bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Combining complex networks and data mining: why and how

M. Zanin[a,b,*], D. Papo[c], P. A. Sousa[b], E. Menasalvas[c], A. Nicchi[d], E. Kubik[e], S. Boccaletti[f]

[a] *Innaxis Foundation & Research Institute, José Ortega y Gasset 20, 28006 Madrid, Spain*
[b] *Departamento de Engenharia Electrotécnica, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal*
[c] *Center for Biomedical Technology, Universidad Politécnica de Madrid, 28223 Pozuelo de Alarcón, Madrid, Spain*
[d] *Università degli Studi Guglielmo Marconi, Centro Studi in Scienze della Vita, Via Plinio 44, 00193 Roma, Italy*
[e] *Columbia University, School of International and Public Affairs, 420 W 118th St #1411, New York, NY 10027, United States*
[f] *CNR- Institute of Complex Systems, Via Madonna del Piano, 10, 50019 Sesto Fiorentino, Florence, Italy*

Abstract

The increasing power of computer technology does not dispense with the need to extract meaningful information out of data sets of ever growing size, and indeed typically exacerbates the complexity of this task. To tackle this general problem, two methods have emerged, at chronologically different times, that are now commonly used in the scientific community: data mining and complex network theory. Not only do complex network analysis and data mining share the same general goal, that of extracting information from complex systems to ultimately create a new compact quantifiable representation, but they also often address similar problems too. In the face of that, a surprisingly low number of researchers turn out to resort to both methodologies. One may then be tempted to conclude that these two fields are either largely redundant or totally antithetic. The starting point of this review is that this state of affairs should be put down to contingent rather than conceptual differences, and that these two fields can in fact advantageously be used in a synergistic manner. An overview of both fields is first provided, some fundamental concepts of which are illustrated. A variety of contexts in which complex network theory and data mining have been used in a synergistic manner are then presented. Contexts in which the appropriate integration of complex network metrics can lead to improved classification rates with respect to classical data mining algorithms and, conversely, contexts in which data mining can be used to tackle important issues in complex network theory applications are illustrated. Finally, ways to achieve a tighter integration between complex networks and data mining, and open lines of research are discussed.

Keywords: Complex networks; data mining; Big Data
2010 MSC: 05C82, 62-07, 92C42

Contents

1 Introduction **2**

2 Some basic concepts **4**
2.1 Complex networks 4
2.2 Data Mining 9
2.3 What is "Big Data"? 18
2.4 The limits of data mining in physics 19
2.5 Hands-on: analysing EEG data 21

*Corresponding author
Email address: mzanin@innaxis.org (M. Zanin)

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3 Classification **21**
3.1 Statistics vs. classification . . . 22
3.2 Are network metrics really necessary? . . . 23
3.3 Classifying with complex networks: Neuroscience . . . 24
3.4 Classifying with complex networks: Brain-Computer Interfaces . . . 26
3.5 Classifying with complex networks: -omics data . . . 27
3.6 Hands-on: classification . . . 28

4 Improving network significance **30**
4.1 Selecting the synchronisation metric . . . 30
4.2 Selecting the threshold . . . 31
4.3 Selecting the topological metrics . . . 32
4.4 Hands-on: improving significance . . . 32

5 Feature selection **34**
5.1 Selecting the nodes of a network . . . 34
5.2 Selecting the links of a network . . . 36
5.3 Feature selection through complex networks . . . 39
5.4 Hands-on: feature selection . . . 39

6 Additional bridges **40**
6.1 Clustering and community detection . . . 40
6.2 Link prediction . . . 41
6.3 Evaluating and validating generative models . . . 43
6.4 Complex networks in Big Data: semantic representations by graphs . . . 43

7 Conclusions **45**
7.1 Problems to overcome . . . 46
7.2 Open lines of research . . . 46

Appendix A List of Acronyms **58**

1. Introduction

As much as they may sometimes seem chaotic or random, large-scale natural and man-made phenomena show characteristic complex structure. Parts or components are related to each other in a non-random way, with identifiable regularities and recurring patterns, ultimately making for internally coherent entities. The ability to extract regularities of all sorts from the environment is already functional within the first few months of humans' life. In fact, the ability to recognise and use patterns is likely the gateway to some important function acquisition such as that of language, and appears to be key to the adaptive capabilities of living organisms [1, 2].

Scientific knowledge can be seen as an endeavour to extract and quantify reproducible regularities of phenomena. Up until relatively recently, save for a few notable *Gedankenexperiment* examples in physics, science generally progressed through similar inductive rule extraction cycles of carefully sampled phenomena, then reordered in a compact, *e.g.* mathematical, formulation. However powerful and general, the classical theory-based approach encounters limits more often than scientists were initially ready to admit. Some systems simply turn out to be too complex to be modelled in a synthetic form. Furthermore, the relevant mechanisms may be hidden. In some other cases, the seemingly intuitive notion of system may be of little help in identifying the underlying structure of a constellation of separate entities.

