advance. Second, establishing a more general link between the waiting time and interevent times could also be of significant value.

Non-human activity patterns: Heavy tailed interevent time distributions do not occur only in human activity, but emerge in many natural and technological systems. For example, Omori's law on earthquakes [249, 250] records heavy tailed interevent times between consecutive seismic activities; measurements indicate that the fishing patterns of seabirds also display heavy tailed statistics [251];

plasticity patterns [252] and avalanches in lungs [253] show similar power law interevent times. While a series of models have been proposed to capture some of these processes individually, there is also a possibility that some of these modeling frameworks can be reduced to various queuing processes. Some of the studied queuing models show close relationship to several models designed to capture self-organized criticality [254, 255, 256, 257, 258, 259]. Could the mechanisms be similar at some fundamental level? Even if such higher degree of universality is absent, understanding the mechanisms and queuing processes that drive human dynamics could help us better understand other natural phenomena as well

Network effects: In searching for the explanation for the observed heavy tailed human activity patterns we limited our study to the properties of *single queues*. In reality none of our actions are performed independently — most of our daily activity is embedded in a web of actions of other individuals [260]. An important goal is to understand how the various human activities and their timing is affected by the fact that the individuals are embedded in a network environment. The next Chapter aims to develop this aspect.

Chapter 6

Model of interactions on human dynamics

In the previous Chapter we used queueing theory as a framework to model the heavy tailed statistics of human activity patterns. The main predictions are the existence of a power law distribution for the interevent time of human actions and two universality classes, with decay exponents $\alpha = 1$ and $\alpha = 3/2$. The proposed models lack, however, a key aspect of human dynamics, *i.e.* several tasks require, or are determined by, interactions between individuals. Here we introduce a minimal queueing model of human dynamics that already takes into account human-human interactions. To achieve large scale simulations we obtain a coarse-grained version of the model, allowing us to reach large interevent times and reliable scaling exponents estimations. Using this coarse-grained version, we show that the interevent distribution of interacting tasks exhibit the scaling exponents $\alpha = 2$, $3/2$ and a series of numerable values between $3/2$ and 1. This work demonstrates that, within the context of queueing models of human dynamics, interactions change the universality class. Beyond the study of human dynamics, these results are relevant to systems where the event of interest consists of the simultaneous occurrence of two (or more) events.

6.1 Introduction

In the recent years we have experienced an increased research activity in this area motivated by the increased availability of empirical data [27, 116, 219, 220, 222]. Thanks to this data we are in a position to investigate the laws and patterns of human dynamics using a scientific approach.

Within the framework of queueing theory [240, 241], the *to do list* of an individual is modeled as a finite length queue with a task selection protocol, such as highest priority first. The main predictions are the existence of a power law distribution of interevent times $P_\tau \sim \tau^{-\alpha}$ and two universality classes characterized by exponents $\alpha = 1$ [219, 221, 215] and $\alpha = 3/2$ [220, 215]. These universality classes have been corroborated by empirical data for email [219, 215] and regular mail communications [220, 215], respectively, motivating further theoretical research [261, 119, 262, 263, 264].

The models proposed so far have been limited, however, to single individual dynamics. In practice people are connected in social networks and several of their activities are not performed independently. This reality leads us to model human dynamics in the presence of interactions between individuals. Our past experience with phase transitions has shown us that interactions and their nature are a key factor determining the universality classes and their corresponding scaling exponents [223].

Furthermore, beyond the study of human dynamics, there are several systems where the event of interest consists of the simultaneous occurrence of two (or more) events. For example, collective phenomena in disordered media, such as the interaction of two (or more) particles in cluster formation [265].

6.2 The model of interacting queues

To investigate the impact of human-human interactions on the timing of their activities we consider a minimal model consisting of two agents, A and B (Fig. 6.1). Each agent is modeled by a priority list containing two tasks, interacting task (I) and aggregate non-interacting task (O). The interacting task models a common activity such as meeting each other, requiring the simultaneous execution of that task by both agents. On the other hand, the non-interacting task represents an aggregate meta-activity accounting for all other tasks the agents execute, which do not require an interaction

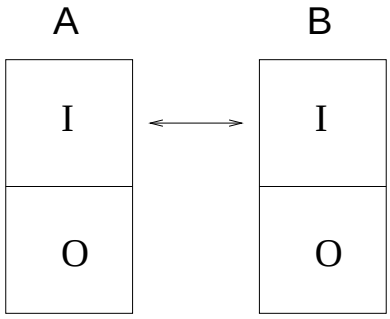

Figure 6.1: System of two agents with a common interacting task I and an aggregate task O representing a set of individual tasks.

between them. To each task we assign random priorities x_{ij} $(i = I, O;\ j = A, B)$ extracted from a probability density function (pdf) $f_{ij}(x)$ (see Fig. 6.1).

The rules governing the dynamics are as follows. *Initial condition:* We start with a random initial condition, assigning a priority to the I and O tasks from their corresponding pdf. *Updating step:* At each time step, both agents select the task with higher priority in their list. If (i) both agents select the interacting task then it is executed, (ii) otherwise each agent executes the O task, representing the execution of any of their non-interacting tasks.

Our aim is to determine the impact of the interaction between the agents and the shape of $f_{ij}(x)$ on the scaling exponent α of the interevent time distribution of the interacting task I. For simplicity, we focus on the following priority distribution. Consider the case where each agent has L_j $(j = A, B)$ tasks, one I task and $L_j - 1$ non-interacting tasks, their priorities following a uniform distribution in the interval $[0, 1]$. The pdf of the highest priority among $L_j - 1$ tasks is in this case given by $(L_j - 1)x^{L_j - 2}$, resulting in

$$f_{ij}(x) = \begin{cases} 1\ , & i = I \\ (L_j - 1)x^{L_j - 2}\ , & i = O\ . \end{cases} \tag{6.1}$$

This example shows that the priorities pdf of task I and O are in general different. All the results shown below were obtained using the pdf in Eq. (6.1).

To investigate the interevent time distribution we perform extensive numerical simulations. Figure 6.2 shows the interevent time distribution as obtained from direct simulations of the model

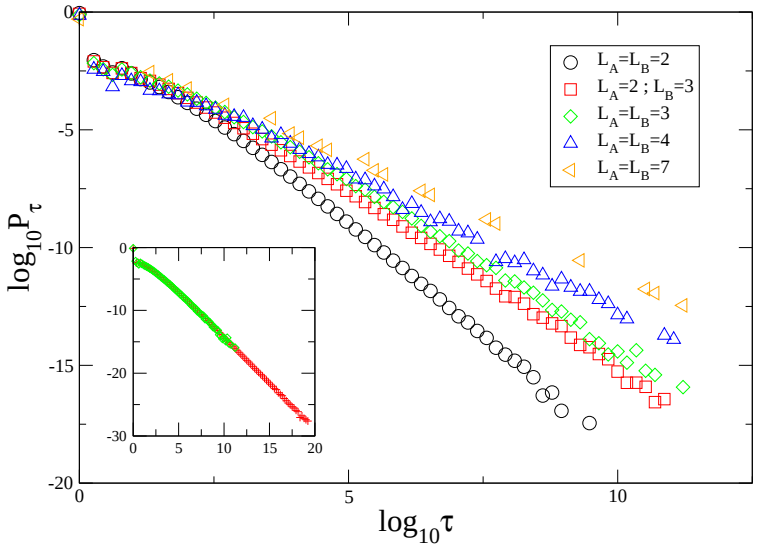

Figure 6.2: Probability distribution of the interevent time τ of the interacting task I, as obtained from the direct numerical simulations of the model. Each dataset was obtained after 10^{11} model time steps, corresponding with total number of I plus O task executions. Note that as L_A and/or L_B increases it becomes computationally harder to have a good estimate of P_τ because the execution of the I task becomes less frequent. The inset shows the distribution for $L = 3$ as obtained from the original model with 10^{12} steps (green diamonds), and the coarse-grained model with $N = 10^9$ (red plus), derived to obtain more reliable estimation of the exponents.

introduced above. It becomes clear that for large L_A and/or L_B we do not obtain a good statistics, even after waiting for 10^{11} updating steps. This observation is a consequence of the behavior of $f_{Oj}(x)$ when L_A and/or L_B are large (Fig. 6.3). Focusing on agent A, as L_A increases $f_{OA}(x)$ gets more concentrated around priority one, while the priority of the I task remains uniformly spread between zero and one. This fact results in increasingly large interevent times between the execution of the I task.

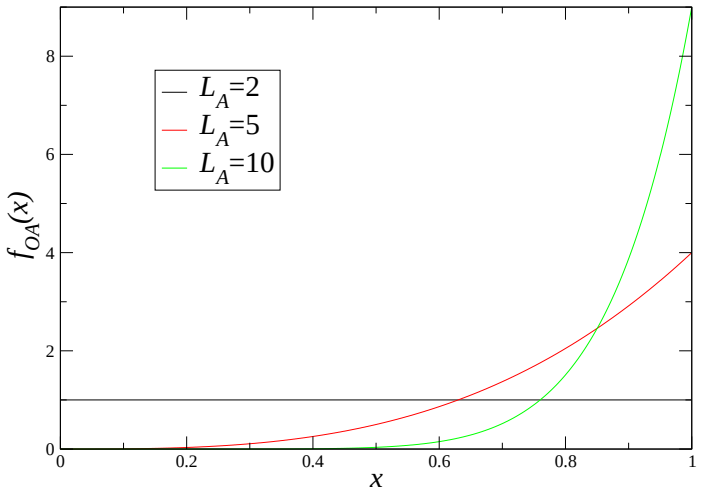

Figure 6.3: Probability density function of the non-interacting aggregate task priority of user A, as obtained from Eq. (6.1). With increasing the queue length L_A, $f_{OA}(x)$ concentrates more and more in the vicinity of $x = 1^-$.

