UNIVERSITÁ DEGLI STUDI DI CATANIA
Dottorato di ricerca in Fisica XXIV ciclo

Roberta Sinatra

HIGH-ORDER MARKOV CHAINS IN COMPLEX NETWORKS: MODELS AND APPLICATIONS

Doctoral Thesis

Supervisor:
Prof. Vito Latora

Co-supervisor:
Dr. Jesùs Gòmez Gardeñes

December 2011

Wir dürfen nicht denen glauben,
die heute mit philosophischer Miene
und überlegenem Tone
den Kulturuntergang prophezeien
und sich in dem Ignorabimus gefallen.
Für uns gibt es kein Ignorabimus,
und meiner Meinung nach auch
für die Naturwissenschaft überhaupt nicht.
Statt des törichten Ignorabimus
heiße im Gegenteil unsere Losung:

Wir müssen wissen - wir werden wissen!

David Hilbert

We must not believe those,
who today, with philosophical bearing
and deliberative tone,
prophesy the fall of culture
and accept the ignorabimus.
For us there is no ignorabimus,
and in my opinion
none whatever in natural science.
In opposition to the foolish ignorabimus
our slogan shall be:

We must know - we will know!

David Hilbert

Contents

Abstract

Various complex systems such as the Internet and the World WideWeb, neural networks, the human society, chemical and biological systems are composed of highly interconnected dynamical units. Such systems can all be described as complex networks where the nodes represent the dynamic units, and two nodes are connected by an edge if the two units interact with each other. For most networks, the complexity arises from the fact that the structure is highly irregular, complex and dynamically evolving in time and that the observed patterns of interactions highly influence the behaviour of the entire system. However, despite this complexity, a large number of networks from diverse fields such as biology, sociology, economics or technology has been found to exhibit similar topological and dynamical features. In this thesis we study different aspects of the structure and dynamics of complex networks by using approaches based on Markov and high-order Markov models. Regarding the structure of complex networks, we address the problem of the presence of three-body correlations between the node degrees in networks. Namely, we introduce measures to evaluate three-body correlations by using a third-order Markov model, and we study them in a wide range of real datasets. Then, we investigate how these correlations influence various dynamical processes. Specifically, we focus on Biased Random Walks (BRW), a class of Markovian stochastic processes which can be treated analytically and which extend the well-known concept of Random Walk on a network. In a BRW, the motion of walkers is biased accordingly to a generic topogical or dynamical node property. In particular, we investigate the connection between node-correlations in a network and the entropy rate that can be associated to the BRWs on the network. We also show how it is possible to rephrase a BRW process on a network as a plain RW on another network having the same topology but different weights associated to the edges, and we propose a number of applications where this conversion proves to be useful. In the final part of the thesis we apply the theory of complex networks and high-order Markov models to analyze and model data sets in three different contexts. First, we introduce a method to convert ensembles of sequences into networks. We apply this method to the study of the human proteome database, to detect hot topics from online social dialogs, and to characterize trajectories of dynamical systems. Second, we study mobility data of human agents moving on a network of a virtual world. We show that their trajectories have long-time memory, and how this influences the diffusion properties of the agents on the network. Finally, we study the topological properties of networks derived by EEG recordings on humans that interact by playing the prisoner's dilemma game.

Publications and Author Contribution

The research reported in this thesis is based on the following publications, resulting from the collaboration between Roberta Sinatra, the author of this thesis, and the authors listed below:

- R. Sinatra, D. Condorelli, and V. Latora,
  Networks of motifs from sequences of symbols,
  Physical Review Letters **105**, 178702 (2010), arXiv:1002.0668.

- F. De Vico Fallani, V. Nicosia, R. Sinatra, L. Astolfi, F. Cincotti, D. Mattia, C. Wilke, A. Doud, V. Latora, B. He and F. Babiloni,
  Defecting or Not Defecting: How to "Read" Human Behavior during Cooperative Games by EEG Measurements,
  PloS One **5**, 12:e14187 (2010), arXiv:1101.5322.

- R. Sinatra, J. Gómez-Gardeñes, R. Lambiotte, V. Nicosia, and V. Latora,
  Maximal-entropy random walks in complex networks with limited information,
  Physical Review E **83**, 030103(R) (2011), arXiv:1007.4936.

- R. Lambiotte, R. Sinatra, M. Barahona, J.-C. Delvenne, T.S. Evans, and V. Latora,
  Flow graphs: interweaving dynamics on structure,
  Physical Review E **84**, 017102 (2011), arXiv:1012.1211.

- M. Szell, R. Sinatra, G. Petri, S. Thurner, and V. Latora
  Understanding mobility in a social petri dish,
  arXiv:1112.1220, *in review.*

- R. Sinatra, J. Gómez-Gardeñes, S. Meloni, V. Nicosia, and V. Latora
  Quantifying three-body degree correlations in complex networks,
  to be submitted.

The publications above correspond to references in the bibliography [1–6].

During her PhD, Roberta Sinatra also published the following works, not included in this thesis:

- S. Thurner, M. Szell, and R. Sinatra,
  Emergence of good conduct, scaling and Zipf laws in human behavioral sequences in an online world, accepted for publication in *PLoS one*, arXiv:1107.0392.
  Reference [7]

- R. Sinatra, J. Iranzo, J. Gómez-Gardeñes, M. Florìa, V. Latora, and Y. Moreno,
  The Ultimatum Game in Complex Networks,
  J. Stat. Mech. P09012 (2009), arXiv:0807.0750v3.
  Reference [8]

- R. Sinatra, F. De Vico Fallani, V. Latora, L. Astolfi, D. Mattia, F. Cincotti, and F. Babiloni
  Cluster structure of functional networks estimated from high-resolution EEG data,
  International Journal of Bifurcation and Chaos (IJBC), Vol. 19, No. 2 (2009) 665676, arXiv:0806.2840v1.
  Reference [9]

- R. Sinatra, J. Gómez-Gardeñes, Y. Moreno, D. Condorelli, L. M. Florìa, V. Latora
  Structural and dynamical properties of cellular and regulatory networks,
  in *Statistical Mechanics of Cellular Systems and Processes*, edited by M. H. Zaman, Cambridge University Press (2009).
  Chapter in [10]

- F. De Vico Fallani, R. Sinatra, L. Astolfi, D. Mattia, F. Cincotti, V. Latora, S. Salinari, M.G. Marciani, A. Colosimo, and F. Babiloni,
  Community structure of cortical networks in spinal cord injured patients,
  Conf. Proc. IEEE Eng. Med. Biol. Soc. 2008, 3995-8 (2008).
  Reference [11]

- R. Sinatra,
  Ecological Clustering Reveals Topological Structures of Complex Networks,
  HPC-Europa2 Final Report, 2010.

- R. Sinatra, C. Gokhale, E. Fille Legara,
  Evolutionary dynamics of fitness driven walkers on a graph,
  Proceeding paper for the *Complex Systems Summer School 2010*, Santa Fe Institue (USA),

Introduction

Science is a wonderful thing if one does not have to earn one's living at it.

Albert Einstein

Several systems in nature and in technology astonish us for their remarkable complexity. For example even the simplest form of life relies on hundreds of intricate biochemical reactions, with the product of one reaction acting as a substrate for another one. Larger organisms are characterized by thousands of cells communicating with each other, and the complexity becomes overwhelming if one considers the human brain, made up of approximately $10^{15} - 10^{17}$ connections between billions of neurons. Other examples of natural systems where complexity is a striking feature are ecosystems, where vegetal and animal species depend on the existence of each other, or the genetic information encoded in their DNA. Complexity however is not just a feature of life, but can be found also in many manmade systems. Among these systems there is for example the WorldWideWeb, consisting of webpages interlinked in a nontrivial, complex structure and whose complexity every internet surfer experiences daily. Eventually, the most intriguing and amazing complex system is probably the one where we, human beings, are the fundamental constituents: the human society.

Complex systems are commonly understood as systems composed of a large number of elementary units which as a whole exhibit properties not obvious from the properties of the individual parts. The microscopic interactions in the system lead to the *emergence* of macroscopic properties. The complexity of a system emerges from the behaviors of the numerous interacting simple elements, and the behavior of one element is usually different in isolation from when it is part of the larger system. Although many systems around us are complicated, not all are necessarily complex. For example, a car or an airplane are both objects made of many interacting parts, or the system of chemical reactions to produce a cleanser, but these systems are perfectly controlled since the functioning of each constituent, as well as the interactions between their different parts, are completely understood. What makes a system complex, or what at least all complex systems share, is the fact that an organization is present without any external organizing principle being applied.

The Importance of Being Correlated

Because of the importance of the interactions between the constituents of such systems, and because of the intrinsic strong correlations between the elements of the system, powerful tools for the understanding of complexity are provided by *complex network theory* and *information theory*.

Indeed, one of the simplest approaches to understand complex systems is to model them as graphs whose nodes represent the dynamical units (e.g. the neurons in a brain, individuals in a social system) and the links stand for the interactions between the units. Since the properties of these graphs are much different from the properties of regular lattices and random graphs, the standard models usually studied in mathematical graph theory, it is common to refer to *complex networks* when talking about the graph-backbone of complex systems. Complex network theory is the discipline studying the topological and dynamical features of these complex networks. Of course, modelling a system as a complex network can be a very strong approximation, since it means translating the interaction between two dynamical units, which is usually depending on time, space and many other details, into a simple binary number: the existence or absence of a link between the two corresponding nodes. Nevertheless, in many cases of practical interest, such an approximation provides a simple but still very informative representation of the entire system. Also, the large number of studies of real-world complex networks has revealed the important finding that, despite the inherent different nature of many networks, most of them are characterized by the same topological properties, suggesting the existence of universal mechanisms underlying many different complex systems.

Information theory, instead, is the mathematical framework which helps us to tackle the amount of information present in a system, making it possible to find hidden regularities in what, from a first approach, can appear to be random or highly disordered. The starting point of information theory relies on the definition of the concepts of random variables, stochastic processes and entropy. As we will see in more detail later, for example the strings of letters appearing in a text, the chain of aminoacids in a protein or the sequences of cities a salesman visits can all be seen as stochastic processes. One goal of information theory is trying to quantify how much memory on the past is needed in order to predict what will appear next. For instance, information theory is at work when our mobile phone suggests us how to complete a word while we are writing an SMS. In the case of complex systems, where interactions between constituents play a fundamental role, information theory allows us to understand how much information is encoded in these interactions. We can answer for example questions like "Which is the range and strength of correlations between aminoacids in proteins?" or "How random are the connections between nodes in a network?" or also "How much memory is there in patterns of human movements?", allowing us to gain a deeper understanding in many processes and structures underlying complex systems.

The contribution of this thesis to the field of complex systems is twofold: (i) by

widely using the framework of information theory, with emphasis on the concepts of high-order Markov chain and of entropy, we focus and develop theoretical models and analytical tools for the characterization of specific structural and dynamical properties of complex networks, and (ii) we apply some of these models, the framework of complex networks and in general concepts of stochastic processes, with emphasis on information theory, for the study and analysis of data of different real systems. In particular, by means of high-order Markov chains, we can develop tools able to detect particular patterns of correlations not explored so far, in many different contexts and kinds of data. High-order correlations, namely third-order correlations, are found in the patterns of connectivity in a large number of real-world networks (chapter 3). A particular "mix" of short- and long-range correlations are present in proteomes (chapter 5). Long memory processes, hence long-range correlations, are at the base of human movements (chapter 6). At the same time, concepts derived from information theory, like the entropy rate, were at the base of our investigation of the interplay between structure and dynamics for a class of processes on networks named biased random walks (chapter 4). Finally, we take advantage of the poweful framework of complex network theory to relate correlations in EEG data from people engaged in cooperative games to selfish behavior (chapter 7).

Outline of the Thesis

The thesis is divided into three parts.

The first part, made up of two chapters, provides the background concepts we use in the second and third part of the thesis. Chapter 1 is an introduction to complex networks. We give first the formal definition of graph, the mathematical representation of a network, and introduce the most important measures to characterize the topological properties of graphs. Successively, we discuss typical properties of real-world networks, and we conclude with a review of the main models to construct networks. In chapter 2, we provide the basics of information theory. First we introduce the notions of stochastic process, as well as of joint, conditional and relative entropy. We then illustrate the concepts of Markov chain and of high-order Markov chain and focus on their properties. In particular, we explain the key-concept of ergodicity and show how this impacts the growth of entropy, as expressed by the so-called entropy rate.

In the second part of the thesis, we focus on the study of the structure and dynamics of complex networks from a theoretical point of view. In particular, in chapter 3 we describe the formalism to study correlations between pairs of nodes in a network. As these two-body correlations have been found to be a particular feature in most real networks and play an important role in many processes, we address in this chapter the problem of measuring two-body correlations as well as higher order correlations. In particular, we illustrate how to extend the mathematical formalism to the study of three-body correlations, namely correlations in triplets of nodes, by using high-

order Markov chains. We also investigate the existence of three-body correlations in a number of real-world networks, assessing that they are present in most of them and that they are not negligible in respect to the correlations of lower order. The dynamics of networks are instead studied in chapter 4. We focus on a particular class of dynamical processes on networks, called random walks. After introducing the definition of random walk, and distinguishing between unbiased and biased random walks, we characterize them as markovian stochastic processes, and we use information theory to derive their properties. We also show how to associate an entropy rate to a random walk on a graph, and we prove the possibility of designing biased random walks with maximal entropy rate for a given graph by solely using information locally available on the graph constituents. In the end of the chapter, we also mention how to rephrase the problem of biased random walks in terms of unbiased random walks by redefining the underlying graph.

Finally, in the third and last part of the thesis we move to the analysis of real-world datasets by using approaches based on complex networks and information theory. In chapter 5, we first propose a method to convert ensembles of symbolic sequences into networks. This conversion, based on the use of high-order Markov chains, retains the short- and long-range correlations in the sequences and is shown to have the advantage to compact the most important information contained in the original data into an easier to handle object. The usefulness of the method is illustrated by means of applications in different contexts, namely to the collection of human proteins where we are able to uncover details on the protein biological function, to a set of short messages in a social platform with the aim of detecting "hot topics", and to quantifying chaos in trajectories of dynamical systems. Further, in chapter 6, we study a data set on mobility of players in an online game. The analysis of these data provides evidence that mobility is influenced not only by spatial constraints, but also by socio-economic factors, and that human mobility patterns exhibit long-range correlations. We construct a long-term memory model that captures the statistical properties observed in the data. Finally, in chapter 7, we present results from an experimental study on human interaction, conducted in collaboration with a neuroscience laboratory. The experiment consists in recording EEG data of electrical brain activities between pairs of individuals who are playing the Prisoner's Dilemma game. Performing an analysis on this EEG data within the framework of complex networks, we demonstrate the striking possibility to predict whether a player will cooperate with his/her co-player.

Chapter 1

The ABC of Complex Networks

The important thing in science is not so much to obtain new facts, as to discover new ways of thinking about them.

William Bragg

Networks can be represented as graphs, and graph theory [12–14] offers the framework for the exact mathematical treatment of complex networks. In the first part of this chapter, we will introduce definitions and notations for the study of graphs. We will discuss the most important properties characterizing the graphs of many real-world networks in the second part of the chapter, while the most important models to construct graphs will be revised in the third part.

1.1 Definitions and measures

1.1.1 Notation

An *undirected (directed) graph* $G = (\mathcal{N}, \mathcal{L})$ consists of two sets $\mathcal{N}$ and $\mathcal{L}$, such that $\mathcal{N} \neq \emptyset$ and $\mathcal{L}$ is a set of unordered (ordered) pairs of elements of $\mathcal{N}$. The elements of $\mathcal{N} \equiv \{n_1, n_2, ..., n_N\}$ are called *nodes* or *vertices*, while the elements of $\mathcal{L} \equiv \{l_1, l_2, ..., l_K\}$ are *links* or *edges*. The number of elements in $\mathcal{N}$ and $\mathcal{L}$ is usually denoted by N and K, respectively. We assume that the graph has no multiple links, meaning that all the elements of the set $\mathcal{L}$ are different from each other.

A node is usually denoted with a number $i = 1, \cdots, N$, which is the order of the node in the set $\mathcal{N}$. In a undirected graph, a link is individuated by a couple of nodes i and j, and it is denoted as (i, j) or l_{ij}. The link is said to connect the two nodes i and j, or also to be incident in i and j.

Two nodes connected by a link are said to be *adjacent* or *neighboring* nodes. In a directed graph, the order of the two nodes forming a links matters: (i, j) stands for a

link from i to j, which is different from (j, i), standing for a link from j to i. Two links between the same pair of nodes but with different directions may occur simultaneously. Often links of a directed graph are referred to as *arcs*. In an undirected graph with N nodes, the total number of links K is a number between 0 and $N(N-1)/2$. In the case of a directed graph, the maximum number of links is equal to $N(N-1)$. When a graph has a number of links equal to the maximum, the graph is said to be a *fully connected* graph.

1.1.2 Representations of a graph

A graph can be visually represented by drawing dots or small circles, corresponding to nodes, and by connecting two dots by a line if between the corresponding nodes a link occurs. In Fig. 1.1, an example of an undirected graph and of a directed graphs are drawn in panels (a) and (d) respectively. Although the visual representation can be very helpful to get a first idea of the structure of a graph, it can be used only in a few cases, when the graph has a small number of nodes and of links. A more powerful representation is provided by the so-called *adjacency* matrix $\mathcal{A}$. This is a square matrix of dimension N, whose generic entry a_{ij} is either one, if a link exists between i and j, or zero, if no link occurs. In the case of a directed graph, a_{ij} is 1 if there is a link going from i to j and zero otherwise. In panel (b) of Fig. 1.1 the adjacency matrix of the graph in (a) is shown, while the adjacency of the graph in (d) is reported in panel (e). Notice that the adjacency matrix of the undirected graph is symmetric, while it is not for the directed graph.

While the adjacency matrix is very convenient for analytical calculations and for theoretical proofs, it is not practical for numerical computations. In fact, since the adjacency matrix of most real-world network is sparse, meaning that the number of non-zero entries of the matrix is of the same order of N, the computer representation of the graph in terms of adjacency matrix stores also many useless zero-entries. A more compact representation, saving considerable storage space, is the *ij-form*, also called *edge list*. In this representation, the graph is encoded as a $K \times 2$ matrix, where each row contains two entries corresponding to the ending nodes i and j of one of the link (i, j) of the graph. For an undirected graph, the edge list has dimension $2K \times 2$ if one considers pairs of nodes has ordered. This is a less compact form, but it avoids misunderstanding if nothing about the directed/undirected nature of the links is a priori specified. Notice that in the edge list representation, there is no information about the presence of isolated nodes, i.e. nodes with no links. In Fig. 1.1(c) and (f), examples of edge lists, associated to the graphs of panels (a) and (d), are shown.

1.1.3 Path, walk, cycle

A central concept in graph theory is that of reachability of two different nodes of a graph. In fact, two nodes that are not adjacent may nevertheless be reachable from

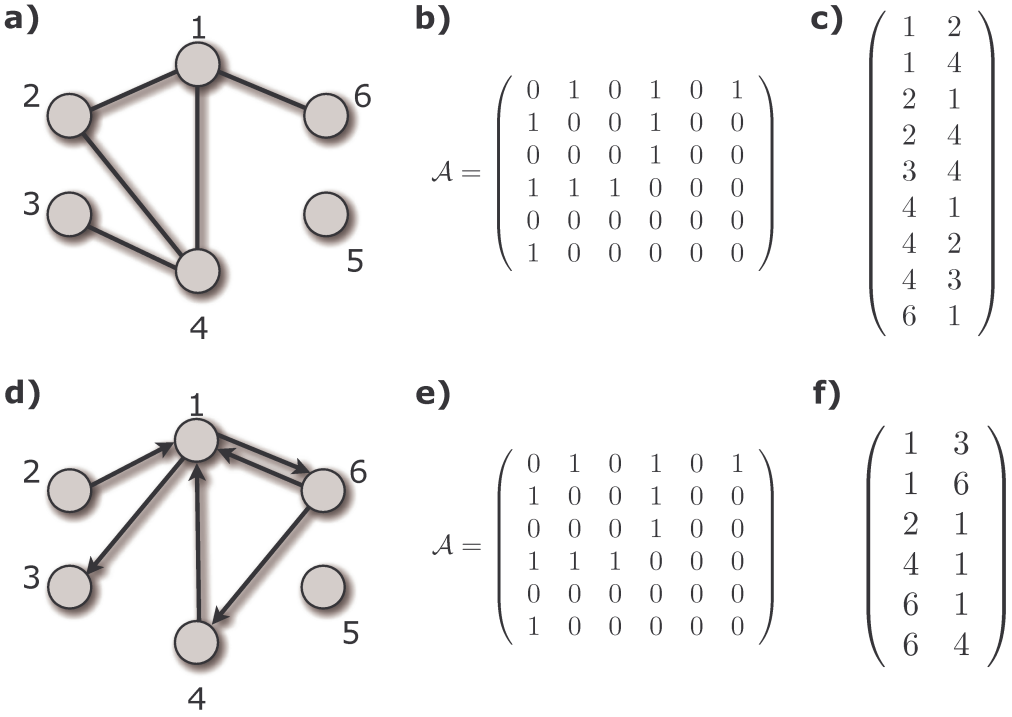

Figure 1.1: Representations of an undirected (upper panels) and of a directed (bottom panels) graph. In a) an undirected graph of 6 nodes and 5 links is shown, while in (b) we report the corresponding adjacency matrix and in (c) the corresponding edge list. In this edge list, each pairs of connected nodes is reported twice, as ordered pairs are considered. This representation is somehow redundant since a link of an undirected graph can be in principle individuated just by its ending nodes, without specifying their order. A more compact representation, where each link is specified only once, can be used if one knows a priori that the graph is undirected. In (d) a directed graph of 6 nodes and 6 arcs is shown. The corresponding adjacency matrix is reported in panel (e), the corresponding edge list in panel (f). Notice that in both the graphs node 5 is an isolated node (i.e. a node without links), which is reflected in the 5th column and 5th row of the adjacency matrixes being made of zeros.

one to the other. A *walk* from node i to node j is an alternating sequence of nodes and edges (a sequence of adjacent nodes) that begins with i and ends with j. The length of the walk is defined as the number of edges in the sequence. A *trail* is a walk in which no edge is repeated. A *path* is a walk in which no node is visited more than once. The walk of minimal length between two nodes is known as *shortest path* or *geodesic* (see also Sec. 1.1.4). A *cycle* is a closed walk, of at least three nodes, in which no edge is repeated. A cycle of length k is usually said a k-cycle and denoted as C_k. C_3 is a triangle ($C_3 = K_3$), C_4 is called a quadrilateral, C_5 a pentagon, and so on. A graph is said to be *connected* if, for every pair of distinct nodes i and j, there is a path from i to j, otherwise it is said *unconnected* or *disconnected*. A *component* of the graph is a maximally connected induced subgraph. A *giant component* is a component whose

size is of the same order as N.

1.1.4 Shortest Path

In a graph, the distance between two elements of the network, d_{ij}, is defined as the length of the geodesic, i.e. the shortest path, that goes from node i to j. In a unconnected graph, the geodesic path between two nodes belonging two different component is said to be infinite. One can then construct the distance matrix D so that its ij-entry is equal to d_{ij}. This matrix is symmetric in the case of an undirected graph, while it is in general asymmetric for directed graphs. Based on the distance matrix D, two global graph measures can be defined: *diameter* and *average path length*. The diameter of a graph is the longest geodesic between any pair of nodes in the graph for which a path actually exists. The average path length, usually denoted with L, is the mean of the geodesic path lengths between all the pairs of nodes in the graph, hence

$$L = \frac{1}{N\left(N-1\right)} \sum_{i,j \in \mathcal{N}} d_{ij}. \tag{1.1}$$

1.1.5 Degree of a node

The *degree* or *connectivity* k_i of a node i is defined as the number of edges incident in i. In terms of the adjacency matrix $\mathcal{A}$, the degree can be defined as:

$$k_i = \sum_{j \in \mathcal{N}} a_{ij} \ . \tag{1.2}$$

In the undirected graph of Fig. 1.1, for example, $k_1 = 3$ and $k_5 = 0$. If the graph is directed, the degree of the node is of two different kinds: the out-degree $k_i^{out} = \sum_j a_{ij}$, i.e. the number of links outgoing from the node, and the in-degree $k_i^{in} = \sum_j a_{ji}$, i.e. the number of links incoming in the node. Then, a total degree k_i can be defined as the sum of the in- and out-degree $k_i = k_i^{out} + k_i^{in}$. In the directed graph of Fig. 1.1, $k_1^{out} = 2$, $k_1^{in} = 3$ and $k1 = k_1^{out} + k_1^{in} = 5$.

In the case of undirected graphs, another quantity of interest is the average connectivity of the first neighbors of a node i, denoted as $k_{nn}\left(i\right)$. Formally, this quantity can be written as:

$$k_{nn}\left(i\right) = \frac{\sum_{j=1}^{N} a_{ij} \sum_{m=1}^{N} a_{jm}}{\sum_{j=1}^{N} a_{ij}} = \frac{1}{k_i} \sum_j a_{ij} k_j \ . \tag{1.3}$$

In the undirected graph of Fig. 1.1 the average connectivity of the first neighbors of node 1 is $k_{nn}\left(1\right) = \frac{k_2 + k_4 + k_6}{3} = 2$.

1.1.6 Degree distribution

One of the most basic topological properties of a graph G is its *degree distribution* $P(k)$. It is defined as the probability that a randomly chosen node of the graph has degree k or, equivalently, as the fraction of nodes in the graph having degree k. In this thesis we will also indicate the degree distribution, with the symbol P_k, which stresses the fact that k is a discrete variable.

In the case of directed networks, one has to consider two distributions, $P(k^{in})$ and $P(k^{out})$. To compute the degree distribution P_k of a real network, one has to count the number of nodes N_k which have the same connectivity k. Then, it will be $P_k = N_k/N$, where N is, as usual, the total number of nodes in the graph.

As we will see in following sections of this chapter and in other parts of this thesis as well, important information on the graph can be derived from the statistical moments of the degree distribution $P(k)$. The n-moment of $P(k)$ is defined as:

$$\langle k^n \rangle = \sum_k k^n P(k). \tag{1.4}$$

The first moment $\langle k \rangle$ is the mean degree of G. The second moment measures the fluctuations of the connectivity distribution.

1.1.7 Degree-degree correlations

The degree distribution completely determines the statistical properties of uncorrelated networks. However a large number of real networks are *correlated* in the sense that the probability that a node of degree k is connected to another node of degree, say k', depends on k. In these cases, it is necessary to introduce the *conditional* probability $P(k'|k)$, being defined as the probability that a link from a node of degree k points to a node of degree k'. $P(k'|k)$ satisfies the normalization $\sum_{k'} P(k'|k) = 1$, and the degree detailed balance condition $kP(k'|k)P(k) = k'P(k|k')P(k')$ [15]. For uncorrelated graphs, in which $P(k'|k)$ does not depend on k, the detailed balance condition and the normalization give $P(k'|k) = k'P(k')/\langle k \rangle$.

In general, the computation of the matrix $P(k'|k)$ yields noisy results, mainly due to finite size effects. Because of this, in order to investigate the presence of correlations, it is convenient to compute other quantities, such as the average degree of the neighbours of nodes with connectivity k, $\langle k_{nn} \rangle (k)$. This quantity is in fact related to the conditional probability $P(k'|k)$ by means of the following formal definition:

$$\langle k_{nn} \rangle (k) = \sum_{k'} k' P(k'|k) \tag{1.5}$$

Although the rigorous definition of $\langle k_{nn} \rangle (k)$ is the one given above, the average degree of the neighbours of nodes with connectivity k is nothing else than the average

connectivity of the neighbors of a node, as expressed by Eq. 1.3, averaged over all nodes belonging to the same degree class, i.e. having the same connectivity k:

$$\langle k_{nn} \rangle (k) = \frac{1}{N_k} \sum_i k_{nn} (i) \, \delta (k_i, k)$$

where $\delta (k', k'')$ is the kronecker delta and N_k is the number of nodes with the same degree k.

When a graph has no degree correlations, making use of the relation $P(k'|k) = k'P(k')/\langle k \rangle$ and of Eq. 1.4, one can easily prove that $\langle k_{nn} \rangle (k) = \langle k^2 \rangle / \langle k \rangle$ and thus that $\langle k_{nn} \rangle (k)$ does not depend on k. However this is not the case in most real-networks, which do have degree-degree correlations. In particular, it has been shown [13, 16] that in many real networks one finds that $\langle k_{nn} \rangle (k) \sim k^{-\nu}$. When $\langle k_{nn} \rangle (k)$ is an increasing (decreasing) function of k, happening for $\nu < 0$ ($\nu > 0$), the network is said to be assortative (dissasortative). The assortativity denotes the tendency of nodes of similar degree to connect with each other, while in dissasortative networks highly connected nodes tend to be linked to low degree ones. It has been observed that social networks tend to be assortative, while technological and biological networks show usually dissasortative patterns.

Sometimes it is useful to measure the overall degree-degree correlations in a network, meaning that one summarizes with one value the presence and magnitude of degree-degree correlations. Such a measure is provided by the *assortativity mixing coefficient* r [14, 17], which is defined as:

$$r = \frac{\sum_{k_i,k_j} k_i k_j \left(E_{k_i,k_j}/2K - q_{k_i} q_{k_j} \right)}{\sigma_q^2} \tag{1.6}$$

where E_{k_i,k_j} is the number of edges connecting nodes of degree k_i and k_j, q_k is the distribution of the so-called remaining degree and is $q_k = \frac{(k_j+1)P_{k+1}}{\sum_{k_i} k_i p_{k_i}}$, P_k is the degree distribution and σ_q^2 is the variance of the distribution q_k. With this definition, $r \in [-1, 1]$. In a network with no assortative (or disassortative) correlations $E_{k_i,k_j}/2K$ takes the value $q_{k_i} q_{k_j}$, and the coefficient $r = 0$. If there are correlations instead, $E_{k_i,k_j}/2K$ will differ from this value and r, which is nothing else than the Pearson correlation coefficient of the degrees at either ends of an edge, will indicate how (dis)assortative a network is. A positive value of r, indicates that the presence of assortativity, i.e. an overall tendency of nodes with high (low) degrees connecting to nodes with high (low) degrees. A negative value of r means instead that the network is disassortative, i.e. there is a tendency for nodes with high (low) degrees to connect to nodes with low (high) degrees.

1.1.8 Clustering coefficient

Clustering, also known as transitivity, is a typical property of many social networks, like the acquaintance network, where two individuals with a common friend are likely to know each other [18]. In terms of a generic graph G, transitivity means the presence of a high number of triangles. This can be quantified by defining the *transitivity* T of the graph as the relative number of transitive triples, i.e. the fraction of connected triples of nodes (triads) which also form triangles [19]:

$$T = \frac{3 \times \#\text{ of triangles in G}}{\#\text{ of connected triples of vertices in G}} \tag{1.7}$$

The factor 3 in the numerator compensates for the fact that each complete triangle of three nodes contributes three connected triples, one centered on each of the three nodes, and ensures that $0 \leq T \leq 1$, with $T = 1$ for K_N.

An alternative possibility is to use the graph *clustering coefficient* C [20], defined as follows. A quantity c_i (the local clustering coefficient of node i) is first introduced, expressing how likely $a_{jm} = 1$ for two neighbors j and m of node i. Its value is obtained by counting the actual number of edges (denoted by e_i) in G_i (the subgraph made up of the neighbors of i). Notice that G_i can be, in some cases, unconnected. The local clustering coefficient is defined as the ratio between e_i and $k_i(k_i - 1)/2$, the maximum possible number of edges in G_i [20]:

$$c_i = \frac{2e_i}{k_i(k_i - 1)} = \frac{\sum_{j,m} a_{ij} a_{jm} a_{mi}}{k_i(k_i - 1)} \ . \tag{1.8}$$

The clustering coefficient of the graph is then given by the average of c_i over all the nodes in G:

$$C = \langle c \rangle = \frac{1}{N} \sum_{i \in \mathcal{N}} c_i \ . \tag{1.9}$$

By definition, $0 \leq c_i \leq 1$, and $0 \leq C \leq 1$. It is also useful to consider $c(k)$, the clustering coefficient of a connectivity class k, which is defined as the average of c_i taken over all nodes with a given degree k.

1.2 Topological properties of real networks

1.2.1 The small-world property

The study of several dynamical processes over real networks has pointed out the existence of shortcuts, i.e. bridging links that connect different areas of the networks, thus speeding up the communication among otherwise distant nodes.

In regular lattices in D dimensions, the mean number of vertices one has to pass by in order to reach an arbitrarily chosen node, grows with the lattice size as $N^{1/d}$.

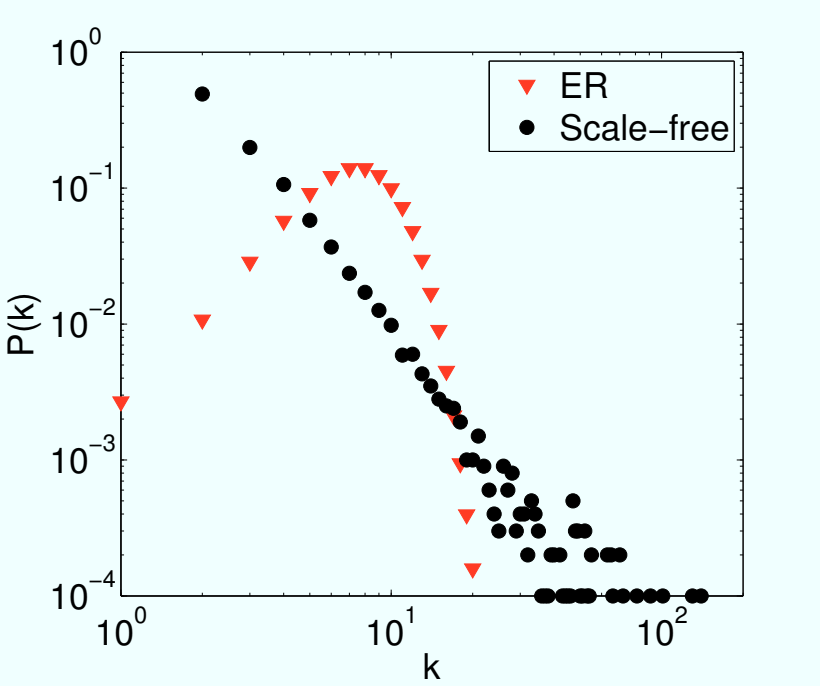

Figure 1.2: An important graph property is the degree distribution function $P(k)$, expressing the probability that a randomly chosen node of the graph has k edges. A *random graph*, described in Sec. 1.3.1, has a Poissonian degree distribution $P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$, where $\langle k \rangle$ is the average connectivity in the network. A *scale-free graph* is instead characherized by a power-law degree distribution ($P(k) = Ak^{-\gamma}$ usually with $2 < \gamma < 3$), as described in Sec. 1.2.2 and 1.3.4. A power-law distribution appears as a straight line in a double-logarithmic plot. In the figure, we show with red triangles the degree distribution of an Erdős-Rényi (ER) random graph, while with black circle the degree distribution of a scale-free graph is reported. Both graphs have the same number of nodes ($N = 10^4$) and the same average connectivity ($\langle k \rangle = 8$). One can easily notice that in the case of a scale-free graph, low degree nodes are the most frequent ones, but there are also a few highly connected nodes, usually called *hubs*, not present in a ER random graph.

Conversely, in most of the real networks, despite their often large size, any two nodes are usually connected by a relatively short path. This feature is known as the *small-world* property and is characterized by an average shortest path length L, defined as the average length of the shortest paths between any two pairs of nodes in the graph, that depends at most logarithmically on the network size N [20]. Historically, the small-world property was first observed in social networks by Milgram [18, 21, 22], who conducted a series of social experiments to estimate the actual number of links in a chain of acquaintances. Milgram's surprising result was that the number of links needed to connect two individuals taking part in his experiments and sampled from the USA population, had an average value of just six. The small-world property has been later observed in many other real networks, including biological and technological ones, and is a fundamental mathematical property in some network models, as for instance in random graphs. However, at variance with random graphs, the small-

world property in real networks is often associated to the presence of high values of the clustering coefficient, defined as in equation (1.9). For this reason, Watts and Strogatz, in their seminal paper, have proposed to define small-world networks as those networks having both a small value of L, like random graphs, and a high clustering coefficient C, like regular lattices [20]. In another formulation, the small-word properties can be reformulated in terms of the so-called *local* and *global efficiency* [23, 24].

1.2.2 Scale-free degree distributions

Until a few years ago, before the extensive exploration of data about real-world networks, it was common idea that networks were "homogeneous". Homogeneity in the interaction structure means that almost all nodes are topologically equivalent, like in regular lattices or in random graphs. In these latter ones, for instance, each of the $N(N-1)/2$ possible links is present with equal probability, and thus the degree distribution is binomial or Poisson in the limit of large graph size (see sec. 1.3.1). It is not startling then that, when the scientists approached the study of real networks from the available databases, it was considered reasonable to find degree distributions localized around an average value, with a well-defined average of quadratic fluctuations. In contrast with all the expectancies, it was found that most of the real networks display power law shaped degree distribution $P(k) \sim Ak^{-\gamma}$, with exponents varying in the range $2 < \gamma < 3$. The average degree $\langle k \rangle$ in such networks is therefore well defined and bounded, while the variance $\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2$ is dominated by the second moment of the distribution that diverges with the upper integration limit k_{max} as:

$$< k^2 >= \int_{k_{min}}^{k_{max}} k^2 P(k) \sim k_{max}^{3-\gamma} \ . \tag{1.10}$$

Such networks have been named *scale-free* networks [25, 26], because power-laws have the property of having the same functional form at all scales. In fact, power-laws are the only functional form $f(x)$ that remains unchanged, apart from a multiplicative factor, under a rescaling of the independent variable x, being the only solution to the equation $f(\alpha x) = \beta f(x)$. Power-laws have a particular role in statistical physics because of their connections to phase transitions and fractals. In the following, when referring to scale-free networks, we will denote the class of graphs with power-laws in the degree distribution. Of course, this does not necessarily implies that such graphs are scale-free with respect to other measurable structural properties. These networks, having a highly heterogenous degree distribution, result in the simultaneous presence of a few nodes (the *hubs*) linked to many other nodes, and a large number of poorly connected elements.

1.2.3 Community structure

Real networks display large inhomogeneities, revealing a high level of order and organization. The distribution of edges is not only globally inhomogeneous, as shown by fat-tailed degree distributions, but also locally, with high concentrations of edges within special groups of nodes, and low concentrations between these groups. This feature of real networks is called *community structure*[1] [27]. Communities, also called clusters or modules, are groups of nodes which probably share common properties and/or play similar roles within the graph. In Fig. 1.3 a schematic example of a graph with communities is shown.

