Massimiliano Zanin

Licenciado

Complex Networks and Data Mining: Toward a new perspective for the understanding of Complex Systems

Dissertação para obtenção do Grau de Doutor em
Engenharia Electrotécnica e de Computadores

Orientadores : Pedro Sousa, Professor Doutor, Universidade Nova de Lisboa
Stefano Boccaletti, Senior scientist, CNR - Institute of Complex Systems, Florence, Italy

Júri:

Presidente: Prof. Camarinha da Matos

Arguentes: Prof. João Paulo Branquinho Pimentão
Prof. Ruedi Stoop

Vogais: Prof. Ginestra Bianconi
Prof. Ernestina Menesalvas
Prof. Pedro Alexandre da Costa Sousa

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Dezembro, 2014

Complex Networks and Data Mining:
Toward a new perspective for the understanding of Complex Systems

It is a great pleasure, now that this voyage is reaching a conclusion, to look back and remember all the people who helped me through the research work that led to this PhD thesis. Gratitude should clearly go to all those friends and colleagues whose advice and wisdom have positively marked my career; but also to those who have just unintentionally increased my determination to walk through the path of research.

Going back in time, I should start with those people who have guided (and trusted) me through the first steps: Javier Buldú, of the Universidad Rey Juan Carlos in Madrid; Alexander Pisarchik, then at the Center for Optic Research in Mexico; and Francisco Mancebo, of the Universidad Politécnica di Madrid. After these somehow improvised beginnings, things got more serious, with my work at Innaxis and at the Center for Biomedical Technology in Madrid. The latter has been a place where my vocation for research has been confirmed, especially thanks to the positive interactions with excellent professionals like Juan Almendral, Inmaculada Leyva, Adrián Navas, David Papo, Irene Sendiña-Nadal; but also to not-so-positive experiences, which, as many profess, are part and parcel of the academic world.

I should mention a few ones of people with whom I've interacted during these years. Ricardo Sevilla, with whom I shared many a bottle of tequila - but just for research purposes! Joan Serrà and Pedro Cano, who introduced me to the world of music and audio technology; Daniel Ramos, whose skill as a singer is second only to his skill as a researcher; Joaquín Medina and Jesús Vicente, who taught me that plant genetics is easier to understand with Gin Tonic; Regino Criado and Miguel Romance, always ready to explain complex mathematical concepts in simple terms. My fellow reader, if you are not included here, please excuse my horrible memory, as no offence is meant in the omission!

Finally, I can hardly find the words to express my gratitude to the three people who made this Thesis possible: Pedro Sousa, Stefano Boccaletti and Ernestina Menasalvas. Besides teaching me so many things, both actively and by their own example, that a thesis would only be a short summary, they have been instrumental in overcoming all the problems (scientific and bureaucratic) that we have encountered - in spite of my stubborn attempts at ruining my own scientific career. Thanks Pedro, Stefano and Ernestina for demonstrating that these attempts have so far been unsuccessful!

Abstract

Complex systems, *i.e.* systems composed of a large set of elements interacting in a non-linear way, are constantly found all around us. In the last decades, different approaches have been proposed toward their understanding, one of the most interesting being the *Complex Network* perspective. This legacy of the 18th century mathematical concepts proposed by Leonhard Euler is still current, and more and more relevant in real-world problems. In recent years, it has been demonstrated that network-based representations can yield relevant knowledge about complex systems. In spite of that, several problems have been detected, mainly related to the degree of subjectivity involved in the creation and evaluation of such network structures. In this Thesis, we propose addressing these problems by means of different *data mining* techniques, thus obtaining a novel hybrid approximation intermingling complex networks and data mining. Results indicate that such techniques can be effectively used to *i)* enable the creation of novel network representations, *ii)* reduce the dimensionality of analyzed systems by pre-selecting the most important elements, *iii)* describe complex networks, and *iv)* assist in the analysis of different network topologies. The soundness of such approach is validated through different validation cases drawn from actual biomedical problems, *e.g.* the diagnosis of cancer from tissue analysis, or the study of the dynamics of the brain under different neurological disorders.

Keywords: Complex systems, complex networks, data mining.

Resumo

Os sistemas complexos, *i.e.* sistemas compostos por um vasto conjunto de elementos que interagem de forma não linear, são comuns e abundantes. Nas últimas décadas, diferentes aproximações têm sido tentadas com vista a fazer a sua interpretação, sendo que o uso de Redes Complexas é um dos mais eficazes. O legado de conceitos matemáticos propostos por Leonhard Euler (matemático do século XVIII) continuam actuais e confirmam a sua aplicabilidade. Os últimos anos têm confirmado que representações baseadas em redes conseguem descrever conhecimento relevante sobre os sistemas complexos. Contudo estão identificadas diversas limitações, sobretudo relacionadas com a o grau de subjectividade na criação e avaliação das estruturas em rede. Nesta dissertação, abordam-se estes assuntos com o recurso a técnicas de *data mining*, resultando assim numa aproximação híbrida que interliga a aproximação das redes complexas com o *data mining*. Os resultados obtidos indicam que as aproximações sugeridas são eficazes *i)* na criação de novas representações em rede; *ii)* na redução da dimensionalidade dos sistemas analisados pela pré-selecção dos elementos mais relevantes, *iii)* na descrição de redes complexas e *iv)* no suporte à analise de diferentes topologias de redes. A robustez da aproximação foi validada através de diversos estudos de casos da área da biomédica, *e.g.* o diagnóstico do cancro na análise de tecidos, ou no estudo da dinâmica do cérebro afectado por diversas patologias neurológicas.

Palavras-chave: Sistemas complexos, redes complexas, data mining.

Contents

List of Figures

List of Tables

1

Introduction

In the last decades, the scientific community has realized that there are some systems, both natural and manmade, which cannot be fully understood by a reductionist approach, *i.e.* by analyzing their constituting elements in an isolated way. On the contrary, their macroscopic properties seem to be defined by the structures of interactions between these elements. Such systems are now called *complex systems*. Examples have been found in many scientific fields, *e.g.* in social or technical (transport networks, Internet, *et caetera*) contexts. Probably, one of the most astonishing examples of a complex system is the brain; it is composed of more than 100 billions neurons, each one of them showing a very simple dynamics; the human capacity for reasoning (or for writing a PhD Thesis) only emerges when these simple dynamics start to interact.

Nowadays, two are the approaches used to extract information from complex systems (see Figure 1): classical data mining techniques, and complex networks. Born within physics, with substantial inputs from mathematics and statistics, the theory of *complex networks* has proven to be a powerful tool for the analysis of complex systems; it allows reducing them into simple structures of interactions, which can easily be studied by means of mathematical (algebraic) tools, while removing all unnecessary details. Following this idea, some important results have been obtained, as, for instance, the detection of critical genes in an organism [BO04], or the definition of the best strategies to stop the spreading of an infectious disease [PSV01]. On the other hand, *data mining* refers to the process of discovering patterns in large data sets, in order to automatically extract information and transform it into an understandable structure [FPSS96]. Born within computer science, it involves methods drawn from applied mathematics and statistics.

As is schematically represented in Figure 1.1 (Left), both approaches (data mining on one side, complex network analysis on the other side) are now applied independently,

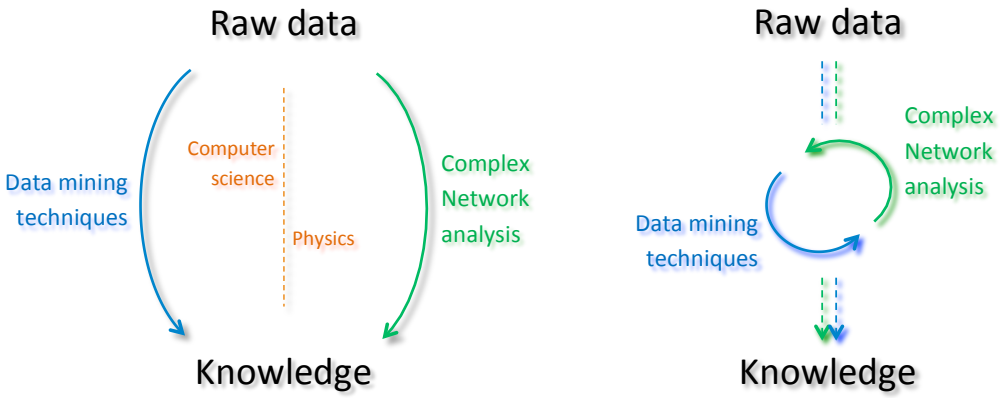


Figure 1.1: **Interaction between complex networks and data mining.** (Left) The nowadays approach to the study of Complex Systems. (Right) Creating interactions between data mining and complex networks, as proposed in this Thesis.

usually by two communities of researchers, *i.e.* computer scientists and physicists, that had little contact in the past. The main challenge is, therefore, to see how the state of the art and the most recent advancements from both communities can be integrated, with the ultimate aim of better understanding complex systems surrounding us.

In spite of some initial and basic attempts to join both fields (see, for instance, Ref. [HW12]), these undertakings have been mainly focused on confined problems, and a more comprehensive approach has not yet been pursued. Indeed, each field can yield new ideas and techniques that can strongly contribute to the improvement of the state of the art of the other. On one side, knowledge discovery and data mining techniques may improve the creation and analysis of complex networks by means of: *i)* identification and selection of the most relevant features in the initial data, *ii)* standard methods for data pre-processing (like, for instance, creation of new features), and *iii)* analysis of the significance of network-based results. On the other side, complex networks analysis is mainly expected to provide a new way of representing and extracting information about the structure of systems characterized by interacting elements, thus providing a new point of view to classical data mining tasks like classification or regression.

Consequently, in this PhD thesis we are going to tackle the problem of the integration of knowledge discovery and complex network analysis (see Figure 1.1 Right), in order to improve the output of classification tasks performed on complex systems, thanks to the improved information provided by a complex network analysis. As this will require stronger and sounder methods for the extraction of knowledge from complex systems, such integration will also improve our understanding of how they are organized and evolve. From an engineering point of view, such additional knowledge can also open new doors toward the optimization, repair or forecasting of complex systems and of their dynamics, being them technological, social, or biomedical.

In order to validate such proposal, the presented case studies are drawn from contemporary biological and biomedical problems. Several are the reasons supporting this

election. First, the social relevance of this field: any small improvement on the state of the art can yield important benefits toward the understanding, and hence the treatment, of deadly diseases. Additionally, biomedical data sets are intrinsically challenging, due to the cost of obtaining a relevant quantity of information, and to the always-present measurement noise; they thus represent a perfect tool for validating any new methodology. Last, but not least, the interaction of the PhD candidate with the Center for Biomedical Technology in Madrid, which ensured the access to a large collection of different biomedical data sets.

In spite of this primary focus, the techniques here developed have a general applicability, and thus potential applications to social or technological systems will be taken into account.

1.1 Objectives and hypothesis

In this Section, we review the main objectives of the Thesis, along with their corresponding research questions and hypothesis, as developed in the Thesis Plan Proposal.

Objective 1: *Use of feature selection techniques in the pre-processing phase of network reconstruction.*

> **Research Question**: *How can features be pre-selected and ranked, in order to reduce the computational cost of the network reconstruction and analysis phases, without reducing results significance?*
>
> In the creation of the network representation of a complex system, it is common to map all the elements composing such system into nodes, without any beforehand evaluation of their significance; on the contrary, it is believed that the true significance will be defined by the network itself. For instance, in the analysis of genetic networks, all available genes are usually included in the structure, with the idea that the network analysis will be able to detect if some of them are irrelevant. Clearly, this approach has a major drawback: a significant increment in the computational cost. The complexity of the analysis may be reduced by eliminating irrelevant features from the initial data set; furthermore, flexibility may be improved by ranking features according to their relevance, such that the final network size can arbitrarily be tuned. In both cases, an improvement of the score associated to data mining tasks may also be detected.
>
> **Hypothesis:** If the initial data are processed, such that only relevant features are included in the analysis, both an improvement of the score associated to data mining tasks, and a significant reduction of the computational cost should be detected.

Objective 2: *Network reconstruction through data mining techniques.*

> **Research Question**: *How can data mining techniques be used to create novel network*

representations of data sets?

One of the most important points in the creation of a network representation of a given data set is the definition of the meaning of *links*, *i.e.* the connections between pairs of nodes. It is not uncommon to find that physical or virtual relationships between the elements of the system, *e.g.* hyper-links between the pages of a web site, constrain the way a link is defined. When such relationships are not explicit, *functional* links can still be built, providing that nodes are described by time evolving observables (*e.g.* the time evolution of a stock price, or of brain activity in a given region). Here we will tackle this problem for data sets that do not fulfill these requirements, as the case of (static) sets of biomedical measurements. The use of data mining techniques should allow the creation of non-conventional network representations, thus enabling the application of complex network analysis to previously off-limits problems.

Hypothesis: If data mining techniques are applied to the creation of network representations, the resulting networks should be more representative of the systems under study, and this should reflect in an improvement of the performance of subsequent data mining tasks.

Objective 3: *Use of network representations for improving data mining tasks.*

Research Question: *How can complex networks be transformed into a set of features, which can be used to feed a data mining algorithm?*

Classical data mining tasks can be of utmost importance in the analysis of complex systems; a classification algorithm can be used, for instance, to detect people suffering from a disease, or to automatically classify different variants of a tumor. Nowadays, data mining techniques are applied directly to the raw pre-processed data: here we propose the use of a network representation as an intermediate step. The interactions between the elements composing the system may be represented as a complex network, whose topological characteristics may then be used to feed a classifier. Such explicit use of the structure of interactions is expected to reduce the complexity of the data mining task, as structural features are promptly provided, thus improving the score.

Hypothesis: If a description of the structure of interactions, characterizing the complex system under analysis, is used to feed data mining algorithms, there should be a significant increase in the output scores associated to different knowledge discovery tasks.

Objective 4: *Use of data mining tools to validate network representations.*

Research Question: *How can a network representation be validated, i.e. how can the quantity of information codified in it be estimated, and its relevance for a given knowledge*

discovery task be assessed?

The accepted methodology for complex network reconstructions includes several steps in which the experience of the researcher comes at play. This creates the problem of validating the significance of the network representation, that is, estimate how much information about the system is indeed encoded in the network, and assess whether such information is enough to perform a given data mining task. Our contribution consists in turning the problem around: specifically, in using the score of a data mining task to provide metrics assessing the significance of the representation in an objective way. This will also allow an automation of the network reconstruction process, in which specific parameters are optimized in order to achieve the highest score.

Hypothesis: If the score obtained in a data mining task, by using information extracted from complex networks, is improved with respect to comparable studies in the Literature, then we can conclude that the network representation considered is valid.

1.2 Research methodology

The use of a robust methodology is one of the most important requirements of scientific research, in that, by defining the general direction and the specific steps one ought to follow, it allows reaching the expected results in an efficient and safe way. Throughout this Thesis, the scientific methodology proposed by Quivy and Van Campenhoudt [QVC98] has been used, adapted to the characteristics and challenges of the problem in hand.

As proposed in Ref. [QVC98], any scientific activity should be constructed around three *acts*. First, the researcher needs to *break* pre-established ideas; this requires identifying a question, for then analyzing the answers that have already been given in the Literature and preparing a set of objectives to be pursuited. Second, these pieces of information should be used to *construct* an analysis model, that is, a theory that can systematically be analyzed by means of real data. Finally, such model must be *validated*: the observed reality should be used to assess its usefulness or, if this does not happen, should be used as a starting point for the construction of a new analysis model.

The reader will recognize these acts in the different parts composing this document: from the identification of the scientific questions and associated hypotheses in Section 1.1, the discussion of the current state of knowledge in Chapter 2, up to the development and validation of several analysis models in Chapters 3 - 5 - see the following Section 1.4 for a description of the structure of this Thesis.

Three important aspects of this methodology should here be discussed. First of all, this Thesis deals with the application of data mining techniques to complex network analysis; the latter requires a specific methodology, *i.e.* a set of steps leading from the raw data, as recorded in the real world, to the final knowledge about the system. The analysis

process here followed corresponds to the best practices accepted in the research community, for the analysis of both complex networks [FCOTRBAVR11] and data in general [FPSS96]. Second, the validation of all developed analysis models has been performed with the help of real data. Validation case studies have been drawn from biomedical and biological problems, mainly due to the associated complexity and social relevance. As one of the main drawbacks of complex network analysis is its subjectivity, the main topic of Chapter 5, it was important to ground all validations with objective measures. This has been achieved by means of classification problems, *e.g.* discriminating between healthy subjects and patients, or the identification of the elements responsible for a disease, which can then be validated against the literature. Finally, the reader should notice that, although any validation process should include analyses of both accuracy, reliability and usefulness, here we have focused only on the first of them. Obtaining biomedical data sets from real experiments was out of the scope of this PhD, due to its complexity and the need of specific knowledge; therefore, the reliability was ensured by the use of well-known public data sets, or of data sets already used in peer-reviewed publications. Similarly, the usefulness of the obtained results was assessed by experts in the pertinent biomedical areas.

1.3 Main contributions and publications

In this Thesis, we present the first description of a **global process for the use of data mining techniques in the study of complex networks**, covering all its phases: from the definition of nodes and links, to the final analysis of topological metrics. Starting from a set of raw data representing a complex system, our first contribution is **a novel technique for mapping them on a network structure**. While different methods have been proposed in the past, none of them was targeting situations in which the elements of the system are described by static data, *i.e.* not evolving with time; furthermore, they lacked a collaborative approach: each system is usually analyzed in an individual fashion, without creating a global picture of the differences and similarities between groups of systems. The algorithm we propose maps deviations from an expected behavior into links, being such normal behaviors detected through standard data mining techniques. The possibility of creating network representations from static data sets opens doors toward addressing problems that were outside the range of complex network applications, as for instance the analysis of biomedical data like genetic or metabolic expression levels.

We also **compare feature selection strategies for the reduction of complex network dimensionality**. While the physics community considers complex networks as indissoluble objects, the deletion of superfluous information is a standard problem in data mining, providing benefits such as reduced computational costs and increased statistical significance. For the first time we here apply algorithms, based on the assessment of the information encoded by each node, to demonstrate that the network size can be reduced up to a 50% without affecting the score of a classification task. The application of feature

selection strategies to *Obstructive nephropathy* (Section 4.2) and cancer (Section 4.3) have been respectively published in *Networks and Heterogeneous Media* (Impact Factor of 0.909) [ZMSB12] and *PLoS One* (Impact Factor of 4.092) [ZMBS13].

Furthermore, we **provide a methodology for selecting the network topological features most relevant for describing a system**. This is performed by assessing the information encoded in networks representing different conditions, *e.g.* healthy subjects and patients, through the score obtained in a classification task. Thus, starting from a ground truth, the proposed methodology yields criteria for an optimal network representation relative to a given problem. We also show how this approach can be extended to related problems, *e.g.* the identification of the best parameters in the network reconstruction phase, or the estimation of the severity of a pathological condition. This methodology, that will be extensively discussed in Section 5.1, has been published in *Nature Scientific Reports* (Impact Factor of 2.927) [ZSPBGPPMB12].

Due to the young age of complex network theory, it is acknowledge that a lot of work has still to be done, both from the theoretical (*i.e.* the design of metrics able to detect relevant topological structures) and applied point of view (as, for instance, the reduction of the computational cost of network analysis, which would yield a deeper understanding of the properties of large-size systems). In order to cope with problems arisen by the novelty of the proposed methodology, a **fast motif detection algorithm** and a **meso-scale topological metric** have been developed. The former reduces the computation time by up to two orders of magnitude, making possible the analysis of motifs in medium and large networks; this has allowed us, for the first time, to highlight their importance in the dynamics of the brain - something that was hypothesized, but never demonstrated with real data. The latter allows the representation of a large number of meso-scale topological structures with a single number, thus allowing the use of such structures as an input in data mining tasks. This method for assessing the presence of meso-scale structures has been presented in *Europhysics Letters* (Impact Factor of 2.260) [ZSM14].

Finally, we **validate these new concepts by applying them to seven biomedical problems**, ranging from the analysis of human and plant genetic expressions, up to the identification of pathological patterns in brain dynamics. This has lead to the publication of several manuscripts. Specifically, results obtained from the analysis of metabolic data, as will be presented in Sections 3.3 and 5.5, have been published in *Metabolites* [ZPSEFRASEJRBMS13]. Furthermore, the identification of genes involved in the Arabidopsis Thaliana response to osmotic stress has been published in *Nature Scientific Reports* (Impact Factor of 2.927) [ZMVGPSMB14].

Beside these six published papers, two more have been prepared and are now under consideration in different journals. Specifically, the application of data mining techniques for the comparison of different synchronization metrics in the analysis of brain activities, as reported in Section 5.3, is under consideration in *Neuron* (Impact Factor of 15.766). Furthermore, the algorithm for the fast motif enumeration in dense graphs, presented in Annex 6.1, is being considered in *Bioinformatics* (Impact Factor of 5.468).

Some selected results have been also presented in international conferences, and in several closed-doors meetings with research groups involved in biomedical studies. Among them, the most important have been given at *Net-Works 2011*, October 26-28, in El Escorial (Spain); *9th AIMS Conference*, July 1-5, 2012, in Orlando (USA); *Dynamics Days Europe 2013*, June 3-7, 2013, Madrid (Spain); *European Conference on Complex Systems '13*, 16-20 September 2013, Barcelona (Spain); and *ESMRMB '13*, 3-5 October 2013, Toulouse (France). Presentations have also been given at the University College of London, UK (group of Alexey Zaikin), and at the Università di Milano Bicocca, Italy (group of Costanza Papagno).

1.4 Structure of the document

We begin this Thesis by discussing the current state of knowledge in Chapter 2. This chapter is intended to provide relevant background material on the two main fields relevant for this work, *i.e.* complex systems and complex network science and data mining. Subsequently, the main work is developed following the structure depicted in Fig.1.2. It is based on the logical flow that allows the researcher to move from raw data to knowledge about the system under study. Thus, Chapter 3 tackles the problem of creating network representations starting from real-world data sets, and presents a novel technique for reconstructing network representations that builds up from sets of scalar data representing groups of pre-labeled subjects; Chapter 4 presents the application of standard *feature selection* techniques to different types of biomedical data, and their effectiveness in improving the accuracy of network representations; finally, Chapter 5 deals with the problem of extracting useful knowledge from a network representation, and how to assess the quantity of information codified in a network. Each one of these Chapters first presents an introduction to the problem and the novel solution we propose, for then demonstrate its relevance with a set of validation cases drawn from the biomedical field.

As a topic transversal to all the work here proposed, Chapter 6 presents two novel instruments for complex networks analysis, which have been developed to tackle specific problems encountered throughout this Thesis: the fast enumeration of motifs in dense networks, and the assessment of the presence of meso-scale structures.

Finally, we summarize the insight gained in Chapter 7, and discuss directions for extending the proposed methodologies in the future. For the sake of completeness, a final Annex presents a description of the complex network topological metrics used throughout this document.

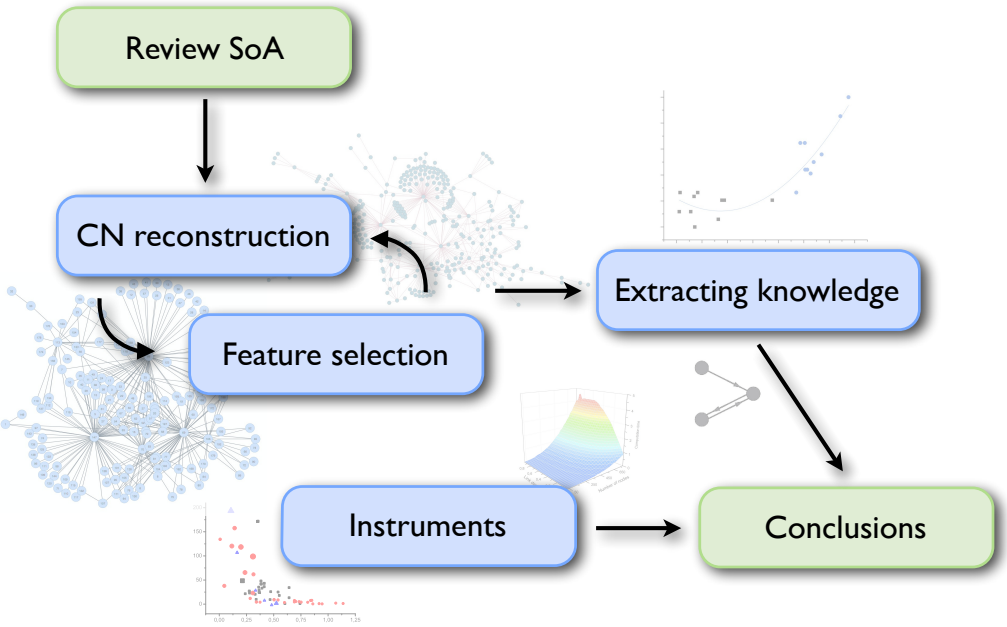
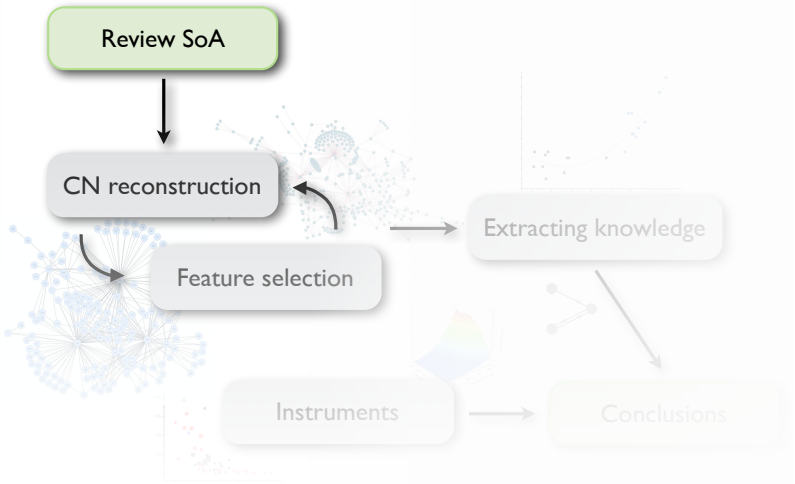


Figure 1.2: **Structure of the Thesis.**

2

Review of the State of the Art



The work expounded in this PhD Thesis knits two different inter-related research fields, *i.e. data mining* and *complex networks* theory. The field of data mining focuses on the design of algorithms that enable computers to learn how to recognize non-trivial patterns and make intelligent decisions based on empirical data. On the other hand, the complex networks theory deals with the study of network representations of complex systems, *i.e.* systems composed of a high number of highly interconnected elements. In the intersection of both fields we can find the work proposed in this PhD Thesis, which aims at creating bridges between both frameworks, for enriching our understanding of complex systems.

In this Chapter, we review the relevant Literature of these two fields. First of all,

Section 2.1 introduces the concept of complex systems, and how their study has been historically tackled. Afterwards, moving to more topical lines of research, Section 2.2 presents the main concepts associated to complex networks, the metric used for describing their structure, and the most important types of networks known from the analysis of real-world complex systems. Finally, Section 2.3 reviews the main concepts associated to data mining.

2.1 The birth of Complex Systems

The decade of the seventies witnessed the birth of a new fundamental concept in the scientific community. At that time, almost all researchers were accepting without question the reductionist hypothesis, according to which all systems could be defined just by their composing elements; therefore, any phenomena could be understandable by characterizing their individual constituents, and summing up these individual effects. Ultimately, this meant that any system could be explained in terms of some few fundamental laws, *e.g.* the performance of a car by applying quantum mechanics to its atoms. Nevertheless, at the same time, it was becoming clear some systems escape this principle. Although, for instance, cell biology is based on the ideas of chemistry, it also requires brand new laws and generalizations, especially if one wants to make even simple problems tractable in feasible time [And72].

The behaviors of some systems cannot be explained just by extrapolating the properties of their constituting elements, as important information is codified in the interactions between these elements. Such interactions generate behaviors at the macro-level, *i.e.* the level of the whole system, which cannot be explained by simply studying the constituting elements: such behaviors are known as *emergent phenomena*.

These systems are called *Complex Systems*: systems composed of a large number of elements, interacting between them in a non-linear fashion, and giving birth to emergent behaviors. In the last decade, thanks to the advancements in information management and in available computational power, the ubiquity of such systems has been observed in different fields of knowledge. From biology to economy and sociology, interactions between the elements of a system have been recognized as important as the elements themselves.

If one is to clarify what a complex system is by means of an example, one of the most paradigmatic would be the human brain (see Figure 2.1). According to the last estimations, it is composed of 100 billion neurons, which are the basic computation units [HH09]. Each neuron is very simple, and its dynamics is very well known: no computation can arise from a single neuron, nor from a small group of neurons. Yet, when a huge number of them are connected following some specific structures, the human intelligence appears as an emergent phenomenon. Once the special nature of complex systems has been recognized, the problem of their characterization, or, in other words, of the extraction of useful information about their structure and dynamics, aroused. This knowledge

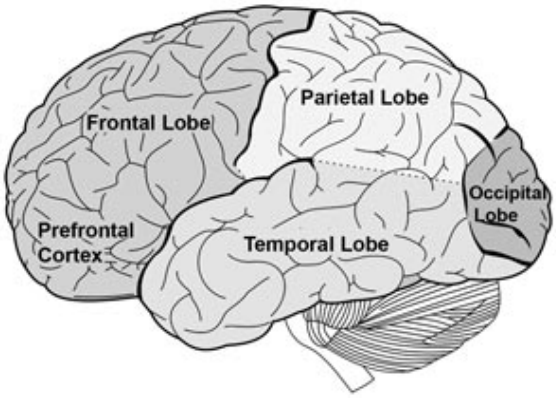


Figure 2.1: **The human brain:** one of the most astonishing known complex systems.

discovery process should take into account the specificities of the systems under analysis:

- Most of the information is encoded in the interactions, and not in the individual elements.
- Such interactions are not always evident, and their presence should be inferred from the dynamics of the elements composing the system.
- Finally, the quantity of information to be processed grows with the number of potential interactions, which, in turns, grows with the square of the number of elements. Therefore, the computational cost can become an important limiting factor, especially in the study of large socio-technical systems.

The extraction of useful information from complex systems has traditionally followed two parallel and almost independent paths:

1. The use of standard data mining tools and techniques, therefore partly neglecting the *complex* nature of the system.
2. A physical approach, by applying concepts drawn from the disciplines of applied mathematics and statistical mechanics.

Up to now, these two approaches have walked independently: the objective of this PhD Thesis is the creation of a bridge between both, *(i)* by improving the understanding of these systems by supporting the physical analysis with data mining techniques, and *(ii)* by improving the outcomes of data mining tasks with the help of physical concepts. In the next two Sections, a review of the most important concepts in complex network theory and knowledge discovery is presented.

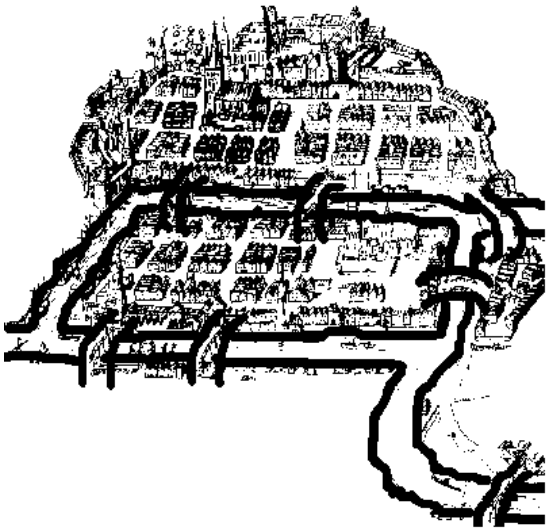


Figure 2.2: **Representation of the city of Königsberg, and of its seven bridges.**

2.2 Complex networks

Any description of the complex networks theory should start from its origin, namely from the graph theory developed by Leonhard Euler and the Königsberg bridges problem. In 1735, the city of Königsberg (nowadays Kaliningrad, Russia) laid on both sides of the Pregel River, with two islands in the middle; connecting all the regions of the city were seven bridges (see Figure 2.2). A problem was formulated: was it possible, starting from one zone of the city, to visit all other zones, by crossing all bridges exactly once, *i.e.* crossing all the bridges, but not crossing a bridge twice?

Euler tackled this problem by firstly eliminating any unessential information. In other words, he realized that the actual structure of the land masses, or of paths inside them, was not necessary toward the resolution of the problem; what was really relevant was the structure of connections created by the bridges. The map of the city was then transformed into a *virtual* representation: land masses were represented as nodes (or vertices), and bridges as links (or edges). Figure 2.3 shows this abstract representation.

Beyond the Königsberg problem, Figure 2.3 represents the first representation of a real

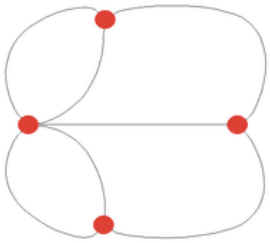


Figure 2.3: **Graph representation of the city of Königsberg.** Nodes represent land masses, and links bridges between them.

system done with the help of a graph. There is one concept that is important to stress: the graph of Fig. 2.3 is a representation of the system that only considers the interactions between its constituting elements, avoiding any information about the nature of the elements themselves. In other words, by means of a graph it is possible to represent in an abstract form the structure behind a system, independently of the nature of the system itself. As we will see below, this can be applied from biology (for instance, with nodes representing genes, and links co-expressions between pairs of them) up to sociotechnical systems (persons and friendship relations, or computers and communication networks).

In the 18th century, the possibilities offered by this approach were not recognized: without doubt, this was due to the limited sources of information about complex systems, and on the null computational capabilities available at that time. As a consequence, graph theory was initially developed as a pure mathematical subject, far away from any application. Yet, in recent years, the advancements in information storage and computation has permitted the analysis of many real-world systems [FCOTRBAVR11], and hence the development of a large set of measures describing their structure [CRTV07].

2.2.1 Characterizing networks

As a first step toward the analysis of complex networks [1] and their classification in families, it is necessary to define some methods for characterizing their structure. Indeed, and as may be expected, network measurements are an essential ingredient for many tasks, as network representation, characterization, classification or modeling. Nowadays, hundreds of different structural network metrics have been defined; while, in what follows, an overview of the most important is presented, the interested reader may refer to Annex A for the definition of all metrics used in this PhD Thesis. Furthermore, the interested reader may also refer to the different reviews available in the Literature, like for instance Refs. [BLMCH06; CRTV07; FCOTRBAVR11].

Metrics describing complex networks are generally classified into three families, depending on their focus: *micro-*, *meso-* and *macro*-scale metrics.

The first of these families, *i.e.* **micro-scale metrics**, focuses on the properties of a single node; when the whole network is analyzed, these metrics are usually averaged over all nodes composing the network. The simplest example of a micro-scale metric is the *degree*, defined as the number of connections arriving or departing from a node. While this measure unveils some information about the structure of the network (for instance, it is possible to define the most central node as the one with more connections), even more knowledge can be extracted from the aggregation of all degrees. In other words, a *degree distribution* $P(k)$ can be constructed, which expresses the fraction of nodes in the network with degree k.

The second important metric that can be extracted from the analysis of the degrees is the assortativity, defined as the presence of correlation between the degree of connected,

[1] While *graph* and *network* are usually used as synonymous, the former refers to simple objects, *i.e.* having a regular or random structure, while a network is characterized by a more complex backbone of connections.

i.e. neighboring, nodes; in highly assortative networks, central nodes (that is, those having a high degree) are connected with other central nodes with a probability higher than what expected in a random configuration. By analyzing different real-world networks, it has been discovered that social networks tend to be assortative, while biological and technological networks are often disassortative [BLMCH06].

When shifting the focus from a single node to a group of them, it is possible to define **meso-scale metrics** [ACLBSN11] [2]; the three most important concepts here are the *clustering coefficient*, *motifs* and *communities*.

The *clustering coefficient* measures the presence of loops of order three in the network, *i.e.* the density of triangular structures. It is also known as *transitivity* [New01], and is defined as follows:

$$C = \frac{3N_{\triangle}}{N_3}, \tag{2.1}$$

$N_{\triangle}$ being the number of triangles in the network, and N_3 being the number of connected triples. Notice that the former measures the number of sets composed of three nodes fully connected between them, while the latter considers sets of three nodes connected by at least 2 links. The factor three accounts for the fact that each triangle can be seen as consisting of three different connected triplets, one with each vertices as central vertex; this, in turn, ensures that $0 \leq C \leq 1$.

An extension of the concept of clustering, named *motifs*, was proposed in 2002 by Milo and coworkers [MSOIKCA02; SOMMA02]. Motifs are subgraphs (usually composed of three or four nodes) that appear more frequently than what could be statistically expected. In order to find these motifs in a real network, one should calculate the number of occurrences of each subgraph in the network, and compare it with the number expected in an ensemble of random equivalent networks. In order to quantify the significance of a given motif, a *Z-score* is usually calculated as:

$$Z_i = \frac{N_i^{(real)} - \langle N_i^{(rand)} \rangle}{\sigma_i^{(rand)}}. \tag{2.2}$$

Here, $N_i^{(real)}$ is the number of occurrences of motif i in the real network, $\langle N_i^{(rand)} \rangle$ the ensemble average, and $\sigma_i^{(rand)}$ the ensemble standard deviation. From this definition of motifs, it can be seen that the clustering coefficient is just the frequency of occurrence of one of all the possible motifs of the network, specifically, of complete triangles.

To conclude this review of meso-scale characteristics of networks, it is worth noticing that most real networks present an inhomogeneous structure of connections, *i.e.* it is possible to identify groups of nodes more densely interconnected to one other than with the rest of the network. Such groups are called modules, or *communities*, and their presence

[2] A good definition of *meso-scale* can be found in Ref. [BS10]: *It is intermediate between a microscopic level, when one studies elements of a system separately (e.g. molecules of a fluid), and a macroscopic one, when an entire system is considered as a whole (e.g. in terms of some averaged characteristics).*

has been accounted for from social systems [ADDGGG04] to metabolic [GA05] networks. Intuitively, the identification of community structures in real networks is of great importance, as nodes belonging to the same community are usually expected to share some characteristics. This insight has been successfully used in biology, and specifically for the identification of the function of new genes and metabolites, based on the function of neighboring nodes [GA05].

In spite of its importance, the concept of community presents two main drawbacks: there is no consensus about its definition, and the problem of identifying communities in networks is NP-complete. For the former problem, *i.e.* defining what a community is, we may cite three relevant definitions: the *weak* definition of Radicchi *et al.*, according to which a subgraph is a community if the sum of all node degrees inside the subgraph is greater than outside it [RCCLP04]; the *modularity* Q, introduced by Newman and Girvan [NG04]; and the definition of communities in terms of information entropy [ZMW05]. The problem of identifying communities in real networks, on the other hand, has been tackled by mean of many different approaches, each one of them having its specific advantages in terms of computational cost, scalability, and precision. For a complete review, the reader may refer to [LF09; For10].

The third family of metrics that can be extracted from complex networks refers to their **macro-scale**, in that they account for the overall structure of the network. In this case, what is analyzed is the movement of information through the network: for instance, how many jumps are needed to move from one side of the network to the other, or what is the importance of a node with respect to the movement of information within the network. In what follows, we will briefly review three of such metrics, which will be repeatedly used throughout this Thesis.

The first metric, and the most simple, is the *average geodesic distance*, defined as the mean number of jumps needed to travel between two nodes of the network:

$$l = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij} \tag{2.3}$$

While this definition has the advantage of being intuitive, it also presents an important problem: it diverges when the network is disconnected, *i.e.* when the distance between two pairs of nodes is $d_{ij} = \infty$. In order to avoid this divergence, Latora and Marchiori proposed a closely related measure, called *global efficiency* [LM01]:

$$E = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}} \tag{2.4}$$

This measurement quantifies the efficiency of the network in transmitting information, supposing that the cost of such transmission if proportional to the distance between the sender and the receiver nodes.

Finally, the overall structure of the network can also be studied by means of spectral measurements. Specifically, it is of interest the calculation of the set of eigenvalues λ_i ($i =$

$1, 2, \ldots, N$) of the matrix encoding the connections between nodes (also called *adjacency matrix* A). The eigenvalues, and their associated eigenvectors, are related to multiple macro-scale properties of the network: from connectivity properties [DA05], up to the importance of individual nodes [Ruh01] or the effects of the topology on the dynamics taking place on top of the network [ADGKMZ08].

2.2.2 Classes of networks

In the last decade, the analysis of a large number of real-world networks [FCOTRBAVR11] has revealed that some structures, or networks topologies, are ubiquitous across many natural and man-made systems. Although an extensive analysis of their properties, and of the models developed to explain how they appear, is out of the scope of this PhD Thesis, in what follows we will shortly review four main types of complex networks, as this constitutes the basis of any classification of real systems.

After the initial work of Euler in 1735, the network (then called graph) theory received little attention. After him, the person mainly responsible for the theoretical advancements in graph theory has been Paul Erdös. Between his numerous contributions, probably the most important has been the introduction of the concept of random graphs, also known as Erdös-Rényi graphs. Given a set of n disconnected vertices, links between all pairs of nodes are created with a probability p. Many theoretical results were obtained in random graphs, as, for instance, the expected size of the largest component (groups of nodes connected between them), or the critical value of p for which the graph was connected [ER59]. A comprehensive review of all results obtained in random graph analysis can be found in Ref. [Bol01].

If random graphs are characterized by a complete absence of structure, the other extreme is represented by regular graphs, *i.e.* networks where all nodes have the same number of connections. Both extremes are of limited applicability in describing real-world networks, as usually natural and man-made systems present a trade-off between regularity and more complex structures. Yet, random graphs are usually used to normalize the properties found in a real network; for instance, the frequency of appearance of a motif is usually compared with the frequency expected in equivalent (same number of nodes and links) random graphs, in order to assess its statistical significance.