6.3 The coarse-grained model

To speed-off the numerical simulations we derive a coarse-grained version of the model, allowing us to analyze the scaling behavior of the interevent time distribution over several orders of magnitude (inset of Fig. 6.2). We start by noticing that, given (x_{IA}, x_{IB}), the joint pdf of (x_{OA}, x_{OB}) factorizes and the probability $q(x_{IA}, x_{IB})$ that both agents execute I right after O is given by

$$q(x_{IA}, x_{IB}) = \int_0^{x_{IA}} dx f_{OA}(x) \int_0^{x_{IB}} dx f_{OB}(x) \ . \tag{6.2}$$

This factorization is possible because the execution of the I task requires its priority to be the largest for both agents. In turn, with probability $1 - q(x_{IA}, x_{IB})$ both agents continue to execute O. Thus, the probability distribution $Q_\tau(x_{IA}, x_{IB})$ that I waits $\tau > 1$ steps before being executed follows the geometric distribution

$$Q_\tau(x_{IA}, x_{IB}) = q(x_{IA}, x_{IB})[1 - q(x_{IA}, x_{IB})]^{\tau-2} \ . \tag{6.3}$$

Once the I task is executed it can be executed again resulting in interevent times of one step ($\tau = 1$). The overall interevent time distribution of the I task is given by

$$P_\tau = \begin{cases} P_1 \ , & \tau = 1 \\ (1 - P_1)\langle Q_\tau(x_{IA}, x_{IB})\rangle \ , & \tau > 1 \end{cases} \tag{6.4}$$

where

$$P_1 = \frac{S_1}{S_1 + 1} \ , \tag{6.5}$$

S_1 is the expected number of consecutive executions of the I task and $\langle \cdots \rangle$ denotes the expectation over different realizations of (x_{IA}, x_{IB}), just at the step of switching from task I to O. Finally, at the step of switching from O to I, the O task priority of both agents must fall below that of the I task. Therefore, the pdf of x_{Oj} $(j = A, B)$ just after the switch from O to I is given by

$$f^*_{Oj}(x|x_{Ij}) = \frac{f_{Oj}(x)}{\int_0^{x_{Ij}} f_{Oj}(x')dx'} \ , \tag{6.6}$$

where $0 \leq x < x_{Ij}$. This later result together with Eq. (6.3) allow us to condense all steps with consecutive executions of the O task into a single coarse-grained step. More important, this mapping is exact.

Putting all together the coarse-grained model runs as follows. *Initial condition:* We start with random initial priorities extracted from the pdfs $f_{ij}(x)$. *Updating step:* At each step, (i) if for both agents the I task priority is larger than that for the O task we run the model as defined above, both agents executing the I task and updating their I task priorities using the pdfs f_{Ij} $(j = A, B)$. (ii) Otherwise, we generate a random interevent time τ from the probability distribution (6.3) and a new O task priority for each agent using the pdf $f^*_{Oj}(x|x_{Ij})$ (6.6). This second step avoids going over

successive executions of the O task which, for a large number of non-interacting tasks, significantly slow down the simulations.

The second step of the coarse grained model requires us to extract a random number from a geometric distribution. This can be achieved very efficiently exploiting the fact that the integer part of a real random variable with an exponential distribution follows a geometric distribution. Using this fact, when $\tau > 1$, we extract τ exactly from the distribution in Eq. (6.3), which differs from the corresponding branch of Eq. (6.4). Normalization by the total number of task I executions, including those with $\tau = 1$, provides $\tau > 1$ distributed according to Eq. (6.4).

The I task interevent time distribution obtained from simulations of the coarse-grained model is plotted in Fig. 6.4a. When $L_A = L_B = L = 2$ it follows a power-law tail with exponent $\alpha = 2$. As L increases α approaches one. A guess to this dependence, in good agreement with the measured values, is given by $\alpha = 1 + 1/\max(L_j - 1)$ (inset of Fig. 6.4a). The numerical results indicate that there are several numerable universality classes parameterized by L_A and L_B. Notice that the second largest value of α (obtained when $L_A = 2$ and $L_B = 3$, or vice-versa) is close to 3/2 and, therefore, our results do not show universality classes with exponent α between 3/2 and 2 (unless we assume real valued queue lengths).

6.4 Scaling of the interevent time distribution

The power laws in Fig. 6.4a exhibit a cutoff at a certain value of τ. To investigate if this is a natural cutoff or just a finite size effect, we investigate the shape of the interevent time distribution as a function of the observation time window T. The later is defined as the total number of steps considering both the I and O task and satisfy

$$T = \sum_{i=1}^{N} \tau_i \ , \tag{6.7}$$

where N is the number of executions of the I task within the time window T and τ_i $(i = 1, \ldots, N)$ is the sequence of interevent times between executions of the I task. We assume that the cutoff is determined by the finite time window and that the interevent time distribution follows the scaling form

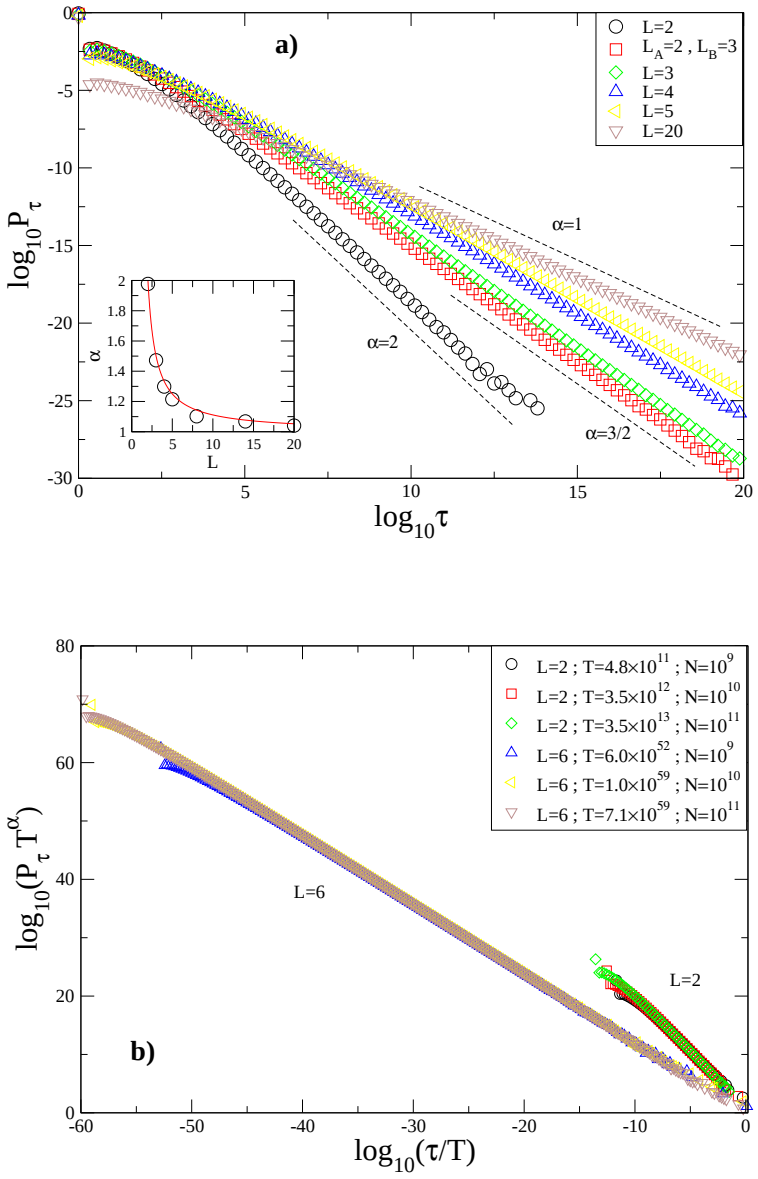

Figure 6.4: a) Probability distribution of the I task interevent time for several values of the number of tasks on each queue (L_A, L_B), as obtained from simulations of the coarse-grained model. When $L_A = L_B$ we denote this number by L. The inset shows the exponent α as measured from the power law tails (black circles) and the guess function $\alpha = 1 + 1/\max(L_j - 1)$ (red curve) in good agreement; to avoid confusion we only plot the case when $L_A = L_B = L$, but we checked for the general case as well. b) Scaling plot of the I task interevent time distribution. Note that, for a given α, the symbols corresponding to different time windows T collapse into a single plot.

$$P(\tau) = A\tau^{-\alpha} g\left(\frac{\tau}{T^z}\right) \tag{6.8}$$

where A is a constant, $z > 0$ is a scaling exponent and $g(x)$ is a scaling function with the asymptotic behaviors $g(x) \approx 1$ when $x \ll 1$ and $g(x) \ll 1$ when $x \gg 1$. Under this assumption $P(\tau) \sim \tau^{-\alpha}$ when $T \to \infty$, with $1 < \alpha \leq 2$. Given this power law tail and exponent, the number of interevent times N necessary to cover the window T is of the order of magnitude of $T^{\alpha-1}$ [60]. In turn, the mean intervent time is of the order of

$$\langle \tau \rangle = \frac{1}{N} \sum_{i=1}^{N} \tau_i \sim T^{2-\alpha} \ . \tag{6.9}$$

From Eqs. (6.8) and (6.9) it follows that $z = 1$.