The aim of community detection in graphs is to identify the modules and, possibly, their hierarchical organization, by only using the information encoded in the graph topology. However, finding communities within an arbitrary network can be a difficult task. The number of communities, if any, within the network is typically unknown and the communities are often of unequal size and/or density. However, despite these difficulties, in the last years, tons of methods for community finding have been proposed, and many of them have been proved to be successful at various levels. Below we describe just two of the several methods to detect communities, which will be recalled and applied later on in this thesis: the *Markov Clustering* (MCl) and *modularity optimization*. A complete review on community structure in networks, together with a detailed analysis of different algorithms to detect communities, is provided in [27].

MCl Algorithm

The Markov Clustering algorithm [28], shortened as MCl, is based on the behavior of random walkers (see Sec. 4.1 for more information about the concept of random walk) moving on the network. For this reason, this algorithm can be used also for directed and weighted graphs. The algorithm works as follows: (i) start constructing the operator $B = A + I$, where A is the adjacency matrix of the network and I is the identity operator, and normalize each column of B to obtain a stochastic transition matrix $\Pi = \{\pi_{ij}\}$: $\pi_{ij} = \frac{b_{ij}}{\sum_l b_{il}}$; (ii) compute Π^2 (this operation is also called expansion); (iii) take the r_{th} power (with $r > 1$) of every single entry p_{ij} of Π (this operation is also denoted as inflation), then normalize again to one each column of the new computed matrix; and (iv) go back to step ii. After several iterations MCL converges to a matrix $\Pi^{MCl}(r)$ which is invariant under expansion and inflation transformations. Only a few rows of $\Pi^{MCl}(r)$ have some nonzero entries: the non-zero entries belonging to the same row correspond to the nodes belonging to the same community. The role of the expansion operation is to let random walkers to explore the network, moving in each expansion step from one node to its neighbors[2]. The inflation operation, instead,

[1] Sometimes, as an alternative to the expression community structure, also the term *clustering* is used. However, we prefer not to use this term to avoid misunderstanding with the concept of clustering coefficient, introduced in Sec. 1.1.8, which has nothing to do with the idea of community structure.

[2] Notice that the addition of the identity matrix in (i) makes a node neighbor of itself.

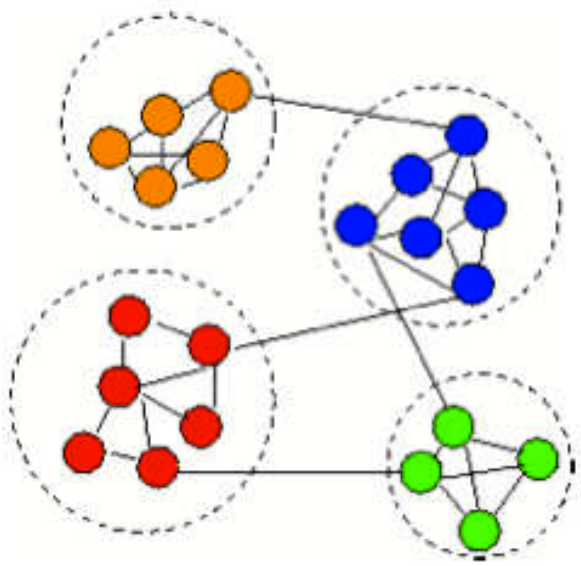

Figure 1.3: Example of a graph made up of 4 communities, indicated in the figure by dashed circles. Nodes belonging to the same community have the same colour. The density of links between nodes within the same community is larger than that of links between members belonging to different communities.

reinforces the high-probability walks at the expense of the low-probability ones. The parameter r tunes then the granularity of the clustering. If r is large, the effect of step becomes stronger and the random walks are likely to end up in small "basins of attraction" of the network, resulting in several small clusters. On the other hand, a small r produces larger clusters. In the limit of $r \to 1$, only one cluster is detected.

Modularity and its optimization

One of the most widely used methods for community detection is *modularity maximization*. The *Modularity* of a graph partition, i.e. the division of the graphs into modules, is the fraction of the edges that fall within the given modules minus the expected fraction if edges were distributed at random. For a given division of the network's nodes into some modules, modularity reflects the concentration of nodes within modules compared with random distribution of links between all nodes regardless of modules. The modularity maximization method detects communities by searching over possible divisions of a network for one or more that have particularly high modularity. There are different methods for calculating modularity. In the most common version of the concept, the randomization of the edges is done so as to preserve the degree of each vertex. For an undirect graph, to detect communities by maximizing modularity

Q, the usual functional form used is the following:

$$Q = \frac{1}{2K} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2K} \right) \delta\left(c_i, c_j\right) \tag{1.11}$$

where c_i is the group or community to which node i belongs in the graph partition considered for the evaluation of the modularity and $\delta\left(m, n\right)$ is the Kronecker delta.

The value of the modularity lies in the range $[-1, 1]$. It is positive if the number of edges within groups exceeds the number expected on the basis of chance.

Since exhaustive search of the value of modularity over all possible divisions is usually intractable, practical algorithms are based on approximate optimization methods such as greedy algorithms, simulated annealing, or spectral optimization, with different approaches offering different balances between speed and accuracy.

1.3 Network models

In the last twenty years, the analysis of many different kinds of real-world networks, highlighting a number of different topological features, has stimulated the introduction and study of many models. Most of these models try to capture the properties of real-graphs, and helped to construct a series of methods and algorithms to create "synthetic" networks. In the following we describe some mathematical models of networks which are either milestones in the network science or which will be used later in this thesis.

1.3.1 Erdős-Rényi random graphs

In 1959, Erdős and Rényi published a seminal article in which they introduced the concept of a random graph [29]. The term random graph refers exactly to the disordered nature of the arrangement of links between different nodes. Erdős and Rényi proposed a model to generate random graphs with N nodes and K links, henceforth called *Erdős and Rényi (ER) random graphs* and denote as $G^{ER}_{N,K}$. Starting with N disconnected nodes, ER random graphs are generated by placing a number K of edges randomly between pairs of nodes, taking care to avoid multiple connections between the same pair of nodes [29]. With this procedure, one gets a graph which is only one of many possible realizations. In other words, the generated graph is only one element of the statistical ensemble of all possible combinations of K connections in a graph of N nodes.

An alternative model for constructing ER random graphs consists in connecting each of the $\frac{N(N-1)}{2}$ couples of nodes with a probability $0 < p < 1$. This defines a different ensemble, indicated as $G^{ER}_{N,p}$, whose elements can have different number of links, that is to say not exactly K links. It can be proved that the two models $G^{ER}_{N,K}$ and $G^{ER}_{N,p}$ coincide in the limit of large N. In this limit, one can switch from one model to the other using the relation $p = \frac{2K}{N(N-1)}$.

For large N, and fixed $\langle k \rangle = 2K/N$, it is easy to prove that the degree distribution of ER graphs is well approximated by a Poisson distribution:

$$P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!} \ . \tag{1.12}$$

This also reflects the idea that all the nodes in a random graph are statistically equivalent, as all nodes in the graph have "similar" degree, in the sense that a randomly chosen node it is likely to have a degree close to the average value $\langle k \rangle^k$. Because of their degree distribution, ER graphs are sometimes referred to as *Poisson random graphs*. ER random graphs are, by definition, uncorrelated graphs, since the edges are connected to nodes regardless of their degree. Consequently, $P(k'|k)$ and $k_{nn}(k)$ are independent of k.

ER random graphs, for their simplicity, are the most studied among graph models. However, as a model of a real-world network, it has some serious shortcomings. Perhaps the most serious is its degree distribution, which is quite unlike those seen in most real-world networks.

1.3.2 Generalized Random Graphs: Configuration Model

The poissonian distribution of ER random graph is very different from most degree distributions found in most real networks. Indeed, the degree distribution of many real networks has been found to be fat-tailed, and, in many cases, to follow a power-law degree distribution (see Sec. 1.2.2). Therefore, if we want to model real-world networks as random graphs, or if we want to extract interesting properties of real networks from a comparison with an appropriate null model, one should extend ER models to consider random graphs in which the degree of each node is arbitrarily assigned to take a precise value. This model represents the generalization of the ER model, and allow us to generate graphs with a given arbitrary degree distribution $P(k)$. In the mathematical literature, this kind of model is known as the *configuration* model, to describe ensembles of random graphs with N nodes, K edges, and a given degree sequence [12, 30]. The configuration model is defined as follows: assign a number of nodes N, a number of links K and a degree sequence $D = \{k_1, k_2, \cdots, k_N\}$, i.e. a sequence of N integer numbers such that $\sum_i k_i = 2K$; the configuration model, denoted as $G^{conf}_{N,D}$, then consists in the ensemble of all graphs of N nodes and K edges, in which vertex i has the specified degree k_i, with $i = 1, 2, \cdots, N$, and where each graph has the same probability to be generated.

To generate one graph of the ensemble defined by a given degree sequence D, one can assign to each node i, with $i = 1, \cdots, N$ a number of half-edges (also called stubs) equal to its degree k_i. Then, one matches randomly with uniform probability pairs of stubs together, until all the K edges of the graph are created. Obtaining a graph with a given degree distribution $P(k)$ is very simple: it is sufficient to extract the N integer numbers forming the degree sequence with a probability distribution identical to the

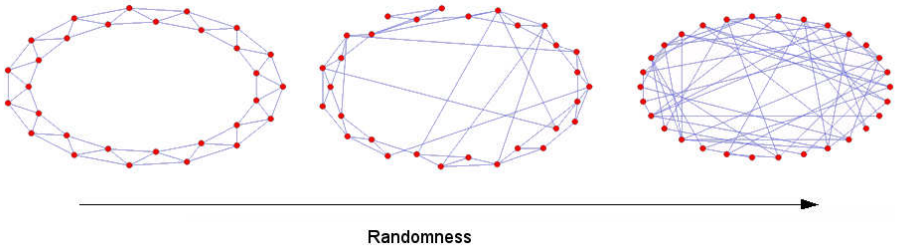

Figure 1.4: Small-world networks [20] have intermediate properties between regular lattices, first graph in figure, and random networks, third graph in figure. A regular lattice has high clustering but also a large average path length, while a random graph is characterized by a short path length together with a low clustering. A small-world network, middle graph in figure, borrows a high clustering coefficient from the former and a short average path length from the latter.

desired degree distribution $P(k)$.

1.3.3 Watts and Strogatz model

The *Watts and Strogatz (WS)* model is a method to construct graphs having both the small-world property and a high clustering coefficient [20]. The model is based on a rewiring procedure of the edges implemented with a probability p. The starting point is a N nodes ring, in which each node is symmetrically connected to its $2m$ nearest neighbors for a total of $K = mN$ edges. Then, for every node, each link connected to a clockwise neighbor is rewired to a randomly chosen node with a probability p, and preserved with a probability $1 - p$. Notice that for $p = 0$ we have a regular lattice, while for $p = 1$ the model produces a random graph with the constraint that each node has a minimum connectivity $k_{min} = m$. For intermediate values of p the procedure generates graphs with the small-world property and a non-trivial clustering coefficient. The small-world property results from the immediate drop in $L(p)$ as soon as p is slightly larger than zero. This is because the rewiring of links creates long-range edges (shortcuts) that connects otherwise distant nodes. The effect of the rewiring procedure is highly nonlinear on L, and not only affects the nearest neighbors structure, but it also opens new shortest paths to the next-nearest neighbors and so on. Conversely, an edge redirected from a clustered neighborhood to another node has, at most, a linear effect on C. That is, the transition from a linear to a logarithmic behavior in $L(p)$ is faster than the one associated with the clustering coefficient $C(p)$. This leads to the appearance of a region of small (but nonzero) values of p, where one has both small path lengths and high clustering.

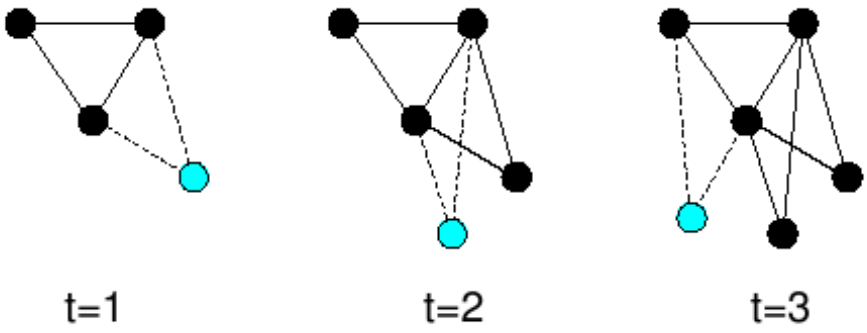

Figure 1.5: Illustration of the BA algorithm for $m_0 = 3$ and $m = 2$. At $t = 0$ we start with a complete graph of m_0 nodes. At every timestep a new node j is added, which is connected to $m = 2$ vertices, preferentially to the vertices with high connectivity, determined by the rule of Eq. 1.13. Thus, at time t there are $m_0 + t$ vertices and $\binom{m_0}{2} + mt$ edges. At each time step, the new node n is in cyan, and the two new edges are drawn with dashed lines.

1.3.4 Barabási-Albert model

The *Barabási-Albert* (BA) model is a model of different kind in respect to those presented before. In fact, it is a model aiming at reproducing how a network grows, instead of modelling a network in its final, "equilibrium" state. The BA model was inspired to the formation of the World Wide Web, a scale-free graph, and is based on two basic ingredients: growth and preferential attachment [26]. The basic idea is that in the World Wide Web, sites with high connectivity obtain new links at higher rates than low-degree nodes. More precisely, a BA undirected graph of N nodes and with average degree $\langle k = 2m \rangle$ is constructed, starting with a complete graph with a small number $N(t = 0) = m0$ of nodes and $K(t = 0) = \binom{m_0}{2}$ links. The graph grows according to the following two steps:

- At each time step t ($t = 1,\ 2,\ 3, \dots$) a new node j is added. The new node has $m \leq m_0$ edges, that link j to m different nodes, already present in the system;
- When choosing the nodes to which the new node j connects, it is assumed that the probability $\Pi_{j \to i}$ that n will be connected to node i is linearly proportional to the degree k_i of node i, i.e.:

$$\Pi_{j \to i} = \frac{k_i}{\sum_l k_l} \ . \tag{1.13}$$

After t time steps, the algorithm results in a graph with $N(t) = N(t = 0) + t$ nodes and $K(t) = K(t = 0) + mt$ edges, which for very large values of t corresponds to a graph with an average degree $\langle k \rangle = 2m$ (see also Fig. 1.5). The procedure is iterated

until the desired final number of nodes N is reached. In the limit $t \to \infty$, the model produces a degree distribution $P(k) \sim k^{-\gamma}$, with an exponent $\gamma = 3$.

Chapter 2

Elements of information theory

An absolute can only be given in an intuition, while all the rest has to do with analysis.

HENRI BERGSON

In this chapter we provide the basics of information theory. In particular we introduce the notion of stochastic process and the definitions of joint, conditional and relative entropy. We then illustrate the concepts of Markov chain and of high-order Markov chain and focus on their properties. In particular, we explain the key-concept of ergodicity and show how this impacts on the growth of entropy, as expressed by the so-called entropy rate.

2.1 Stochastic processes

Let X be a *discrete random variable* with state space $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$ and *probability distribution* $p_i \equiv \text{Prob}[X = s_i]$, $i = 1, \ldots, N$. In other words, p_i is the probability that the random variable X assumes the value s_i, i.e. that the system is in state s_i. Unless otherwise specified, we shall assume that the number N of states is finite.

A *stochastic process* is a sequence of n random variables $(X_1, X_2, \ldots, X_n)$. The integer n is the length of the stochastic process. Since there can be an arbitrary dependence among the random variables, the stochastic process is characterized by the *joint probability distribution*:

$$p_{i_1,i_2,\ldots,i_n} \equiv \text{Prob}[(X_1, X_2, \ldots, X_n) = (s_{i_1}, s_{i_2}, \ldots, s_{i_n})] \tag{2.1}$$

or by the *conditional probability distribution*:

$$p_{i_{n+1}|i_1,\ldots,i_n} = \text{Prob}[X_{n+1} = s_{i_{n+1}}|(X_1, \ldots, X_n) = (s_{i_1}, \ldots, s_{i_n})] \tag{2.2}$$

where $i_1, i_2, \ldots, i_n, i_{n+1}$ are indices that can take integer values between 1 and N. The conditional probability $p_{i_{n+1}|i_1,\ldots,i_n}$ can be written as:

$$p_{i_{n+1}|i_1,\ldots,i_n} = \frac{p_{i_1,i_2,\ldots,i_{n+1}}}{p_{i_1,i_2,\ldots,i_n}} \tag{2.3}$$

in terms of the joint probabilities.

A stochastic process can be viewed as a *dynamical system*, i.e. as a system whose state changes in time:

$$X_1 \to X_2 \to X_3 \to \cdots \tag{2.4}$$

The subscript on the random variables here represents a time index: by X_t we mean the state of the system at time t, with $t = 1, 2, \ldots$. At each time t the system can assume one of the states from set $\mathcal{S}$. The sequence $i_1, \ldots, i_n$ represents a possible trajectory of time length n. This means that the dynamical system is in state s_{i_1} at time $t = 1$, then it moves to state s_{i_2} at time $t = 2$ and so on. In principle, we have N^n different sequences of length n. Not all of them are in general possible, and some of them happens with a probability higher than others. All of this is described in terms of the joint probability distributions in Eq. (2.1) or by the conditional probability distributions in Eq. (2.2). The joint probability $p_{i_1,i_2,\ldots,i_n}$ represent how frequent is the time sequence of states $s_{i_1}, \ldots, s_{i_n}$, while the conditional probability $p_{i_{n+1}|i_1,\ldots,i_n}$ gives the probability that the dynamical system is at state $s_{i_{n+1}}$ at time $t = n + 1$, after the sequence $s_{i_1}, \ldots, s_{i_n}$.

2.1.1 Markov chains

A discrete stochastic process $(X_1, X_2, X_3, \ldots)$ is said to be a *Markov process*, or equivalently a *Markov chain* if, for $n = 1, 2, \ldots$, the conditional probability distribution has the form:

$$p_{i_{n+1}|i_1,\ldots,i_n} = p_{i_{n+1}|i_n}(n) \equiv \text{Prob}[X_{n+1} = s_{i_{n+1}}|X_n = s_{i_n}] \tag{2.5}$$

for all i_1, i_2,..., i_n, $i_{n+1} \in [1, \ldots, N]$. This means that the state of the nth random variable depends only on the previous one, namely the $(n-1)$th, and not on the entire previous sequence. Such a process is said to be a short memory process, since the "history" of the first $n-2$ steps has no influence on the nth state. Eq. 2.5 implies that the joint probability distribution $p_{i_1,i_2,\ldots,i_n}$ of a Markov chain has the form:

$$p_{i_1,i_2,\ldots,i_n} = p_{i_1} p_{i_2|i_1}(1) \cdots p_{i_n|i_{n-1}}(n-1). \tag{2.6}$$

The conditional probability $p_{i|j}(t)$ is called the *transition probability* of the Markov chain at time t.

From now on we shall consider only Markov chains with time-invariant transition

probabilities, i.e., Markov chains whose transition probabilities

$$p_{i|j}(t) = \pi(s_i|s_j) \tag{2.7}$$

do not depend explicitly on time. Such Markov chains are called *time-invariant Markov chains*. The probabilities $\pi(s'|s)$ satisfy the relation

$$\sum_{s' \in \mathcal{S}} \pi(s'|s) = 1 \tag{2.8}$$

because from a given state at time $t-1$, the system goes to one of the possible states at the next time with probability one.

The transition probabilities $\pi(s'|s)$ can be written in the form of a $N \times N$ *transition matrix*:

$$\Pi = \begin{pmatrix} \pi(s_1|s_1) & \pi(s_1|s_2) & \cdots & \pi(s_1|s_N) \\ \pi(s_2|s_1) & \pi(s_2|s_2) & \cdots & \pi(s_2|s_N) \\ \vdots & & \ddots & \vdots \\ \pi(s_N|s_1) & \pi(s_N|s_2) & \cdots & \pi(s_N|s_N) \end{pmatrix}. \tag{2.9}$$

With such a definition we have that $\pi_{ij} = \pi(s_i|s_j)$, i.e. that the matrix entry π_{ij} represents the probability to go from state j to state i. Because of condition (2.8), Π is a stochastic matrix, i.e. a matrix each of whose columns consist of nonnegative real numbers, with each column summing to 1. By writing the probability $p_i(t) = \text{Prob}[X_t = s_i]$, $t = 1, 2, \ldots, n$, as a vector $\mathbf{p}(t)$:

$$\mathbf{p}(t) \equiv \begin{pmatrix} p_1(t) \\ p_2(t) \\ \vdots \\ p_N(t) \end{pmatrix}, \tag{2.10}$$

the dynamical evolution of the Markov chain is ruled by the equation

$$\mathbf{p}(t+1) = \Pi \mathbf{p}(t). \tag{2.11}$$

The solution of the equation is given by

$$\mathbf{p}(t) = \Pi \cdot \Pi \cdot \ldots \cdot \Pi \; \mathbf{p}(0) = \Pi^t \mathbf{p}(0).$$

A time-invariant Markov chain is therefore characterized by the transition matrix Π, and by the initial distribution $\mathbf{p}(0)$.

2.1.2 High-order Markov chain

Markov chains are short-memory processes. Not all stochastic processes are Markov chains, since in general the probability of moving to state i_{n+1} depends on the whole history, and equations (2.5,2.6) are not valid. The simplest case of a non-Markov stochastic process is when the state at time $t+1$ depends on both the states of the system at time t and at time $t-1$. In this case we have to work with the conditional probabilities $p_{i_3|i_1,i_2}$. Markov processes are also called *first order Markov processes*, while processes determined by conditional probabilities $p_{i_3|i_1,i_2}$ are called *second order Markov processes*. Note that a second order Markov process can be represented as a first order Markov process by extending the state vectors to pair of states.

A particular example of Markov process is the *random walk* on a graph. As we will see in details in chapter 4, a random walk on a network is a dynamical process where particles or walkers move from node to node. The random walk is a sequence $\{X_m\}$, where $X_m \in \{i_1, i_2, \cdots, i_N\}$ with i_k representing the k-th node in the graph. Given $X_m = i$, the next node in the sequence is chosen with uniform probability from the neighbors of node i. Such a random walk is known as an *plain random walk* (see also Sec. 4.1.1). In a more general random walk, named *biased random walk*, a neighbor of the node i, say j, is chosen with a probability which depends on a (time-independent) property of j. These dynamical rules yield processes which can be described by an equation like 2.11 with a transition matrix Π which is a function of the adjacency matrix $\mathcal{A}$ of the graph and of the node properties biasing the walkers movements.

Examples of stochastic processes that have long memory, hence being high-order Markov chains, are provided by written texts or by sequences of DNA. In fact, in a text, the probability of finding a letter at a given point depends usually not only on the previous letter in the sequence, but also on a number of other previous letters. Similarly, in the DNA information is encoded in strings of nucleotides, which can be represented by four letters A, C, G, T, and an analysis equivalent to the one performed on written texts can be done. In chapter 5 we will discuss about the application of high-order Markov chains to extract meaning from linguistic and biological sequences.

2.2 Characterization of Markov Chains

From now on we shall restrict to first order Markov chains.

2.2.1 Classification of states

The states of a Markov chain fall into distinct types according to their limiting behaviour and are characterized by the following definitions.

Suppose that the chain is initially in state s_i.

1. state s_i is said *recurrent* if the chain returns to s_i with probability 1. In this case the time of first return will be a random variable called the *recurrence time*, and

the state is called *positive-recurrent* or *null-recurrent* according to whether the mean recurrence time is finite or infinite.

2. state s_i is said *transient* if it is not recurrent (i.e. if the probability that the chain returns to s_i is less than one).

States can be distinguished also in periodic and aperiodic. Suppose that the chain is initially in the state s_i. A state s_i has period T if any return to state i must occur in multiples of T time steps. Formally, the period of a state is defined as

$$T = \gcd\{n > 0 : \pi_{ii}^{(n)} > 0\}$$

where gcd denotes the greatest common divisor. If $T = 1$, state s_i is said to be *aperiodic*. If $T > 1$, state s_i is said to be *periodic* with period T. A state s_i is said to be *ergodic* if it is aperiodic and positive recurrent.

2.2.2 Accessibility and communicating states

Having defined the basic types of states, it is possible to show that only states of the same type are "accessible" to each other and can hence "communicate". A state s_i is said to be *accessible* from s_j if it is possible to reach s_i from s_j in a finite number of transitions, i.e. if there is an integer k such that $\pi_{ij}^{(k)} > 0$. If state s_i is accessible from s_j and s_j is accessible from s_i then states s_i and s_j are said to *communicate*. Hence if s_i and s_j are communicating, they must be both transient or both null recurrent, or both positive recurrent, and furthermore, they must have the same period [31].

2.2.3 Classification of chains

From the definitions of the states, it is possible to introduce different classes of Markov chains. A Markov chain is said *irreducible* iff all pairs of states communicate, i.e. if it is possible to go from any state of the Markov chain to any other state in a finite number of steps. Note that an irreducible chain has the property that all its states are of the same type, and therefore it is possible to speak of an irreducible Markov chain as being transient, recurrent, and so on.

2.3 Finite size Markov chains

The definitions given in Sec. 2.2 are in general valid for every kind of Markov chains, therefore also for infinite ones, i.e. when the number of states N is infinite. Here our main interest is in finite Markov chain, described by a finite transition matrix. In such a case the Markov chain is irreducible if and only its transition matrix is an irreducible matrix [32].

The probability distribution $\mathbf{p}^*$ such that

$$\mathbf{p}^* = \Pi \mathbf{p}^*. \tag{2.12}$$

is said the *stationary* or *invariant distribution*. That is to say, $\mathbf{p}^*$ is an eigenvector of Π with eigenvalue 1. By means of the Perron-Frobenius theorem [32] it is possible to prove that for irreducible non negative matrices there is only one eigenvalue equal to 1, and therefore that the stationary distribution $\mathbf{p}^*$ is unique.

A sufficient condition for a vector p^* to be a stationary distribution is the *detailed balance condition*:

$$\pi_{ij} p_j^* = \pi_{ji} p_i^*. \tag{2.13}$$

Indeed, summing over j on both sides of this condition and keeping in mind the normalization condition for Π, we obtain

$$\sum_j \pi_{ij} p_j^* = \sum_j \pi_{ji} p_i^* = p_i^*. \tag{2.14}$$

2.3.1 Ergodic Markov chains

If all states in an irreducible Markov chain are ergodic, then the chain is said to be ergodic. In this case, one is assured that

$$\lim_{n\to\infty} \mathbf{p}_n = \lim_{n\to\infty} \Pi^n \mathbf{p}_0 = \mathbf{p}^*, \tag{2.15}$$

for any initial distribution $\mathbf{p}_0$.

Note that in a finite Markov chain, we cannot have null recurrent states, therefore states can be either transient of positive-recurrent. Moreover, it is obvious that not all states can be transient. Therefore in an irreducible finite Markov chain, the states have to be all positive-recurrent (because they have to be all of the same type, and not all transient). We can conclude that a finite irreducible Markov chain is ergodic if and only if it is aperiodic. A Markov chain having a symmetric transition matrix Π is ergodic. Its stationary distribution is the uniform distribution.

2.4 Joint entropy and conditional entropy

Entropy is a key concept in physics [33] and in information theory [34]. In general words, the entropy of a system is a measure of its disorder or, equivalently, of the amount of information needed to describe it. In the case of a random variable, the entropy is a measure of the uncertainty of the random variable. The *entropy* of a

discrete random variable X is usually defined as:

$$H(X) = -\sum_x p(x) \ln p(x) \tag{2.16}$$

where $p(x)$ is the probability distribution of the random variable X [34].

The concept of entropy can be introduced also for two random variables X and Y, if they can be described by a joint probability distribution $p(x, y)$. In this case the *joint entropy* $H(X, Y)$ of the two random variables X and Y is defined as:

$$H(X, Y) = -\sum_x \sum_y p(x, y) \ln p(x, y) \tag{2.17}$$

It is also possible to define the *conditional entropy* of a random variable given another as the expected values of the entropies of the conditional distributions, averaged over the conditioning random variable. If X and Y are two random variables, and $p(x, y)$ their joint probability distribution, the conditional entropy $H(Y|X)$ is defined as:

$$H(Y|X) = -\sum_x \sum_y p(x, y) \ln p(y|x) \tag{2.18}$$

$p(y|x)$ is the probability of y if x occurs, and is referred to as the conditional probability of y given x. It can be proven that $p(x, y) = p(x) + p(y|x)$ and that in general $p(y|x) \neq p(x|y)$.

The naturalness of the definition of joint entropy and conditional entropy is exhibited by the fact that the entropy of a pair of random variables is the entropy of one plus the conditional entropy of the other:

$$H(X, Y) = H(X) + H(Y|X) \tag{2.19}$$

This can be simply proven remembering that $p(x, y) = p(x) + p(y|x)$ and using the properties of the logarithms

In general, it is possible to generalize the concept of joint and conditional entropy to the case of N random variables $X_1, X_2, \cdots, X_n$ whose probability distribution function is $p(x_1, x_2, \cdots, x_n)$. The joint entropy reads:

$$H(X_1, X_2, \cdots, X_n) = -\sum_{x_1, x_2, \cdots, x_n} p(x_1, x_2, \cdots, x_n) \ln p(x_1, x_2, \cdots, x_n), \tag{2.20}$$

while the conditional entropy reads:

$$H(X_n|X_1, X_2, \cdots, X_{n-1}) = -\sum_{x_1, x_2, \cdots, x_n} p(x_1, x_2, \cdots, x_n) \ln p(x_n|x_1, x_2, \cdots, x_{n-1}), \tag{2.21}$$

In the case of N random variables, it can be proven that the following chain rule

holds:

$$H\left(X_1, X_2, \cdots, X_n\right) = -\sum_{i=1}^{n} H\left(X_i | X_1, X_2, \cdots, X_{i-1}\right) \tag{2.22}$$

2.5 Relative entropy

The *relative entropy* $D_{KL}(P|Q)$ of the random variable P with respect to random variable Q, also known as *Kullback-Leibler distance* between P and Q, is a measure of the amount of extra information required to represent P by using only information about Q. It is defined as the average of the logarithmic distance between P and Q, weighted by the probability P, i.e.:

$$D_{KL}(P|Q) = \sum_{x} p\left(x\right) \ln \frac{p\left(x\right)}{q\left(x\right)} \tag{2.23}$$

where $p\left(x\right)$ and $q\left(x\right)$ are the probability distributions characterizing P and Qrespectively. $D_{KL}(P|Q)$ represents the number of extra bits of information required to reconstruct P starting from Q [34], and it can be considered as a measure of the inaccuracy of assuming that the distribution is $q\left(x\right)$ when the true distribution is $p\left(x\right)$. The relative entropy is always non negative and is zero if and only if $p\left(x\right) = q\left(x\right)$. However, it is not a true distance since it not symmetric and does not satisfy the triangle inequality.

2.6 Entropy rate

When one has a sequence of n random variables, it is useful to define a measure that quantifies how the entropy of the sequence grows with n. A measure of this growth is given by the *entropy rate* h, defined as:

$$h = \lim_{n \to \infty} \frac{H(X^n)}{n}, \tag{2.24}$$

provided this limit exists. The entropy rate h is a measure of the average description length for the stochastic process. This means that we can practically represent the typical sequences of length n generated by the stochastic process by using approximately $n \cdot h$ bits. For example, one can consider the case of a typewriter typing on a keyboard with m letters, each of them equally likely to appear, and calculate the entropy rate of the sequences of symbols the typewriter generates. Since the typewriter can produce m^n sequences of length n, all of them being equally likely, the entropy of sequences of length n is $H(X_1, X_2, \cdots, X_n) = \ln m^n$, while the entropy rate is $h = \ln m$. This is also the bits of information per symbol in the sequence.

2.6.1 Entropy rate of Markov chains

The entropy rate is very easy to calculate for Markov chains, moreover if the they are ergodic and time invariant. In fact, it can be proven [34] that the entropy rate h of an ergodic Markov chain with stationary distribution $p^*(x)$ and time invariant transition probability $\pi(x'|x)$ is given by:

$$h = -\sum_{x',x} \pi(x'|x) p^*(x) \ln \pi(x'|x). \tag{2.25}$$

Notice that, while for a general stochastic process the entropy rate might not be defined because the limit in Eq. 2.24 might not exist, the entropy rate for an ergodic Markov chain with a finite number of states is always defined, as the sum in Eq. 2.25 contains always a finite number of terms.

All the definitions and concepts we have presented in this and in the previous chapters will be used in the following parts of the thesis to address different theoretical problems, like the issue of the maximization of the entropy rate for a biased random walk, or to study different kinds of real-world datasets, such as patterns of human mobility or the statistical properties of aminoacid sequences.

Chapter 3

Three-body degree correlations in complex networks

Not everything that can be counted counts,
and not everything that counts can be counted.

Albert Einstein

Many physical phenomena can be fully understood only by considering the effects of high order correlations [35]. Up to now, the connectivity of complex networks has been described and modeled solely on the basis of node degrees and of two-body degree correlations, as mentioned in the first chapter of this thesis. In this chapter we describe a formalism based on second-order Markov chains to study and detect genuine three-body degree correlations. By comparing a network to the ensemble of graphs with the same degree-degree correlations, we give empirical evidence that non-trivial three-body degree correlations do occur in a number of real-world systems.

The presented analysis reveals that three-body correlations have marked effects on some network properties, such as the average connectivity of second neighbors of a node or the rich-club ordering [36–38], which will be introduced below, and can play a role in network dynamical processes, such as random walks [3], the topic of the next chapter. As a consequence, a consistent theory of complex networks should properly take three-body correlations into account.

3.1 More on degree-degree correlations

Before introducing the formalism for the study of three-body degree correlations, we recall the definitions introduced in Sec. 1.1.7, we provide more details on how to derive the relations shown in that section, and we deepen some concepts regarding the degree-degree correlations in graphs.

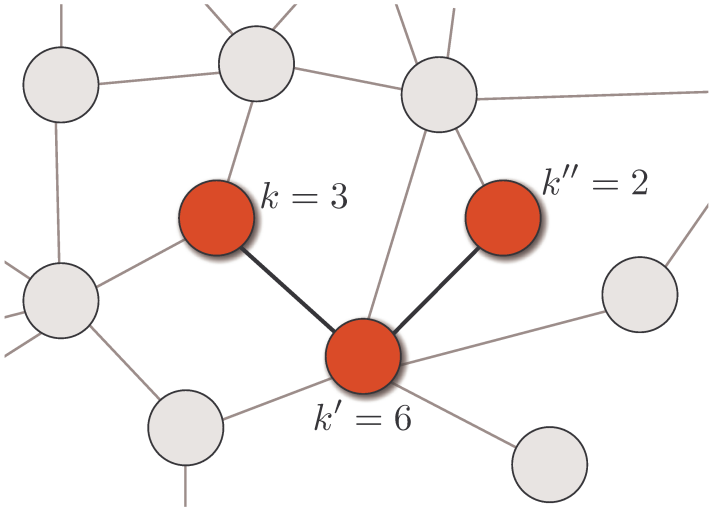

Figure 3.1: Wedges are objects embedded in the structure of networks. We show here an example of a wedge (k, k', k'') with $k = 3$, $k' = 6$ and $k'' = 2$, whose nodes have been coloured red and the links drawn with bold black.

Let us consider an undirected graph of N nodes, K links and with degree distribution P_k. In order to study the degree-degree correlations, sometimes also called two-body degree correlations, we need to evaluate the probability that, starting from a node of degree k and following one of its link, we end up on a node of degree k'. We evaluate this probability in terms of the number of the number of links $E_{kk'}$ between nodes of degree k and k'. More specifically, we define [39]:

$$E_{kk'} = \begin{cases} \text{if } k \neq k', \ \# \text{ of edges connecting a node of degree } k \text{ and a node of degree } k'; \\ \text{if } k = k', \ 2 \cdot \# \text{ edges between } k \text{ and } k'. \end{cases}$$

Notice that each link of the graph with the definition above is counted two times. For example, a link connecting a node of degree 2 and of degree 3, will contribute to the term $E_{2,3}$ and to the term $E_{3,2}$. This is also the reason of the factor 2 in the case $k = k'$: a link between two nodes of the same degree, for example 3, will contribute twice to the term $E_{3,3}$.

The following properties hold:

$$\begin{aligned} \textstyle\sum_{k'} E_{kk'} &= kN_k = kNP_k \\ \textstyle\sum_{kk'} E_{kk'} &= \langle k \rangle N = 2K \end{aligned}$$

where N_k is the number of nodes of degree k, $\langle k \rangle$ the average degree and K the total number of edges in the graph. To derive first property, it is sufficient to notice that the sum of $E_{kk'}$ over all possible k' returns the total number of edges that are incident on a node of degree k. This number than amounts to the number of nodes N_k with

degree k times the connectivity k. The second property is obtained simply by noticing that the sum of all possible degrees k and k' returns twice the total number of links.

According to the previous definitions, the probability $P_{kk'}$ of having a node of degree k connected to a node of degree k' is given by:

$$P_{kk'} = \frac{E_{kk'}}{\sum_{kk'} E_{kk'}} = \frac{E_{kk'}}{2K} \tag{3.1}$$

Similarly, once selected a node of degree k', the probability to find a neighbor of degree k', $P_{k'|k}$, is:

$$P_{k'|k} = \frac{E_{kk'}}{\sum_{k'} E_{kk'}} = \frac{E_{kk'}}{kN_k} \tag{3.2}$$

By equating the expression of $E_{kk'}$ of 3.1 and 3.2, we obtain:

$$P_{kk'} = \frac{kN_k}{2K} P_{k'|k} = \frac{kP_k}{\langle k \rangle} P_{k'|k} = q_k P_{k'|k} \tag{3.3}$$

where q_k is the probability of taking a link which is connected to a node of degree k.

If a network is uncorrelated, the probability 3.2 does not depend on the degree k of the starting node. In fact, the conditional probability $P^{u.c.}_{k'|k}$ in an uncorrelated network can be derived simply by counting the number of possibilities to connect to a node of degree k', since connections between two nodes are random. Since in the network there are $N_{k'}$ nodes with degree k', and in each of these nodes we can arrive from k' links, the probability that $P^{u.c.}_{k'|k}$ reads:

$$P^{u.c.}_{k'|k} = \frac{k'N_{k'}}{\sum_{k'} k'N_{k'}} = \frac{k'P_{k'}}{\sum_{k'} k'P_{k'}} = \frac{k'P_{k'}}{\langle k \rangle} = q_{k'}$$

In this case the joint probability function of an uncorrelated network $P^{u.c.}_{kk'}$ factorizes into two functions of k and k': $P^{u.c.}_{kk'} = q_k q_{k'}$.