In 1998, Watts and Strogatz realized that real networks were not regular, nor completely random graphs, but that they lie somewhere between these two extremes. Specifically, random graphs are characterized by a low mean geodesic distance, and by a low clustering coefficient; on the other hand, regular graphs show high mean geodesic distance and high clustering. By analyzing social and biological networks, they discovered that most of them are characterized by a low mean geodesic distance, but also by a high clustering coefficient. In order to explain how such combination can emerge in real systems, they proposed a hybrid model: starting from a regular graph, few links are deleted at random, and replaced by random (and therefore, not regular) long-range connections.

By tuning the number of links affected by such *rewiring*, it is possible to create a large family of networks, all of them maintaining the high clustering of the regular initial graph, but also showing a reduced distance between nodes. The two combined properties are now widely known as the *small-world effect* [WS98].

Finally, a fourth class of network topologies emerged when another fact about real-world networks was observed. Instead of having homogeneous nodes, *i.e.* nodes with approximately the same number of connections, real-world networks are characterized by some highly important nodes, usually called *hubs*. A clear example can be found in transportation networks: for instance, the air transport network is characterized by few airports connecting with most of the network, *e.g.* Paris, London, or Madrid, and by many small (secondary) airports [ZL13]. Mathematically, these networks are called *scale-free*, as their degree distribution follows a power law, and thus have no characteristic scale.

In 1999, Barabási and Albert developed a simple model explaining how such scale-free networks can naturally emerge, based on the concept of *preferential attachment* [BA99]. The process begins with an initial network of m nodes, m being usually small. Afterwards, new nodes are added to the network, one at a time; each one of these new nodes is connected to existing nodes with a probability that is proportional to the degree of the latter:

$$p_i = \frac{k_i}{\sum_{j \in G} k_j} \tag{2.5}$$

p_i being the probability of connecting with node i, and k the degree of nodes. Due to this biased attachment mechanism, highly connected nodes rapidly gain more links, thus becoming hubs of the system: an effect known as *rich gets richer*.

2.2.3 Recent trends in network theory

The two previous Sections have presented, in a concise form, the main elements of complex network theory, *i.e.* the core concepts that are by and large used in the analysis of real systems. Nevertheless, these concepts have been extended in the last years, in order to include situations that cannot be directly described by the standard framework: the result has been the creation of *temporal* and *multi-layer* networks.

The former, *i.e.* **temporal networks**, are composed of edges that are not continuously active. As an example, in networks of communication via e-mail, text messages, or phone calls, edges represent sequences of instantaneous or practically instantaneous contacts. In some cases, edges are active for non-negligible periods of time: for instance, the proximity patterns of inpatients at hospitals can be represented by a graph, where individuals are pairwise connected while they are at the same ward. Clearly, the temporal structure of edge activations can affect dynamics of systems interacting through the network, from disease contagion on the network of patients to information diffusion over an e-mail network. The interested reader may refer to some reviews on the topic that can be found in

the Literature [HS12; HS13].

Beside the temporal aspect, it should be noticed that the traditional complex network approach has mostly been limited to the representation of node interactions by means of a (generally, real) number, quantifying the weight of the corresponding graph's connection (or link). Nevertheless, considering all links as instances of a single object can be an important constraint. It may occasionally result in not fully capturing the details present in some real-life problems, leading even to incorrect descriptions of the corresponding phenomena. In what follow, three examples are considered, representative of the major limitation of that approach.

The first one is borrowed from sociology. Social networks analysis is one of the most used paradigms in behavioral sciences, as well as in economics, marketing, and industrial engineering [KY08], but some questions related to the real structure of social networks have been not properly understood. A social network can be described as a set of people (or groups of people) with some pattern of contacts or interactions between them [Joh00]. At a first glance, it seems natural to assume that all the connections or social relationships between the members of the network take place at the same level: the real situation is quite different, though. Indeed, social interactions seldom develop on a single channel, and more than one relationship can bind pairs of people. Such idea was probably firstly analyzed by Erving Goffman in 1974, along with the theory of frame analysis [Gof74]. According to this research method, any communication between individuals (or organizations) is constructed (or framed) in order to maximize the probability of being interpreted in a particular manner by the receiver. Such framing may differ according to the type of relations between the involved individuals, several of them potentially overlapping in a single communication.

A second paradigmatic example of intrinsically multirelational systems can be found in transportation networks, as for instance the Air Transportation Network (ATN) [ZL13]. The traditional study of this infrastructure is based on representing it as a single-layer network, where nodes represent airports and links stand for direct flights between them. Yet, it is clear that a more accurate mapping can be reached if airlines are considered, as passengers cannot easily connect two flights operated by different airlines, at least by airlines belonging to different alliances [CZGGRAB13].

Moving on to biology, the third example is the effort of scientists to understand the role of specific components in a biological system. For instance, the *Caenorhabditis elegans* (or *C. elegans*) is a small nematode, the first organism for which the entire genome was successfully sequenced. Recently, biologists were able to get a full mapping of the C. elegans' neural network, consisting of 281 neurons and around two thousand connections [WSTB86]. The result is not a single network, as neurons can be connected by chemical and electrical (ionic) link, having these two types of connections completely different dynamics. Therefore, a correct representation should include two independent *layers* of connections.

These three examples explain the last years efforts for generalizing the traditional

network theory, by developing a novel framework for the study of **multi-layer networks**, *i.e.* graphs where several different layers of connections are taken into account [BBCGGGRSNWZ14]. Multi-layer networks explicitly incorporate multiple channels of connectivity and constitute the natural environment to describe systems interconnected through different categories of connections: each channel (relationship, activity, category) is represented by a layer, and the same node may have different kinds of interactions (different set of neighbors in each layer). For instance, in social networks, one can consider several types of different actors' relationships: friendship, vicinity, kinship, membership of the same cultural society, partnership or coworker-ship, etc.

2.3 Data mining

The idea of extracting some knowledge from a set of data is not recent, but has existed, in its manual form, since the beginning of civilization. For instance, one may think in the *Kabbalah*, a school of thoughts inside Judaism, born in the 12th- to 13th-century Southern France and Spain, which uses different methods to analyze hidden meanings and messages inside the Torah [Dan05]; or the analysis of the astronomical observations performed by Johannes Kepler in the 16th century. Yet, the proliferation, ubiquity and increasing power of computer technology have dramatically increased data collection, storage, and manipulation ability. This, in turn, has created a new need for automatic data analysis, classification, and understanding.

Throughout this Thesis, several concepts and techniques drawn from data mining will be used. Yet, it should be noticed that data mining is a particular step, involving the analysis of data, of a more general concept called *Knowledge Discovery in Databases* (KDD) [3] For the sake of completeness, the formed will be described in the light of the latter.

2.3.1 Knowledge Discovery in Databases

The name Knowledge Discovery in Databases (KDD) was introduced in the first KDD workshop in 1989 [PS90], and since then it has become popular in the artificial intelligence and machine learning fields. KDD has been defined as the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [FPSS96]. In other words, KDD is a process that transforms a set of raw data into other representations that might be more compact, more abstract, or more useful. This process is performed by combining methodologies and techniques from different fields, such as statistics, databases, machine learning and visualization, and it is composed of the different steps depicted in Figure 2.4 [FPSS96]. It should be noticed that data mining refers to the design of algorithms that enable computers to recognize non-trivial patterns

[3] In the last decade, the new concept of *data science* has emerged, whose goals include extracting knowledge from data and creating data products. KDD and data mining are usually considered as tools inside data science - see, for instance, Ref. [SBGMPT11].

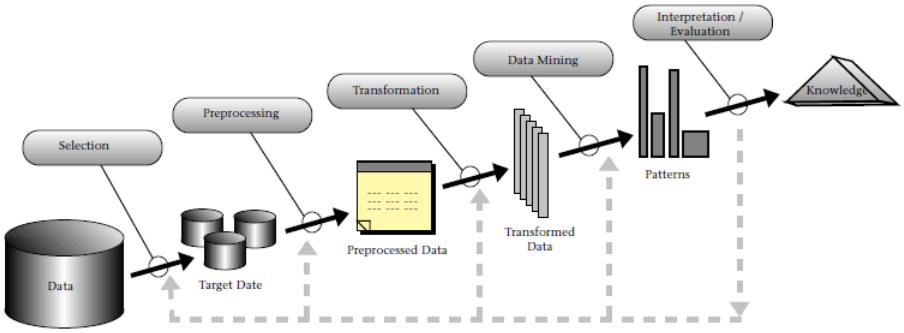


Figure 2.4: **Overview of the steps composing the KDD process.** Source: www.crisp-dm.org.

and make intelligent decisions based on empirical data, and as such, it is one of the elements required in a KDD study. The terms “Knowledge Discovery in Databases" and “data mining" are frequently used interchangeably, with the latter becoming more popular in the business and press communities in the last decade. Yet, KDD refers to the overall process of discovering useful patterns from data, while data mining refers to a particular step in this process, *i.e.* its analysis step [FPSS96].

KDD is usually organized in phases, although most of the time the knowledge discovery process is not linear and presents loops and feedbacks - such characteristic will be depicted in Fig. 2.5. In 1997, an industry group called the *Cross-Industry Standard Process for Data Mining* (CRISP-DM) proposed a methodology for organizing KDD processed in standard steps, as illustrated in Fig. 2.5, which are described in what follows [CCKKRSW].

- *Business (or Problem) Understanding*: Initial phase that focuses on understanding the project objectives and requirements from a business perspective. This first phase requires several sub-objectives to be fulfilled. Business objectives, *i.e.* what the client really wants to achieve, should be clearly understood; resources, assumptions, constraints and other important factors (collectively known as the current situation) should be analyzed; both business objectives and the current situation should be transformed into data mining goals; finally, a good data mining plan has to be established. In resume, this first step creates the initial knowledge about the problem that will guide all subsequent phases.

- *Data Understanding*: this phase mainly deals with the acquisition of the data and their initial analysis, with the aim of getting familiar with them, and of understanding their main features. Of special relevance is the analysis of the quality of the data, that is, the identification of missing or wrong information within the set.

- *Data preparation*: against the general belief, this phase is one of the most important of the whole process, as the success of the final analysis strongly depends on it, and may consume up to the 90% of the time devoted to the process. Once the

suitable data sources have been identified, it is necessary to select, clean, construct and format them into the desired form. Good surveys are available on this topic: see, for instance, Ref. [ZZY03], where firstly the importance of data preparation for data analysis is motivated, and secondly a review of the research achievement on this field are presented, along with future directions and open problems.

- *Modeling*: this phase of the KDD is commonly known as data mining. Once data are ready to be analyzed, different machine learning and data mining algorithms are selected and applied, and their parameters fine-tuned; in the next Section, some of these techniques will be reviewed in detail. Typically, several techniques can be chosen to solve the same modeling problem, each one of them having their own specific requirements on the format of input data; due to this, a step back to the data preparation phase is often required.

- *Evaluation*: once the models have been executed against the data, it is necessary to review the output patterns obtained in the light of the business requirements identified in the first phase. This is because the final aim of the KDD process is to gain deeper understanding of the business under study (as described in the first phase), and therefore the discovery of a specific pattern in the data may rise new questions, which should be analyzed in detail. Due to this, the phases included between Business Understanding and Evaluation should be seen as an iterative process, as depicted in Fig. 2.5. Only when all relevant questions have been addressed, one can move to the final phase, *i.e.* the deployment of the product.

- *Deployment*: once all the knowledge about the business under study has been gathered, such knowledge should be organized and presented in a way the customer can understand and use. Depending on the context, this can be as simple as creating a report, or as complicated as a repeatable data mining process across the organization.

From this short review of the KDD process, and in the light of what was described in Section 2.2, it should be noticed that a joint data mining and complex network approach may go beyond specific contributions to the Modeling phase. Instead, new synergies can also be created for the Data Preparation and Evaluation phases. In what follows, we will review the three phases, with special attention for the Modeling phase.

2.3.2 Feature selection

High dimensionality of the feature space can make the learning problem more difficult, even when the model does only depend on a reduced number of variables. Indeed, even if many data mining algorithms attempt to automatically detect which features are important, and which features can be eliminated, both theoretical and experimental studies indicate that many algorithms scale poorly with a large number of irrelevant features are

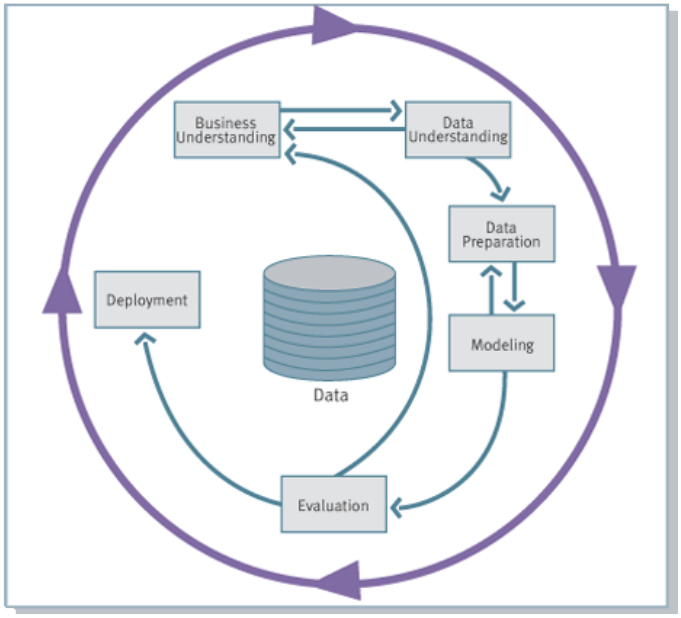


Figure 2.5: **KDD as a non-linear process.** Source: www.crisp-dm.org.

included [Lan96]. The goal of feature extraction procedures is then threefold: reducing the amount of data to be analyzed, center the analysis only on relevant data, and improve the quality of the data set. Feature selection is a particularly important step in those domains that entail a large number of measured variables but a very low number of samples, like, for instance, biological and medical domains: gene and protein expressions, magnetoencephalographic and electroencephalographic records, *et caetera*.

Many feature selection algorithms have been described in the Literature [LM98; LM07]. Broadly speaking, they can be classified into three different families:

- *Wrappers*, using an algorithm to search through the space of possible features and evaluate each subset by running a model on the subset. Wrappers can be computationally expensive and have a risk of overfitting the model.
- *Filters* are similar to Wrappers in the search approach, but instead of evaluating against a model, an independent filter is evaluated.
- *Embedded techniques*, which are embedded in and specific to a model.

Furthermore, the construction of the subsets to be evaluated can follow different strategies: *exhaustive approaches*, where all possible subsets of features are analyzed, which is impractical for large sets of features due to its extreme computational cost; *heuristic approaches*, *i.e.* based on a greedy algorithm that adds the best feature (or deletes the worst feature) at each round; *nondeterministic approaches*, randomly generating feature subsets; and *instance-based approaches*, where features are weighted according to their role in differentiating instances of different classes for a data sample.

2.3.3 Data Mining tasks

Problems to be solved by data mining techniques can be divided in two main families:

- *Descriptive problems.* Descriptive Modeling is one of the two main branches of Data Modeling, also called sometimes Exploratory Analysis. The main purpose of descriptive modeling is to describe, in the form of patterns, the information encoded the data set under study. This type of techniques has an exploratory nature, in that they allow a better characterization of the existing data, without providing any forecast. Clustering and Association Rules techniques and algorithms are used in this family of problems.

- *Predictive problems.* The main goal pursued in this case is to find a model, constructed on top of the information already labeled in the data set, that can be used in the future to predict information, *i.e.* to label a non-labeled subject. It is important to mention that both in the case of predictive and descriptive problems, the methods used to extract knowledge are based on inductive learning, where a model is constructed by generalizing a set of training records. In the particular case of predictive models, it aims at predicting the value of a particular attribute (the target variable) based on the values of other attributes. The underlying assumption of the inductive approach is that the data used for training are representative of the whole universe, *i.e.* of all the possible unknown data that may be encountered, and therefore that the trained model is able to accurately predict the values of future unseen instances.

The work presented in this Thesis is mainly focused on classification methods, which represent a set of supervised learning techniques whose goal is to learn/build a model from training data, composed of discrete labeled instances, and apply this model to predict the class of new/future unlabeled records. The process of learning is organized in different steps, which are described here below:

1. Split the dataset into test and training subsets. The set of training instances will be used to construct the model, while the test set will be used to validate the model previously obtained. In Section 2.3.5 we will analyze different techniques for the validation of the resulting models.

2. Establish which variables are the input variables (also known as *conditional variables* or *independent variables*), and which one is the output representing the concept to be learnt (also known as *decision variable* or *dependent variable*). One important aspect to take into account is the number of variables and the information they gather to predict the output with accuracy. In cases where the number of variables is too high, it is necessary to apply the feature selection techniques reviewed in Section 2.3.2.

3. Determine the kind of representation of the output. This is also related to selecting the algorithm that will be used to build the model.

4. Execute the algorithm and validate. The model will be built and validated in order to measure the quality of the obtained model (see also Section 2.3.5).

The number of algorithms and techniques for supervised learning is immense: in what follows, the most relevant ones, and those that have been used in this Thesis, are reviewed.

2.3.4 Review of classification algorithms

As previously described, classification algorithms are developed to learn from a set of training data, with the final aim of predicting the class of new unlabeled records. More specifically, let X be the feature space and its possible values, and Y be the space of possible discrete class labels of target variable. The underlying assumption is that it exists a function $f(X) \mapsto Y$ that assigns a record to a class depending on the values of its describing features. In general, it is not possible to know directly f, and therefore all classification algorithms try to create an approximate function $\tilde{f}(X) \mapsto Y$ given a set of correctly labeled instances. The classification algorithm will try to minimize the distance between f and $\tilde{f}$, *i.e.* minimize the expected error. Consequently, classifiers are usually evaluated by assessing their predictive accuracy on a test set, and this is calculated by dividing the number of correctly classified records by the total number of proposed records. Many classification algorithms have been proposed, each one with different advantage and disadvantages, and their own requirements on the format of the data. We list below the most successful and well-known types:

Decision trees techniques aim at generating comprehensible tree structures that classify records by sorting them based on attribute values. Each node in a decision tree represents an attribute in a record to be classified, and each branch represents a value that the attribute can take. The attribute that best divides the training data with respect to the class is the root node of the tree. There are several methods to calculate the goodness of an attribute, such as the *information gain* [HMS66] or the *gini index* [BFOS84]. The most popular decision tree algorithms are ID3 [Qui86], C4.5 [Qui93] and CART [BFOS84]. However, the details of such algorithms are beyond the scope of this thesis. For a more detailed review of works in decision trees we refer the reader to Ref. [Mur98].

Rule induction, aiming at creating the smallest rule-set that is consistent with training data. Ref. [Qui87] proposes the creation of these rules from decision trees, where each path from the root to a leaf represents a rule. However, direct induction algorithms have also been proposed such as the RIPPER [Coh95], which has been shown to be competitive with C4.5 [Qui93]. Ref. [Fur99] provides an extensive overview of existing works in rule induction.

Artificial Neural Networks (ANN), inspired by the structural aspects of biological neural networks. ANNs are represented by a set of connected nodes in which each connection has a weight associated with it. The network learns the classification function adjusting the node weights. The simplest kind of neural network is the single layer perceptron [Ros62], which has two important drawbacks. Firstly, since perceptron-like methods are binary, in the case of multi-class problems the whole classification task must be split into multiple binary sub-problems. Secondly, single layer perceptrons are only capable of learning linearly separable functions, and thus are not suitable for the kind of problems usually found in real KDD applications. To overcome these limitations, it was proposed the use of multiple layers of perceptrons. The problem, in this case, was finding a suitable method for training the network; the solution was proposed in 1986, with the celebrated *back-propagation algorithm* [RHW86]. A good review of ANN works can be found in Ref. [Zha00].

Instance based algorithms, also known as lazy-learning algorithms in that no explicit model is created in advance. When a new record is presented to the system, it is classified based on the class label of training records that have similar properties. This type of algorithms require less computation time during the training phase compared to other algorithms, as no model should be created beforehand. For a comprehensive review of instance-based classifiers, the interested reader may refer to Ref. [Aha97]. kNN [CH67] is one of the most well known examples of lazy-learning algorithms. It classifies a record using the most frequent class of its k nearest records, by means of some distance metric. For this algorithm, the distance metric is the most important parameter, and it should be carefully selected in order to minimize the distance between two similarly classified records, while maximizing the distance between records of different classes.

Support Vector Machines (SVM), are binary linear classifiers that model concepts by creating hyperplanes in a multidimensional space. The axes of this space are given by the features available in the data set, whose values should always have a numerical form. Records are mapped into this space, and the best linear separation between them is then calculated. SVM were proposed by Cortes and Vapnik [CV95], and a comprehensive tutorial can be found in Ref. [Bur98].

While all the previous algorithms were non-probabilistic, in the sense that all records are assigned to just one class, **probabilistic models** have also been proposed. Specifically, these statistics-based approaches build an explicit probabilistic model, which provides a probability for a record to belong to a given class. *Bayesian networks* are amongst the most well known statistical learning algorithms; the reader may refer to Ref. [Jen96] for a good introduction on this topic. Within the family of Bayesian networks algorithms, special attention has been devoted to *Naive Bayes Algorithms*, which assume independence amongst the attributes [CKB87].

2.3.5 Validation

One of the most important steps in the process of Knowledge Discovery in Databases, and yet the one that probably receives less attention, is the validation phase. Its aim is the assessment of the generality of the patterns obtained in the data mining phase, and therefore the understanding if these patterns will still be valid in an independent data set [Koh95].

The criteria used to validate a model usually fall into three categories: accuracy, reliability and usefulness.

The first category, **accuracy**, measures how well the model correlates an outcome with the attributes in the data that has been provided. There are various measures of accuracy, but all these measures depend on the data that are used; in other words, the outcome strongly depends on the quality of data, in terms of number of missing values, or amount of errors. For instance, let us consider the example of a model that predicts some medical property of a set of people using genetic information; in the validation phase, we may detect that the developed model is very accurate: and this in spite the fact that genetic information was retrieved with a wrong protocol. In this example, we succeeded in creating a very accurate model, describing the initial data: the problem is that those data do not reflect the reality. Due to this, measurements of accuracy must always be balanced by assessments of reliability.

The accuracy of a classification task, like the one proposed throughout this Thesis, is usually assessed through several metrics:

- *Accuracy*, defined as the proportion of correct predictions obtained, *i.e.*:

$$AC = \frac{tp + tn}{tp + fp + tn + fn}, \tag{2.6}$$

  tp being the number of true positives, tn the number of true negatives, and fp and fn respectively the number of false positives and negatives.

- *Precision*, *i.e.* the proportion of correct positive forecast:

$$P = \frac{tp}{tp + fp}. \tag{2.7}$$

- *Recall*, *i.e.* the fraction of relevant instances that have been retrieved:

$$R = \frac{tp}{tp + fn}. \tag{2.8}$$

- Finally, *F-measure*, or F_1 measure, is the harmonic mean of precision and recall:

$$F = 2\frac{precision \cdot recall}{precision + recall}. \tag{2.9}$$

Reliability assesses the way a data mining model performs on different data sets. A data mining model is reliable if it generates the same type of predictions or finds the same general kinds of patterns regardless of the test data that is supplied. For instance, the model generated in the previous example would not generalize well to other subjects, whose genetic information is extracted in another laboratory and with a different protocol, and therefore that model would not be reliable.

Usefulness includes various metrics that tell you whether the model provides useful information. Continuing with the previous example, the identification of a large set of genes, *e.g.* 1000, that contribute to the development of a disease does not tell us which are the mechanisms involved in its development, nor gives us clues about the design of a cure. In other words, the basic business questions may not be answered even by an accurate and reliable model.

In the Literature, two are the options available to set up a validation exercise: the use of two or three independent data sets. In the former case, the data mining model is built upon the first data set (called *training set*), and is validated with the second (called *evaluation*, *test*, or *validation set*). The latter case is used when the data mining process includes the induction of partial models, *e.g.* the combination of multiple models; in this case, a training set is used to train the set of initial models, the validation set helps in the selection of the best models, and finally the test set evaluates the performance of the final model as a whole.

It has to be noticed that a single test, *i.e.* the assessment of accuracy against a single validation data set, usually don't reflect how the model will perform in a real environment. To solve this problem, the simplest and most widely used method is the **cross-validation** technique [Sto74; Sto77]. It involves the estimation of the *extra-sample error* (that is, of the expected error that may be found when applying the model to an independent data set) by applying the trained model to several validation sets, and averaging the error rates obtained. Clearly, in order to have meaningful results from a cross-validation process, the training and validation sets should be representative of the whole universe; this can only be achieved by a rigorous selection of the subjects included in the data set, and in their analysis - in other words, in the existence of a sound experimental protocol.

Biomedical applications are characterized by an important limitation: the information available, in term of number of subjects, is limited by the complexity of performing real tests, and it is usually very small: most of the researches should be performed with information about ten or twenty subjects. Due to this, it is not feasible to have independent training and validation sets, and a different strategy should be adopted: this strategy is called ***K*-fold cross-validation**. The available data is divided into K roughly equal-sized parts (both at random, or using any other suitable rule); afterward, for the kth part, the model is trained using the other $K-1$ parts as initial information, and the prediction error is calculated using the kth part as the validation set. This is performed for all $k = 1 \dots K$, for each one of them obtaining an error E_k. The extra-sample error, *i.e.* the error expected for an independent test sample, is estimated by averaging the errors

obtained in each test:

$$CV = \frac{1}{k}\sum_{i=1}^{k} E_k \tag{2.10}$$

Once again, the selection of the parameter of the validation, *i.e.* K, depends on the characteristics of the problem; furthermore, it has to be noticed that the lower is K (and therefore the smaller is the training set), the higher is the probability of overestimate the prediction error. As a rule of thumb, it is usually recommended to set $K = 5$ or $K = 10$.

A special case is when $K = N$, N being the number of observations in the original data set. In this case, a single observation is used as validation set, while all others are used for training the model; this process is repeated N times, once for each observation, and the mean error is calculated. This special case of the *K*-fold cross-validation is known as **leave-one-out cross-validation** (LOOCV).

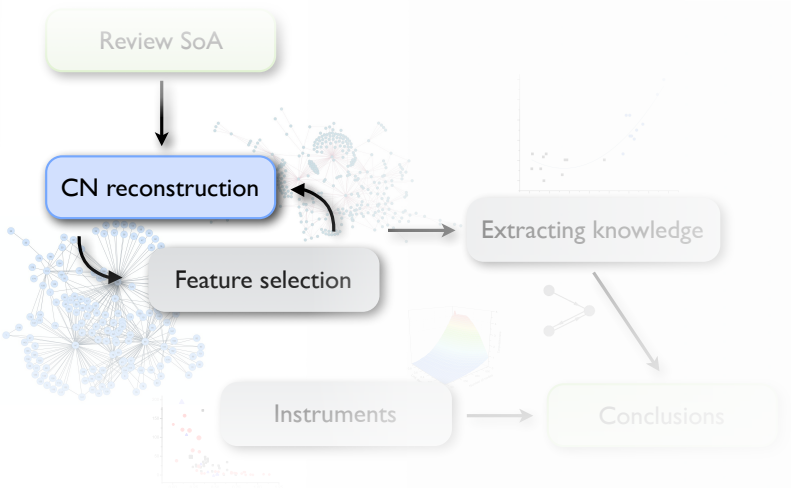
An alternative approach to validation is the **Bootstrap method**, derived from a general tool for assessing statistical accuracy [ET93]. Its basic idea is to randomly draw data sets with replacements from the training data, each one of these data sets of the same size of the original training set. Each one of these *Bootstrap samples* is used to train a model, which is afterward evaluated against a set composed of observations not included in the training Bootstrap sample. The main drawback of the Bootstrap validation method is its bias upward as an estimator of the prediction error. This is due to the fact that each Bootstrap sample includes an average of $0.632N$ different observations, and therefore it behaves like a 2-fold cross-validation. This bias is solved by the so-called **".632 estimator"**. The prediction error $\hat{Err}^{(.632)}$ is defined as:

$$\hat{Err}^{(.632)} = 0.368\bar{err} + 0.632\hat{Err}^{(1)}, \tag{2.11}$$

$\hat{Err}^{(1)}$ being the error of the standard Bootstrap validation, and $\bar{err}$ being the training error (using the whole data set as training set).

3

Representing data sets by means of complex networks



The study of a system by means of complex networks is tantamount to, first, endowing such system with a suitable network representation, and second, to analyzing its structure of connections, *i.e.* its topology. This Chapter deals with the first of these steps.

When relationships between the system elements are defined upon a physical support, their identification is a straightforward task, and the researcher only needs to map them into the network representation. For instance, one may consider the air transportation network; when airports are represented by nodes, links are naturally established

between pairs of them if at least one direct flight is connecting these two airports [ZL13]. In the absence of such relationships, links can still be built, provided a vector of observables can be associated to each node. In this case, each link represents the presence of a *functional relationship* between the data of a pair of nodes, and the resulting networks are called *functional networks*. For instance, if one is to analyze the structure of a stock market, nodes may represent stocks, with pairs of them connected whenever there is a significant correlation in their price evolution through time.

It is important to notice that the requirement of having a vector of observable for each node precludes the use of functional representations for systems whose elements are characterized by a single value. Examples include tissues and organic sample analyses, like spectrography; genetic expression levels of individuals, without evolution through time; biomedical analyses, *e.g.* the study of brain oxygen consumption by means of neuroimaging techniques; or social network analyses, when just a snapshot of its evolution is available.

To overcome this limitation, we here develop a novel method that allows treating collections of isolated, possibly heterogeneous, scalars as networked systems. The method yields a network where each node represents an observable, and links codify the distance between a pair of observables and a model of their typical relationship within the studied population. Thanks to its characteristics, this algorithm allows, for the first time in the Literature, to analyze static data sets, *e.g.* sets of biomedical tests, by means of complex networks.

This Chapter is organized as follows. In Section 3.1 the method for network reconstruction from set of scalar data is presented, with a special focus on cases where one and two classes of subjects are available. Afterward, three validation cases are proposed: the analysis of Obstructive Nephropathy (Section 3.2), of Glomerulonephritis (Section 3.3), and of plant genetic response (Section 3.4) data sets. Finally, some conclusions are drawn in Section 3.5.

3.1 Network reconstruction method

Consider a set of labeled systems $\{s_1, s_2, \ldots, s_n\}$, belonging to a set of classes $C = \{c_1, c_2, \ldots, c_{n_c}\}$. For instance, each system may represent a different person, classified according to his/her health condition: control (or healthy subject), or suffering from some diseases. Each system i is furthermore characterized by a vector of features $\boldsymbol{f}_i = (f_1^i, f_2^i, \ldots, f_{n_f}^i)$ [1] , each one representing an observable of the system: for instance, they may be measurements obtained through blood analyses, or gene expression levels.

The basic assumption behind the proposed method is that each class is defined by a reference *constraint* in the feature space, which describes its characteristics in an ideal way. In other words, and following the previous example, we suppose that it exists

[1] In what follows, we use the standard notation of representing vectors by bold symbols.

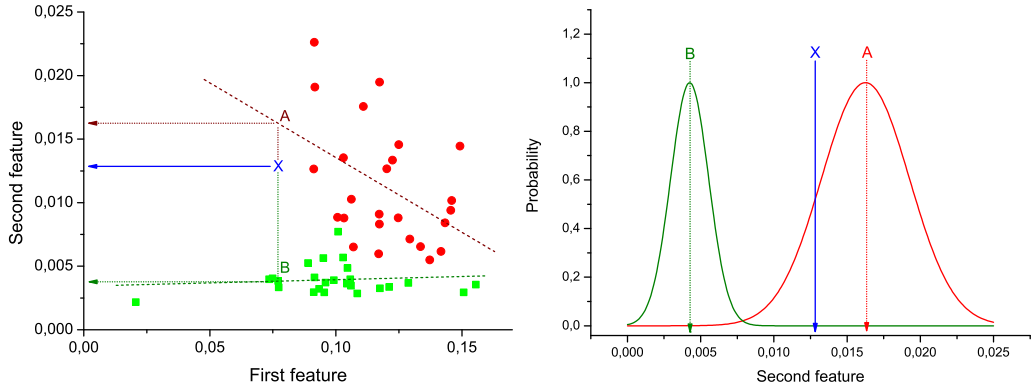


Figure 3.1: **Example of the calculation of link weight for two classes of subjects, using linear fits.** (Left) Lineal fits of two features corresponding to control subjects and patients; (Right) classification of an unlabeled subject (marked as *X*) into one of the two groups. Reprinted with permission from Ref. [ZPSEFRASEJRBMS13]

a function $\mathcal{F}^{c_1}(\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_{n_f}) = 0$ defining the feature combination associated to a perfectly healthy subject, a second function $\mathcal{F}^{c_2}(\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_{n_f}) = 0$ corresponding to disease c_2, *et caetera* [2]. The direct derivation and analysis of such constraints is usually impossible, due to the high dimensionality of the data, and to the presence of noise - recurring characteristics of any biomedical study. On the other hand, the problem can be simplified by analyzing the relationships between pairs of features in an independent way, for then merging all information into a network representation.

The complex network representation is created by projecting the $\mathbb{R}^{n_f}$ space into all possible $\mathbb{R}^2$ planes, each plane corresponding to a pair of features. For each pair of features i and j, the values corresponding to subjects of a given class c are used to create a projected constraint $\tilde{\mathcal{F}}^c_{i,j}(\boldsymbol{f}_i, \boldsymbol{f}_j) = 0$, representing the best constraint defining class c in the $\mathbb{R}^2_{(ij)}$ plane. Such function can be obtained by several methods, like for instance a polynomial fit, or more generally by any data mining method, *e.g.* Artificial Neural Networks.

An example may help clarifying this idea. Without loss of generality, let us consider the case of two groups of subjects, *i.e.* $c_n = 2$: control subjects and patients suffering from a given disease. When a pair of features are considered, the corresponding values can be plot in a plane, as depicted in Fig. 3.1 Left. Afterward, a linear fit can be calculated for both groups, as depicted by the dashed green and red lines. Mathematically, this is equivalent to defining $\tilde{\mathcal{F}}^{c_1}$ and $\tilde{\mathcal{F}}^{c_2}$ respectively as:

[2] An alternative interpretation can be constructed, by considering the evolution of a subject like a chaotic trajectory in a f_{n_f}-dimensional space, and his/her vector of features like a point in such trajectory. The collection of all points corresponding to control subjects, *i.e.* the previously defined constraints, would then become a *strange attractor*, whose characteristics can be studied by means of dynamical systems' theory [Str01]. While thought-provoking, such an approach is outside the scope of this PhD Thesis.

$$\tilde{\mathcal{F}}_{ij}^{c_1}: \quad \boldsymbol{f}_j^{c_1} = \alpha_{ij}^{c_1} + \beta_{ij}^{c_1} \boldsymbol{f}_i^{c_1} + \boldsymbol{\epsilon}_{ij}^{c_1} \tag{3.1}$$

$$\tilde{\mathcal{F}}_{ij}^{c_2}: \quad \boldsymbol{f}_j^{c_2} = \alpha_{ij}^{c_2} + \beta_{ij}^{c_2} \boldsymbol{f}_i^{c_2} + \boldsymbol{\epsilon}_{ij}^{c_2}. \tag{3.2}$$

In the previous equations, i and j represent the two features over whose plane the data are projected, while $\boldsymbol{f}^{c_1}$ ($\boldsymbol{f}^{c_2}$) represents the data corresponding to healthy subjects (patients). Furthermore, α is the slope of the lineal fit, β the intercept, and $\boldsymbol{\epsilon}$ the vector with the residuals of the fits. Notice that these two constraints $\tilde{\mathcal{F}}_{ij}^{c_1}$ and $\tilde{\mathcal{F}}_{ij}^{c_2}$ are just 2-dimensional approximations of the true constraints $\mathcal{F}^{c_1}$ and $\mathcal{F}^{c_2}$ as projected in the $\mathbb{R}^2_{(ij)}$ plane.

Once the reference constraints have been obtained (in the general case, n_c constraints have to be considered), the next step of the network reconstruction process deals with the analysis of a new unlabeled subject, and the identification of the constraint to which he/she is closer to. This information will then be used to create the network representation of the data corresponding to this subject; nodes will represent features, and links between pairs of them will be created if the subject is close to one of the identified constraints.

In what follows, we will consider two cases that are especially relevant in biomedical applications: situations in which two classes of subjects are present, *i.e.* $n_c = 2$, and those in which just one reference is known, *e.g.* when data about control subjects only is known in advance.

Scenarios with two classes

Let us first suppose that a set of subjects, belonging to two different classes, is available for training the reconstruction method. These may be control subjects and patients, as in the previous example; but also patients suffering from two different diseases, when characterizing the differences between these diseases is the aim of the study.

When the two constraints have been identified for each pair of features, it is necessary to assess the position of a new unlabeled subject in each plane, and specifically the likelihood of his/her belonging to one of the two classes. The more the relationship between the two features of the subject under analysis follows the pattern found in patients, the higher the likelihood for that subject to belong to the patient group. In order to reflect such relationship in the network representation, each pair of nodes should be connected with a *weight* proportional to such likelihood.

In Fig. 3.1 the position of an hypothetical subject is marked by a blue *X*. As the constraints have been estimated by means of linear regressions, two projections can be calculated, corresponding to the expected value of the second feature (*i.e.* j) given the value of the first (i). These two values are used to construct the graph of Figure 3.1 (Right). Two normal distributions are plotted, centered on the expected values calculated in the previous step, and with widths equal to the standard deviation of the corresponding vectors

of residuals ($\epsilon_{ij}^{c_1}$ and $\epsilon_{ij}^{c_2}$).

Taking into account the expected value of the second feature in both models and the corresponding expected error in the lineal fit, given by the standard deviation of residuals, the probability $\tilde{p}_{c_1}$ ($\tilde{p}_{c_2}$) for subject X of pertaining to the control (patient) group is proportional to the value of the corresponding normal distribution at the point defined by the second feature. As X must be classified into one of the two classes, the final probability of pertaining to the patient class is given by the normalization:

$$p_{c_2} = \frac{\tilde{p}_{c_2}}{\tilde{p}_{c_1} + \tilde{p}_{c_2}}. \tag{3.3}$$

This probability p_{c_2} is then used to set the weight of the link between nodes i and j. When such procedure is repeated for each possible pair of features, the result is a complete weighted network representing the health condition of the subject under analysis. We therefore now move from a feature representation (features of all subjects represented in the $\mathbb{R}^{n_f}$ space) to a subject representation, where one network is constructed for each subject, with nodes representing features.

Notice that such procedure can be easily extended to other techniques for the definition of the constraints $\tilde{\mathcal{F}}$. For instance, they can be detected by means of Probabilistic Neural Networks; in that case, the neural network will directly provide the probability for a new subject of belonging to one of the two constraints. Independently of the specific technique used, the result will be equivalent, *i.e.* a weighted network representing the health condition of the unlabeled subject.

Scenarios with a single class

A second scenario is also of interest from a biomedical point of view. Instead of being composed of subjects belonging to two different classes, the initial data used for training may represent only one type of subjects: for instance, it may include only data describing control subjects, without any information about specific diseases. When a new subject is analyzed, the network representation should be able to *(i)* assess whether the subject is healthy, and, if not, *(ii)* identify which are the elements (features) responsible for the pathological condition.

In this case, only one reference constraint can be extracted from the data, *i.e.* $n_c = 1$, and the weight of links should be defined as the distance of unlabeled subjects from that reference. An example of the construction of a network representation within this scenario is depicted in Fig.3.2; for the sake of clarity, this example has been constructed by including only three features.

Following the example of Eq. 3.1, we will here suppose that the projected constraints are estimated by means of linear regressions. For each pair of features i and j, $\alpha_{ij}^{c_1}$, $\beta_{ij}^{c_1}$ and ϵ_{ij} (as in Eq. 3.1) are the variables needed to describe the reference constraint. Now, suppose that a new subject p is available, whose features are $\boldsymbol{f}_p = (f_1^p, f_2^p, \ldots, f_{n_f}^p)$. The

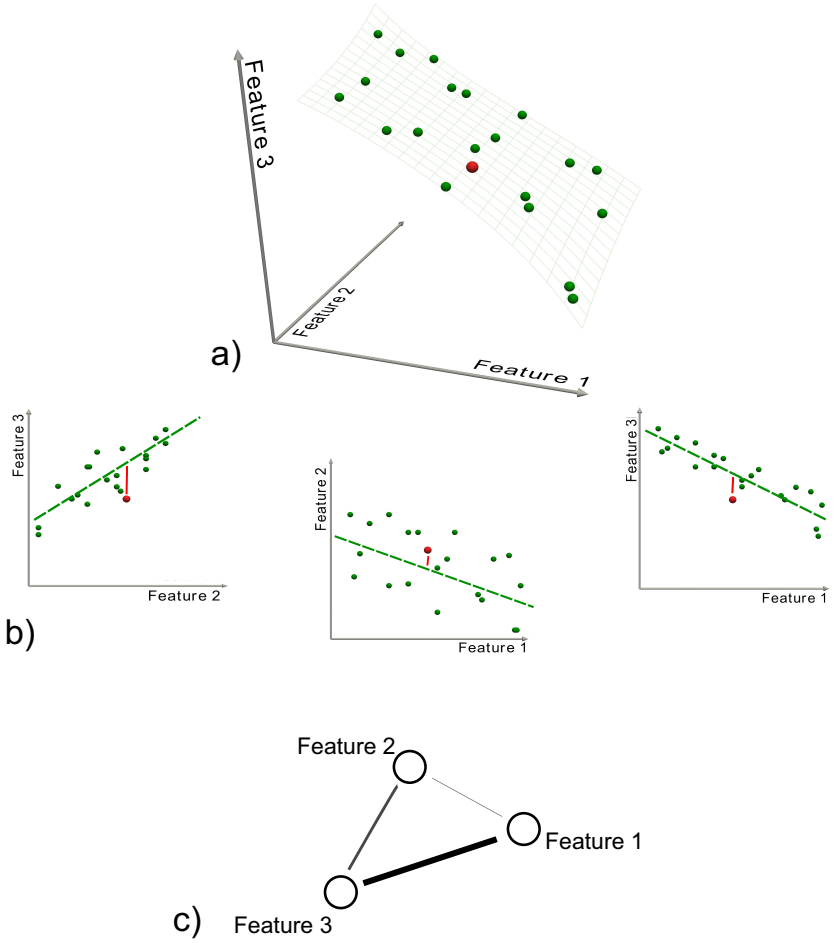


Figure 3.2: **Example of the network reconstruction procedure with one constraint.** (Top) Initial $\mathbb{R}^3$ space of features. Green (red) spheres represent the training set (the subject whose condition is to be analyzed). The gray hyperplane represents the ideal constraint $\mathcal{F}$. (Center) Projections of the 3-dimensional space into the three possible planes, with the green dashed line representing the estimated constraints $\tilde{\mathcal{F}}$. (Bottom) Resulting network representation. Link weight is represented by the thickness of the line.

distance of f_i^p and f_j^p from the reference constraint is given by:

$$e_{ij} = f_j^p - \tilde{f}_j^p, \tag{3.4}$$

$\tilde{f}_j^p$ being the expected value of f_j^p in the estimated constraint, *i.e.*,

$$\tilde{f}_j^p = \alpha_{ij} + \beta_{ij} f_i^p. \tag{3.5}$$

A value of e_{ij} higher than what found in ϵ_{ij} indicates that the relation between the two features i and j in the new instance p is different from what found in the reference group. In order to quantify this difference in a more exact way, a Z-Score is calculated as

follows:

$$Z_{ij}^{p} = \frac{e_{ij}}{\sigma_{ij}}, \tag{3.6}$$

σ_{ij} being the standard deviation of the elements composing ϵ_{ij}. Therefore, extreme values of Z (for instance, $|Z| > 2$) represent relations between pairs of features outside the normal expected range.