To check our scaling assumption we plot $P_\tau T^\alpha$ as a function of τ/T (Fig. 6.4b). The symbols corresponding to different time windows T clearly overlap into a single curve, demonstrating that the scaling assumption in Eq. (6.8) is correct with $z = 1$. Thus, in the $T \to \infty$ limit the I task intervent time distribution exhibits a true power law tail $P_\tau \sim \tau^{-\alpha}$.

6.5 Discussion

Within the context of queuing models of human dynamics, only two universality classes were previously identified, corresponding to the single queue models of Cobham [220, 245] ($\alpha = 3/2$) and Barabási [219] ($\alpha = 1$) — see Chapter 5. The analysis of the two interacting agents model reveals that the interaction between agents results in a richer set of exponents. Although we have attempted to solve the model analytically, the asymmetry between the interacting and non-interacting task, turns this model more challenging than the Barabási model (Appendix B.1) and thus the exact analytical solution has not yet been found. Our numerical results provide, however, evidence of a new universality class with exponent $\alpha = 2$ and exponents between 3/2 and 1. It is worth noticing that the exponents 2 and 1 may also result from a Poisson process with a time dependent rate [266, 267].

Because the exponent α depends on the systems details, here represented by the agent's queue lengths L_A and L_B, we conclude that the model with two interacting agents exhibits non-universal behavior. Interestingly, the exponent $\alpha = 1$ is asymptotically reached when the number of tasks of

one or both agents becomes large. As humans get engaged in several tasks this later asymptotic behavior may explain the ubiquitous observation of the exponent $\alpha = 1$ [215].

We use the number of non-interacting tasks as a mean to modulate the distribution of the non-interacting aggregate task priority. Yet, it is the distribution shape the primary factor determining the scaling exponent α. The effect of increasing L_A and/or L_B is a concentration of the non-interacting aggregate task priority around priority one, resulting in values of α that approaches one. This means that the limit $\alpha = 1$ is achieved for low priority interacting tasks that remain most of the time in the queue without being executed, at expenses of the execution of tasks which in general have a higher priorities.

Considering the interaction between agents we also solve one of the standing problems of the original Barabási model, related to the stationarity of the interevent time distribution [221, 262]. In the Barabási single queue model the task with highest priority is executed with a probability p, otherwise a task is selected at random for execution. When p is close to one the interevent time distribution exhibits a peak at one step and $P_1 \to 1$ when $p \to 1$. When $p = 1$ the distribution is non-stationary and $P_1 \to 1$ when time $t \to \infty$. In contrast, in the model considered here there is no need to introduce the random selection rule and the corresponding model parameter p. The interacting task interevent time distribution is stationary even when the - highest priority first - selection rule is applied. In turn, the exponent α is not exactly one, but reaches one asymptotically with increasing the number of tasks. Finally, the interevent time distribution of the Barabási model exhibits a natural cutoff determined by the parameter p, while for the model introduced here it is a true power law up to finite size effects.

This work represents the first step in understanding how interactions among agents affect their activity pattern. Based on recent works using queueing theory we describe the model in the context of human dynamics. It can be generalized to consider a larger number of agents connected by a specific social network. Also, the model can potentially be used more generally to study the time statistics of events requiring the simultaneous occurrence of two events.

Chapter 7

Conclusions, outlook and list of publications

In this thesis, after the introduction to network theory in Chapter 1, we started in Chapter 2 with a study of structural properties of complex networks. The main results of this Chapter are, in Section 2.1, the finding of the logarithmic $k-$dependence (Eq. 2.3) of the geodesic $\ell(k)$ in networks with power-law degree distribution, and of the linear k-dependence (Eq. 2.4) in networks with exponential distribution; in Section 2.2, we find the existence of two subgraph classes in scale-free networks with power law degree-dependent clustering coefficient: The Type I subgraphs whose density increases with the network size, and the Type II subgraphs whose density is independent of N (Eq. 2.28). Also in Section 2.2 we find two kinds of cycles: Those with length $h > h_c$, whose density increases with N, and those (with length $h < h_c$) having N-independent density (Eq. 2.30). The results of this Section were analytically obtained and empirically verified for several real-world networks.

In Chapter 3 we have analyzed the real-world network of collaborations between universities and industry related entities promoted by the 5th Framework Programme in European Union. The main results are that it is a scale-free, highly correlated network, for which the analytical result of the previous Chapter (Eq. 2.3) is verified. Also, by splitting the network in two, one whose vertices are Universities, and another whose vertices are Companies, we find that the former is more tightly connected than the latter and conjecture some reasons for this as well as possible implications.

In Chapter 4, serving as joint between the first part of the thesis and the second, we studied the frequency of numbers on the World-Wide Web documents, finding an heterogeneous, heavy-tailed distribution, with certain numbers occurring much more frequently than others. This study generalizes results obtained long time ago by Newcomb and Benford for the frequency of numbers in human documents.

The statistics of the timing of human activities was the object of study of Chapter 5. By resorting to empirical data describing the temporal dynamics of several activities (such as email usage, Web browsing or the surface-mail correspondence of Darwin, Einstein and Freud) we find that the time between two consecutive actions by a single individual is heavy-tailed distributed, with periods of intense activity (short interevent times) separated by time gaps of no activity (long interevent times). Two universality classes are observed, one for the Darwin, Einstein and Freud correspondence, characterized by the power-law exponent $\alpha = 3/2$ and another for the other studied activities characterized by $\alpha = 1$. These observations are explained by resorting to single, priority queue models, based on the mathematical queueing theory.

In Chapter 6 we devise a generalized queueing model to account for interactions between individuals on social networks. It is a first model of two interacting priority queues, which may explain other possible exponents α in the timing of human dynamics (like, possibly, the stock broker activity of the previous Chapter). It thus represents a first step in understanding how interactions may affect the patterns observed in the previous Chapter, and can be easily generalized to N interacting queues by means, for example, of a social network. The model may also be potentially used in other areas where the object of study involves the simultaneous occurrence of two, or more, events, like cluster aggregation in disordered media, or synchronization studies in dynamical systems.

In overall, the work presented in this thesis considered only undirected, unweighted networks. As was seen, there is still a lot to be explored in this simplest case of graphs, which signals the many possibilities of research that graph theory presents, considering also that empirical network studies are within the scope of this mathematical theory and for which physics can be of great importance. The cases of directed, weighted networks, and also where intrinsic properties of vertices (or edges) are considered, open up even more possibilities to be investigated. For example, as we said above, the model of Chapter 6 can be generalized to be applied in a (social) network of N interacting

agents. More generally, the edges (representing acquaintances) of the network should be weighted, with weights depending on the time (normally decreasing) since the last interaction (or meeting). Even more generally, the model can affect the network's structure, with new edges (acquaintances) appearing and others whose weight vanishes, and thus practically disappear. In this way, the model can potentially be generalized to the whole society, for example. Of course, its results should be confronted with reality, if not we would be just working on the grounds of speculation. This is where, for example, social networking websites (beyond e-mail, instant messaging services, or telephone) data can be useful (as discussed in Section 1.3.1), not only to avoid relationships to disappear — the reason why we are in a connected age [55] — but also to allow for easier and faster analysis of social data.

List of Publications

- Books
  - *Science of Complex Networks*, Editors: J. F. F. Mendes, S. N. Dorogovtsev, A. Povolotsky, F. Abreu & J. G. Oliveira, AIP Conference Proceedings 776, June 2005.
- Journal Articles
  - A. Vázquez, J. G. Oliveira & A.-L. Barabási, *"The inhomogeneous evolution of subgraphs and cycles in complex networks"*, Phys. Rev. E **71**, 025103(R) (2005).
  - J. G. Oliveira & A.-L. Barabási, *"Human dynamics: Darwin and Einstein correspondence patterns"*, Nature **437**, 1251 (2005).
  - J. G. Oliveira and A.-L. Barabási, *"Correspondence patterns - Mechanisms and models of human dynamics - Reply"*, Nature **441**, E5-E6 (2006).
  - A. Vázquez, J. G. Oliveira, Z. Dezső, K.-I. Goh, I. Kondor & A.-L. Barabási, *"Modeling bursts and heavy tails in human dynamics"*, Phys. Rev. E **73**, 036127 (2006).
  - S. N. Dorogovtsev, J. F. F. Mendes & J. G. Oliveira, *"Frequency of occurrence of numbers in the World Wide Web"*, Physica A **360**, 548 (2006).

- S. N. Dorogovtsev, J. F. F. Mendes & J. G. Oliveira, *"Degree-dependent intervertex separation in complex networks"*, Phys. Rev. E **73**, 056122 (2006).
- J. A. Almendral, J. G. Oliveira, L. López, M. A. F. Sanjuán & J. F. F. Mendes, *"The interplay of universities and industry through the FP5 network"*, New J. Phys. **9**, 183 (2007).
- J. A. Almendral, J. G. Oliveira, L. López, M. A. F. Sanjuán & J. F. F. Mendes, *"The network of scientific collaborations within the European framework programme"*, Physica A **384**, 675 (2007).

• Preprints

- J. G. Oliveira & A. Vázquez, *"Impact of interactions on human dynamics"*, arXiv:0710.4916, submitted to Physica A.

Appendix A

Classification of FP5 participants

The Framework Programme (FP) sets out the priorities for the European Union's research and technological development. These priorities are defined following a set of criteria which pursue an increase of the industrial competitiveness and the quality of life for European citizens. A fact which shows the effort made by the European Union to promote this global policy for knowledge is the budget devoted to these programmes. For example, the FP5 (1998-2002) was implemented by means of 13,700 million euros and the FP6 (2002-2006) has assigned a budget of 17,883 million euros.