3.1.1 Average degree of nearest neighbors

The probability $P_{k'|k}$, as k and k' change, is represented by a matrix which in general is not easy to visualize. In order to detect the presence of degree-degree correlations in a graph, it is useful to study other topological properties which are related to the conditional probability $P_{k'|k}$. A useful quantity in this sense is $\langle k_{nn} \rangle (k)$, the average degree of the neighbours of nodes with degree k, already introduced in Eq. 1.5, which we report again here:

$$\langle k_{nn} \rangle (k) = \sum_{k'} k' P_{k'|k} \tag{3.4}$$

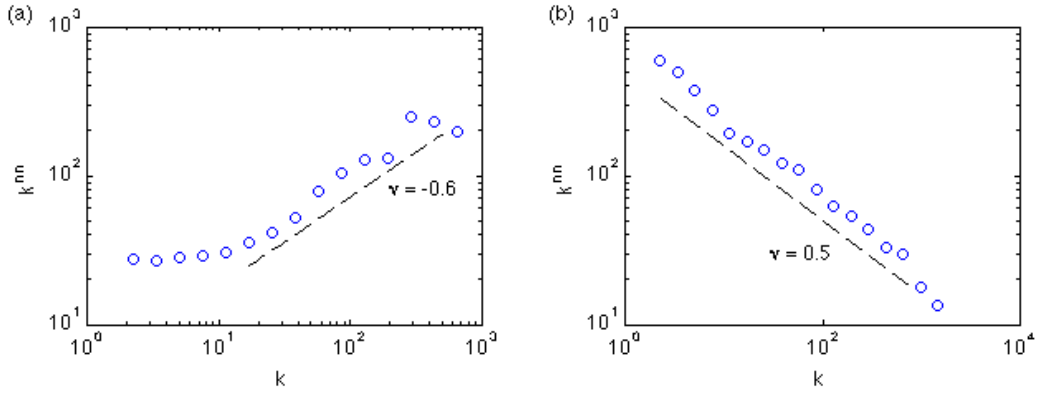

Figure 3.2: The tendency of low degree nodes to be connected to low degree nodes, and of high degree nodes to be connected to high degree nodes is called assortativity. This non-trivial correlation often leads to a relation $k^{nn} \sim k^{-\nu}$ with $\nu < 0$ and is typically observed in collaboration networks [40], see panel (a). On the other hand, the tendency of low degree nodes to be connected to high degree nodes and vice versa is disassortativity, with an exponent $\nu > 0$. Panel (b) shows one example of such a disassortative network, the autonomous system of the internet [16].

Using the equality 3.2, the previous equation turns to be:

$$\langle k_{nn} \rangle (k) = \frac{1}{kN_k} \sum_{k'} k' E_{kk'}$$

If the graph has no degree-degree correlations, i.e. $P_{k'|k} = P^{u.c.}_{k'|k}$, from the previous equation we get:

$$\langle k^{u.c.}_{nn} \rangle (k) = \frac{1}{\langle k \rangle} \sum_{k'} k'^2 P_{k'} = \frac{\langle k^2 \rangle}{\langle k \rangle} \tag{3.5}$$

which turns out to be a constant, as expected, since $\langle k_{nn} \rangle (k)$ will not depend on k if the graph is uncorrelated.

On the other hand, if the graph has degree-degree correlations, then $\langle k_{nn} \rangle (k)$ will show depend on k. In particular, it has been shown that a wide range of graphs is characterized by degree-degree correlations which are well described by the expression $\langle k_{nn} \rangle (k) = k^{-\nu}$, where ν can be positive or negative. When $\nu < 0$ ($\nu > 0$), networks are said to be *assortative* (*disassortative*). As example of assortative network and one of disassortative network are shown in Fig. 3.2. Although the characterization of two-body degree correlations has already provided valuable insights in the study of the structure and dynamics of complex networks, most real networks, as we will show below, exhibit also higher order correlations, more specifically three-body degree correlations. Our aim in the following sections is to provide the tools to measure exactly correlations of the third order.

3.2 How to quantify three-body degree correlations

3.2.1 Definition of wedge

In order to find out whether a network has higher order degree correlations, namely three-body correlations, we need to study the statistical properties of triples of connected nodes, namely pairs of edges having a node in common or, in other words, paths of length 2. Recalling the symbol of the vector product, $\wedge$, we name such objects *wedges*. In Fig. 3.1, we show an example of a wedge, embedded in a network. The nodes of the wedge are coloured in red and its links are bolded. The wedge has a central node with degree $k' = 6$ and two external nodes respectively with $k = 3$ and $k'' = 2$ links. We can classify this wedge as a $(k = 3, k' = 6, k'' = 2)$ wedge, meaning that it has a central node of degree $k' = 6$ to which a node of degree $k = 3$ and a node of degree $k'' = 2$ are connected. Notice that such an object can be defined only if $k' \geq 2$. Therefore, in the following, we implicitly assume that we are looking at (k, k', k'') with $k' \geq 2$ and all the sums over k' should be interpreted as sums over $k' \geq 2$. We now need to count how many wedges are present in a graph. More formally we define:

$$W_{kk'k''} = \begin{cases} \text{if } k \neq k'', \ \# \text{ of wedges } (k, k', k'') \text{ with a node of degree } k' \text{ as centre,} \\ \qquad \text{and nodes of degree } k \text{ and } k'' \text{ as branches;} \\ \text{if } k = k'', \ 2 \cdot \# \text{ of wedges with a node of degree } k' \text{ as centre,} \\ \qquad \text{and two nodes of degree } k \text{ as branches.} \end{cases}$$

The following normalizations properties hold:

$$\sum_{kk'k''} W_{kk'k''} = \sum_{k'} k'(k'-1)N_{k'} \equiv N_W \tag{3.6}$$

$$\sum_{k''} W_{kk'k''} = E_{kk'}\,(k'-1) \tag{3.7}$$

The quantity $N_W = \sum_{k'} k'(k'-1)N_{k'}$ is nothing else that the total number of wedges in the graph. This is easily derived by noticing that, if $N_{k'}$ is the number of nodes of degrees k', then there will be $k'(k'-1)N_{k'}$ different wedges whose central node has degree k'. Finally, by summing over all possible k', the total number of wedges in the graph N_W is obtained. The quantity $\sum_{k''} W_{kk'k''}$ of Eq. 3.7 amounts to the number of wedges in the graph whose first and second node have respectively degree k and k', and whose third node can have any degree. Property 3.7 can be obtained by observing that, once the first two nodes are fixed, the first one with degree k and the second with degree k', one can create $(k'-1)$ different wedges. Then, $\sum_{k''} W_{kk'k''}$ is equal to the total number of edges connecting two nodes of degree k and k', i.e. $E_{kk'}$, times $(k'-1)$.

3.2.2 Joint and conditional probability

Given the above definitions and the relative normalization properties, the probability $P_{kk'k''}$ of finding a wedge (k, k', k'') can be defined as:

$$P_{kk'k''} = \frac{W_{kk'k''}}{\sum_{kk'k''} W_{kk'k''}} = \frac{W_{kk'k''}}{\sum_{k'} k'\,(k'-1)\,N_{k'}} \tag{3.8}$$

Similarly, the conditional probability $P_{k''|kk'}$, i.e. the probability that given an edge (k, k'), this is part of a wedge (k, k', k''), can be defined as:

$$P_{k''|kk'} = \frac{W_{kk'k''}}{\sum_{k''} W_{kk'k''}} = \frac{W_{kk'k''}}{E_{kk'}\,(k'-1)} \tag{3.9}$$

By equating the term $W_{kk'k''}$ from Eqs. 3.8 and 3.9, we obtain:

$$P_{kk'k''} = \frac{E_{kk'}\,(k'-1)}{\sum_{k'} k'\,(k'-1)\,N_{k'}} P_{k''|kk'}$$

Then, by using the two-body joint and conditional probability, 3.1 and 3.2 respectively, from previous equation we can get:

$$\begin{aligned} P_{kk'k''} &= \frac{(k'-1)\,2K}{\sum_{k'} k'\,(k'-1)\,N_{k'}} P_{kk'} P_{k''|kk'} = \\ &= \frac{kN_k\,(k'-1)}{\sum_{k'} k'\,(k'-1)\,N_{k'}} P_{k'|k} P_{k''|kk'} \end{aligned}$$

Notice that, since in undirected graphs we have $E_{kk'} = E_{k'k}$, Eq. 3.2 yields the equality $kN_k P_{k'|k} = k'N_{k'} P_{k|k'}$. Plugging this in the previous relations, $P_{kk'k''}$ can be expressed as:

$$P_{kk'k''} = \frac{k'\,(k'-1)\,N_{k'}}{\sum_{k'} k'\,(k'-1)\,N_{k'}} P_{k|k'} P_{k''|kk'} = \omega_{k'} P_{k|k'} P_{k''|kk'} \tag{3.10}$$

where $\omega_{k'} = \frac{k'N_{k'}(k'-1)}{\sum_{k'} k'(k'-1)N_{k'}} = \frac{k'N_{k'}(k'-1)}{N_W}$ is the probability that, selecting randomly a wedge, it will have a central node of degree k'.

Furthermore, since $W_{k,k',k''} = W_{k'',k',k}$, from definition 3.8 $P_{kk'k''} = P_{k''k'k}$. Hence, Eq. 3.10 can be also written as:

$$P_{kk'k''} = \omega_{k'} P_{k''|k'} P_{k|k''k'}$$

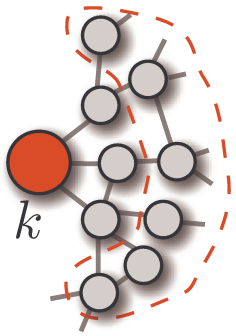

Figure 3.3: Schematic example to visualize the second neighbors of a node of degree $k = 3$, red-colored in the figure. This node is connected to three other nodes, its first neighbors, that are in turn connected to other nodes, that are then the second neighbors of the red node. The second neighbors of the red node, encircled in figure with a red dashed curve, made up together with the red node the branches of wedges the first neighbors of the red node are centers of. It is the clear that the average connectivity of the second neighbors of a node is governed by the three-body degree correlations in the graph.

3.2.3 Markovian networks

A network which is completely described by the $P_{kk'}$, meaning that is only characterized by two-body correlations and is otherwise random, with no higher order correlations, is usually referred to as a *markovian* network [39]. The expressions derived up to Eq. 3.10 characterize the three-body degree correlations in a general network. However, we can obtain simpler expressions if we assume that the network is markovian. In this case the conditional probability $P_{k''|k'k}$ does not depend on k. Thus, denoting the three-body joint probability of markovian networks as $P^{(2)}_{kk'k''}$, we have:

$$P^{(2)}_{kk'k''} = \omega_{k'} P_{k|k'} P_{k''|k'}.$$

In the equation above, all the terms on the right side depend only on two degree classes, indicating that there are no correlations between the degrees of the nodes at the extremes of a wedge. However, the joint probability $P^{(2)}_{kk'k''}$ still depends on three variables, namely the degrees k, k' and k'', indicating that degree-degree correlations induce spurious three-body degree correlations. To detect the real three-body correlations, i.e. the correlations not induced by those of lower order, we have to compare particular topological properties of real-world networks to those of null models having the same degree-degree correlations and otherwise random.

3.3 Average connectivity of the second neighbors

In section 3.1.1 we have mentioned the difficulty to get information about two-body degree correlations from the simple visualization of the matrix $P_{k'|k}$. Visualizing the conditional probability $P_{k''|kk'}$ is even harder since we have to deal with a three-dimensional hyper-matrix. Analogously to the case of two-body degree correlations, for which the average degree of the neighbors of nodes with degree k, $\langle k_{nn}(k)\rangle$, is studied, we introduce the quantity $\langle k_{nnn}\rangle(k)$, defined as the average degree of the second-nearest-neighbors of nodes of degree k, and we study its behavior to check the existence of three-body degree correlations. We can get a better idea by looking at Fig. 3.3. In the figure a node with degree k, shown in red, is directly connected to some nodes, that are hence its first neighbors. Its first neighbors are in turn connected to others nodes, surrounded by a red dashed curve in the figure, which are the second neighbors of the red node. The average of their degrees then represent the k_{nnn} of the red node. To get $\langle k_{nnn}\rangle(k)$, one has to average over all the nodes having the same degree k. More formally, $\langle k_{nnn}\rangle(k)$ can be written in terms of conditional probabilities as:

$$\langle k_{nnn}\rangle(k) = \sum_{k'} P_{k'|k} \sum_{k''} k'' P_{k''|kk'} \ . \tag{3.11}$$

To derive this formula, one has to think of a node of degree k as being the branching node of a certain number of wedges. Its second neighbors will be all the nodes on the other branch of those wedges. Therefore, with the sum $\sum_{k''} k'' P_{k''|kk'}$ one takes the average degree of all nodes forming a wedge with edges $E_{kk'}$. Then, to get the average connectivity of the second neighbors of a node with degree k, one has to sum over all the possible degrees k' weighting each of term of the sum with the probability that a node of degree k is connected to a node of degree k'. In the case of markovian networks $\langle k_{nnn}\rangle(k)$ of Eq. 3.11 has a simpler form. In fact, the the three-body joint probability 3.9 turns into a two-body one, i.e. $P_{k''|kk'} = P_{k''|k'}$. This means that the probability of forming a wedge (k, k', k'') starting from an edge (k, k'), only depends on the correlations between the degrees k' and k''. In this (null) case, formula 3.11 reduces to:

$$\langle k^{exp}_{nnn}\rangle(k) = \sum_{k'} P_{k'|k} \sum_{k''} k'' P_{k''|k'} \ . \tag{3.12}$$

where the superscript ^{exp} indicates the expected value of $\langle k_{nnn}\rangle(k)$ if there are no correlations of order higher than two. Notice that, while $\langle k_{nn}\rangle(k)$ is constant when there are no degree-degree correlations, $\langle k^{exp}_{nnn}\rangle(k)$ is in general a function of k even if there are not three-body degree correlations. Such a dependence is indeeed induced by the two-body correlations.

	N	$\langle k \rangle$	$\langle k^2 \rangle$	r
E-Mail in URV [41]	1133	9.62	179.82	0.073
Scientific collaboration network (cond-mat) [42]	12722	6.28	80.39	0.147
Scientific collaboration network (HepPh) [40]	12008	19.74	2564.78	0.63
Patents [40]	230686	4.81	59.04	0.146
Internet AS [16]	11174	4.19	1112.82	-0.194
WWW [43]	325729	6.69	1878.69	-0.054
PGP [44]	10680	4.55	85.98	0.232
Neural network *C. Elegans* [45]	297	14.46	376.78	-0.241
Protein interactions *S. Cerevisiae* [46]	4626	6.40	155.15	-0.142
Protein interactions *S. Pombe* [47]	2361	6.08	102.27	-0.084
Private messages in Pardus [48]	5877	36.57	4997.58	-0.058
Friendship in Pardus [48]	4313	9.79	281.08	-0.002
Enmity in Pardus [48]	2906	13.77	1129.66	-0.240
Attacks in Pardus [48]	7992	13.41	1027.74	-0.163
Jazz musicians collaboration [49]	198	27.60	1070.24	0.066

Table 3.1: Basic features of the networks used in this chapter to investigate the presence of three-body degree correlations. The number of nodes N, the average connectivity $\langle k \rangle$, the second moment of the degree distribution $\langle k^2 \rangle$, and the assortativity coefficient r (see Eq. 1.6 and [17]) are reported.

3.4 Three-body correlations in real-world networks

At this point we can tackle the question about the existence of non-trivial three-body degree correlations in real complex networks. As for the case of two-body degree correlations, in which the observed value of the average connectivity of the first neighbors of a node k, which we indicate now as $\langle k_{nn}^{obs} \rangle(k)$ is compared to the null-case of a network with the same degree distribution P_k, $\langle k_{nn}^{exp} \rangle(k) = \langle k^2 \rangle / \langle k \rangle$, we will compare the observed patterns for $\langle k_{nnn}^{obs} \rangle(k)$, with the null case prediction $\langle k_{nnn}^{exp} \rangle(k)$. In Fig. 3.4 we report the comparisons of these quantities for the Internet network at the Autonomous System (AS) level [16]. As it can be observed from the left panel, this network presents non-trivial two-body degree correlations of disassortative nature, i.e., $\langle k_{nn} \rangle(k)$ is a decreasing function of k. However, the right panel show that three-body correlations do also exist as the observed patterns $\langle k_{nnn}^{obs} \rangle(k)$ differs from what expected if solely two-body correlations were at work. Morever, the three-body degree correlations show an assortative character, as $\langle k_{nnn}^{obs} \rangle(k)$ increases with k, at variance with the nature of the two-body degree correlations. A number of previous studies have shed light on the nature of two-body degree correlations in real-world networks yielding a classification between assortative and disassortative network that unveils a stricking dependence on the nature of the network. On one hand, technological and biological networks are seen to usually found to display disassortative correlations, while social networks are mainly as-

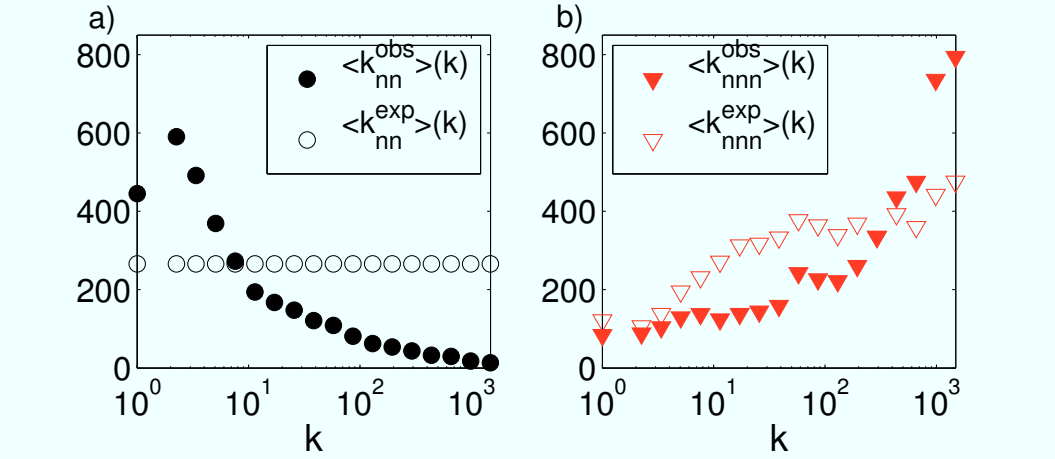

Figure 3.4: We show the average connectivity of the first neighbors and of the second neighbors of a node, as a function of its degree k, for the Internet network at the level of autonomous system [16]. In (a) the measured average connectivity of the first neighbors, $\langle k_{nn}^{obs} \rangle (k)$, is compared to $\langle k_{nn}^{exp} \rangle (k)$, the value of the first neighbors' average degree expected in a random uncorrelated network with the same degree sequence, which is equal to the constant $\frac{\langle k^2 \rangle}{\langle k \rangle}$, as proven in Sec. 3.1.1. In (b), the measured average connectivity of the second neighbors $\langle k_{nnn}^{obs} \rangle (k)$ is compared to $\langle k_{nnn}^{exp} \rangle (k)$, the value calculated for the corresponding markovian networks, i.e. for networks having the same degree sequence and the same degree-degree correlations as measured by the joint probability distribution $P_{kk'}$.

sortative. Such a classification points out to a possible functional origin of correlations. We have further investigated correlations, by looking also at the presence of non-trivial degree correlation in a number of networks of social, technological and biological nature. In Fig. 3.5, this typical behavior of the two-body degree correlations is shown for four networks of different nature (black symbols). In addition to this, we report in the same figure the ratio $\langle k_{nnn}^{obs}(k) \rangle / \langle k_{nnn}^{exp}(k) \rangle$, which indicates the nature of the three-body correlations (red symbols). Looking at both two- and three-body correlations, three main behaviors for the networks can be distinguished: assortative-assortative, null-null and dissasortative-assortative. By assortative-assortative we intend a situation where both the ratios $\langle k_{nn}^{obs}(k) \rangle / \langle k_{nn}^{exp}(k) \rangle$ and $\langle k_{nnn}^{obs}(k) \rangle / \langle k_{nnn}^{exp}(k) \rangle$ increase with k. In the null-null case, both the ratios are constant with k, while in the dissasortative-assortative case $\langle k_{nn}^{obs}(k) \rangle / \langle k_{nn}^{exp}(k) \rangle$ increases with k while $\langle k_{nnn}^{obs}(k) \rangle / \langle k_{nnn}^{exp}(k) \rangle$ decreases. However, despite the presence of these three main trends, it is hard to make an association between each trend and a specific network nature, e.g. social, technological, biological. Therefore, looking at three-body correlations in terms of average connectivity of second neighbors reveals not only that real networks are far from being markovian, but also that the usual classification based on degree-degree correlations probably fails to capture fundamental organizing principles of networks that are encoded in their topology.

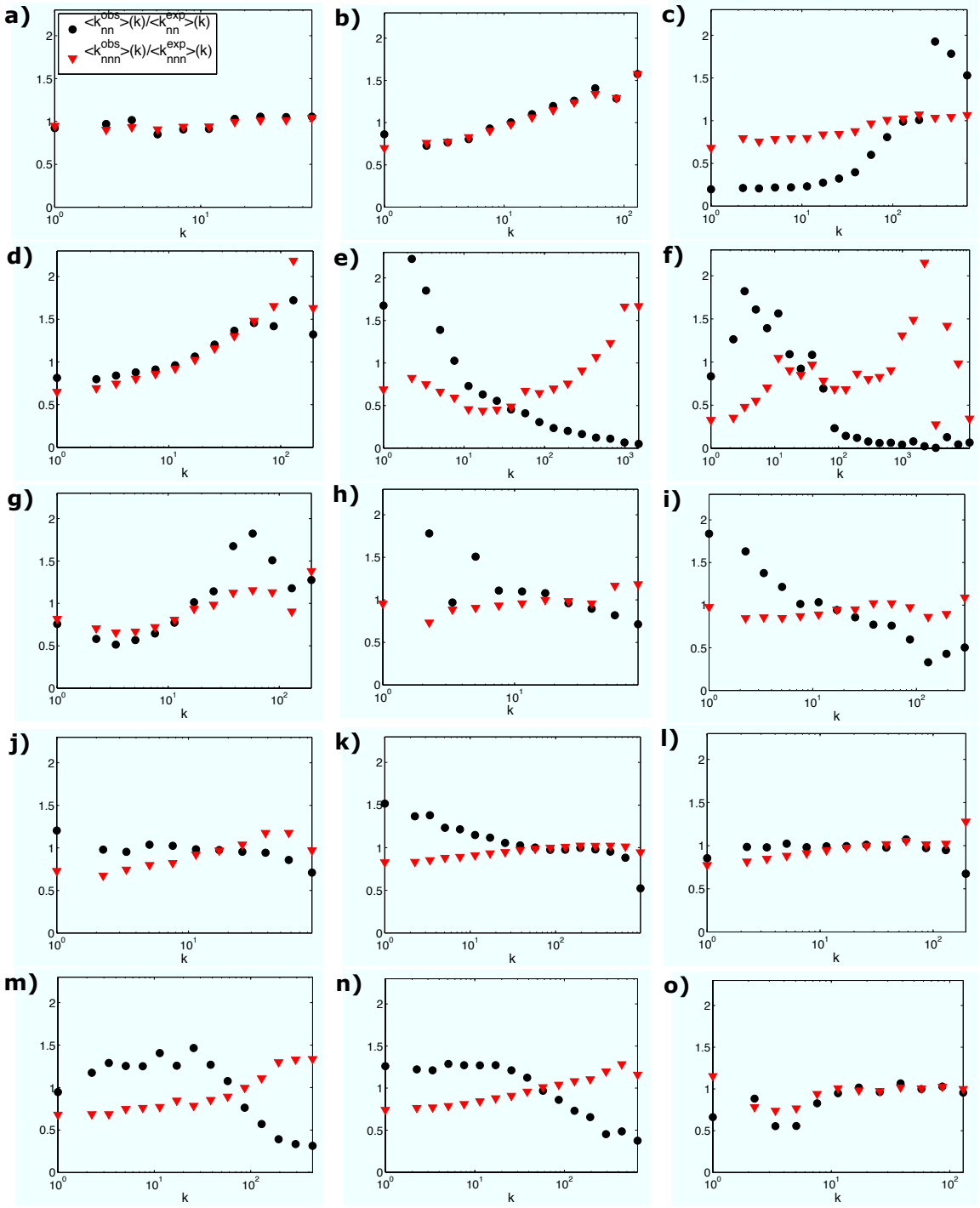

Figure 3.5: We compare the ratio between the observed average connectivity and the expected average connectivity of the first neighbors of a node, $\langle k_{nn}^{obs} \rangle (k) / \langle k_{nn}^{exp} \rangle (k)$ (black circles), with the ratio between the observed average connectivity and the expected average connectivity of the second neighbors of the same node, $\langle k_{nnn}^{obs} \rangle (k) / \langle k_{nnn}^{exp} \rangle (k)$ (red triangles), as a function of the degree k of the node in the following networks: (a) e-mail exchange at URV [41], (b) scientific collaboration (cond-mat)[42], (c) scientific collaboration (HepPh) [40], (d) Patents [40], (e) Internet AS [16], (f) WWW [43], (g) PGP [44], (h) Neural network *C. Elegans* [45], (i) Protein interactions *S. Cerevisiae* [46], (j) Protein interactions *S. Pombe* [47], (k) Private messages in Pardus [48], (l) Friendship in Pardus [48], (m) Enmity in Pardus [48], (n) Attacks in Pardus [48], (o) jazz musicians collaboration [49].

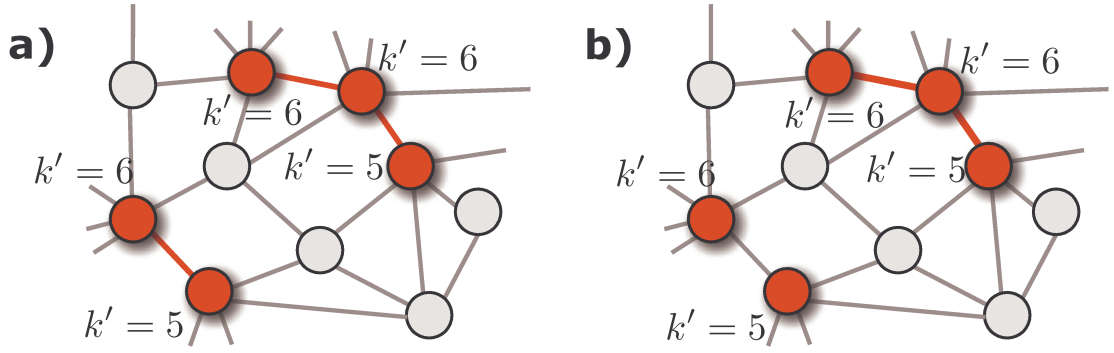

Figure 3.6: A schematic example to understand the rich-club phenomenon at the level of (a) two-body degree correlations and (b) three-body correlations. In the figure, we consider nodes that have degree k' higher than 4 to be rich. These nodes have been colored red in figure. Then, in the case we want to calculate the rich-club coefficient $\rho^{(2)}$ at the level of two-body correlations, we need to count the number of edges connecting rich nodes. In panel (a) these edges are bolded in red and amount to 3. Instead, to calculate the rich-club coefficient $\rho^{(3)}$ at the level of three-body correlations, we have to count the number of wedges connecting triplets of rich nodes. In panel (b) the edges forming the only "rich wedge" of this toy network are red bolded. Finally, to normalize properly $\rho^{(2)}$ and $\rho^{(3)}$, the number of rich edges and of rich wedges need to be counted in the corresponding uncorrelated network and markovian network respectively.

3.5 Revisiting the rich-club phenomenon

Another quantity where correlations have been found to play an important role is the so-called *rich-club ordering*. The rich-club phenomenon, a concept introduced first by Zhou and Mondragon [36] and then fully investigated by Colizza et al. [37], is the tendency of nodes with high degree, usually dominant elements of the system, to form tightly interconnected subgraphs. Actually, up to now the proposed measurements for this feature have been based only on the presence of particular connected pairs of nodes, hence being indirectly related to the presence of two-body correlations. However, as we will see below, the rich-club phenomenon can also manifest in the patterns of connectivity between triplets of nodes and then be associated to the presence of three-body correlations.

3.5.1 Rich-club and degree-degree correlations

The quantitative measurement of the rich-club phenomenon proposed by Colizza et al. [37] aims first at defining a rich-club coefficient $\phi(k)$, based on the presence of the "rich edges", and then compares it to $\phi^{u.c.}(k)$, the same coefficient measured in a network with the same degree distribution but with randomized connections. The rich-club

coefficient is defined as:

$$\phi(k) = \frac{\sum_{k'>k} \sum_{k''>k} E_{k'k''}}{\left(\sum_{k'>k} N_{k'}\right)\left(\sum_{k'>k} N_{k'} - 1\right)/2} \tag{3.13}$$

where $\sum_{k'>k} \sum_{k''>k} E_{k'k''}$ are all the "rich edges", being the richness determined by the presence of connected couples of nodes both with a degree higher than a threshold k; the denominator accounts for the maximum possible number of edges between nodes with degree higher than k (see also Fig. 3.6, panel (a) to understand how to individuate rich edges in a network). $\phi^{u.c.}(k)$, the rich-club coefficient for an uncorrelated network, can be derived analitically. Given the degree distribution P_k or, equivalently, the number of nodes N_k with the same degree k, $\phi^{u.c.}(k)$ reads:

$$\phi^{u.c.}(k) = \frac{\left(\sum_{k'>k} k' N_{k'}\right)^2}{\left(\sum_{k'>k} N_{k'}\right)\left(\sum_{k'>k} N_{k'} - 1\right)/2} \tag{3.14}$$

Finally a normalized function proposed by Colizza et al. [37] to highlight the presence of the rich-club phenomenon can be the following:

$$\rho^{(2)}(k) = \frac{\phi(k)}{\phi^{u.c.}(k)} = \frac{\langle k \rangle^2 \sum_{k'>k} \sum_{k''>k} P_{k'k''}}{(\sum_{k'>k} k' P_{k'})^2}. \tag{3.15}$$

Here we have used the superscript $^{(2)}$ to indicate that here the rich-club effect is connected to the presence of two-body degree correlations, and to distinguish this function from the one we will define in the next session using three-body degree correlations. Notice that the second term of Eq. 3.15 is easily obtained from 3.13and 3.14 using the relations $P_k = N_k/N$ and $P_{kk'} = E_{kk'}/2K$. A ratio $\rho^{(2)}(k)$ (Eq. 3.15) larger than 1 is the actual evidence of a rich-club phenomenon at the level of two-body relations as it is due to a larger number of edges between high-degree nodes than in the random case. In contrast, a ratio $\rho^{(2)}(k) < 1$ is a signature of an opposite organizing principle that leads to a lack of links among high-degree nodes in respect to the uncorrelated model.

3.5.2 Rich-club and three-body degree correlations

In order to investigate the role of three-body degree correlations in the emergence of the rich-club phenomenon, we define a quantity, $\rho^{(3)}(k)$, in the same spirit of Eq. 3.15. The definition of $\rho^{(3)}(k)$ is based on comparing the number of "rich" wedges, i.e. of wedges where all the nodes have degree higher than the threshold k, between a network and its markovian null model (see also Fig. 3.6, panel (b) to understand how to individuate rich wedges in a network). The normalized rich-club coefficient $\rho^{(3)}(k)$ at the level of

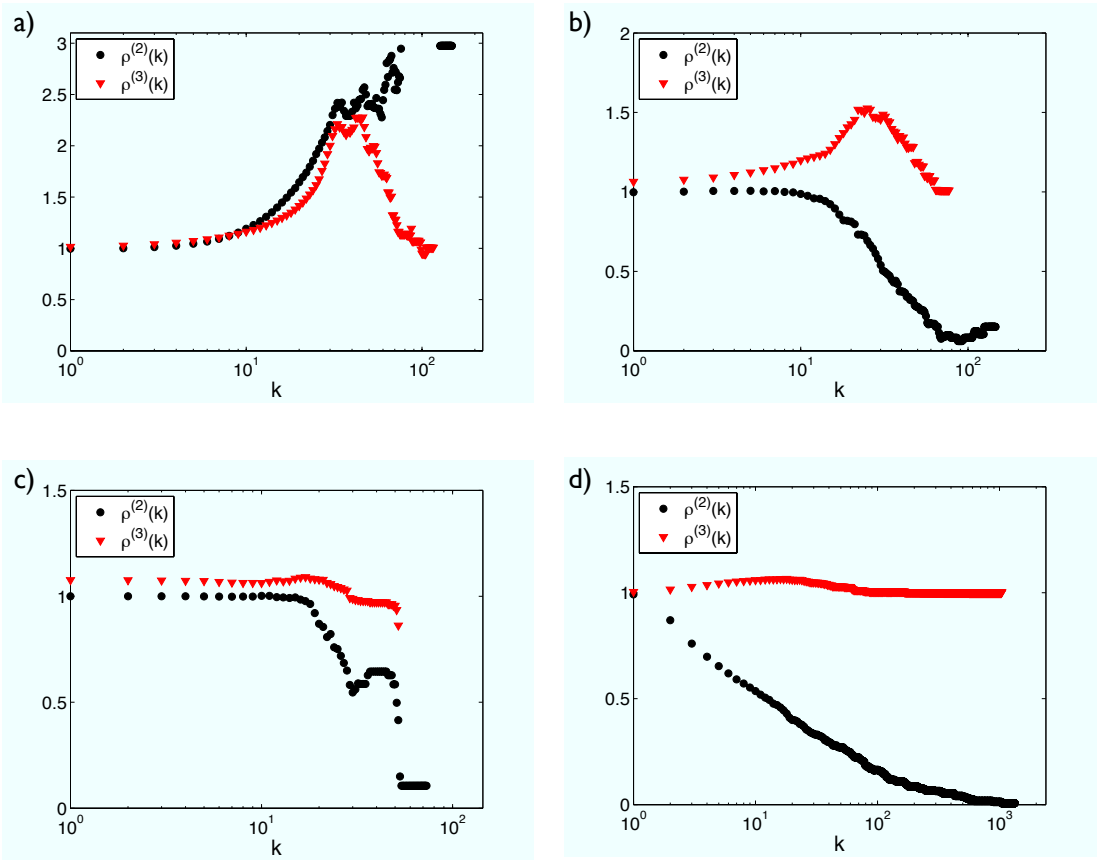

Figure 3.7: Study for the presence of the rich-club effect in (a) the patents collaboration network [40], (b) the protein interaction network of the yeast *Saccharomices Cerevisiae* [46], (c) the e-mail network at URV [41], and in the (d) Internet network at the level of autonomous system [16]. We plot the normalized rich-club coefficient based on two-body correlations, $\rho^{(2)}(k)$ (black circles), and our extension to three-body degree correlations, $\rho^{(3)}(k)$ (red triangles).

three-body correlations can be formally defined as follows:

$$\begin{aligned}\rho^{(3)}(k) &= \frac{\sum_{k'>k}\sum_{k''>k}\sum_{k'''>k}\left(P_{k'k''k'''}\right)}{\sum_{k'>k}\sum_{k''>k}\sum_{k'''>k}\left(P^{exp}_{k'k''k'''}\right)} = \\ &= \frac{\sum_{k'>k}\sum_{k''>k}\sum_{k'''>k}\left(P_{k'k''k'''}\right)}{\sum_{k'>k}\sum_{k''>k}\sum_{k'''>k}\left(\omega_{k''}P_{k'|k''}P_{k'''|k''}\right)}\end{aligned} \tag{3.16}$$

where the superscript $^{(3)}$ indicates that the formula accounts for three-body correlations and where $P^{exp}_{k'k''k'''}$ must be intended as the joint probability in markovian networks. The rich-club phenomenon can appear also at the level of three-body correlations. In fact, $\rho^{(3)}(k)$ (Eq. 3.16) larger than 1 indicates that there is a tendency of nodes of high-degree to be connected to each other in the form of wedges if compared to networks

with the same pattern of degree-degree correlations. Conversely, $\rho^{(3)}(k) < 1$ indicates that there is a sort of "anti rich-club" effect, in the sense that nodes of high degree avoid to be connected in wedge structures altogether in respect to the markovian null model. Notice that the presence of rich-club or anti rich-club at the level of three-body cannot be due to $\rho^{(2)}(k)$ being different from 1, since the markovian network used as null model encloses already the rich-club induced by degree-degree correlations. s

3.5.3 Rich-club phenomenon in real-world networks

We have investigated the presence of the rich-club phenomenon both at the level of two- and three-body degree correlations in a number of networks, the same networks where the average connectivity of first and second neighbors of nodes of degree k were analyzed. In Fig. 3.7 we report the plots of $\rho^{(2)}(k)$ and $\rho^{(3)}(k)$, as defined in Eqs. 3.15 and 3.16 respectively, for four representative networks of different nature. Clearly different scenarios emerge: there are cases where the rich-club appears at both at the level of two- and three-body correlations (panel (a)), in other situations there is a sort of anti rich-club at the level of degree-degree correlations while rich-club appears at the level of three-body correlations (panel (b)). Finally, there are cases where rich-club or anti rich-club appear at the level of two-body correlations while there is no effect at the level of three-body (panel (c) and (d)).

A strong presence of rich-club at the level of both two- and three-body degree correlations has been found in many social networks, especially in collaboration networks like the patents [40] and the scientific [42] collaboration networks. This provides support to the idea that the elite formed of influential people, for example prominent scientists in the scientific collaboration network, tends to form collaborative groups within specific domains. Surprisingly instead we find that protein interaction networks, like the one of the Saccharomices Cerevisiae [46], exhibits rich-club at the level of three-body. In fact, the decreasing behavior of the rich-club spectrum at the level of two-body correlations in the protein interaction networks in the past has been explained to be due to specific biological mechanisms [37, 50]: it seems that high-connected proteins preside over different functions and thus coordinate specific functional modules. However, the emergence of a clear rich-club phenomenon at the level of three-body provides new insights in the role of high-connected proteins in these networks, and requires further investigation of the biological mechanisms that are at work. Other kind of networks instead, like the technological ones, do not appear to be affected by the rich-club phenomenon at the level of three-body correlations, while there is a strong rich-club effect when considering two-body correlations.

In this chapter we have provided some novel tools to quantify three-body degree correlations in graphs and we have investigated their presence in a number of complex networks. We have been able to show that in most of the cases real-world networks have non-negligible three-body degree correlations. This can be fundamental in the study of

many dynamical process taking place of networks, such as in the case of random walks on graphs, which will studied in details in the following chapter. Also, the extension of the concept of rich-club ordering in terms of three-body degree correlations we have provided in this chapter will allow to uncover new underlying mechanisms in the organization of complex networks.

Chapter 4

Entropy rate of random walks on graphs

Scientist can't prove their theories;
they can only disprove, or improve, them.

JOHN HARTE

In the last decade an increasing attention has been devoted to the study of random walks on complex topologies [13, 51, 52]. Random walks are for example of fundamental importance for all searching processes. When googling a keyword, or searching a file in a peer-to-peer network, or even when looking for a hotel in a city we do not have a map of, the underlying process is a random walk on a graph.