Finally, a network can be created for this instance p, by repeating the described process for each pair of nodes i and j - see Fig. 3.2 Bottom. All pairs of nodes are then connected with a link, whose weight is given by the Z-Score calculated for the corresponding pair of features.

3.2 Validation: Obstructive Nephropathy

Introduction

The previously described network reconstruction methodology has been applied to the characterization of vectors of features of subjects suffering from congenital *Obstructive Nephropathy* (*ON*) [Che99; WFJD98]. This pathology is the most frequent cause of renal failure in infants and children: the presence of an obstacle in the urinary tract prevents a normal urine flow, with a consequent accumulation within the kidney and progressive alterations of the renal parenchyma, development of renal fibrosis and loss of renal function. Fetal screening detects ON in 1 of 100 births, with at least 20% being clinically significant. In spite of the impact of this disease, pathological mechanisms are not yet fully understood, and no common therapy has been developed [Che06; MGTG75]; there is also a need for better biomarkers, measuring the severity of the obstruction and the response to medical or surgical interventions [Che04].

A dataset of genetic and metabolic features for control and ON subjects is here considered. This dataset was obtained through the *e-LICO multi-omics prediction challenge with background knowledge on Obstructive Nephropathy* [On], and includes information about 10 control and 10 ON subjects. Data provided cover both metabolism and genetic. Metabolic processes are represented by the levels of 852 metabolites, small molecules that are the intermediates and products of metabolism; the genetic information on individuals is represented by the expression levels of 834 miRNAs, *i.e.* short ribonucleic acid RNA molecules involved in the negative regulation of almost all biologic processes. Although they convey different information, both are indirectly related: miRNA is responsible for the repression of mRNA, which in turns defines the level of expression of different proteins, and consequently the levels of cellular metabolome.

Related work

While data mining has been extensively used for the analysis of biomedical data, few are the studies applying such techniques to Obstructive Nephropathy; this is probably caused by the scarcity of suitable data sets, and by the cost associated with their generation, as invasive analyses should be performed in newborns. Among the few available examples that can be found in the Literature, two of them [Kra11; VMC13] tackle the problem of classifying subjects by mining mRNA expression levels. Specifically, in Ref. [Kra11] Support Vector Machines with Gaussian Radial Basis kernels are used, reaching a classification precision of 89.1%.

Some other works are worth noticing. Ref. [VCMKSC10] applies a *k*-means clustering algorithm for the identification of relevant antibodies in urine; yet, the aim was just exploratory, and no classification is performed. In Ref. [DGFMCM10] authors propose a software for automated characterization of kidney biopsy images. Finally, in Ref. [MVKMSC11] data from several biological layers (transcriptomic, post-transcriptomic and proteomic) are represented by means of networks, the objective being an enhanced representation of data.

Results

The ON data set has been analyzed by means of the proposed network reconstruction method. Just one class was considered, composed of all control subjects, in order to highlight the elements (metabolites and miRNA) responsible for the disease.

In the ideal case, for networks corresponding to control subjects, we would expect all links to have a neglegible weight: as all data should perfectly fit the estimated constraints, the Z-Score obtained for any link in the network should be close to zero (see Eq. 3.6). Clearly, this will seldom happen, as measurement errors or small imperfections in the control group may introduce noise to the data. The consequence of this noise is that some links may gain a higher weight. Yet, if no pattern is present in the noise, these promoted links should form a random structure. On the other hand, when analyzing ON networks, the disease is expected to manifest mainly around some miRNA or metabolites, implying that most links, *i.e.* relationships characteristic of the disease, should gather around few hubs.

This intuition is confirmed in Fig.3.3, where four example networks are represented. Top networks correspond to two control subjects, and bottom networks to ON subjects. For the sake of clarity, only the 500 strongest links are displayed. ON subjects are usually associated with star-type graphs, where the central node is the most abnormal miRNA or metabolite. It is interesting to note that this central node is not constant, and changes according to the subject, suggesting that the cause of the disease may be different from person to person.

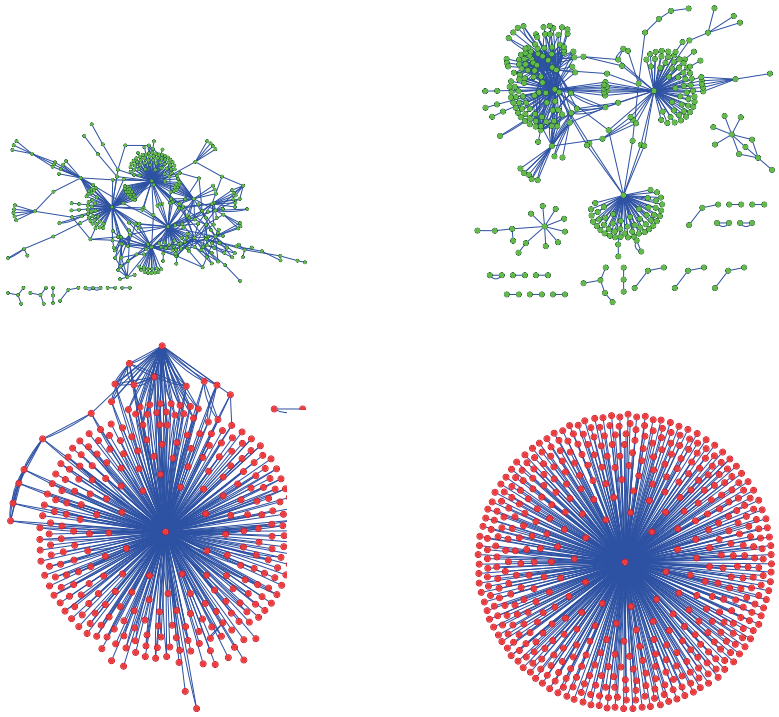


Figure 3.3: **Examples of four networks built from genetic and metabolic profiles.** Top green networks correspond to control subjects, bottom red networks to ON subjects; left images correspond to miRNA networks, while images on the right to metabolic networks. The different topologies reflect the presence or absence of the pathology. Reprinted with permission from Ref. [ZB11].

3.3 Validation: Glomerulonephritis

Introduction

As a second example of the application of this network reconstruction algorithm, we consider a data set of metabolic spectral measurements, corresponding to 25 control subjects and 25 patients suffering from *Glomerulonephritis* (*GN*). GN designates a group of renal diseases, characterized by an inflammation of glomerular capillaries, leading to a strong reduction in the renal function [Cou99]. These data are a subset of the information considered in Refs. [PKDSSB07; XPYW09; XMSBW12]. They include, for each subject of the two groups, a proton-NMR spectrum calculated from a urine sample; spectra have been filtered by removing water regions and drug peaks, and subsequently binned into 200 bins of 0.04 ppm width.

Unlike Obstructive Nephropathy, several works can be found in the Literature where data mining techniques are applied to the study of Glomerulonephritis. While a complete review is outside the scope of this Thesis, some applications are worth of note: identification of the causes of the disease [PWASMSPJF04; Wiw06], analysis of medical tests

with the aim of producing prognoses [EA08], evaluation of the effect of drugs [SBKFFS-ABM08], treatment monitoring [BLMBC02], or survivability analysis [YSLL10].

Results

The proposed reconstruction technique has here been applied by considering two classes of subjects, *i.e.* control and GN subjects, and linear fits for the identification of the constraints. The resulting network representations for four subjects, two of them of the control group (upper part, in green) and two of the patient group (lower part, in red), are shown in Figure 3.4. In order to simplify the image, only links with weight higher than 0.65 are represented.

By analyzing the four networks of Fig. 3.4, two features can be easily recognized. Firstly, the two networks of the control group have less links (lower link density) than the other two. This effect is to be expected, as data corresponding to GN patients should be closer to the disease constraints (see Fig. 3.1 Left), and therefore the weight associated to their links should be higher.

Secondly, while control subject networks lack of a clear structure, in the GN networks there is one or a few nodes with a central position, *i.e.*, concentrating most of the connections. This relationship between the network topology and the health condition of the subject is a consequence of the way the network is created, as previously explained when analyzing ON results.

To further analyze the differences between networks representing both groups of subjects, Fig. 3.5 presents the histograms corresponding to several network topological metrics: link density, Clustering Coefficient, and Efficiency [3].

The structural features that have been manually identified in Fig. 3.4 are now confirmed in the three histograms of Fig. 3.5. Specifically, all networks associated to GN subjects have a higher link density, with the value of 0.4 being a natural threshold for the classification of both groups. Furthermore, the clustering coefficient and the efficiency are lower in control subjects, indicating a more random structure; on the contrary, the *star-like* topology associated with GN subjects has a very high efficiency, as most nodes are connected by a path of length 2. Such results indicate that control subjects and patients must respectively be associated to *Poisson-like* and *scale-free* degree distributions, two important families of graphs that have been extensively studied in the last decades.

The presence of these two structures can also automatically be detected by analyzing the distribution of node centralities. *Centrality* is a general term that refers to the importance of a node in the network. Clearly, both in random graphs and regular lattices, each node is essentially equivalent to all other nodes, but when more complicated structures appear, one node may become especially important for the system. In this example we focus on the *eigenvector centrality*, which considers that a node has high importance if it is itself connected to other central positions [BL01]. Fig. 3.6 reports the four centrality

[3] For an exhaustive definition of the topological metrics used in this Thesis, the reader is referred to Annex A.

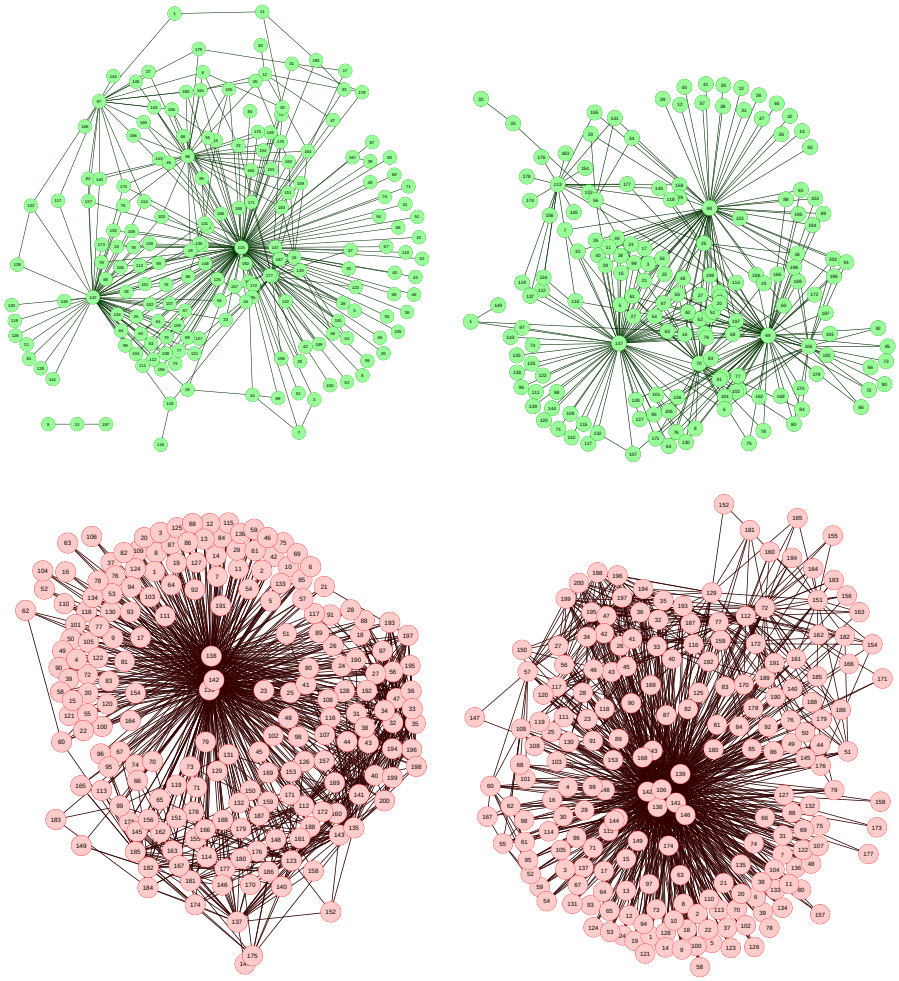


Figure 3.4: **Four examples of network representations of Glomerulonephritis spectral data.** Upper (bottom) networks represent control subjects (patients suffering from Glomerulonephritis). Notice how the two bottom graphs have a marked star structure, *i.e.* all nodes are directly connected with a central one, thus indicating that the latter is the responsible of the disease. On the other hand, the networks corresponding to control subjects have a more random topology, result of the presence of biological noise. Reprinted with permission from Ref. [ZPSEFRASEJRBMS13].

histograms corresponding to the four networks depicted in Fig.3.4. It can be noticed that control subject networks are characterized by a flatter centrality distribution, in which all nodes have medium importance; on the other hand, nodes in networks corresponding to GN patients have an average lower centrality, except for some highly important nodes.

The classification of subjects can be easily performed by any of the standard data mining algorithms available in the Literature, by using the extracted network features, *e.g.*, link density or efficiency, as inputs of the model. By using Support Vector Machines with link density and efficiency as input features, and a *leave-one-out* validation technique, a 100% score is easily achieved. It is worth noticing that the perfect classification has been reached directly from the raw spectral data, without the use of any of the standard

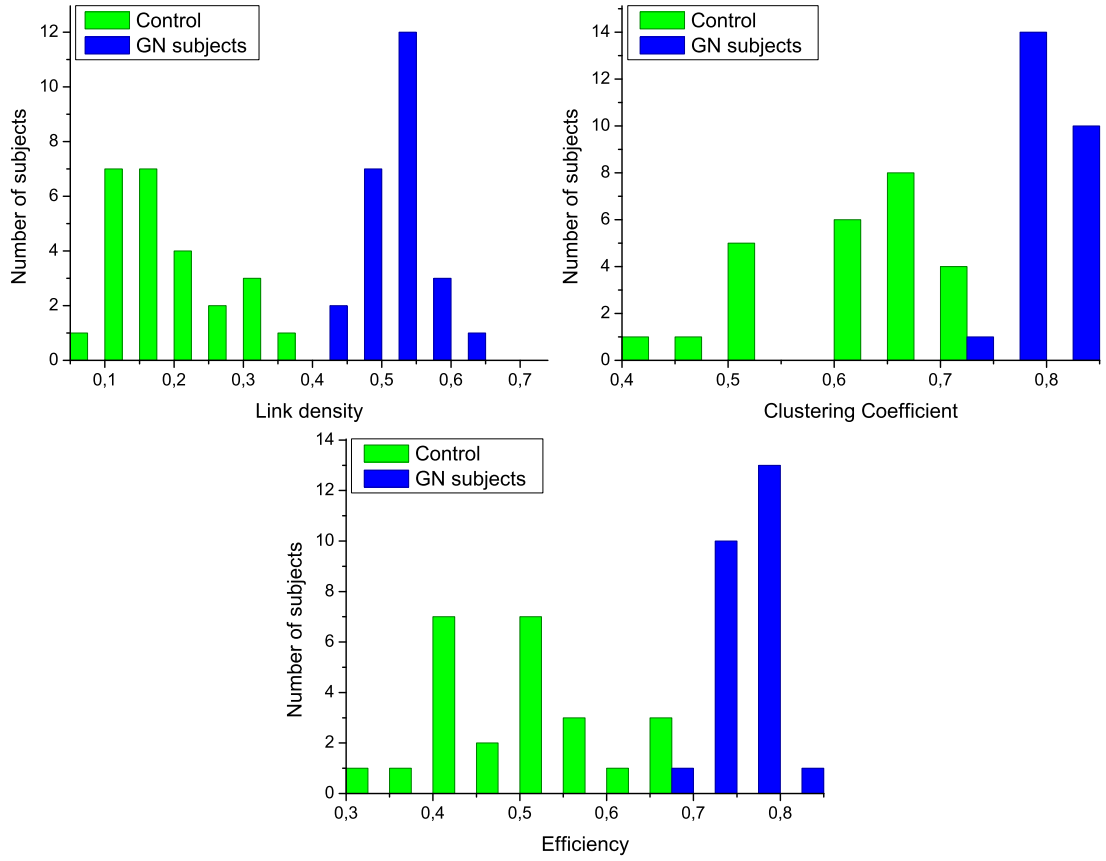


Figure 3.5: **Analysis of the structural characteristics of networks for control subjects (green) and patients (blue).** The three plots represent the histograms for (top left) link density, (top right) clustering coefficient, and (bottom) efficiency (see Annex A for definitions). Reprinted with permission from Ref. [ZPSEFRASEJRBMS13].

pre-processing techniques, like smoothing, baseline correction, or Principal Component Analysis [WJS98]. Indeed, one of the advantages of this complex network approach is that noise and other errors in the data may locally affect the structure of the network, but they do not affect the global properties of the resulting structure.

Furthermore, the analysis of the most central nodes in each network provides valuable information about the bins defining the health status of the subject. Specifically, in the networks here reconstructed, the most important nodes are number 138, 139 and 142, corresponding to segments centered in δ^1H 9.44–9.6 ppm (associated to CH and CHO signals).

To check the sensitivity of the proposed algorithm to the presence of noise, an ensemble of 100 modified data sets has been created, by polluting the original measurements with an additive noise drawn from a normal distribution centered in zero. Fig. 3.7 presents the average classification score of several algorithms as a function of the standard deviation of the noise. The proposed network representation outperforms the other three considered classification techniques, *i.e.* naive Bayes, decision trees and multilayer perceptrons, thus showing a great robustness against noise contamination.

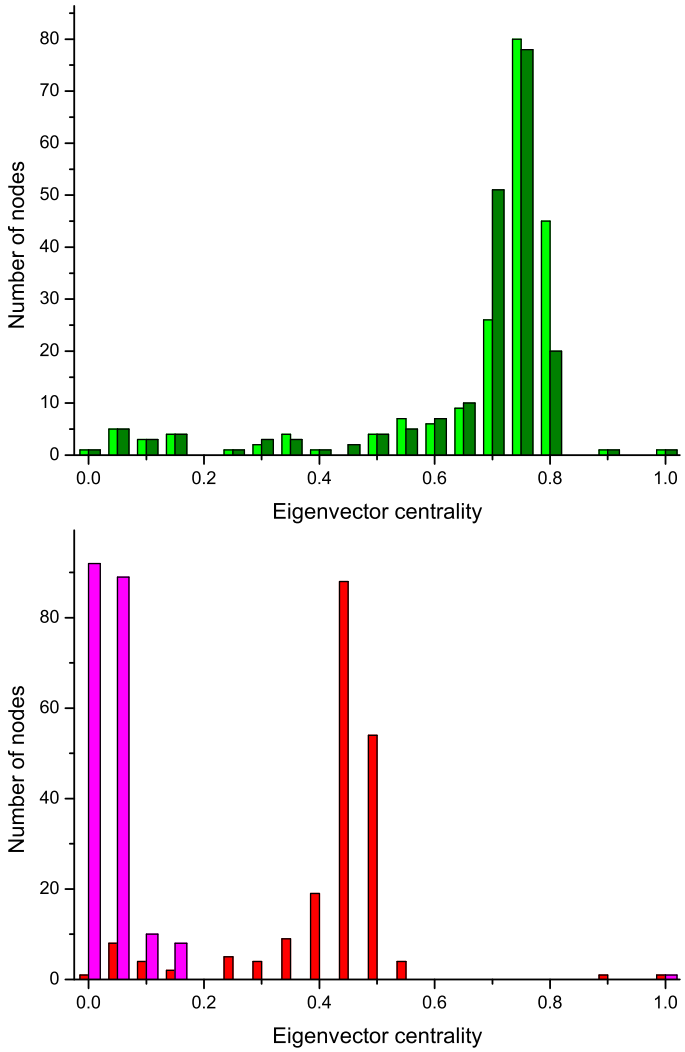


Figure 3.6: **Histograms of eigenvector centrality.** Results reported correspond to the nodes of the networks represented in Figure 3.4, *i.e.*, two control subjects (Top) and two GN patients (Bottom). Reprinted with permission from Ref. [ZPSEFRASEJRBMS13].

3.4 Validation: analysis of plant genetic responses

Introduction

As a third validation case, we here present the use of the methodology described in Section 3.1 to analyze gene expressions of the plant *Arabidopsis thaliana* under osmotic stress, with the objective of identifying those genes orchestrating the plant response under this specific condition. This is of particular relevance, as abiotic stresses represent the primary cause of crop loss worldwide, lowering by more than 50% the average yields of many crop plants. Therefore, a better understanding of the mechanisms behind plant responses to such stresses, starting from the genetic level, is essential.

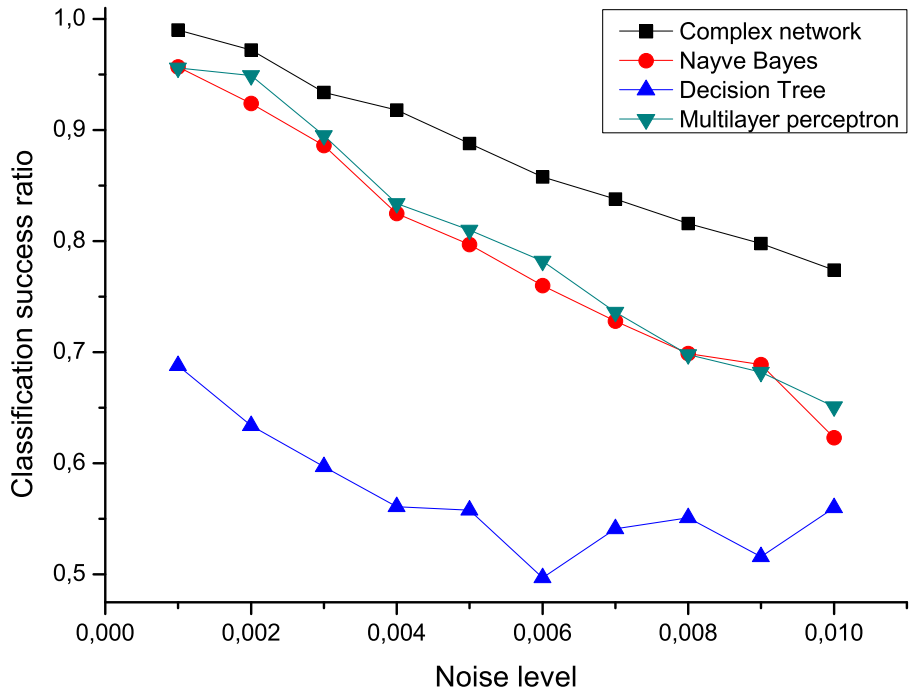


Figure 3.7: **Mean classification scores obtained by four algorithms for data sets polluted with additive noise.** The proposed network-based representation is indicated by black squares. Reprinted with permission from Ref. [ZPSEFRASEJRBMS13].

Similar data sets have been studied in the last decade by means of different techniques, *e.g.* co-expression networks [CLPSEMDW05; MVHDD09; BGBHRDS12] and differential-expression analysis [Bra02; SNINFOKNESSATYSCKHS02; KWWCESPDD02; KWHWWBDBBKH07]. Yet, we expect the proposed approach to yield complementary results. Specifically, differential-expression analyses only focus on the evolution of expression levels through time, considering each gene as independent from the others. Co-expression networks analyze similarities between the evolutions of pairs of expression levels. On the contrary, the network representation of Section 3.1 focuses on pairs of genes whose expressions depart from a reference model, thus it concentrates on differences. Furthermore, in marked contrast with classical approaches where a single network is obtained reflecting similarities across stages, in this representation the construction of a different network for each time step allows tracking the plant response through time.

Network reconstruction

The original data set corresponds to the *AtGenExpress project* [KWHWWBDBBKH07], including expression levels of $22,620$ genes under 8 different abiotic stresses (*i.e.*, cold, heat, drought, osmotic, salt, wounding and UV-B light) and at six different moments of time (30 min, 1 h, 3 h, 6 h, 12 h and 24 h after the onset of the stress treatment). Of these, only the osmotic stress is considered in this work, and the analysis is limited to the $n_f = 1,922$ genes composing the transcription factors of Arabidopsis represented in the ATH1 array

[GHLBGWL05].

Following the method previously described, the set of systems under analysis is here composed of the status of the plant at a given time step, each one described by a set of features representing the genetic expression of the plant. The objective of the study is the creation of a network representing the genes with an abnormal expression at each time step. In other words, when analyzing data at time τ, we create the $n_f(n_f-1)$ reference models $\{\tilde{\mathcal{F}}=0\}$ with the data corresponding to all other time steps, and we generate links according to the distance from that reference.

During the network reconstruction process it is necessary to define the general form of the reference model $\tilde{\mathcal{F}}$. Here, we have chosen the use of a simple linear regression between the expression levels of genes i and j, such that:

$$\tilde{f}_j^\tau = \alpha_{ij} + \beta_{ij} f_i^\tau, \tag{3.7}$$

$\tilde{f}_j^\tau$ being the expected value of gene j at time τ, f_i^τ the known expression levels of gene i, and α_{ij} and β_{ij} two free model parameters. These two coefficients are calculated by means of a linear fit of all values corresponding to other time steps, *i.e.*, minimizing the error of the relation:

$$f_j^{t\neq\tau} = \tilde{\mathcal{F}}(f_i^{t\neq\tau}) = \alpha_{ij} + \beta_{ij} f_i^{t\neq\tau}. \tag{3.8}$$

While more complex functions cound have been used for $\tilde{\mathcal{F}}$, the choice of a linear regression has been motivated by two considerations. First, genetic expression levels are customary transformed in order to have a linear behavior, and the calculation of linear correlations between them is a common procedure in the Literature [CLPSEMDW05; MVHDD09; BGBHRDS12]. Second, the reduced number of points available to fit the function $\tilde{\mathcal{F}}$ precludes the use of higher-order expressions, as this would result in an overfitting.

Furthermore, the reader should notice that the analysis here presented considers instantaneous interactions between genes, *i.e.* that the value of f_j^t (at time t) is only function of f_i^t, and not of the historical expression of gene i. In other words, when the 24 h expression levels are analyzed, we suppose that they are independent on the expression levels at 12 h. While this is clearly a simplification, the low temporal resolution of the available data set prevents a detailed analysis of the delayed influence of gene expressions.

The distance between the expected (corresponding to the model $\tilde{\mathcal{F}}(f_i^{t\neq\tau})$) and the real value of gene j is then used to weight the link connecting nodes i and j in the network. More specifically, the weight of the link is the absolute value of the Z-Score of the distance $\left|\tilde{f}_j^\tau - f_j^\tau\right|$.

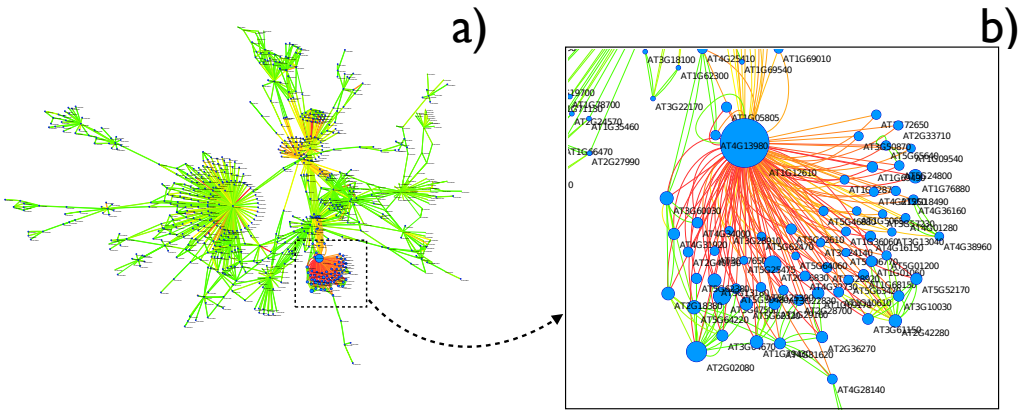


Figure 3.8: **Network for the response of Arabidopsis thaliana to osmotic stress after 3 h.** (a) Representation of the giant component of the network; for the sake of clarity, links with weight lower than 3 are not depicted. (b) Magnification of the neighborhood of the most central node, *AT1G12610*. Notice that labels are positioned in the lower right corner of each node - thus *AT4G13980* is the label of the small one on the left. In both cases, color represents the link weight (from green to red), and node size is associated with the corresponding value of α-centrality. Reprinted with permission from Ref. [ZMVGPSMB14].

Network analysis and validation

An example of the obtained networks is shown in Fig. 3.8. Namely, Fig. 3.8 (a) depicts the giant component of the network corresponding to 3 h after the onset of the osmotic treatment. The color of links accounts for their weights, with green (red) shades indicating low (high) Z-Scores, and the size of nodes is proportional to their $\alpha - centrality$ [BL01] - see Annex A for more details about this centrality metric. The resulting network topologies are characterized by a highly heterogeneous structure, dominated by a small number of *hubs* - as can be appreciated from the zoom reported in Fig. 3.8 (b). Such nodes with high centrality indicate that, at 3 h., the expression levels of the corresponding genes strongly deviate from the relationships generally established at other times. This suggests that hubs are performing some specific task at this time point, and therefore that they are key actors in regulating the overall plant response to that particular stress. The network representation allows identifying novel candidate genes, the full list of which is reported in Table 3.1, that were either previously unknown or were not considered to be related to the response to osmotic stress.

To confirm these predictions, we performed an *in vivo* screening, in which genes corresponding to the most central nodes of each graph were artificially induced in transgenic plants, and the derived phenotype after a stress response was monitored in a typical essay by measuring the length of the root of each plant. More specifically, the *Arabidopsis thaliana* inducible lines from Transplanta collection [Tra] were used, with the ecotype Columbia (Col-0) as the Wild Type. Each one of the transgenic *Arabidopis* lines of the

Time step	Gene	Name	Centrality
30 m.	AT1G13300	*HRS1*	0.88111
30 m.	AT5G51910	TCP family transcription factor	0.729679
30 m.	AT4G23750	*CRF2*, Cytokinin response factor 2	0.507826
1 h.	AT1G44830	*DREB*	1.0
1 h.	AT3G12820	*MYB10*	0.236686
3 h.	AT2G46830	*CCA1*, Circadian clock associated 1	0.271497
3 h.	AT5G62320	*MYB99*	0.177404
3 h.	AT1G29160	*COG1*	0.148112
6 h.	AT4G16610	C2H2-like zinc finger protein	0.767785
6 h.	AT2G44910	*ATHB-4*	0.689358
12 h.	AT3G61910	*NST2*	0.264721
24 h.	AT1G09540	*MYB61*	0.709785
24 h.	AT2G40950	*BZIP17*	0.551008
24 h.	AT5G62320	*MYB99*	0.482752
24 h.	AT5G04410	*ANAC078*	0.438538

Table 3.1: **List of new identified genes involved in osmotic stress responses revealed by the network representation.** Gene function was previously unknown in the Literature, and here experimentally proven to develop a statistically significant phenotype in response to osmotic stress. The right most column reports the corresponding α-centrality values. Reprinted with permission from Ref. [ZMVGPSMB14].

collection expresses a single *Arabidopsis* transcription factor under the control of the β-stradiol inducible promoter. For osmotic stress screening, seeds from control plants (Col-0) and at least two independent T3 homozygous transgenic lines (Transplanta collection [Tra]) of each transcription factor were sterilized, vernalized for 2 days at $4°$C and plated onto Petri dishes containing $\frac{1}{2}$ MS medium [MS62] supplemented with 10μM β-Stradiol. After 5 days, seedlings were transferred to vertical plates containing $\frac{1}{2}$ MS medium supplemented with 300 mM Mannitol, 10μM β-stradiol and transferred to a growth chamber at $21°$C under long-day growth conditions (16/8h light/darkness). After 12 days pictures were taken to record the phenotypes, and root elongation measurements were performed with ImageJ software [AMR04].

As an example, Fig. 3.9 reports the results obtained with seven transgenic lines, *i.e.* seven groups of plants in which the expression of one gene, corresponding to a hub, was artificially induced. Specifically, Fig. 3.9 (a) reports the mean length of roots for the seven lines, as compared to the normal root length in the wild type (black column) grown under osmotic stress conditions. The figure clearly visualizes the fact that, in all the seven examples, the induction of the corresponding gene leads to a significant functional response in the development of the plant. The results of the *in vivo* screening are summarized in Fig. 3.10. For each of the six networks analyzed, only the 20 most central genes at each time step were considered. This figure reports the number of genes already known to be relevant for the osmotic response of the plant, and the number of previously unknown genes disclosed by the network representation that have successfully been confirmed.

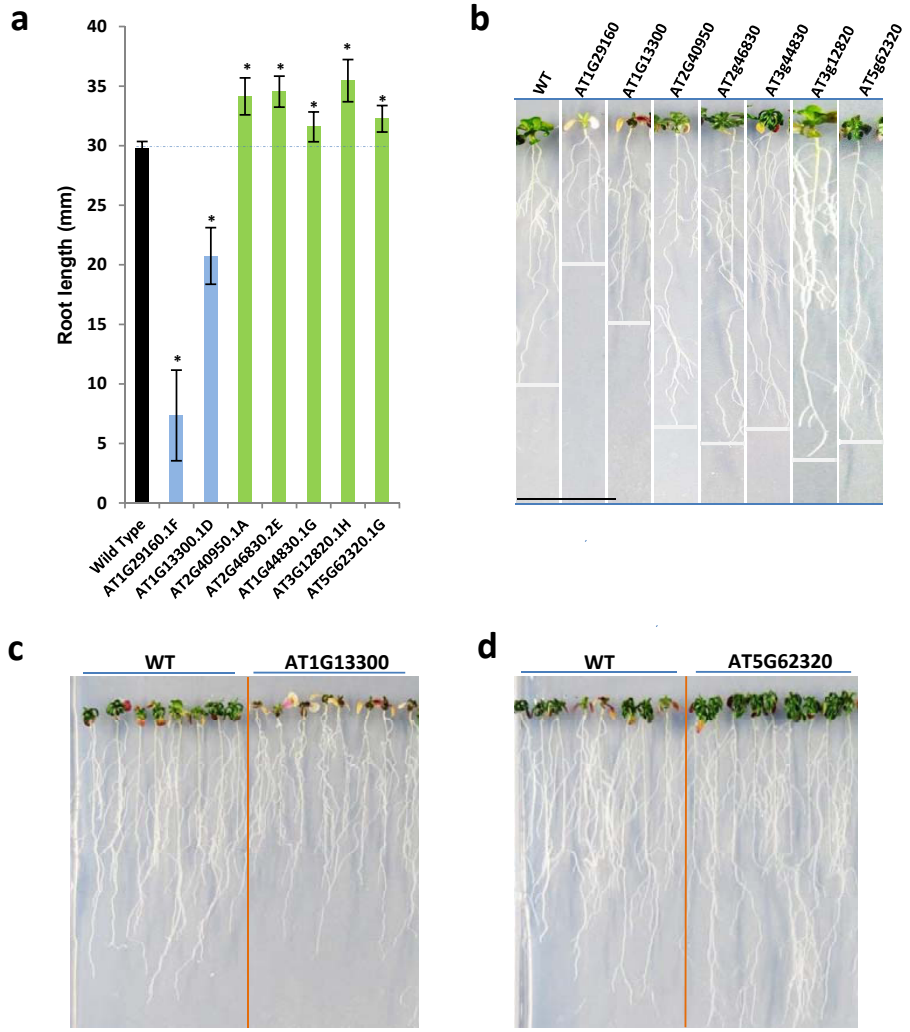


Figure 3.9: **In vivo experimental verification of the predictions.** (a) Mean root length corresponding to the wild type (WT, black column) and to 7 other transgenic lines in which a specific gene has been artificially induced. Whiskers represent the standard deviation corresponding to each group. Asterisks denote groups for which the distribution of root lengths is different with respect to the wild type with a 0.01 significance level. (b) Photos of one plant of each of the 8 lines, at the end of the full development process. (c) and (d) Photos of two vertical plates where plants are grown. In both cases, the left (right) photos refer to wild phenotypes (to phenotypes developed by the transgenic line). Reprinted with permission from Ref. [ZMVGPSMB14].

Thus, the use of a network representation allows the prediction -and further experimental confirmation- of key transcription factors that were not detected using alternative methodologies.

3.5 Conclusions

In this Chapter we have developed a novel way of constructing network representations starting from collections of isolated scalars. Values corresponding to unlabeled subjects

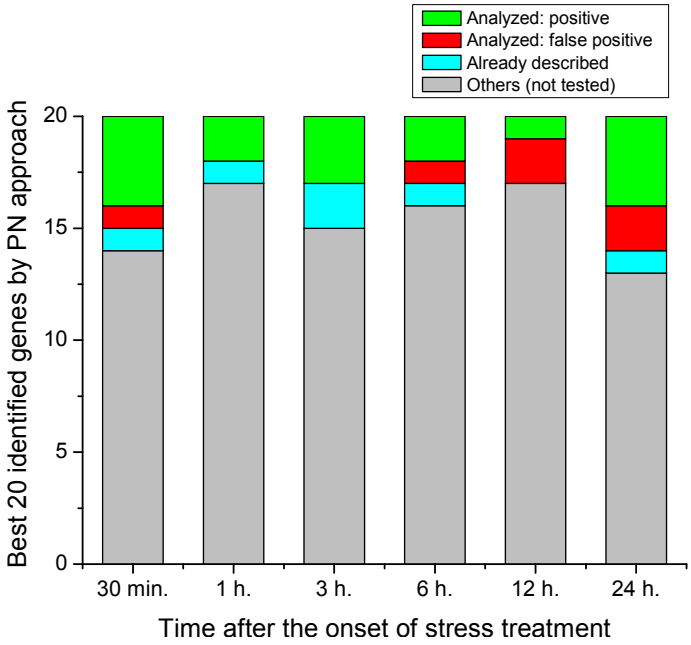


Figure 3.10: **Outcome of the experimental results.** Bars account for the 20 most central genes at each time step. For the six time steps considered, bar colors are coded according to the following stipulations: genes previously considered not to be involved in the plant's response to osmotic stress, that were respectively experimentally proven to develop (green) or to fail to develop (red) a statistically significant difference in the phenotype with respect to the wild-type phenotype; (cyan) genes predicted by the parenclitic analysis that were previously associated with the stress response in the Literature; and (gray) previously unknown genes, which could not be tested experimentally, due to their unavailability in the TRANSPLANTA collection. Reprinted with permission from Ref. [ZMVGPSMB14].

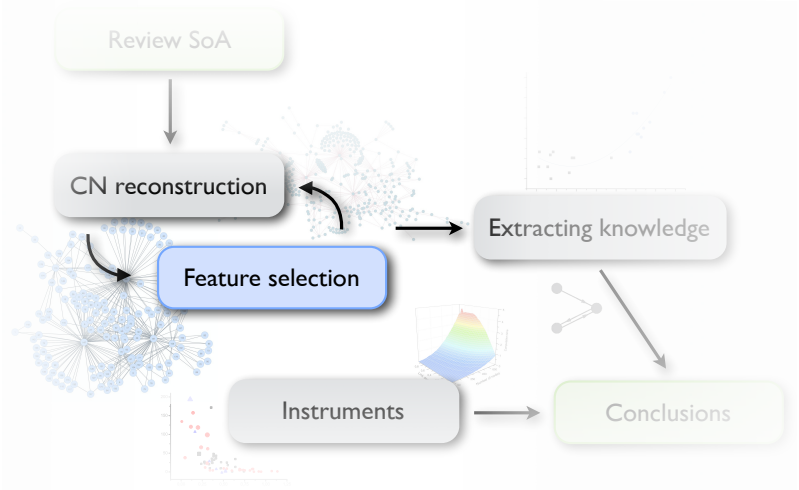
are compared with *constraints* representing the expected relationships between pairs of features in the analyzed population. The resulting networks display some clear topological characteristics that make possible a straightforward identification of control subjects and patients, and that allow highlighting which are the element responsible for the disease.