All projects in the FP5 are organized in eight specific programmes which can be classified as follows. There are five focused Thematic Programmes implementing research, technological development and demonstration activities:

- QOL: Quality of life and management of living resources (2,524 projects).
- IST: User–friendly information society (2,382 projects).
- GROWTH: Competitive and sustainable growth (2,014 projects).
- EESD: Energy, environment and sustainable development (1,772 projects).
- NUKE: Research and Training in the field of Nuclear Energy (1,032 projects).

And there are three Horizontal Programmes to cover the common needs across all research areas:

- INCO: Confirming the international role of Community research (1,034 projects).

- SME: Promotion of Innovation and encouragement of small and medium enterprises participation (142 projects).

- HPOT: Improving human research potential and the socio–economic knowledge base (4,876 projects).

The data to analyze the FP5 as a complex network were obtained from the web pages of CORDIS[1] with a robot implemented in Perl[2]. The result was a database with 15,776 records as follows:

Programme | Year | Participant1 - Nation - Dedication | Participant2 - Nation - Dedication | ...

The first field refers to the specific programme to which the project belongs and the second field informs us about the year in which it started. The following fields are the participants in the project with their corresponding nationality and dedication ('research', 'education', 'industry'...). We then have a bipartite graph [42, 43] since there are two kinds of vertices (participants and projects) and each edge links a participant with a project. To obtain the graph with 25,287 participants (nodes) and 329,636 collaborations (edges) used throughout the text, we have only to project it onto the participants.

The names of the participants were not free of typos since we collected them as they were in the web. The consequence of this fact was that sometimes the same participant appeared in two projects with different names and, consequently, it was recorded twice in the data. For instance, 'François Company of Something, Ltd.' and 'Francois Company of SOMETHING LTD' would be recorded as different. To avoid these duplications, we used a parser covering many possibilities which could lead to false entries. Nevertheless, despite our efforts, not all duplications have been eliminated. However, after a visual inspection of the data, we estimate that the error is below 10%.

To split the participants in Universities and Companies, we considered the organization type reported in the project. This information is encoded in the field 'Dedication', where we found 11 levels: 'Commission External Service', 'Commission Service', 'Consultancy', 'Education', 'Industry', 'Non Commercial', 'Not Available', 'Other', 'Research', 'Technology Transfer' and ⟨Void⟩.

[1] Community Research and Development Information Service: `http://cordis.europa.eu`

[2] http://www.perl.org/

The level 'Not available' means that the FP itself was not able to obtain the information and this absence is shown in this manner. In addition, the level ⟨Void⟩ means that no information at all is given, i.e. our robot found nothing (not even 'Not Available').

The first step to define only two groups was to reduce the number of levels in 'Dedication'. We found that eight levels could be merged to define a new one, called 'Non Companies'. It was not homogeneous since we found consultancies, universities, hospitals, institutes, laboratories, observatories, museums, technological parks even cities. However, they all were participants involved in some type of research for whom results do not necessarily return income. This new level was, basically, the union of 'Research' and 'Education' since the other six levels appeared few times in the data: 'Commission External Service' (4 records), 'Commission Service' (8 records), 'Consultancy' (49 records), 'Non Commercial' (389 records), 'Technology Transfer' (1 record) and ⟨Void⟩ (1 record). The record with ⟨Void⟩ was identified as 'Non Company' by direct inspection.

Therefore, all records could be classified in one of the following levels: 'Non Companies' (41,317), 'Industry' (6,447), 'Other' (17,588) and 'Not Available' (12,346). The total number of records (77,698) is larger than the number of participants (25,287) since many of them collaborate in several projects. Then, it was necessary to verify if repeated records were always classified in the same level of 'Dedication'.

We found that many participants were classified in different levels, thus we had to define a set of rules which eliminated this ambiguity. Hence, the following step was to study each level to understand their composition. For every level, we chose 100 records randomly to check by direct inspection their dedication. The result was that all selected records in 'Industry' were companies, any in 'Non Companies', 95 in 'Other' and 55 in 'Not Available'.

With the former information, we proceeded as follows. We first defined for each participant a vector D={'Non Companies', 'Industry', 'Other', 'Not Available'}, where the components are the number of times that it is classified in that level. For instance, D={17, 0, 8, 4} means that the participant appears 17 times as 'Non Company', 8 as 'Other' and 4 as 'Not Available'. Then, we decided that vectors in the form {a, 0, 0, 0} or {a, 0, 0, d} were Universities and vectors in the form {0, b, c, d}, {0, b, c, 0}, {0, b, 0, d} and {0, b, 0, 0} were Companies. With only these sensible rules, we managed to classify 22,001 participants (87%).

In order to confirm this result and to classify the remaining 3,286 entities, we defined a filter based in keywords relative to the Universities group, such as 'univer', 'schule', 'laborato'... When we focused our attention in the group of 22,001 participants classified using 'Dedication', we found that those classified as Universities according to the filter were also Universities according to 'Dedication'. Since the filter was a completely different manner of splitting the dataset, we could use it for the rest of the entries. Note that we only believed the result of the filter if it was University, not if the result was Company. This is reasonable since the filter was designed to identify terms related to Universities, not to Companies.

By means of the filter we classified all participants but 309. To place these entities, we paid attention to which value was higher: 'Non Companies' or 'Industry', independently of the other two values. If the value 'Non Companies' was higher, it was a University, otherwise it was a Company.

Appendix B

Results on the single queue models

B.1 Exact solution of the priority queue model with $L = 2$

Consider the model discussed in Section 5.6 [219] with $L = 2$ [221]. The task that has been just selected and its priority has been reassigned will be called the new task, while the other task will be called the old task. Let $\eta(x)$ and $R(x) = \int_0^x dx\tilde{\eta}(x)$ be the priority probability density function (pdf) and distribution function of the new tasks, which are given. In turn, let $\tilde{\eta}(x,t)$ and $\tilde{R}(x,t) = \int_0^x dx\tilde{\eta}(x,t)$ be the priority pdf and distribution function of the old task in the t-th step. At the $(t+1)$-th step there are two tasks on the list, their priorities being distributed according to $R(x)$ and $\tilde{R}(x,t)$, respectively. After selecting one task the old task will have the distribution function

$$\tilde{R}(x,t+1) = \int_0^x dx'\tilde{\eta}(x',t)q(x') + \int_0^x dx'\eta(x)\tilde{q}(x',t) \ , \tag{B.1}$$

where

$$q(x) = p[1 - R(x)] + (1-p)\frac{1}{2} \tag{B.2}$$

is the probability that the new task is selected given the old task has priority x, and

$$\tilde{q}(x) = p[1 - \tilde{R}(x,t)] + (1-p)\frac{1}{2} \tag{B.3}$$

is the probability that the old task is selected given the new task has priority x. In the stationary state, $\tilde{R}(x,t+1)=\tilde{R}(x,t)$, thus from (B.1) we obtain

$$\tilde{R}(x)=\frac{1+p}{2p}\left[1-\frac{1}{1+\frac{2p}{1-p}R(x)}\right] . \tag{B.4}$$

Next we turn our attention to the waiting time distribution. Consider a task with priority x that has just been added to the queue. The selection of this task is independent from one step to the other. Therefore, the probability that it waits τ_w steps is given by the product of the probability that it is not selected in the first τ_w-1 steps and that it is selected in the τ_w-th step. The probability that it is not selected in the first step is $\tilde{q}(x)$, while the probability that it is not selected in the subsequent steps is $q(x)$. Integrating over the new task's possible priorities we obtain

$$P(\tau_w)=\begin{cases} \int_0^\infty dR(x)\left[1-\tilde{q}(x)\right] , & \tau_w=1 \\ \\ \int_0^\infty dR(x)\tilde{q}(x)\left[1-q(x)\right]q(x)^{\tau_w-2} , & \tau_w>1 \end{cases} \tag{B.5}$$

Using (B.2)-(B.4) and integrating (B.5) we finally obtain

$$P(\tau_w)=\begin{cases} 1-\frac{1-p^2}{4p}\ln\frac{1+p}{1-p} , & \tau_w=1 \\ \\ \frac{1-p^2}{4p(\tau_w-1)}\left[\left(\frac{1+p}{2}\right)^{\tau_w-1}-\left(\frac{1-p}{2}\right)^{\tau_w-1}\right] , & \tau_w>1 \end{cases} \tag{B.6}$$

Note that $P(\tau_w)$ is independent of the $\eta(x)$ pdf from which the tasks are selected. Indeed, what matters for task selection is their relative order with respect to other tasks, resulting that all dependences in (B.2)-(B.4) and (B.5) appears via $R(x)$.