Various features of random walks on networks, such as passage times [52–54] and spectral properties [55, 56] have been investigated, and random walks have also been used to detect communities [57, 58], to evaluate centrality of nodes [52, 59, 60] and to coarse–grain graphs [55]. Another quantity recently considered is the entropy rate, a measure to characterize the mixing properties of a stochastic process [34]. In particular, attention has been focused on designing random walks with *maximal entropy rate* on a given graph [61–65], i.e. choosing the transition probabilities of the random walk in such a way that the random walkers are maximally dispersing in the graph, exploring every possible walk with equal probability.

In this chapter, we introduce the concept of random walk and biased random walk on a graph. We formulate the random walk processes in terms of ergodic Markov chains, already introduced in chapter 2. We address the problem of the maximization of the entropy rate for biased random walks on a graph and we show how almost maximal-entropy random walks can be obtained with a limited and local knowledge of the network, a result reported in the publication [3]. In the last part of the chapter, we show an alternative formalism which allows to rephrase the problem of a biased

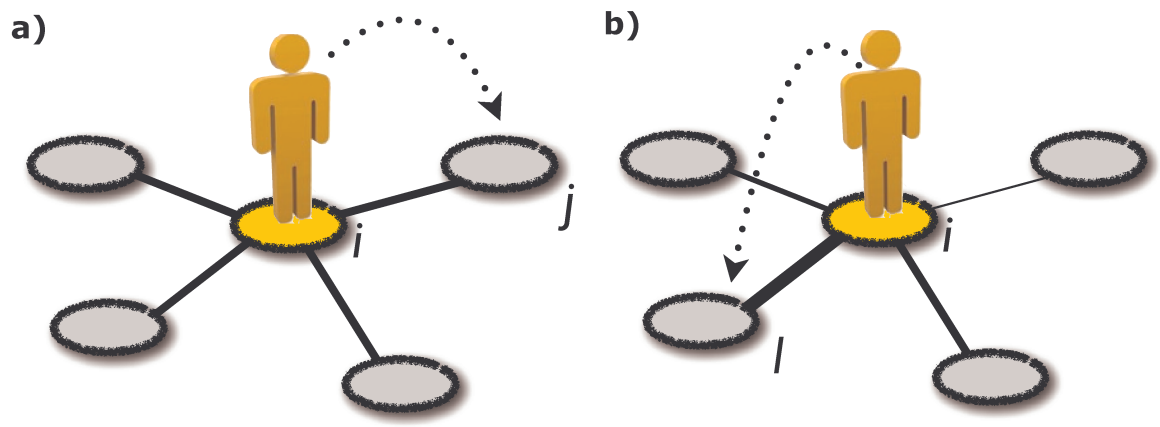

Figure 4.1: We report a schematic representation of (a) a plain random walk and (b) a biased random walk. In panel a) a walker, placed on node i decides to move to a neighbor node j with a probability $1/4$, i.e. with a probability which is equal for all neighbors of node i. In panel (b) different probabilities to be chosen are associated to the neighbors of i, expressed by different thickness of the links between i and its neighbors. In this case a walker performs a biased random walk, meaning that the walker will consider the neighbors of i with different probability to decide where to move. In (b) for example, the neighbor l has an higher probability to be the next node visited by the walker currently on node i.

random walk on graph as a plain random walk on a graph having the same topology but different weights. This latter study has been published in [4].

4.1 Random walks

A random walk, sometimes denoted as RW, is a mathematical formalisation of a trajectory that consists of taking successive random steps. One can define a random walk on a graph. Let us consider a connected, undirected and unweighted graph with N nodes and K links, described by the adjacency matrix $A = \{a_{ij}\}$. A random walk on a graph is a process where at each time step a walker moves producing a sequence of graph nodes: $\{i_0, i_1, i_2, \ldots, i_t\}$. If the walker at time t is at node i, at time $t+1$ it moves to one of its neighbors, say j, with a transition probability $\pi(j|i)$ (see Fig. 4.1 for a visual representation of a random walk). All the $\pi(j|i)$ are the entries of the probability transition matrix Π, having dimension N as the number of nodes in the graph. The generic element $\pi(j|i)$ is usually a function of the entry a_{ij} of the adjacency matrix and of a time-independent property of the node j. In this case, a random walk on a graph is a stochastic process which belongs to the special class of the invariant ergodic Markov chain, hence all the properties shown for the invariant ergodic Markov chain in Sec. 2.3 also hold for random walks on a graph.

4.1.1 Plain random walk

Given a graph, one can define different kinds of random walks, the simplest being the unbiased (or plain) random walk. We consider in this section this simplest version of the random walk and we illustrate some properties of the transition matrix, which also hold for other kinds of random walks as it will be remarked in the following paragraphs.

In a unbiased random walk, a walker currently on node i chooses to move to a neighbor node j with a probability $\pi(j|i)$ which is the same for all neighbors j. The probability $\pi(j|i)$ reads:

$$\pi(j|i) = \frac{a_{ij}}{\sum_l a_{il}} = \frac{1}{k_i} \tag{4.1}$$

Even if the adjacency matrix $\mathcal{A}$ is symmetric, the transition matrix Π is not in general symmetric, unless in the special case in which the graph is regular. We can write $\Pi = A^T D^{-1}$ where D is the diagonal matrix with $(D)_{ii} = k_i$. Since the walker must move from a node to somewhere, the normalization condition $\sum_j \pi_{ji} = 1$ must hold. We say that matrix Π is *stochastic*. In fact, it satisfies the following properties of stochastic matrices. A real square matrix S of order N is said *stochastic* iff:

1. all entries are numbers from the interval $[0, 1]$,
2. the sum of each column is 1.

Property 2 can be written as: $1^T \Pi = 1^T$, indicating that 1 is an eigenvalue of the transition matrix Π with left eigenvector 1^T. This proves that stochastic matrices have always 1 as eigenvalue. In addition to this, stochastic matrices does not admit eigenvalues with absolute value greater than 1. This is immediately shown by the general property that the spectral radius ρ of a matrix [1] is less or equal all natural norms of the matrix. The L^1 norm of a stochastic matrix Π is given by:

$$\|\Pi\|_1 = \max_j \sum_i |\pi_{ij}| = \max_j \sum_i \pi_{ij} = 1$$

and therefore $\rho(\Pi) \leq 1$. Note that 1 is an eigenvalue of Π with multiplicity one, because matrices A and Π are irreducible (the graph is connected) [32].

For an unbiased random walk, the trajectories are defined in terms of transition probabilites, and also the node i_t occupied by the walker at time t, is given in probabilistic terms, as we have seen also in Sec. 2.1.1. Let us denote by $p_j(t)$ the probability that at time t the walker is at node j (with $j = 1, 2, \ldots, N$):

$$p_j(t) = Prob(i_t = j) \tag{4.2}$$

We can imagine to calculate such probabilities by successively repeating various realizations of the walker motion, each realization starting at the same initial node. This

[1] The spectral radius $\rho(A)$ of a $N \times N$ matrix A with eigenvalues λ_i, with $i = 1, \ldots, N$, is defined as: $\rho(A) = \max_{1 \leq i \leq N} |\lambda_i|$

is equivalent to consider an ensemble of M independently moving walkers. If by $M_j(t)$ we indicate the number of walkers at node j at time t, the probability $p_j(t)$ can be approximated as $p_j(t) \approx M_i(t)/M$, when M is very large. Probabilities $p_j(t)$ satisfy the normalization condition: $\sum_{j=1}^{N} p_j(t) = 1$ at each time t. Being $p_j(t)$ the probability that our random walker is at node j at time t, then the probability $p_i(t+1)$ of its being at i one step later is:

$$p_i(t+1) = \sum_j \pi_{ij} p_j(t) \tag{4.3}$$

This equation is equivalent to the evolution equation Eq. 2.11 written for a time-invariant Markov chain in chapter 2. It is easy to verify that, if $\sum_j p_j(t) = 1$, then we also have $\sum_j p_j(t+1) = 1$. Writing the probabilities $p_j(t)$ as a N-dimensional column vector $\mathbf{p}(t)$:

$$\mathbf{p}(t) \equiv \begin{pmatrix} p_1(t) \\ p_2(t) \\ \vdots \\ p_N(t) \end{pmatrix} \tag{4.4}$$

the rule of the walk can be expressed in matricial form as a first order difference equation:

$$\mathbf{p}(t+1) = \Pi \mathbf{p}(t). \tag{4.5}$$

The solution of the equation is given by

$$\mathbf{p}(t) = \Pi \cdot \Pi \cdot \ldots \cdot \Pi \; \mathbf{p}(0) = \Pi^t \mathbf{p}(0). \tag{4.6}$$

or, in components,

$$p_j(t) = \sum_i \pi_{ji}^{(t)} p_i(0), \tag{4.7}$$

where $\pi_{ji}^{(t)} = (\Pi^t)_{ji}$ gives the probability $P_{i\to j}(t)$ that a walker starting from node i reaches node j in t steps.

A fixed point solution of Equation 4.5 is a probability distribution $\mathbf{p}^*$ such that:

$$\mathbf{p}^* = \Pi \mathbf{p}^*. \tag{4.8}$$

Vector $\mathbf{p}^*$ is the right eigenvector of Π with eigenvalue 1, and is called the *stationary* or *invariant distribution*, because it corresponds to a probability distribution that is mapped into itself by the time evolution.

Since the graph is connected (matrix Π is irreducible [32]), then the eigenvalue 1 is a simple root of the characteristic equation of Π and the stationary distribution $\mathbf{p}^*$ is *unique*. In this case, the equilibrium distribution for un unbiased random walk is given by:

$$p_i^* = \frac{k_i}{2K} \tag{4.9}$$

which is easily proved by plugging it in Eq. 4.8 and seeing that the equality is verified. The meaning of Eq. 4.9 is that if we explore through a random walk a connected undirected graph, the walker visits nodes with a probability proportional to their degree. An alternative statement of the same result is that the random walk visits the edges of the graph uniformly.

4.1.2 Biased random walk

A more general class of random walks are the so-called *biased random walks* (BRWs). In this process, at each time step, a walker currently at node i chooses to move to one of the first neighbors of i, say j, with a probability $f_j \equiv f(x_j)$ depending on the node property x_j. The node property x can be topological (like degree, betweenness, clustering coefficient, etc.) or any other quantity relevant to a diffusion dynamics (for example node congestion, healthy state, etc.). Such random walks can also be described with a transition probability matrix Π which is also a stochastic matrix as in plain random walk (see previous section). The entries of Π for a biased random walk of the kind described above read:

$$\pi(j|i) = \frac{a_{ij} f_j}{\sum_l a_{il} f_l} \tag{4.10}$$

Since a BRW is also an ergodic Markov chain, there is a unique stationary distribution $\mathbf{p}^*$, expressing the probability of nodes to be occupied at equilibrium by the walkers. This distribution is in general different from the one found for unbiased random walks in 4.9, and reads [65]:

$$p_i^* = \frac{c_i f_i}{\sum_j c_j f_j} \tag{4.11}$$

where $c_i = \sum_l a_{il} f_l$.

Degree biased random walk

A special and interesting case of biased random walk is that where f_j has a power law dependence on the property x_j and where the bias is represented by the degree of neighbor nodes [65]. In this case, we will have $f_j = k_j^\alpha$ and the entries of the transition probability matrix will read:

$$\pi(j|i) = \frac{a_{ij} k_j^\alpha}{\sum_l a_{il} k_l^\alpha} \tag{4.12}$$

α is a real number. In the case $\alpha > 0$, a walker has a bias to move to nodes with higher degree, while when $\alpha < 0$ a walker avoids high degree nodes preferring low degree ones. For $\alpha = 0$ the unbiased random walk is recovered. In this case, the stationary distribution $\mathbf{p}^*$ is immediately derived from Eq. 4.11 plugging in the bias $f_j = k_j^\alpha$.

We obtain:

$$p_i^* = \frac{k_i^\alpha \sum_j a_{ij} k_j^\alpha}{\sum_l k_l^\alpha \sum_j a_{lj} k_j^\alpha} \tag{4.13}$$

4.2 Maximal-entropy random walk

In the last decade, particular attention has been focused on designing random walks with *maximal entropy rate* on a given graph [61–65], i.e. choosing the transition probabilities of the random walk in such a way that the random walkers are maximally dispersing in the graph, exploring every possible walk with equal probability and hence maximizing the entropy rate h defined in Sec. 2.6.1. Practical examples where the maximization of entropy rate is important are diffusion processes which aim at well-mixing, such as spreading information about a node's state (its healthy or infected condition, its availability or congestion, etc.) [65], mixing in meta-populations models [66], or global synchronization of moving agents by local entrainment [67].

4.2.1 Entropy rate and random walks

The optimal random walk on a given graph can be rigorously determined on mathematical grounds by considering the properties of entropy rate h of the stochastic processes, introduced in Sec. 2.6 associated to different random walks [34]. A trajectory of t steps generated by a random walk starting at a fixed node i is described by the sequence of occupied nodes $i, i_1, i_2, \ldots, i_t$, where i_1,..., i_t are all indices that can take integer values between 1 and N. This means that the walker first moves from i to node i_1, then it jumps to node i_2 and so on. In practice, there is a maximum of $M(t)$ different allowed sequences of length t, corresponding to all possible walks of length t (and starting at node i) on the graph under study. Depending on the rules of the random walk, not all possible sequences will appear, while some of them will occur with a probability higher than the others. If we denote as joint probability $p(i, i_1, i_2, \ldots, i_t)$ the probability that the sequence $i, i_1, i_2, \ldots, i_t$ is generated by a given random walk, then the entropy rate of the random walk, h, is defined as:

$$h = \lim_{t\to\infty} \frac{H_t}{t} \,, \tag{4.14}$$

where H_t is the joint entropy of the set of trajectories of length t starting at node i: $H_t = -\sum_{i_1,i_2,\ldots,i_t} p(i, i_1, \ldots, i_t) \ln p(i, i_1, \ldots, i_t)$. The value of the entropy rate in Eq. 4.14 can be calculated directly from matrix π, as for any ergodic Markov chain (see Sec. 2.3.1), from:

$$h = -\sum_{i,j} \pi(j|i) \cdot p^*(i) \ln \left[\pi(j|i)\right] \,. \tag{4.15}$$

where $p^*(i)$ is the i^{th} component of the stationary distribution.

4.2.2 Maximum entropy rate

The minimum possible value of the entropy rate, $h_{\min} = 0$, is obtained when, for large time t, only one trajectory dominates. On the other hand, the maximum possible value is obtained when, for large time t, all the $M(t)$ allowed trajectories have equal probability to occur, i.e. $p(i, i_1, \ldots, i_t) = 1/M(t)$ if $i, i_1, \ldots, i_t$ is a walk on the graph originating in i, and $p(i, i_1, \ldots, i_t) = 0$ otherwise. The maximum value of the entropy is equal to: $h_{\max} = \lim_{t\to\infty} \frac{M(t)}{t}$. The number of trajectories of length t between a given pair of nodes (i, j) is simply given by the ij entry of the t_{th} power of the adjacency matrix A of the graph. Therefore, the number of all possible trajectories of given length t on a graph, between any pair of nodes, is obtained by summing all the entries of the matrix A^t:

$$M(t) = \sum_{i,j} \left(A^t\right)_{ij}$$

Therefore, the maximum entropy rate on a given graph, can be expressed as:

$$h_{\max} = \lim_{t\to\infty} \frac{\sum_{i,j} \left(A^t\right)_{ij}}{t}$$

Now, due the diagonalization properties of the adjacency matrix A, the limit for $t \to \infty$ of $\frac{A^t}{t}$ will converge to λ_1, the largest eigenvalue of A. This implies that the maximum value of the entropy rate one can obtain is:

$$h_{\max} = \ln \lambda_1 \tag{4.16}$$

4.3 Exact solution for the maximal-entropy random walk

In principle, the optimization of entropy rate could require the definition of transition probabilities relying on the history of the walker's positions. However, it has been proven that allowing a long-term memory of the past is not needed in order to construct maximal-entropy random walks, since it turns out that there always exists an optimal set of transition probabilities that is Markovian [61–64].

Namely, the maximum entropy rate Eq. 4.16 can be obtained with a Markov random walk in which the probability to step from node i to node j is equal to:

$$\pi(j|i) = \frac{a_{ij} u_j}{\lambda_1 u_i}, \tag{4.17}$$

where $\mathbf{u}$ is the eigenvector of the adjacency matrix A associated to the largest eigenvalue λ_1, or in other words: $A\mathbf{u} = \lambda_1 \mathbf{u}$. By using the formula for the entropy rate of an ergodic Markov chain Eq. 4.15, it is immediate to prove that a random walk with

transition matrix 4.17 yields a process with maximum entropy rate equal to $\ln \lambda_1$ [64]. This random walk process has the interesting property to be biased, in the sense that a walker follows a link (i, j) with a probability proportional to the importance of its end j, as measured by its eigenvector centrality u_j [68].

4.4 Maximal-entropy random walk with local information

The main problem with a real implementation of the random walk described by Eq. 4.17 is that, at each time step, the walker needs to have a global knowledge of the network: it needs to know the adjacency matrix of the entire graph. Such global information is very oftenunavailable. A walker at a node i usually has only a local information, in the sense that it knows the first neighbors of node i, and possibly some of their topological properties, such as their degree [65]. We will show that it is possible to construct random walks that have almost maximal entropy rate and that make use only of local information or of information which is locally available.

In the previous sections we have mentioned that, in order to have a process with maximum entropy rate, one needs to sample with the same probability all the $M(t)$ allowed trajectories of the same length t, i.e. $p(i, i_1, \ldots, i_t) = 1/M(t)$ if $i, i_1, \ldots, i_t$ is a walk on the graph originating in i, and $p(i, i_1, \ldots, i_t) = 0$ otherwise.

In the most general case, the probability of having a sequence of t nodes originating at a given node i can be written (for any $t > 1$) in terms of conditional probabilities as:

$$p(i, i_1, \ldots, i_t) = p(i_1|i)p(i_2|i, i_1) \ldots p(i_t|i, i_1, \ldots, i_{t-1}).$$

Summing both ends over $i_2, i_3, \ldots, i_t$, and by using the normalization conditions $\sum_{i_t} p(i_t|i, i_1, i_2, \ldots, i_{t-1}) = 1$ [2] for $t \geq 2$, we get an expression for the conditional probability at the first step as a function of the t-times joint probabilities:

$$p(i_1|i) = \sum_{i_2, i_3, \ldots, i_t} p(i, i_1, \ldots, i_t) \, . \tag{4.18}$$

This means that, no matter how long is the memory in the random walker, we can always describe it as a Markov random walker, provided that we define the transition matrix of the Markov chain $\pi(i_1|i)$ in terms of the joint probabilities $p(i, i_1, \ldots, i_t)$ as in Eq. 4.18. In particular, if we want to construct a maximal-entropy random walk, we have to set $p(i, i_1, i_2, \ldots, i_t) = 1/M(t)$ iff $i, i_1, i_2, \ldots, i_t$ is a walk on the graph, and $p(i, i_1, i_2, \ldots, i_t) = 0$ otherwise. The number of walks of length t originating in i can be written in terms of the adjacency matrix as: $M(t) = \sum_{i_1, i_2, \ldots, i_t} a_{ii_1} a_{i_1 i_2} \ldots a_{i_{t-1} i_t}$.

[2] By definition, $\sum_B P(B|A)$. In fact given that we are on a node A, the probability that we move to node B, summed over all possibilities B must be 1, i.e. the certain event.

Hence, the joint probability of a trajectory $i, i_1, i_2, \ldots, i_t$ reads:

$$p(i, i_1, \ldots, i_t) = \frac{a_{ii_1} a_{i_1 i_2} \ldots a_{i_{t-1} i_t}}{\sum_{i_1, i_2, \ldots, i_t} a_{ii_1} a_{i_1 i_2} \ldots a_{i_{t-1} i_t}} \,, \tag{4.19}$$

and the transition matrix of the Markov random walker with the maximal entropy is finally given by:

$$\pi(i_1|i) = \lim_{t\to\infty} \frac{a_{ii_1} \sum_{i_2} a_{i_1 i_2} \ldots \sum_{i_t} a_{i_{t-1} i_t}}{\sum_{i_1} a_{ii_1} \sum_{i_2} a_{i_1 i_2} \ldots \sum_{i_t} a_{i_{t-1} i_t}} \,. \tag{4.20}$$

From Eq. 4.20 it is clear that, in the most general case, in order for a walker at a node i to select one of its first neighbors to step on, the walker needs to know not only which node is in $\mathcal{N}_i$, but also the neighborhood of first neighbors, the neighborhood of second neighbors, and so on. In practice, the local choice of moving from i to one particular neighbor i_1, depends on the whole adjacency matrix of the graph. However, as we demonstrate below, this global information is not necessary in most of the cases.

4.4.1 Maximal-entropy random walk on uncorrelated networks

Uncorrelated graphs can be described only by means of the degree sequence of the nodes $\{k(1), k(2), \ldots, k(N)\}$, corresponding to a degree distribution P_k, since the degree of a node does not depend on the degree of its first neighbors (see also Sec. 1.3.2). In mathematical terms, this means that the conditional probability $P_{k'|k}$ does not depend on k, and can be written in terms of the degree distribution as: $P^{\text{unc}}_{k'|k} = k' P_{k'} / \langle k \rangle$ where the right hand side is the probability to end up in a node of degree k' by choosing an edge at random with uniform probability. Consequently, the average degree of the neighbors of node j, $k_{nn}(j) = 1/k(j) \sum_l a_{jl} k(l)$, does not depend on the degree of j, $k_{nn}(j) = k_{nn}\ \forall j$, and the last two summations in the numerator and in the denominator of Eq. 4.20, namely $\sum_{i_{t-2}} a_{i_{t-3} i_{t-2}} \sum_{i_{t-1}} a_{i_{t-2} i_{t-1}} k(i_{t-1}) = \sum_{i_{t-2}} a_{i_{t-3} i_{t-2}} k(i_{t-2}) k_{nn}(i_{t-2})$ can be written as $k_{nn} \sum_{i_{t-2}} a_{i_{t-3} i_{t-2}} k(i_{t-2})$. The constant k_{nn} at the numerator and at the denominator cancels out, so that the same argument can be repeated again and again. Finally, the formula factorizes into:

$$\pi^1(i_1|i) = \frac{a_{ii_1} k(i_1)}{\sum_{i_1} a_{ii_1} k(i_1)} \,. \tag{4.21}$$

where, by the symbol π^1 we mean the first order approximation to the transition matrix π in Eq. 4.20. This formula tells us that the best diffusion process on a uncorrelated graph is a random walk whose motion is linearly biased on node degrees. Thus, a walker at a given node, only needs to have information on its first neighbors and their degree. Since the degrees of different nodes are not correlated, local information of the degree of first neighbors is, in this case sufficient to construct the diffusion process with maximal

entropy. Such information is "locally available" to the walkers, meaning that a walker at node i has complete information on the degree of each node in its neighborhood $\mathcal{N}_i$. Now, it is intuitive that a random walk choosing a node j proportionally to $k(j)$, so that all the trajectories of length 2 starting in i will occur with the same probability, will be more random than a walker selecting uniformly the first neighbors of i. Formula 4.21 is a special case of the more general degre-biased random walk of Eq. 4.12, which becomes Eq. 4.21 for $\alpha = 1$.

Of course, if all nodes have the same degree, as in a regular graph, the transition matrix reduces to that of an unbiased walker:

$$\pi^0(i_1|i) = \frac{a_{ii_1}}{\sum_{i_1} a_{ii_1}} \, . \tag{4.22}$$

This is the lowest possible approximation for π in Eq. 4.20: in the case of no available information, each neighbor has the same probability to be selected. The values of h obtained numerically with transition matrices π^0 and π^1 in different models of uncorrelated networks are reported in Table 4.1. In agreement with our predictions, in regular lattices and in random regular graphs, $h(\pi^0)$ is equal to the maximal possible entropy $h_{\max} = \ln \lambda$. In Erdős-Rényi (ER) random graphs not all nodes have the same degree, so that a random walk linearly biased on degree has an entropy $h(\pi^1)$ that is much closer to the maximum, than $h(\pi^0)$. This effect is even more evident in scale-free graphs, i.e. in graphs with a very heterogeneous degree distribution. This is the case of Barabasi-Albert graphs (see Sec. 1.3.4) and of configuration graphs (see Sec. 1.3.2) constructed starting from a power-law degree distribution. Both these graphs have an heterogenous degree distribution since there are many nodes with low degree and just a few very highly connected nodes.

4.4.2 Maximal-entropy random walk on networks with degree-degree correlations

Graphs with degree-degree correlations are described in terms of their degree distribution P_k, and of a non-trivial $P_{k'|k}$. This is because the probability that a link from a node of degree k arrives at a node of degree k' does not simply factorize in terms of the degree distribution. In such graphs the average degree of the first neighbors of a node j, $k_{nn}(j)$, does depend on $k(j)$. Therefore, in analogy with Eq. 4.21 we can define a second order approximation of Eq. 4.20:

$$\begin{aligned} \pi^2(i_1|i) &= \frac{a_{ii_1} \sum_{i_2} a_{i_1 i_2} k(i_2)}{\sum_{i_1} a_{ii_1} \sum_{i_2} a_{i_1 i_2} k(i_2)} = \\ &= \frac{a_{ii_1} k(i_1) k_{nn}(i_1)}{\sum_{i_1} a_{ii_1} k(i_1) k_{nn}(i_1)} \, , \end{aligned} \tag{4.23}$$

	$\frac{h(\pi^0)}{h(\pi)}$	$\frac{h(\pi^1)}{h(\pi)}$	$\frac{h(\pi^2)}{h(\pi)}$	$h_{\max} = h(\pi)$
Regular lattice	1.000	1.000	1.000	1.79
Random regular graph	1.000	1.000	1.000	1.79
ER random graph	0.954	0.993	0.998	1.98
Uncorr. scale-free $\gamma = 1.5$	0.886	0.992	0.996	2.36
BA model	0.825	0.976	0.996	2.52
Assort. scale-free $\gamma = 1.5$	0.876	0.991	0.999	2.44
Disassort. scale-free $\gamma = 1.5$	0.937	0.990	0.997	2.18
Internet AS [16]	0.744	0.900	0.980	4.10
US Airports [66]	0.879	0.990	0.997	3.88
E-Mail [41]	0.881	0.983	0.997	3.03
SCN (cond-mat) [42]	0.694	0.867	0.946	3.17
SCN (astro-ph) [42]	0.784	0.941	0.973	4.41
PGP [44]	0.597	0.92	0.976	3.75

Table 4.1: The entropies of random walks with no information, $h(\pi^0)$, and with local information respectively on nearest, $h(\pi^1)$, and next-nearest neighbors, $h(\pi^2)$, are compared to the maximal possible entropy $h_{\max} = h(\pi) = \ln \lambda$ on different graph models with $N = 500$ and average degree $\langle k \rangle = 6$ and on various real networks.

describing a Markov walker that, at each time step, selects a first neighbor, i_1, of the current node, with a probability proportional to the sum of the degrees of the first neighbors of i_1. This is equivalent to make equiprobale all the walks of length 3 originating in i. In conclusion, to construct high-entropy random walks on correlated graphs, a walker at a given node needs to know the degree of first and second neighbors of the current node, which is still local information.

In Table 4.1 we report $h(\pi^2)$ for various models and for real networks. In models of uncorrelated graphs $h(\pi^2)$ is not very different from $h(\pi^1)$, while in models of correlated graphs, in lattices with defects and in most of the networks from the real world $h(\pi^2)$ is a much better approximation of $h(\pi)$ than $h(\pi^1)$. As we have seen in Sec. 1.1.7 and 3.1.1, in most real-world networks degree-degree correlations are such that the average degree of the first neighbors of a node exhibits a clear power-law dependence on degree: $k_{nn}(j) \sim [k(j)]^{-\nu}$, with $\nu > 0$ ($\nu < 0$) for disassortative (assortative) networks [13]. For instance, as shown in the inset of Fig. 4.2, $\nu \simeq 0.4$ for the Internet at the autonomous systems level [16]. Plugging this dependence in Eq. 4.23, we get an approximate form for the maximal-entropy random walk in a correlated random graph in terms of degree-biased random walks:

$$\pi^2(i_1|i) \simeq \frac{a_{ii_1}[k(i_1)]^{1-\nu}}{\sum_{i_1} a_{ii_1}[k(i_1)]^{1-\nu}} \,. \tag{4.24}$$

In practice, on a correlated network, an approximation for the maximal-entropy random walk can be obtained by considering a random walk whose motion is biased as a power

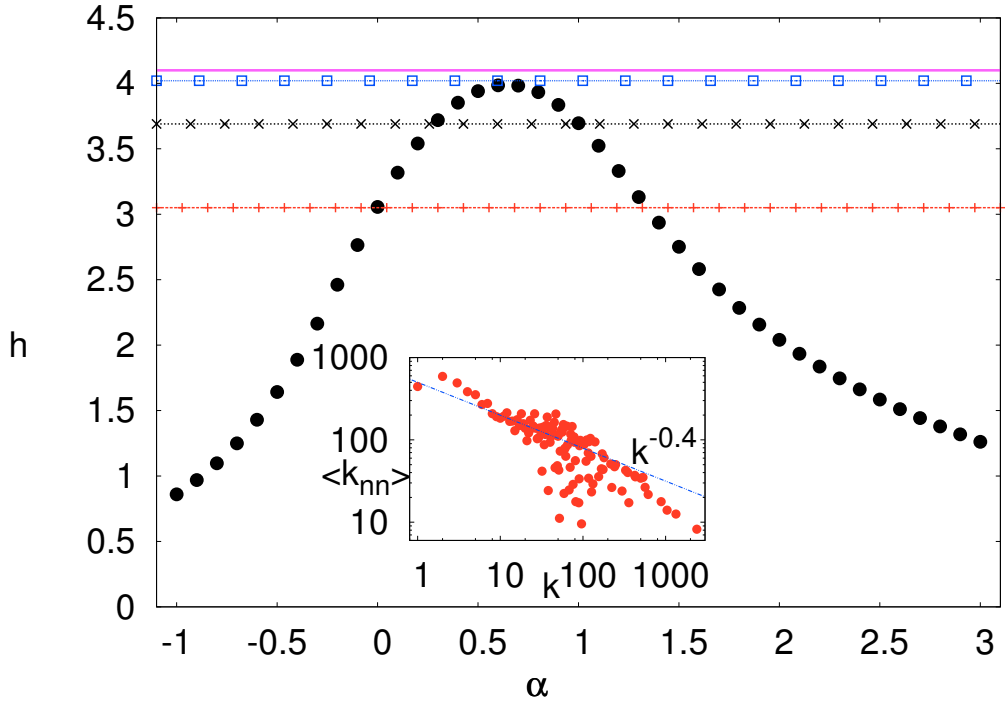

Figure 4.2: Entropy rate of power-law biased random walks as a function of the degree exponent α for network Internet AS [16]. Horizontal lines correspond to, from bottom to top, $h(\pi^0)$, $h(\pi^1)$, $h(\pi^2)$ and $h_{\max} = h(\pi)$. (Inset) Average degree k_{nn} of the first neighbors of nodes of degree k, as a function of k, with fit $k^{-0.4}$.

of the target node degree, with an exponent $\alpha = 1 - \nu$. Hence, the optimal bias α_{opt} is larger (smaller) than 1 for assortative (disassortative) networks, meaning that we have to prefer a super-linear (sub-linear) bias on the node degree. As an example, in Fig. 4.2 we report the entropy rate of a biased random walk as a function of the exponent α on a disassortative real-world network. We found $\alpha_{opt} = 0.6$ for Internet AS, which is perfectly in agreement with the value $\nu = 0.4$ in the inset, through the relation $\alpha_{opt} = 1 - \nu$. We have also checked that this relation holds for the other real networks in Table 4.1.

4.4.3 Maximal-entropy random walk on networks with higher-order degree-correlations

Similar arguments can be repeated for networks with higher-order correlations. This procedure generates a class of biased random walks defined by the transition matrices π^0, π^1, π^2, etc, incorporating more and more information about the system structure. In Sec. 4.4.4 we studied how this sequence of transition matrices converges to π in different networks. In the limit case in which a graph has correlations at all orders,

we have to rely on the full transition matrix of Eq. 4.20, which can be also expressed by means of the eigenvalues and eigenvectors of the adjacency matrix of the graph. In fact, the numerator and the denominator of Eq. 4.20 can be rewritten in terms of powers of the adjacency matrix, respectively as $a_{ii_1}\sum_{i_t}(A^{t-1})_{i_1 i_t} = a_{ii_1}(A^{t-1}\cdot\mathbf{1})_{i_1}$ and $\sum_{i_t}(A^t)_{ii_t} = (A^t\cdot\mathbf{1})_i$, where $(A^t)_{ij}$ indicate the entry i,j of matrix A^t, and $\mathbf{1}$ is a vector of ones. By making use of the power method for $t\to\infty$, we finally get:

$$\pi(i_1|i) = \frac{a_{ii_1}u_{i_1}}{\lambda u_i} = \frac{a_{ii_1}u_{i_1}}{\sum_j a_{ij}u_j}\ . \tag{4.25}$$

where λ and vector $\mathbf{u}$ are respectively the largest eigenvalue and its associated eigenvector of the adjacency matrix [3]. Eq. 4.25 represents a Markov walk whose transition probability is linearly biased by the components of eigenvector $\mathbf{u}$, also known as the eigenvector centrality of the node [68], and it is indeed the same transition matrix proposed in [64] as the process with the maximum possible entropy rate $h_{\max} = \ln\lambda$ [61–64].

4.4.4 Kullback–Leibler divergence

In order to test the quality of the approximations of different orders we considered the Kullback-Leibler divergence between the transition matrix π in Eq. 4.17 and the transition matrices which use only local information. As explained in Sec. 2.5, given two discrete distributions $P=\{p_i\}$ and $Q=\{q_i\}$, the Kullback–Leibler divergence $D_{KL}(P|Q)$, measures the amount of extra information required to represent P by using only information about Q. It is calculated by averaging the logarithmic distance between P and Q with a weight given by the probability P:

$$D_{KL}(P|Q) = \sum_i p_i \ln\frac{p_i}{q_i} \tag{4.26}$$

[3] For uncorrelated graphs, $u_j \sim k(j)$ [69], and Eq. 4.25 reduces to Eq. 4.21.

$D_{KL}(\cdot)$	$(\pi\|\pi^0)$	$(\pi\|\pi^1)$	$(\pi\|\pi^2)$	$(\pi\|\pi^3)$	$(\pi\|\pi^4)$
Internet AS	0.784	0.163	0.089	0.032	0.031
US airports	0.928	0.176	0.072	0.011	0.001
E-mail	0.724	0.137	0.045	0.019	0.009
SCN (cond-mat)	1.796	0.900	0.737	0.576	0.471
SCN (astro)	2.499	1.167	0.805	0.570	0.417
PGP	1.529	0.729	0.529	0.387	0.282

Table 4.2: Kullback–Leibler divergence between π and successive approximations π^k for different real and synthetic networks.

In a sense, D_{KL} measures how much P and Q are different. The Kullback–Leibler divergence can be calculated also in the case P and Q represent two matrices. Indeed, if $P = \{p_{ij}\}$ and $Q = \{q_{ij}\}$, the Kullback–Leibler divergence in this can be written as:

$$D_{KL}(P|Q) = \sum_{i,j} p_{i,j} \ln \frac{p_{i,j}}{q_{i,j}}$$

We have computed the Kullback–Leibler divergence between the transition matrix π in Eq. 4.17 and the transition matrices π^0, π^1, π^2, π^3, π^4 corresponding to local approximations of increasing order. The expressions for π^0, π^1, π^2 are respectively given in Eq. 4.22, Eq. 4.21 and Eq. 4.23, while π^3 and π^4 are defined as follows:

$$\pi^3(i_1|i) = \frac{a_{ii_1} \sum_{i_2} a_{i_1 i_2} k(i_2) k_{nn}(i_2)}{\sum_{i_1} a_{ii_1} \sum_{i_2} a_{i_1 i_2} k(i_2) k_{nn}(i_2)} \tag{4.27}$$

$$\pi^4(i_1|i) = \frac{a_{ii_1} \sum_{i_2 i_3} a_{i_1 i_2} a_{i_2 i_3} k(i_3) k_{nn}(i_3)}{\sum_{i_1} a_{ii_1} \sum_{i_2 i_3} a_{i_1 i_2} a_{i_2 i_3} k(i_3) k_{nn}(i_3)} \tag{4.28}$$

Notice that the choice of transition matrix π^k guarantees that all the walks of length $k+1$ are equiprobable. Therefore, the values of $D_{KL}(\pi|\pi^k)$ measure the inaccuracy in using the process π^k, which makes equiprobable walks of length $k+1$, with respect to using process π, which makes equiprobable walks of infinite length. In Table 4.2 we report the values of $D_{KL}(\pi|\pi^k), k = 0, 1, 2, 3, 4$, obtained for the six real networks considered in Table I of the main text.

In the first three networks in the table, $D_{KL}(\pi|\pi^2)$ is lower than 0.1 bits. For these networks, the entropy rate $h(\pi^2)$ is about 99% of the maximal entropy rate $h(\pi)$. Conversely, for the last three networks, the divergence $D_{KL}(\pi|\pi^2)$ is always approximately 1 bit, and in fact the entropy rate $h(\pi^2)$ is around 96% of $h(\pi)$. As expected the divergence rapidly decreases as we include walks of higher length.

In order to perform a maximal-entropy Markov walk on a graph, at each time step, a walker needs a global knowledge of the whole network and has to compute $\mathbf{u}$, which has $O(K)$ computational complexity. However, global information is in practice always unavailable in real systems. As we have shown in this chapter, this global knowledge is not necessary since in many real-world networks long-range interactions are weak and can be neglected. It is therefore possible to construct almost maximal-entropy random walks with only local information on the graph structure. This can be done with $O(\langle k \rangle)$ complexity, a dramatic improvement which opens up to practical applications in social, biological and technological systems.

4.5 Flow graphs

Dynamical processes on a graph, in particular the unbiased random walk, have been used to propose flow-based metrics to characterize complex network properties. As

mentioned in the introduction to this chapter, unbiased random walk have been used to detect communities [27], to evaluate centrality of nodes [52, 59, 60], to coarse–grain graphs [55], just to make some examples. However, the unbiased random walk most of the times might not represent a good description for the process taking place on the graph under scrutiny. We propose here a mathematical framework which allows to analyze the structure of complex networks using wider class of dynamical processes. We introduce the concept of *flow graphs*, namely weighted networks where dynamical flows, like the one represented by random walks, are embedded into the link weights. Flow graphs provide an integrated representation of the structure and dynamics of the system, which can then been analyzed with standard tools from network theory. In other words, given a graph where a certain process, say a biased random walk, has a crucial role, one can define another graph with the same links, but different weights associated to them so that an unbiased random walk on the latter graph is equivalent to a biased random walk on the original graph

4.5.1 Unbiased random walk in weighted graphs

Let $\mathcal{G}$ be an undirected graph with N nodes and K links. In addition to this, each link (i, j) has associated a weight w_{ij}, which can be a real positive number. Then, the generalization of the adjacency matrix $\mathcal{A}$ for this graph is a matrix W whose entry (i, j) is equal to the weight w_{ij} of the link between i and j, and is zero if between i and j no link is present . We assume that a walker at node i chooses one of the nearest neighbors of i with a probability proportional to the weight of the corresponding edge. The transition probability from node i to its neighbor j is then:

$$\pi(j|i) = \frac{w_{ij}}{\sum_\ell w_{i\ell}} = \frac{w_{ij}}{s_i} \tag{4.29}$$

where $s_i = \sum_l w_{il}$ is the *strength* of node i. The stationary distribution is given in this case by:

$$p_i^* = \frac{s_i}{\sum_j s_j} \tag{4.30}$$

namely, the larger strength a node has, the more often it will be visited by a random walker.