It is worth noticing the relevance and the novelty of such approach: for the first time, genetic and metabolic expression levels can be represented as a network structure, thus allowing unleashing all the power of complex network theory for the study of diseases like Obstructive Nephropathy and Glomerulonephritis. Even when network theory was already used, the proposed approach makes possible the analysis of the evolution of these systems through time, as this information is not destructed in the reconstruction phase: it thus represent a complementary approach to classical functional networks.

In the next Chapters, several data mining techniques will be applied to such resulting network representations, with the aim of improving the quantity of knowledge that can be extracted from them.

4

Reducing the dimensionality of the system



Chapter 3 presented a new network reconstruction methodology, aimed at handling systems described by sets of scalar observables. The most important drawback of such approach is its computational cost, especially because most biomedical problems are characterized by a high dimensionality. For instance, genetic information of different subjects usually includes thousands of expression levels; spectral analyses of biological tissues also yield thousands of measurements.

Before proceeding to the network analysis, which will be the focus of Chapter 5, we

here tackle the problem of reducing the dimensionality of the initial data set by means of the application of *iterative feature selection* algorithms [GE03; GGNZ06]. While reducing the number of nodes is not considered a normal praxis, from a data mining point of view it yields important benefits. First, a significant reduction in the computation cost: as most network metrics complexity scales with the square, or with the cube of the number of nodes, even small reductions in the initial information size have important consequences in the processing time. Furthermore, the elimination of *noisy* nodes may result in an overall improvement of the accuracy of the method and in a reduction of the dimensionality of the problem, thus enhancing the statistical significance of results. The contribution of this Chapter is therefore twofold. On one side, demonstrate that the use of feature selection algorithms is feasible and useful in the analysis of complex networks, specifically when the aim is to perform a classification task; on the other side, propose a methodology to select the best algorithm among the vast set available in the Literature, using the final classification score as an indicator of the quantity of information lost in the process.

This Chapter is organized as follows. In Section 4.1 three different feature selection algorithms are contemplated, chosen according to the characteristics of the data sets analyzed in this work. The effectiveness of these algorithms is assessed through two validation cases: the analysis of data sets corresponding to Obstructive Nephropathy (Section 4.2) and to different types of cancer (Section 4.3). Finally, some conclusions are drawn in Section 4.4.

4.1 Feature selection methods

Three different feature selection strategies have been examined in this work: *i)* binning the data according to sequential, non-overlapping regions; *ii)* analyzing the goodness of fit in the network creation process; and *iii)* filtering features according to their *Mutual Information*. These strategies have been selected due to their relevance for the data under study and for the network reconstruction procedure described in Chapter 3. While in the former method all bins are equivalent in their relevance, in the two latter cases the selection process starts by creating a *ranking* of features, from which a desired number of elements are then drawn.

4.1.1 Binning the data

The first feature selection technique here considered is the *binning* of the data set, a technique widely used in the analysis of metabolic spectra [GWSCRN01; BEBEKAHLN03]. The original spectra are divided into sequential, non-overlapping regions; each one of these regions is converted into a new feature, whose value corresponds to the average of all measurements included in it. Fig. 4.1 reports an example of such binning process, including the original process (Left) and the resulting feature set (Right).

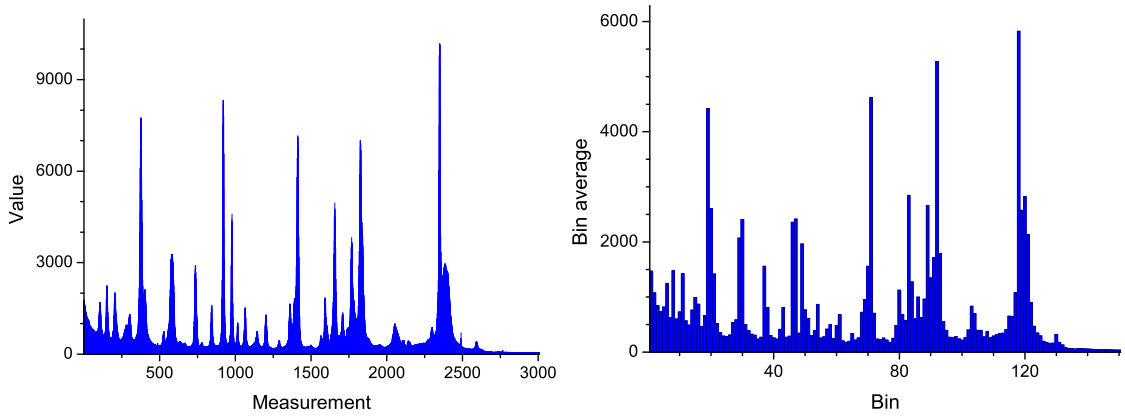


Figure 4.1: **Example of the spectra binning process.** (Left) Original spectrum, composed of 3000 measurements. (Right) Result of the binning process, where 150 non-overlapping windows are considered.

While this strategy represents a very simple way of reducing the dimensionality of the system, it does not guarantee that important information is retained; indeed, if such information is codified by a single measurement, it may be lost, or at least diluted, in the binning process.

4.1.2 Goodness of constraint models

The second method here considered has been specifically designed to improve the network reconstruction technique presented in Chapter 3, and is based on a measure of the goodness of the constraint models obtained for each class of subjects. When such constraints are calculated by means of linear fits (see for instance Eq. 3.2), we can derive their goodness by means of their corresponding Pearson's coefficient of determination R^2 [ST60]:

$$R^2_{ij} = 1 - \frac{\sum_s \left(f^s_j - \tilde{f}^s_j \right)^2}{\sum_s \left(f^s_j - \overline{\boldsymbol{f}}_j \right)^2}, \tag{4.1}$$

where i and j are the two features being analyzed, $\overline{\boldsymbol{f}}_j$ the average value of feature j, and $\tilde{f}^s_j$ is the value of feature j for subject s as estimated by the linear fit, *i.e.*:

$$\tilde{f}^s_j = \alpha_{ij} + \beta_{ij} f^s_i. \tag{4.2}$$

R^2 usually lays between zero and one [1], with $R^2 = 1$ meaning that the two features are perfectly described by a line.

[1] R^2 may be negative only in the case of wrong linear fits.

Using this metric, a value S is assigned to each feature, defined as:

$$S_i^G = \frac{1}{n}\sum_k R_{ik}^2. \tag{4.3}$$

S^G represents the quantity of information that is codified by the feature under analysis. Specifically, suppose that the feature i just codifies noise; in such case, the R^2 corresponding with the linear fit of feature i with any other feature will be close to zero, yielding a low S_i^G. On the other hand, perfectly correlated features in the reference group may be successfully used to detect anomalies in unlabeled subjects. Therefore, the ranking is created according to the value of S^G, and networks are constructed by including features with the highest S_i^G.

4.1.3 Mutual information

Finally, we here consider the application of the *Mutual Information* (MI for short), a well-known measure of mutual dependence between random variables [Kar03], which has extensively been used for the selection of relevant features in a data set - see, for instance, Refs. [YP97; Fle04; PLD05]. Given two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$, their two marginal probabilities distribution functions, $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$, and the joint probability distribution function $p(\boldsymbol{x}, \boldsymbol{y})$, the mutual information I between $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as:

$$I_{\boldsymbol{x},\boldsymbol{y}} = \sum_{l=1}^{m}\sum_{k=1}^{m} p(x_l, y_k)\log_2\left(\frac{p(x_l, y_k)}{p(x_l)p(y_k)}\right). \tag{4.4}$$

I measures, in bits, how much information is shared by the two variables, *i.e.* how much the knowledge of one of them reduces the uncertainty about the other. In order to rank each feature included in the original data set, we create a metric assessing the average information shared by one feature with all other features:

$$S_i^{MI} = \frac{1}{n}\sum_k I_{i,k}. \tag{4.5}$$

At this point, there are two different possible approaches for selecting features based on their value of S^{MI}. The first one, also known as the principle of *minimal redundancy* [PLD05], states that selected features should share the minimum amount of information between them, thus ensuring that the addition of a new feature provide new information to the classification process. This is equivalent to selecting features with small S^{MI}, or to sorting them in an increasing order of S^{MI}. On the other hand, one may expect several features to include a high quantity of noise, like for instance in the case of measurements obtained through mass spectrometry. When a measurement is representing noise, and thus no valuable knowledge for the analysis, the quantity of information it shares with other measurements is expected to be small. Therefore, features with low S^{MI} may codify no relevant information, while those associated with high S^{MI} may form groups of

highly correlated, and yet meaningful features.

Following these criteria, two different strategies are here compared for selecting features based on Mutual Information: respectively selecting nodes with highest S^{MI}, and nodes with lowest S^{MI}.

4.2 Validation: Obstructive Nephropathy

Introduction

As a first validation case, we analyze again the *miRNA* expression levels data set corresponding to subjects suffering from *Obstructive Nephropathy*, as described in Section 3.2. In this case, instead of using the whole data set for reconstructing the network representations, genes are selected according to two of the feature selection strategies previously described, *i.e.* according to their S^G and S^{MI} rankings.

Due to the high dimensionality of genetic expressions data sets, there is a vast Literature dealing with the application of feature selection algorithms to such information. Most of the works deal with the comparison of different feature selection strategies, and with their performance in different classification tasks - see for instance Refs. [XJK01; LLW02; LZO04; LCJM04; ILBC04; DP05; BMB07].

Results

Fig. 4.2 presents the quantity of information codified in the reconstructed networks as a function of the number of features included in the analysis, *i.e.* the number of nodes in each network. Such quantity of information is assessed through the correlation found between the network structure, measured by means of its *efficiency* (see Annex A for the definition), and the pelvic diameter, a proxy for the severity of the disease. The higher such correlation, thus, the more effective is the network in representing the health condition of the patient.

The complete evaluation of both algorithms is performed as follows. First, for a given set of features, a network is created for each person (both healthy and ON); afterwards, the *efficiency* of each network is computed, and these latter values are fitted against the pelvic diameter by means of a second-order polynomial. Finally, the goodness of all the process (and, thus, the relevance of the selected features) is estimated through the coefficient of determination R^2 of the fit.

As can be noticed from Fig. 4.2, both algorithms perform well in selecting the relevant features under which the results of the analysis are significant. An optimal result is obtained with a small number of features: the maximum of the R^2 corresponds to 300 features for S^G, and 280 for S^{MI}. Two thirds of the initial features have been eliminated, thus reducing the computational cost by a factor of 10. Furthermore, and not surprisingly, the highest score achieved by both methods is higher than the score obtained by analyzing the whole dataset. This is due to the nature of the feature selection, as the least

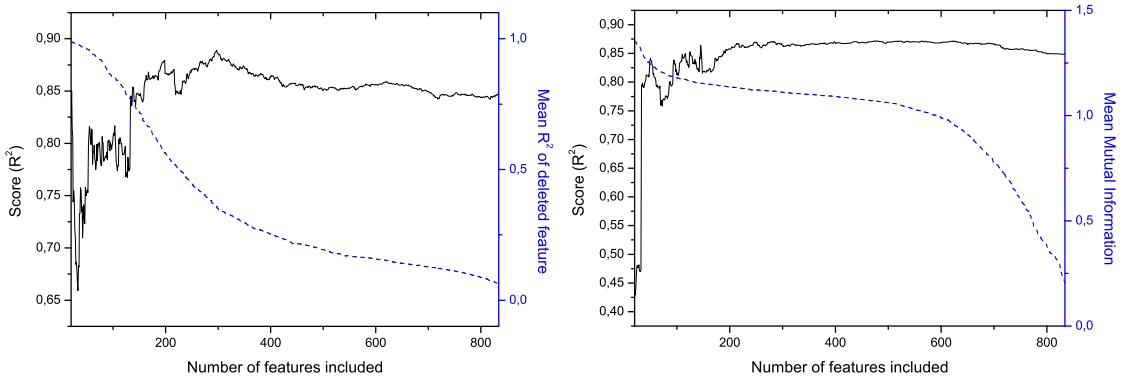


Figure 4.2: **Performance of the two considered feature selection algorithms for the ON data set.** Left (Right) graph shows the results corresponding to a S^G (S^{MI}) selection strategy. Black solid lines represent the quantity of information codified in the network as a function of the number of features included in the analysis (see text for definitions). Blue dashed lines indicate the value of the metric (R^2 and Mutual Information MI) associated with the feature included in each step. Reprinted with permission from Ref. [ZMSB12].

important features, which are not codifying relevant information, are excluded from the analysis.

In Fig. 4.3 we represent the second-order polynomial fits corresponding to networks created with five different numbers of nodes, selected according to the mutual information criterion. It can be noticed that the last two plots, corresponding respectively to 280 and 834 nodes, depict a clear relation between the characteristics of the networks and the severity of the disease, *i.e.* the pelvic diameter.

4.3 Validation: ARCENE data set

Introduction

A further validation of the effectiveness of feature selection algorithms in the reconstruction of network representations has been performed against the *ARCENE* data set, as used in the NIPS 2003 feature selection challenge [GGBHD04]. The training part of this data set included information for 100 subjects, 56 of them being control (healthy) subjects and 44 corresponding to subjects suffering from different kinds of cancers. Each one of them is described by a vector of 10.000 measurements, representing mass-spectra obtained with the SELDI technique [IVCF02]. Besides of the large number of measurements available for each subject, the challenge behind this data set resides in the presence of different types of cancers, specifically ovarian and prostate cancers [PIAHLFSMSFKL02; POPAHHVTWWSLMEBSKL02; AQDWCCSSYFW02]. While its study may yield results that are generic of the separation cancer *vs.* control across various cancers, it also requires the classification method to take into account potential differences in disease, gender, and sample preparation.

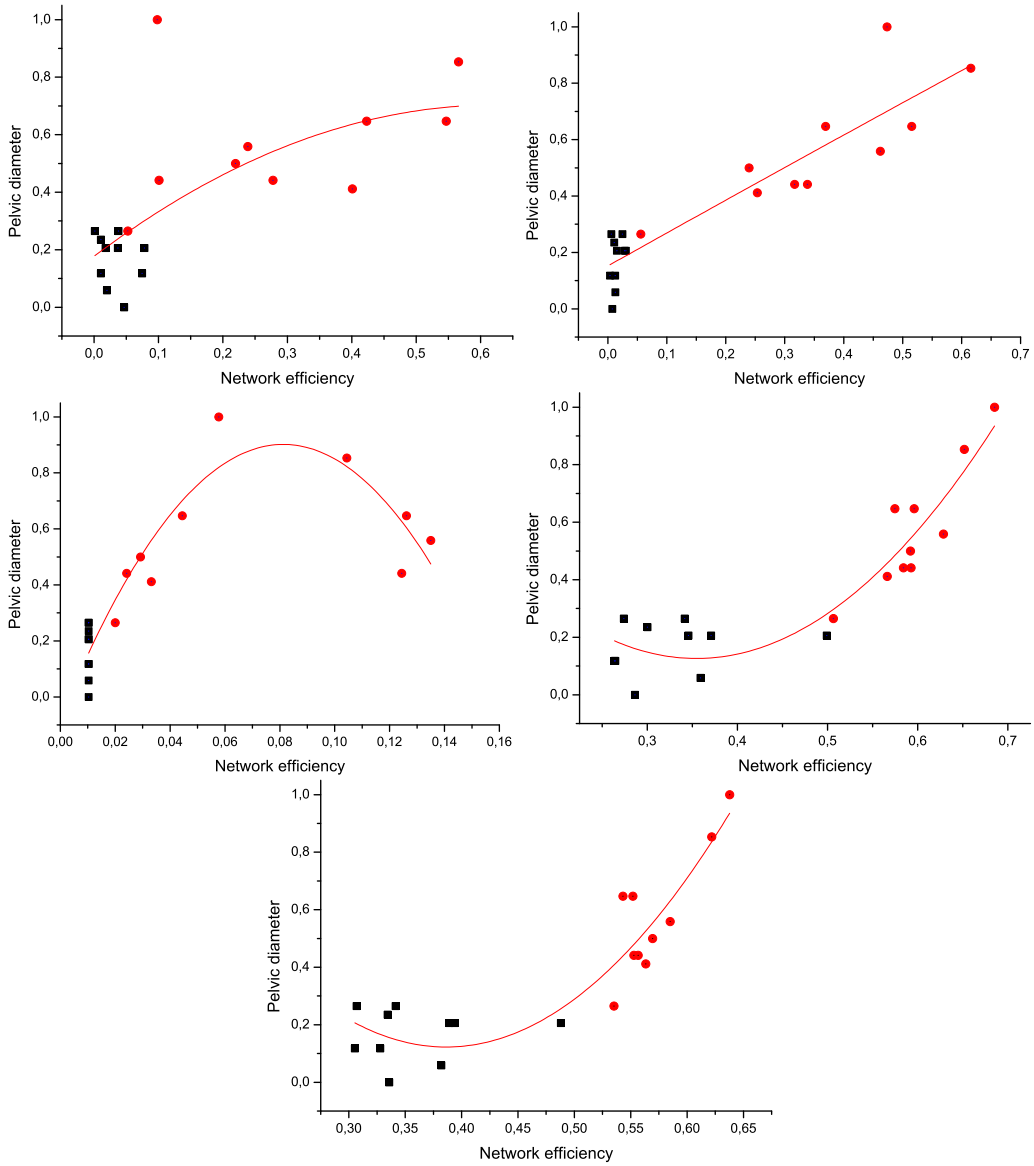


Figure 4.3: **Relation between ON severity and the network structure for different numbers of features.** From left to right, top to bottom, 28 ($R^2 = 0.482$), 50 ($R^2 = 0.841$), 145 ($R^2 = 0.864$), 280 ($R^2 = 0.871$), and 834 (the full dataset, $R^2 = 0.850$) nodes. Black squares (red circles) represent values of control (ON) subjects. Reprinted with permission from Ref. [ZMSB12].

In this validation case, we assess the relevance of the obtained networks by means of a classification task. For each network, a set of topological features are extracted, and different classification algorithms are then trained to discriminate control subjects from patients. Notice that the use of network characteristics for classification tasks will be the scope of Chapter 5.

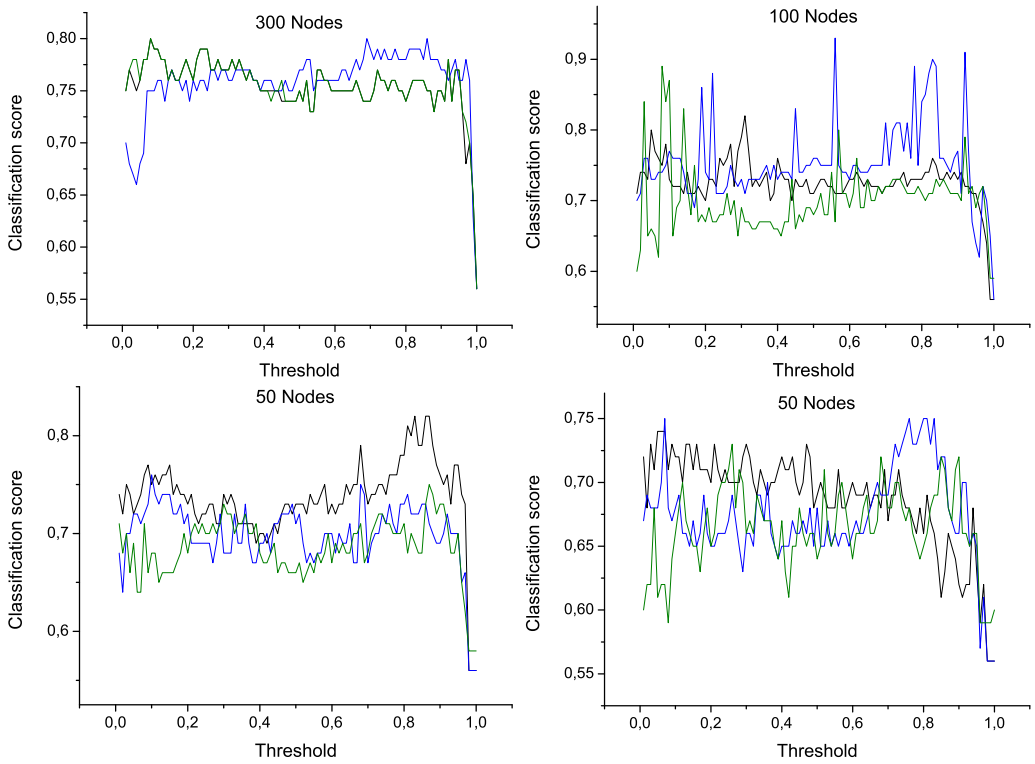


Figure 4.4: **Classification scores as a function of the number of nodes and of the applied threshold.** Black, blue and green lines respectively represent the classification score (*precision*) obtained by averaged bins, and by measurements selected in decreasing and increasing S^{MI} order. Reprinted with permission from Ref. [ZMBS13].

Results

Fig. 4.4 shows the classification score, expressed by means of the *precision* of the classification, as a function of the applied threshold, *i.e.* the number of links included in the networks [2]. Specifically, each image composing Fig.4.4 reports the results corresponding to the four network sizes here examined: from left to right, top to bottom, 300, 100, 50 and 25. Furthermore, inside each graph the three lines represent the score associated to the network representation created by means of the three feature selection algorithms contemplated: average binning, measurements with high S^{MI}, and measurements with low S^{MI}. In this case, the selection has been performed by means of a *Support Vector Machine* algorithm [Bur98; Ham11], due to its simplicity and its effectiveness in identifying relevant network metrics [ZSPBGPPMB12].

Fig. 4.5 reports the quality of the classification expressed in terms of the *F-measure* [Pow11], defined as:

$$F\text{-}measure = 2\frac{precision \cdot recall}{precision + recall}, \tag{4.6}$$

[2]The importance of the threshold applied in the network reconstruction phase will be deeply studied in Chapter 5

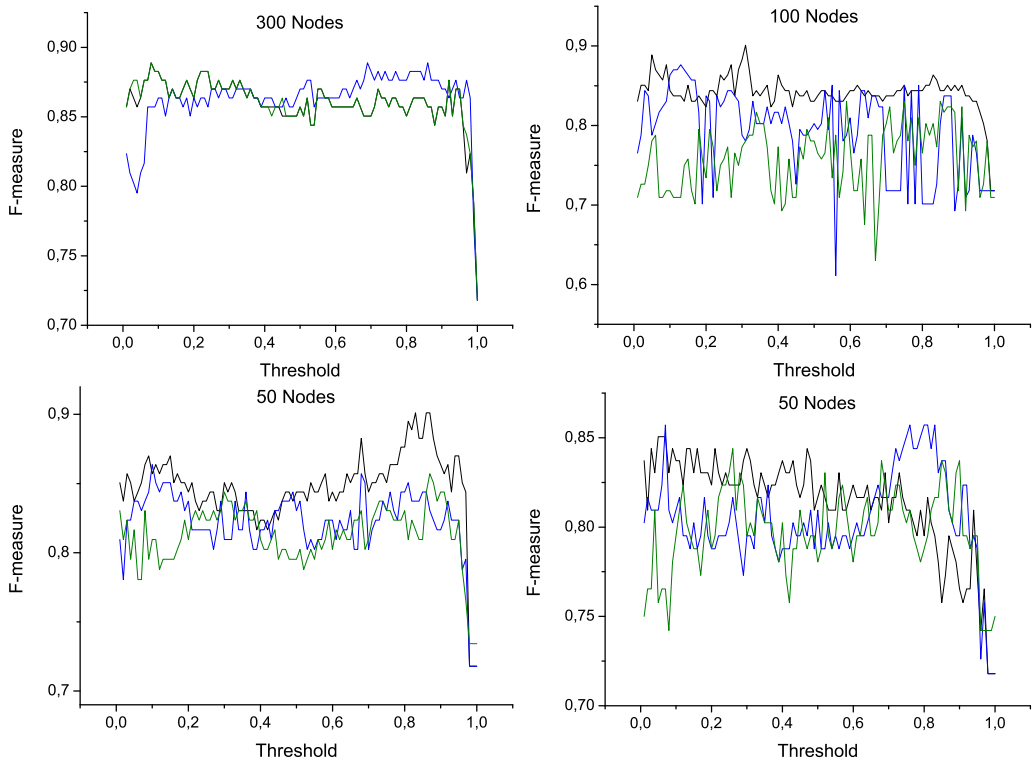


Figure 4.5: **F-measure as a function of the number of nodes and of the applied threshold.** Black, blue and green lines respectively represent the classification score (*precision*) obtained by averaged bins, and by measurements selected in decreasing and increasing Mutual Information order. See main text for the definition of the *F-measure*. Reprinted with permission from Ref. [ZMBS13].

recall being the number of correct results divided by the number of results that should have been returned. While some minor differences can be detected, especially in the behavior of the classification with 100 nodes, a general agreement between Figs. 4.4 and 4.5 is observed.

In order to validate such results, and exclude their dependence on the chosen classification algorithm, Fig.4.6 represents the classification score obtained by means of *Probabilistic Neural Networks* [Cla88; Spe90]. In this case, the result is given as the area under the ROC curve [ZC93], representing the performance of binary classifier systems whose output is expressed as a probability.

Finally, Table 4.1 reports a resume of the results, *i.e.* the best classification score obtained as a function of the number of features included in the analysis and of the feature selection algorithm applied. Several relevant conclusions can be drawn from these results.

First of all, reducing the number of measurements included in the analysis improves the classification score. Reducing the dimensionality of the data set under analysis allows limiting the quantity of noise, *i.e.* of irrelevant information, included in it, thus simplifying the classification task. Furthermore, reducing the number of features beyond

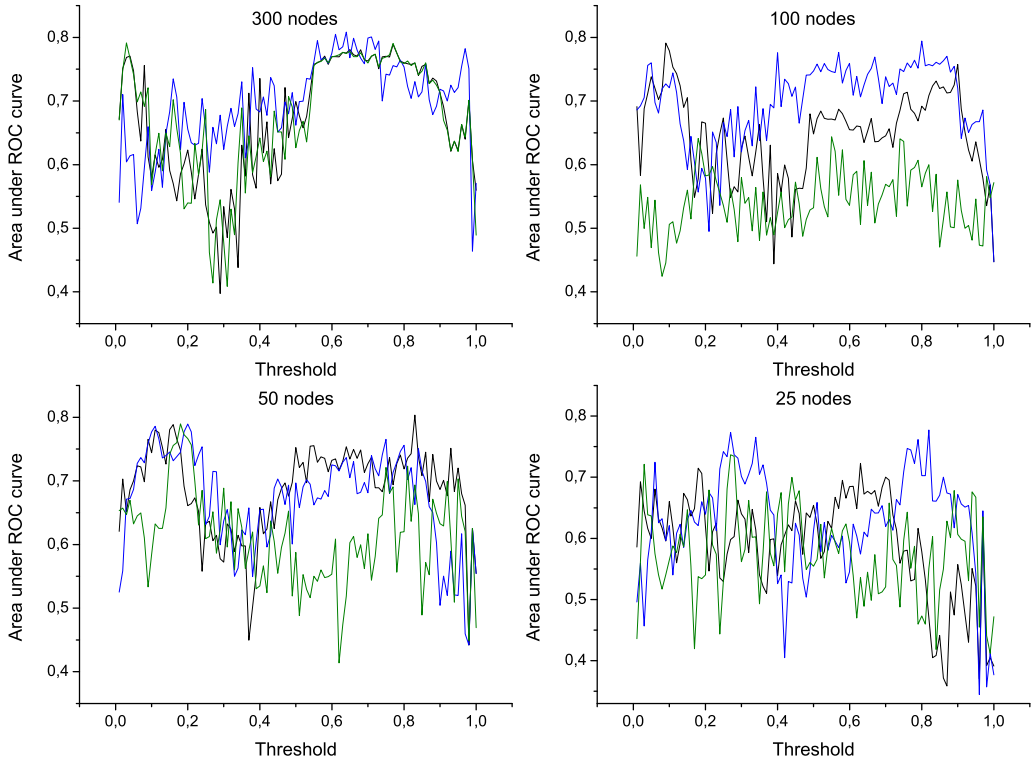


Figure 4.6: **Area under the ROC curve, as a function of the number of nodes and of the applied threshold.** Black, blue and green lines respectively represent the classification score (*precision*) obtained by averaged bins, and by measurements selected in decreasing and increasing S^{MI} order. Reprinted with permission from Ref. [ZMBS13].

Table 4.1: **Resume of classification results for the ARCENE data set.** Reprinted with permission from Ref. [ZMBS13].

	Binning	**High** S^{MI}	**Low** S^{MI}
300 nodes	0.8	0.8	0.8
100 nodes	0.82	0.93	0.89
50 nodes	0.82	0.76	0.75
25 nodes	0.74	0.75	0.73

a given threshold results in a drop in the effectiveness of the classification; this also has to be expected, in that important information for the task may be deleted. Such threshold is higher in the case of MI-based feature selection algorithms, which display their maximum for networks of 100 nodes.

MI-based feature selection algorithms are more effective than a feature reduction based on binning, as shown by the higher classification scores (0.93 vs. 0.82). This indicates that creating bins by averaging the measurements inside sequential regions, while

a common practice in the study of biological spectra, may result in the deletion of important information, which can be codified in very small windows or even in single measurements. While MI-based feature selection strategies always yield better results, the best solutions are obtained by selecting measurements with higher S^{MI}. Therefore, the important information is codified within few measurements that are highly correlated between them; on the contrary, selecting measurements according to a *minimal redundancy* strategy seems to introduce a high amount of noise in the classification task, reducing the discrimination power.

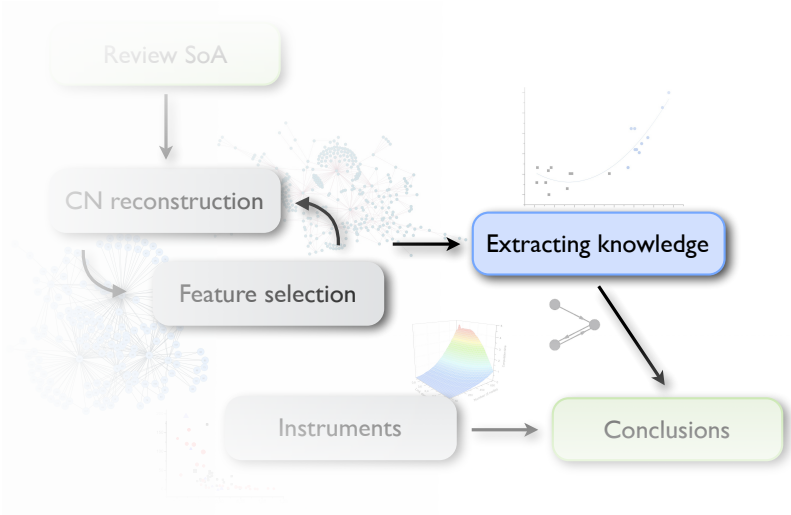
4.4 Conclusions

This Chapter proposed a methodology to compare different feature selection algorithms in terms of their effectiveness for the reconstruction of network representations from biomedical data sets. Especially in the case of mass-spectrometrics data, comprising 10.000 different measurements for each subject, a direct network representation would be unfeasible, by reason of the extremely high computational cost associated to the analysis of graphs with thousands of nodes. Furthermore, it is known that biomedical data usually contain a high quantity of redundant and noisy information that can be safely eliminated, and whose presence may even reduce the discrimination capability of a classification algorithm.

Results suggest that the application of information science-based measures, *e.g.* of *Mutual Information*, can be used to safely reduce the number of measurements, and therefore of nodes in the network representation, in up to two orders of magnitude. On the other hand, binning the spectrum, *i.e.* considering the average of sequential non-overlapping regions, while commonly performed in the Literature, yields to a destruction of relevant information. This is therefore a good example of how data mining techniques can be used to optimize a network representation, reducing the dimensionality of the problem and, at the same time, minimizing the information lost.

5

Extracting knowledge from a complex network representation



As discussed throughout this Thesis, the study of a system by means of complex networks is composed of two steps: first, represent such system by a suitable network, and second, analyze its structure of connections, *i.e.* its topology. Whether the network representation is assembled by mapping physical connections, by constructing a *functional* representation, or by using the techniques proposed in Chapters 3 and 4, the last step requires extracting a set of metrics describing some topological characteristics, and interpreting the results in the light of the problem being studied. This Chapter deals with this

last step.

The analysis of the structure of a network presents several challenges, which will here be tackled by dint of data mining techniques. An example may help shedding light on this point. When studying brain dynamics, it is common to create networks such that nodes represent individual sensors, or brain volumes, and links between pairs of them are established if a relationship, *e.g.* correlation or synchrony, is detected in the signals recorded by such sensors [ECCBA05; BS09]. Such a technique has allowed characterizing important features of functional brain activity in healthy brains [BS09; BBMLAWC09; MLB10] and in neurological and psychiatric diseases [SJNBS07; SDHD-JMWMVDMVDBS09; BBMCLGSNANDPB11].

The result of the previous process, as well as the result of the network reconstruction algorithm proposed in Chapter 3, is a weighted clique, *i.e.* a fully connected network whose information is codified in the *weight* associated to each link. The analysis of such cliques is plagued by three elements of arbitrariness. First, there is no objective criterion determining which metrics ought to be used (out of the great number of available ones) for the quantification of the relationships among nodes. Second, the transformation of the clique into a structured unweighted network generally requires a thresholding process that ultimately leads to an adjacency matrix, which crucially depends on the value of the adopted threshold. Finally, there is no criterion for establishing which feature, or set of features, of the resulting network should be looked for and taken into account to extract the best information from the original system.

In this Chapter, we propose a new methodology for addressing these three issues, based on the application of data mining techniques. By starting from an external classification as ground truth (in our specific case the association of each network to a healthy status or an illness), we propose the use of the output of a data mining classification task as a proxy for the relevance of the network representation under study. This yields criteria for an optimal network representation relative to a given problem.

This Chapter is organized as follows. In Section 5.1 the proposed methodology is described in mathematical terms, for then being validated in four different cases: analysis of MEG data (Section 5.2), comparison of different synchronization metrics in the study of brain activity (Section 5.3), analysis of neuroimage data of Alzheimer's patients (Section 5.4), and diagnosis of *leukemia* from blood spectroscopy (Section 5.5).

5.1 Optimizing the network representation

Fig. 5.1 depicts the three steps that are usually performed to analyze a set of multivariate data by means of complex networks, as for instance in the case of the analysis of brain dynamics. They include *i)* the creation of a weighted clique by means of some synchronization metrics, *ii)* the application of a threshold, in order to keep only those links that codify a strong relationship, and *iii)* the extraction of a set of topological metrics that are considered relevant for the problem being studied. Each one of these steps includes

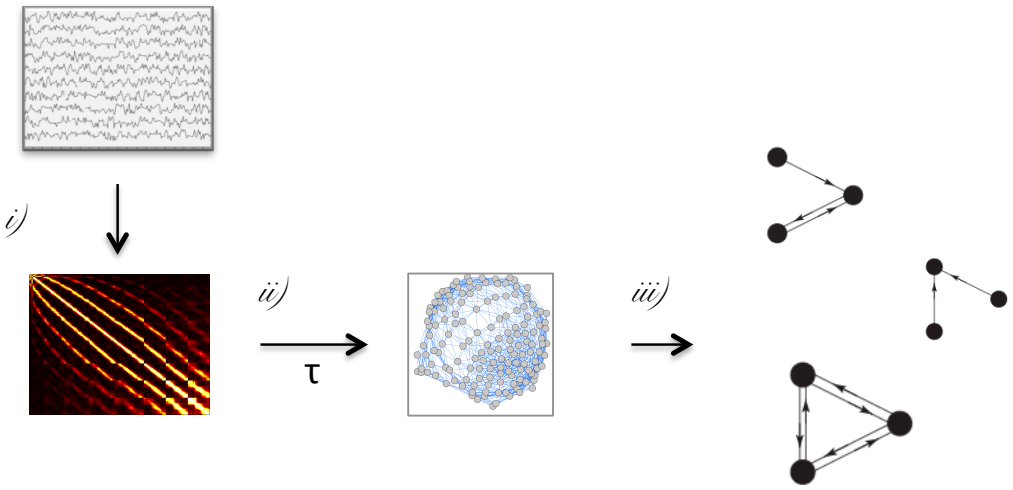


Figure 5.1: **Schematic representation of the classical network analysis steps.** Sketch of the three phases of the process, starting from a set of multivariate data (Top Left). *i)* Detection of relationships between the different time series, yielding a weighted clique; *ii)* transformation of the weighted clique into an unweighted adjacency matrix by means of a threshold τ; *iii)* extraction of a set of features from the network.

an element of arbitrariness, *i.e.* the selection of an appropriate synchronization metrics, the determination of a suitable threshold, and finally the election of the metrics to be analyzed. It has to be noticed that such problem is quite general: for instance, in the case of the technique presented in Chapter 3, the result is a weighted clique that must be processed according to steps *ii)* and *iii)*.

The methodology we here propose is based on the idea that the quality of the results obtained by a data mining task, which uses as input the features extracted from a network representation, is proportional to the quantity of information codified in the networks. In other words, if we are able to discriminate healthy people from patients using features extracted from their corresponding networks, such networks are indeed codifying relevant information for describing the disease under study. Consequently, the process can be reversed: for instance, instead of selecting *a priori* a single threshold, different thresholds can be tested, for then *a posteriori* selecting the one yielding the best results in a data mining task.

Following this approach, we propose the four steps depicted in Fig. 5.2 to transform raw data sets into functional network representations: *i)* creation of a set of weighted cliques by means of different metrics; *ii)* transformation of the cliques into sets of structured networks by applying different threshold values; *iii)* analysis of the resulting network topologies and extraction of a set of features for each of them; finally, *iv)* selection of the best metric, of the best threshold and of the most significant set of features according to the results of a data mining task. In what follows, these four steps are described in detail; without loss of generality, we consider the classification of healthy subjects *vs.* patients as the target data mining task.

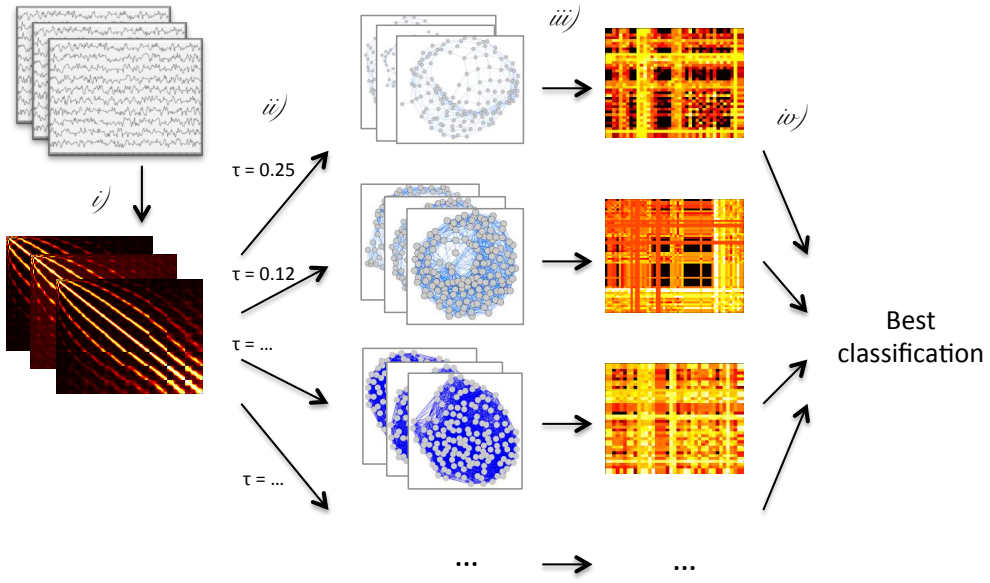


Figure 5.2: **Schematic representation of the process for the optimization of network reconstruction.** Sketch of the three phases of the identification process, starting from a weighted clique for each instance (left). *i)* Creation of weighted cliques by means of different metrics. *ii)* transformation of the weighted cliques to a set of unweighted adjacency matrices by dint of different thresholds; *iii)* extraction of a set of features for each network, and execution of a data mining task; *iv)* selection of the best metric (step *i*), threshold (step *ii*) and features (step *iii*). Reprinted with permission from Ref. [ZSPBGPPMB12].

The initial information about subjects under study should include raw data characterizing his/her constituting elements, *e.g.* n time series for each one of the N subjects under study. In the standard approach, the output of the analysis would be given by N networks, each composed of n nodes. We also assume prior knowledge of the initial labeling of each subject i to two non-overlapping classes $c_i = \{0, 1\}$. For instance, subjects may be categorized as control and MCI [1] subjects, or according to any other suitable categories.

The first step involves the creation of different weighted cliques W for each subject, by means of a set of functional metrics F that are considered relevant for the problem at hand. The result of this step is therefore $N \cdot |F|$ different cliques. Notice that, when networks are reconstructed following the approach developed in Chapter 3, this step is not necessary, as the output has already the form of a weighted clique.

Instead of applying a single pre-determined threshold τ on W (*i.e.* defining an associated adjacency matrix A with elements $a_{i,j} = 1$ whenever $w_{i,j} > \tau$, and $a_{i,j} = 0$ otherwise), we propose the use of a set of thresholds $T = \{\tau_1, \tau_2, \ldots\}$, covering the whole range of applicable thresholds. Therefore, step *ii)* yields $|T|$ structured networks for each of the $N \cdot |F|$ cliques, ranging from sparsely to densely connected graphs.

Finally, the analysis of the topological properties of the resulting networks is usually

[1] *Mild Cognitive Impairment*, a neurodegenerative disease that will be further tackled in Section 5.2.

performed by calculating and comparing a specific topological indicator, often chosen by the investigator based on his/her own experience. Instead, step *iii)* of our procedure involves the extraction of a large set of measures M from each network, including the most relevant macro-, meso- and micro-scale topological features of a complex network (see Annex A for the complete list of the features taken into account). At the end of the third step, the initial raw data are therefore converted into a large set of measures, specifically into $N \cdot |F| \cdot |T| \cdot |M|$ metrics, that represents a wide sample of the possible analyses that may be performed from a complex network perspective.