While methods to extract patterns from data have a long history – Bayes' theorem (1700s) and regression analysis (1800s) arguably belong to this tradition – over the past few decades, the increasing power of computer technology has dramatically increased data collection, storage, and manipulation ability. While

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

uncovering more information, ever more precise instruments with constantly improved sampling capabilities do not yield compact representations of the entities that are being studied, and therefore do not dispense with the need to extract and model regularities, patterns or rules from the data that these instruments generate. Often, quite the contrary, in fact, as the multiplication of samples makes up for overwhelming and intricate jumbles of always increasing size.

As data sets have grown in size and complexity, direct data analysis has started been augmented with indirect, automated data processing methods including neural networks, cluster analysis, genetic algorithms (1950s), decision trees and decision rules (1960s) and, more recently support vector machines (1990s) [3, 4]. *Data mining* was born: an interdisciplinary subfield of computer science applying these methods with the intention of uncovering hidden patterns in large data sets, and transforming them into understandable structures for further use. Data mining bridges the gap from applied statistics and artificial intelligence to database management, to execute learning and discovery algorithms more efficiently, allowing such methods to be applied to ever larger data sets. When data sets are so complex that managing their size is a challenge in itself, data mining is also called *Big Data* [5].

Over the last two decades, *complex network theory* has emerged as a new way to understand the structure of relationships characterising complex systems. A statistical physics understanding of graph theory (a much older branch of pure mathematics), complex networks have been used by scientists from very different fields to describe systems ranging from power grids to social interactions and the brain [6, 7, 8, 9, 10].

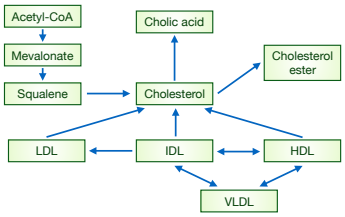
Complex network analysis and data mining have similar goals: given some data, usually representing a complex system, the aim is to extract (or synthesise) some information from them, and create a new representation (either a complex network, or a data mining model) that can be used in successive steps of the analysis. Likewise, many of the problems encountered in both fields are similar, although bearing different names: for instance, selecting the most relevant nodes in a network is mirrored by the feature selection problem; evaluating differences between groups of networks in classification tasks; and so forth. At the same time, complex network theory and data mining also present some important differences. Networks aim at *describing* the relations between the elements of a single (complex) system, by means of patterns that have a statistical nature and that are free from any topological space. On the other hand, data mining characterises the relations between the observables associated to different instances (not of a single system), by means of patterns pinned to a set of well-identified elements/features that are used to *explain* reality.

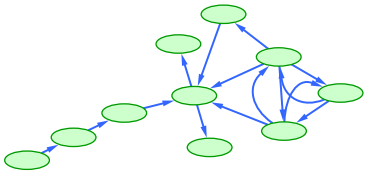
Given the similarity in the general purpose and sometimes even in procedures, it may be natural to think that both approaches can be used interchangeably. Nevertheless, the differences between network theory and data mining may, in some situations, provide an added value when both of them are used in combination. The question is then: why and when would one want to resort to both data mining and complex networks? In what kind of problems do the differences between both of them result in a synergic effect? Note that the primary goal of this review is not to make network-minded individuals fluent in the language of data mining or vice versa, but to convince either side of this field divide that there is reason enough to resort to both.

Some examples of such synergy are easy to identify. Probably the clearest ones can be drawn from biology and biomedicine, two fields in which it is simple to identify different conditions (*e.g.* control subjects and patients suffering from some disease) that have to be diagnosed and prognosed. Data mining is well suited for constructing models allowing the evaluation of the health condition of a person, for instance by detecting a pattern between two or more genes that is symptomatic of a disease; yet, it requires the target elements and the corresponding relations to be regular across instances - in other words, the same pattern, but involving other genes, would not be correctly identified. Complex network theory may solve the problem, by identifying target structures even if they vary across subjects (provided they are topological invariant). Similar situations can be found in neuroscience, where complex networks can be used to describe the structure of relationships between different regions of the brain, and data mining can be used to create diagnostic models. This is not limited to biology: many interdisciplinary fields face similar problems, as for instance econophysics, where networks may represent the economical structure of countries, which can then be modelled through data mining techniques.

This review aims at creating bridges between two communities that have so far had little contact and cross-fertilisation, mainly due to contingent and background differences: Section 2 thus starts by providing an overview of both fields, laying down some fundamental concepts, and providing references to allow the

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.





Metric	Value
Number of nodes	10
Average degree	2.4
Clustering coefficient	0.243
Characteristic path length	1.829

Figure 1: Creating a complex network representation. (Left) Simplified cholesterol pathway. (Center) Network representation of the same pathway. (Right) Some topological metrics of the network.

interested reader to dig deeper into them. Next, in Section 3 we focus on one of the most important tasks in data mining, *i.e.* classification, to understand how complex networks can be used to provide a richer representation of some complex systems. Section 4 changes the point of view, by dealing with how data mining can be used to tackle one of the most important problems in the application of complex network theory, that is, the reconstruction of meaningful functional networks. This latter section is complemented by Section 5, in which we discuss the potential applications of *feature selection* in network analysis. Finally, Section 6 discusses some additional problems and scenarios in which complex network theory and data mining have been (or could be) jointly used. The review is closed by Section 7, in which conclusions are drawn about two important topics: what are the problems to overcome for a tighter integration of complex networks and data mining, and open lines of research. Additionally, hands-on sections will illustrate with worked examples the main points discussed in each chapter.