4.5.2 Biased random walks and flow graphs

Consider on $\mathcal{G}$ a BRW as the one introduced in Sec. 4.1.2, defined by the transition matrix

$$\pi(j|i)_{BRW} = \frac{w_{ij} f_j}{\sum_\ell w_{i\ell} f_\ell}. \tag{4.31}$$

It is possible to interpret this process as an unbiased random walk on an opportunely defined graph $\mathcal{G}'$. In order to prove this, let us define a non negative symmetric matrix

W', whose entries $w'_{ij}$are

$$w'_{ij} = f_i w_{ij} f_j. \tag{4.32}$$

W' is the weighted adjacency matrix of the graph $\mathcal{G}'$, whose edges are the same as in $\mathcal{G}$ but with different weights. An unbiased random walk on $\mathcal{G}'$ is then characterized by the transition matrix

$$\pi'(j|i) = \frac{w'_{ij}}{\sum_\ell w'_{i\ell}}. \tag{4.33}$$

By substituting the expression for the entries w'_{ij}, we get $\pi'(j|i) = \frac{w_{ij} f_j}{\sum_\ell w_{i\ell} f_\ell}$, which yields $\pi'(j|i) = \pi(j|i)_{BRW}$. An unbiased random walk on $\mathcal{G}'$ is driven by the same transition matrix of a biased random walk on $\mathcal{G}$, which implies that the stationary distribution for both process is the same and is given by applying Eq. 4.30:

$$p^{*'}_j = \frac{s'_i}{\sum_j s'_j} = \frac{f_j \sum_i w_{ij} f_i}{\sum_{i\ell} f_i w_{i\ell} f_\ell}. \tag{4.34}$$

This result also shows that w'_{ij} is proportional to the flow of probability from j to i at equilibrium. Similar results can be proven also for other dynamical processes, such as continuos time random walks or consensus processes [4].

In general, the equivalence between trajectories of a biased random walker on $\mathcal{G}$ and those of an unbiased random walker on $\mathcal{G}'$ has important implications as it makes possible to use theoretical results known for unbiased random walks for the analysis of BRWs. An important context where for example this formalism proves useful is in community detection. In fact, many community detection methods are based on plain random walk processes. The notion of flow graph allows for detection of modules taking in consideration dynamical processes that are more adapt to the graph under scrutiny.

Chapter 5

Networks of motifs from sequences of symbols

Prediction is very difficult,
especially of the future.

Niels Bohr

There are many examples in biology, in linguistics and in the theory of dynamical systems, where information resides and has to be extracted from corpora of raw data consisting of sequences of symbols. For instance, a written text in English or in another language is a collection of sentences, each sentence being a sequence of the letters from a given alphabet. Not all sequences of letters are possible, since the sentences are organized on a lexicon of a certain number of words. In addition to this, different words are used together in a structured and conventional way [70–73]. Similarly, in biology, DNA nucleotides or aminoacid sequence data can be seen as corpora of strings [74–77]. For example, it is well known that proteomes are far from being a random assembly of peptides, since clustering of aminoacids [78] and strong correlations among proteomic segments [79] have been clearly demonstrated. These results give meaning to the metaphor of protein sequences regarded as texts written in a still unknown language [74, 80]. Sequences of symbols can also be found in time series generated by dynamical systems. In fact, a trajectory in the phase space can be transformed into sequence of symbols, by the so-called "symbolic dynamic" approach [81]. The basic idea is to partition phase space into a finite number of regions, each of which is labelled with a different symbol. In this way, each initial condition gives rise to a sequence of symbols representing the initial cell, the cell occupied at the first iterate, the cell occupied at the second iterate, and so forth.

In all the examples mentioned above, the main challenge is to decipher the message contained in the corpora of data sequences, and to infer the underlying rules that govern their production. In order to do this, one needs: *i)* to detect the fundamental units carrying information, like words do in language, and *ii)* to study their combination

syntax in the ensemble of sequences. In fact, information in its general meaning is located not only at the level of strings, but also in their correlation patterns [82, 83]. In this chapter, we introduce a method to transform a generic corpus of strings, such as written texts, protein sequence data, sheet music, a collection of dance movement sequences [84], into a network representing the significant and fundamental units of the original message together with their relationships. The method relies on a statistical procedure to detect patterns carrying relevant information, and works as follows. We first construct a dictionary of the recurrent strings of k letters, called k-motifs. Recurrent strings play, in this more general context, the same role as words in written or spoken languages. We then construct a k-motif network, a graph in which each node is one entry of the dictionary, and a directed arc between two nodes is drawn when the ordered co-occurence of the two motifs is statistically significant in the dataset analyzed. We will show how the analysis of topological properties of networks of k-motifs, such as the detection of community structures [13, 27], allows to extract important information encoded in the original data. In particular, we will consider the application of the method to datasets in three different domains, namely, biological sequences of proteins, messages from online social networks, and sequences of symbols generated by the trajectories of a dynamical system.

5.1 High-order Markov chains and motifs in ensembles of sequences

Let us consider an ensemble $\mathcal{S}$ of S sequences of symbols. Each sequence s ($s = 1, 2, \ldots, S$) is a string of letters from an alphabet $\mathcal{A}$ of A letters, $\mathcal{A} \equiv \{\sigma_1, \sigma_2, ..., \sigma_A\}$. In general, the strings can have different lengths. We indicate by l_s the length of sequence s, and by $L = \sum_{s=1}^{S} l_s$ the total length of the ensemble. An example is provided by proteomes. A proteome is a collection of $S \approx 10^4$ proteins of a species. Each protein is a sequence of length l_s, ranging from 10^2 to 10^3, made of symbols from an alphabet $\mathcal{A}$ with $A = 20$ letters, $\mathcal{A} \equiv \{\sigma_1, \sigma_2, ..., \sigma_{20}\}$, where each σ labels one of the aminoacids a protein can be made of. We define as *k-string* a segment of k contiguous letters $x_1 x_2 \ldots x_k$, where $x_i \in \mathcal{A}\ \forall i$. The number of all possible k-strings is A^k, while from the ensemble of sequences $\mathcal{S}$ we can select only $L - S \cdot (k-1)$ overlapping k-strings, so that some of the possible k-strings do not occurr, some of them occur once, others more than once, either in the same or in different sequences of symbols. We define as:

$$p^{obs}(x_1 x_2 \cdots x_k) = \frac{c(x_1 x_2 \cdots x_k)}{\sum_{(x_1, x_2, \cdots, x_k) \in \mathcal{A}^k} c(x_1 x_2 \cdots x_k)} \tag{5.1}$$

the *observed probability* of a string $x_1 x_2 \ldots x_k$. This probability is obtained by counting the total number of times, $c(x_1 x_2 \ldots x_k)$, the string actually occurs in the sequences of the ensemble. To assess for the statistical significance of the string, the probability in

Eq. 5.1 has to be compared with the *expected probability* $p^{exp}(x_1x_2\cdots x_k)$ of the string occurrence. The latter can be evaluated under different assumptions. In fact, the joint probability $p(x_1x_2\cdots x_k)$ can be written as:

$$p(x_1x_2\cdots x_k) = p(x_1x_2\cdots x_{k-1})p(x_k|x_1x_2\cdots x_{k-1}),$$

and different approximations for the conditional probabilities $p(x_k|x_1x_2\cdots x_{k-1})$ lead to different values of the expected probability $p^{exp}(x_1x_2\cdots x_k)$. Namely, if we assume that the occurrence of a letter does not depend on any of the previous letters, i.e. $p(x_k|x_1x_2\cdots x_{k-1}) = p(x_k)$, the expected probability is simply given by the product of the relative frequencies of the string's component letters: $p^{exp}(x_1x_2\cdots x_k) = p^{obs}(x_1)\cdots p^{obs}(x_k)$ [85, 86]. By using instead a first order Markov approximation, i.e. $p(x_k|x_1x_2\cdots x_{k-1}) = p(x_k|x_{k-1})$, the expected probability can be expressed in the form: $p^{exp}(x_1x_2\cdots x_k) = p^{obs}(x_1)p^{obs}(x_2|x_1)\cdots p^{obs}(x_k|x_{k-1})$, where $p^{obs}(x_j|x_i)$ is extracted from the countings as: $p^{obs}(x_j|x_i) = c(x_ix_j)/\sum_{x_j} c(x_ix_j) = p^{obs}(x_ix_j)/p^{obs}(x_i)$. This latter assumption is based on the fact that there is a minimal amount of memory in the sequence: a symbol of the sequence is correlated to the previous one only. Here, we go beyond the approximation of Markov chains of order 1, by retaining as much memory as possible [75]. We assume:

$$\begin{aligned} p^{exp}(x_1x_2\cdots x_k) = p^{obs}(x_1x_2\cdots x_{k-1})\cdot \\ \cdot p^{obs}(x_k|x_2\cdots x_{k-1}) \end{aligned} \tag{5.2}$$

where the conditional probabilities can be evaluated from countings as:

$$p^{obs}(x_k|x_2\cdots x_{k-1}) = \frac{c(x_2x_3\cdots x_k)}{\sum_{x_k} c(x_2x_3\cdots x_k)} \tag{5.3}$$

or can be expressed in terms of the observed probability for shorter sequences as:

$$p^{obs}(x_k|x_2\cdots x_{k-1}) = \frac{p^{obs}(x_2\cdots x_k)}{p^{obs}(x_2\cdots x_{k-1})} \tag{5.4}$$

By using the latter expression, we can finally write the expected probabilities in a more compact form:

$$\begin{aligned} p^{exp}(x_1) &= p^{obs}(x_1) \\ p^{exp}(x_1x_2) &= p^{obs}(x_1x_2) \\ p^{exp}(x_1x_2x_3) &= p^{obs}(x_1x_2)\frac{p^{obs}(x_2x_3)}{p^{obs}(x_2)} \\ \ldots &= \ldots. \\ p^{exp}(x_1x_2\cdots x_k) &= p^{obs}(x_1\cdots x_{k-1})\cdot\frac{p^{obs}(x_2\cdots x_k)}{p^{obs}(x_2\cdots x_{k-1})} \end{aligned} \tag{5.5}$$

This way, the expected probability of a given k-string is evaluated based on observations for strings of up to $(k-1)$ symbols. Therefore, by predicting the probability of appearance with a high order Markov model, our method allows to highlight the true k-body correlations subtracting from them the effects due to $(k-1)$ and lower order correlations. Based on observed and expected probabilities, a test of statistical significance, for instance a Z-score[1], is then performed for each k-string. We define *k-motifs* or *recurrent k-strings*, the statistically-relevant strings whose observed and expected number of occurrences are such as to validate the statistical test adopted, and we indicate as $\mathcal{Z}_k$ the dictionary composed by all the selected k-motifs [2].

5.2 Networks of motifs

Once we have constructed a lexicon of fundamental units, the next goal is to represent in a graph the way they are combined together. Recurrent k-strings can be distributed differently along the sequences: they can appear in a single sequence or in more than one sequence, alone or in clusters. To extract the non trivial patterns of correlated appearance of k-motifs, we need to evaluate the probability for the random co-occurrence of two motifs, when these are uncorrelated. We estimate first the expected probability that motif X is followed by motif Y within a generic sequence of the ensemble $\mathcal{S}$, then we sum over all the sequences of $\mathcal{S}$. We denote as $p(X)$ and $p(Y)$ the probabilities of finding the two motifs in $\mathcal{S}$. In sequence s, motif X can occupy positions ranging from the first to the (l_s-2k)th site, where l_s is the length of s, and k is the length of the motif. We have assumed that the two motifs cannot overlap. For each fixed position i of X on s, with $i=1,...,(l_s-2k)$, there are $(l_s-2k+1-i)$ possibilities for Y to appear in the sequence. Hence, the number of expected co-occurences of X and Y within s is given by: $\sum_{i=1}^{l_s-2k}(l_s-2k+1-i)p(X)p(Y)$. In order to obtain the expected number of co-occurrences, we have to sum over all the sequence in the ensemble $\mathcal{S}$. We finally get:

$$N^{exp}(Y|X) = p(X)p(Y)\sum_{s=1}^{S}\sum_{i=1}^{l_s-2k}(l_s-2k+1-i) = \frac{1}{2}p(X)p(Y)\sum_{s=1}^{S}(l_s-2k+1)(l_s-2k+2) \tag{5.6}$$

For each value of k, we are now able to construct the *k-motif network* of the ensemble $\mathcal{S}$, i.e. a directed network whose nodes are motifs in the dictionary $\mathcal{Z}_k$, and an arc point from node X to node Y if the number of times Y follows X in the ensemble of sequences is statistically significant. Furthermore, a weight can be associated to the

[1] The Z-score indicates how many standard deviations an observation or datum is above or below the mean. It is a dimensionless quantity derived by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation.

[2] The term *motif* is chosen in analogy with the concept of network *motifs*, i.e. recurrent patterns of nodes and links in a graph [87]

arc from X to Y, based on the extent to which the co-occurrence of the two motifs deviates from expectation.

This approach is able to represent the correlation patterns encrypted in the ensemble of sequences into a single object, the k-motif network. Then, graph theory allows to extract information from the structural properties of the network, and to retrieve the main message encoded in the original sequences. In particular, it is interesting to study the components of the k-motif network or, if the graph is connected, its community structures, i.e. those groups of nodes tightly connected among themselves and weakly linked to the rest of the graph [27].

5.3 Applications

In the following, we will consider the application of the method to three different datasets, belonging to three contexts as diverse as biology, social dialogs and dynamical systems. We will show how the community analysis of the related k-motif networks enables to extract functional domains in proteomes, social cascades and hot topics in Twitter, and the increase of chaoticity in deterministic maps.

5.3.1 Biological sequences

Methods to study over- or under-representation of particular motifs in a complete genome [75, 88, 89] or in a proteome [90], have already been proposed, and the results have been used to make functional deductions. Although the information contained in strings deviating from expectancy is useful for the analysis of many biological mechanisms [86], it turns out to be not sufficient for a complete and exhaustive interpretation of the genomic and proteomic message. A fundamental key to its comprehension is in fact hidden in the correlations among recurrent patterns of strings. The spatial structure of proteins provides an example: when a protein folds, segments distant on the sequence come to be close to each others in the space. This can happen because two (or more) segments need to physically interact in order to perform the biological function the protein is supposed to go through. Such a mechanism translates into a statistical correlation between short motifs of aminoacids, which is well captured by an analysis in terms of k-motif networks.

Human proteome

In our application, we have considered the ensemble of sequences relative to the human proteome [3]. It consists of 34180 aminoacidic sequences of variable size, with an average length of 481 letters. For this dataset, we have computed the probabilities p^{obs} and p^{exp} for each of the $20^3 = 8000$ possible strings of three aminoacids, and we have selected

[3] Data downloaded from [91]

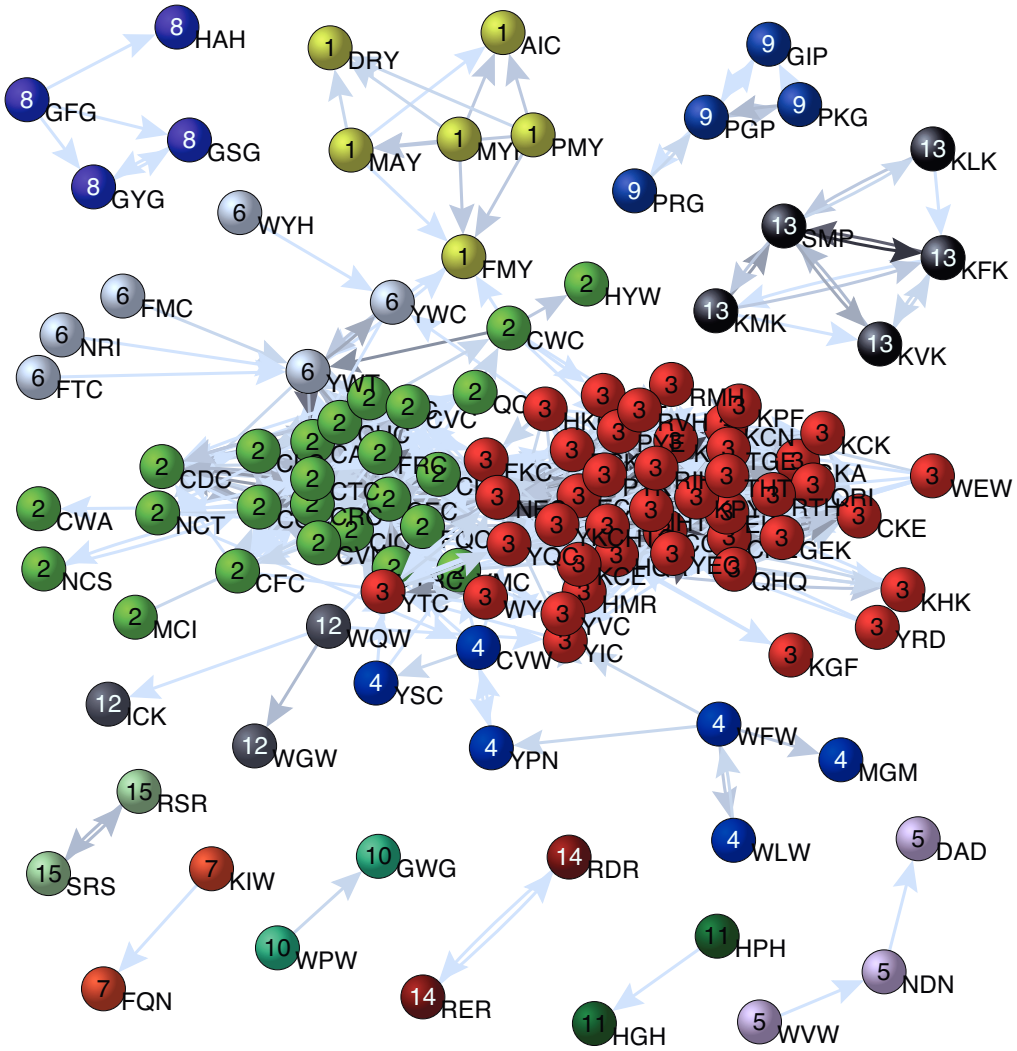

Figure 5.1: The 3-motifs network of the human proteome. Nodes belonging to the same community are labeled by the same number and share the same colour. Most of the communities can be associated to a functional domain as described in table 5.1.

as 3-motifs the strings satisfying $\frac{p^{obs}}{p^{exp}} > \left\langle \frac{p^{obs}}{p^{exp}} \right\rangle + 2\sigma$, hence creating the dictionary Z_3 [4]. The entries of the dictionary are the nodes of the 3-motif network. The node X is then linked to Y with a directed arc if the number of times that motif Y follows motif X within the same protein is statistically significant, according to the relation: $\frac{p^{obs}(Y|X)}{p^{exp}(Y|X)} > \left\langle \frac{p^{obs}(Y|X)}{p^{exp}(Y|X)} \right\rangle + 2\sigma$. The statistical significance $\frac{p^{obs}(Y|X)}{p^{exp}(Y|X)}$ is also the weight

[4] With the notation $\langle p\,(x) \rangle$, we denote the average of $p(x)$ over all the possible configurations of x and with σ the standard deviation of the distribution

Table 5.1: List of communities in the 3-motif network of the human proteome. Community labels as in Fig. 5.1, number of nodes, total internal weight, associated domain, and the domain specificity are reported.

	# nodes	Internal weight	Domain	Domain recognition
1	6	83,30%	Olfactory receptor	171/175
2	25	74,91%	—	
3	43	94,13%	Zinc Finger	1345/1364
4	6	55,42%	G-protein and CUB-Sushi	9/11
5	3	100%	Cadherin	330/347
6	4	100%	Lipoproteins	16/19
7	2	100%	Homeobox	65/84
8	4	100%	—	
9	4	100%	Collagen	271/482
10	2	100%	Serine protease	22/51
11	2	100%	—	
12	3	60,30%	C-type proteins	3/4
13	5	100%	—	
14	2	100%	—	
15	2	100%	—	

of the arc. In this way we obtain the 3-motif graph of 199 nodes and 1302 directed links, shown in Fig. 5.1. The graph has 86 isolated nodes (not displayed in Figure), while the remaining 113 nodes are organized into 10 weak components. The largest component of the graph contains 5 clusters, detected by means of the MCl algorithm [28]. Therefore, 15 different communities are present in the graph. In Table 5.1 we report, for each community, the number of nodes and its total internal weight, defined as the sum of the weights of links between nodes of the communities normalized by the sum of the weights of links incident in nodes of the community. By submitting a query to the Prosite database [92] we have obtained, for each couple of connected motifs belonging to the same community, the list of all proteins, classified by domain, where the two motifs co-occur. The results show that linked couples of motifs belonging to the same community, all co-occur in the same kind of domains. In addition to this, one can associate 9 of these 15 communities just to one protein domain, since the majority of co-occurrences emerge in proteins matching a well-defined function. In Table 5.1 we

report, when possible, the association to a single protein domain, together with the ratio between the number of times the couple of motifs with the highest weight occurred in that specific domain, and the total number of co–occurrences in the database.

Analogous results were also found for the 4-motif graph, while it is not possible to derive the same kind of information by using lower order Markov models to construct dictionaries. For example, the 3-motif network constructed with a dictionary based on a lower order approximation rather than on a 2-bodies Markov chain, exhibits a community structure with just four communities, none of which could be identified with a functional protein domain.

5.3.2 Social networks and microblogging

By means of k-motif networks, information can also be retrieved from datasets of social dialogs and microblogging websites. Although in these cases, in principle, a dictionary is a-priori known, not all terms used in the Internet language are always listed in a dictionary: abbreviations, puns, leet language words [93], names of websites or names of public figures, are just some examples. Moreover, some expressions or combinations of terms appear more frequently in some periods or contexts due to the interest to some hot topics. In addition to this, the method of k-motif networks turns to be very useful in all those contexts where it is necessary to process and compact information from large amount of symbolic data. This is the case of Internet, where the amount of text data provided by blogs, dialogs in social networks, forums, etc. is growing and growing.

In the following, we provide details on how network of motifs are able to deduce information about hot topics and cascades [94, 95] in a dataset extracted from Twitter, a well-know platform for social networking and microblogging.

Twitter and the case of the 2010 UK election

Twitter [96] is a social networking and microblogging service which allows users to send short messages known as *tweets*. Tweets are composed only of text, with a strict limit of 140 characters: they are displayed on the author's profile page and delivered to the author's subscribers, who are also known as "followers". The dataset we have analyzed is a collection of 28143 tweets, crawled on two days, from the 23rd to 24th April 2010, and selected through the Twitter Streaming API [97] if they contained the string *#leadersdebate*. The choice of such a keyword, called in Twitter also *hashtag*, was aimed to select all those tweets concerning electoral campaign in UK, where general election to elect the members of the House of Commons would have taken place two weeks later. We have analyzed the dataset removing all blank spaces between words and all symbols that where not numbers or letters (punctuation, symbols like $, @, *, etc.) and not distinguishing between lower- and upper-case letters. From these sequences, dictionaries of motifs $\mathcal{Z}_3$ and $\mathcal{Z}_4$ have been extracted, selecting respectively

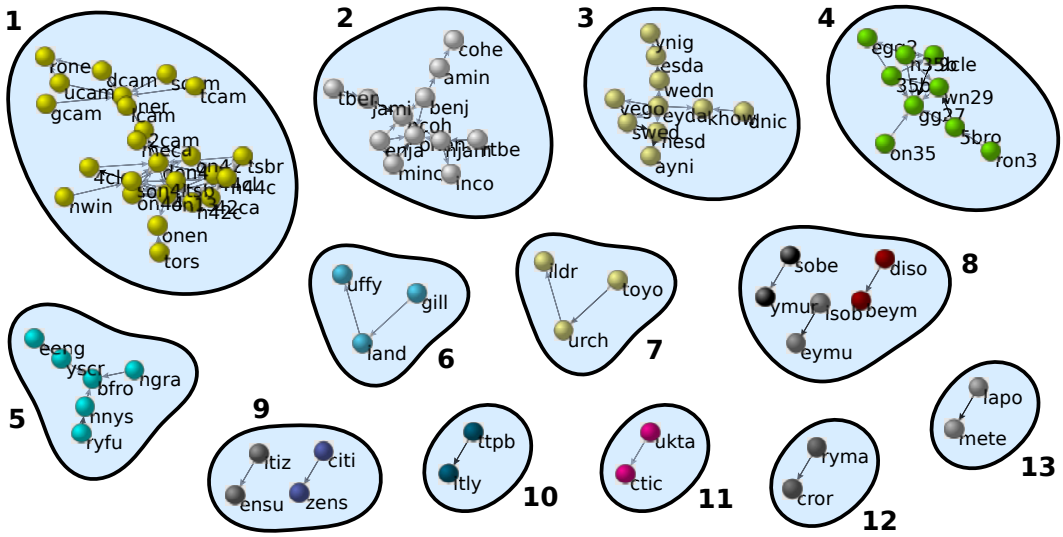

Figure 5.2: Components of the 4-motifs network of the twitter dataset. Each component and its associated topic are described in table 5.3.

the 10% and 1% of most significant strings of 3 and 4 letters. As described in the main text, we have constructed networks whose nodes represent the entries of a dictionary, and an arc is drawn from the node representing string X to the node standing for string Y, if $p^{obs}(Y|X)/p^{exp}(Y|X)$ is greater than a certain threshold. In Fig. 5.2, we show the 4-motifs network when the threshold is set equal to 400 (isolated nodes not reported). Such a high threshold is chosen to have a small network that can be easily visualized and studied. More information can be obtained by setting the threshold to lower values or analyzing networks made up of motifs of different length k. Searching in the original dataset the connected motifs, it is possible to associate each component to a particular tweet which generated a cascade or with a specific expression, related to a specific hot topic discussed by users of the microblogging platform. For all components of Fig. 5.2, we report in Table 5.3 the tweet or expression associated and its meaning. For example, component 1 and 4 can be associated to two exit polls disclosed on those days by two different journals, or component 6 to the name "Gillian Duffy", a 65-years old pensioner involved in a political scandal with British PM Gordon Brown during the election tour (Brown's remarks of her as a "bigoted woman" were accidentally recorded and broadcast).

5.4 Symbolic dynamics

Symbolic dynamics is a general method to transform trajectories of dynamical systems into sequences of symbols. The distinct feature in symbolic dynamics is that time is measured in discrete intervals. So at each time interval the system is in a particular

Table 5.2: The first ten most significant links between motifs, belonging to 7 different communities in the Twitter dataset. Each community corresponds to a specific tweet or expression that generated a topic cascade.

motif 1	motif 2	$\frac{p^{obs}}{p^{exp}}$	Expression or Tweet	Topic
9cle 5bro	gg27 wn29	955.3 894.8	*GUARDIAN ICM POLL Cameron 35% Brown 29% Clegg 27%*	poll results from various websites, journals, tv channels, etc
son4 don4	4cle 2cam	924.3 881.7	*Brown wins on 44%, Clegg is second on 42%, Cameron 13% None of them 1%*	
lapo	mete	892.3	www.slapometer.com	A funny website on the election
swed nesd	nesd ayni	864.7 826.1	*hey Dave, Gordon and Nick : how about a 4th debate on Channel 4 this wednesday night without the rules?!*	Proposal for a 4th debate among leaders, made by a journalist on his twitter page
jami minc	ncoh ohen	842.0 764.9	Benjamin Cohen	Journalist of Channel 4 News
isob	eymu	831.4	#disobeymurdoch	hashtag

state. Each state is associated with a symbol and the evolution of the system is then described by a sequence of symbols. The method turns to be very useful in all those cases where system states and time are inherently discrete. In case the time scale of the system or its states are not discrete, one has to set a coarse-grained description of the system. Different initial conditions usually generate different trajectories in the phase space, which map onto different sequences of symbols. A large number of initial conditions produces an ensemble of sequences whose analysis can be addressed with the method based on networks of motifs.

In the following, we will describe the application of the method to the standard map, and we will show how the related networks of motifs shape according to its chaotic behavior.

Standard Map

The standard map, also known as Chirikov map, is a bidimensional area-preserving chaotic map. It maps a square with side 2π onto itself [102]. It is described by the

Table 5.3: In relation to Fig. 5.2, we report the number of nodes, links, the tweet or the expression containing the motifs and the topic associated to each of the 13 communities

Comm.	Nodes	Links	Expression or Tweet	Topic
1	25	33	*Brown wins on 44%, Clegg is second on 42%, Cameron 13% None of them 1%*	poll results from various websites, journals, tv channels, etc
2	12	14	Benjamin Cohen	Journalist of Channel 4 News [98]
3	10	11	*hey Dave, Gordon and Nick : how about a 4th debate on Channel 4 this wednesday night without the rules?!*	Proposal for a 4th debate among leaders, made by a journalist on his Twitter page
4	9	13	*GUARDIAN ICM POLL Cameron 35% Brown 29% Clegg 27%*	poll results from various websites, journals, tv channels, etc
5	6	5	*Very funny screengrab from the LeadersDebate*	About a funny picture of the leaders debate on BBC [99]
6	3	2	Gillian Duffy	Woman branded a 'bigot' by Gordon Brown in general election campaign [100]
7	6	5	*Cameron: I believe that if you've inherited hard all your life you should pass it on to your children*	Electoral campaign from David Cameron
8	6	3	#disobeymurdoch	Twitter hashtag
9	4	2	#citizensuk	Twitter hashtag
10	2	1	http:// ... .ly	Format of shortened weblinks in twitter
11	2	1	Tactical voting	*Strategy that when a voter misrepresents his or her sincere preferences in order to gain a more favorable outcome [101]*
12	2	1	Henry Macrory	Head of press for the Conservatives, owner of a twitter account
13	2	1	www.slapometer.com	A funny website on the election

equations:

$$\begin{cases} x_{t+1} = p_t + a \sin x_t & \mod 2\pi \\ p_{t+1} = p_t + x_{t+1} & \mod 2\pi \end{cases} \tag{5.7}$$

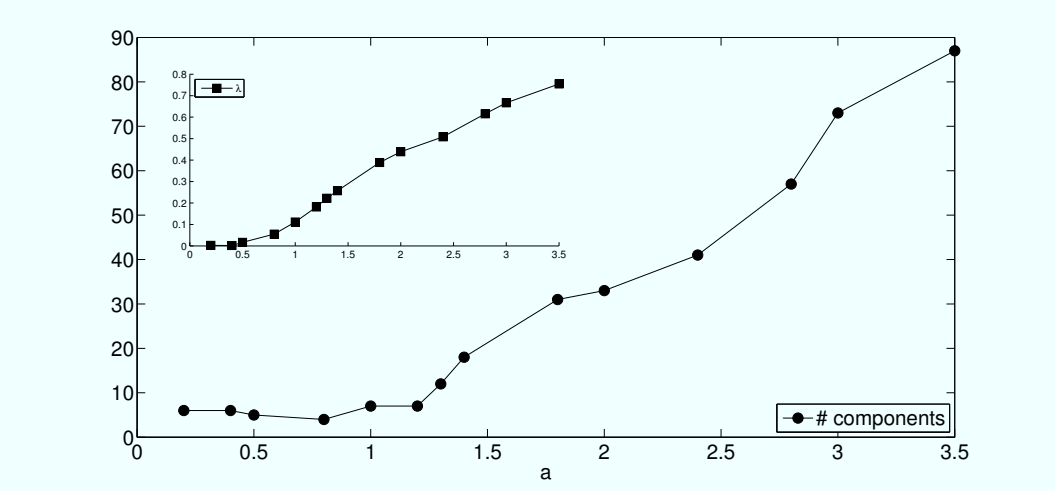

Figure 5.3: Standard map: number of components in the 3-motifs networks (main figure), and the Lyapunov exponent (inset), as a function of the non-linearity parameter a.

where t represents time iteration and a is a parameter assuming real values. The map is increasingly chaotic as a increases (see inset of Fig. 5.3 to see a plot of the Lyapunov exponent as a function of the parameter a). For $a = 0$, the map is linear and only periodic and quasiperiodic orbits are allowed. When evolution of trajectories are plotted in the phase space (the xp plane), periodic orbits appear as closed curves, and quasiperiodic orbits as necklaces of closed curves whose centers lie in another larger closed curve. Which type of orbit is observed depends on the map's initial conditions. When the nonlinearity of the map increases, for appropriate initial conditions it is possible to observe chaotic dynamics.

In order to obtain sequences from the standard map (5.7) by means of the symbolic dynamic approach [103], one needs to make a coarse graining of the phase space, defining a discrete and finite number of possible states the trajectory can occupy. This way it is possible to associate a symbol to each of the possible states and derive a sequence according to the trajectory originating from an initial condition. We have coarse-grained the phase space into 25 (5×5) squares of equal size and we have derived for different values of the parameter a, 10^4 sequences of 10^3 symbols. In other words, this means to follow for 10^3 time steps the trajectories originating from 10^4 different initial conditions.

The idea is that closed orbits or quasi periodic-ones correspond to correlations between motifs and therefore in links of the graph of motifs. When the map becomes more and more chaotic, closed orbits disappear and, correspondingly, the networks break in many components, see Fig. 5.4. In the extreme limit of map highly chaotic ($a > 3$), the network of motifs are completly disconnected, with all nodes isolated. Nevertheless, this scenario is different from the one generated by stochastic sequences,

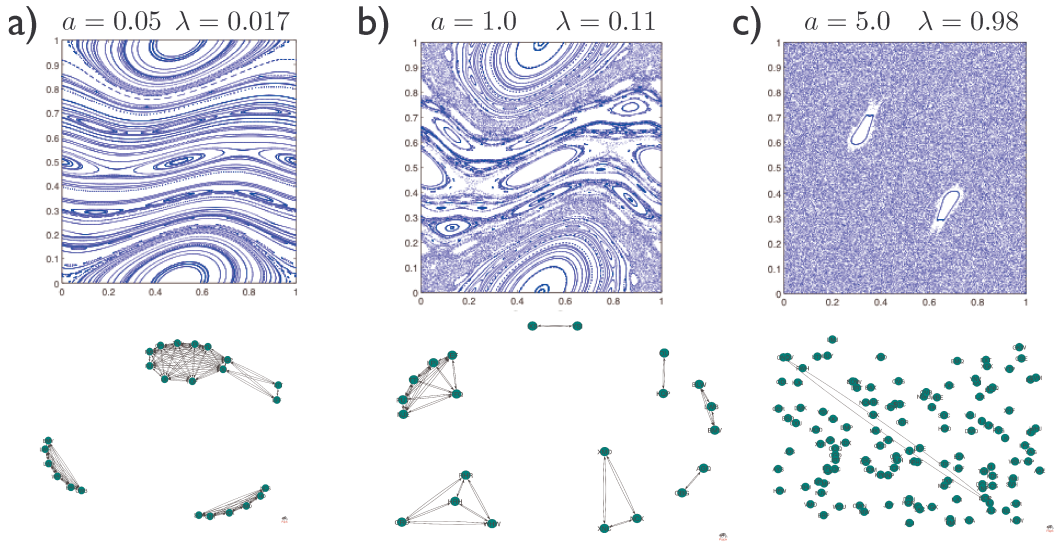

Figure 5.4: Phase space of the standard map (top) and relative network of motifs (bottom) corresponding to a) $a = 0.5$, b) $a = 1.0$, and c) $a = 5.0$, where a is the map parameter of Eqs. 5.7. Patterns of connectivity in the graphs correspond to tori of the phase space. When the map becomes more and more chaotic, as indicated by the value of the Lyapunov exponent λ, the closed orbits (tori) disappera and, correspondingly, the networks break in many components. However, even when the map is very chaotic, a dictionary - corresponding to the nodes in the graph - is still present, meaning that some short-range correlations are preserved.

since in this case motifs would not be detected, while this still happens in the chaotic map, although only for small values of k. This result is well depicted in Fig. 5.3, where the number of components of the 3-motif graphs is plotted as a function of the value a of the map generating the ensemble. This curve is shown to have the same behavior of the Lyapunov exponent, as reported in the inset of the same figure.

5.5 Conclusion and perspectives

We have introduced a general method to construct networks out of any symbolic sequential data. The method is based on two different steps: first it extracts in a "natural" way motifs, i.e. those recurrent short strings which play the same role words do in language; then it represents correlations of motifs within sequences as a network. Important information from the original data are embedded in such a network and can be easily retrieved as shown with different applications (a biological system, a social dialog and a dynamical system). With respect to previous linguistic methods, our approach does not need the a priori knowledge of a given dictionary, and also allows to compare different ensembles, corresponding, for example, to different values of control parameters in dynamical systems. All this makes the method very general and

opens up a wide range of applications from the study of written text, to the analysis of sheet music or sequences of dance movements. Moreover, the method does not use parameters on the position of motifs in order to correlate them, since co-occurrences are computed within sequences, which represent natural interruptions of a corpora of data (proteins in a proteome, posts in a blog, different initial conditions in a symbolic dynamics, etc.).

Chapter 6

A high-order Markov model for the study of mobility

Simplicity is the ultimate sophistication.