The problem is now that of identifying the optimal combination of functional metric, threshold and topological metrics that better characterize the system. This problem is tackled in step *iv)* by means of a data mining classification task. Specifically, for each functional metric F, threshold τ_i, and for each pair (or triplet) of metrics, subjects are classified; the percentage of subjects correctly classified is then used as a proxy of the relevance of such set of parameters. Indeed, if a good classification is achieved, the considered parameters and network metrics correctly represent the structural differences between the two classes of subjects. Thus, the best classification corresponds to both the best set of metrics and to the best threshold.

5.2 Validation: MEG data

To demonstrate the validity of the proposed approach, we firstly consider a set of magnetoencephalographic data (MEG), and identify the features that better differentiate healthy subjects from patients suffering from a cognitive impairment, specifically from *Mild Cognitive Impairment* (*MCI*). In this first validation case, step *i)* of the proposed method has been omitted, *i.e.* the analysis has been only focused to the detection of the best thresholds and topological metrics; the selection of the best functional metric will be the focus of Section 5.3

Principles behind magnetoencephalography

Magnetoencephalography (*MEG* in short) is a non-invasive neurophysiological technique that measures the magnetic fields generated by the activity of neurons in the brain. Neurons mostly interact by means of trains of electrical peaks, resulting from the flow of electrically charged ions through the cell. As a result, electromagnetic fields are generated; while the magnitude of individual fields is negligible, multiple neurons acting together generate a measureable magnetic field outside the head. Even during highly sinchronized activity, the generated neuromagnetic signal is extremely small, a billionth of the strength of the earths magnetic field: therefore, MEG scanners must be composed of superconducting sensors called *SQUID* (Superconducting QUantum Interference Device).

Among the advantages of MEG analyses, it has to be noticed that MEG data is a direct

measure of brain function, unlike functional measures such as fMRI, PET and SPECT that reflect brain metabolism. Also, MEG presents very high temporal and spatial resolutions, respectively measured in milliseconds and millimeters. Finally, MEG is completely non-invasive, as it does not require the injection of isotopes or exposure to X-rays or magnetic fields.

While a complete description of the MEG technology is beyond the scope of this Thesis, the interested reader may refer to the multiple review papers available in the Literature [Coh72; HHIKL93; VR01].

Initial data

The data set comprises recordings from nineteen patients and nineteen healthy volunteers during a modified Sternberg's letter-probe task. All subjects were right-handed elderly volunteers recruited from the Geriatric Unit of the Hospital Universitario San Carlos, Madrid. The nineteen patients were classified as multi-domain MCI, according to the criteria proposed in Ref. [Pet04]. Nineteen age-matched, healthy elderly volunteers, without memory complaints, recruited for a project called *Aging with Health*, consented to participate in the study. In order to avoid possible differences due to different educational records, patients and controls were chosen so that the resulting average number of years of education was similar: 10 years for patients and 11 years for controls. Before the MEG recording, all participants or legal representatives gave informed consent to participate in the study. The study was approved by the local ethics committee.

Participants were asked to memorize a set of five letters presented on a computer screen. After the presentation of the five letter set, a series of single letters (1000 ms in duration with a random ISI [2] between 2-3 s) was presented one at a time, and the participants were asked to press a button with their right hand when a member of the previous set was detected. All participants completed a training session before the actual test, which did not start until the participant demonstrated that he/she could remember the five letter set. The MEG signal was recorded with a 254 Hz sampling rate, and a band pass filter between 0.5 to 50 Hz; the recording was performed using 148-channel whole head magnetometer, confined in a magnetically shielded room (MSR). An environmental noise reduction algorithm using reference channels was applied to the data. Thereafter, single trial epochs where visually inspected by an experienced investigator and those containing visible blinks, eye movements or muscular artifacts were excluded from further analysis. Artifact-free epochs from each channel were then classified into four different categories according to the subject performance in the experiments: hits, false alarms, correct rejections and omissions, of which only hits were considered for further analysis. 35 1-second-long epochs were randomly chosen from each of participant.

A correlation matrix $C\{\omega_{ij}\}$ of size 148 x 148 was computed for each participant using the MEG time series. The correlation between each pair of sensors was calculated by

[2] Interstimulus interval, *i.e.* the temporal interval between the offset of one stimulus to the onset of another.

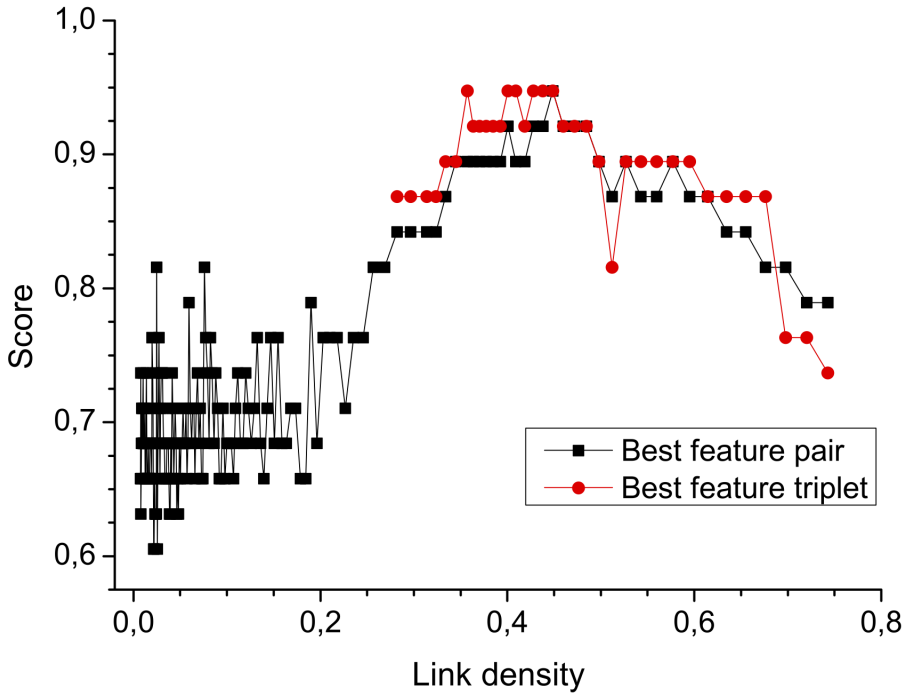


Figure 5.3: **Classification score as a function of link density.** Black (red) points indicate the best classification score obtained using pairs (triplets) of features. Best classification results appear at high link density, *i.e.* at rather dense networks, indicating that links associated to low correlations are actually codifying relevant information. Reprinted with permission from Ref. [ZSPBGPPMB12].

means of a Synchronization Likelihood (SL) algorithm, as proposed in Ref. [SD02] (see also Section 5.3 for further details).

Results

Following the methodology proposed in Section 5.1, $|T| = 178$ networks have been created for each subject, corresponding to the number of different thresholds considered. From each one of these networks, $|M| = 72$ different topological metrics have been calculated. A classification task was ultimately performed for each pair and triplet of considered features, using a Support Vector Machine algorithm. Fig. 5.3 reports the *precision* (percentage of correctly classified subjects) for the most representative pair (triplet, in red) of features, corresponding to each threshold. Specifically, at each threshold value, we consider the density of links resulting in the corresponding functional networks, and report the best score (Figure 5.3) when adopting pairs (in black) and triplets (in red) of the metrics. Classification was also attempted with other algorithms, including Naive Bayes and neural networks, producing qualitatively comparable results.

Several relevant conclusions can be derived from Figure 5.3. Firstly, the best classification rate (95%) is obtained for sufficiently low threshold values, *i.e.* including a great

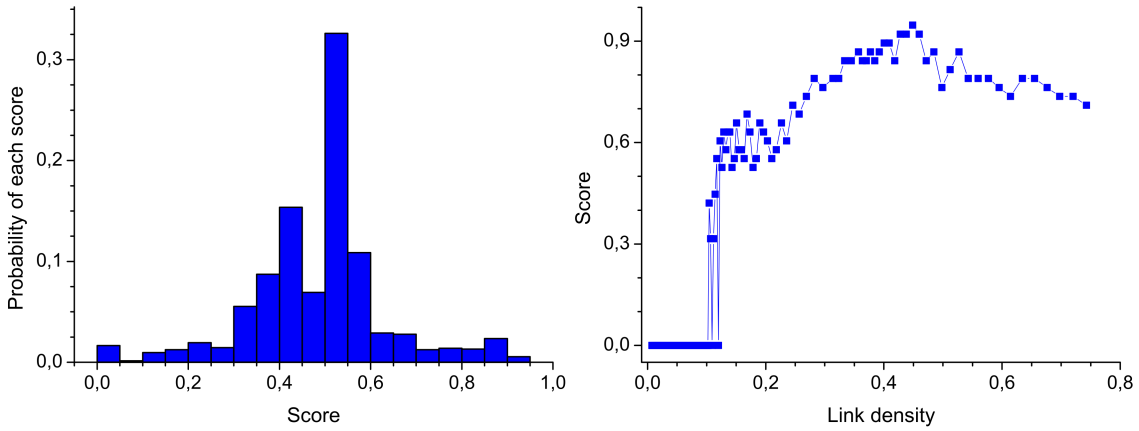


Figure 5.4: **Relevance and stability of classification results.** (Left Panel) Histogram of the score values obtained with pairs of features for the best threshold value (0.069). (Right Panel) Score obtained with the best feature triplet (*i.e.*, *small-worldness*, *Motif 1 ZScore*, and entropy of centrality distribution) at different thresholds. Reprinted with permission from Ref. [ZSPBGPPMB12].

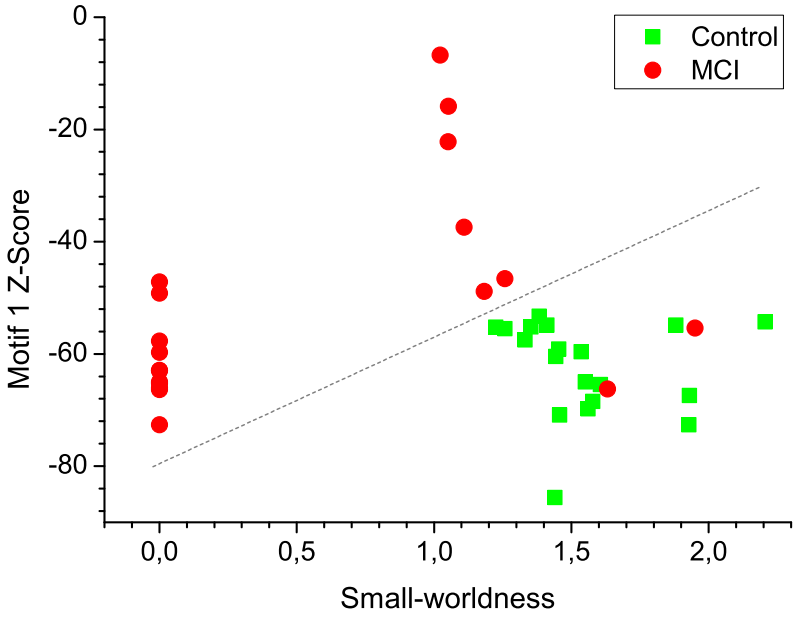


Figure 5.5: **Classification of MCI and healthy patients.** Green (red) points represent the position in the space of features of healthy (MCI) patients. The graph depicts the best classification obtained with the selected pair of features. Reprinted with permission from Ref. [ZSPBGPPMB12].

quantity of low-correlated links inside the analysis. Specifically, the maximum score corresponds to including about 40% of the links. Remarkably, the functional brain network literature typically considers networks with a 5% link density [BS09; ECCBA05]. The increase in the number of links, as suggested by the proposed methodology, has a major consequence: allowing a better consideration of meso-scale structures, *e.g.* of motifs,

whose role is much less prominent at higher threshold values. As a consequence, the best classification is always obtained when explicitly including the frequency (the Z-score) of motifs inside the set of analyzed topological indicators.

Secondly, a comparison of the scores obtained by a two-feature strategy and a three-feature approach (from Figure 5.3) reveals no relevant increase. Therefore, in this particular example, one can define a $|T| \times |M| \times |M|$ tensor of scores, where the first variable is the threshold value, and the following two are a suitable combination of two measures extracted from the topological quantities definable on the reconstructed functional network. One then may simply look for the highest tensor component.

Thirdly, results corresponding to low link densities are much more unstable, as demonstrated by the leftmost part of the plot in Figure 5.3. Clearly, the addition, or deletion, of a few links has a major effect in the topology, changing the meaning of all metrics calculated on the top of it. Therefore, our results invite to reconsider many studies made in the Literature about functional brain network reconstruction.

We now discuss the relevance and stability of the obtained results. Figure 5.4 Left shows the histogram of the score values obtained at the best threshold ($\tau_{best} = 0.069$) for all possible pairs of measures considered. The Figure clearly shows that the best score (~ 0.95) only occurs for a *very specific selection* of the pair of features (namely, the $Z-$score of Motif 1 and the *small-worldness*), whereas a generic choice of a pair of measures leads to a much worse classification performance, with scores just above random classification level. This demonstrates the ability of the method to unveil which specific topological information one has to look for in order to gather the best information on the system under study, and the best classification capability. This is confirmed by Figure 5.4 Right, where we report the value of the scores obtained when adopting the individuated best triplet (*small-worldness*, *Motif 1 ZScore*, and entropy of centrality distribution) for different values of the threshold τ. The classification ability of such a triplet is reflected by the stability of the score within a huge region of link densities around the optimal one.

Finally, Figure 5.5 helps visually understanding the power of the used classification technique. When adopting the best obtained threshold ($\tau = 0.069$), and the best corresponding pair of topological features, each patient is then represented by a point in the corresponding plane of values, with green (red) points corresponding to healthy (MCI) patients.

5.3 Validation: comparing different synchronization metrics

As previously introduced, one critical step of brain functional network reconstruction is the choice of the metric used to create links between nodes, *i.e.* the assessment of the degree of synchronization between different brain regions. As a first approximation, linear correlation was used. Yet, it was soon clear that the dynamics of the brain was not linear, and that more information could be extracted if the concept of causality was included in the analysis. Due to this, a large number of different metrics were born, some of them

imported from other fields of research, other constructed *ad-hoc* taking into account the characteristics of the problem at hand. The selection of the best synchronization metric is even more complex if one takes into account that the performance of some metrics may be pathology-dependent, due to the characteristics of the underlying dynamics.

Here we apply the strategy previously introduced with the aim of assessing the quantity of information that different metrics are able to codify. The strategy relies on creating different complex network representations of a set of subjects, presenting different neurological pathologies, using each one of the available metrics. A set of features is then extracted from each network, characterizing its topological structure, and used to train a classification model. The success in performing the classification is then used as a proxy of the relevance of each synchronization metric: the higher this score, the larger the quantity of information codified in the network, and thus the relevance of the considered metric. Furthermore, the analysis of the features used in the classification task will unveil information about the network characteristics best encoded by each metric, *e.g.* macro-scale *vs.* micro-scale features.

Description of the data set

The data set was created by processing MEG recordings of fifty-six, right handed, elderly participants recruited from the Geriatric Unit of the *Hospital Universitario San Carlos Madrid* and the *Centro de Prevención del Deterioro Cognitivo, Ayuntamiento de Madrid*. Participants were divided into three groups based on their clinical profiles: 19 participants were considered as multidomain MCI patients, 25 as elderly control participants and 12 as SMC participants.

The SMC group was composed of elderly participants (average 72.5 years old) who came, on their own initiative, to the *Centre for the Prevention of Cognitive Decline*, a public health center in Madrid (Spain) running memory training programs for both healthy elders and MCI patients, and reporting experiencing memory deficits. Participants for the SMC group were selected following the criteria proposed by Ref. [AH08]: *(1)* Patient stating that their memory function has deteriorated compared to earlier stages in life; *(2)* time of onset being in adulthood; *(3)* providing a valid example; *(4)* memory deterioration confirmed by an informant (close relative or friend); *(5)* normal objective memory performance. The assessment was based on structured interviews and neuropsychological tests. To ensure that memory complaints were not caused by a psychiatric condition, all patients were interviewed by an experienced psychiatrist and had to score below 9 in the geriatric depression scale [YBr91]. Additionally, to confirm the memory complaints, participants from this group had to score higher than 13 (mean 27.6) in the memory failures of everyday test [SHB83]. Given that the association between subjective ratings and future cognitive decline is stronger when complaints have been confirmed by an informant [FMJ05], we required confirmation from relatives or close friends. None of these

patients met the criteria for MCI and had no history of psychiatric or neurological disorders. Most SMC patients were following educational courses at local social centers.

MCI diagnosis was established according to the criteria proposed by Petersen *et al.* [Gru+04; Pet04]. Thus, MCI patients fulfilled the following criteria: *(1)* cognitive complaint corroborated by an informant, *i.e.* a person who stays with the patient at least for half a day at least 4 days a week; *(2)* objective cognitive impairment, documented by delayed recall in the logical memory II subtest of the revised Wechsler Memory Scale (score lower than 16/50 for patients with more than 15 years of education; lower than 8/50 for patients with 8 - 15 years of education) [Wec87]; *(3)* normal general cognitive function, as assessed by a clinician during a structured interview with the patient and an informant and, additionally, a mini mental state examination (MMSE) score greater than 24; *(4)* relatively preserved daily living activities as measured by the Lawton scale; *(5)* not sufficiently impaired, cognitively and functionally to meet criteria for dementia. Age and years of education were matched to the SMC group. According to their clinical and neuropsychological profile, all patients in this group were considered multi-domain MCI patients [Pet04]. As for the geriatric depression scale, none of the MCI showed depression (score lower than 9) [YBr91].

Twenty-five age-matched healthy elderly participants were included as a control group, whose age and years of education were matched to the SMC group. To confirm the absence of memory complaints, a score of 0 was required in a four question questionnaire [Mit08]. None of the participants had a history of neurological or psychiatric condition.

To summarize, MCI patients showed both subjective and objective memory impairment, SMC participants presented only with memory complaints with a normal score on the memory test and healthy elders showed neither subjective nor objective memory impairments. MCI patients, SMC subjects and healthy participants underwent a neuropsychological assessment, in order to establish their cognitive status with respect to multiple cognitive functions. Specifically, memory impairment was assessed by the logical memory test (immediate and delayed) from the Wechsler Memory Scale-III-R. Two scales of cognitive and functional status were applied as well: the Spanish version of the MMSE [LEGSS79], and the Global Deterioration Scale/Functional Assessment Staging GDS/FAST. Participants were selected so that the number of years of education was as similar as possible for the three groups (MCI patients 8.5, SMC patients 8.3 and control participants 8.9 on average). Before the MEG recording, all participants or their legal representatives gave written informed consent to participate in the study, which was approved by the local ethics committee.

MEG scans were obtained in the context of a modified version of the Sternberg's letter-probe task [DMEHMGF91], in which a set of five letters was presented and participants were asked to keep the letters in mind. After the presentation of the five-letter set, a series of single letters (500 *ms* in duration with a random ISI between 2 and 3 *s*) was introduced one at a time, and participants were asked to press a button with their right hand when a member of the previous set was detected. The list consisted of 250 letters in

which half were targets (previously presented letters) and half distracters (not previously presented letters). Participants undertook a training series before the actual test, which did not start until the participant demonstrated that he/she remembered the five-letter set.

The MEG signal was recorded with a 254 Hz sampling frequency and a band pass of $0.5 - 50$ Hz. An environmental noise reduction algorithm using reference channels was applied to the data. Thereafter, single-trial epochs were visually inspected by an experienced investigator, and epochs containing visible blinks, eye movements or muscular artifacts were excluded from further analysis. Artifact-free epochs from each channel were then classified into four different categories according to the subject's performance in the experiment: hits, false alarms, correct rejections and omissions. Only hits were considered for further analysis because we were interested in evaluating the functional connectivity patterns that support recognition success. Thirty-five epochs of 1 s each were used to calculate the seven synchronization metrics here considered. This lower bound was determined by the participant with least epochs. To have an equal number of epochs across participants, 35 epochs were randomly chosen from each of the other participants.

Synchronization metrics included in the analysis

The following synchronization metrics have been considered in this study:

Correlation (*COR*). Pearson's correlation coefficient, measuring the linear correlation in the time domain between two signals at zero lag. It is defined as

$$R = \frac{1}{N}\sum_{k=1}^{N} x(k)y(k), \tag{5.1}$$

x and y being the two signals. *COR* is defined between -1 and 1, respectively representing a complete inverse and direct correlation.

Coherence (*COH*). It measures the linear correlation between two time series at a given frequency. Denoting the cross power spectral density at a given frequency f as $P_{xy}(f)$, and the power spectral densities of the two signals as $P_{xx}(f)$ and $P_{yy}(f)$, *COH* is defined as

$$k_{xy}^2 = \frac{|\langle P_{xy}(f)\rangle|^2}{|\langle P_{xx}(f)\rangle|\,|\langle P_{yy}(f)\rangle|}. \tag{5.2}$$

Synchronization Likelihood (*SL*). *SL*, arguably one of the most popular index for assessing the presence of generalized synchronization, returns a normalized estimate of the dynamical interdependencies between two or more time series [SD02]. In a way similar to the generalized mutual information, it relies on the detection of simultaneously occurring patterns, even when they are different in the two signals.

Phase-Locking Value (*PLV*). The *PLV*, also known as *Mean Phase Coherence* [MLDEE00], estimates how relative phase differences are distributed over the unit circle [LRMV99]. When there is strong Phase Synchronization (PS) between x and y, the relative phase occupies a small portion of the circle and the *PLV* is close to 1. But if the systems are not synchronized, the relative phase spreads out all over the unit circle and the *PLV* remains low. Mathematically, the *PLV* is defined as:

$$\begin{aligned} PLV &= \left|\left\langle e^{i\Delta\varphi_{rel}(t)}\right\rangle\right| = \left|\frac{1}{N}\sum_{n=1}^{N} e^{i\Delta\varphi_{rel}(t_n)}\right| = \\ &= \sqrt{\langle\cos\Delta\varphi_{rel}(t)\rangle^2 + \langle\sin\Delta\varphi_{rel}(t)\rangle^2} \end{aligned} \tag{5.3}$$

Weighted Phase-lag Index (*wPLI*). The *wPLI* is an improved version of the Phase Lag Index, which solves some problems that may appear due to the presence of volume-conduction, noise and sample-size bias. It is defined as [VOWBP11]:

$$wPLI = \frac{|\langle\Im\{X\}\rangle|}{\langle|\Im\{X\}|\rangle} = \frac{|\langle|\Im\{X\}|\,\mathrm{sgn}\,(|\Im\{X\}|)\rangle|}{\langle|\Im\{X\}|\rangle}, \tag{5.4}$$

$\Im\{X\}$ being the imaginary part of the cross-spectrum between both signals.

Granger Causality (*GC*). A time series is called *causal* to a second one if one can improve the prediction of the evolution of the latter by incorporating information about the past dynamics of the former [Gra69]. Such relationship can be tested by means of bivariate autoregressive models (AR), which, supposing we are checking whether time series y is causal to x, reads:

$$x(n) = \sum_{k=1}^{P} a_{x|x,k}\, x(n-k) + \sum_{k=1}^{P} a_{x|y,k}\, y(n-k) + u_{xy}(n), \tag{5.5}$$

$u_{xy}(n)$ being the residuals associated to the model. The *Granger Causality* from y to x is defined as:

$$GC_{y\to x} = \ln\frac{\mathrm{var}(u_x)}{\mathrm{var}(u_{xy})}, \tag{5.6}$$

where u_x are the residuals of the AR model created using only past information of x.

Partial Directed Coherence (*PDC*). This metric provides a frequency domain version of the causality correlation assessed by *GC* [SB99; BS01]. Given a frequency f, the *PDC* represents the relative coupling strength of the interaction of a given source (signal y), with regard to some signal x, as compared to all of the y's connections to other signals.

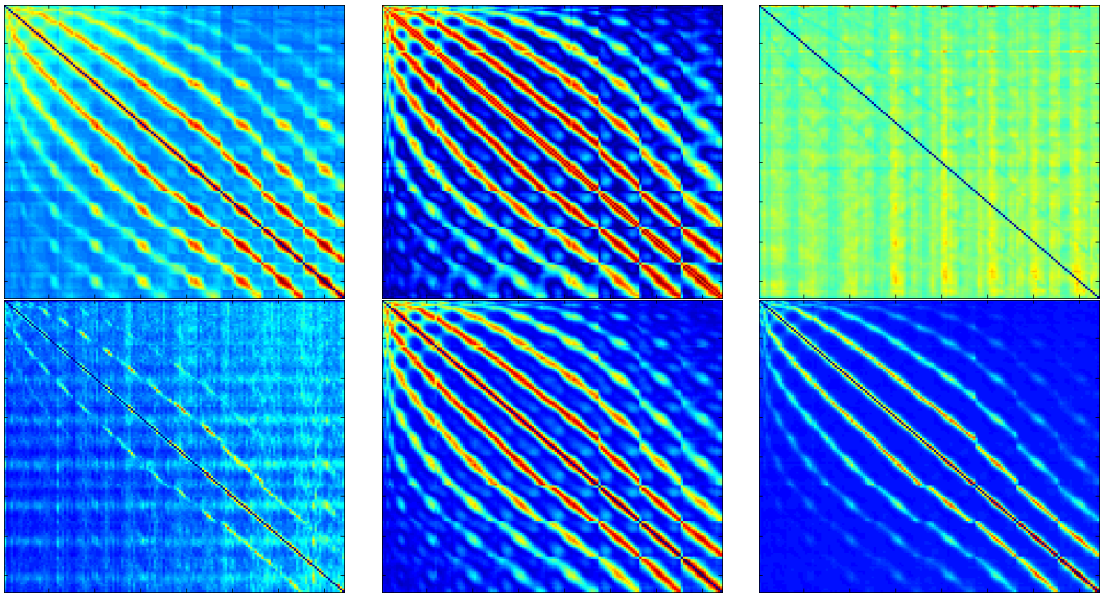


Figure 5.6: **Examples of correlation matrices for different synchronization metrics.** From left to right, top to bottom, Coherence, Linear Correlation, Granger Causality, Partial Directed Coherence, Phase-Locking Value and Synchronization Likelihood. The color of each point represents the synchronization strength, from low (blue) to high (red).

Analysis of results

Once the different synchronization metrics have been applied to the MEG data set, results were composed of 343 different weighted fully-connected networks, one for each subject-metric pair, the weight of each link representing the degree of synchronization between pairs of nodes. Fig. 5.6 reports one example, corresponding to a control subject, for all the six considered metrics. Even a simple visual inspection reveals that the results are not equivalent; notably, the two causality metrics (*i.e.* Granger Causality and Partial Directed Coherence) report lower synchronization strengths near the main diagonal, suggesting that they are less sensitive to conductivity problems [3].

In order to simplify the analysis, such weighted cliques were converted to 200 different unweighted networks by varying the applied threshold, thus ranging from disconnected to dense graphs. From each one of these unweighted networks, 47 topological features are calculated, assessing the most important micro- and macro-scale properties defined in the Literature. Six of them are represented in Fig. 5.7: maximum degree, degree-degree correlation, entropy of the degree distribution, clustering coefficient, efficiency and small-worldness - see Annex A for definitions. As in the case of Fig. 5.6, their behavior is not homogeneous, suggesting that each synchronization metric is assessing a different aspect of the brain dynamics.

[3] *Conductivity* here refers to the fact that brain tissues have a high electrical conductivity, due to their large water content. Due to this, one MEG (or EEG) sensor can receive signals from adjacent ones, thus creating spurious correlations between near regions of the brain [BRDP98; GA04].

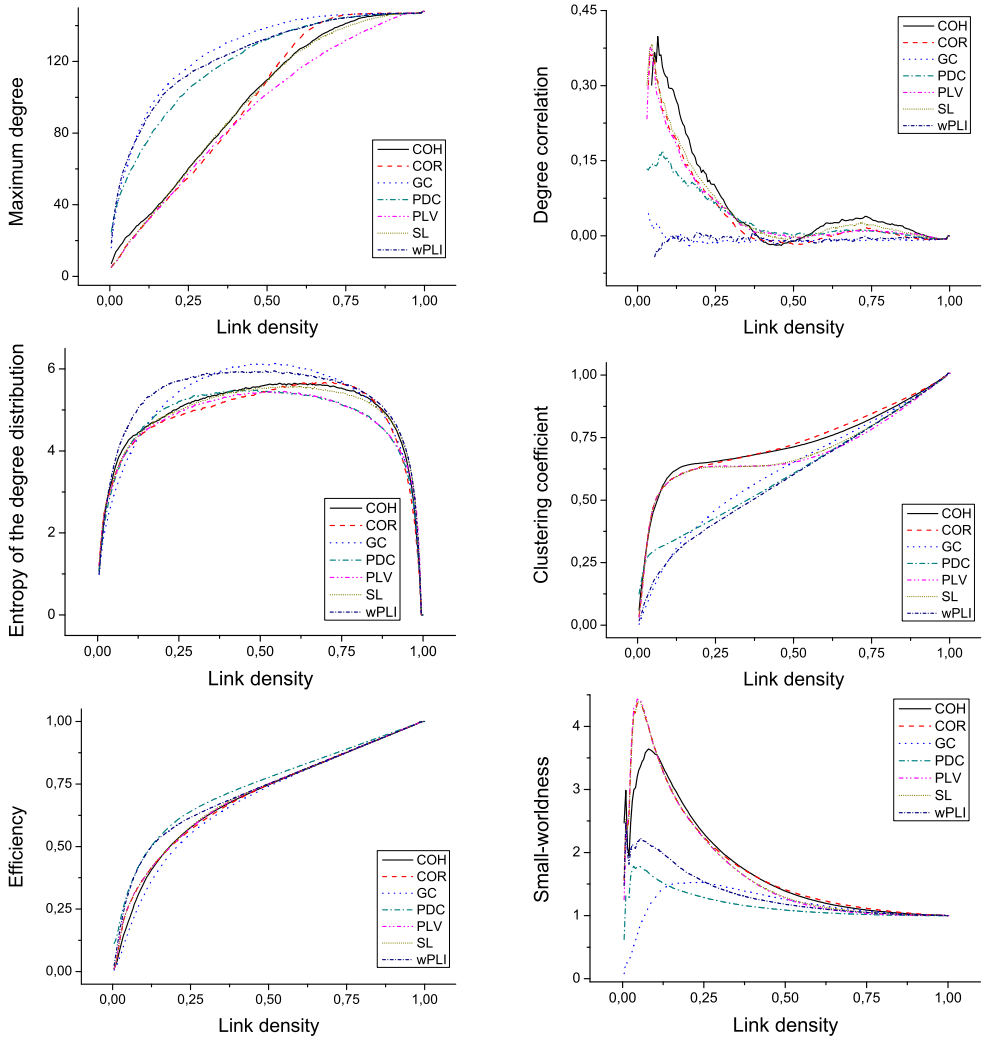


Figure 5.7: **Evolution of network topology for different synchronization metrics.** Results correspond to a single control subject.

The assessment of the significance of each synchronization metric can be further performed by means of a classification task. Using a *leave-one-out* validation strategy, a model tries to predict the category of each subject by learning from the features corresponding to all other subjects. This is performed for each synchronization metric and threshold. For a given synchronization metric, the threshold corresponding to the best classification determines the link density at which the most information is encoded in the networks. Furthermore, the synchronization metrics with higher classification scores are those that best analyze the problem at hand.

Two different cases are studied in this work. Firstly, the global classification is performed with all the three classes of subjects at the same time, with the aim of obtaining a global performance indicator for each synchronization metric. In order to validate results, four well-known classification algorithms have been considered: *Multi-Layer Perceptrons* (MLP) [Hay07], *Probabilistic Neural Networks* (PNN) [Hay07], *Decision Trees* (DT)

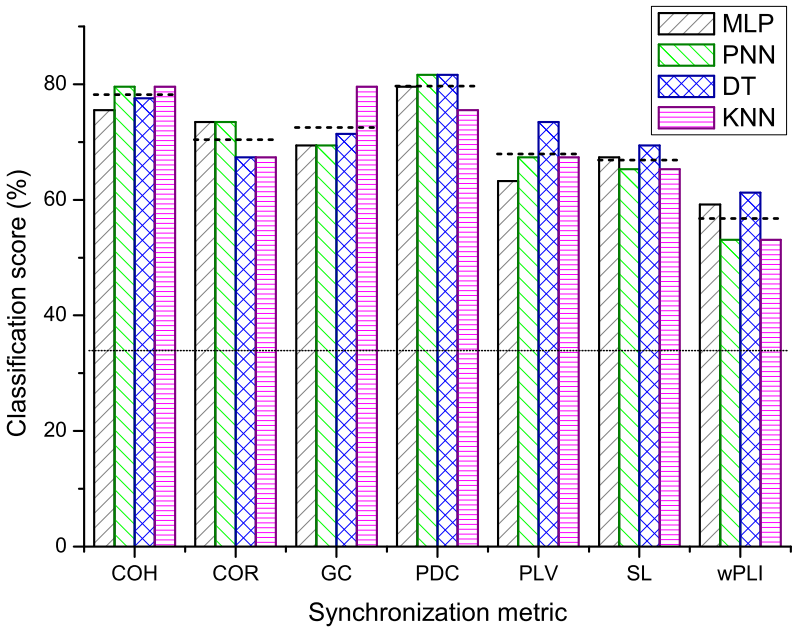


Figure 5.8: **Classification score for the seven synchronization metrics considered.** The dashed lines report the average score obtained for each metric, while the dotted line represents the score expected from a random classification of subjects (34.52%).

[Bis06; HTFJ01] and *K-Nearest Neighbors* (KNN) [Bis06; HTFJ01]. Secondly, synchronization metrics are compared for the three pairwise classification task, namely control-MCI, MCI-SMC and control-SMC. In this second case, the classification has been performed using a Support Vector Machine (SVM) with linear kernel [Ham11], taking only two network features at the time as input. This has been motivated by the reduced number of subjects in each class, and thus by the associated reduction in the predictive power - effect known as Hughes's effect [Hug68] or *curse of dimensionality*.

Fig. 5.8 reports the scores obtained for the task of classifying all three classes of subjects at the same time, by using as input network features created by the seven considered synchronization metrics. Results suggest that metrics can be grouped into three families:

- *PDC* and *COH* outperform all other metrics, with an average score of 79.59% and 78.06% respectively.
- *wPLI* yields the worst result, with a classification score of 56.63%.
- All other metrics lies in between, from a 72.44% of *GC*, down to a 66.83% of *SL*.

It is worth noticing that the difference between the best (*PDC*, 79.59%) and the worst (*wPLI*, 56.63%) metric is of the same order of the difference between the latter and the score expected in a random classification (dotted line of Fig. 5.8, 34.52%). This result highlights the great amount of information that is lost by using *wPLI* in functional networks reconstruction.

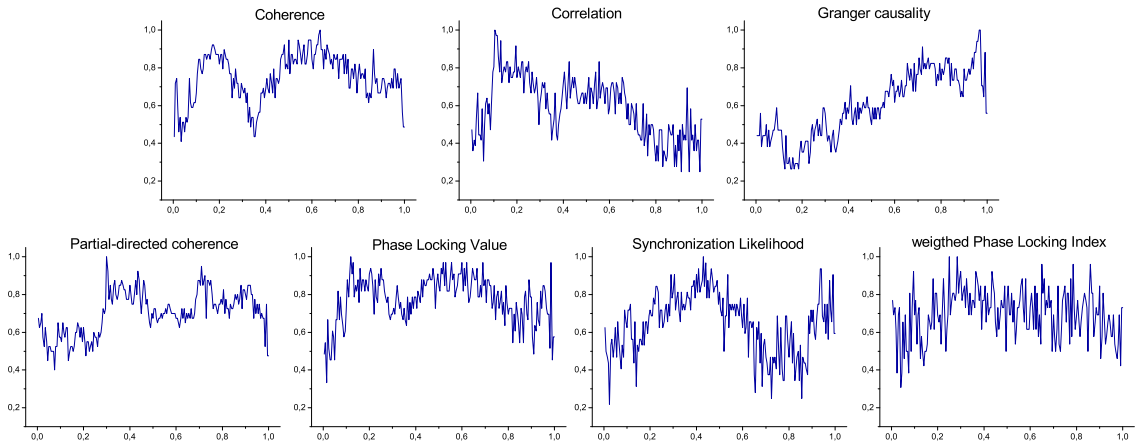


Figure 5.9: **Classification score as a function of the link density.** In order to make these graphs comparable, each one of them has been normalized, so that the maximum classification score is reported as 1.0

In order to further shed light on the differences between synchronization metrics, Fig. 5.9 reports the classification score as a function of the link density, *i.e.* the proportion of links included in the unweighted network. For the sake of clarity, scores have been normalized, so that the best classification is reported as 1.0 in all graphs. In most cases, two similar maxima can be identified, one for low (between 0 and 0.2) and one for high (between 0.4 and 0.6) link densities. Different network characteristics and scales are associated to such maxima. On one side, sparse networks best represent the global structure of the system being analyzed, with metrics like *efficiency* and *small-worldness* yielding the best scores - see Annex A for definitions. On the other side, dense networks allow the study of local structures, for instance the presence of motifs. While these maxima are usually similar, *e.g.* for *COH*, *PLV* and *wPLI*, some synchronization metrics seem to be more efficient in detecting one of these two scales. For instance, *COR* has its maximum for a link density of 0.105, with the network features most important for the classification being the *efficiency ZScore* and the *mean geodesic path length ZScore*. On the other hand, *GC* requires an almost fully-connected network (link density of 0.967), with the most important features being the ZScores of *connectivity* and *clustering coefficient*.

Fig. 5.10 reports the best score obtained by each synchronization metric, for the three possible pairwise classification tasks: control subjects *vs.* MCI, control subjects *vs.* SMC, and SMC *vs.* MCI. It should be noticed that the global performance of each metric (Fig. 5.8) is not independent on the performance in the three individual tasks. Therefore, it is not surprising that *COH* and *PDC* perform well in the three tasks, while *wPLI* obtains the worst score in two tasks out of three. Nevertheless, it is interesting to analyze the behavior of some metrics that perform unevenly in the three tasks.

The two most interesting cases are *COR* and *GC*. The former performs remarkably well in the control *vs.* SMC task, but loses ground in the SMC *vs.* MCI one, and scores in

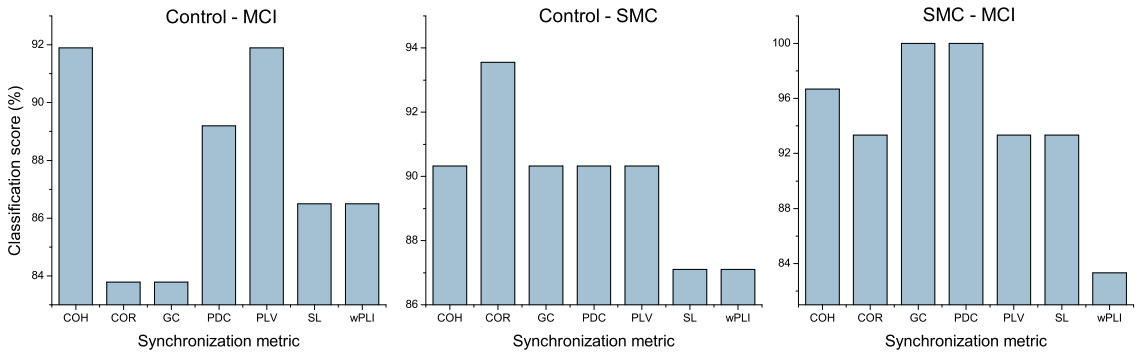


Figure 5.10: **Classification score for the three classification task considered:** control-MCI (left), control-SMC(center) and SMC-MCI (right).

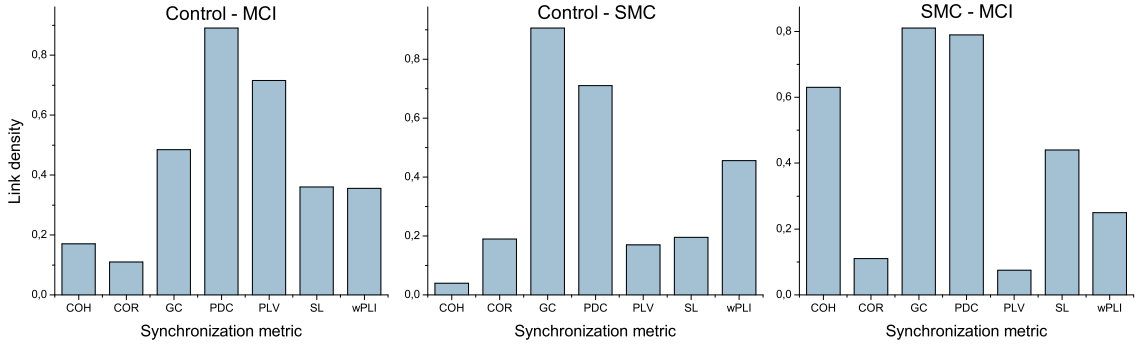


Figure 5.11: **Link density associated to the best score**, obtained by each synchronization metric in the three classification tasks considered: control-MCI (left), control-SMC(center) and SMC-MCI (right).

the lower end in the control *vs.* MCI task. On the other hand, *GC* also scores low in the control *vs.* MCI task, but obtains its best result in the SMC *vs.* MCI.

Results here reported confirm that metrics are not equivalent, and that their election should be considered as an important step in any brain activity analysis; such election can be performed by means of a data mining approach, where the usefulness of each metric is assessed through a classification task. The reason behind the uneven behavior of some of them, *i.e. COR* and *GC*, in different classification tasks is still an open problem, which will be tackled in future works.