B.2 The asymptotic characteristics of $P(\tau_w)$

In Section 5.6 we focused on a model with fixed queue length L, demonstrating that it belongs to a new universality class with $\alpha=1$. Next we derive a series of results that apply to any queuing model that has a *finite queue length*, and is characterized by an *arbitrary task selection protocol* [221]. In each time step there are L tasks in the queue and one of them is executed. Therefore

$$\sum_{i=1}^{t} \tau_i + \sum_{i=1}^{L-1} \tau_i' = Lt \ , \tag{B.7}$$

where τ_i is the waiting time of the task executed at the i-th step and τ_i', $i = 1, \ldots, L-1$, is the time interval that task i, that is still active at the t-th step, has already spent on the queue. The first term in the l.h.s. of (B.7) corresponds to the sum of the waiting times experienced by the t tasks that were executed in the t steps since the beginning of the queue, while the second term describes the sum of the waiting times of the $L-1$ tasks that are still on the queue after the t step. Given that in each time step each of the L tasks experience one time step delay, the sum on the l.h.s. should equal Lt. From (B.7) it follows that

$$\langle \tau_w \rangle \equiv \lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} \tau_i = L - \lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{L-1} \tau_i' \ . \tag{B.8}$$

If all active tasks have a chance to be executed sooner or later, like the case for the model studied in Section 5.6 in the $0 \leq p < 1$ regime [219], we have $\langle \tau_w' \rangle \leq \langle \tau_w \rangle$ and the last term in (B.8) vanishes when $t \to \infty$. In contrast, for $p = 1$ the numerical simulations [219] indicate that after some transient time the most recently added task is always executed, while $L-1$ tasks remain indefinitely in the queue. In this case $\tau_i' \sim t$ in the $t \to \infty$ limit and the last term in (B.8) is of the order of $L-1$. Based on these arguments we conjecture that the average waiting time of executed tasks is given by

$$\langle \tau_w \rangle = \begin{cases} L \ , & 0 \leq p < 1 \\ 1 \ , & p = 1 \ , \end{cases} \tag{B.9}$$

which is corroborated by numerical simulations (see Fig. 5.5b).

It is important to note that the equality in (B.8) is independent of the selection protocol, allowing us to reach conclusions that apply beyond the model discussed in Section 5.6. From (B.8) we obtain

$$\langle \tau_w \rangle \leq L \ . \tag{B.10}$$

From this constraint follows that $P(\tau_w)$ must decay faster than τ_w^{-2} when $\tau_w \to \infty$, otherwise $\langle \tau_w \rangle$ would not be bounded. Indeed, it is easy to see that for any $\alpha < 2$ the average waiting time $\langle \tau_w \rangle$ diverges for Eq. (5.2). Thus, when $\tau_w \to \infty$, we must either have

$$P(\tau_w) \sim a\tau_w^{-\alpha} \ , \alpha > 2 \tag{B.11}$$

or

$$P(\tau_w) = \tau_w^{-\alpha} f\left(\frac{\tau_w}{\tau_0}\right) \ , \tag{B.12}$$

where $\tau_0 > 0$ and $f(x) = \mathcal{O}(bx^{\alpha-2})$ when $x \to \infty$, where b is a constant. That is, each time an $\alpha < 2$ exponent is observed (as it is for the empirical data discussed in Section 5.3), an exponential cutoff must accompany the scaling. For example, for the model discussed above with $L = 2$ and $0 \leq p < 1$ we have $\alpha = 1$ and $f(x)$ decays exponentially (5.10), in line with the constraint discussed above.

B.3 Transitions between the two universality classes

A basic difference between the models discussed in Section 5.5 and Section 5.6 is the capacity of the queue. Our results indicate that the model without limitation on the queue length displays $\alpha = 3/2$, rooted in the fluctuations of the queue length. In contrast, the model with fixed queue length (Section 5.6) has $\alpha = 1$, rooted in the queuing of the low priority tasks on the priority list. If indeed the limitation in the queue length plays an important role, we should be able to develop a model that can display a transition from the $\alpha = 3/2$ to the $\alpha = 1$ universality class as we limit the fluctuations in the queue length. In this section we study such a model, interpolating between the two observed scaling regimes. We start from the model discussed in Section 5.5, and impose on it a maximum queue length L. This can be achieved by altering the arrival rate of the tasks: when there are L tasks in the queue no new tasks will be accepted until at least one of the tasks is executed. Mathematically this implies that the arrival rate depends on the queue length as

$$\lambda_l = \begin{cases} \lambda \ , & 0 \leq l < L \\ 0 \ , & l = L \ . \end{cases} \tag{B.13}$$

In the stationary state the queue length distribution $P(l)$ satisfies the balance equation

$$\lambda_{l-1} P(l-1) + \mu_{l+1} P(l+1) = (\lambda_l + \mu_l)\, P(l) \ , \tag{B.14}$$

where

$$\mu_l = \begin{cases} 0 \ , & l = 0 \\ \mu \ , & 0 < l \leq L \ . \end{cases} \tag{B.15}$$

From (B.14) we obtain the queue length distribution as

$$P(l) = \frac{1-\rho}{1-\rho^{L+1}} \rho^l \ , \tag{B.16}$$

suggesting the existence of three scaling regions.

Subcritical regime, $\rho \ll 1$: If the arrival rate of the tasks is much smaller than the execution rate, the fact that the queue length has an upper bound has little significance, since l will rarely reach its upper bound L, but will fluctuate in the vicinity of $l = 0$. This regime can be reached either for $\rho \ll 1$ and L fixed or for $\rho < 1$ and $L \gg 1$. Therefore, in this case the waiting time distribution is well approximated by that of the model with an unlimited queue length, displaying the scaling predicted by Eq. (5.5), *i.e.* either exponential, or a power law with $\alpha = 3/2$, coupled with an exponential cutoff (see Fig. B.1a).

Critical regime: For $\rho = 1$ we observe an interesting interplay between the queue length and L. Normally in this critical regime $l(t)$ should follow a random walk with the return time probability density scaling with exponent 3/2. However, the limitation imposed on the queue length limits the power law waiting time distribution predicted by Eq. (5.5), introducing a cutoff (see Fig. B.1a). Indeed having the number of tasks in the queue limited allows each task to be executed in a finite time.

Supercritical regime: When $\rho \gg 1$ from (B.16) follows that

$$\mathcal{L}_l = \begin{cases} \mathcal{O}(\rho^{-1}) \ , & 0 \leq l < L \\ 1 - \mathcal{O}(\rho^{-1}) \ , & l = L \ , \end{cases} \tag{B.17}$$

i.e. with probability almost one the queue is filled. Thus, in the supercritical regime $\rho \gg 1$ new tasks are added to the queue immediately after a task is executed. If we take the number of executed tasks as a new reference time then this model corresponds to the one discussed in Section 5.6, displaying $\alpha = 1$ [219], as supported by the numerical simulations (see Fig. B.1b).

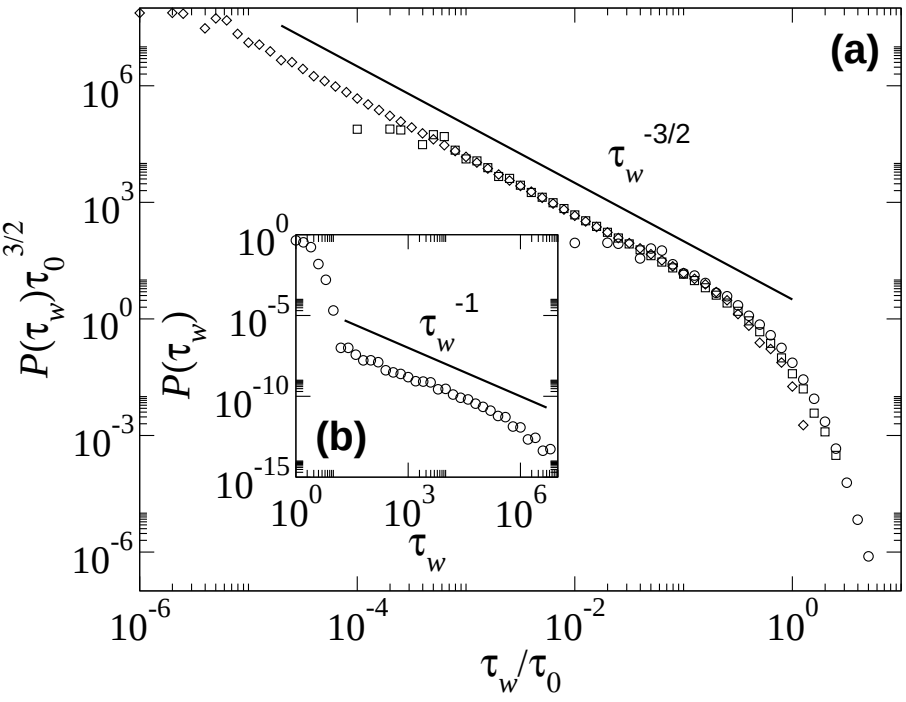

Figure B.1: Waiting time distribution for tasks in the queueing model discussed in Section B.3, with a maximum queue length L. The waiting time distribution is plotted for three L values: $L = 10$ (circles), $L = 100$ (squares) and $L = 1000$ (diamonds). The data has been rescaled to emphasize the scaling behavior $P(\tau_w) = \tau_w^{-3/2} f(\tau_w/\tau_0)$, where $\tau_0 \sim L^2$. In the inset we plot the waiting time for $\rho = 10^6$, showing the crossover to the model discussed in Section 5.6 in the limit $\rho \to \infty$ and L fixed.

Bibliography

[1] R. Solomonoff and A. Rapoport. Connectivity of random nets. Bulletin of Mathematical Biophysics **13**, 107 (1951).

[2] E. N. Gilbert. Random graphs. Ann. Math. Statist. **30**, 1141 (1959).

[3] P. Erdős and A. Rényi. On random graphs. Publicationes Mathematicae **6**, 290 (1959).

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

[5] S. Milgram. The small world problem. Psychology Today **2**, 60 (1967).

[6] D. J. de S. Price. Networks of scientific papers. Science **149**, 510 (1965).

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

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

[9] B. Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled random graphs. Eur. J. Comb. **1**, 311 (1980).

[10] B. Bollobás, *Random graphs* (Academic Press, London, 1985).