Leonardo Da Vinci

6.1 Studying human mobility

Understanding the statistical patterns of human mobility, predicting trajectories and uncovering the mechanisms behind human movements [104] is a considerable challenge with important practical applications to traffic management [105, 106], planning of urban spaces [107, 108], epidemics [109–112], information spreading [113, 114], and geo-marketing [115, 116]. In the last years, advanced digital technologies have provided huge amounts of data on human activities, allowing to extract information on human movements. For instance, observations of banknote circulation [117, 118], mobile phone records [119], online location-based social networks [120, 121], GPS location data of vehicles [122], or radio frequency identification traces [104, 108, 123], have all been used as proxies for human movements. These studies have provided valuable insights into several aspects of human mobility, uncovering distinct features of human travel behavior such as scaling laws [117, 124], predictability of trajectories [125], and impact of motion on disease spreading [110–112, 126]. However, from a comparative analysis of the different works it emerges clearly that a "unified theory" of human mobility is still outstanding, since results, even on some very basic features of the motion, often appear to be contrasting [104]. One example is the measured distribution of human trip lengths in various types of transportation: some studies agree that mobility is generally characterized by fat-tailed distributions of trip lengths [117, 124], while others

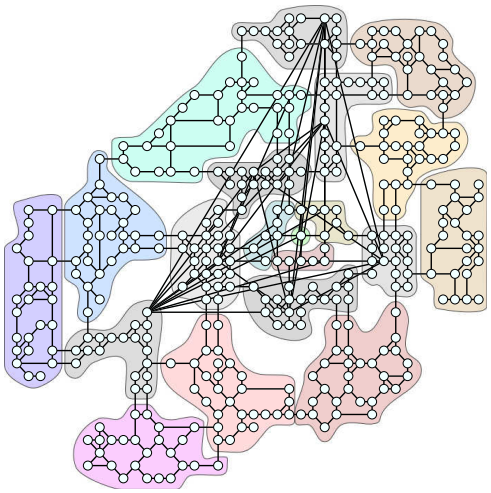

Figure 6.1: **The universe map of the massive multiplayer online game *Pardus*.** The universe of Pardus can be represented as a network with $N = 400$ nodes, called sectors (playing the role of cities), and $K = 1160$ links. Sectors are organized into 20 different regions, called clusters, shown in the figure as different color-shaded areas. There is no explicit set of goals in the game. Players are free to interact in a number of ways to e.g. increase their virtual wealth or status. Players move between sectors to interact with other players, e.g. to trade, attack, wage war, or to explore the virtual world.

report exponential or binomial forms [104, 108, 122]. The discrepancies arise due to the different mobility data sets used, where mobility is indirectly inferred from some specific human activity in a particular context. For instance, mobile phone records typically provide location information only when a person uses the phone [124], while radio frequency identification traces like the ones of Oyster cards in the London subway [108] only log movements based on public transportation systems. Analyses of these data sets can then result in a possibly biased view of the underlying mobility processes. Furthermore, most of the analyzed data sets have poor information on how socio-economic factors influence human mobility patterns. More generally, the lack of an all-encompassing record set with positional raw data including complete information on the socio-economic context and on the behavior of all members of a human society, has so far limited the possibilities for a comprehensive exploration of human mobility.

6.2 A new approach to the study of mobility

In this chapter and in [5], we address the issue of mobility from a novel point of view by analyzing, with unprecedented precision, the movements of a large number of individuals, the players of a self-developed massive multiplayer online game (MMOG). Such online platforms provide a fascinating new way of observing hundreds of thousands of interacting individuals who are simultaneously engaged in social and economic activities. The potential of online worlds as large-scale "socio-economic laboratories" has

been demonstrated in a number of previous studies [48, 127–129]. For the MMOG at hand [130], we have access to practically all actions [7], including movements, accumulated over several years. This MMOG can therefore be considered as a "socio-economic petri dish" to study mobility in a completely controlled way. We can in fact observe the long-time evolution of a social system at the scale of an entire human society, having a perfect knowledge of all the spatio-temporal and socio-economic details. In contrast to traditional studies in social science which are typically biased by well-known "interviewer-effects", in MMOGs the socio-economic measurements are objective and unobtrusive, since subjects are not consciously aware of being observed.

Using positional data of the players in the game universe [1], in combination with other socio-economic information from the game, we uncover various fundamental features of mobility, and we provide a complete description of the mechanisms causing the observed anomalous diffusion. Two are the main results we will show in this chapter. First, we find the emergence of different spatial scales, due to the strong tendency of the players to limit their economic activities to some specific areas over long time periods and to avoid crossing the borders between different areas. Making use of this observation, we propose an efficient method to identify socio-economic regions by means of community detection algorithms based solely on the measured movement dynamics. Our second result unveils the driving mechanism behind the movement patterns of players: Locations are visited in a specific order, leading to strong long-term memory effects which are essential to understand and reproduce the observed trajectories. Finally, we provide large-scale evidence that neglecting either of these spatial or temporal constraints may obstruct the possibility of understanding the processes behind human mobility.

6.3 A social arena: the online game Pardus

Pardus is a massive multiplayer online game running since 2004, with a worldwide player base of more than 350,000 individuals. It is an open-ended game whose players live in a virtual, futuristic universe and interact with each other in a multitude of ways. The topology of the universe can be represented as a network with 400 nodes, called *sectors*, embedded in a two-dimensional space, the so-called *universe map* shown in Fig. 6.1. Each sector is like a city where players can have social relations (establish new friendships, make enemies and wage wars), and entertain economic activities (trade and production of commodities). Typically, sectors adjacent on the universe map, as well as a few far-apart sectors, are interconnected by links which allow players to move from sector to sector. At any point in time, each sector is typically attended by a large

[1] Whenever we address the position or the movement of 'a player' in the game universe, this is meant as a short form for referring to the virtual avatar which is uniquely associated to and controlled by the player. This abbreviation is consistent with the tendency of players to identify themselves with their avatars.

N	400
K	1160
$\bar{k}$	2.9
C	0.089
C/C_{r}	12.33
L	11.89
L/L_{r}	2.11
E_{loc}	0.80
E_{glob}	0.03
D	27
D_{eff}	18

Table 6.1: Network properties of the Pardus universe: number of nodes N, number of (undirected) links K, average degree $\bar{k}$, clustering coefficient C, clustering coefficient to corresponding coefficient of random graph C/C_{r}, average geodesic L, average geodesic to corresponding average geodesic of random graph L/L_{r}, local efficiency E_{loc}, global efficiency E_{glob}, diameter D, effective diameter (0.9-quantile) D_{eff}.

.

number of players. The network is sparse and, similarly to other spatial networks, is not a small world. It has a characteristic path length $L = 11.89$ and a diameter $d_{\max} = 27$, which means that, on average, players have to move through a non-negligible number of sectors to traverse the universe. See table 6.1 for a detailed characterization of the universe network structure.

The sectors have been originally organized by the developers of the game into 20 different *clusters*, which are perceived by the players as different political or socio-economic regions such as countries. For example, a player who is member of a political faction in the game is provided some game-relevant protection in all clusters which are controlled by the faction, and has the opportunity of social promotion when accomplishing certain tasks within these clusters. Each cluster is shown in Fig. 6.1 with a different background color. All clusters contain about 20 sectors each, with the exception of the central cluster, consisting of just one sector, and its surrounding three clusters having only 6-7 sectors. Sectors belonging to the same cluster are geographically close on the map, meaning that the distance between any two sectors in the same cluster is small, with an average distance around 3. Players typically have a "home cluster" where they focus their socio-economic activities over long time periods. Occasionally, they also move to sectors belonging to other clusters in order to explore the universe, to relocate their home (migrate), or during extreme game events such as wars.

In Pardus, players are free to pursue whichever role they like to take. Many of them focus on expanding their social relations or political influence, some play the role of "scientists" exploring the universe, while others choose their main goal in trade and optimizing the amount of virtual money earned [48]. The large variety of complex socio-

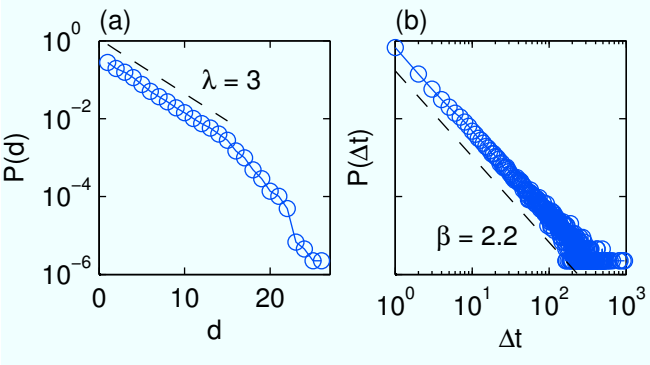

Figure 6.2: Distribution of jump distances and of waiting times. To each player a time series consisting of the sector positions over 1000 days is associated. A *jump* is said to occur when the sector position in the time series changes from one day to the following. The length d of a jump is measured in terms of graph distance and can take an integer value between 1 and $d_{\max} = 27$, the diameter of the network. (a) The probability distribution of jump distances is reported in a semi-log plot. For $d \leq 15$, the distribution follows an exponential $P(d) \sim e^{-\frac{d}{\lambda}}$ with a characteristic length $\lambda \approx 3$. Players can also remain in the same sector for more days, without moving to other sectors. We define as waiting time Δt the number of consecutive days a player spends in only one sector. (b) We show the probability distribution of waiting times Δt in a log-log plot, which is well fitted by a power-law $P(\Delta t) \sim \Delta t^{-\beta}$, with $\beta \approx 2.2$.

economic behaviors emerging in this online society, results in high heterogeneity in the mobility patterns, such as observed in real human motion. However, differently from other empirical studies on human movements, mobility in Pardus can be investigated in a controlled way, since complete information on actions of players is available [48, 127]. Here we consider a data set consisting of movements in the network universe of all players who were active over a period of 1,000 days, as well as of socio-economic information about their environment. This opens the possibilty of investigating motion in relation to other social and economic factors. Note that we do not have to address the common issues of relying on incomplete data, on data that are only a proxy of mobility, or on data that are aggregates of different types of transportation [112].

6.4 Basic features of the motion

The position of each player in the universe, namely the ID number of the sector where the player is currently situated, is logged once a day. In this way the motion of each player becomes a time series of 1,000 sector positions. A *jump* occurs when a player's sector position changes from one day to the following. The associated length d of a jump is measured in terms of graph distance, an integer value between 1 and $d_{\max} = 27$. The probability distribution of jump distances, computed for all players over the whole observation period, is reported in Figure 6.2 (a). For $d \leq 15$, the distribution is

well-fitted by an exponential:

$$P(d) \sim e^{-\frac{d}{\lambda}}, \tag{6.1}$$

with a characteristic jump length $\lambda \approx 3$. The existence of a typical travel distance, as also recently found in other mobility data [108, 122], is related to the use of a single transportation mode in *Pardus* [131]. This allows to disentangle the intrinsic heterogeneity of the players from the effects due to the presence of different means of transportation [112], which might be the cause of the scale-free distributions found in mobile phone or other mobility data sets [117, 119]. It has in fact been suggested that power laws in distance distributions of movement data may emerge from the coexistence of different scales [104, 132].

In some cases, players stay in the same sector for a number of consecutive days. For instance, 11 of the 1458 considered players, although being active in the game, never jump within the entire observation period. On average, a player does not change sector in approximately 75% of the days. To better characterize the motion, we computed the waiting times Δt (measured in terms of number of days) between all pairs of consecutive jumps, over all players. The distribution of these waiting times, shown in Fig. 6.2 (b) follows a power-law distribution:

$$P(\Delta t) \sim \Delta t^{-\beta} \tag{6.2}$$

with an exponent $\beta \approx 2.2$, in agreement with other recent measurements on human dynamics [133]. In addition, we found that the average waiting times of individual players are distributed as a power-law. This implies a strong heterogeneity in the motion of different players, which is related to the heterogeneity in their general activity.

6.5 Mobility reveals socio-economic clusters

Mobility patterns are influenced by the presence of the socio-economic regions in the network, highlighted in colors in Fig. 6.1. The typical situation is illustrated in Fig. 6.3 (a), with jumps within the same cluster being preferred to jumps between sectors in different clusters. In order to quantify this effect, we report in Fig. 6.3 (b), blue circles, the observed number of jumps of length d within the same cluster, divided by the total number of jumps of length d. This ratio is a decreasing function of the distance d, and reaches zero at $d = 12$, since no sectors at such distance do belong to the same cluster. As a null model we report the fraction of sector pairs at distance d which belong to the same cluster, see red squares in the same figure. The significant discrepancy between the two curves indicates that players indeed tend to avoid crossing the borders between clusters. For example, a jump of length $d = 8$ from one sector to another sector in the same cluster is expected only in 3% of the cases, while it is observed in about 20% of the cases.

Now, the propensity of a player to spend long time periods within the same cluster

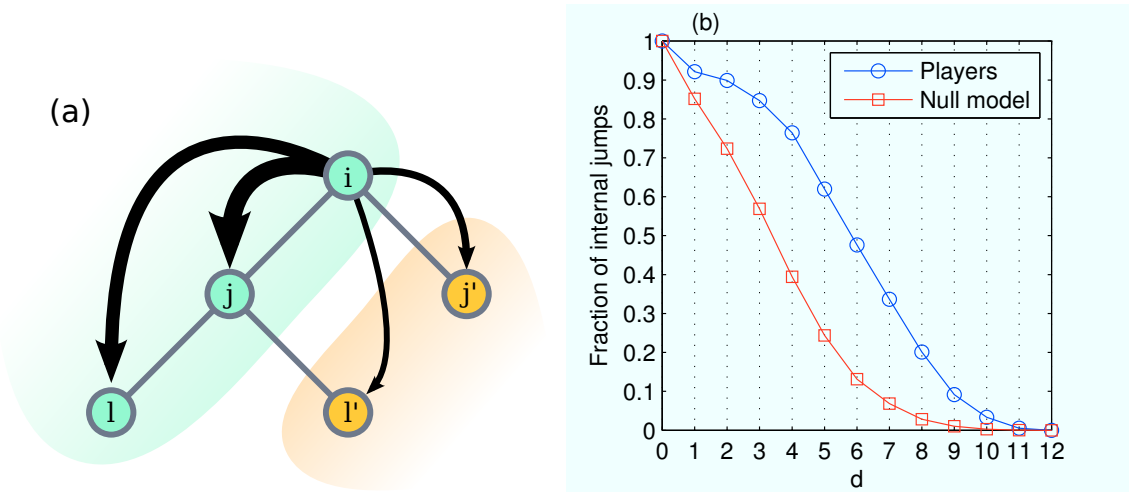

Figure 6.3: Influence of socio-economic clusters on mobility. (a) Sketch of jump patterns from a sector i to sectors within the same cluster, j and l, and to sectors in a different cluster, j', l'. Although sectors j' and l' have the same graph distance from sector i as sectors j and l respectively, transitions across cluster border have smaller probabilities. (b) Quantitative evidence of the tendency of players to avoid crossing borders. Red squares show the null model, i.e. the fraction of all pairs of sectors at a given distance d being in the same cluster. Blue circles show the fraction of measured jumps leading into the same cluster, per distance. Coincidence of the two curves would indicate that clusters have no effect on mobility. Clearly this is not the case – there is a strong tendency of players to avoid crossing the borders between clusters.

might be simply related to the topology of the network, as in the case of random walkers whose motions are constrained on graphs with strong community structures [27]. Nodes belonging to the same cluster are in fact either directly connected or are at short distance from one another. This proximity is reflected in the block-diagonal structure of the adjacency matrix A and of the distance matrix D, respectively shown in Fig. 6.4 (a) and (b). We have therefore checked whether the presence of the socio-economic clusters originally introduced by the developers of the game can be derived solely from the structure of the network. For this reason we adopted standard community detection methods based on the adjacency and on the distance matrix [134, 135]. The results, reported respectively in Fig. 6.4 (d) and (e), show that detected communities deviate significantly from the clusters, implying that in our online world the socio-economic regions cannot be recovered merely from topological features. In comparison we considered the player transition count matrix M, shown in Fig. 6.4 (c), which displays a similar block-diagonal structure as A and D, but with the qualitative difference that it contains *dynamic* information on the system. The entry m_{ij} of the transition count matrix M is equal to the number of times a player's position was on sector i and then, on the following day, on sector j. This number is cumulated for all players. The entry π_{ij} of the transition probability matrix Π corresponds to the probability that a player moves to a sector j given that on the previous day the player's location was sector i. It

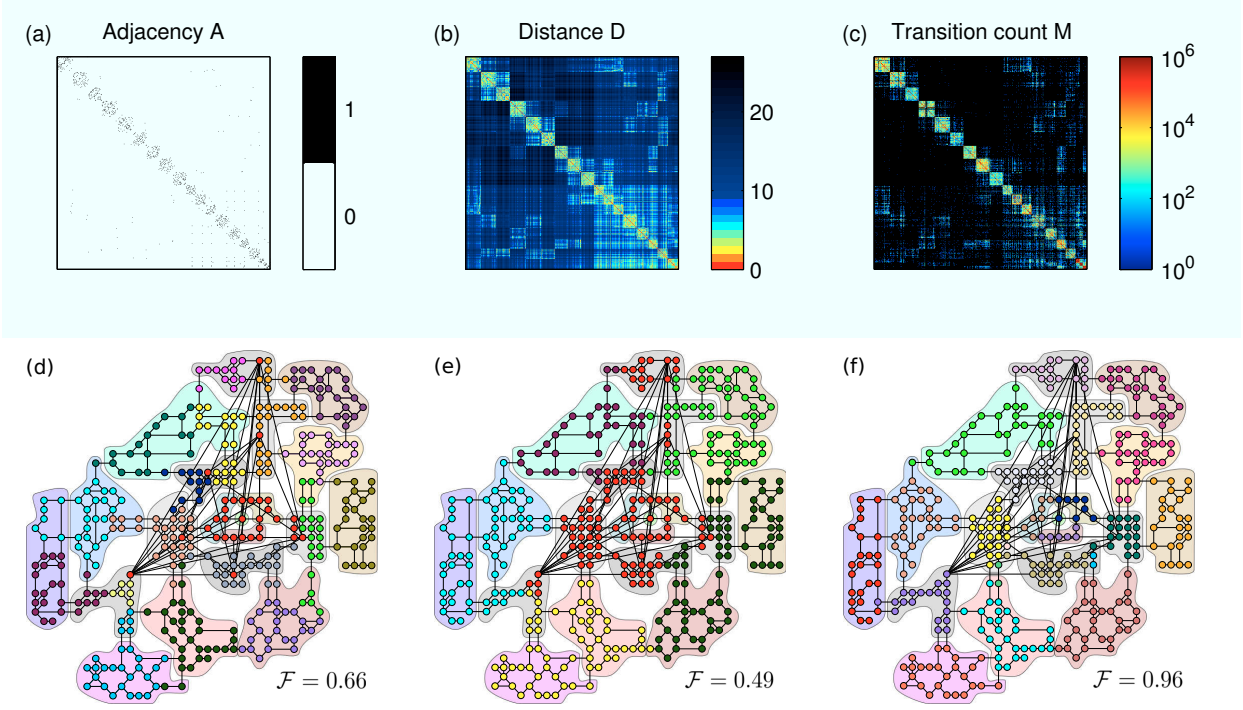

Figure 6.4: Extracting communities from network topology and from mobility patterns. (a) The adjacency matrix A of the universe network, (b) the matrix D of shortest path distances, and (c) the matrix M of transition counts of player jumps. Each of the three matrices contains 400×400 entries, whose values are color-coded. Sector IDs are ordered by cluster, resulting in the block-diagonal form of the three matrices. We have used modularity-optimization algorithms to extract community structures from the information encoded in the three matrices. Different node colors represent the different communities found, while the 20 different color-shaded areas indicate the predefined socio-economic clusters as in Fig. 6.1. The displayed Fowlkes and Mallows index $\mathcal{F} \in [0, 1]$ quantifies the overlap of the detected communities with the predefined clusters. The closer $\mathcal{F}$ is to 1, the better the match. (d) Although information contained in the adjacency matrix A allows to find 18 communities, a number close to the real number of clusters, the communities extracted do not correspond to the underlying color-shades areas ($\mathcal{F} = 0.66$). (e) Extracting communities from the distance matrix D only results in 6 different groups ($\mathcal{F} = 0.49$). (f) The 23 communities detected using the transition count matrix M reproduce almost perfectly the real socio-economic clusters ($\mathcal{F} = 0.96$), with only a few mismatched nodes detected as additional clusters.

reads: $\pi_{ij} = \frac{m_{ij}}{\sum_l m_{il}}$, where m_{ij} is the number of observed player movements from sector i to sector j, and the sum over l is over all sectors of the universe. The matrix Π is a stochastic matrix, i.e. it has the property that the entries of each row sum to one, as it is the Markov approximation of the process underlying mobility.

Figure 6.4 (f) shows that community detection methods applied to the transition count matrix M reveal almost perfectly all the socio-economic areas of the universe. This finding demonstrates that mobility patterns contain fundamental information on the socio-economic constraints present in a social system. Therefore, a community

detection algorithm applied to raw mobility information, as the one proposed here, is able to extract the underlying socio-economic features, which are instead invisible to methods based solely on topology.

In the rest of this section, we give a detailed treatment of adopted community detection methods and measures. Since the Pardus universe is divided in a number of *clusters*, defining political, independent units in the game universe, we have shown that the mobility patterns of players are influenced by such borders. At the same time, the topology of the Pardus universe itself might affect the mobility patterns. In order to investigate the importance of these two elements, one needs to compare the topological modules that can be extracted from the adjacency matrix A or distance matrix D, with the dynamical communities emerging from the collective movement behaviour of players.

At the sector level, the Pardus universe is a directed weighted network with $L = 1160$ links and $N = 400$ nodes. The majority of links are *wormholes* ($\sim 95\%$), mutual links that connect nearby nodes (see Fig. 6.1) and have a small traveling cost (in terms of APs). The long-range links in Fig. 6.1 instead represent *X-holes* and *Y-holes*. Players moving along such links incur significantly higher traveling costs than in the case of wormholes. Since X/Y-holes may be only scarcely used in-game, in addition to studying the complete directed weighted adjacency matrix, A, we also study the adjacency matrix A^{wh} where X/Y-holes were removed, yielding a symmetric and unweighted network. Finally, we consider the weighted network D^{inv}, defined element-wise as $d_{ij} = d(i,j)^{-1}\ \forall\ i \neq j$ and $d_{ii} = 0\ \forall i$, where $d(i,j)$ is the shortest path distance on the Pardus network.

The player dynamics was studied at the aggregate level through the transition count matrix M and the normalized transition matrix Π. Each element m_{ij} of M corresponds to the total number of times a player was found at position i at a time t and at position j at time $t+1$. The transition matrix $\Pi = (\pi_{ij})$ is obtained by row-normalizing M so that $\pi_{ij} = \frac{m_{ij}}{\sum_l m_{il}}$. Hence, for all rows i, $\sum_j \pi_{ij} = 1$ and Π is a well-defined probability matrix for the transitions between pair of nodes in the network. Notice that for both M and Π the diagonal elements can be significantly different from zero and therefore the resulting networks display self-loops. Moreover, both matrices M and Π correspond to directed, weighted networks, and therefore can be thought as representing flows across the networks. For completeness, we also define the symmetrized versions of the matrices above, namely the symmetrized jump matrix $M^{\mathrm{symm}} = (M + M^{\mathrm{T}})/2$, Π and the symmetrized transition matrix $\Pi^{\mathrm{symm}} = (\Pi + \Pi^{\mathrm{T}})/2$. The corresponding weighted networks are undirected and represent a first coarse-graining of the information contained in the dynamical flows. It is thus interesting to compare these two to understand how much information is lost in the coarse-graining.

We performed community detection algorithms by optimizing modularity [134, 135]. To ensure consistency, we checked the results under different heuristics and repeated detections [136, 137]. Figure 6.5 shows the communities extracted from the M, M^{symm}, Π and Π^{symm} matrices. The coloured hulls are included for comparison and indicate

the Pardus cluster to which each sector belongs. For comparison, in figure 6.6 we plot the communities obtained from the topological quantities, namely the directed weighted adjacency matrix A, the undirected unweighted matrix A^{wh} and the inverse distance matrix D^{inv}. One can easily see that the communities extracted from the transition matrices appear to reproduce much better the cluster structure as opposed to the topological communities.

Notice also that the partitions obtained for the dynamical transition matrices contain communities composed of a single node. Although unusual in community detection, this result is consistent with the mobility patterns. In fact, we measure the positions of players at the same time every day. Then, the presence of non zero values on the diagonal of M, M^{symm}, Π and Π^{symm} simply means that there is a positive probability for a player to be found again on the same node after 24 hours, implying that the player either stayed still on the node or traveled but came back to its original position within 24 hours. These self-loops are responsible for the presence of single-node communities in the dynamical matrices and for their absence in the topological ones, where there are no self-loops.

We find a different number of communities for different matrices, making it hard to come to a conclusion regarding which one is the closest to the Pardus cluster structure. To quantify the relative goodness of the partitions obtained from the various matrices, we calculate three measures of clustering similarity: the Fowlkes and Mallows index $\mathcal{F}$ [138], the Rand's criterion $\mathcal{R}$ [139] and the normalized information variation (NVI) [140]. Consider a set of nodes $\mathcal{T}$ of cardinality n and two partitions $\mathcal{C}$ and $\mathcal{C}'$ of $\mathcal{T}$, then the set of all unordered pairs of elements of $\mathcal{T}$ is the union of the sets [141, 142]:

- t_{11} is the set of pairs the same community under $\mathcal{C}$ and $\mathcal{C}'$;
- t_{01} is the set of pairs not in the same community under $\mathcal{C}$ but under the same community in $\mathcal{C}'$;
- t_{10} is the set of pairs in the same community under $\mathcal{C}$ but not under the same community in $\mathcal{C}'$;
- t_{00} is the set of pairs not in the same community under $\mathcal{C}$ and $\mathcal{C}'$;

and n_{11}, n_{01}, n_{10}, n_{00} are their respective cardinalities (and $n_{11} + n_{01} + n_{10} + n_{00} = n(n-1)/2$). The $\mathcal{F}$ and $\mathcal{R}$ indices are then given by:

$$\mathcal{F} = \frac{n_{11}}{\sqrt{(n_{11} + n_{10})(n_{11} + n_{01})}} \qquad \mathcal{R} = \frac{2(n_{11} + n_{00})}{n(n-1)} \tag{6.3}$$

which are essentially two ways of quantifying how well the partitions match pairs of nodes. Therefore a perfect match between two partitions will have $\mathcal{F}, \mathcal{R} = 1$. The *Variation of Information* (VI) is a measure based on information theoretical concepts

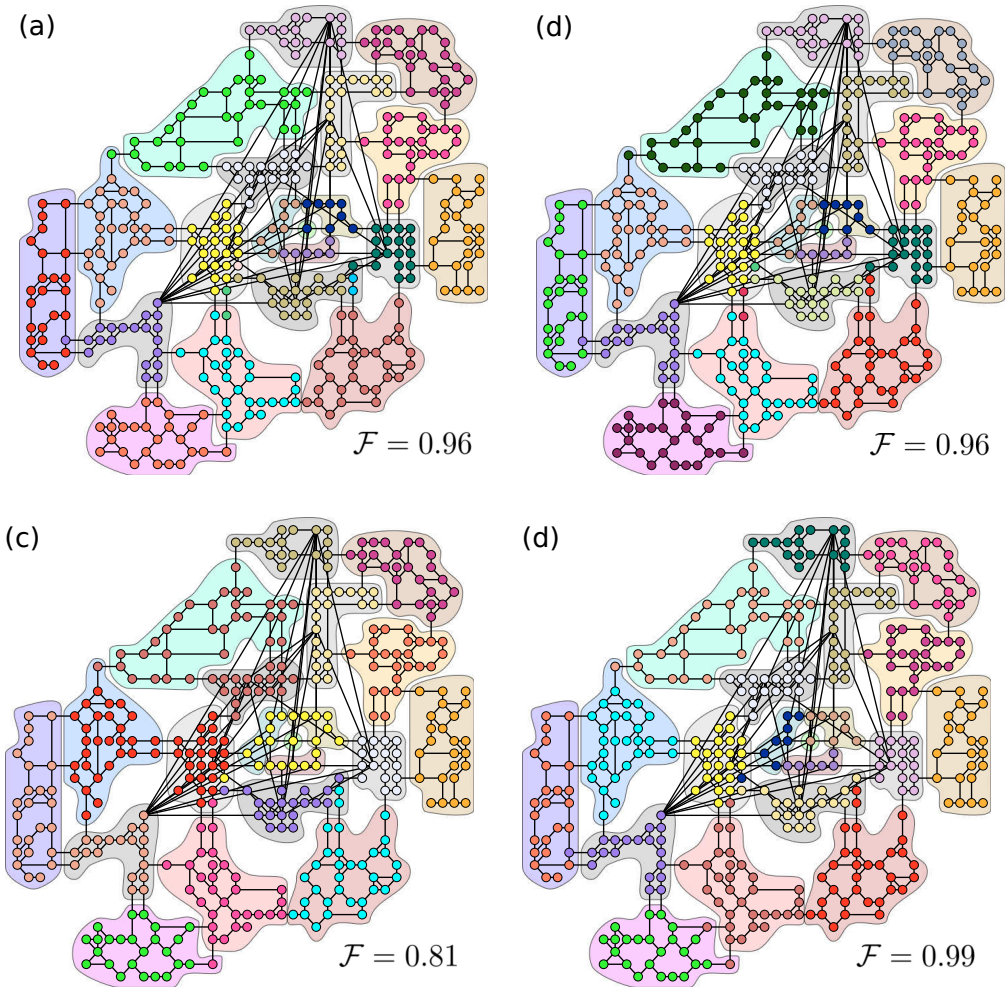

Figure 6.5: Extracting communities from mobility patterns. Communities found for (a) the jump matrix M, (b) the symmetrized jump matrix $M^{\text{symm}} = (M + M^{\text{T}})/2$, (c) the transition matrix Π and (d) the symmetrized transition matrix $\Pi^{\text{symm}} = (\Pi + \Pi^{\text{T}})/2$. Different node colours represent the different communities found, while the 20 different colour-shaded areas indicate the predefined socio-economic clusters as in Fig. 6.1. The communities found through the information of motions reproduces well the bulk of the Pardus cluster structure, with a few exceptions along borders where some nodes are assigned to wrong clusters. The Fowlkes and Mallows index $\mathcal{F}$ is close to 1 for all detected partitions, reflecting the good match. For more measures, see Table 6.2.

and represents the informational distance between two partitions. Therefore, if the VI is large, the two partitions are very dissimilar. The VI of a partition is bounded by $\log_2 n$, hence it is possible to normalize it, obtaining the *Normalized Variation of Information* (NVI $\in (0, 1)$):

$$\mathcal{NVI}(\mathcal{C}, \mathcal{C}') = \frac{\mathcal{VI}(\mathcal{C}, \mathcal{C}')}{\log_2 n} \tag{6.4}$$

where

$$\mathcal{VI}(\mathcal{C},\mathcal{C}') = \mathcal{H}(\mathcal{C}) + \mathcal{H}(\mathcal{C}') - 2\mathcal{I}(\mathcal{C},\mathcal{C}') \tag{6.5}$$

The terms in equation (6.5) are the entropy $\mathcal{H}(\mathcal{C})$ of partition $\mathcal{C}$ and the mutual information between two partitions $\mathcal{C}$ and $\mathcal{C}'$ [142]:

$$\mathcal{H}(\mathcal{C}) = -\sum_{i=1}^{k} P(i)\log_2 P(i) \qquad \mathcal{I}(\mathcal{C},\mathcal{C}') = \sum_{i=1}^{k}\sum_{j=1}^{l} P(i,j)\log_2 \frac{P(i,j)}{P(i)P(j)} \tag{6.6}$$

where $P(i) = \frac{|C_i|}{n}$ is the probability that an element of $\mathcal{T}$ chosen at random belongs to community $C_i \in \mathcal{C}$, and $P(i,j) = \frac{|C_i \bigcap C'_j|}{n}$ the probability that an element belongs to $C_i \in \mathcal{C}$ and to $C'_j \in \mathcal{C}'$.

Table 6.2 reports the values obtained for the studied matrices. The values of the Fowlkes-Mallows and Rand indices for the dynamical communities are much closer to 1 than the ones for the topological communities. The result is confirmed also by the NVI values, where we measured very small values for the dynamical partitions, indicating that player mobility follows closely the Pardus cluster structure. It could be argued that this similarity emerges from the topological structure of the network. However, we also found a difference of almost one order of magnitude between the dynamical and topological partitions and thus such hypothesis is not supported, that is the topological properties (e.g. adjacency matrix, distance matrix) produce partitions that are very different from the dynamical ones and the Pardus cluster one and cannot therefore be considered as the underlying mechanism of the mobility patterns. Moreover, this result is robust under different measures of player movement, as shown by the remarkable stability of the values of the clustering similarity measures for the other dynamical cases, M^{symm}, Π and Π^{symm}, which stay close to the ones obtained for M. Therefore, our conclusions cannot be considered an artifact of the particular measure we adopted.

6.6 Anomalous diffusion and a long-term memory model

In order to characterize the diffusion of players over the network, we have computed the mean square displacement (MSD) of their positions, $\sigma^2(t)$, as a function of time. The MSD is defined as $\sigma^2(t) = \langle (r(T+t) - r(T))^2 \rangle$, where $r(T)$ and $r(T+t)$ are the sectors a player occupies at times T and $T+t$ respectively, and where $(r(T+t) - r(T))$ denotes the distance between the two sectors. The average $\langle\cdot\rangle$ is performed over all windows of size t, with their left boundaries going from T=0 to T=1000-t, and over all the 1458 players in the data set. If σ^2 has the form $\sigma^2(t) \sim t^\nu$ with an exponent $\nu < 1$, the diffusion process is subdiffusive, in the case $\nu > 1$ it is super-diffusive. An exponent of $\nu = 1$ corresponds to classical brownian motion [143, 144].

Results reported in Fig. 6.7 (a) indicate that, for long times, the MSD increases as

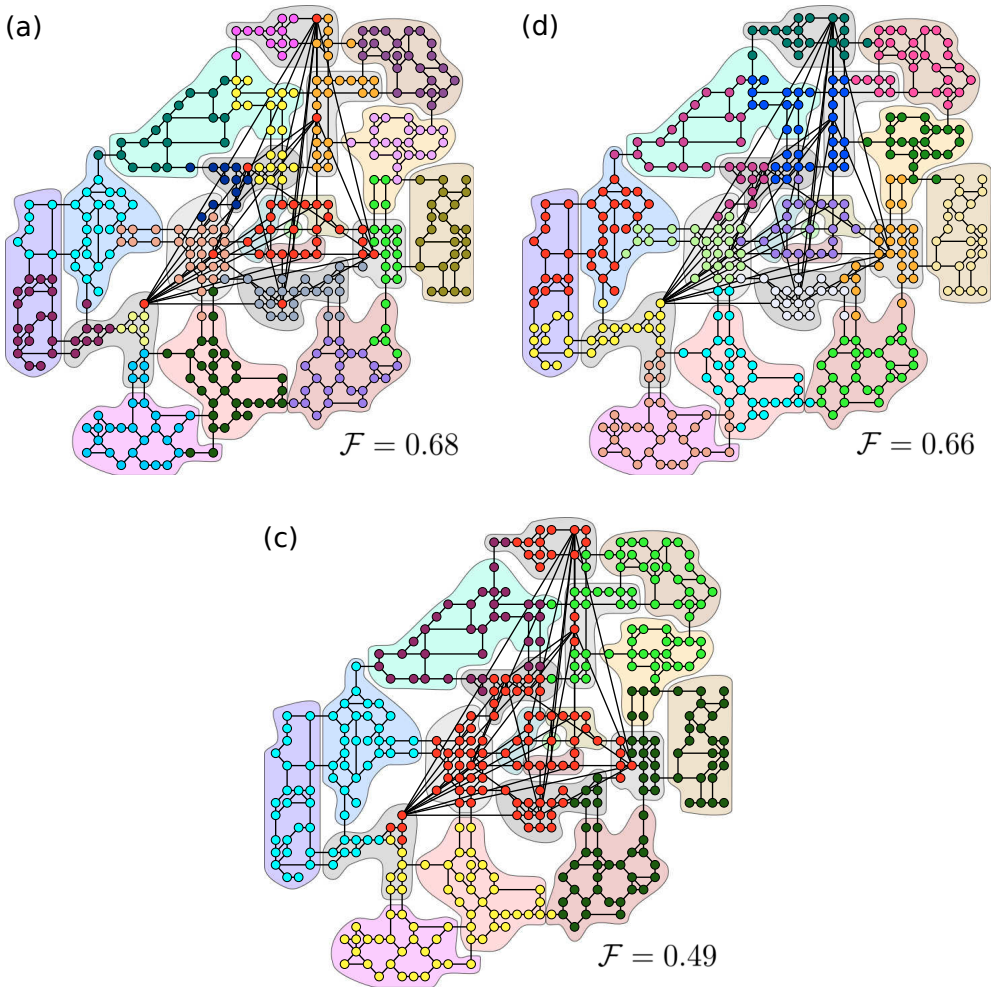

Figure 6.6: Extracting communities from topological information. Communities found for (a) the adjacency matrix A, (b) the adjacency matrix A^{wh} in which the X/Y-holes were removed yielding an undirected unweighted network, and (c) the distance matrix D^{inv}. Different node colours represent the different communities found, while the 20 different colour-shaded areas indicate the predefined socio-economic clusters as in Fig. 6.1. The partitions obtained from the adjacency matrices produce communities that cross over the borders between clusters and therefore do not recover the clusters well. This is particularly evident in the case of D^{inv} where only 6 communities are found. The Fowlkes and Mallows index $\mathcal{F}$ is not close to 1 for all detected partitions, reflecting the bad match. For more measures, see Table 6.2.

a power-law:

$$\sigma^2(t) \sim t^{\nu} \tag{6.7}$$

with an exponent $\nu \approx 0.26$. This anomalous subdiffusive behavior is not a simple effect of the topology of the Pardus universe. In fact, as shown in Fig. 6.7 (b), gray stars, the simulation of plain random walks on the same network produces a standard diffusion with an exponent $\nu \approx 1$ up to $t \approx 100$ days, and then a rapid saturation

Matrix	Network Properties	n_{comm}	$\mathcal{F}$	$\mathcal{R}$	NVI
Clusters	—	20	1	1	0
A	directed, unw.	18	0.678	0.963	0.179
A^{wh}	undirected, unw.	17	0.655	0.957	0.180
D^{inv}	directed, weight.	6	0.489	0.864	0.271
M	directed, weight.	23	0.963	0.996	0.025
M^{symm}	symmetrized, weight.	22	0.957	0.995	0.026
Π	directed, weight.	14	0.812	0.973	0.075
Π^{symm}	symmetrized, weight.	19	0.999	0.993	0.036

Table 6.2: Overview of community detection results for the studied matrices. From left to right, the columns correspond to: the studied matrix, the properties of the corresponding network, the number of communities found n_{comm}, the scores for the Fowlkes-Mallows index $\mathcal{F}$ [138], the adjusted Rand's criterion $\mathcal{R}$ [139] and, finally, the normalized information variation (NVI) [140]. For reference, the first row contains the values for the Pardus cluster structure. The closer the indices $\mathcal{F}$ and $\mathcal{R}$ are to 1, and the closer the NVI is to 0, the better a found partition matches the clusters. The values reported clearly indicate that the Pardus cluster structure is faithfully reproduced by the player mobility. On the other hand, the topological, non-dynamic properties (e.g. adjacency matrix, distance matrix) produce partitions that are very different from the Pardus cluster structure.

effect which is not present in the case of the human players. Insights from the previous section suggest that the anomalous diffusion behavior might be related to the tendency of players to avoid crossing borders. We have therefore considered a Markov model in which each walker moves from a current node i to a node j with a transition probability $\pi_{ij} = m_{ij}/\sum_l m_{il}$, where m_{ij} is the number of jumps between sector i and sector j, as expressed by the transition count matrix M of Fig. 6.4 (c). The probabilities π_{ij} are the entries of the transition probability matrix Π, which contains all the information on the day-to-day movement of real players, such as the preference to move within clusters, the length distribution of jumps, as well as the tendency to remain in the same sector. Despite this detailed amount of information used (the matrix Π has 160,000 elements), the Markov model fails to reproduce the asymptotic behavior of the MSD, see magenta diamonds in Fig. 6.7 (b). Since the model considers only the position of the individual at its current time to determine its position at the following time, deviations from empirical data appear presumably due to the presence of higher-order memory effects. For this reason we have considered the recently proposed preferential return model [124] which incorporates a strong memory feature. The model is based on a reinforcement mechanism which takes into account the propensity of individuals to return to locations they visited frequently before. This mechanism is able to reproduce the observed tendency of individuals to spend most of their time in a small number of locations, a tendency which is also prevalent in the mobility behavior of Pardus players. However, the implementation of the preferential return model on the Pardus universe network is not able to capture the scaling patterns of the MSD, as shown in Fig. 6.7 (b).