5.4 Validation: analysis of neuroimage data

Introduction

Neuroimaging is a set of techniques aimed at creating images of the brain for clinical purposes, *i.e.* for diagnosis and diseases characterization, without the need of shots or

surgery. The most important of these techniques is probably the *Magnetic Resonance Imaging* (MRI in short). A MRI scanner is based on the interaction of two magnetic fields. A first one is used to align the magnetization of atomic nuclei in the part of the body under analysis, in this case the brain; a second one, tuned at radio frequencies, systematically alters the alignment of this magnetization. As a result, nuclei produce a rotating magnetic field, which is detected and analyzed by the scanner, and which allows the identification of different soft tissues. Results are usually presented in 2D images or 3D volumes [PLRZ90; HSM04; Atl09].

In this validation case, we propose the use of the technique presented in Chapter 3 to create network representations of neuroimage data, for then analyzing their characteristics according to the approach described in this Chapter.

Description of the data set

The data set under study comprises the volume of different regions of the brain of 44 control subjects, 30 amnesic MCI patients, 29 multi-domain MCI patients, 13 people suffering from mild Alzheimer's disease, and 48 fully developed Alzheimer's. Regions available include different parts of the sub-cortical brain, *e.g. Thalamus*, *Caudate*, *Hippocampus*, *et caetera*. The outstanding feature of this data set is that it comprises homogeneous information for different diseases that are connected between them, *e.g.* MCI and Alzheimer's, the former being usually considered a prodromal stage of the latter; it thus allows searching for trends in the evolution of the diseases. Consequently, the objective of the analysis here performed was twofold: *i)* classify subjects in the correct category, and *ii)* look for some indicators (or biomarkers) describing the progress of the disease.

Analysis of results

As for the first objective, *i.e.* classify subjects according to their disease, network representations of available data were created by using the technique described in Section 3.1 for each possible pair of conditions, *e.g.* control *vs.* amnesic MCI, control *vs.* multi-domain MCI, and so forth. The best classification scores are presented in Table 5.1, which are consistent with results available in the Literature.

In order to identify indicators representing the progress of the disease, network representations were created, by using control subjects and Alzheimer's groups as labeled data, such that each network represents the closeness of a subject to one of these two conditions. Afterwards, the link density of each network has been used as a biomarker, thus representing the degree of development of the disease. Histograms of the different stages of the disease are presented in Fig. 5.12.

Results are consistent with what should be expected: control subjects are mostly in the left part of the graph (green bars), while final stages of the disease, *i.e.* multi-domain MCI and Alzheimer's, are located in the right part. Yet, there are some interesting cases that need further analysis. Specifically, many amnesic MCI patients are found in the left part,

Table 5.1: **Neuroimage data best classification scores.**

	Control	Amnesic MCI	Multi-domain MCI	Mild Alzheimer's	Alzheimer's
Control	—	78.38%	86.30%	94.74%	92.31%
Amnesic MCI	78.38%	—	79.66%	90.70%	83.21%
Multi-domain MCI	86.30%	79.66%	—	83.33%	71.05%
Mild Alzheimer's	94.74%	90.70%	83.33%	—	83.33%
Alzheimer's	92.31%	83.21%	71.05%	83.33%	—

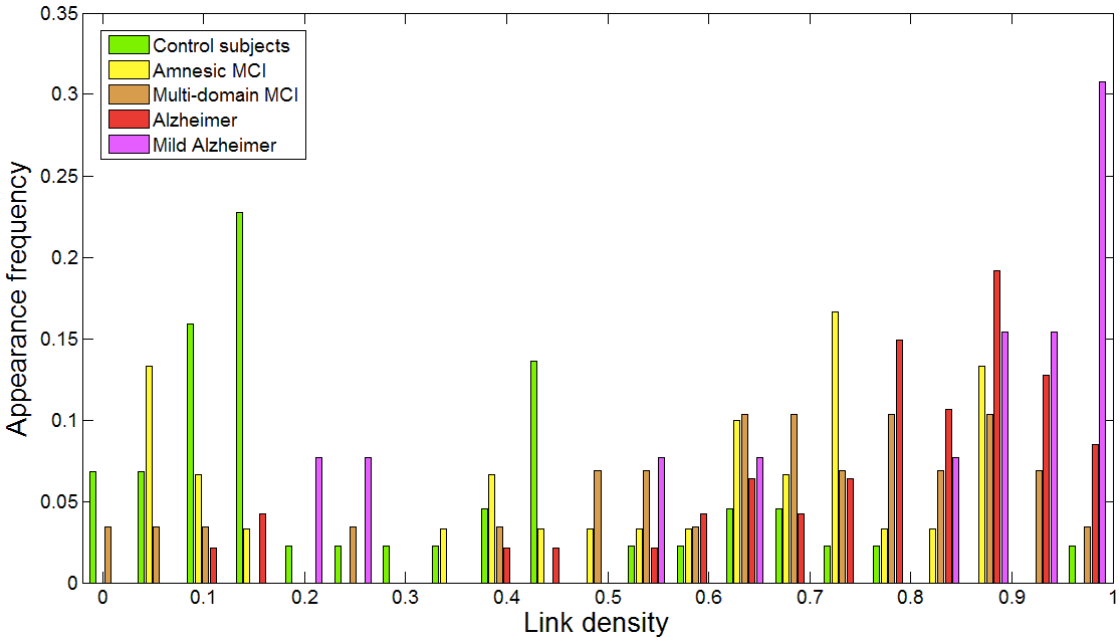


Figure 5.12: **Evolution of Alzheimer's disease.** Each color represents the histogram of link density for networks corresponding to people presenting different stages of the disease, from control subjects (green) to Alzheimer's (red).

suggesting that their symptoms are not related to Alzheimer's; furthermore, many mild Alzheimer's patients are located in the far right part, indicating stronger symptoms than people classified as Alzheimer's. Both results can be explained by a wrong diagnosis of the conditions of these subjects; in other words, it is possible that their memory problems are not caused by this specific disease. While it is known that 20% of people which have been diagnosed with Alzheimer's are indeed developing other pathologies, there is no known biomarker able to detect such situations, except for a post-mortem analysis of the brain tissues.

While further studies are needed to validate these results, and specifically a screening of patients through time to determine their real health condition, it seems clear that a complex network analysis of the brain may help identifying wrong diagnostic, as well as

characterizing this group of intermingled diseases.

5.5 Validation: diagnosis of *leukemia* from blood spectroscopy

Introduction

In this last validation case, we present a study aimed at the diagnosis of *leukemia* by means of complex network representations of blood analysis data.

The data set here considered has been constructed by collecting Raman spectra [LE01; GSMESRPA14] from 133 blood samples, 102 of them corresponding to control subjects, and 31 to people suffering from *leukemia*. A Raman spectrum is a fingerprint of biological sample, its bands providing information about the conformation of macromolecules, such as proteins, nucleic acids, or lipids. Control samples correspond to people who presented themselves as potentially blood donors at the *Hospital Regional de Alta Especialidad del Bajío* (HRAEB), Guanajuato, Mexico; data include both male and female with a mean age of 25, who had no alcohol nor drugs in the last 72 hours. The second group comprises people suffering from three different types of leukemia, *Acute Lymphoblastic Leukemia* (ALL), *Acute Myeloid Leukemia* (AML) and *Chronic Myelogenous Leukemia* (CML), with ages spanning from 5 to 80 years. In all cases, blood samples were centrifuged in order to separate the plasma and then analyzed using a *Horiba Jobin Yvon LabRAM HR800* micro-Raman System. Samples were radiated with an 830 nm laser diode of 17 mW power, and the resulting spectra were recorded with an 800 mm focal length spectrograph. Written consent was obtained from the subjects and the study was conducted according to the Declaration of Helsinki.

Analysis of results

As a first step, both sets of data, corresponding to control subjects and patients, have been used to train the model, *i.e.* to calculate the models as in Figure 3.1. The same set of subjects have then been classified using a Support Vector Machine algorithm and *leave-one-out* validation technique, and the feature selection method proposed in Section 5.1. In order to reduce the computational cost, the original measurements have been binned by using different widths. The two most relevant network characteristics for this task, corresponding to a bin size of 10 measurements, are shown in Figure 5.13 (left); the clear separation between the two groups graphically confirms the classification score of 100%. Additionally, Figure 5.13 (right) presents the evolution of the classification score as a function of the bin size; results indicate that the classification is robust, and that the task score lowers only when the number of nodes in the network is drastically reduced.

For the second phase, we got access to the Raman spectra corresponding to an additional patient, who was diagnosed with leukemia and underwent chemotherapy treatment. Notably, several measurements were available, for the day before the start of the therapy, and for each treatment days (*i.e.* after the chemotherapy sessions), grouped in

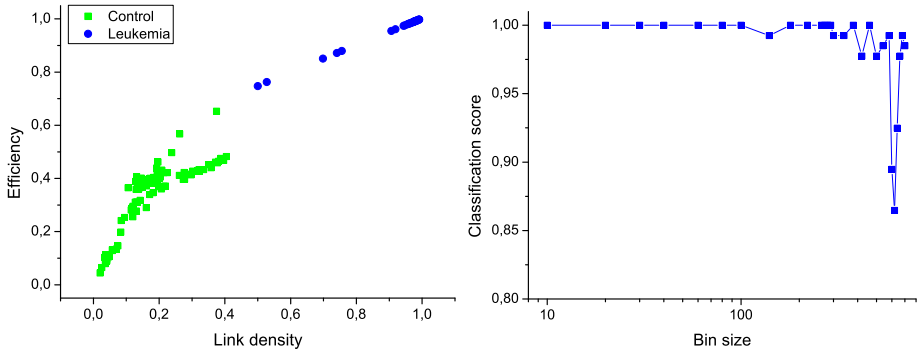


Figure 5.13: **Classification of control and leukemia subjects.** (Left) Representation of the position of control subjects (green squares) and leukemia patients (blue points) in the space of network features. (Right) Classification score as a function of the binning size. Reprinted with permission from Ref. [ZPSEFRASEJRBMS13].

three sessions. By analyzing the Raman spectra after each chemotherapy session, this time-dependent information allows us tracking the evolution through time of the structure of the network. The results corresponding to *link density* are reported in Figure 5.14.

Chemotherapy seems to globally decrease the link density of the network, therefore representing an improvement in the status of the patient. Nevertheless, the network corresponding to September 11^{th}, which has been measured after a long pause of two weeks in the treatment, suggests a return to the initial status. While no other biomedical information was available for this subject, and thus no definitive conclusions can be drawn from this example, Figure 5.14 suggests that the proposed network representation of spectral data may be used for tracking patient status, providing a new tool for decision making processes in different treatments, as suggested in Ref.[GSMESRPA14]. For instance, doctors may have decided to anticipate the start of the second session to avoid relapses.

5.6 Conclusions

While a network representation may be useful to characterize data corresponding to different biomedical problems, *e.g.* different diseases or the evolution through time of a single illness, it is not always clear what is the best way of constructing such representation, or which are the best metrics that ought to be used. In this Chapter we have proposed the use of data mining techniques to tackle such problems. Specifically, the precision of the output of a data mining task can be used as a proxy of the quantity of information codified in the initial networks, thus providing an objective metric for guiding in their construction.

The proposed approach has been validated by means of four different cases, covering the most relevant topical issues in biomedicine. Results of Sections 5.2 and 5.3 are especially relevant from the neuroscience point of view. Specifically, they allow to resolve some of the most important questions in functional network reconstruction, as which

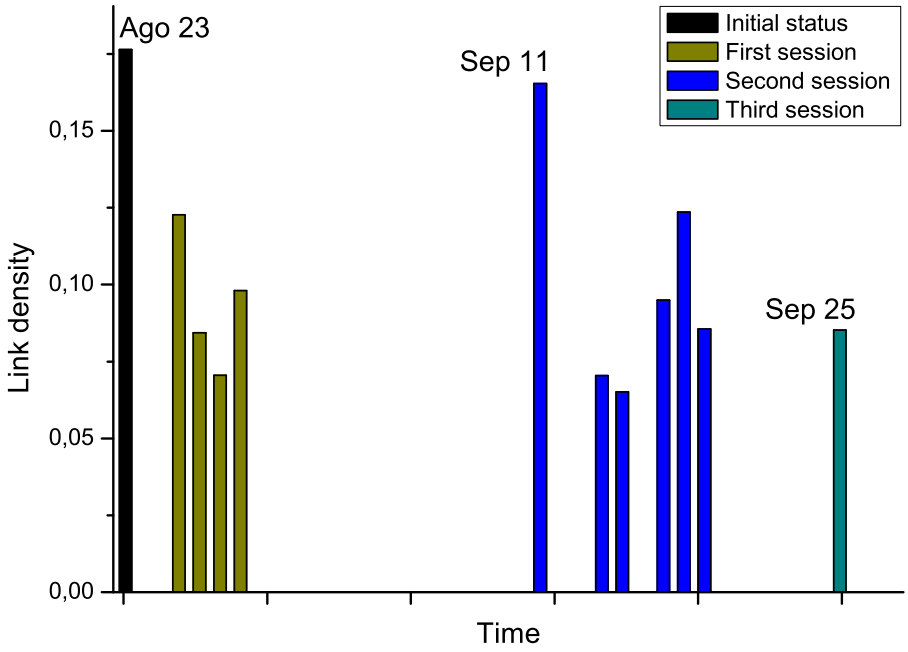


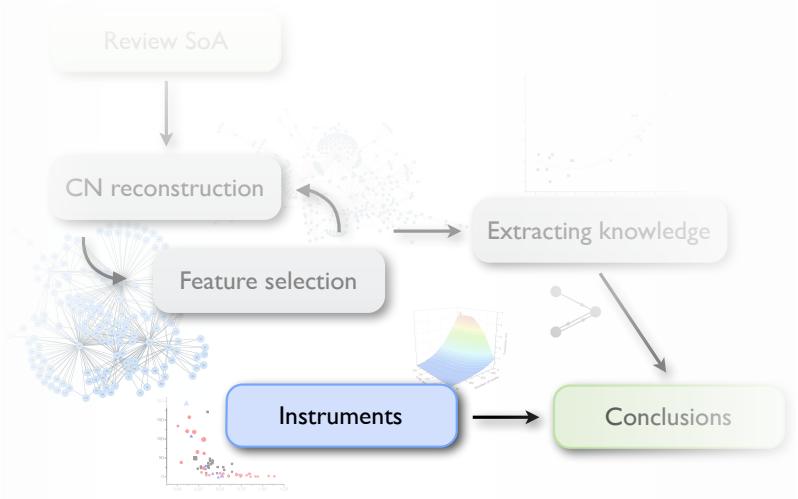
Figure 5.14: **Evolution through time of the link density of networks representing a leukemia patient under chemotherapy treatment.** Reprinted with permission from Ref. [ZPSEFRASEJRBMS13].

are the best link densities and synchronization metrics that ought to be used. It is worth noticing that it was believed that such questions were of a general nature; on the contrary, here we show how the answer strongly depends on the problem (*i.e.* on the pathology) being studied, and that both problems are strongly intermingled. Furthermore, and for the first time, we have shed light on the importance of motifs in functional brain networks, structure that have been largely disregarded in the Literature.

Additionally, Sections 5.4 and 5.5 illustrated how networks can be used to study systems that have no explicit network structure. While the constituting elements of fMRI images and Raman spectra are expected to be interconnected, the lack of prior knowledge about such connections precluded the safe use of a complex network approach. This problem is here once again solved by using the score of a data mining task as a proxy of the relevance of the reconstructed networks, allowing an estimation of the quantity of information codified in them. Furthermore, solutions based on complex networks are shown to be more efficient than standard data mining algorithm, potentially improving our diagnostic capabilities.

6

Novel instruments for complex networks analysis



Throughout the development of this PhD Thesis, the characteristics of the network analyses performed have required the development of new tools, with two main objectives: reducing the computational cost, and providing topological metrics suitable for data mining tasks. In this Chapter, we review two of the main outcomes: (*i*) a new algorithm that allows computing motifs [MSOIKCA02] in dense network, several orders of magnitude faster than its competitors; and (*ii*) a new topological metric able to detect the presence of meso-scale structures [ACLBSN11].

6.1 Fast enumeration of 3-nodes motifs

6.1.1 The need of a new motif enumeration program

Among the vast set of topological features that have been used to characterize graphs and networks, *motifs* are one of the simplest and yet of the most powerful. As explained in Annex A, motifs are defined as specific patterns of interconnections created by a small number of connected nodes [MSOIKCA02]. The importance of motifs resides in that each one of them codifies a specific function inside the network, thus creating a bridge between structure and function. For instance, in the transcriptional regulation network of *E. Coli*, it has been found that feedforward loops act as a circuit that rejects transient activation signals from a transcription factor [SOMMA02; MA03]. Motifs analysis has also been successfully applied to the study of human brain: specifically, it has been shown that only a small number of structural motifs are expressed between neurons, in such a way to minimize the biological cost [SK04]. For other examples of the use of motifs in biomedical applications, the interested reader may refer to Refs. [WOB03; BO04; HKBS07]

The analysis of the presence of motifs in networks can be performed by means of several free software programs, which can be classified in two families. On one side, FANMOD [WR06], MAVisto [SS05], GraphCrunch [KSHP11] and Pajek [BM02] are characterized by user-friendly *Graphical User Interfaces* that favor an exploratory analysis of networks. On the other side, MFINDER [KIMA02] and RAGE [MS12] are command-line tools, having the advantage of being easily integrable in other analysis tools, as the software can be called from any custom program - See Table 6.1 for a comparison of different motif-related programs functionalities.

Table 6.1: **Resume of motif detection software functionalities.**

Software name	**Command-line / GUI**	**Subgraphs counting or motifs detection**
FMotifs	Command-line	Subgraphs
MFINDER	Command-line	Both
RAGE	Command-line	Subgraphs
FANMOD	GUI	Both
GraphCrunch	GUI	Both
Pajek	GUI	Subgraphs
Software name	**Colored network**	**Computation time** 150 **nodes,** 11000 **links**
FMotifs	No	0.10 seconds
MFINDER	No	10.6 seconds
RAGE	No	n.a.
FANMOD	Yes	1.41 seconds
GraphCrunch	No	n.a.
Pajek	Yes	n.a.

In spite of their different focus, all previously mentioned programs are based on similar algorithms. They exploit chains of connected links, which makes the computational complexity scale polynomially with the mean degree of nodes, thus making these software unfit for the exhaustive enumeration of motifs in dense networks. As most of the analyses performed in this Thesis are based on the characterization of dense graphs, a new tool was needed to speed up calculations.

Here we present a new algorithm specifically designed for the fast enumeration of 3-nodes motifs in medium-size dense networks. It is based on the extraction of the adjacency matrices of subnetworks created by all triplets of nodes of the original network. The resulting 3×3 matrices are codified as discrete numbers of 9 bits, representing the connectivity pattern among the triplet of nodes. These numbers are then used to increment values in an array, which therefore represents the occurrence frequency of each pattern. Motifs are finally extracted from this array, by taking into account all possible permutations of nodes.

The result is a software, called *FMotifs*, outperforming other programs for the types of networks analyzed in this study.

6.1.2 Description of the algorithm

A complex network can be mathematically represented as a graph $G = (\mathcal{N}, \mathcal{L})$; $\mathcal{N}$ and $\mathcal{L}$ are two sets that respectively list the *nodes* (or *vertices*) $\{n_1, n_2, \ldots, n_N\}$ and the *links* (or *edges*) $\{l_1, l_2, \ldots, l_K\}$ composing the graph. The number of elements in $\mathcal{N}$ and $\mathcal{L}$ are denoted by N and K, thus representing the number of nodes and links. In what follows, the matricial representation will be considered, according to which a graph G is codified by an *adjacency* (or *connectivity*) matrix $\mathcal{A}$, of size $N \times N$, whose element a_{ij} is equal to 1 when a link connecting nodes i and j exists, and zero otherwise.

The focus of this work is the detection of motifs in a graph, *i.e.* of patterns of interconnections created by triplets of nodes. As depicted in Fig. A.3, each one of the 13 possible motifs that can occur between groups of three nodes can be interpreted as a 3-nodes graph being a subgraph of G. It follows that motifs may be represented by a set of 3×3 adjacency matrices, accounting for all possible permutations of nodes within the subgraph. For instance, motif 1 can be represented by the following adjacency matrices:

$$A_{1(1)} = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad A_{1(2)} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \quad A_{1(3)} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} \tag{6.1}$$

The proposed algorithm dwells on the codification of each motif, and therefore of each one of the corresponding subgraph adjacency matrices, into a number. The following

Table 6.2: **Association of motifs to numbers.** Motifs are labeled according to Fig. A.3.

Motifs	**Associated numbers**					
Motif 1	192	40	6			
Motif 2	96	132	136	34	66	12
Motif 3	224	196	168	42	70	14
Motif 4	72	130	36			
Motif 5	200	194	104	44	134	38
Motif 6	46	198	232			
Motif 7	164	162	100	76	138	74
Motif 8	228	170	78			
Motif 9	140	98				
Motif 10	204	172	78	102	106	226
Motif 11	202	108	166			
Motif 12	236	230	174	234	206	110
Motif 13	238					

convention about the numbering of the elements of the matrix is adopted:

$$\begin{bmatrix} 8 & 7 & 6 \\ 5 & 4 & 3 \\ 2 & 1 & 0 \end{bmatrix}. \tag{6.2}$$

Following Eq. 6.2, each one of the 3×3 adjacency matrices representing a given motif can be associated to a number as follows:

$$I = \sum_{i=0}^{8} a_i 2^i, \tag{6.3}$$

a_i being the i-th element of the matrix, as in Eq. 6.2. Notice that this is equivalent to unroll the matrix, and consider the corresponding 9-bits number in a binary notation. As an example, the three adjacency matrices of motif 1 (see Eq. 6.1) are respectively associated to 192, 40 and 6. Table 6.2 reports all the numbers associated to each one of the 13 possible motifs.

Once this convention has been established, the identification of motifs becomes trivial. For each triplet of nodes $\{n_i, n_j, n_k\}$, the subgraph created by them is extracted and stored in an integer variable p_{ijk}, such that p_{ijk} encondes the specific motif created by nodes i, j and k. Each bit of p_{ijk} is set according to the corresponding element of the adjacency matrix, following the numbering convention of Eq. 6.2. Hence,

$$p_{ijk} = a_{ij}2^7 + a_{ik}2^6 + \ldots + a_{ki}2^2 + a_{kj}2^1. \tag{6.4}$$

Notice that no self-loops are allowed, and thus bits 8 (corresponding to a_{ii}), 4 (a_{jj}) and 0 (A_{kk}) of the variable are always set to zero.

Motif frequency can be determined by means of a vector m of size 2^9. For each triplet of nodes $\{n_i, n_j, n_k\}$, the corresponding motif is annotated into m, by incrementing by

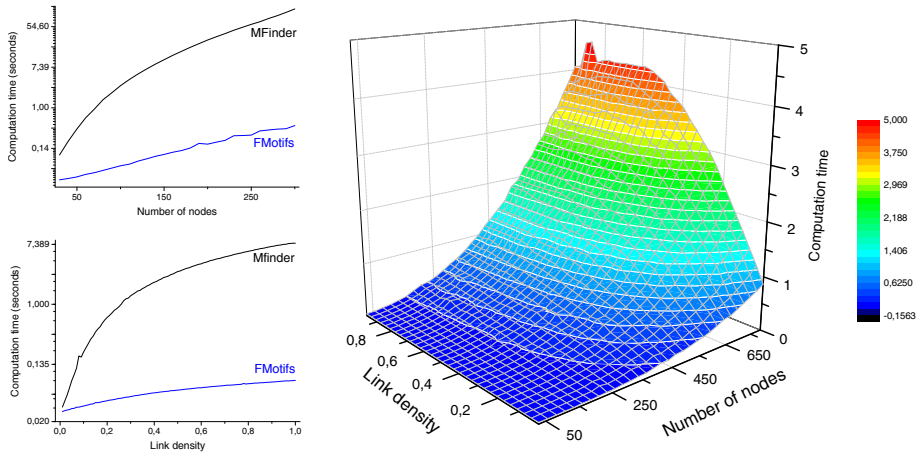


Figure 6.1: **Comparison of computation time between FMotifs and MFINDER.** (Left top) Computation time of both algorithms for random Erdös-Rényi networks [ER60] with a link density of 0.5 as a function of the number of nodes. (Left bottom) Comparison of both algorithms for random networks of 100 nodes as a function of the link density. (Right) Plot of the computation time of FMotifs as a function of the number of nodes and link density. All results have been obtained in a Intel® Xeon® E5335 CPU at 2.00 GHz.

one the value of the element $m_{p_{ijk}}$. In order to take into account all possible permutations of nodes, the values included in Table 6.2 should be used. For instance, the total number of triplets connected according to motif 1 will be $m_{192}+m_{40}+m_6$, to motif 2 $m_{96}+m_{132}+m_{136}+m_{34}+m_{66}+m_{12}$, and so forth.

6.1.3 Computational cost comparison

While this algorithm has a complexity of $O(n^3)$, n being the number of nodes in the network, each iteration is extremely fast, as it is based on the manipulation of bits in integer variables. Figure 6.1 (Left) reports the comparison of velocities of FMotifs and MFINDER, as a function of the number of nodes and of the link density of analyzed networks. As an example, for a network of 150 nodes and link density of 0.5, the computation time is reduced from 10.6 to 0.10 seconds. In general, improvements of two order of magnitude in the calculation time can be achieved for high link densities. Networks of 700 nodes can be analyzed with FMotifs in less than 5 seconds, independently on the link density, while with MFINDER this would require several minutes.

6.1.4 Availability

The algorithm has been programmed in C++. It is available as a C++ class, for its integration in any other custom development, and as a command-line tool under Windows® environment. In order to foster cross-compatibility, the tool expects the input network in a file whose format is the same as MFINDER, *i.e.* with each line of the file codifying a link as a start and destination node, plus a disregarded weight. Furthermore, the input

network can be described directly as an adjacency matrix, thus simplifying the integration with programs like MATLAB®. The output is yielded in a file as a 2×13 matrix, one row for each one of the 13 possible 3-nodes motifs. The program is also capable of taking advantage of multi-core processors, reducing the computation time when such hardware is available.

FMotifs, along with its source code and documentation, is freely available at http://www.mzanin.com/FMotifs.

6.2 Network Information Content

During the first years of complex network theory, researchers focused their attention towards two global, *macro-scale* network structures, *i.e. small-world* and *scale-free* topologies, which are ubiquitous among many real-world systems. But it was soon found that complex networks typically possess non-trivial patterns of connectivity at a *meso-scale* level, *i.e.* in between micro and macroscopical scales [ACLBSN11], which have been shown to have an important impact on, for instance, spreading [GZ06; WL08] and synchronization [ADGPV06; ADGKMZ08] processes.

Among the different types of meso-scale structures that have been described, one has received most of the attention: *communities*, that is, the organization of nodes in clusters, with many links connecting nodes belonging to the same cluster and comparatively few joining nodes of different clusters [NG04; LF09; For10]. In spite of their importance, the attentive reader would have noticed that communities have not been included in the network analysis strategy proposed throughout this Thesis. The reason behind this decision is to be found in the complexity of measuring the presence of such feature.

The pervasiveness of a community structure can, in principle, be characterized by quantifying the network *modularity* [New06]. Yet, this metric suffers from two main drawbacks: first of all, it is *a posteriori* metric, in that it can only be calculated after a community structure has been defined. Furthermore, modularity is not robust to the presence of different topological scales, *e.g.* when one community is much smaller than the others [DDGA06; FB07]. While the concept of modularity can be generalized to include other meso-scale structures, as for instance bipartite networks [GSPA07], it still inherits the previously discussed drawbacks. The same problem can be found when analyzing other types of meso-scale structures: for instance, this is the case of *core-periphery* topologies, composed of a densely connected inner core and a set of peripheral nodes sparsely connected with the core [Hol05].

In this Section we address the following question: is it possible to define a single metric able to detect the presence of different kinds of meso-scale structures? We propose a novel metric, called *Information Content*, which is simultaneously *(i)* capable of detecting generic *regularities* in the adjacency matrix of a network, *(ii) a priori* metric, *i.e.* not requiring any previous computation like community detection, and *(iii)* robust to different topological scales.

The guiding hypothesis here is that important meso-scale structures are associated with regularities in the corresponding adjacency matrix. For instance, in the simplest case of a network with a perfect modular structure, nodes connect to all peers belonging to the same community: the resulting adjacency matrix is composed of four blocks, two containing only ones, two only zeros (see Eq. 6.9 below). In this case, erasing nodes within one community causes no loss of information, as their connections are equivalent; thus, measuring the information lost when pairs of nodes are merged can be used as a way of detecting such kind of regularities - and hence meso-scale structures.

Given an initial network, the proposed algorithm identifies the pair of nodes whose merging would suppose the smallest information loss, a quantity which is a function of the number of common links to / from other nodes shared by the pair. Once the best pair has been detected, both nodes are merged (thus yielding a network one node smaller), and the quantity of information I lost in the process is calculated. When this process is iteratively repeated, the *Information Content* IC of the network is defined as the sum of all Is, *i.e.* of all information contained in the network. The lower IC, the more regular is the link arrangement, indicating the presence of meso-scale structures.

As such, the calculation of the *Information Content* can be seen as a type of network renormalization procedure [RRBF08; RSM10], characterized by two specific features. First of all, the objective is the estimation of the quantity of information lost in the process, while classical renormalization focuses on how some properties of the system are conserved at different scales. Furthermore, the renormalization transformation is guided by information theory criteria, instead of geometrical (topological) rules.

6.2.1 Information Content calculation

The calculation of the *Information Content* starts with a network of n nodes fully defined by its adjacency matrix $\mathcal{A}$, whose elements a_{ij} are equal to one when a link exists between nodes i and j, and zero otherwise. The amount of information that would be lost if two nodes were merged together is first estimated for each pair of nodes k, l (with $k \neq l$). This is performed by comparing the connections departing from and arriving at both nodes, *i.e.* the vectors $a_{k\cdot}$, $a_{\cdot k}$, $a_{l\cdot}$ and $a_{\cdot l}$, and by creating a new vector $\boldsymbol{m}$ of size $2n$, representing the links that should be modified to recover the connections of node l given the connections of node k, and thus the information lost when both nodes are merged together. In the first half of $\boldsymbol{m}$, the i-th element (with $i \in [1, n]$) is defined as one if $a_{ki} \neq a_{li}$, and zero otherwise, thus accounting for different outgoing links; the second half of $\boldsymbol{m}$ accounts for different incoming links: thus m_{i+n} (again with $i \in [1, n]$) is set to one when $a_{ik} \neq a_{il}$, and zero otherwise. In the two extreme situations, when two nodes either share all links or none, $\boldsymbol{m}$ will either take all values 0 or 1 respectively.

Once the vector $\boldsymbol{m}$ is constructed, the probability of finding an element equal to one (zero) is given by

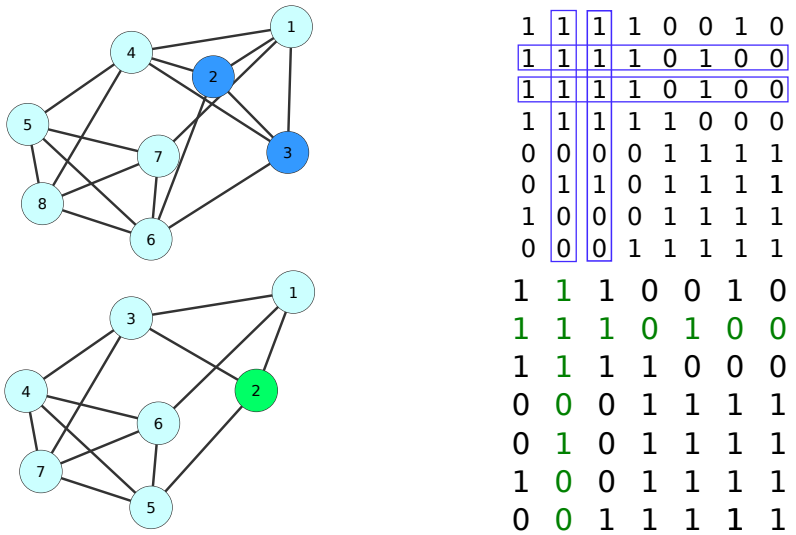


Figure 6.2: **Example of one iteration of the** *Information Content* **assessment process.** (Top Left) Initial network, composed of 8 nodes arranged in two communities (respectively composed of nodes $1-4$ and $5-8$). Notice that nodes 2 and 3 (in blue) share the same links. (Top Right) Adjacency matrix of the initial network; blue boxes highlight the four vector of incoming and outgoing links for nodes 2 and 3. (Bottom Left) The network after the merging process; the new node 2 (in green) is the result of merging the old nodes 2 and 3. (Bottom Right) Adjacency matrix of the resulting network. Reprinted with permission from Ref. [ZSM14].

$$p_1 = \frac{1}{2n}\sum_{i=1}^{2n} m_i, \tag{6.5}$$

$$p_0 = 1 - p_1. \tag{6.6}$$

Finally, the information contained in $\boldsymbol{m}$ is assessed through the Shannon's entropy [Sha48]:

$$I_{kl} = 2n\left(-p_0 \log_2 p_0 - p_1 \log_2 p_1\right). \tag{6.7}$$

I_{kl} is defined in $[0, 2n]$, being $I_{kl} = 0$ when $p_0 = 1$ or $p_1 = 1$, meaning that all links are respectively equal or different, and $I_{kl} = 2n$ when there is no correlation between the links of nodes k and l.

Once I has been assessed for all possible pairs of nodes, the algorithm identifies the pair whose merging will suppose minimum information loss. Such pair is then merged by deleting one of its nodes, and the original network is transformed into a new one composed of $n-1$ nodes (see Fig. 6.2 for an example). The whole process is then repeated iteratively, until one single node remains.

Each merging step supposes some loss of information (previously denoted by $I_{k,l}$): the *Information Content* IC is given by the total amount of information lost as a result of

the merging steps leading from the initial network to a single node. Conversely, it can be seen as the amount of information needed to reconstruct the full topology of the network, once it is reduced to a single node, by the merging process.

Two aspects of this metric should be clarified. Firstly, the information included in IC is not complete, as for instance at each step it would be necessary to track which pair of nodes has been merged: yet, the quantity of information required for this is constant, as does not depend on the topology of the network, and is thus discarded. Secondly, the Shannon entropy only provides a lower bound to the quantity of information required to encode vector $\boldsymbol{m}$, which may be lower than what required in real applications.

6.2.2 The meaning of Information Content

For a network with a completely random structure, no correlation is expected on average between incoming and outgoing links of any pair of nodes: thus, merging pairs of nodes will result in a nearly maximal I, and a maximal IC is expected. This can be used to normalize the *Information Content* of any network, such that:

$$IC_{norm} = IC/\langle IC_{random}\rangle, \tag{6.8}$$

$\langle IC_{random}\rangle$ being the average IC obtained for an ensemble of random networks with the same number of nodes and links of the original graph.

If $\langle IC_{random}\rangle$ provides the upper bound of IC, it is easy to find regular structures that will result in a very low *Information Content*. Clearly $IC = 0$ both for empty ($a_{ij} = 0, \forall i, j$) and fully connected networks ($a_{ij} = 1, \forall i, j$), as merging two nodes would suppose no information loss. More interestingly, the same will occurs with a fully modular network, such that

$$A = \begin{bmatrix} 1 & 1 & & 0 & 0 \\ 1 & 1 & & 0 & 0 \\ & & \cdots & & \\ 0 & 0 & & 1 & 1 \\ 0 & 0 & & 1 & 1 \end{bmatrix}. \tag{6.9}$$

The fact that all pairs of nodes have either the same or the opposite connections, thus either $p_1 = 0$ or $p_1 = 1$ and $I_{kl} = 0$ for any k and l, and $IC = IC_{norm} = 0$, can be used to assess the modularity of a network: moving from a perfectly modular to a random structure, the IC_{norm} smoothly raises from zero to one. Contrary to traditional community detection algorithms, IC_{norm} is unaffected by the presence of multiple, widely separated, scales. Both ideas are demonstrated in Fig. 6.3, in which different rewiring probabilities are applied to an initial network of 400 nodes, comprising two communities of different sizes.

More generally, IC can be used to assess the presence of any regular mesoscale structure. Consider for instance a bipartite network, *i.e.* networks where nodes belong to two

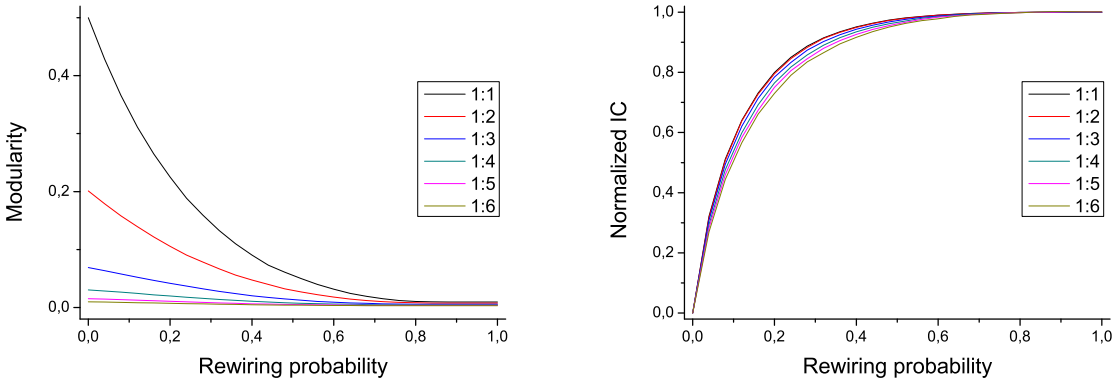


Figure 6.3: **Modularity vs. IC_{norm}.** (Left) Modularity (as calculated with the Blondel's community detection algorithm [BGLL08]) for a network of 400 nodes organized in two communities. The different lines represent different sizes of the two communities: $1:1$ (black line) two communities of 200 nodes, $1:2$ (red line) 134 and 266 nodes respectively, and so forth. (Right) Normalized *Information Content* for the same networks. Reprinted with permission from Ref. [ZSM14].

groups, with nodes belonging to one of them being connected only to nodes of the other. The resulting adjacency matrix would thus have the following structure:

$$A = \begin{bmatrix} 0 & 0 & & 1 & 1 \\ 0 & 0 & & 1 & 1 \\ & & \cdots & & \\ 1 & 1 & & 0 & 0 \\ 1 & 1 & & 0 & 0 \end{bmatrix}. \tag{6.10}$$

Similar results can also be obtained for networks showing a *core-periphery* structure, with a densely connected inner core, and a set of peripheral nodes sparsely connected with the core [Hol05]. In this case, merging nodes in the network core will result in low information loss, with a IC_{norm} lower than expected for random graphs.

The previously described meso-scale structures are mainly defined at a global level, in that they affect the overall topology of the network; thus, one may ask if the proposed IC is also effective in detecting more local meso-scales, *i.e.* those defined slightly above the single node level. Toward this aim, we test the measure against networks with high global Clustering Coefficient CC, defined as the number of closed triplets (or triangles) over the total number of (both open and closed) triplets. Networks were constructed following the classical method proposed by Watts and Strogatz [WS98], *i.e.* by starting from regular lattices of fixed degree (thus maximizing the Clustering Coefficient) and by applying a random rewiring process. Results are reported in Fig. 6.4, for networks of 200 nodes and initial degrees of 4, 6 and 8; a clear correlation can be found between IC_{norm} and CC, such that the higher the latter, the more regular is the resulting topology, thus yielding low IC_{norm} values.

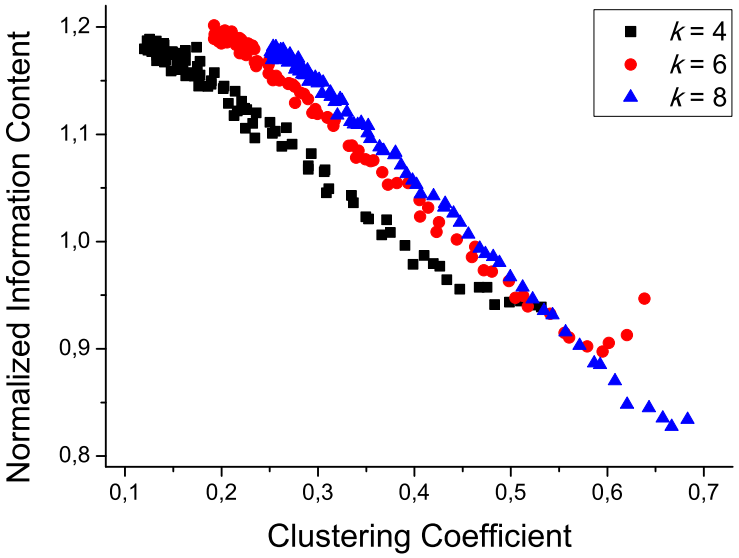


Figure 6.4: IC_{norm} **and Clustering Coefficient.** Evolution of the IC_{norm} as a function of the Clustering Coefficient. Black squares, red circles and blue triangles respectively correspond to networks with mean degree of 4, 6 and 8. Reprinted with permission from Ref. [ZSM14].

The Clustering Coefficient can be seen as a special case of *motifs*, *i.e.* sub-graphs recurring within a network with a higher than expected frequency [MSOIKCA02]. Their importance resides in the fact that they can be understood as basic building blocks, each one of them associated with specific functions within the global system [SOMMA02]. The main difference with complete triangles is that motifs are not necessarily symmetric nor complete, thus one expects a lower contribution toward creating regular structures in the adjacency matrix. By analyzing the IC_{norm} in random networks as a function of the frequency of appearance of different 3-nodes motifs, a significant correlation can be found with motifs 3 ($\rho = -0.7970$, $r^2 = 0.6194$), 5 ($\rho = -0.7557$, $r^2 = 0.5711$), 7 ($\rho = -0.7888$, $r^2 = 0.6222$) and 9 ($\rho = -0.7415$, $r^2 = 0.5498$) - for the enumeration of 3-nodes motifs, refer to Fig. A.3 or Fig. 1B of Ref. [MSOIKCA02].