[11] R. J. Baxter. *Exactly solved models in statistical mechanics* (London, Academic Press, 1982).

[12] H. A. Bethe. Statistical Theory of Superlattices. Proc. Royal Soc. London A **150**, 552 (1935).

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

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

[15] A.-L. Barabási and R. Albert. Emergence of Scaling in Random Networks. Science **286**, 509 (1999).

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

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

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

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

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

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

[22] B.A. Huberman, P.L. Pirolli, J.E. Pitkow, and R.M. Lukose. Strong regularities in World Wide Web surfing. Science **280**, 95 (1998).

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

[24] B. A. Huberman, L. A. Adamic. Growth dynamics of the World-Wide Web. Nature **401**, **131**, (1999).

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

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

[27] H. Ebel, L.-I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. Phys. Rev. E **66**, R35103 (2002).

[28] J. P. K. Doye. Network topology of a potential energy landscape: A static scale-free network. Phys. Rev. Lett. **88**, 238701 (2002).

[29] Z. Toroczkai, and K. E. Bassler. Network dynamics: Jamming is limited in scale-free systems. Nature **428**, 716 (2004).

[30] Z. Toroczkai, B. Kozma, K. E. Bassler, N. W. Hengartner, and G. Korniss. Gradient networks. E-print: arXiv:cond-mat/0408262

[31] A. Scala, L. A. N. Amaral, and M. Barthélémy. Small-world networks and the conformation space of a short lattice polymer chain. Europhys. Lett. **55**, 594 (2001).

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

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

[34] J. Scott. *Social Network Analysis: A Handbook* (Sage, London, 2nd ed., 2000).

[35] M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. Proc. Nat. Acad. Sci. **99**, 2566 (2002).

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

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

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

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

[40] E. Almaas, B. Kovacs, T. Vicsek, Z. N. Oltvai, and A.-L. Barabási. Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature **427**, 839 (2004).

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

[42] R. Albert and A.-L. Barabási. Statistical Mechanics of complex networks. Rev. Mod. Phys. **74**, 47 (2002).

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

[44] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. Critical phenomena in complex networks. E-print: arXiv:0705.0010, accepted for publication in Rev. Mod. Phys.

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

[46] T. S. Evans. Complex networks. Contemporary Physics **45**, 455 (2004).

[47] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. Phys. Rep. **424**, 175 (2006).

[48] S. N. Dorogovtsev and J. F. F. Mendes. *Evolution of Networks: from Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).

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

[50] R. Durret. *Random Graph Dynamics* (Cambridge University Press, Cambridge, 2006).

[51] G. Caldarelli. *Scale-Free Networks* (Oxford University Press, Oxford, 2007).

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

[53] D. J. Watts. *Six Degrees: The Science of a Connected Age* (Norton, New York, 2003).

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

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

[56] M. Buchanan. *Nexus: Small Worlds and the Groundbreaking Science of Networks* (Norton, New York, 2002).

[57] *The Structure and Dynamics of Networks*, Editors: M. E. J. Newman, A.-L. Barabási, and D. J. Watts (Princeton University Press, Princeton, 2006).

[58] J. Clark and D. A. Holton. *A First Look at Graph Theory.* (World Scientific, Singapore, 1991).

[59] B. Bollobás. *Random Graphs* (Cambridge University Press, second edition, 2001).

[60] W. Feller. *An introduction to probability theory and its applications* Vol. 2 (Wiley, 2nd Edition, New York, 1971).

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

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

[63] G. Szabó, M. Alava, and J. Kertész. Shortest paths and load scaling in scale-free trees. Phys. Rev. E **66**, 026101 (2002).

[64] B. Bollobás and O. M. Riordan. *Mathematical results on scale-free random graphs*, in Handbook of Graphs and Networks: From the Genome to the Internet, S. Bornholdt and H.G. Schuster, eds., Wiley-VCH, Berlin, 2003, pp. 1-32.

[65] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Metric structure of random networks. Nucl. Phys. B **653**, 307 (2003).

[66] R. Sedgewick. *Algorithms* (Addison-Wesley, Reading, MA, 1988).

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

[68] M. A. Serrano and M. Boguñá. Clustering in complex networks. I. General formalism. Phys. Rev. E **74**, 056114 (2006).

[69] M. A. Serrano and M. Boguñá. Clustering in complex networks. II. Percolation properties. Phys. Rev. E **74**, 056115 (2006).

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

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

[72] A. Barrat and M. Weigt. On the properties of small-world networks. Eur. Phys. J. B, **13**, 547 (2000).

[73] T. Schank and D. Wagner. Approximating Clustering Coefficient and Transitivity. J. of Graph Algorithms and Applications **9**(2), 265 (2005).

[74] A. Vázquez. *Degree correlations and clustering hierarchy in networks: measures, origin and consequences*, PhD Thesis, (SISSA, 2002).

[75] A. Vázquez. Growing networks with local Rules: Preferential attachment, clustering hierarchy and degree correlations, Phys. Rev. E, **67**, 056104 (2003).

[76] A. Vázquez and Y. Moreno. Resilience to damage of graphs with degree correlations. Phys. Rev. E **67**, 015101 (2003).

[77] M. E. J. Newman. Assortative mixing in networks. Phys. Rev. Lett. **89**, 08701 (2002).

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

[79] A. Vázquez, R. Dobrin, D. Sergi, J.-P. Eckmann, Z. Oltvai, and A.-L. Barabási. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. Proc. Nat. Acad. Sci. **101**, 17940 (2004).

[80] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of Escherichia coli. Nat. Genet. **31**, 64 (2002).

[81] R. Milo, S. S. Shen-Orr, S. Itzkovitz, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. Science **298**, 824 (2002).

[82] S. Wuchty, Z. Oltvai, and A.-L. Barabási. Evolutionary conservation of motif constituents in the yeast protein interaction network. Nat. Genet. **35**, 118 (2003).

[83] R. E. Ulanowicz. Identifying the structure of cycling in ecosystems. Math. Bioscienc. **65**, 219 (1983).

[84] P. Bearman. Generalized Exchange. Am. J. Soc. **102**, 1383 (1997).

[85] M. Bernstein, Structural Patterns and Hypertext Rhetoric. ACM Computing Surveys **31** (1999).

[86] E. Marinari and R. Monasson. Circuits in random graphs: from local trees to global loops. J. Stat. Mech. **9**, 09004 (2004).

[87] L. C. Freeman. A set of measures of centrality based on betweenness. Sociometry **40**, 35-41 (1977).

[88] G. Sabidussi. The centrality index of a graph. Psychometrika **31**, 581 (1966).

[89] A. Rapoport and W. J. Horvath. A study of a large sociogram. Behavioral Sci. **6**, 279 (1961).

[90] T. J. Fararo and M. Sunshine. *A Study of a Biased Friendship Network* (Syracuse University Press, Syracuse, NY, 1964).

[91] J. Moody. Race, school integration, and friendship segregation in America. Amer. J. Sociol. **107**, 679 (2001).

[92] Kristina Lerman. Social Browsing & Information Filtering in Social Media. E-print: arXiv:0710.5697.

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

[94] M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. Proc. Nat. Acad. Sci. **101**, 5200 (2004).

[95] L. M. Branscomb, F. Kodama, and R. L. Florida. *Industrializing Knowledge: University-Industry Linkages in Japan and the United States* (MIT Press, Cambridge, 1999).

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

[97] A. Vázquez, J. G. Oliveira, and A.-L. Barabási. The inhomogeneous evolution of subgraphs and cycles in complex networks. Phys. Rev. E **71**, 025103(R) (2005).

[98] J. A. Almendral, J. G. Oliveira, L. López, M. A. F. Sanjuán, and J. F. F. Mendes. The interplay of universities and industry through the FP5 network. New J. Phys. **9**, 183 (2007).

[99] J. A. Almendral, J. G. Oliveira, L. López, M. A. F. Sanjuán, and J. F. F. Mendes. The network of scientific collaborations within the European framework programme. Physica A **384**, 675 (2007).

[100] 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).

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

[102] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Aberg. The web of human sexual contacts. Nature **411**, 907 (2001).

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

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

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

[106] W. Willinger, R. Govindan, S. Jamin, V. Paxson, and S. Shenker. Scaling phenomena in the Internet: Critically examining criticality. Proc. Nat. Acad. Sci. **99**, 2573 (2002).

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

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

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

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

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

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

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

[114] F. Benford. The law of anomalous numbers. Proc. Amer. Phil. Soc. **78**, 551 (1938).

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

[116] J.-P. Eckmann, E. Moses and D. Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. Proc. Nat. Acad. Sci. **101**, 14333 (2004).

[117] J. Balthrop, S. Forrest, M. E. J. Newman, and M. M. Williamson. Technological Networks and the Spread of Computer Viruses. Science **304**, 527 (2004).

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

[119] A. Vázquez, B. Rácz, A. Lukács, and A.-L. Barabási. Impact of Non-Poissonian Activity Patterns on Spreading Processes. Phys. Rev. Lett. **98**, 158702 (2007).

[120] S. Redner. Citation Statistics from 110 Years of Physical Review. Physics Today **58**, 49 (2005).

[121] A. Vázquez. Statistics of citation networks. E-print: arXiv: cond-mat/0105031 (2001).

[122] A. E. Motter and Y.-C. Lai. Cascade-based attacks on complex networks. Phys. Rev. E **66**, 065102 (2002).

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

[124] A. E. Motter. Cascade Control and Defense in Complex Networks. Phys. Rev. Lett. **93**, 098701 (2004)

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

[126] W. Aiello, F. Chung, and L. Lu. *Random evolution of massive graphs*, in Handbook of Massive Data Sets, J. Abello, P. M. Pardalos, and M. G. C. Resende, eds., pp. 97-122, Kluwer, Dordrecht (2002).