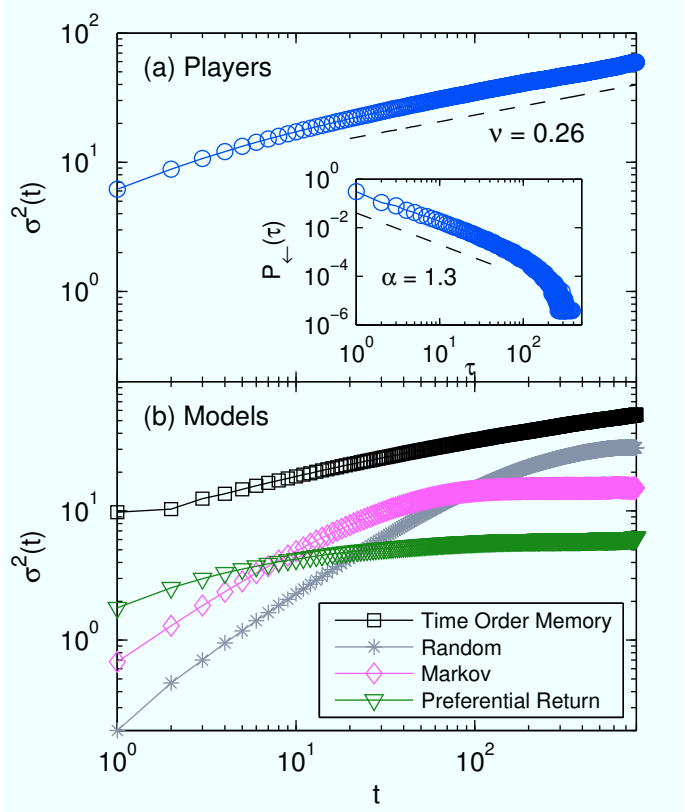

Figure 6.7: Diffusion scaling in empirical data and simulated models. (a) The mean square displacement (MSD) of the positions of players follows a power relation $\sigma^2(t) \sim t^{\nu}$ with a subdiffusive exponent $\nu \approx 0.26$. The inset shows the average probability $P_{\hookleftarrow}(\tau)$ for a player to return after τ jumps to a sector previously visited. The curve follows a power law $P_{\hookleftarrow}(\tau) \sim \tau^{-\alpha}$ with an exponent of $\alpha \approx 1.3$ and an exponential cutoff. We report, for comparison, (b) the MSD for various models of mobility. For random walkers and in the case of a Markov model with transition probability $\pi_{ij} = m_{ij} / \sum_j m_{ij}$ we observe an initial diffusion with an exponent $\nu \approx 1$ and then a rapid saturation of $\sigma^2(t)$, due to the finite size of the network. A preferential return model also shows saturation and does not fit the empirical observed scaling exponent ν. Conversely, a model with long-time memory (Time Order Memory) reproduces the exponent almost perfectly. Such a model makes use of the empirically observed $P_{\hookleftarrow}(\tau)$ while the Markov model and the preferential return model over-emphasize preferences to locations visited long ago and does not recreate the empirical curve well. Curves are shifted vertically for visual clarity.

The reason is that in the model the probability for an individual to move to a given location does not depend on the current location, nor on the order of previously visited locations. Instead, we observe that in reality individuals tend to return with higher probability to sectors they have visited recently and with lower probability to sectors visited a long time before. Consequently a sector that has been visited many times but with the most recent visit dating back one year has a lower probability to be visited again than a sector that has been visited just a few times but with the last visit dating

back only one week.

To highlight this mechanism we measured the return time distribution in the jump-time series, which is the transformation of the time-series of daily sector IDs occupied by the players from real-time to jump-time, in order to be able to compare time-series of different length and to focus on the movements between sectors. An example of this conversion is provided: a time series

$$[5, 5, 5, 32, 32, 104, 5, 5, 104, 104, 104, 32, 337, 337, 32 \ldots]$$

becomes in jump-time

$$[5, 32, 104, 5, 104, 32, 337, 32, \ldots] .$$

We denote jump-time by the greek letter τ, that is, at jump-time τ a player has performed exactly τ jumps. We use τ in the computation of the first return time distribution. In the hypothetical time series of sectors $[5, 32, 104, 5, 104, 32, 337, 32]$ a first return to a sector lying $\tau = 1$ jumps back happens 2 times $(104, 5, 104$ and $32, 337, 32)$, for $\tau = 2$ this happens once $(5, 32, 104, 5)$, for $\tau = 3$ also once $(32, 104, 5, 104, 32)$. Hence, in this example, $P_{\hookleftarrow}(1) = 0.5$, $P_{\hookleftarrow}(2) = P_{\hookleftarrow}(3) = 0.25$, where the sum over all $P_{\hookleftarrow}(\tau)$ is equal to 1. In particular, we extracted the probability $P_{\hookleftarrow}(\tau)$ for an individual to return again (for the first time) to the currently occupied sector after τ jumps. As shown in the inset of Fig. 6.7 (a), we found that the return time distribution reads

$$P_{\hookleftarrow}(\tau) \sim \tau^{-\alpha} \tag{6.8}$$

with an exponent $\alpha \approx 1.3$. We used this information for constructing a model which takes into account the higher re-visiting probability of recently explored locations. In this way we can capture the long-term scaling properties of movements. Exactly these asymptotic properties are fundamentally relevant for issues of epidemics spreading or traffic management.

This "Time Order Memory" (TOM) model incorporates a power-law distribution of first return times, together with a power-law distribution of waiting times and an exponential distribution of jump distances, as those observed empirically in Fig. 6.2. We show below that these ingredients are sufficient to reproduce the subdiffusive behavior reported in Fig. 6.7 (a). The model works as follows: an individual stands still in a given sector for a number of days drawn from the waiting time distribution, Eq. (6.2). Then, the individual jumps. There are two possibilities: (i) with a probability v she returns to an already visited sector, (ii) with the probability $1 - v$ she jumps to a so far unexplored sector. In case (i), one of the previously visited sectors is chosen according to Eq. (6.8). In the exploration case (ii), the individual draws a distance d from the distance distribution, Eq. (6.1), and jumps to a randomly selected, unexplored sector at that distance. The model has four parameters. The parameters λ, β and α of equations (6.1), (6.2) and (6.8) respectively, are fixed by the data. Further,

averaging over all jumps and players, the probability of returning to an already visited location is $v \approx 0.83$. Similarly to the measured data, the MSD of the TOM model, black squares in Fig. 6.7 (b), exhibits no saturation effects and displays an exponent $\nu_{\mathrm{TOM}} = 0.23 \pm 0.02$ in full agreement with the exponent observed for the players.

The flat slope of $\nu = 0.26$ and the lack of saturation of the MSD of the players over the whole observation period exposes the significant level of subdiffusivity in the motions of individuals, consistent with previous findings [77, 124, 143–145]. However, the mere tendency of individuals to return to already visited locations is not sufficient to capture these subdiffusive properties of the MSD, but it is fundamental to consider a mechanism that takes into account the temporal order of visited locations, as achieved by the TOM model. Moreover, the TOM model is realistic in the sense that, in contrast to Markov models, it takes into account the tendency of individuals to develop a preference for visiting certain locations. At the same time it allows for the possibility that a previously preferred location becomes not frequented anymore. This view provides an alternative to recently suggested reinforcement mechanisms in preferential return models [124]. The possibility for individuals to "change home" is relevant when the model should be able to account for migration, which is an important feature in the long-time mobility behavior of humans.

Finally, we discuss to which extent the findings from our "social petri dish" are valid also for human populations unrelated to the game. Previous analyses of human social behavior in Pardus [48, 127] have shown agreement with well-known sociological theories and with properties on comparable behavioral data. Examining the preference of players to move *within* socio-economic regions is of obvious importance for clearing up the role of political or socio-economic borders on the movement and migration of humans, where the presence of borders has a strong influence on mobility [118, 146–148]. Online societies as the one of Pardus have the evident potential to serve as "socio-economic laboratories", where the complete knowledge of activities, social relations, and positions of all individuals can significantly advance our understanding of large-scale human behavior, in particular of mobility.

Chapter 7

Understanding cooperative behavior with functional brain networks

Reason has always existed,
but not always in a reasonable form.

Karl Marx

7.1 Neuroscience and Game Theory

Game theory provides a mathematical framework to study decision-making processes in groups of individuals. In a game, the players adopt one among a set of possible actions (strategies), and the reward or penalty for each player crucially depends on the actions taken by all players [149]. Game theory has proven useful in the investigation of the neural basis of social interactions and social decision-making. In particular, researchers have investigated what happens in the brain of subjects involved in games where each player can choose between cooperative and non-cooperative behaviors, or between altruistic and selfish behaviors, with the aim of understanding the modification of brain activity related to the selected strategy [150].

In this chapter we discuss an experiment to investigate the connection between brain activity, as measured by EEG recordings, and social interaction, modelled using the framework of game theory. We first introduce some notions about classical game theory, focusing on the Prisoner's Dilemma game, which is played by the participants of the experiment. Then, we describe the experimental setup and we discuss the methodology used to analyze the data from the experiment. This methodology relies on a complex networks approach: from the correlations between signals of different brain areas, one can construct what is called a functional brain network. These networks can thus be analyzed using tools of complex network theory. As a major result, we show that it is possible to predict the cooperative behavior of individuals by looking at the topological properties of their functional brain networks.

7.2 Classical Game Theory

Game theory is a unifying paradigm behind many scientific disciplines. It is a set of analytical tools and solution concepts, which provide explanatory and predicting power in situations involving interactive decisions, when the aims, goals and preferences of the participating players are potentially in conflict. Classical (rational) game theory is based upon a number of assumptions about the structure of a game. It assumes that agents (players) have well defined goals and preferences which can be described by a utility function. The utility is the measure of satisfaction the player derives from a certain outcome of the game, and the player's goal is to maximize his/her utility. Another key assumption in the classical theory is that players are perfectly rational and that this fact is common knowledge of all players. "Perfect rationality" means that the players have well defined payoff functions, and they are fully aware of their own and the opponents' strategy options and payoff values. They have no cognitive limitations in deducing the best possible way of playing whatever the complexity of the game is. In this sense computation is costless and istantaneous. "Common knowledge" implies that beyond the fact that all players are rational, they all know that all players are rational.

The intriguing predicted outcome, based on these hypotheses and using the mathematical formulations of different games that model conflict (see for example Sec. 7.2.1 for a description of the Prisoner's Dilemma), is that the solutions to the game consist in each player defecting with each other, meaning that each player tries to get his/her maximum benefit regardless what happens to the other player. However, this is in contrast with many empirical observations from the world around us, where it is clear that defection is not the common scenario. Quite the opposite, in many contexts individuals tend to cooperate. A pletora of mechanisms have been proposed to explain the emergence of cooperation, some of these stating that individuals that play iteratively the same game can turn to cooperation in the long term [151].

Here we do not want to go into the details of game theory, but just present the main characteristics of a game extensively used for the study of conflict and for modelling social interaction: the Prisoner's Dilemma.

7.2.1 The Prisoner's Dilemma

The Prisoner's Dilemma is a strategic game used as the standard metaphor to conceptualise the conflict between mutual support and selfish exploitation among interacting non-relatives in biological communities. Its name comes from a thought experiment involving suspects in a crime, and this is also the first formulation we want to provide before giving the rigorous mathematical one. The dilemma, as told by Osborne [149], is the following:

> Two suspects in a major crime are held in separate cells. There is enough evidence to convict each of them of a minor offense, but not enough evidence

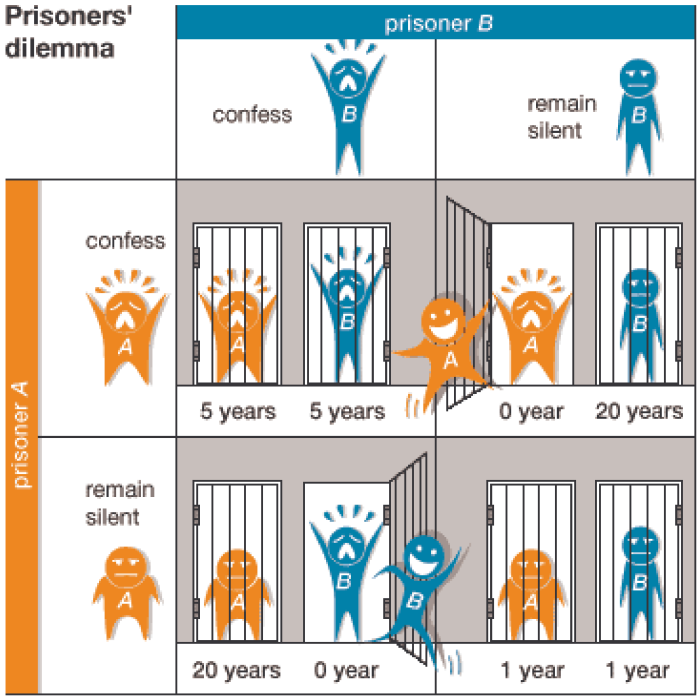

Figure 7.1: Representation of the Prisoner's Dilemma. The two rows correspond to the two possible actions of prisoner A, the two columns correspond two the two possible actions of prisoner B. Depending on their actions, either both go to prison for 5 years (both confess), one is set free but the other one goes to prison for 20 years (one confesses, the other one remains silent), or both go to prison for one year (both remain silent). If both players are rational and assume that the other one is too, the decision of both players is to confess, which brings both of them 5 years in prison. The dilemma is that regardless of what the other prisoner chooses, one will always gain a greater payoff by betraying the other, leading to a situation where both confess and serve a longer time in prison (5 years) compared to the case where both cooperate (1 year).

> to convict either of them of the major crime unless one of them acts as an informer against the other (confesses). If they both stay silent, each will be convicted of the minor offense and spend one year in prison. If one and only one of them confesses, he/she will be freed and used as witness against the other, who will spend twenty years in prison. If they both confess, each will spend five years in prison.

Here, regardless of what the other decides, each prisoner gets a higher pay-off (less years in jail) by betraying the other. For example, Prisoner A can, with close certainty, state that no matter what prisoner B chooses, prisoner A is better off 'ratting him out' (defecting) than staying silent (cooperating). As a result, solely for his own benefit, prisoner A should logically betray him. On the other hand, if prisoner B acts the same way, then they both have acted this way, and both receive a lower reward than if both were to stay quiet. Seemingly logical decisions result in both prisoners being worse off

than if each chose to diminish the sentence of his accomplice at the cost of spending more time in jail himself. This solution to the Dilemma corresponds to the so-called Nash Equilibrium of the game (see also below in this section).

The game can be summarized in a visual form as shown in Fig. 7.1. In general, games like the Prisoner's Dilemma can be formulated in a rigorous way by using a matrix representation. In the case of the Prisoner's Dilemma, where a player can choose between two possible strategies, either to cooperate ($\mathcal{C}$) with the other player or to defect ($\mathcal{D}$), the outcomes of the game are summarized in the following matrix:

$$\Pi = \begin{pmatrix} P^{\mathcal{CC}} & P^{\mathcal{CD}} \\ P^{\mathcal{DC}} & P^{\mathcal{DD}} \end{pmatrix} = \begin{pmatrix} (R,R) & (S,T) \\ (T,S) & (P,P) \end{pmatrix} \tag{7.1}$$

where rows correspond to the strategies player A can adopt ($\mathcal{C}$ for first row, $\mathcal{D}$ for the second), and columns correspond to the strategies of player B ($\mathcal{C}$ for first column, $\mathcal{D}$ for the second). Each entry of the matrix contains two values, the first being the payoff of player A, the second the payoff of player B. These values are usually denoted by the capital letters R for Reward (for mutual cooperation), S for Sucker, T for Temptation (to defect), and P for Punishment (for mutual defection). In such a matrix form, the Prisoner's Dilemma is a game where the following relation between the values of the payoff holds:

$$T > R > P > S$$

In particular, the temptation to defect is higher than the reward to cooperate. In the experiment we describe in Sec. 7.3, we adopt the following payoff matrix:

$$\Pi = \begin{pmatrix} P^{\mathcal{CC}} & P^{\mathcal{CD}} \\ P^{\mathcal{DC}} & P^{\mathcal{DD}} \end{pmatrix} = \begin{pmatrix} (2,2) & (0,3) \\ (3,0) & (1,1) \end{pmatrix} \tag{7.2}$$

A *Nash equilibrium* is defined as a set of strategies for which no player can do better by unilaterally changing his or her strategy. Thus, each strategy in a Nash equilibrium is a best response to all other strategies in that equilibrium. In the case of the Prisoner's Dilemma, both players defecting is a Nash Equilibrium since in no situation can a player can gain a higher benefit if he/she switches to cooperation, provided that the strategy of the other player remains fixed. This set of strategies is also the only Nash Equilibrium of the game.

7.2.2 The Iterated Prisoner's Dilemma

When the Prisoner's Dilemma is played iteratively the situation becomes more complicated, since a player remembers previous actions of the opponent and can change the strategy accordingly [152]. This is the actual situation in the experimental setting studied in this chapter. In the case of an Iterated Prisoner's Dilemma, we are interested in the iterative strategy of the player. If a player plays iteratively N games, his/her

t	A B	$t+1$: $\mathcal{C}$	$t+1$: $\mathcal{D}$
	$\mathcal{CC}$	C	D
	$\mathcal{CD}$	C	T
	$\mathcal{DC}$	T	D
	$\mathcal{DD}$	C	D

Figure 7.2: A schematic representation of the classification of the iterated prisoner's dilemma strategies for a player A. The classification of the iterated strategies is based on the pair of strategies adopted by the players in the simple Prisoner's Dilemma at trial t and on the strategy player A adopts in the game in the following trial $t+1$. In the first column in the table, all the possible outcomes of the game at trial t are shown, where the first symbol, $\mathcal{C}$ or $\mathcal{D}$, refers to the strategy player A adopted, while the second symbol refers to the strategy player B chose. In the first row, the possible strategies of player A in the trial $t+1$ are shown. In this way, it is possible to classify the iterated strategy of player A as cooperation (C), Defection (D) or Tit-for-Tat (T).

iterative strategy can be classified by looking at the entire sequence composed of the total of the N single game outcomes. For our experiment, we simplify the classification of possible strategies taking place. We base the classification of the iterative strategy of a player just by looking at the outcome of a single game (what both players do) and at the strategy the player adopts in the following single game (see also Fig. 7.2). The strategy a player adopts can be of three different kinds: i) cooperative strategy (C), when a player who is playing defection $\mathcal{D}$, starts to cooperate $\mathcal{C}$ as soon as the other player defects, or when a player who is playing cooperation $\mathcal{C}$, continues to do so for all the possible actions of the opponent; ii) defector strategy (D), when a player who is playing cooperation, starts to defect $\mathcal{D}$ as soon as the other player cooperates $\mathcal{C}$, or when a player who is playing defection $\mathcal{D}$, continues to do so for all the possible actions of the opponent; iii) tit-for-tat strategy (T), when a player who is cooperating $\mathcal{C}$ switches to defection $\mathcal{D}$ if the opponent defects $\mathcal{D}$, or when a player who is defecting $\mathcal{D}$ switches to cooperation $\mathcal{C}$ if the opponent cooperates $\mathcal{C}$.

Considering the pair of iterative strategies of two players, the outcome of each round, or trial, of the game can be one of the six possible combinations of the individual

actions. Three of them, namely the cases in which players play both either cooperation (indicated as CC), or defection (indicated as DD) or tit-for-tat (indicated as TT), are called here "pure" strategies because the two players adopt the same action, while the other cases, indicated as CD, CT, and DT or equivalently DC, TC, and TD, are called "mixed" strategies since the players adopt different iterative strategies.

7.3 Design of the experiment

Fifty-two voluntary and healthy subjects (age ranging from 23 to 33 years) participated in our experiment. These volunteers were combined in 26 couples and played an iterated Prisoner's Dilemma game of at least 200 rounds. In every round, or trial, one single game, described in Sec. 7.2.1, is played. Each of the two players were asked to choose either to cooperate (C) or to defect (D) and to enter their decision through a special keyboard. Their "reward" in each single game was assigned according to the payoff matrix Eq. 7.2. A trial (t) consists of two distinct time intervals. During the first interval, players have to communicate their strategies on the base of the outcome at the previous trial ($t-1$). Typically, this interval ranged from 0.5 seconds to 2 seconds. After communicating their choice, a report summarizing the strategy and the score at the trial (t) is displayed for 4 seconds. At the beginning of this second interval, the two subjects make a new decision to be communicated in the next trial ($t+1$). A scheme of the experimental setup is provided in Fig. 7.3.

7.3.1 EEG recordings and cortical activity

During all the trials, both the players wore an electrode cap composed of 64 sensors, registering the cortical activity of the players. Cortical activity from scalp EEG recordings was estimated by using an average realistic head model (MNI template, http://www.loni.ucla.edu/ICBM/) consisting of four concentric surfaces: scalp, inner skull, outer skull and cortex. Each surface is composed of approximately 3000 uniformly disposed vertices, each corresponding to one current dipole. By means of standard methods [153–155] in neuroscience for the solution of the so-called electromagnetic linear inverse problem, from the 3000 dipoles the activity of a total of six regions of interest (ROIs) of the scalp of each subject was estimated. The ROIs used are standard regions according to the Brodmann classification [156] and are: 10_L for the left hemisphere and 10_R for the right hemisphere, the Anterior Cingulate Cortex (ACC), the Cingulate Motor Area (CMA), the Brodmann area 7_L for the left hemisphere and 7_R for the right one.

After all the trials were played, all the cortical signals of the single players in each trial were classified as Cooperation (C), Defection (D), or as Tit-for-Tat (T) according to the rules specified in the Sec. 7.2.2. Thus, three different subsets of trials C, D and T were collected for each subject.

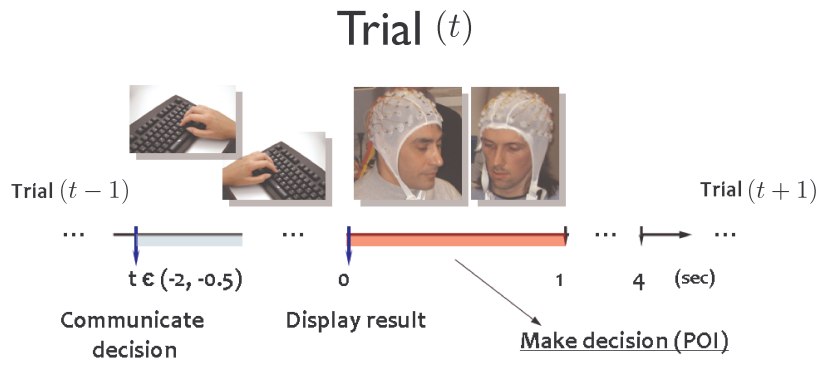

Figure 7.3: **Timeline of the experiment.** At each round, or trial, players are asked to choose either to cooperate ($\mathcal{C}$) or defect ($\mathcal{D}$) through a special keyboard. A trial (t) consists of two distinct time intervals. During the first interval, players have to communicate their strategies on the base of the outcome at the previous trial (t-1). Typically, this interval ranged from 0.5 seconds to 2 seconds. After communicating their choice, a report summarizing the strategy and the score at the trial (t) is displayed for 4 seconds. At the beginning of this second interval, the two subjects make a new decision to be communicated in the next trial (t+1). In particular, we considered the first second (i.e., 1 s of EEG recordings) as period of interest (POI) for the initial decision-making processes.

7.4 The concept of hyper-brain

Most of the approaches used so far in the literature to characterize brain responses during social interaction have the major limitation of measuring signals from just one player at a time. The functional connectivity between the brain activities of two interacting individuals is thus not measured directly, but inferred from independent observations subsequently aggregated by statistical models which associate observed behaviors and neural activation. In the experiment presented, instead the cortical activity of players has been recorded simultaneously by means of EEG hyperscanning. To get all the potentialities from the simultaneous recordings, we have devised a method to create a merged dataset considering data from the six cortical regions of the two subjects, thus obtaining a set of 12 cortical signals. The cortical signals of the merged data set were then processed to construct a functional brain network [157], whose $N = 12$ nodes correspond to the ROIs individuated. For defining the links of the network, in neuroscience literature many different methods have been proposed[157]. Here, we use a standard method named Partial Directed Coherence (PDC) [158]. The method is based on the Granger causality index, a measure that quantifies whether a time series, a cortical signal in our case, can predict another one. The measure is not symmetric, meaning that the Granger causality of timeseries A to predict timeseries B can be different from the value obtained to predict A based on B. To each ordered pair of nodes of the functional brain network, we can then associate a value provided by

the PDC. In order to consider only functional links between ROIs that are not due to chance, we used a statistical significance threshold on the measure associated to each pair of nodes. All the (ordered) pairs of nodes whose PDC value is above the threshold are then connected by a directed weight link, with weight being equal to the estimated statistical significance. This weight also indicates the degree of interaction between the two ROIs.

Concerning the spectral properties of the EEG signals, we selected four frequency bands of interest (Theta 4-7 Hz, Alpha 8-12 Hz, Beta 13-29 Hz and Gamma 30-40 Hz) and we gathered the corresponding cortical networks by averaging the values within the respective range. Finally, for each band, different functional brain networks were constructed. More specifically, for each band six different graphs were produced, corresponding to the six different possible pairs of strategies CC, DD, TT, CD, CT, DT (see Sec. 7.2.2). The first six nodes of the graph correspond to ROIs of the first player, while the remaining 6 nodes correspond to the ROIs of the second player. In practice, we represented the functional connectivity of the two brains altogether in the same graph: a link in the graph can be either an intra-brain or an inter-brain connection, according to the fact that it expresses the relationship between two ROIs belonging to the same brain, or between a region of one brain and a region of the other brain. We call these graphs *hyper-brain networks*, since they represent at the same time the correlations between ROIs in the same brain and correlations across the two brains.

Fig. 7.4 illustrates, for a representative couple of subjects, the hyper-brain networks associated to the pure strategies CC, DD, and TT, in the Alpha (8-13 Hz) frequency band. Each network consists of twelve nodes representing the six specific ROIs considered in this study for each subjects' brain. Note that the selected ROIs are the same for each player. To highlight the inter-brain connectivity, only links between the two brains are illustrated in the figure.

7.5 Is it possible to predict social behavior?

We show here the results obtained from the analysis of the hyper-brain networks for the 26 couples of subjects studied. First, we report a series of graph indexes, some of them already introduced in chapter 1, that have been used to characterize the hyper-brain graphs. With these measures, we prove that the structure of networks corresponding to situations in which individuals play cooperatively is significantly different from cases of couples playing in a "selfish" way. Moreover, we test the possibility to predict the outcome of a game from the structural analysis of the hyper-brain network obtained from the signals recorded during the decision-making process. This also suggests that EEG hyper-scanning and hyper-brain networks allow the direct observation of neural signatures of human social interactions.

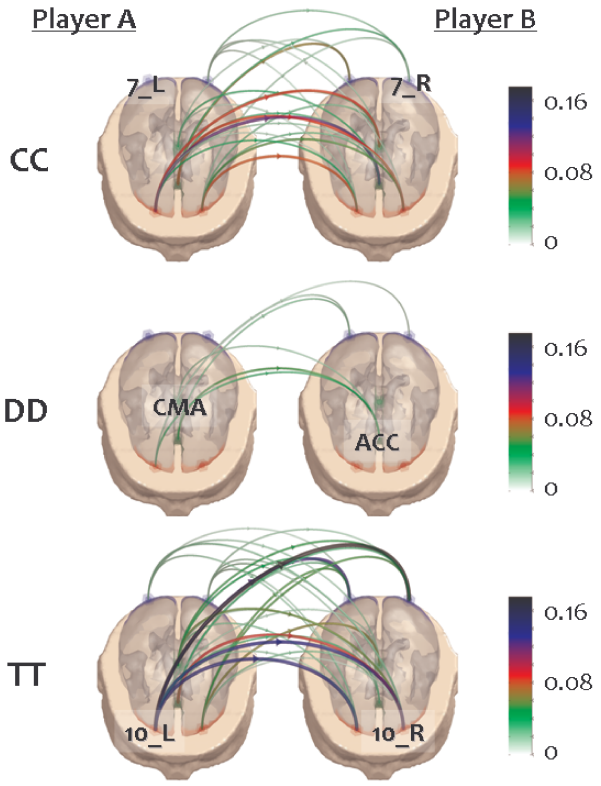

Figure 7.4: **Inter-brain connectivity for pure strategies in the Alpha band.** Two generic players are represented by the realistic head models used to estimate the cortical activity in the same six regions of interest (ROIs). Different colored points indicate the barycenters of these ROIs on the semi-transparent cortex. For the sake of simplicity, we did not label the ROIs of each subplot, but just two for the CC (7_L, 10_L), TT (7_R, 10_R) and DD (CMA, ACC) subplot. Only links between the two brains are illustrated in each hyper-brain network, i.e. the inter-brain connections. The size and the color of each directed connection represent the PDC values of a representative couples of subjects in the Alpha (8-13 Hz) frequency band.

7.5.1 Graph indexes

The hyperbrain networks we analyzed are directed weighted graphs consisting of $N = 12$ nodes. In general, they can be represented by a $N \times N$ weighted adjacency matrix $W = \{w_{ij}\}$, where $w_{ij} > 0$ is the weight associated to the directed arc from node i to node j, and in general $w_{ij} \neq w_{ji}$. As we have seen in chapter 1, the most intuitive index of a graph is its total number of links, which measures the overall level of connectivity within the system. The respective weighted version is the total network weight W that is the sum of all arc weights in the graph:

$$W = \sum_{i,j} w_{ij} \tag{7.3}$$

In weighted networks one can also define the *strength*, which is the extension of the concept of degree. The strength of a node is equal to the sum of the weights of the links incident in the node. In the case of directed weighted networks, we can define for each node the in- and out-strength. More rigorously, the out-strength s_i^{out} of node i is defined as:

$$s_i^{out} = \sum_j w_{ij},$$

while the in-strength s_i^{in} is:

$$s_i^{in} = \sum_j w_{ji},$$

As already mentioned in Sec. 1.2.1, the performance of a network can be measured by assuming that information flows along shortest paths and that the efficiency in the communication between two nodes i and j is inversely proportional to their shortest distance d_{ij}, i.e. the smallest sum of arc weights of all possible paths from i to j in the case of weighted networks. Namely, the efficiency index E of a graph, is defined as [23]:

$$E = \frac{1}{N(N-1)} \sum_{i \neq j=1}^{N} \frac{1}{d_{ij}} \tag{7.4}$$

If there is no path from i to j, $d_{ij} = \infty$ and the couple (i, j) does not contribute to the graph efficiency. Large distances imply small efficiency, while short distances imply high efficiency, with the efficiency being maximal in a fully connected graph.

We have also implemented two measures to quantify how well the graph G can be divided into two sets of nodes B_1 and B_2, corresponding to the brains of the two players. The divisibility D is defined as:

$$D = \frac{W}{\sum w_{ij} \left(1 - \delta\left(C_i, C_j\right)\right) + \epsilon} \tag{7.5}$$

where C_i indicates the community to which the node i belongs (in our case there are only two communities: $C_1 = B_1$ or $C_2 = B_2$); the δ function yields 1 if vertices i and j are in the same community (i.e. in the same brain), and 0 otherwise; ϵ is a positive constant (here set equal to W) to avoid possible divergence of D. The divisibility D is actually the inverse of the cut size [159] extended to weighted graphs. Modularity Q, originally defined for unweighted graphs (see Sec. 1.2.3), measures the difference between the fraction of arcs connecting nodes belonging to the same community in the actual graph and its expected value in a random graph. Modularity Q in the case of directed weighted graphs reads [160]:

$$Q = \frac{1}{W} \sum_{ij} \left(w_{ij} - \frac{s_i^{out} s_j^{in}}{W} \right) \delta\left(C_i, C_j\right) \tag{7.6}$$

where the δ function has the same meaning as for the divisibility D. As a result, in the expression of Q, the only contributions come from couples of nodes belonging to the same brain. Hence, the higher is the value of modularity, the better is the partition of the networks into the two communities B_1 and B_2. In order to compare network measures for different strategies τ (τ = CC, DD, TT, CD, CT, DT) of the same couple k ($k = 1, \ldots, 26$), we introduce the Z-score, $Z_\tau^k(x)$, of a generic network measure x (x being the efficiency E, the divisibility D, or the modularity Q) as:

$$Z_\tau^k(x) = \frac{x_\tau^k - \bar{x}^k}{\sigma^k} \tag{7.7}$$

The averages $\bar{x}^k$ and the standard deviations σ^k are evaluated, for each value of k, over all strategies τ. Finally, the average Z-score, $\langle Z_\tau(x) \rangle$, is evaluated, for each strategy τ, by averaging the Z-scores $Z_\tau^k(x)$, over all couples k:

$$\langle Z_\tau(x) \rangle = \langle \frac{x_\tau^k - \langle x^k \rangle}{\sigma^k} \rangle_k \tag{7.8}$$

7.5.2 Inter-brain connectivity discovers selfish behaviors

The novelty of the study we present in this chapter consists in classifying different social behaviors by comparing, for each pair of individuals, the six hyper-brain networks relative to CC, DD, TT, CD, CT, DT strategies. For each of the 26 couples involved, we have considered the graph efficiency E, and computed two measures, the divisibility D and the modularity Q, which give a quantitative estimation of how well the hyper-brain network can be separated into two subsets of nodes, corresponding respectively to the network of cortical regions of the two players. A comparison of the six values of E, D and Q, obtained for each couple of players and for each frequency band, allowed successful discrimination of selfish behavior from other behaviors, as reported in the pie diagrams of Fig. 7.5 for the Theta band. The first pie diagram shows that 50% of the cases (13 couples) display the minimal value of efficiency in the DD hyper-brain networks, 11.6% (3 couples) in the CC hyper-brain networks, and 19.2% (5 couples) in the TT hyper-brain networks. The remaining 19.2% (5 couples) exhibits the lowest efficiency in mixed-strategies (CD, CT and DT) hyper-brain networks. For any frequency band, the DD connectivity pattern has the lower efficiency with respect to the other five networks in approximately 50% of the couples. Similarly, modularity and divisibility are maximal for DD strategies in about 75% and 62% of the couples, respectively.These results indicate that hyper-brain networks corresponding to DD have longer paths between ROIs (lower global efficiency) and a small number of links between the two brains (high divisibility). this number being much lower than expected in a random graph with the same number of nodes and links (high modularity). Conversely, as shown in the bottom panels of Fig. 7.5, the efficiency is maximal for TT (resp. CC)

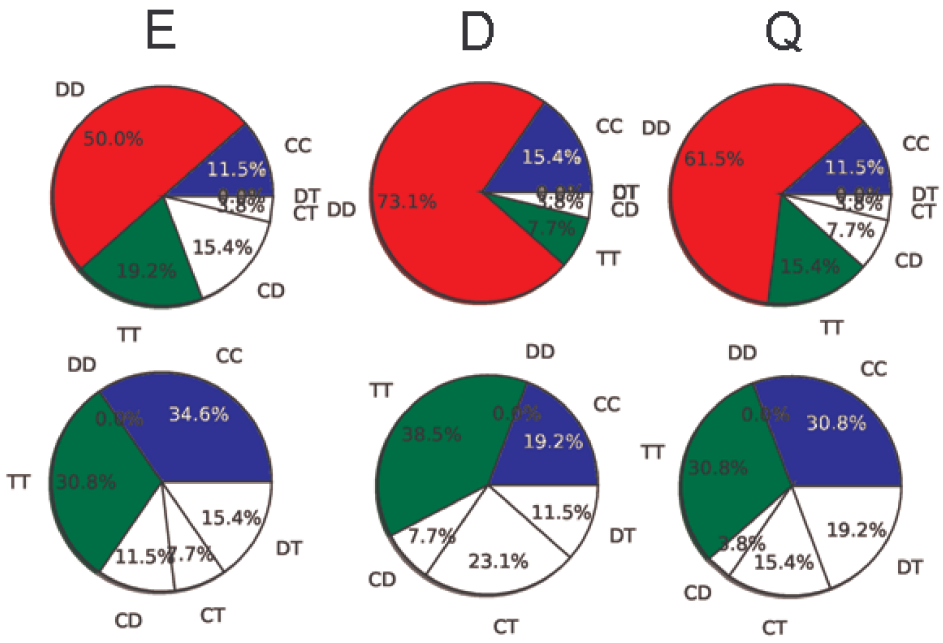

Figure 7.5: **Pie diagrams of efficiency E, divisibility D and modularity Q in the Theta band.** Top panels: from left to right the diagrams represent the percentage of cases - over the 26 couples - in which graph efficiency E is minimal, whilst the divisibility D and modularity Q are maximal. Bottom panels: percentage of cases - over the 26 couples - in which E is maximal and D and Q are minimal. Blue areas represent pure cooperation CC, red areas represent pure defection DD, green areas represent pure tit-for-tat TT. Mixed situations CD, CT, and DT are represented by white areas. The results are reported for the Theta band (4-7 Hz).

in the 30% (resp. 34%) of couples, while the modularity and the divisibility are minimal for TT and CC with similar percentages. Analogous results were observed in all the other frequency bands.