6.2.3 Application to real networks

In summary, a low value of IC_{norm} indicates the presence of some kind of meso-scale regularity, although it gives no information about the specific type of structure detected; in other words, one knows that a structure is present, but not if it is a modular structure, a bipartite one, *etc.* Thus it is natural to complement the information yielded by IC with other common topological metrics. In order to stress this point, Fig. 6.5 presents four different phenospace of 55 real networks, covering social, biological and technological systems [KB76; HD89; WF94; BZCXZLZSLZ03; Lus03; MB04; OAS10]. Each network is represented as a point in the plane, whose coordinates are given by the IC_{norm} and by the value of a second topological metric, drawn from the following: ZScore of the maximum node degree, slope of the exponential fit of the degree distribution, modularity (as

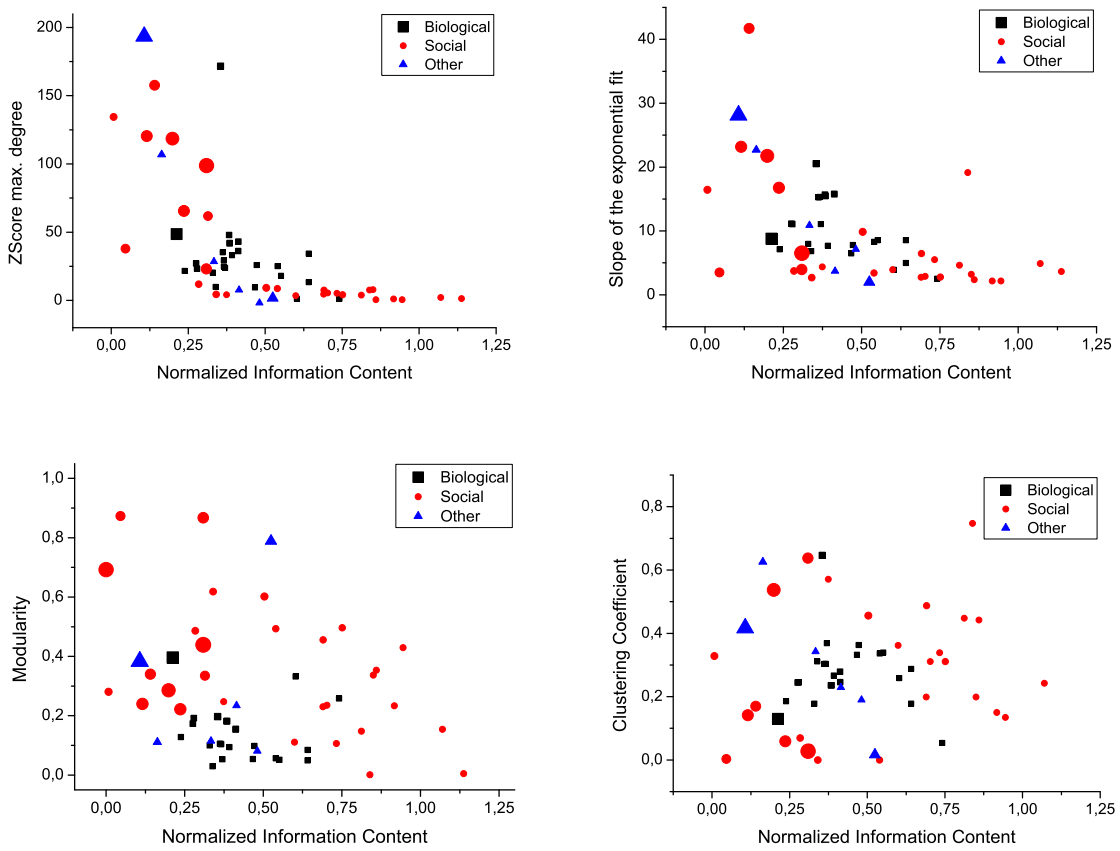


Figure 6.5: **Phenospaces of** 55 **real networks.** In the four panels, each network is represented by a point, whose coordinates are given by IC_{norm} and a second topological metric (*i.e.* from left to right, top to bottom, ZScore of the maximum node degree, slope of the exponential fit of the degree distribution, modularity and clustering coefficient). Colors encode the type of system represented by each network: black squares for biological systems, red circles for social, and blue triangles for other types of systems (mainly technological); the size of each point represents the size of the corresponding network. Reprinted with permission from Ref. [ZSM14].

calculated with the Blondel's community detection algorithm [BGLL08]) and clustering coefficient. If the pair of topological metrics considered in each phenospace were equivalent, one should expect all points to lay on a line. On the contrary, the four panels of Fig. 6.5 display a large variety of relationships. First, an inverse relationship between IC_{norm} on the one hand, and ZScore of the maximum node degree (top left panel) and the slope of the exponential fit (top right) on the other, can be appreciated; second, modularity and clustering coefficient yield graphs in which points cover the whole plane, indicating that the information they provide is not redundant. Thus, a low *Information Content* cannot immediately be associated to a given meso-scale feature, but it should be complemented with different phenospace analyses. It is also worth noticing the different behaviors corresponding to the different types of networks: social networks (red circles) cover the whole parameter space, while biological networks (black squares) seem to be bounded inside specific regions.

Information Content can also be used to assess the presence of different structures in weighted networks, by applying different thresholds and track how the IC_{norm} evolves. As a test case, here we consider three brain functional networks [BS09], obtained through magneto-encephalographic (MEG) recordings of three healthy subjects performing a Sternberg's letter-probe task. For each subject, a weighted clique of size 148×148 was computed using the MEG time series, where the weights are given by the correlation between each pair of sensors as calculated by means of a Synchronization Likelihood (SL) algorithm [SD02] - see Section 5.2 for further details.

Fig. 6.6 reports the evolution of the modularity and of the normalized *Information Content* for the three subjects as a function of the applied threshold. While the former has a monotonous behavior (except for high thresholds, where the reduced amount of links results in strong fluctuations), the IC_{norm} presents a clear maximum corresponding to a threshold of $0.2 - 0.25$. This region of reduced topological regularity points to a change in the structure of the networks, which is consistent with the varying fractal topology of the human brain at different synchronization thresholds [BMLADB06; GMS12].

6.2.4 Information Content for feature selection

If one considers the algorithm through which *Information Content* is calculated, it appears natural to make a parallelism with a feature selection problem in complex networks. Specifically, the previously described process iteratively eliminates nodes, selecting those whose deletion supposes the minimum information loss; thus, at each iteration, the reduced network should maintain most of the characteristics of the initial one. In order to test this hypothesis, we have performed again the classification problem described in Section 5.2, and calculated the resulting score as a function of the number of nodes maintained in the network. Fig. 6.7 clearly shows that the network size can be reduced up to a 40% without affecting the discrimination power.

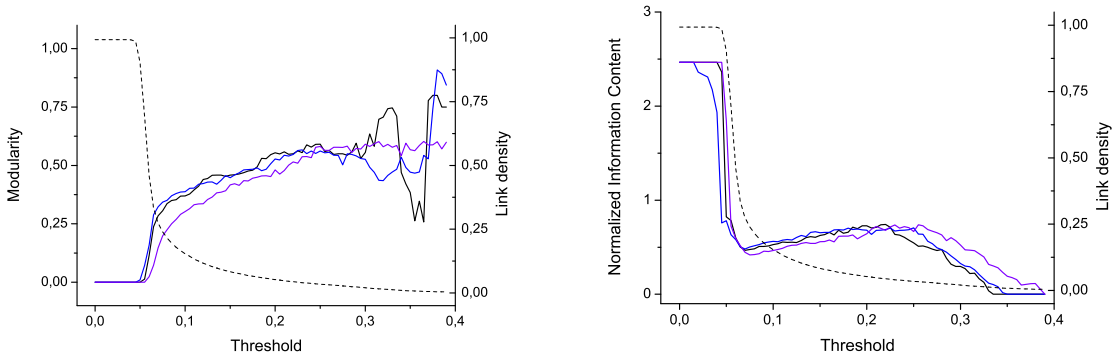


Figure 6.6: **Modularity and IC_{norm} in weighted functional brain networks.** Evolution of the modularity (Left) and of the normalized *Information Content* (Right) for three human brain functional networks, as a function of the applied threshold. Dotted gray lines represent the corresponding link density (right axes). Reprinted with permission from Ref. [ZSM14].

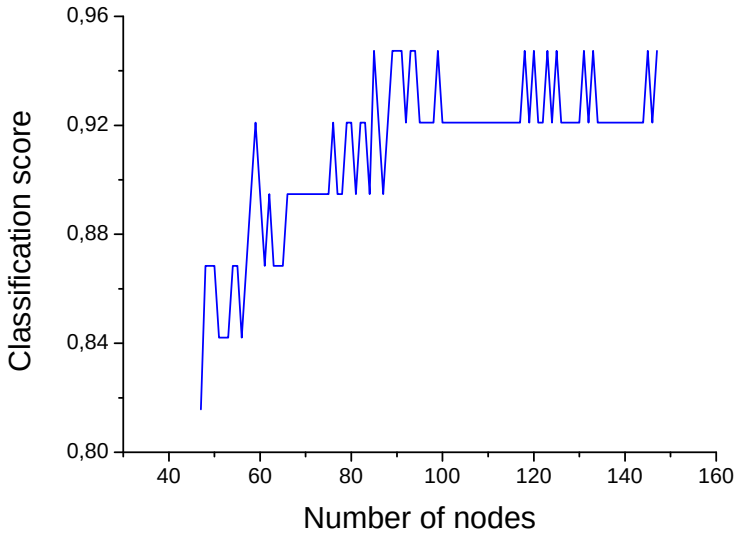


Figure 6.7: **Information Content and feature selection.** Classification score as a function of the number of nodes composing the network, for the problem described in Section 5.2.

6.2.5 Conclusions

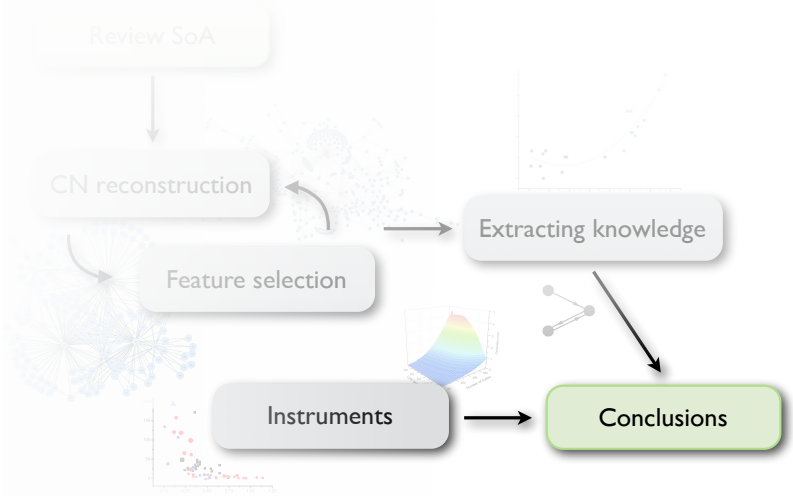
In conclusion, we have here reported on the definition of a new metric designed to assess the presence of regular meso-scale structures in complex networks. While other metrics, *e.g.* modularity, are defined *a posteriori*, that is the community structure should be detected before the calculation of the modularity of a network, *Information Content* can be obtained directly from the adjacency matrix. Furthermore, it is an exact metric, whose output does not depend on initial conditions or specific algorithm implementations. Finally, it enables the simultaneous assessment of different meso-scale structures, providing information complementary to standard measures.

Specifically, results corresponding to the analysis of different real-world networks have shown that classical metrics, *e.g.* modularity or clustering coefficient, are not enough to fully characterize them. *Information Content* is able to detect the presence of regular structures that are neglected by standard techniques, and deliver such information in a way that is compatible with standard data mining analyses. Additionally, such meso-scale structures seem to play an important role in feature selection tasks, as they indicate sets of node creating redundant structures. For all this, *Information Content* is expected to provide important benefits in tasks requiring the systematic and automatized analysis of large sets of networks, as in the case of classification tasks.

The algorithm has been programmed in MATLAB®, and is freely available at http://www.mzanin.com/IC.

7

Conclusions and future lines of research



7.1 Toward a new perspective for the understanding of complex systems

When one wants to understand a complex system, *i.e.* a system where interactions between its constituting elements are expected to be as important as the elements themselves, it is usually recognized that a complex network representation is the most appropriate tool, as it is able to explicitly describe the structure created by such interactions in an elegant form. Yet, as shown in this Thesis, complementing such representation with data mining techniques yields more significant results, and thus a more profound knowledge about the problem.

Starting from an initial set of raw data, a process can be defined that ends up with the attainment of new knowledge about the system. The three main steps of such process are synthesized here below.

1. Firstly, a data set describing some problem, biomedical or of any other nature, should be represented by means of complex networks. This can be done in several ways: directly mapping physical or abstract connections between the elements of the system, if these connections are available; creating links when a correlation is detected between the evolution of element characteristics, if such evolution is present; or by using the technique proposed in Section 3.1, when the starting point is a set of static multivariate data.

   This PhD Thesis has been focused on complex network analyses, and has therefore not tackled other steps that should be accomplished in advance, and that are essential for obtaining meaningful results: they are defined in knowledge discovery as *business understanding*, *data understanding* and *data preparation*. Without a good understanding of the problem at hand and of its specific characteristics, results obtained in subsequent phases may be meaningless. Thus, one should never forget that domain knowledge is a prerequisite, not just in network analysis, but in science in general.

2. When the number of elements composing the system is too large, or when some elements are suspected to be irrelevant for the analysis, a feature selection step is of advice. There is a large literature on how to perform such step, and here we have illustrated how some techniques are specially befitting when representing a system by means of complex networks.

3. Finally, some metrics should be extracted from each one of the reconstructed networks. Instead of selecting a set of them *a priori*, it is advisable to extract a large collection of topological features, for then using some data mining task, *e.g.* a classification, to detect which are the most meaningful for describing the problem.

By following the three previous steps, one can move from a raw set of data, up to a small collection of network metrics describing the initial problem. Clearly, there is one

last issue to be addressed, *i.e.* what is called *evaluation* in KDD: studying the obtained metrics in the light of the *business requirements* set at the beginning. In other words, such metrics should be used to gain deeper understanding of the problem under study, which in turn may result in new questions arising.

7.2 Review of the Thesis objectives

In this document we have presented the main results obtained within this PhD Thesis. As demonstrated through a large number of validation cases drawn from actual biomedical problems, data mining and complex network techniques can be fruitfully merged together, enabling the improvement of our comprehension of complex systems, as the human brain or the onset of specific diseases. The strong point of the latter, *i.e.* the complex network capability of synthesizing large structures of interactions in simple matrices, naturally blends with the ability of data mining for managing large sets of data. As a result, knowledge can be extracted from complex systems, and in this specific case from the human body, in a way not possible before.

Specifically, we have reported how such interaction between complex networks and data mining can be beneficial in the light of the four research questions proposed in Section 1.1:

Use of feature selection techniques. While the customary approach in complex systems analysis recommends mapping all the constituting elements into nodes of a network, this presents some important disadvantages. Besides of an increased computational cost, the inclusion of noisy nodes, *i.e.* of nodes codifying irrelevant information, may reduce the performance of any subsequent analyses.

As shown in Chapter 4, by using established *feature selection* techniques, it is possible to identify a subset of elements representing most of the relevant information, thus reducing the size of the resulting network in up to two orders of magnitude. In addition of drastically cutting down the computational cost of network analysis, the compression of the number of features yields an increase of the statistical significance of results.

These improvements are of special relevance in the case of the analysis of spectral data, as the one presented in Section 4.3: starting from vectors of 10.000 measurements, whose characterization would be a challenging task in a normal computer, it is possible to obtain a subset of 100 nodes, still detecting the presence of the disease mark. Notice that reducing the size of the network in two orders of magnitude usually implies reducing the analysis' computational cost by four orders, due to the way topological metrics are calculated. Such results thus confirm the validity of the first hypothesis of this Thesis (see Section 1.1).

Network reconstruction through data mining techniques. The reconstruction of network representations of a system has usually been limited to two scenarios: systems with

a clear physical background, *e.g.* transportation networks, or *functional networks*, based on the analysis of the relationships between the time evolution of some features. This precluded the analysis of systems whose elements are described by *static* features, *i.e.* scalar measurements, like gene expression levels.

In Section 3.1 we presented a new methodology for dealing with such situations, based on the identification of characteristic patterns between pairs of features of the population under study. In Sections 3.2 and 3.3 such methodology has been validated by means of two different data sets, respectively composed of genetic and spectral data. In both cases, it was possible to characterize the disease under study through the characteristics of the resulting network topology. It is worth stressing that this is the first time such types of data have been represented by dint of complex networks. Furthermore, the scores obtained in the corresponding classification tasks were higher than the ones obtained by standard data mining techniques, thus validating the second hypothesis.

Use of network representations for improving data mining tasks. Complex networks are mathematical objects that cannot directly be used as the input of data mining algorithms, *i.e.* they cannot directly feed an Artificial Neural Network or a Support Vector Machine. This barrier can be avoided by firstly transform a network into a set of topological indicators, describing in simple numbers several features of its structure, for then secondly use such indicators as inputs of the data mining process. Data mining techniques can then be used to select those topological features that are relevant for the problem at hand, thus ensuring optimality and objectivity in the analysis. Through Chapters 3 and 5 this process has been repeatedly performed on different validation cases, demonstrating its feasibility and its superior performance when compared to more classical approaches.

Use of data mining tools to validate network representations. Finally, the analysis of network representations has been mainly performed on a subjective basis. For instance, the election of some parameters, like the synchronization metrics in the study of the brain activity, or the threshold for the creation of unweighted networks, was manually performed by the researcher. Furthermore, the selection of the best topological metrics for describing a set of networks was also left to individual judgment.

In this Thesis we have shown how data mining can be used to automatize such steps. Specifically, when a labeled set of subjects is available, the score of a data mining task represents the quantity of information successfully codified in the networks. Therefore, the optimal set of parameters for the network reconstruction can easily be found by scanning all combinations, and selecting the one achieving the best classification. Section 5.2 applies this approach to the case of magnetoencephalographic data, challenging some established conventions, like the use

of sparse network representations. Furthermore, in Section 5.3, this methodology allowed an objective comparison of different synchronization metrics for brain dynamics analysis, showing that some of them, widely used in the Literature, are not suited for this aim.

In resume, the novel contributions of this Thesis to the state of the art, *i.e.* to the analysis of data sets by means of complex networks and data mining techniques, are the following:

- A new algorithm for reconstructing networks from data sets, especially designed to detect deviations from standard patterns.
- A methodology for applying feature selection techniques in complex networks, aimed at reducing the number of nodes in the graph, and thus the dimensionality of the problem.
- A methodology for optimizing the network representation of a complex system, *i.e.* to choose the best parameters in the reconstruction phase.
- A methodology for using networks as input for a data mining task. In order to optimize the implementation of such methodology, two novel algorithms have been developed, aimed at assessing two network topological properties, *i.e.* the fast enumeration of motifs and the presence of meso-scales.

7.3 Future lines of research

While several questions have been successfully tackled in this work, some new challenges have also emerged, which we hope will be the target of future research beyond this PhD. Among them, some of the most important are:

Multi-layer networks. Throughout the work presented in this document, a fundamental assumption has been the homogeneity of links, *i.e.* that all connections between the elements of the system were of a same kind. While this simplification enables the elegant matricial analysis of the problem, it is known that many real-world systems require a more detailed modeling. This is the case, for instance, of social networks, where connections between individuals may represent different social relationships: two people may be friends, colleagues, partners, and so forth. Representing such different relationships by the same type of links is an oversimplification that may distort the conclusions drawn from the analysis. As described in Section 2.2.3, the solution is the creation of *multi-layer* (or *multiplex*) networks, *i.e.* of a structure composed of multiple layers, usually with the same nodes, each one of them comprising information about one kind of relationship [BBCGGGRSNWZ14].

Such a richer representation may yield important benefits for biomedical analyses. For instance, each network may represent the status of the subject with respect

to a given disease; when multiple networks are stacked, the resulting multi-layer structure would represent the global health status of the person. Another example may include the study of the brain, in which different analyses, *e.g.* connectome, magnetoencephalograms or electroencephalograms, may be merged together in a single representation.

Yet, the theory behind multi-layer networks is still far from mature. While some proposals have been made, as for instance in Refs. [KT06; MRMPO10; BPPSH10; CGGZRPPB13; BBCGGGRSNWZ14], there is still no consensus on how to apply classical topological metrics to a multi-layer network.

Interpretation of network topologies. As a result of the optimization methodology presented in Section 5.1, motifs have appeared as one of the best metrics to analyze different data sets, from gene expression levels up to tissue spectra. Yet, it is not clear how to interpret those results. While in genetic networks some motifs have been related to circuits performing specific functions, like filtering transients in a regulatory signal [SOMMA02], the meaning of a motif in spectral data is still to be uncovered.

Extracting biomedical knowledge from network topologies is the required step to move from a merely theoretical work, in which several characteristics are detected, up to real applications, *i.e.* the development of prognosis techniques or treatments.

Analysis of large data sets. During the development of this PhD, a large effort has been devoted to the reduction of the computational cost associated with network analysis. Several techniques have been developed: from the parallelization of some steps, the pre-calculation of network metrics, up to the development of a specific algorithm for the enumeration of motifs in dense networks (see Section 6.1). This has allowed the reduction of the computational time associated to the analysis of a medium-size data set from weeks to hours. Yet, this is not enough for the study of large-scale systems, or the real-time study of evolving systems, like for instance social networks or complete genetic data sets.

Several strategies may be explored in this context: for instance, streaming processing of data, when new subjects are from time to time added to the initial set, or when new biomedical analyses are added to the records; or the massive parallelization of network topology analyses, *e.g.* by means of GPU hardware [SK10; WWSWHXY10].

7.4 Acknowledgments

During the development of this PhD Thesis, large computational capabilities were required. The Candidate acknowledges the following institutions for the use of their infrastructures and their technical support:

Amazon AWS. Part of the computations have been performed in the Amazon Web Service, through its Elastic Cloud Computing infrastructure [MVM10].

ENEA. The Candidate acknowledge the computational resources, facilities and assistance provided by the Centro computazionale di RicErca sui Sistemi COmplessi (CRESCO) of the Italian National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA).

CESVIMA. Some validation tests have been performed in the Universidad Politécnica de Madrid's CeSViMa (Madrid Supercomputing and Visualization Center) infrastructure.

CIEMAT. The computing facilities of Extremadura Research Centre for Advanced Technologies (CETA-CIEMAT), funded by the European Regional Development Fund (ERDF), have been the basis of several parallelization tests. CETA-CIEMAT belongs to CIEMAT and the Government of Spain.

Bibliography

[On] http://tunedit.org/challenge/ON.

[Tra] http://bioinfogp.cnb.csic.es/transplanta_dev/.

[AH08] K. Abdulrab and R. Heun. "Subjective Memory Impairment. A review of its definitions indicates the need for a comprehensive set of standardised and validated criteria". In: *European Psychiatry* 23.5 (2008), pp. 321–330.

[AMR04] M. D. Abràmoff, P. J. Magalhães, and S. J. Ram. "Image processing with ImageJ". In: *Biophotonics international* 11.7 (2004), pp. 36–43.

[AQDWCCSSYFW02] B.-L. Adam, Y. Qu, J. W. Davis, M. D. Ward, M. A. Clements, L. H. Cazares, O. J. Semmes, P. F. Schellhammer, Y. Yasui, Z. Feng, and G. L. Wright. "Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men". In: *Cancer Research* 62.13 (2002), pp. 3609–3614.

[Aha97] D. W. Aha. "Lazy learning: Special issue editorial". In: *Artificial Intelligence Review* 11 (1997), pp. 7–10.

[AB02] R. Albert and A. Barabási. "Statistical mechanics of complex networks". In: *Review of Modern Physics* 74 (2002).

[ACLBSN11] J. A. Almendral, R. Criado, I. Leyva, J. M. Buldú, and I. Sendina-Nadal. "Introduction to focus issue: Mesoscales in complex networks". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 21.1 (2011), p. 016101.

[And72] P. W. Anderson. "More is different". In: *Science* 177.4047 (1972), pp. 393–396.

[ADDGGG04] A. Arenas, L. Danon, A. Díaz-Guilera, P. M. Gleiser, and R. Guimerà. "Community analysis in social networks". In: *Eur. Phys. J. B* 38 (2004), pp. 373–380.

[ADGKMZ08] A. Arenas, A. Díaz-Guilera, J. Kurths, Y. Moreno, and C. Zhou. "Synchronization in complex networks". In: *Physics Reports* 469.3 (2008), pp. 93–153.

[ADGPV06] A. Arenas, A. Díaz-Guilera, and C. J. Pérez-Vicente. "Synchronization reveals topological scales in complex networks". In: *Physical review letters* 96.11 (2006), p. 114102.

[Atl09] S. W. Atlas. *Magnetic resonance imaging of the brain and spine*. Vol. 1. Lippincott Williams & Wilkins, 2009.

[BS01] L. A. Baccala and K. Sameshima. "Partial directed coherence: a new concept in neural structure determination". In: *Biological cybernetics* 84.6 (2001), pp. 463–474.

[BMB07] M. Banerjee, S. Mitra, and H. Banka. "Evolutionary rough feature selection in gene expression data". In: *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 37.4 (2007), pp. 622–632.

[Bar09] A. Barabási. "Scale-free networks: a decade and beyond". In: *Science* 325.5939 (2009), pp. 412–413.

[BA99] A.-L. Barabási and R. Albert. "Emergence of scaling in random networks". In: *Science* 286.5439 (1999), pp. 509–512.

[BO04] A.-L. Barabási and Z. N. Oltvai. "Network biology: understanding the cell's functional organization". In: *Nature Reviews Genetics* 5.2 (2004), pp. 101–113.

[BGBHRDS12] G. W. Bassel, A. Gaudinier, S. M. Brady, L. Hennig, S. Y. Rhee, and I. De Smet. "Systems analysis of plant functional, transcriptional, physical interaction, and metabolic networks". In: *The Plant Cell Online* 24.10 (2012), pp. 3859–3875.

[BBMLAWC09] D. S. Bassett, E. T. Bullmore, A. Meyer-Lindenberg, J. A. Apud, D. R. Weinberger, and R. Coppola. "Cognitive fitness of cost-efficient brain functional networks". In: *Proceedings of the National Academy of Sciences* 106.28 (2009), pp. 11747–11752.

[BMLADB06] D. S. Bassett, A. Meyer-Lindenberg, S. Achard, T. Duke, and E. Bullmore. "Adaptive reconfiguration of fractal small-world human brain functional networks". In: *Proceedings of the National Academy of Sciences* 103.51 (2006), pp. 19518–19523.

[BM02] V. Batagelj and A. Mrvar. "Pajek: Analysis and visualization of large networks". In: *Graph Drawing*. Springer. 2002, pp. 8–11.

[BEBEKAHLN03] O. Beckonert, M. E. Bollard, T. Ebbels, H. C. Keun, H. Antti, E. Holmes, J. C. Lindon, and J. K. Nicholson. "NMR-based metabonomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches". In: *Analytica Chimica Acta* 490.1 (2003), pp. 3–15.

[BLMBC02] R. Bellazzi, C. Larizza, P. Magni, R. Bellazzi, and S. Cetta. "Intelligent Data Analysis techniques for Quality assessment of hemodialysis services". In: (2002).

[BS10] B. P. Bezručko and D. A. Smirnov. *Extracting Knowledge from Time Series: An Introduction to Nonlinear Empirical Modeling*. Springer, 2010.

[Bia06] G. Bianconi. "Degree distribution of complex networks from statistical mechanics principles". In: *arXiv:cond-mat/0606365* (2006).

[Bis06] C. M. Bishop. *Pattern recognition and machine learning*. Vol. 1. Springer New York, 2006.

[BGLL08] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008), P10008.

[BBCGGGRSNWZ14] S. Boccaletti, G. Bianconi, R. Criado, C. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin. "The structure and dynamics of multilayer networks". In: *Physics Reports* (2014).

[BLMCH06] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang. "Complex networks: Structure and dynamics". In: *Physics Reports* 424 (2006).

[Bol01] B. Bollobás. *Random graphs*. Cambridge University Press, 2001.

[BL01] P. Bonacich and P. Lloyd. "Eigenvector-like measures of centrality for asymmetric relations". In: *Social Networks* 23 (2001), pp. 191–201.

[Bra02] E. A. Bray. "Classification of genes differentially expressed during water-deficit stress in Arabidopsis thaliana: An analysis using microarray and differential expression data". In: *Annals of Botany* 89.7 (2002), pp. 803–811.

[BFOS84] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Chapman & Hall / CRC, 1984.

[BP98] S. Brin and L. Page. "The anatomy of a large-scale hypertextual Web search engine". In: *Computer networks and ISDN systems* 30.1 (1998), pp. 107–117.

[BRDP98] S. P. van den Broek, F. Reinders, M. Donderwinkel, and M. J. Peters. "Volume conduction effects in EEG and MEG". In: *Electroencephalography and clinical neurophysiology* 106.6 (1998), pp. 522–534.

[BZCXZLZSLZ03] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, and N. Zhang. "Topological structure analysis of the protein–protein interaction network in budding yeast". In: *Nucleic acids research* 31.9 (2003), pp. 2443–2450.

[BBMCLGSNANDPB11] J. M. Buldú, R. Bajo, F. Maestú, N. Castellanos, I. Leyva, P. Gil, I. Sendiña-Nadal, J. A. Almendral, A. Nevado, F. Del-Pozo, and S. Boccaletti. "Reorganization of functional networks in mild cognitive impairment". In: *PloS one* 6.5 (2011), e19584.

[BPPSH10] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin. "Catastrophic cascade of failures in interdependent networks". In: *Nature* 464.7291 (2010), pp. 1025–1028.

[BS09] E. Bullmore and O. Sporns. "Complex brain networks: graph theoretical analysis of structural and functional systems". In: *Nature Reviews Neuroscience* 10.3 (2009), pp. 186–198.

[Bur98] C. J. C. Burges. "A tutorial on support vector machines for pattern recognition". In: *Data mining and knowledge discovery* 2.2 (1998), pp. 121–167.

[CGGZRPPB13] A. Cardillo, J. Gómez-Gardeñes, M. Zanin, M. Romance, D. Papo, F. del Pozo, and S. Boccaletti. "Emergence of network features from multiplexity". In: *Scientific Reports* 3 (2013).

[CZGGRAB13] A. Cardillo, M. Zanin, J. Gómez-Gardeñes, M. Romance, A. J. G. del Amo, and S. Boccaletti. "Modeling the multi-layer nature of the European Air Transport Network: Resilience and passengers re-scheduling under random failures". In: *The European Physical Journal Special Topics* 215.1 (2013), pp. 23–33.

[CKB87] B. Cestnik, I. Kononenko, and I. Bratko. "Assistant 86: A knowledge elicitation tool for sophisticated users". In: *Progress in machine learning* (1987), pp. 31–45.

[CCKKRSW] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. *CRISP-DM 1.0*. Tech. rep.

[Che99] R. L. Chevalier. "Molecular and cellular pathophysiology of obstructive nephropathy". In: *Pediatr. Nephrol.* 13 (1999), pp. 612–619.

[Che04] R. L. Chevalier. "Biomarkers of congenital obstructive nephropathy: past, present and future". In: *The Journal of Urology* 172.3 (2004), pp. 852–857.

[Che06] R. L. Chevalier. "Obstructive nephropathy: towards biomarker discovery and gene therapy". In: *Nature Reviews Nephrology* 2 (2006), pp. 157–168.

[Cla88] J. W. Clark. "Probabilistic neural networks". In: *Evolution, Learning and Cognition* (1988), pp. 129–180.

[CLPSEMDW05] R. Clifton, R. Lister, K. L. Parker, P. G. Sappl, D. Elhafez, A. H. Millar, D. A. Day, and J. Whelan. "Stress-induced co-expression of alternative respiratory chain components in Arabidopsis thaliana". In: *Plant molecular biology* 58.2 (2005), pp. 193–212.

[Coh72] D. Cohen. "Magnetoencephalography: detection of the brain's electrical activity with a superconducting magnetometer". In: *Science* 175.4022 (1972), pp. 664–666.

[Coh95] W. W. Cohen. "Fast effective rule induction". In: *Proceedings of the Twelfth International Conference on Machine Learning*. 1995, pp. 115–123.

[CV95] C. Cortes and V. Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297.

[CRTV07] L. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. "Characterization of complex networks: A survey of measurements". In: *Advances in Physics* 56 (2007).

[Cou99] W. G. Couser. "Glomerulonephritis". In: *The Lancet* 353 (1999), pp. 1509–1515.

[CH67] T. Cover and P. Hart. "Nearest neighbor pattern classification". In: *IEEE Transactions on Information Theory* 13.1 (1967), pp. 21–27.

[Dan05] J. Dan. *Kabbalah: A very short introduction*. Oxford University Press, USA, 2005.

[DDGA06] L. Danon, A. Díaz-Guilera, and A. Arenas. "The effect of size heterogeneity on community identification in complex networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2006.11 (2006), P11010.

[DM04] L. Demetrius and T. Manke. "Robustness and network evolution - an entropic principle". In: *Physica A* 346.3-4 (2004), pp. 682–696.

[DMEHMGF91] L. DeToledo-Morrell, S. Evers, T. J. Hoeppner, F. Morrell, D. C. Garron, and J. H. Fox. "A'Stress' Test for Memory Dysfunction: Electrophysiologic Manifestations of Early Alzheimer's Disease". In: *Archives of neurology* 48.6 (1991), p. 605.

[DP05] C. Ding and H. Peng. "Minimum redundancy feature selection from microarray gene expression data". In: *Journal of bioinformatics and computational biology* 3.02 (2005), pp. 185–205.

[DGFMCM10] C. Doukas, T. Goudas, S. Fischer, I. Mierswa, A. Chatzioannou, and I. Maglogiannis. "An open data mining framework for the analysis of medical images: Application on Obstructive Nephropathy microscopy images". In: *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*. IEEE. 2010, pp. 4108–4111.

[DA05] J. Duch and A. Arenas. "Community detection in complex networks using extremal optimization". In: *Phys. Rev. E* 72 (2005), p. 027104.

[ET93] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[ECCBA05] V. M. Eguiluz, D. R. Chialvo, G. A. Cecchi, M. Baliki, and A. V. Apkarian. "Scale-free brain functional networks". In: *Physical review letters* 94.1 (2005), p. 18102.

[EA08] L. Elfangary and W. A. Atteya. "mining medical databases using Proposed Incremental Association Rules Algorithm (PIA)". In: *Digital Society, 2008 Second International Conference on the*. IEEE. 2008, pp. 88–92.

[ER59] P. Erdös and A. Rényi. "On random graphs". In: *Publicationes Mathematicae* 6.290–297 (1959).

[ER60] P. Erdös and A. Rényi. "On the evolution of random graphs". In: *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5 (1960), pp. 17–61.

[FCOTRBAVR11] L. da F. Costa, O. N. Oliveira, G. Travieso, F. A. Rodrigues, P. R. V. Boas, L. Antiqueira, M. P. Viana, and L. E. C. Rocha. "Analyzing and modeling real-world phenomena with complex networks: a survey of applications". In: *Advances in Physics* 60.3 (2011), pp. 329–412.

[FMJ05] S. T. Farias, D. Mungas, and W. Jagust. "Degree of discrepancy between self and other-reported everyday functioning by cognitive status: dementia, mild cognitive impairment, and healthy elders". In: *International journal of geriatric psychiatry* 20.9 (2005), pp. 827–834.

[FPSS96] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. *Advances in knowledge discovery and data mining*. MIT Press, 1996.

[Fle04] F. Fleuret. "Fast binary feature selection with conditional mutual information". In: *The Journal of Machine Learning Research* 5 (2004), pp. 1531–1555.

[For10] S. Fortunato. "Community detection in graphs". In: *Physics Reports* 486.3-5 (2010), pp. 75–174.

[FB07] S. Fortunato and M. Barthelemy. "Resolution limit in community detection". In: *Proceedings of the National Academy of Sciences* 104.1 (2007), pp. 36–41.

[Fro12] G. Frobenius. "Ueber Matrizen aus nicht negativen Elementen". In: *Sitzungsber. Königl. Preuss. Akad. Wiss.* (1912), pp. 456–477.

[Fur99] J. Furnkranz. "Separate-and-conquer rule learning". In: *Artificial Intelligence Review* 13.1 (1999), pp. 3–54.

[GMS12] L. K. Gallos, H. A. Makse, and M. Sigman. "A small world of weak ties provides optimal global integration of self-similar modules in functional brain networks". In: *Proceedings of the National Academy of Sciences* 109.8 (2012), pp. 2825–2830.

[GA04] N. G. Gençer and C. E. Acar. "Sensitivity of EEG and MEG measurements to tissue conductivity". In: *Physics in medicine and biology* 49.5 (2004), p. 701.

[GZ06] S. Gil and D. H. Zanette. "Coevolution of agents and networks: Opinion spreading and community disconnection". In: *Physics Letters A* 356.2 (2006), pp. 89–94.

[Gof74] E. Goffman. *Frame analysis: An essay on the organization of experience.* Harvard University Press, 1974.

[GSMESRPA14] J. L. González-Solís, J. C. Martínez-Espinosa, J. M. Salgado-Román, and P. Palomares-Anda. "Monitoring of chemotherapy leukemia treatment using raman spectroscopy and principal component analysis". In: *Lasers in medical science* 29.3 (2014), pp. 1241–1249.

[Gra69] C. W. J. Granger. "Investigating causal relations by econometric models and cross-spectral methods". In: *Econometrica: Journal of the Econometric Society* (1969), pp. 424–438.

[GWSCRN01] J. L. Griffin, H. J. Williams, E. Sang, K. Clarke, C. Rae, and J. K. Nicholson. "Metabolic Profiling of Genetic Disorders: A Multitissue ^{1}H Nuclear Magnetic Resonance Spectroscopic and Pattern Recognition Study into Dystrophic Tissue". In: *Analytical Biochemistry* 293.1 (2001), pp. 16–21.

[Gru+04] M. Grundman et al. "Mild cognitive impairment can be distinguished from Alzheimer disease and normal aging for clinical trials". In: *Archives of neurology* 61.1 (2004), p. 59.

[GA05] R. Guimerà and L. A. N. Amaral. "Functional cartography of complex metabolic networks". In: *Nature* 433 (2005), pp. 895–900.

[GSPA07] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral. "Classes of complex networks defined by role-to-role connectivity profiles". In: *Nature physics* 3.1 (2007), pp. 63–69.

[GHLBGWL05] A. Guo, K. He, D. Liu, S. Bai, X. Gu, L. Wei, and J. Luo. "DATF: a database of Arabidopsis transcription factors". In: *Bioinformatics* 21.10 (2005), pp. 2568–2569.

[GE03] I. Guyon and A. Elisseeff. "An introduction to variable and feature selection". In: *The Journal of Machine Learning Research* 3 (2003), pp. 1157–1182.

[GGBHD04] I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror. "Result analysis of the NIPS 2003 feature selection challenge". In: *Advances in Neural Information Processing Systems* 17 (2004), pp. 545–552.

[GGNZ06] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh. *Feature extraction: foundations and applications*. Vol. 207. Springer, 2006.

[HHIKL93] M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa. "Magnetoencephalographytheory, instrumentation, and applications to noninvasive studies of the working human brain". In: *Reviews of modern Physics* 65.2 (1993), p. 413.

[Ham11] L. H. Hamel. *Knowledge discovery with support vector machines*. Wiley-Interscience, 2011.

[HW12] E. R. Hancock and R. C. Wilson. "Pattern analysis with graphs: Parallel work at Bern and York". In: *Pattern Recognition Letters* 33.7 (2012), pp. 833–841.

[HTFJ01] T. Hastie, R. Tibshirani, J. Friedman, and H. Jerome. *The elements of statistical learning*. Vol. 1. Springer New York, 2001.

[Hay07] S. S. Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall Englewood Cliffs, NJ, 2007.

[HH09] S. Herculano-Houzel. "The human brain in numbers: a linearly scaled-up primate brain". In: *Frontiers in Human Neuroscience* 3 (2009), p. 31.

[Hol05] P. Holme. "Core-periphery organization of complex networks". In: *Physical Review E* 72.4 (2005), p. 046111.

[HS12] P. Holme and J. Saramäki. "Temporal networks". In: *Physics reports* 519.3 (2012), pp. 97–125.

[HS13] P. Holme and J. Saramäki. *Temporal networks*. Springer, 2013.

[HKBS07] C. J. Honey, R. Kötter, M. Breakspear, and O. Sporns. "Network structure of cerebral cortex shapes functional connectivity on multiple time scales". In: *Proceedings of the National Academy of Sciences* 104.24 (2007), pp. 10240–10245.

[HSM04] S. A. Huettel, A. W. Song, and G. McCarthy. *Functional magnetic resonance imaging*. Vol. 1. Sinauer Associates Sunderland, MA, 2004.

[Hug68] G. Hughes. "On the mean accuracy of statistical pattern recognizers". In: *Information Theory, IEEE Transactions on* 14.1 (1968), pp. 55–63.

[HD89] N. P. Hummon and P. Dereian. "Connectivity in a citation network: The development of DNA theory". In: *Social Networks* 11.1 (1989), pp. 39–63.

[HG08] M. D. Humphries and K. Gurney. "Network 'small-world-ness': a quantitative method for determining canonical network equivalence". In: *PLoS One* 3 (2008), e0002051.

[HMS66] E. B. Hunt, J. Marin, and P. J. Stone. *Experiments in induction*. Academic Press, 1966.

[ILBC04] I. Inza, P. Larrañaga, R. Blanco, and A. J. Cerrolaza. "Filter versus wrapper gene selection approaches in DNA microarray domains". In: *Artificial intelligence in medicine* 31.2 (2004), pp. 91–103.