[127] Jukka-Pekka Onnela, Jari Saramäki, J. Hyvönen, G. Szabó, M. Argollo de Menezes, K. Kaski, A.-L. Barabási, and J. Kertész. Analysis of a large-scale weighted network of one-to-one human communication. New J. Phys. **9**, 179 (2007).

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

[129] S. N. Dorogovstev and J. F. F. Mendes. Language as an Evolving Word Web. Proc. Royal Soc. London B **268**, 2603 (2001).

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

[131] R. Ferrer i Cancho and R. V. Solé. Least effort and the origins of scaling in human language. Proc. Nat. Acad. Sci., **100**, 788 (2003).

[132] M. Steyvers and J. B. Tenenbaum. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model for Semantic Growth. Cognitive Science **29**, 41 (2005).

[133] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles. Metabolic net- work structure determines key aspects of functionality and regulation. Nature **420**, 190 (2002).

[134] R. Guimerà and L. A. N. Amaral. Functional cartography of complex metabolic networks. Nature **433**, 895 (2005).

[135] E. Ravasz, S. Gnanakaran, and Z. Toroczkai. Network Structure of Protein Folding Pathways. E-print: arXiv:0705.0912

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

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

[138] R. V. Solé and R. Pastor-Satorras. *Complex networks in genomics and proteomics*, in Handbook of Graphs and Networks: From the Genome to the Internet, S. Bornholdt and H.G. Schuster, eds., Wiley-VCH, Berlin, 2003, pp. 145-167.

[139] N. Guelzim, S. Bottani, P. Bourgine, and F. Kepes. Topological and causal structure of the yeast transcriptional regulatory network. Nat. Genet. **31**, 60 (2002).

[140] I. J. Farkas, H. Jeong, T. Vicsek, A.-L. Barabási, and Z. N. Oltvai. The topology of the transcription regulatory network in the yeast, Saccharomyces cerevisiae. Phys. A **381**, 601 (2003).

[141] S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of Escherichia coli. Nat. Genet. **31**, 64 (2002).

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

[143] Victor M. Eguiluz, Dante R. Chialvo, Guillermo A. Cecchi, Marwan Baliki, and A. Vania Apkarian. Scale-free brain functional networks. Phys. Rev. Lett. **94**, 018102 (2005).

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

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

[146] J. Camacho, R. Guimerà, and L. A. N. Amaral. Robust Patterns in Food Web Structure. Phys. Rev. Lett. **88**, 228102 (2002).

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

[148] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási. The human disease network. Proc. Nat. Acad. Sci. **104**, 8685 (2007).

[149] J. Loscalzo. I. Kohane, and A.-L. Barabási. Human disease classification in the postgenomic era: A complex systems approach to human pathobiology. Molecular Systems Biology **3**, 179 (2007).

[150] V. Emilsson, G. Thorleifsson, B. Zhang *et. al.* Genetics of gene expression and its effect on disease. Nature **452**, 423 (2008).

[151] M. A. Yildirim, K.-I. Goh, M. E. Cusick, A.-L. Barabási, and M. Vidal. Drug-target network. Nat. Biotechnol. **25**, 1119 (2007).

[152] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. Random Structures and Algorithms **6**, 161 (1995).

[153] M. Molloy and B. Reed. The size of the giant component of a random graph with a given degree sequence. Combinatorics, Probability and Computing **7**, 295 (1998).

[154] G. U. Yule. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis. Phil. Trans. Royal Soc. London B **213**, 21 (1925).

[155] A.-L. Barabási, E. Ravasz, and T. Vicseck. Deterministic scale-free networks. Physica A **299**, 559 (2001).

[156] R. Cohen and S. Havlin. Scale-free networks are ultrasmall. Phys. Rev. Lett. **90**, 058701 (2003).

[157] A. Fronczak, P. Fronczak, and J. A. Holyst. Average path length in uncorrelated random networks with hidden variables. Phys. Rev. E **70**, 056110 (2004).

[158] J. A. Holyst, J. Sienkiewicz, A. Fronczak, P. Fronczak, and K. Suchecki. Universal scaling of distances in complex networks. Phys. Rev. E **72**, 026108 (2005).

[159] K. Malarz and K. Kulakowski. Dependence of the average to-node distance on the node degree for random graphs and growing networks. Eur. Phys. J. B **41**, 333 (2004).

[160] K. Nakao. Distribution of measures of centrality: Enumerated distributions of Freeman's graph centrality measures. Connections **13**, 10 (1990).

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

[162] J. D. Noh. Exact scaling properties of a hierarchical network model. Phys. Rev. E **67**, 045103 (2003).

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

[164] F. Comellas, G. Fertin, and A. Raspaud. Recursive graphs with small-world scale-free properties. Phys. Rev. E **69**, 037104 (2004).

[165] H. D. Rozenfeld, J. E. Kirk, E. M. Bollt, and D. ben-Avraham. Statistics of Cycles: How loopy is your network? J. Phys. A **38**, 4589 (2005).

[166] J. S. Andrade Jr., H. J. Herrmann, R. F. S. Andrade, and L. R. da Silva. Apollonian networks. Phys. Rev. Lett. **94**, 018702 (2005).

[167] J. P. K. Doye and C. P. Massen. Self-similar disk packings as model spatial scale-free networks. Phys. Rev. E **71**, 016128 (2005).

[168] S. N. Dorogovtsev and J. F. F. Mendes. Minimal models of weighted scale-free networks. E-print: arXiv:cond-mat/0408343 (2004).

[169] E. M. Bollt and D. ben-Avraham. What is special about diffusion on scale-free nets? New J. Phys. **7**, 26 (2005).

[170] B. Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled random graphs. Eur. J. Comb. **1**, 311 (1980).

[171] A. Bekessy, P. Bekessy, and J. Komlos. Asymptotic enumeration of regular matrices. Stud. Sci. Math. Hungar. **7**, 343 (1972).

[172] E. A. Bender and E. R. Canfield. The asymptotic number of labelled graphs with given degree sequences. J. Combinatorial Theory A **24**, 296 (1978).

[173] N. C. Wormald. The asymptotic connectivity of labelled regular graphs. J. Combinatorial Theory B **31**, 156 (1981).

[174] N. C. Wormald. The asymptotic distribution of short cycles in random regular graphs. ibid. **31**, 168 (1981).

[175] H. S. Wilf. *Generatingfunctionology*, (Academic Press, second edition, London, 1994).

[176] P. L. Krapivsky and S. Redner. Organization of growing random networks. Phys. Rev. E **63**, 066123 (2001).

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

[178] K. A. Eriksen, I. Simonsen, S. Maslov, and K. Sneppen. Modularity and extreme edges of the Internet. Phys. Rev. Lett. **90**, 148701 (2003).

[179] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Size-dependent degree distribution of a scale-free growing network. Phys. Rev. E **63**, 062101 (2001).

[180] G. Bianconi, G. Caldarelli, and A. Capocci. Number of h-cycles in the Internet at the Autonomous Systems level. E-print: arXiv:cond-mat/0310339.

[181] G. Bianconi, G. Caldarelli, and A. Capocci. Loops structure of the Internet at the Autonomous System Level. E-print: arXiv:cond-mat/0408349.

[182] J. G. Oliveira (unpublished).

[183] D. Sergi. Random graph model with power-law distributed triangle subgraphs. E-print: arXiv:cond-mat/0412472.

[184] S.-H. Yook and H. Jeong, (unpublished).

[185] G. Bianconi and A. Capocci. Number of Loops of Size h in Growing Scale-Free Networks. Phys. Rev. Lett. **90**, 078701 (2003).

[186] T. Petermann and P. de los Rios. Role of clustering and gridlike ordering in epidemic spreading. Phys. Rev. E **69**, 066116 (2004).

[187] D. Amidon, Baltic Dynamics 2004 Conference, Riga (2004).

[188] Y. Caloghirou, A. Tsakanikas, and N. S. Vonortas. University-Industry Cooperation in the Context of the European Framework Programmes. Journal of Technology Transfer **26**, 153 (2001).

[189] F. Meyer-Krahmer and U. Schmoch. Science-based technologies: university-industry interactions in four fields. Research Policy **27**, 835 (1998).

[190] H. Wigzell. Framework programmes evolve. Science Editorial **295**, 443 (2002).

[191] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. Contemporary Physics **46**, 323 (2005).

[192] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. E-print: arXiv:0706.1062.

[193] S. N. Dorogovtsev and J. F. F. Mendes. *Accelerated growth of networks*, in Handbook of Graphs and Networks: From the Genome to the Internet, S. Bornholdt and H.G. Schuster, eds., Wiley-VCH, Berlin, 2003, pp. 318-341.

[194] S. N. Dorogovtsev, J. F. F. Mendes, and J. G. Oliveira. Degree-dependent intervertex separation in complex networks. Phys. Rev. E **73**, 056122 (2006).

[195] S. Chaoming, S. Havlin, and H. E. Makse. Self-similarity of complex networks. Nature **433**, 392 (2005).

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

[197] M. E. J. Newman. Fast algorithm for detecting community structure in networks. Phys. Rev. E **69**, 066133 (2004).

[198] The software package PAJEK can be found at `http://pajek.imfm.si/doku.php`.

[199] M. Boguñá, R. Pastor-Satorras, and A. Vespignani. Cut-offs and finite size effects in scale-free networks. Eur. Phys. J. B. **38**, 205 (2004).