In other words, the relationship between the brains of two-defector couples (DD) decreases significantly (i.e. the ROIs of the two brains are better separated) with respect to two-cooperator (CC) couples or tit-for-tat couples (TT). The average Z-scores computed for the three graph measures give a clearer picture of the relations between strategies across the couples. They are reported in Fig. 7.6, which provides a compact visualization of the results obtained for different frequency bands. As illustrated by the figure, DD hyper-brain networks are well separated from networks corresponding to other strategies. In particular, the four points relative to the DD strategy cluster together at the upper-left corner of the panel (a), indicating a relatively high divisibility and, at the same time, a relatively low efficiency with respect to the other hyper-brain networks of the same couple. In addition, the four DD points in panel (b) cluster together at the upper-right region revealing that the DD hyper-brain network modularity

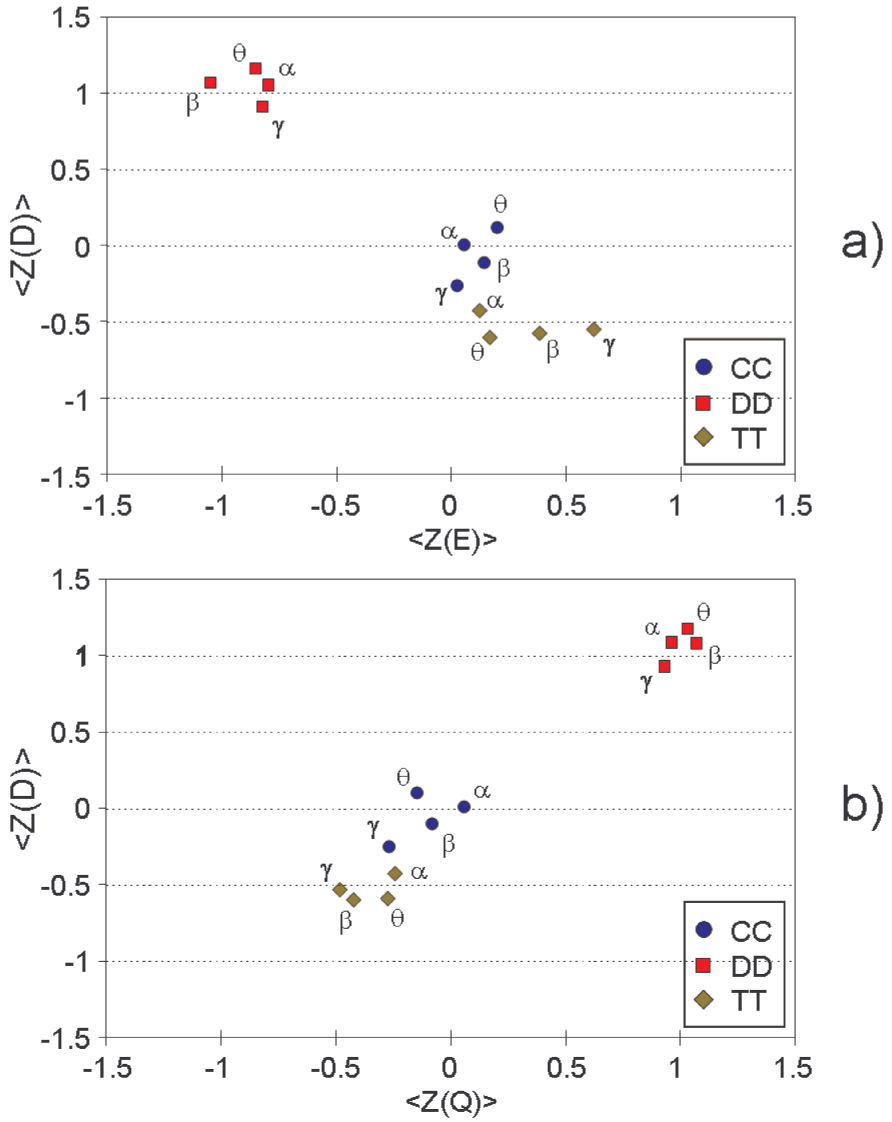

Figure 7.6: **Scatter plot of efficiency E, divisibility D and modularity Q during cooperation (CC), defection (DD) and tit-for-tat (TT).** For each couple x, and each strategy τ, the Z-scores are computed as in formula (7.7). Then $\langle Z_\tau(x)\rangle$ is evaluated as an average of $\langle Z^k_\tau(x)\rangle$ over all the 26 couples. For each strategy, and each frequency band, we report in panel (a), the average Z-score for the measure of divisibility, $\langle Z_\tau(D)\rangle$, vs. the average Z-score of the efficiency, $\langle Z_\tau(E)\rangle$, and in panel (b), the average Z-score of divisibility, $\langle Z_\tau(D)\rangle$, vs. the average Z-score of the modularity, $\langle Z_\tau(Q)\rangle$. Red squares represent DD values; blue circles represent CC values and green diamonds TT values. The Greek letter next to each symbol indicates the considered frequency band.

is usually higher than the modularity of TT or CC connectivity patterns.

7.5.3 On-line classification

Hyper-brain networks corresponding to a given couple's DD strategy have peculiar topological features, such as lower efficiency, higher divisibility and higher modularity with respect to hyper-brains corresponding to the other strategies of the same couple. Such differences can be exploited in order to make predictions on the strategy that a player is going to adopt, based on the on-line analysis of hyper-brain networks constructed from data recorded in the decision-making process. For each frequency band, we have implemented a non-linear classifier, more specifically a Multi-Layer Perceptron, using 21 couples for training (6 networks per couple, each graph corresponding to one of the 6 different strategies, for a total number of 126 networks), and the remaining 5 couples (30 networks in total) for validation. The classification is based on the values of the Z-scores of efficiency, divisibility and modularity, and not on the actual values of the measures themselves. In fact, for each couple, the Z-score of a graph measure provides its deviation from the average value computed over all hyper-brains of the same couple. The accuracies obtained by the classifiers during the validation process, i.e. the number of hyper-brain networks classified correctly as DD or non-DD out of the 30 validation patterns, are respectively: 27, 22, 26, 24 for the Theta, Alpha, Beta and Gamma frequency band.

Since EEG recordings provide high temporal resolution, they can be used in real-time for the construction of the hyper-brain networks, the relative computation of graph measures, and the on-line prediction of the outcome for each trial of the game. In particular, all the parameters needed for source reconstruction, signal ROI estimation and PDC computing can be obtained before the actual EEG session. For instance, they could be obtained in a training session during which the players learn how to play the game, or during a rest condition where the two players are exposed to the same environment that they will experience later. In such a way, all the computations can be reduced basically to a sequence of matrix multiplications. The results presented here indicate that a non-linear classifier is able to discriminate the DD strategy with up to 90% of accuracy. Therefore, the proposed classification process is able to predict the defection strategy of the two players before they press the keyboard buttons to communicate their choices. In principle, a similar approach can be used to train non-linear classifiers to predict CC and TT strategies as well. Such an extension would probably require only a larger dataset, i.e. more than 26 couples.

7.6 Conclusion

Neuroimaging techniques have recently provided strong evidence of a close link between mind and brain. It is well known that the action of concentrating on a specific object or performing a given sensory, cognitive or motor task is reflected in different patterns of brain activity. However, it is not clear whether the decoding of mental states, or brain reading [161, 162], i.e. inferring what an individual is thinking from his brain activity,

can be practically achieved with current neuroimaging methods. The task becomes even harder if one wants to identify neural patterns corresponding to social interactions, such as the choice to cooperate or to defect in the Iterated Prisoner's Dilemma. Results reported in this chapter show quantitatively that the non-cooperative behavior of a pair of players is usually associated with peculiar brain connectivity patterns, and in general with a much lower interaction between the activities of the cortical areas of the two players. The DD hyper-brain network is radically different from the other pure strategies (CC and TT), in which the selected cortical regions of the two players are highly interconnected. In fact, there are only a few inter-brain links in the DD case, giving simultaneously a "picture" and a physical interpretation of the selfish behavior of the subjects. Each player in the couple tends to maximize his own outcome and to minimize at the same time the opponent's outcome. This evidence is coded in the hyper-brain network: cooperation requires areas corresponding to the two brains to be intermingled, while cortical areas of selfish players are almost uncoupled. This outcome indicates the possibility of "reading" mental states, and inferring social behavior from the brain activity of couples of individuals. In particular, these results suggest that:

i) with current neuroimaging techniques, it is possible to estimate in healthy subjects patterns of functional connectivity between cortical areas, which are active in decision-making processes. In the specific case of cooperation or defection strategies in social games, such patterns appear to be linked to the decisions that were made successively by the subjects, and cannot be confused with normal cerebral activity. That is because the operative conditions for the subjects are unchanged during the whole experiment.

ii) the patterns of functional connectivity among cortical areas sub-serving the decision of cooperating or defecting, estimated from data recorded in the decision-making process, produce different hyper-brain networks for different observed outcomes of the game. In particular, for all the frequency bands analyzed, the level of connectivity between the ROIs of the two brains significantly decreases in the case of DD strategies, while hyper-brain networks of TT and CC trials are more tightly connected and intermingled.

In conclusion, we have presented an application of complex network theory to the analysis of functional brain connectivity and to the study of its correlation with observed social behaviors. Indeed, many of the theoretical results obtained in the last few years in the field of complex networks are still waiting to be exploited in the field of neuroscience, and could potentially give us a better insight into the structure and meaning of complex biological systems, as they have already done with social and technological networks. The fact that graph theoretical indexes can also be used to better understand how the human brain works [9, 157, 163–165], suggests that hyper-brain networks can be adopted in the near future as a valuable reference model for further investigations of the mechanisms that are the bases of social empathy [166].

The experiments were conducted by the Neuroelectrical Imaging and Brain Computer Interface laboratory (NEILab) at the Scientific Institute for Research, Hospitalization and Health Care, "Fondazione Santa Lucia" in Rome (Italy) and by the Department of Biomedical Engineering in Minneapolis (USA). All the subjects involved in the experiment were recruited by advertisement. Written informed consent was obtained from each subject after the explanation of the study, which was approved by the local institutional ethics committee of the Scientific Institute for Research, Hospitalization and Health Care, "Fondazione Santa Lucia" in Rome and by the Institutional Review Board of the University of Minnesota.

Conclusion

I have seen too much not to know that the impression of a woman may be more valuable than the conclusion of an analytical reasoner.

Arthur Conan Doyle

In this thesis we have presented several novel results on the study of structure and dynamics of complex networks, on how to construct networks encoding non-trivial statistical features of real-world data, on the characterization of human mobility patterns, and on the relation between functional brain networks obtained during cooperative games and selfish behavior.

As regards the structure of networks, we have studied how to measure exactly an additional order of degree correlations in networks, namely three-body degree correlations. To do this, we used a third-order Markov model. Counterintuitively, we showed that these correlations are not negligible in respect to the two-body ones. In fact, the effects of three-body degree correlations on many topological measures are comparable to those of the correlations between pairs of node degrees. We have shown, for example, that in a wide range of real networks three-body degree correlations (i) alter considerably the average connectivity of the second neighbors of a node of degree k in respect to the expectation given by two-body degree correlations, and (ii) are also responsible for the rich-club phenomenon.

Successively, we have investigated an important kind of dynamics on graphs, the so-called Biased Random Walks (BRW), a class of markovian stochastic processes which can be treated analytically and which extend the well-known concept of Random Walk (RW) on a network. In particular, we investigated the connection between correlations in the connectivity patterns of the network and the entropy rate that can be associated to the BRWs. We have demonstrated that it is possible to design maximal-entropy random walks with only local information on the graph structure, according to the order of the correlations present in the graph. We also showed how it is possible to rephrase a BRW process on a network as a plain RW on another network having the same topology but different weights associated to the edges. We name this network flow graph, since it embeds in its link the dynamical flows of the original graph. The concept of flow graph is useful in many applications, as in community detection algorithms, and for deriving many theoretical results.

Regarding the applications of complex network theory and of information theory to real datasets, we have introduced and developed a method to convert ensembles of sequences of symbols into weighted directed networks whose nodes are motifs, while the directed links and their weights are defined from statistically significant co-occurences of two motifs in the same sequence. To our knowledge, this is also the first method in literature that allows to construct a network encoding information about short- and long-range correlations of a complex system. We have then applied the networks of motifs method to the study of the human proteome database, to detect hot topics from online social dialogs, and to characterize trajectories of dynamical systems.

We have used concepts of complex networks and of information theory also for the study of mobility of human players exploring a network of a virtual world. After characterizing the basic features of the motion, such as waiting time and trip length distributions, we have focused on the analysis of the diffusion properties of player movements. We have showed that the players' trajectories are highly subdiffusive, exhibit long-time memory, and that it is necessary to incorporate in a model the information about the order of locations an agent visits to recover the correct scaling properties of the diffusion of players. We also investigated how socio-economic factors influence the trajectories of players. We found that players significantly avoid to cross borders between communities, and prefer locations that are within the community even if at high distance.

Finally, we have analyzed the structure of the functional brain networks derived from EEG recorded during cooperative games. We have first introduced the approach relying on the definition of hyperbrain, a network whose connectivity patterns represent at once the correlation of EEG signals among the cortical regions of a single brain as well as the correlations among the areas of the brains of two distinct individuals that are socially interacting. Then, we have studied the structural properties of hyperbrain networks of pairs of individuals playing an Iterated Prisoner's Dilemma, and we found that networks of two-defector couples have significantly less inter-brain links and overall higher modularity i.e., the tendency to form two separate subgraphs, than couples playing cooperative or tit-for-tat strategies. Furthermore, we found that graph analysis of the hyperbrain obtained during the decision making process allows to predict in advance the defection of a player.

Acknowledgements

Gratefulness is the most exquisite form of courtesy.

François de la Rochefoucauld

What I have learnt, experienced, acquired during these the years as PhD student, what I have become from a "professional" point of view, cannot be summarized in a publication list or in the compilation of a dissertation. During the last three years many people have directly or indirectly helped me to reach this goal, and, even more important, they influenced the way I came to it.

My first and deepest thanks go to Prof. Vito Latora: he has been for me much more than a supervisor. He has been a continuos source of inspiration and motivation, a constant but never intrusive guide, an example to refer to in each situation. Thanks, Vito, for teaching me so much and, at the same time, for treating me not like a student, but more like a young colleague. And, above all, thanks for your friendship: this is what I value most of all the time spent together. Sincere gratitude goes to my second "model", Jesús Gómez-Gardeñes, whose role in my career went much more beyond the co-supervision of this work. He constantly transmitted me enthusiasm for this job, and had always the right advise to make me solve a problem, major or minor. With him I enjoyed great discussions, most of the times over a beer and a *panino con polpetta*. I would also like to thank Enzo Nicosia, for all the precious suggestions about all the possible computer and computational issues (computers and numerics have no secret for him), for all the cups of tea drunk together and, yes, also for all the animated discussions we had together at the blackboard (this is also healthy!).

Many of the results I have presented in this thesis or achieved in my career would not be there without the priceless contribution of all the people I collaborated with during these years: thanks then to Prof. Daniele Condorelli, Dr. Fabrizio De Vico Fallani, Prof. Henrik Jensen, Prof. Renaud Lambiotte, Dr. Yamir Moreno, Dr. Hugo Touchette, Prof. Stefan Thurner. A special mention goes also to Sandro Meloni and Giovanni Petri, much more than simple collaborators, rather unique friends the PhD years have gifted me. Along with my collaborators, I am grateful for the hospitality I received during the months spent abroad in various academical institutions, which I would like to mention: BIFI, Imperial College London, Queen Mary University London,

Medical University of Vienna and the Santa Fe Institute. A special mention goes to those persons that have constantly promoted me, even long before I was a PhD student: Prof. Giuseppe Angilella, Prof. Giovanni Piccitto, Prof. Emanuele Rimini. Thank you, I have always appreciated your belief in my skills.

A special spot is deserved by Federico and Emanuele. They have covered a fundamental role in the last years of my life: not only unconditioned friends, with whom I spent most enjoyable hours, but also junior scientists I always referred to, asked for advice, considered as point of reference.

My PhD has also given me the opportunity to meet wonderful people, or to stay in touch with others with whom ties would have been to weak: Giovanna, Paul, Nicky, Salvo S., Tassos. I enjoyed very much the time spent with you around Europe and USA, attending conferences, meetings, workshops, schools, etc. etc. Immense hugs go instead to Alessio and Maria: office hours would have been incredibly much longer without their presence, and after working hours much more boring. Also during the time spent working at the Department of Physics in Catania, I could not have had more lively lunch breaks without our "pranzo gang": thanks Maria, Lucia, Isodiana, Elena, Giorgia, Paolo, Pietro, Gabriele, Giuseppe, Massimiliano, Daniele, Stefano, Ciccio P,. Salvo C.. And sorry if I forgot to mention someone! Impossible to forget are my friends Filippo, Marinella, Daniele, Alessia, Alessandra, Rachele, Eugenio: for the laughing, the hugs, the fun.

My thoughts go to my wonderful family, Mutti, Claudia and Emma. They know me better than anyone else on the Earth. They understand me, they have always got interested in everything I do. And, most importantly, they always granted me unquestionable support. Thank you!

Last but not least, it was during my first year as a PhD when I travelled to a Complex Systems Summer School in Budapest, where I had the most unexpected and invaluable meeting these years could have reserved me: Michi. Since then, he stands by me and walks with me along the way. He laughs with me and comforts me. We marvel together and get mad together. I cherish all the moments passed with him and cannot wait for all the adventures to come.

I also acknowledge grateful support from Scuola Superiore di Catania, INFN - Sez. di Catania, HPC-EUROPA 2 (Project No. 228398), and European Cooperation in Science and Technology (COST) Action MP0801.

Bibliography

[1] R. Sinatra, D. Condorelli, and V. Latora. Networks of motifs from sequences of symbols. *Physical review letters*, 105(17):178702, 2010.

[2] F. De Vico Fallani, V. Nicosia, R. Sinatra, L. Astolfi, F. Cincotti, D. Mattia, C. Wilke, A. Doud, V. Latora, B. He, et al. Defecting or not defecting: How to "read" human behavior during cooperative games by eeg measurements. *PloS one*, 5(12):e14187, 2010.

[3] R. Sinatra, J. Gómez-Gardeñes, R. Lambiotte, V. Nicosia, and V. Latora. Maximal-entropy random walks in complex networks with limited information. *Physical Review E*, 83(3):030103(R), 2011.

[4] R. Lambiotte, R. Sinatra, J.C. Delvenne, TS Evans, M. Barahona, and V. Latora. Flow graphs: interweaving dynamics and structure. *Physical Review E*, 84:017102, 2011.

[5] M. Szell, R. Sinatra, G. Petri, S. Thurner, and V. Latora. Understanding mobility in a social petri dish. *Arxiv preprint 1112.1220v1*, 2012.

[6] R. Sinatra, J. Gómez-Gardeñes, S. Meloni, V. Nicosia, and V. Latora. Quantifying three-body degree correlations in complex networks. *in review*, 2012.

[7] S. Thurner, M. Szell, and R. Sinatra. Emergence of good conduct, scaling and zipf laws in human behavioral sequences in an online world. *in press in PLoS one, arXiv:1107.0392v1*, 2011.

[8] R. Sinatra, J. Iranzo, J. Gómez-Gardeñes, L.M. Floría, V. Latora, and Y. Moreno. The ultimatum game in complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2009:P09012, 2009.

[9] R. Sinatra, F. De Vico Fallani, L. Astolfi, F. Babiloni, F. Cincotti, V. Latora, and D. Mattia. Cluster structure of functional networks estimated from high-resolution eeg data. *International Journal of Bifurcation and Chaos*, 19:665–676, 2009.

[10] M.H. Zaman and Cambridge University Press. *Statistical mechanics of cellular systems and processes*. Cambridge University Press, 2009.

[11] F. De Vico Fallani, R. Sinatra, L. Astolfi, D. Mattia, F. Cincotti, V. Latora, S. Salinari, MG Marciani, A. Colosimo, and F. Babiloni. Community structure of cortical networks in spinal cord injured patients. In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pages 3995–3998. IEEE, 2008.

[12] B. Bollobás. *Modern graph theory*, volume 184. Springer Verlag, 1998.

[13] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.U. Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4-5):175–308, 2006.

[14] M. Newman. *Networks: an introduction*. Oxford Univ Pr, 2010.

[15] M. Boguná and R. Pastor-Satorras. Epidemic spreading in correlated complex networks. *Physical Review E*, 66(4):047104, 2002.

[16] R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Dynamical and correlation properties of the internet. *Physical Review Letters*, 87(25):258701, 2001.

[17] M.E.J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, 2002.

[18] S. Wasserman. *Social network analysis: Methods and applications*. Cambridge university press, 1994.

[19] M.E.J. Newman. The structure and function of complex networks. *SIAM review*, pages 167–256, 2003.

[20] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[21] S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[22] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, pages 425–443, 1969.

[23] V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Physical Review Letters*, 87(19):198701, 2001.

[24] V. Latora and M. Marchiori. Economic small-world behavior in weighted networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 32(2):249–263, 2003.

[25] R. Albert and A.L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[26] A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[27] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[28] A.J. Enright, S. Van Dongen, and C.A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30(7):1575, 2002.

[29] P. Erdös and A. Rényi. On random graphs, i. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

[30] E.A. Bender and E.R. Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307, 1978.

[31] D.R. Cox and H.D. Miller. *The theory of stochastic processes*, volume 134. Chapman and Hall, 1977.

[32] C.D. Godsil, G. Royle, and CD Godsil. *Algebraic graph theory*, volume 8. Springer New York, 2001.

[33] K. Huang. *Statistical Mechanics*. John Wiley Inc, 1963.

[34] T.M. Cover and J.A. Thomas. *Elements of information theory*. Wiley Online Library, 1991.

[35] T. Tanaka. *Methods of statistical physics*. Cambridge Univ Pr, 2002.

[36] S. Zhou and R.J. Mondragón. The rich-club phenomenon in the internet topology. *Communications Letters, IEEE*, 8(3):180–182, 2004.

[37] V. Colizza, A. Flammini, M.A. Serrano, and A. Vespignani. Detecting rich-club ordering in complex networks. *Nature Physics*, 2(2):110–115, 2006.

[38] T. Opsahl, V. Colizza, P. Panzarasa, and J.J. Ramasco. Prominence and control: The weighted rich-club effect. *Physical review letters*, 101(16):168702, 2008.

[39] M. Boguñá and R. Pastor-Satorras. Class of correlated random networks with hidden variables. *Physical Review E*, 68(3):036112, 2003.

[40] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM, 2005.

[41] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas. Self-similar community structure in a network of human interactions. *Physical Review E*, 68(6):065103, 2003.

[42] M.E.J. Newman. Scientific collaboration networks. i. network construction and fundamental results. *Physical Review E*, 64(1):016131, 2001.

[43] A.L. Barabási, R. Albert, and H. Jeong. The diameter of the world wide web. *Nature*, 401(9):130–131, 1999.

[44] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas. Models of social networks based on social distance attachment. *Physical Review E*, 70(5):056122, 2004.

[45] V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Physical Review Letters*, 87(19):198701, 2001.

[46] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910, 2002.

[47] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, et al. Topological structure analysis of the protein–protein interaction network in budding yeast. *Nucleic Acids Research*, 31(9):2443, 2003.

[48] M. Szell and S. Thurner. Measuring social dynamics in a massive multiplayer online game. *Social Networks*, 32(4):313 – 329, 2010.

[49] P. Gleiser and L. Danon. Community structure in jazz. *Advances in Complex Systems*, 6(4):565–573, 2003.

[50] R. Guimera and L.A.N. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.

[51] P. Blanchard and D. Volchenkov. *Random Walks and Diffusions on Graphs and Databases: An Introduction*. Springer Pub Co, 2011.

[52] J.D. Noh and H. Rieger. Random walks on complex networks. *Physical review letters*, 92(11):118701, 2004.

[53] S. Condamin, O. Bénichou, V. Tejedor, R. Voituriez, and J. Klafter. First-passage times in complex scale-invariant media. *Nature*, 450(77), 2007.

[54] A. Fronczak and P. Fronczak. Biased random walks in complex networks: The role of local navigation rules. *Physical Review E*, 80(1):016107, 2009.

[55] D. Gfeller and P. De Los Rios. Spectral coarse graining of complex networks. *Physical review letters*, 99(3):38701, 2007.

[56] V. Zlatić, A. Gabrielli, and G. Caldarelli. Topologically biased random walk as a tool for understanding graph spectra. *Arxiv preprint 1003.1883v1*, 2010.

[57] J.C. Delvenne, SN Yaliraki, and M. Barahona. Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences*, 107(29):12755, 2010.

[58] M. Chavez, M. Valencia, V. Navarro, V. Latora, and J. Martinerie. Functional modularity of background activities in normal and epileptic brain networks. *Physical review letters*, 104(11):118701, 2010.

[59] L. da Fontoura Costa, O. Sporns, L. Antiqueira, M.G.V. Nunes, and O.N. Oliveira Jr. Correlations between structure and random walk dynamics in directed complex networks. *Applied Physics Letters*, 91:054107, 2007.

[60] S. Lee, S.H. Yook, and Y. Kim. Centrality measure of complex networks using biased random walks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 68(2):277–281, 2009.

[61] W. Parry. Intrinsic markov chains. *Trans. Amer. Math. Soc*, 112(1):55–65, 1964.

[62] L. Demetrius and T. Manke. Robustness and network evolution–an entropic principle. *Physica A: Statistical Mechanics and its Applications*, 346(3-4):682–696, 2005.

[63] J.C. Delvenne and A.S. Libert. Centrality measures and thermodynamic formalism for complex networks. *Physical Review E*, 83(4):046117, 2011.

[64] Z. Burda, J. Duda, JM Luck, and B. Waclaw. Localization of the maximal entropy random walk. *Physical review letters*, 102(16):160602, 2009.

[65] J. Gómez-Gardeñes and V. Latora. Entropy rate of diffusion processes on complex networks. *Physical Review E*, 78(6):065102, 2008.

[66] V. Colizza, R. Pastor-Satorras, and A. Vespignani. Reaction-diffusion processes and metapopulation models in heterogeneous networks. *Nature Physics*, 3(276), 2007.

[67] N. Fujiwara, J. Kurths, and A. Díaz-Guilera. Synchronization in networks of mobile oscillators. *Physical Review E*, 83(2):025101, 2011.

[68] P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1):113–120, 1972.

[69] J.G. Restrepo, E. Ott, and B.R. Hunt. Approximating the largest eigenvalue of network adjacency matrices. *Physical Review-Section E-Statistical Nonlinear and Soft Matter Physics*, 76(5):56119–56119, 2007.

[70] R.F. i Cancho, R.V. Solé, and R. Köhler. Patterns in syntactic dependency networks. *Physical Review E*, 69(5):051915, 2004.

[71] R.F. i Cancho and R.V. Solé. The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1482):2261, 2001.

[72] A.E. Motter, A.P.S. de Moura, Y.C. Lai, and P. Dasgupta. Topology of the conceptual network of language. *Physical Review E*, 65(6):065102, 2002.

[73] E.G. Altmann, J.B. Pierrehumbert, and A.E. Motter. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS One*, 4(11):e7678, 2009.

[74] D.B. Searls. The language of genes. *Nature*, 420(6912):211–217, 2002.

[75] V. Brendel, J.S. Beckmann, and E.N. Trifonov. Linguistics of nucleotide sequences: morphology and comparison of vocabularies. *Journal of biomolecular structure & dynamics*, 4(1):11, 1986.

[76] CK Peng, SV Buldyrev, AL Goldberger, S. Havlin, F. Sciortino, M. Simons, and HE Stanley. Long-range correlations in nucleotide sequences. *Nature*, 356(6365):168–170, 1992.

[77] N. Scafetta, V. Latora, and P. Grigolini. Lévy scaling: The diffusion entropy analysis applied to dna sequences. *Physical Review E*, 66(3):031906, 2002.

[78] V. Rosato, N. Pucello, and G. Giuliano. Evidence for cysteine clustering in thermophilic proteomes. *Trends in Genetics*, 18(6):278–281, 2002.

[79] H.J. Bussemaker, H. Li, and E.D. Siggia. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proceedings of the National Academy of Sciences*, 97(18):10096, 2000.

[80] Z. Solan, D. Horn, E. Ruppin, and S. Edelman. Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11629, 2005.

[81] C. Beck and F. Schlögl. *Thermodynamics of chaotic systems: an introduction*, volume 4. Cambridge Univ Press, 1995.

[82] L. Lacasa, B. Luque, F. Ballesteros, J. Luque, and J.C. Nuño. From time series to complex networks: The visibility graph. *Proceedings of the National Academy of Sciences*, 105(13):4972, 2008.

[83] B. Tadić and M. Mitrović. Jamming and correlation patterns in traffic of information on sparse modular networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 71(4):631–640, 2009.

[84] E. Bradley, D. Capps, J. Luftig, J.M. Stuart, W.J. Hwang, C.M. Ou, P.C. Hung, C.Y. Yang, T.H. Yu, V.G. Edupuganti, et al. Towards stylistic consonance in human movement synthesis. *Open Artificial Intelligence Journal*, 4:1–19, 2010.

[85] L. Ferraro, A. Giansanti, G. Giuliano, and V. Rosato. Co-expression of statistically over-represented peptides in proteomes: a key to phylogeny? *Arxiv preprint q-bio/0410011*, 2004.

[86] A. Giansanti, M. Bocchieri, V. Rosato, and S. Musumeci. A fine functional homology between chitinases from host and parasite is relevant for malaria transmissibility. *Parasitology research*, 101(3):639–645, 2007.

[87] U. Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007.

[88] M. Caselle, F. Di Cunto, and P. Provero. Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes. *BMC bioinformatics*, 3(1):7, 2002.

[89] D. Corà, F. Di Cunto, P. Provero, L. Silengo, and M. Caselle. Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrep-resented upstream motifs. *BMC bioinformatics*, 5(1):57, 2004.

[90] P. Nicodème, T. Doerks, and M. Vingron. Proteome analysis based on motif statistics. *Bioinformatics*, 18(suppl 2):S161, 2002.

[91] Consensus Coding Sequence database. http://www.ncbi.nlm.nih.gov/ccds/.

[92] ExPASy PROSITE. http://www.expasy.ch/prosite.

[93] Leet Language. http://en.wikipedia.org/wiki/leet.

[94] K. Lerman and R. Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, 2010.

[95] M. Cha, A. Mislove, B. Adams, and K.P. Gummadi. Characterizing social cascades in flickr. In *Proceedings of the first workshop on Online social networks*, pages 13–18. ACM, 2008.

[96] Twitter. www.twitter.com.

[97] Twitter API. http://apiwiki.twitter.com/streaming-api-documentation.

[98] http://en.wikipedia.org/wiki/benjamin_cohen_%28journalist%29.

[99] http://twitpic.com/1jge7b.

[100] http://www.guardian.co.uk/politics/2010/jul/26/gillian-duffy-backs-david-miliband.

[101] http://wiki.electorama.com/wiki/tactical_voting.

[102] B.V. Chirikov. A universal instability of many-dimensional oscillator systems. *Physics reports*, 52(5):263–379, 1979.

[103] V.M. Aleksev and M.V. Jakobson. Symbolic dynamics and hyperbolic dynamical systems. *Physics Reports*, 75:287–325, 1981.

[104] M. Barthélemy. Spatial networks. *Physics Reports*, 499:1–101, 2010.

[105] R. Guimerà, S. Mossa, A. Turtschi, and L.A.N. Amaral. The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. *Proc. Natl. Acad. Sci. (USA)*, 102:7794–7799, 2005.

[106] D. Helbing. Traffic and related self-driven many-particle systems. *Reviews of modern physics*, 73(4):1067, 2001.

[107] H. A. Makse, S. Havlin, and H. E. Stanley. Modelling urban growth patterns. *Nature*, 377:608–612, 1995.

[108] Camille Roth, Soong Moon Kang, Michael Batty, and Marc Barthélemy. Structure of urban movements: Polycentric activity and entangled hierarchical flows. *PLoS ONE*, 6(1):e15923, 01 2011.

[109] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200–3203, 2001.

[110] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences*, 103(7):2015, 2006.

[111] L. Hufnagel, D. Brockmann, and T. Geisel. Forecast and control of epidemics in a globalized worlds. *Proc. Natl Acad. Sci. USA*, 101:15124–15129, 2004.

[112] D. Balcan, V. Colizza, B. Goncalves, H. Hu, J.J. Ramasco, and A. Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, 2009.

[113] G. Miritello, E. Moro, and R. Lara. Dynamical strength of social ties in information spreading. *Physical Review E*, 83(4):045102, 2011.

[114] J.P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.L. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332, 2007.

[115] D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft. Recommending social events from mobile phone location data. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 971 –976, dec. 2010.

[116] P. Jensen. Network-based predictions of retail store commercial categories and optimal locations. *Physical Review E*, 74(3):035101, 2006.

[117] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[118] C. Thiemann, F. Theis, D. Grady, R. Brune, and D. Dirk Brockmann. The structure of borders in a small world. *PLoS one*, 5:e15422, 2010.

[119] M.C. González, C.A. Hidalgo, and A.L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[120] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. *Proceedings of ICWSM*, 11, 2011.

[121] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. Campbell. Nextplace: A spatio-temporal prediction framework for pervasive systems. *Pervasive Computing*, pages 152–169, 2011.

[122] A. Bazzani, B. Giorgini, S. Rambaldi, R. Gallotti, and L. Giovannini. Statistical laws in urban mobility from microscopic gps data in the area of florence. *Journal of Statistical Mechanics: Theory and Experiment*, 2010:P05001, 2010.

[123] C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J.F. Pinton, and A. Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PloS one*, 5(7):e11596, 2010.

[124] C. Song, T. Koren, P. Wang, and A. Barabasi. Modelling the scaling properties of human mobility. *Nature Physics*, 6:818–823, 10 2010.

[125] C. Song, Z. Qu, N. Blumm, and A.L. Barabasi. Limits of predictability in human mobility. *Science*, 327(5968):1018, 2010.

[126] V. Belik, T. Geisel, and D. Brockmann. Natural human mobility patterns and spatial spread of infectious diseases. *Physical Review X*, 1:011001, 2011.

[127] M. Szell, R. Lambiotte, and S. Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31):13636–13641, 2010.

[128] E Castronova. On the research value of large games. *Games and Culture*, 1:163–186, 2006.

[129] W.S. Bainbridge. The scientific research potential of virtual worlds. *Science*, 317(5837):472, 2007.

[130] www.pardus.at.

[131] R. Kölbl and D. Helbing. Energy laws in human travel behaviour. *New Journal of Physics*, 5:48, 2003.

[132] X.P. Han, Q. Hao, B.H. Wang, and T. Zhou. Origin of the scaling law in human mobility: Hierarchy of traffic systems. *Phys. Rev. E*, 83(3):036117, Mar 2011.

[133] Albert-László Barabási. The origin of bursts and heavy tails in humans dynamics. *Nature*, 435:207, Apr 2005.

[134] A Arenas, A Fernández, and S Gómez. Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics*, 10(5):053039, May 2008.

[135] M.E.J. Newman. Analysis of weighted networks. *Physical Review E*, 70(5):056131, 2004.

[136] Thomas Richardson, Peter Mucha, and Mason Porter. Spectral tripartitioning of networks. *Physical Review E*, 80(3), September 2009.

[137] BW Kernighan and S Lin. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, pages 291–307, 1970.

[138] E.B. Fowkes, C.L. Mallows, and E.B. Fowlkes. *A method for comparing two hierarchical algorithms*, volume 78. J. Amer. Statist. Assoc, 1983.

[139] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):pp. 846–850, 1971.

[140] Thomas M Cover and Joy A Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2 edition, July 2006.

[141] Marina Meila. Comparing clusterings–an information based distance. *Journal of Multivariate Analysis*, 98(5):873 – 895, 2007.

[142] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner. On finding graph clusterings with maximum modularity. In *Graph-Theoretic Concepts in Computer Science*, pages 121–132. Springer, 2007.

[143] R. Metzler and J. Klafter. The random walk's guide to anomalous diffusion: a fractional dynamics approach. *Physics Reports*, 339:1–77, 2000.

[144] B.J. West, P. Grigolini, R. Metzler, and T.F. Nonnenmacher. Fractional diffusion and levy stable processes. *Physical Review E*, 55(1):99, 1997.

[145] GM Viswanathan, S.V. Buldyrev, S. Havlin, MGE Da Luz, EP Raposo, and H.E. Stanley. Optimizing the success of random searches. *Nature*, 401(6756):911–914, 1999.

[146] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S.H. Strogatz. Redrawing the map of Great Britain from a network of human interactions. *PLoS One*, 5(12):e14248, 2010.

[147] D. Newman. The lines that continue to separate us: borders in ourborderless' world. *Progress in Human Geography*, 30(2):143, 2006.

[148] R. Lambiotte, V.D. Blondel, C. de Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, 2008.

[149] M.J. Osborne and A. Rubinstein. *A course in game theory*. The MIT press, 1994.

[150] D. Lee. Game theory and neural basis of social decision making. *Nature neuroscience*, 11(4):404–409, 2008.

[151] M.A. Nowak. *Evolutionary dynamics: exploring the equations of life*. Harvard University Press, 2006.

[152] R. Axelrod and W.D. Hamilton. The evolution of cooperation. *Science*, 211(4489):1390, 1981.

[153] C. Babiloni, F. Babiloni, F. Carducci, S.F. Cappa, F. Cincotti, C. Del Percio, C. Miniussi, D.V. Moretti, S. Rossi, K. Sosta, et al. Human cortical responses during one-bit short-term memory. a high-resolution eeg study on delayed choice reaction time tasks. *Clinical neurophysiology*, 115(1):161–170, 2004.

[154] F. Babiloni, F. Carducci, F. Cincotti, C. Del Gratta, GM Roberti, GL Romani, PM Rossini, and C. Babiloni. Integration of high resolution eeg and functional magnetic resonance in the study of human movement-related potentials. *Methods of information in medicine*, 39(2):179–182, 2000.

[155] C. Babiloni, A. Brancucci, F. Babiloni, P. Capotosto, F. Carducci, F. Cincotti, L. Arendt-Nielsen, A.C.N. Chen, and P.M. Rossini. Anticipatory cortical responses during the expectancy of a predictable painful stimulation. a high-resolution electroencephalography study. *European Journal of Neuroscience*, 18(6):1692–1700, 2003.

[156] K. Brodmann. Vergleichende lokalisationslehre der grobhirnrinde. *Barth, Leipzig*, 1909.

[157] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci*, 10(3):186–198, 2009.

[158] L.A. Baccala and K. Sameshima. Partial directed coherence: a new concept in neural structure determination. *Biological cybernetics*, 84(6):463–474, 2001.

[159] M.E.J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104, 2006.

[160] A. Arenas, J. Duch, A. Fernández, and S. Gómez. Size reduction of complex networks preserving modularity. *New Journal of Physics*, 9:176, 2007.

[161] J.D. Haynes, K. Sakai, G. Rees, S. Gilbert, C. Frith, and R.E. Passingham. Reading hidden intentions in the human brain. *Current Biology*, 17(4):323–328, 2007.

[162] J.D. Haynes and G. Rees. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523–534, 2006.

[163] M. Chavez, M. Valencia, V. Navarro, V. Latora, and J. Martinerie. Functional modularity of background activities in normal and epileptic brain networks. *Physical review letters*, 104(11):118701, 2010.

[164] D. Meunier, R. Lambiotte, A. Fornito, K.D. Ersche, and E.T. Bullmore. Hierarchical modularity in human brain functional networks. *Frontiers in neuroinformatics*, 3, 2009.

[165] F. De Vico Fallani, L. Astolfi, F. Cincotti, D. Mattia, A. Tocci, M.G. Marciani, A. Colosimo, S. Salinari, S. Gao, A. Cichocki, et al. Extracting information from cortical connectivity patterns estimated from high resolution eeg recordings: A theoretical graph approach. *Brain topography*, 19(3):125–136, 2007.

[166] M.B. Schippers, A. Roebroeck, R. Renken, L. Nanetti, and C. Keysers. Mapping the information flow from one brain to another during gestural communication. *Proceedings of the National Academy of Sciences*, 107(20):9388, 2010.