[IVCF02] H. J. Issaq, T. D. Veenstra, T. P. Conrads, and D. Felschow. "The SELDI-TOF MS approach to proteomics: protein profiling and biomarker identification". In: *Biochemical and biophysical research communications* 292.3 (2002), pp. 587–592.

[Jen96] F. V. Jensen. *An introduction to Bayesian networks*. UCL press, 1996.

[Joh00] S. John. *Social network analysis: a handbook*. 2000.

[Kar03] J. Karmeshu. *Entropy measures, maximum entropy principle and emerging applications*. Springer-Verlag New York, Inc., 2003.

[KIMA02] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. "Mfinder tool guide". In: *Department of Molecular Cell Biology and Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot Israel, Tech. Rep* (2002).

[KWHWWBDBBKH07] J. Kilian, D. Whitehead, J. Horak, D. Wanke, S. Weinl, O. Batistic, C. DAngelo, E. Bornberg-Bauer, J. Kudla, and K. Harter. "The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses". In: *The Plant Journal* 50.2 (2007), pp. 347–363.

[KB76] P. D. Killworth and H. R. Bernard. "Informant accuracy in social network data". In: *Human Organization* 35.3 (1976), pp. 269–286.

[KWWCESPDD02] E. J. Klok, I. W. Wilson, D. Wilson, S. C. Chapman, R. M. Ewing, S. C. Somerville, W. J. Peacock, R. Dolferus, and E. S. Dennis. "Expression profile analysis of the low-oxygen response in Arabidopsis root cultures". In: *The Plant Cell Online* 14.10 (2002), pp. 2481–2494.

[KY08] D. Knoke and S. Yang. *Social network analysis*. Vol. 154. Sage, 2008.

[Koh95] R. Kohavi. "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *Fourteenth International Joint Conference on Artificial Intelligence*. Vol. 2. 12. 1995, pp. 1137–1143.

[Kra11] B. Krawczyk. "Classifier committee based on feature selection method for obstructive nephropathy diagnosis". In: *Semantic Methods for Knowledge Management and Communication*. Springer, 2011, pp. 115–125.

[KSHP11] O. Kuchaiev, A. Stevanović, W. Hayes, and N. Pržulj. "GraphCrunch 2: Software tool for network modeling, alignment and clustering". In: *BMC bioinformatics* 12.1 (2011), p. 24.

[KT06] M. Kurant and P. Thiran. "Layered complex networks". In: *Physical review letters* 96.13 (2006), p. 138701.

[LRMV99] J.-P. Lachaux, E. Rodriguez, J. Martinerie, and F. J. Varela. "Measuring phase synchrony in brain signals". In: *Human brain mapping* 8.4 (1999), pp. 194–208.

[LF09] A. Lancichinetti and S. Fortunato. "Community detection algorithms: A comparative analysis". In: *Phys. Rev. E* 80 (2009), p. 056117.

[Lan96] P. Langley. *Elements of machine learning*. Morgan Kaufmann, 1996.

[LM01] V. Latora and M. Marchiori. "Efficient behavior of small-world networks". In: *Phys. Rev. Lett.* 87 (2001), p. 198701.

[LE01] I. R. Lewis and H. G. M. Edwards. *Handbook of Raman spectroscopy: from the research laboratory to the process line*. CRC, 2001.

[LZO04] T. Li, C. Zhang, and M. Ogihara. "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression". In: *Bioinformatics* 20.15 (2004), pp. 2429–2437.

[LCJM04] B. Liu, Q. Cui, T. Jiang, and S. Ma. "A combinational feature selection and ensemble neural network method for classification of gene expression data". In: *BMC bioinformatics* 5.1 (2004), p. 136.

[LM98] H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Springer, 1998.

[LM07] H. Liu and H. Motoda. *Computational Methods of Feature Selection*. Chapman & Hall, 2007.

[LLW02] H. Liu, J. Li, and L. Wong. "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns". In: *Genome Informatics Series* (2002), pp. 51–60.

[LEGSS79] A. Lobo, J. Ezquerra, B. F. Gómez, J. M. Sala, and D. A. Seva. "Cognocitive mini-test (a simple practical test to detect intellectual changes in medical patients)." In: *Actas luso-españolas de neurología, psiquiatría y ciencias afines* 7.3 (1979), p. 189.

[Lus03] D. Lusseau. "The emergent properties of a dolphin social network". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270.Suppl 2 (2003), S186–S188.

[MA03] S. Mangan and U. Alon. "Structure and function of the feed-forward loop network motif". In: *Proc. Nat. Acad. Sci. USA* 100 (2003), 11980–11985.

[MVHDD09] L. Mao, J. L. Van Hemert, S. Dash, and J. A. Dickerson. "Arabidopsis gene co-expression network and its functional modules". In: *BMC bioinformatics* 10.1 (2009), p. 346.

[MS12] D. Marcus and Y. Shavitt. "RAGE–A rapid graphlet enumerator for large networks". In: *Computer Networks* 56.2 (2012), pp. 810–819.

[MGTG75] G. Mayor, N. Genton, A. Torrado, and J. P. Guignard. "Renal Function in Obstructive Nephropathy: Long-Term Effect of Reconstructive Surgery". In: *PEDIATRICS* 56.5 (1975), pp. 740–747.

[MB04] C. J. Melián and J. Bascompte. "Food web cohesion". In: *Ecology* 85.2 (2004), pp. 352–358.

[MLB10] D. Meunier, R. Lambiotte, and E. T. Bullmore. "Modular and hierarchically modular organization of brain networks". In: *Frontiers in neuroscience* 4 (2010).

[MVM10] F. P. Miller, A. F. Vandome, and J. McBrewster. *Amazon Web Services*. Alpha Press, 2010.

[MSOIKCA02] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. "Network motifs: simple building blocks of complex networks". In: *Science* 298 (2002), pp. 824–827.

[Mit08] A. J. Mitchell. "Is it time to separate subjective cognitive complaints from the diagnosis of mild cognitive impairment?" In: *Age and ageing* 37.5 (2008), pp. 497–499.

[MLDEE00] F. Mormann, K. Lehnertz, P. David, and C. E Elger. "Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients". In: *Physica D: Nonlinear Phenomena* 144.3 (2000), pp. 358–369.

[MVKMSC11] P. Moulos, I. Valavanis, J. Klein, I. Maglogiannis, J. Schanstra, and A. Chatziioannou. "Unifying the integration, analysis and interpretation of multi-omic datasets: Exploration of the disease networks of Obstructive Nephropathy in children". In: *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE. 2011, pp. 3716–3719.

[MRMPO10] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela. "Community structure in time-dependent, multiscale, and multiplex networks". In: *Science* 328.5980 (2010), pp. 876–878.

[MS62] T. Murashige and F. Skoog. "A revised medium for rapid growth and bio assays with tobacco tissue cultures". In: *Physiologia plantarum* 15.3 (1962), pp. 473–497.

[Mur98] S. K. Murthy. "Automatic construction of decision trees from data: A multidisciplinary survey". In: *Automatic construction of decision trees from data: A multidisciplinary survey* 2.4 (1998), pp. 345–389.

[New01] M. E. J. Newman. "Scientific collaboration networks: I. Network construction and fundamental results". In: *Physical Review E* 64 (2001).

[New02] M. E. J. Newman. "Assortative mixing in networks". In: *Phys. Rev. Lett.* 89.20 (2002).

[New03] M. E. J. Newman. "The Structure and Function of Complex Networks". In: *SIAM Review* 45 (2003).

[NG04] M. E. J. Newman and M. Girvan. "Finding and evaluating community structure in networks". In: *Phys. Rev. E* 69 (2004), p. 026113.

[New06] M. E. J. Newman. "Modularity and community structure in networks". In: *Proceedings of the National Academy of Sciences* 103.23 (2006), pp. 8577–8582.

[OAS10] T. Opsahl, F. Agneessens, and J. Skvoretz. "Node centrality in weighted networks: Generalizing degree and shortest paths". In: *Social Networks* 32.3 (2010), pp. 245–251.

[PSV01] R. Pastor-Satorras and A. Vespignani. "Epidemic Spreading in Scale-Free Networks". In: *Phys. Rev. Lett.* 86.14 (2001), pp. 3200–3203.

[PLD05] H. Peng, F. Long, and C. Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 27.8 (2005), pp. 1226–1238.

[Per07] O. Perron. "Zur Theorie der Matrices". In: *Mathematische Annalen* 64.2 (1907), pp. 248–263.

[Pet04] R. C. Petersen. "Mild cognitive impairment as a diagnostic entity". In: *Journal of internal medicine* 256.3 (2004), pp. 183–194.

[POPAHHVTWWSLMEBSKL02] E. F. Petricoin, D. K. Ornstein, C. P. Paweletz, A. Ardekani, P. S. Hackett, B. A. Hitt, A. Velassco, C. Trucco, L. Wiegand, K. Wood, C. B. Simone, P. J. Levine, W. Marston Linehan, M. R. Emmert-Buck, S. M. Steinberg, E. C. Kohn, and L. A. Liotta. "Serum proteomic patterns for detection of prostate cancer". In: *Journal of the National Cancer Institute* 94.20 (2002), pp. 1576–1578.

[PIAHLFSMSFKL02] E. F. Petricoin III, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta. "Use of proteomic patterns in serum to identify ovarian cancer". In: *The lancet* 359.9306 (2002), pp. 572–577.

[PLRZ90] A. Pfefferbaum, K. O. Lim, M. Rosenbloom, and R. B. Zipursky. "Brain Magnetic Resonance Imaging". In: *Schizophrenia bulletin* 16.3 (1990), pp. 453–476.

[PS90] G. Piatetsky-Shapiro. "Knowledge discovery in real databases: a report on the IJCAI-89 workshop". In: *AI magazine* 11.4 (1990), pp. 68–70.

[Pow11] D. M. W. Powers. "Evaluation: From Precision, Recall and F-Measure to ROC., Informedness, Markedness & Correlation". In: *Journal of Machine Learning Technologies* 2.1 (2011), pp. 37–63.

[PWASMSPJF04] G. A. Preston, I. Waga, D. A. Alcorta, H. Sasai, W. E. Munger, P. Sullivan, B. Phillips, J. C. Jennette, and R. J. Falk. "Gene expression profiles of circulating leukocytes correlate with renal disease activity in IgA nephropathy". In: *Kidney international* 65.2 (2004), pp. 420–430.

[PKDSSB07] N. G. Psihogios, R. G. Kalaitzidis, S. Dimou, K. I. Seferiadis, K. C. Siamopoulos, and E. T. Bairaktari. "Evaluation of Tubulointerstitial Lesions' Severity in Patients with Glomerulonephritides: An NMR-Based Metabonomic Study". In: *Journal of Proteome Research* 6 (2007), pp. 3760–3770.

[Qui86] J. R. Quinlan. "Induction of decision trees". In: *Machine Learning* 1.1 (1986), pp. 81–106.

[Qui87] J. R. Quinlan. "Generating production rules from decision trees". In: *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*. 1987, pp. 304–307.

[Qui93] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann, 1993.

[QVC98] R. Quivy and L. Van Campenhoudt. "Manual de investigação em ciências sociais". In: (1998).

[RCCLP04] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. "Defining and identifying communities in networks". In: *Proc. Nat. Acad. Sci. USA* 101.9 (2004), pp. 2658–2663.

[RRBF08] F. Radicchi, J. J. Ramasco, A. Barrat, and S. Fortunato. "Complex networks renormalization: Flows and fixed points". In: *Physical review letters* 101.14 (2008), p. 148701.

[Ros62] F. Rosenblatt. "A comparison of several perceptron models". In: *Self-Organizing Systems* (1962), pp. 463–484.

[RSM10] H. D. Rozenfeld, C. Song, and H. A. Makse. "Small-world to fractal transition in complex networks: a renormalization group approach". In: *Physical review letters* 104.2 (2010), p. 025701.

[Ruh01] B. Ruhnau. "Eigenvector-centrality: a node-centrality?" In: *Social Networks* 22.4 (2001), pp. 357–365.

[RHW86] D. E. Rumelhart, G. E. Hintont, and R. J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323.6088 (1986), pp. 533–536.

[SB99] K. Sameshima and L. A. Baccalá. "Using partial directed coherence to describe neuronal ensemble interactions". In: *Journal of neuroscience methods* 94.1 (1999), pp. 93–103.

[SK10] J. Sanders and E. Kandrot. *CUDA by example: an introduction to general-purpose GPU programming*. Addison-Wesley Professional, 2010.

[SS05] F. Schreiber and H. Schwöbbermeyer. "MAVisto: a tool for the exploration of network motifs". In: *Bioinformatics* 21.17 (2005), pp. 3572–3574.

[SNINFOKNESSATYSCKHS02] M. Seki, M. Narusaka, J. Ishida, T. Nanjo, M. Fujita, Y. Oono, A. Kamiya, M. Nakajima, A. Enju, T. Sakurai, M. Satou, K. Akiyama, T. Taji, K. Yamaguchi-Shinozaki, P. Carninci, J. Kawai, Y. Hayashizaki, and K. Shinozaki. "Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray". In: *The Plant Journal* 31.3 (2002), pp. 279–292.

[Sha48] C. E. Shannon. "A Mathematical Theory of Communication". In: *Bell System Technical Journal* 27.3 (1948), 379–423.

[SOMMA02] S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. "Network motifs in the transcriptional regulation network of Escherichia Coli". In: *Nature Genetics* 31 (2002), pp. 64–68.

[Spe90] D. F. Specht. "Probabilistic neural networks". In: *Neural networks* 3.1 (1990), pp. 109–118.

[SK04] O. Sporns and R. Kötter. "Motifs in Brain Networks". In: *PLoS Biology* 2 (2004), e369.

[SDHDJMWMVDMVDBS09] C. J. Stam, W. De Haan, A. Daffertshofer, B. F. Jones, I. Manshanden, A. M. v. C. van Walsum, T. Montez, J. P. A. Verbunt, J. C. De Munck, B. W. Van Dijk, H. W.

Berendse, and P. Scheltens. "Graph theoretical analysis of magnetoencephalographic functional connectivity in Alzheimer's disease". In: *Brain* 132.1 (2009), pp. 213–224.

[SD02] C. J. Stam and B. W. van Dijk. "Synchronization likelihood: an unbiased measure of generalized synchronization in multivariate data sets". In: *Physica D* 163.3 (2002), pp. 236–251.

[SJNBS07] C. J. Stam, B. F. Jones, G. Nolte, M. Breakspear, and P. H. Scheltens. "Small-world networks and functional connectivity in Alzheimer's disease". In: *Cerebral Cortex* 17.1 (2007), pp. 92–99.

[ST60] R. G. D. Steel and J. H. Torrie. *Principles and procedures of statistics.* McGraw-Hill Book Company, 1960.

[Sto74] M. Stone. "Cross-validatory choice and assessment of statistical predictions". In: *Journal of the Royal Statistical Society Series B* 36 (1974), pp. 111–147.

[Sto77] M. Stone. "An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion". In: *Journal of the Royal Statistical Society Series B* 39 (1977), pp. 44–47.

[Str01] S. Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry and engineering*. Perseus Books Group, 2001.

[SBKFFSABM08] R. Sukhija, Z. Bursac, P. Kakar, L. Fink, C. Fort, S. Satwani, W. S. Aronow, D. Bansal, and J. L. Mehta. "Effect of statins on the development of renal dysfunction". In: *The American journal of cardiology* 101.7 (2008), pp. 975–979.

[SBGMPT11] M. G. Summa, L. Bottou, B. Goldfarb, F. Murtagh, C. Pardoux, and M. Touati. *Statistical Learning and Data Science*. Chapman & Hall/CRC, 2011.

[SHB83] A. Sunderland, J. E. Harris, and A. D. Baddeley. "Do laboratory tests predict everyday memory? A neuropsychological study". In: *Journal of verbal learning and verbal behavior* 22.3 (1983), pp. 341–357.

[VCMKSC10] I. Valavanis, C. Caubet, I. Maglogiannis, J. Klein, J. Schanstra, and A. Chatziioannou. "Analysis of pediatric obstructive nephropathy using protein antibody arrays and computational techniques". In: *Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on*. IEEE. 2010, pp. 1–5.

[VMC13] I. Valavanis, I. Maglogiannis, and A. Chatziioannou. "Intelligent identification of biomarkers for the study of obstructive nephropathy". In: *Intelligent Decision Technologies* 7.1 (2013), pp. 11–22.

[VOWBP11] M. Vinck, R. Oostenveld, M. van Wingerden, F. Battaglia, and C. Pennartz. "An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias". In: *Neuroimage* 55.4 (2011), pp. 1548–1565.

[VR01] J. Vrba and S. E. Robinson. "Signal processing in magnetoencephalography". In: *Methods* 25.2 (2001), pp. 249–271.

[WTGX06] B. Wang, H. Tang, C. Guo, and Z. Xiu. "Entropy optimization of scale-free networks robustness to random failures". In: *Physica A* 363 (2006), pp. 591–596.

[WF94] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University, Cambridge, 1994.

[WS98] D. J. Watts and S. H. Strogatz. "Collective dynamics of 'small-world' networks". In: *Nature* 393 (1998), pp. 440–442.

[Wec87] D. Wechsler. *WMS-R: Wechsler Memory Scale-Revised: Manual*. Psychological Corporation San Antonio, 1987.

[WFJD98] J. G. Wen, J. Frkir, T. M. Jrgensen, and J. C. Djurhuus. "Obstructive nephropathy: an update of the experimental research". In: *Urological Research* 27.1 (1998), pp. 29–39.

[WR06] S. Wernicke and F. Rasche. "FANMOD: a tool for fast network motif detection". In: *Bioinformatics* 22.9 (2006), pp. 1152–1153.

[WSTB86] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner. "The structure of the nervous system of the nematode Caenorhabditis elegans". In: *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 314.1165 (1986), pp. 1–340.

[Wiw06] V. Wiwanitkit. "Angiotensin-converting enzyme gene polymorphism is correlated to the progression of disease in patients with IgA nephropathy: A metaanalysis". In: *Renal failure* 28.8 (2006), pp. 697–699.

[WJS98] J. Workman Jr and A. Springsteen. *Applied spectroscopy: a compact reference for practitioners*. Academic Press, 1998.

[WWSWHXY10] D. Wu, T. Wu, Y. Shan, Y. Wang, Y. He, N. Xu, and H. Yang. "Making human connectome faster: GPU acceleration of brain network analysis". In: *Parallel and Distributed Systems (ICPADS), 2010 IEEE 16th International Conference on*. IEEE. 2010, pp. 593–600.

[WL08] X. Wu and Z. Liu. "How community structure influences epidemic spread in social networks". In: *Physica A: Statistical Mechanics and its Applications* 387.2 (2008), pp. 623–630.

[WOB03] S. Wuchty, Z. N. Oltvai, and A.-L. Barabási. "Evolutionary conservation of motif constituents in the yeast protein interaction network". In: *Nature genetics* 35.2 (2003), pp. 176–179.

[XMSBW12] J. Xia, R. Mandal, I. Sinelnikov, D. Broadhurst, and D. S. Wishart. "MetaboAnalyst 2.0 - a comprehensive server for metabolomic data analysis". In: *Nucl. Acids Res* 40 (2012).

[XPYW09] J. Xia, N. Psychogios, N. Young, and D. S. Wishart. "MetaboAnalyst: a web server for metabolomic data analysis and interpretation". In: *Nucl. Acids Res* 37 (2009).

[XJK01] E. P. Xing, M. I. Jordan, and R. M. Karp. "Feature selection for high-dimensional genomic microarray data". In: *Machine Learning-International Workshop Then Conference*. 2001, pp. 601–608.

[YSLL10] C. Yang, N. W. Street, D.-F. Lu, and L. Lanning. "A data mining approach to MPGN type II renal survival analysis". In: *Proceedings of the 1st ACM International Health Informatics Symposium*. ACM. 2010, pp. 454–458.

[YP97] Y. Yang and J. O. Pedersen. "A comparative study on feature selection in text categorization". In: *Machine Learning-International Workshop Then Conference*. Morgan Kaufmann Publishers, Inc. 1997, pp. 412–420.

[YBr91] J. A. Yesavage and J. O. Brooks 3rd. "On the importance of longitudinal research in Alzheimer's disease." In: *Journal of the American Geriatrics Society* 39.9 (1991), p. 942.

[ZB11] M. Zanin and S. Boccaletti. "Complex networks analysis of obstructive nephropathy data". In: *Chaos* 21 (2011), p. 033103.

[ZMSB12] M. Zanin, E. Menasalvas, P. Sousa, and S. Boccaletti. "Preprocessing and analyzing genetic data with complex networks: an application to Obstructive Nephropathy". In: *Networks and Heterogeneous Media* 7 (2012).

[ZL13] M. Zanin and F. Lillo. "Modelling the air transport with complex networks: A short review". In: *The European Physical Journal Special Topics* 215.1 (2013), pp. 5–21.

[ZMVGPSMB14] M. Zanin, J. Medina Alcazar, J. Vicente Carbajosa, M. Gomez Paez, D. Papo, P. Sousa, E. Menasalvas, and S. Boccaletti. "Parenclitic networks: uncovering new functions in biological data". In: *Sci. Rep.* 32.3 (2014), pp. 245–251.

[ZMBS13] M. Zanin, E. Menasalvas, S. Boccaletti, and P. Sousa. "Feature selection in the reconstruction of complex network representations of spectral data". In: *PloS one* 8.8 (2013), e72045.

[ZPSEFRASEJRBMS13] M. Zanin, D. Papo, J. L. G. Solís, J. C. M. Espinosa, C. Frausto-Reyes, P. P. Anda, R. Sevilla-Escoboza, R. Jaimes-Reategui, S. Boccaletti, E. Menasalvas, and P. Sousa. "Knowledge Discovery in Spectral Data by Means of Complex Networks". In: *Metabolites* 3.1 (2013), pp. 155–167.

[ZSPBGPPMB12] M. Zanin, P. Sousa, D. Papo, R. Bajo, J. García-Prieto, F. del Pozo, E. Menasalvas, and S. Boccaletti. "Optimizing functional network representation of multivariate time series". In: *Scientific reports* 2 (2012).

[ZSM14] M. Zanin, P. A. Sousa, and E. Menasalvas. "Information content: assessing meso-scale structures in complex networks". In: *Europhysics Letters* 32.3 (2014), pp. 245–251.

[Zha00] G. P. Zhang. "Neural networks for classification: a survey". In: *Systems, Man, and Cybernetics, Part C* 30.4 (2000), pp. 451–462.

[ZZY03] S. Zhang, C. Zhang, and Q. Yang. "Data preparation for data mining". In: *Applied Artificial Intelligence* 17 (2003), pp. 375–381.

[ZMW05] E. Ziv, M. Middendorf, and C. H. Wiggins. "Information-theoretic approach to network modularity". In: *Phys. Rev. E* 71 (2005), p. 046117.

[ZC93] M. H. Zweig and G. Campbell. "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine." In: *Clinical chemistry* 39.4 (1993), pp. 561–577.

Complex networks topological metrics

In this Annex we present a short review of the main metrics used for complex network analysis. The complete list is included in Table A.1. Specifically, column 1 of the Table reports the name of the indicator, as it can be found in the Literature on network theory, and column 2 indicates the symbol that is commonly used for denoting it; the fifth column reports, when necessary, the Manuscript in the Literature where the full description and mathematical expression of the specific indicator can be found. As for columns 3 and 4 of the Table, they contains a tic in all cases in which values were normalized, or Z-scored.

In some cases, it is known that the value of the metric alone is not enough to understand if the associated topological feature is relevant or not. For instance, the *efficiency* E strongly depends both on the structure of the network, and on its number of links. Therefore, in order to have meaningful comparisons of heterogeneous network, it is necessary to normalize the obtained value against a reference model.

This is accomplished by creating a large ensemble of random Erdös-Renyi (ER) graphs [ER60; Bol01], each one of them with the same number of nodes and links as in the original graph; $n = 500$ is the number of random graphs initially chosen. The metric M under study is then calculated on these graphs, resulting in a vector $R^M = (R_1^M, R_2^M, R_3^M, \ldots, R_n^M)$. The average value for this ensemble, *i.e.*

$$\mu^M = \frac{1}{n} \sum_{i=1}^{n} R_i^M, \tag{A.1}$$

represents the expected value of that metric for the considered network parameters (number of nodes and links). The first normalization is straightforward: if we denote the metric calculated in the real network by m, its *normalized* value is given by:

$$m_{norm} = \frac{m}{\mu^M}. \tag{A.2}$$

m_{norm} is defined in the interval $[0, \infty)$; $m_{norm} < 1$ ($m_{norm} > 1$) indicates that the measured metric value is lower (higher) than what expected in random equivalent graphs.

It has to be noticed that this form of normalization gives little information about the significance of the value, as it only defines if it is higher or lower than expected. In order to solve this issue, it is possible to calculate the *ZScore* of the metric as:

$$m_{ZScore} = \frac{m - \mu^M}{\sigma^M}, \tag{A.3}$$

σ^M being the standard deviation of the vector R^M, *i.e.*

$$\sigma^M = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(R_i^M - \mu^M\right)^2}. \tag{A.4}$$

m_{ZScore} is defined in the interval $(-\infty, \infty)$, with values of $m_{ZScore} < 0$ ($m_{ZScore} > 0$) indicating a measurement lower (higher) than expected. Furthermore, $|m_{ZScore}|$ is the number of standard deviations between the expected and the measured values, or, in other words, how abnormal the measured value is with respect to the distribution of R^M. Therefore, if the result of a measurement is $|m_{ZScore}| < 1$, we can conclude that the obtained value is not significantly different from what expected in random equivalent graphs.

In what follows, the definition of each metric, and its significance in terms of the underlying network structure, is presented. The interested reader may find further information in the numerous reviews published within this field, *e.g.* Refs. [New01; AB02; New03; BLMCH06; CRTV07].

Number of nodes, number of links, and link density

The number of nodes and links are the basic metrics defining the size of the network. If we denote the number of nodes as n, any graph can be described by an *adjacency matrix* $\mathcal{A}$, of size $n \times n$, whose element $a_{i,j} = 1$ if it exists a link connecting nodes i and j, and $a_{i,j} = 0$ otherwise. Notice that this matrix may not be symmetric, *i.e.* $a_{i,j} \neq a_{j,i}$, in that it may exist a connection from node i to node j, but the connection $j \to i$ may be missing.

The number of active links l is simply given by:

$$l = \sum_{i,j} a_{i,j}. \tag{A.5}$$

Table A.1: **List of the network topological features considered in this study.** See the main text for the definition of each column.

Name of the metric	Symbol	Normalized	Z-Score	References
Number of nodes	n			
Number of links	l			
Link density	d			
Maximum degree	k_{max}	$\surd$	$\surd$	
Entropy of the degree distribution	H			[WTGX06]
Energy of the degree distribution	$E(\{N_k\})$			[Bia06]
Random network p-value				[ER60]
Exponential fit slope	γ			[AB02]
Degree correlation	r			[New02]
Clustering coefficient	C	$\surd$	$\surd$	
Mean geodesic distance	L	$\surd$	$\surd$	
Efficiency	E	$\surd$	$\surd$	[LM01]
Small-worldness	S			[HG08]
Number of connected components	n_{cc}	$\surd$	$\surd$	
Size of the giant component	s_{gc}			
Dispersion of component sizes	v_c			
Entropy of eigenvector centrality distribution	E_{ec}			
Algebraic connectivity	C_a	$\surd$	$\surd$	
Motifs of 3 nodes	M_i		$\surd$	[MSOIKCA02]

The *link density* is defined as the proportion of links that are active, with respect to the total number of potential links, *i.e.*

$$d = \frac{\sum_{i,j} a_{i,j}}{n^2} = \frac{l}{n^2}, \tag{A.6}$$

and is therefore defined in the interval $[0, 1]$ (0 and 1 indicating a void and a fully connected network, respectively).

Maximum degree

Based on the degree of the vertices, it is possible to derive many measurements for the network. Beside the degree distribution and correlation (see below), one of the simplest is the maximum degree:

$$k_{max} = \max_i k_i, \tag{A.7}$$

k_i being the degree of node i, *i.e.* $k_i = \sum_j a_{i,j}$.

Degree distribution

Some of the most important topological characterizations of a graph can be obtained in terms of the *degree distribution* $P(k)$, defined as the probability that a randomly chosen node has degree k or, equivalently, as the fraction of nodes in the graph having degree k. This function $P(k)$ defines many important properties of the network, like for instance its resilience to attacks and random failures, and has been used since the beginning of graph theory to classify different types of networks.

In spite of its importance, it has to be noticed that the function $P(k)$ cannot directly be used to feed a data mining algorithm: in other words, it should be transformed into a small set of features capturing its characteristics. One of them is the *entropy* of the degree distribution [WTGX06], *i.e.*:

$$H = -\sum_k p(k) \log_2 p(k). \tag{A.8}$$

H provides a measure of the heterogeneity of the network: the maximum value is obtained for a uniform degree distribution, while the minimum $H = 0$ is achieved whenever all vertices have the same degree. As the degree distribution, its entropy has been related to the robustness of networks, and the contribution of vertices to the network entropy is correlated with lethality in protein interactions networks [DM04].

Another metric that can be extracted from the degree distribution is its *energy*, defined as [Bia06]:

$$E(\{N_k\}) = \log(\mathcal{N}_G). \tag{A.9}$$

Here, $\mathcal{N}_G$ is the number of indistinguishable simple networks it is possible to draw given a degree distribution, that is,

$$N_G = \prod_k k!^{N_k}, \tag{A.10}$$

N_k being the number of nodes having degree k. In other words, in a network with a given degree distribution, every permutation of the links departing from each node is equivalent, and the number of such permutations is given by $\prod_k k!^{N_k}$.

Finally, it is worth noticing that most networks belong to two families: random networks, where the probability of having a pair of nodes connected by a link is independent on the nodes themselves, and in which therefore the degree distribution is described by a Poisson distribution $P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$ [ER60; Bol01]; and *scale-free networks*, in which the degree distribution follows a power-law in the form $P(k) \sim k^{-\gamma}$ [AB02; Bar09]. From the point of view of this work, the interest of both types of distributions resides in the type of structures they represent. While in random graphs all nodes are equivalent, and therefore no node is responsible for the resulting topological properties, scale-free networks are characterized by a single node (or few nodes) of high connectivity. Therefore, in the latter case, it is possible to identify a small set of nodes showing an abnormal behavior.

The real degree distribution can be fitted against the two models, *i.e.* Poisson and scale-free distributions, with the goodness of such fits indicating whether the network belongs to one of these two families. The quality of the fit is calculated by means of its *p-value*, that is, the probability of obtaining a test statistic at least as extreme as the one that was actually observed; *p-values* of less than 0.05 thus indicate that the considered distribution is well represented by the model. In the case of a scale-free network, it is also interesting to obtain the slope of the distribution, *i.e.* the value of γ.

Degree correlation

As we have previously seen, the degree distribution is an important way of assessing the structure of real-world networks. This is especially true in the case of uncorrelated networks, *i.e.* networks in which the properties of one node do not depend on other nodes. Nevertheless, a large number of real networks are correlated, in the sense that the probability for a node of degree k of being connected to another node of degree, say k', depends on k. Real-world networks are classified in two families: *assortative networks*, in which nodes tend to connect to their connectivity peers, and *disassortative networks*, where nodes with low degree are more likely connected with highly connected ones.

Expanding the concept of degree distribution, the assortativity can be represented by a coditional probability $P(k'|k)$, defined as the probability for a link from a node of degree k of pointing to a node of degree k'. Here again, the distribution should be transformed into a feature suitable for data mining tasks: this is usually performed by calculating the Pearson correlation coefficient of the degrees at either ends of a link. The formula for

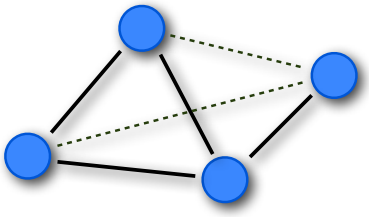


Figure A.1: **Calculation of the clustering coefficient.** Notice that the node in the right side is part of two connected triplets, but forms no triangles, as they would require the links represented by the dashed lines.

calculating the degree correlation r is [New02]:

$$\frac{1}{M}\sum_{j>i}\frac{1}{2}\left(k_i + k_j\right)a_{ij}, \tag{A.11}$$

M being the total number of links in the network.

Clustering coefficient

The *clustering coefficient*, also known as *transitivity*, measures the presence of triangles in the network [New01]. A high clustering coefficient has been historically associated with social networks, where it means that "the friends of my friends are also my friends". Mathematically, it is defined as the relationship between the number of triangles in the network and the number of connected triples:

$$C = \frac{3N_\Delta}{N_3}. \tag{A.12}$$

Here, a triangle is a set of three vertices with edges between each pair of them, while a connected triple is a set of three vertices where each vertex can be reached from each other (directly or indirectly). Both concepts are presented in Fig. A.1: notice that the node on the right side is part of two connected triples, as it is indirectly connected with all nodes, but forms no triangles, which would require the two dashed links. The factor 3 is included to normalize C, as each triangle is equivalent to three different connected triples; this ensures that $0 \leq C \leq 1$. Using the notation of the adjacency matrix, N_Δ and N_3 can be calculated as follows:

$$\begin{aligned} N_\Delta &= \sum_{k>j>i} a_{ij}a_{ik}a_{jk} \\ N_3 &= \sum_{k>j>i} (a_{ij}a_{ik} + a_{ji}a_{jk} + a_{ki}a_{kj}). \end{aligned} \tag{A.13}$$

Mean geodesic distance and efficiency

One important characteristic of many real-world networks, both man-made and natural, is their efficiency in moving goods or information. For instance, both for the Internet

and neural networks, it is essential to send information to the destination in the shortest possible time.

In order to describe this ability, some terms have to be defined. First of all, the number of edges in a path connecting nodes i and j is called the *length* of the path; the shortest possible path connecting i and j is called *geodesic path*, and its associated distance d_{ij} is its *geodesic distance*. The most natural metric is then the *mean geodesic distance*, *i.e.* the average number of links needed to move between two nodes of the network:

$$l = \frac{1}{N\left(N-1\right)} \sum_{i \neq j} d_{ij}. \tag{A.14}$$

This simple definition has one important drawback: when the network is disconnected, it may exist no path connecting two nodes i and j, and therefore $d_{ij} = \infty$ and $l = \infty$. In other words, the mean geodesic path diverges for disconnected networks.

A solution was proposed in Ref. [LM01], called the *efficiency* of a network:

$$E = \frac{1}{N\left(N-1\right)} \sum_{i \neq j} \frac{1}{d_{ij}}. \tag{A.15}$$

Notice that the divergence problem is solved, as $d_{ij} = \infty$ implies $\frac{1}{d_{ij}} = 0$, and thus this pair of nodes does not contribute to the total efficiency. The name of this measure comes from the characteristic it assesses, that is the efficiency of the network in sending information between vertices, assuming that the efficiency for sending information between two vertices i and j is proportional to the reciprocal of their distance.

Small-worldness

At the dawn of the analysis of real-world networks, it was discovered that they possess two specific properties that could not be explained by means of the available models, *i.e.* random graphs and regular lattices. Specifically, most of them show a high clustering coefficient, but a low mean geodesic distance; notice that random graphs (regular lattices) have both low (high) C and l.

In Ref. [HG08] it was proposed the use of the following metrics to capture the degree of *small-worldness* of a network:

$$S = \frac{C/C_{rand}}{l/l_{rand}} = \frac{C_{norm}}{l_{norm}}. \tag{A.16}$$

S is then defined in the interval $[0, \infty)$, with small-world networks having $S \gg 1$.

It should be noticed that this metric is not well defined for disconnected graphs, as $l = \infty$. In order to solve this problem, here we propose the use of the following modification,

$$S^* = \frac{C}{C_{rand}} \frac{E}{E_{rand}} = C_{norm} E_{norm}, \tag{A.17}$$

which uses the fact that $E \approx \frac{1}{l}$.

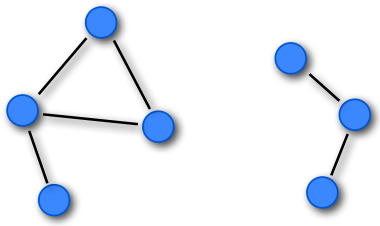


Figure A.2: **Connected components of a graph.** The graph depicted in the image is composed of two components, as it is not possible to reach nodes in the right part starting from nodes on the left.

Components and their characteristics

The *connected components* of a graph are defined as those sets of nodes (or subgraphs) in which any two nodes are connected to each other by at least one path, and which are not connected to any other node of the original graph. Thus, if a graph is not disconnected, there will always be a path between any pair of nodes, and the graph itself will be a single connected component; on the other hand, if the graph is disconnected, there are groups of nodes that cannot be reached - see Fig. A.2 for an example.

Several measurements can be performed on the components structure of a network. For instance, the number of connected components, which will be greater than one in the case of disconnected networks; the size of the *giant component*, *i.e.* of the largest component of the graph; or the dispersion of components size, defined as the relation between the sizes of the largest and smallest components.

Nodes centrality

Centrality is a general term that refers to the importance of a node in the network. Clearly, both in random graphs and regular lattices, each node is essentially equivalent to all other nodes; but when more complicated structures appear, one node may become especially important for the system.

Three classes of centralities can be defined: *(i)* nodes can be important because information passes through them, *(ii)* because they can easily communicate with other members of the network, or *(iii)* because they are themselves connected to other central positions. While the first two depend on the definition of some dynamics on top of the network, *e.g.* information flow, the third one has the advantage of being only topological.

Let us consider with more detail the third type of centrality. If we define a vector $\mathcal{X}$ of centralities, so that x_i is the centrality of the i-th node, we may define the centrality of a node as a linear combination of the centralities of those to whom it is connected:

$$\lambda x_i = \sum_j x_j A_{ij}. \tag{A.18}$$

If we express the previous equation in terms of matrices, we obtain

$$A\mathcal{X} = \lambda\mathcal{X}, \tag{A.19}$$

which is equivalent to an eigenvalue problems, $\mathcal{X}$ being the eigenvector associated to the eigenvalue λ - this is the reason behind the name *eigenvector centrality* [BL01]. In order to have meaningful results, we need all $x_i \geq 0$: for connected graphs, this is guaranteed by the Perron-Frobenium theorem as long as λ is the largest eigenvalue of A [Per07; Fro12].

As for other metrics, the centrality is defined as a spectrum, which has to be converted into a suitable feature for data mining task. As for the degree distribution, here we calculate the entropy of the vector of centralities E_{ec}, indicating the heterogeneity of nodes in the network. Furthermore, the following three metrics are also extracted: *(i)* the relation between the centralities of the two most central nodes; *(ii)* the slope of the centrality distribution, when fitted against a scale-free model, *i.e.* $P(\mathcal{X}) \sim \mathcal{X}^{-\gamma}$; and *(iii)* the *central point dominance*, defined as

$$CPD = \frac{1}{N-1}\sum_i (\max x - x_i), \tag{A.20}$$

where $\max x$ is the largest value of centrality in the network.

The eigenvector centrality, as previous defined, has one main drawback: results are meaningful only when the network is connected, which is the main requirement of the Perron-Frobenium theorem. In order to being able to handle disconnected networks, it is necessary to add weak links between pairs of unconnected nodes: in this way, the network becomes connected, while the centrality vector is not significantly modified. Eqs. A.18 and A.19 are then modified as follows:

$$\lambda x_i = \sum_j x_j(A_{ij} + \alpha); \tag{A.21}$$

$$(A + \alpha)\mathcal{X} = \lambda\mathcal{X}. \tag{A.22}$$

α is a new parameter that controls the strength of the new connections; in order to have meaningful results, its value should be small, specifically smaller than the spectral radius of matrix A. The centrality thus obtained is called *alpha-centrality* [1].

Algebraic connectivity

The *algebraic connectivity* is a metric assessing the modular structure of the network, *i.e.* if the graph is one homogeneous block, or if it is composed of loosely connected groups; notice that a disconnected network, as in Fig. A.2, can be seen an extreme case of a

[1] The *alpha-centrality* is also known as the *PageRank centrality*, as used in the sorting of web pages by Internet search engines [BP98]

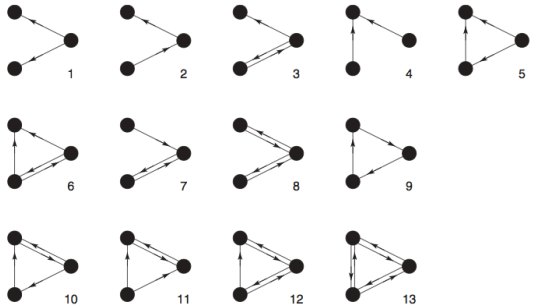


Figure A.3: **3-nodes motifs.** Graphical representation of the 13 motifs that can be obtained by connecting three nodes. Reprinted with permission from Ref.[BLMCH06].

modular structure.

The algebraic connectivity C_a is defined as the second largest eigenvalue of the Laplacian matrix L, calculated as:

$$L = D - A, \tag{A.23}$$

D being the diagonal matrix of vertex degrees, *i.e.* $d_{ii} = \sum_j a_{ij}$, $d_{ij} = 0$ for $i \neq j$.

The smaller C_a, the more modular is the network; if the network is disconnected, *i.e.* it is composed by more than one connected component, $C_a = 0$.

Motifs

A motif M is a pattern of interconnections occurring either in a undirected or in a directed graph at a number significantly higher than in randomized versions of the graph, *i.e.* in graphs with the same number of nodes and links, but where the links are distributed at random [MSOIKCA02]. Fig. A.3 depicts the 13 motifs that can be measured from a 3-nodes subgraph.

The importance of motifs resides in the dynamical processes that they mediate in real-world networks; for instance, in [MSOIKCA02; SOMMA02] it was shown that different networks, *e.g.* transcription networks, neural networks, or electronic circuits) are characterized by the abundance of specific motifs. Section 6.1 presents a novel algorithm for the fast enumeration of motifs in dense networks, which has been developed as part of this PhD Thesis.