[200] S. Newcomb. Note on the frequency of the use of digits in natural numbers. Amer. J. Math. **4**, 39(1881).

[201] R. A. Raimi. The peculiar distribution of first digits. Sci. Amer. **221**, 109 (1969).

[202] R. A. Raimi. The first digit problem. Amer. Math. Monthly **83**, 521 (1976).

[203] L. Pietronero, E. Tosatti, V. Tosatti, and A. Vespignani. Explaining the uneven distribution of numbers in nature: The laws of Benford and Zipf. Physica A **293**, 297 (2001).

[204] G. Levin, M. Wattenberg, J. Feinberg, S. Wynecoop, D. Becker, and D. Elashoff. The secret lives of numbers. ⟨`http://www.turbulence.org/Works/nums/`⟩ (2002).

[205] Google Inc., Google$^{\mathrm{TM}}$ search engine ⟨`http://www.google.com`⟩.

[206] G. K. Zipf. *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Cambridge, 1949).

[207] J. C. Willis. *Age and Area* (Cambridge University Press, Cambridge, 1922).

[208] B. B. Mandelbrot. *The Fractal Geometry of Nature* (Freeman, New York, 1977).

[209] P. Bak. *How Nature Works: The Science of Self-Organized Criticality* (Copernicus, New York, 1996).

[210] L. D. Landau and E.M. Lifshitz. *Statistical Physics*, Part 1 (Pergamon Press, New York 1993).

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

[212] M. Argollo de Menezes and A.-L. Barabási. Separating internal and external dynamics of complex systems. Phys. Rev. Lett. **93**, 068701 (2004).

[213] G A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review **63**, 81 (1956).

[214] N. Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity, Behavioral and Brain Sciences **24**, 87 (2001).

[215] A. Vázquez, J. G. Oliveira, Z. Dezső, K.-I. Goh, I. Kondor, A.-L. Barabási. Modeling bursts and heavy tails in human dynamics. Phys. Rev. E **73**, 036127 (2006).

[216] F. A. Haight. *Handbook of the Poisson Distribution* (Wiley, New York, 1967).

[217] A. K. Erlang. The Theory of Probabilities and Telephone Conversations and Telephone Waiting Times. Nyt. Tidsskrift for Matematik B (1909).

[218] H. R. Anderson. *Fixed Broadband Wireless System Design* (Wiley, New York, 2003).

[219] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. Nature **207**, 435 (2005).

[220] J. G. Oliveira and A.-L. Barabási. Human dynamics: Darwin and Einstein correspondence patterns. Nature **437**, 1251 (2005).

[221] A. Vázquez. Exact results for the Barabási model of human dynamics. Phys. Rev. Lett. **95**, 248701 (2005).

[222] Z. Dezső, E. Almaas, A. Lukács, B. Rácz, I. Szakadát, and A.-L. Barabási. Dynamics of information access on the web. Phys. Rev. E **73**, 066132 (2006).

[223] H. E. Stanley. *Introduction to phase transitions and critical phenomena*. (Oxford University Press, Oxford, 1987).

[224] S.-D. Poisson. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités.* (Bachelier, Paris, 1837).

[225] P. Reynolds. *Call Center Staffing*. (The Call Center School Press, Lebanon, TN, 2003).

[226] J. H. Greene. *Production and Inventory Control Handbook* (McGraw-Hill, New York, 3 ed, 1997).

[227] C. Dewes, A. Wichmann, and A. Feldman, An Analysis of Internet Chat Systems. *Proc. 2003 ACM SIGCOMM Conf. on Internet Measurement (IMC-03)*, ACM Press, New York, 2003).

[228] S. D. Kleban and S. H. Clearwater. Hierarchical Dynamics, Interarrival Times and Performance, *Proc. of SC'03*, November 15-21, 2003, Phonenix, AZ, USA.

[229] V. Paxson and S. Floyd, Wide-Area Traffic: The Failure of Poisson Modeling, *IEEE/ACM Tansactions in Networking* **3**, 226 (1995).

[230] F. Mainardi, M. Raberto, R. Gorenflo, and E. Scalas. Fractional calculus and continuous-time finance II: the waiting-time distribution. Physica A **287**, 468 (2000).

[231] M. Raberto, E. Scalas, F. Mainardi. Waiting-times and returns in high-frequency financial data: an empirical study. E-print: arXiv:cond-mat/0203596.

[232] V. Plerou, P. Gopikirshnan, L. A. N. Amaral, X. Gabaix, and H. E. Stanley. Economic fluctuations and anomalous diffusion. Phys. Rev. E **62**, R3023 (2000).

[233] J. Masoliver, M. Montero, and G. H. Weiss. Continuous-time random-walk model for financial distributions. Phys. Rev. E **67**, 021112 (2003).

[234] T. Henderson and S. Nhatti. Modelling user behavior in networked games, *Proc. ACM Multimedia 2001*, Ottawa, Canada, 212–220, 30 September–5 October (2001).

[235] U. Harder and M. Paczuski. Correlated dynamics in human printing behavior. Physica A **361**, 329 (2005).

[236] B. Rácz and A. Lukács. High density compression of log files. Data compression conference, IEEE Computer Society Press (2004).

[237] *The collected papers of Albert Einstein*, Vol. **1,5,8,9** (Princeton University Press, 1993-2004).

[238] *The correspondence of Charles Darwin*, Vol. **1-14** (Cambridge University Press, Cambridge, 1984-2004).

[239] Freud Museum, London, ⟨`http://www.freud.org.uk/`⟩.

[240] J. W. Cohen, *The Single Server Queue* (North Holland, Amsterdam, 1969).

[241] D. Gross and C. M. Harris. Fundamentals of queueing theory (Wiley, New York, 1998).

[242] M. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: evidence and possible causes. IEEE/ACM Trans. Netw. **5**, 835 (1997).

[243] M. Mitzenmacher. Dynamic models for file sizes and double Pareto distributions. Internet Mathematics **1**, 226 (2004).

[244] A. Baddeley. The Magical Number Seven: Still Magic after All These Years? Psychological Review **101**, 353 (1994); *ibid* **101**, 668 (1994).

[245] A. Cobham. Priority Assignment in Waiting Line Problems. J. Oper. Res. Soc. Amer. **2**, 70 (1954).

[246] J. Abate. Asymptotics for M/G/1 low priority waiting time tail probabilities. Queueing Systems **25**, 173 (1997).

[247] S. Redner. *A guide to first-passage processes* (Cambridge University Press, New York, 2001).

[248] *Diffusion and reactions in fractals and disordered systems*. Editors: D. Ben-Avraham and S. Havlin (Cambridge University Press, Cambridge, 2000).

[249] F. J. Omori. On the aftershocks of earthquakes. J. Coll. Sci. Imper. Univ. Tokyo **7**, 111 (1894).

[250] B.-F. Apostol. Euler's transform and a generalized Omori's law. Phys. Lett. A **351**, 175 (2006).

[251] G. M. Viswanathan, V. Afanasyev, S. V. Buldyrev, E. J. Murphy, P. A. Prince, and H. E. Stanley. Lévy flight search patterns of wandering albatrosses. Nature **381**, 413 (1996).

[252] S. Zapperi, A. Vespignani, and H. E. Stanley. Plasticity and avalanche behaviour in microfracturing phenomena. Nature **388**, 658 (1997).

[253] B. Suki, A.-L. Barabási, Z. Hantos, F. Peták, and H. E. Stanley. Avalanches and power-law behaviour in lung inflation. Nature **368**, 615 (1994).

[254] P. Bak and K. Sneppen. Punctuated equilibrium and criticality in a simple model of evolution. Phys. Rev. Lett. **71**, 4083 (1993).

[255] H. J. Jensen. *Self-Organised Criticality* (Cambridge University Press, Cambridge, 1998).

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

[257] M. Paczuski, S. Maslov, and P. Bak. Avalanche dynamics in evolution, growth, and depinning models. Phys. Rev. E **53**, 414 (1996).

[258] H. Flyvbjerg, K. Sneppen, and P. Bak. Mean field theory for a simple model of evolution. Phys. Rev. Lett. **71**, 4087 (1993).

[259] J. de Boer, B. Derrida, H. Flyvbjerg, A. D. Jackson, and T. Wettlig. Simple Model of Self-Organized Biological Evolution. Phys. Rev. Lett. **73** 906 (1994).

[260] P. Holme. Network reachability of real-world contact sequences. Phys. Rev. E **71**, 046119 (2005).

[261] G. Grinstein and R. Linkser. Biased Diffusion and Universality in Model Queues. Phys. Rev. Lett. **97**, 130201 (2006).

[262] A. Gabrielli and G. Caldarelli. Invasion Percolation and Critical Transient in the Barabási Model of Human Dynamics. Phys. Rev. Lett. **97**, 130201 (2006).

[263] P. Blanchard and M. O. Hongler. Modeling human activity in the spirit of Barabasi's queueing systems. Phys. Rev. E **75**, 026102 (2007).

[264] K.-I. Goh and A.-L. Barabási. Burstiness and memory in complex systems. Eur. Phys. Lett. **81**, 48002 (2008).

[265] A.-L. Barabási and H. E. Stanley. *Fractal Concepts in Surface Growth* (Cambridge University Press, Cambridge, 1995).

[266] C. A. Hidalgo. Conditions for the Emergence of Scaling in the Inter-Event Time of Uncorrelated and Seasonal Systems. Physica A **369**, 877 (2006).

[267] A. Vázquez. Impact of memory on human dynamics. Physica A **373**, 747 (2007).