2. Some basic concepts

This Section is intended to be a primer of several concepts that will be used throughout this review: what complex networks are, how to quantify their structure, up to a first definition of the most important data mining algorithms and concepts. Over the past twenty years, both complex network theory and data mining have undergone changes that are far too dramatic and extensive to be discussed in detail here. What we propose is a first introduction to both topics: some general ideas and concepts of data mining, for the complex network practitioner that has never worked with them; and conversely some basic notions of complex networks for data mining specialists. Throughout this review, we refer the interested reader to excellent reviews and books in both fields. Readers familiar with these concepts can skip the present section, and go straight to Section 3.

2.1. Complex networks

Network theory is a statistical-mechanics understanding of a past branch of pure mathematics: graph theory. In a network representation of a system, all unnecessary details are removed, to extract only its constituent parts and the interactions between them. Graph theory allows characterising a system once its *boundaries*, *constituent parts* and *relationships* are defined, but is indifferent as to what should be treated as a system and how to isolate its constituent parts; a challenge that will be addressed in Section 4. The structure created by such interactions is called the network topology. Most social, biological, and technological networks (including the brain) display substantial non-trivial topological properties, *i.e.* patterns of connections between their elements are neither purely regular nor purely random [11]. These properties can be thought of as features describing the network's structure.

The process of evolving from a complex system to a complex network is depicted in Fig. 1. The left panel represents a simplified version of the *statin pathway* focused around cholesterol. This shows the set of interactions occurring within a human cell that is responsible for many cardiovascular diseases [12]. When the nature of the elements (molecules, metabolites, *etc.*) and of the interactions is disregarded, the result is a network, as shown in the central panel. Finally, the topological structure of the network can be represented as a set of *topological metrics*, as reported in Fig. 1 Right.

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1: The scales of complex networks

Micro-scale	This scale focuses on the properties of a single link or single node. When the whole network is analysed these metrics are usually averaged over each node comprised within the network, in order to create a global picture. *Example: Number of connections a person has in a social network.*
Meso-scale	This is the intermediate level between studying elements of a system separately (micro-) and considering the system as a whole (macro-scale). We may define the meso-scale as any regular structure that affects the connectivity of groups of nodes in the network [14]. *Example: Communities of friends present within a social network.*
Macro-scale	The macro-scale depicts the system as a whole. It accounts for the overall structure of the network, addressing the movement of information throughout. *Example: The average path length to reach a person in a social network.*

Roughly speaking, complex network measures can be divided into three classes, according to the relevant *scale*: micro- (single nodes or links), meso- (groups of few nodes) and macro-scale (the network as a whole). See Table 1 for further examination. In the last decade, tens of different metrics have been proposed for each of these classes and attempts to describe different aspects of the structure. However, some overlap and present different characteristics, like different computational costs and sensitivity to noise.

A few of these metrics will be used throughout this text and are reported in Table 2. Although a complete review of existing complex network metrics is outside the scope of this review, the interested reader can refer to some of the excellent analyses provided in the available literature, such as: Refs. [13, 8, 11, 10].

For the sake of completeness, we will describe three important topics in complex network analysis, which repeatedly appear throughout in this review: *i*) the existence of different *classes* of networks, *ii*) the difference between *structural* and *functional* networks, and *iii*) some recent trends in network analysis.

Classes of networks

In the last decade the analysis of a large number of real-world networks has revealed that some structures, or networks topologies, are ubiquitous across many natural and man-made systems [9]. Such classes are of relevance, as they constitute the basis of the classification of real systems.

From a historical point of view, the start of classifications was the random graphs, also known as Erdös-Rényi graphs after the mathematician who discovered them, Paul Erdös. 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 [20]. A comprehensive review of all results obtained in random graph analysis can be found in Ref. [21]. If random graphs are characterised 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. While these two extrema have limited interest for modelling real-world systems, which seldom present such simple structures, the importance reside in their use as *null models*. 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 [17].

In 1998, Watts and Strogatz realised that real networks were neither regular nor completely random graphs; that real systems lie somewhere between these two extremes. Specifically, random graphs are characterised by both a low mean geodesic distance and clustering coefficient. On the other hand, regular graphs show both high mean geodesic distance and clustering. By analysing social and biological networks, they discovered that most of them are characterised by a low mean geodesic distance, but also by a high clustering coefficient: in this case, networks are said to have the *small-world property* [22]. A third class of networks was born: the small-world ones. Simple models can be designed to synthetically recover such effect

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2: Some basic complex network metrics

Metric	**Scale**	**Description**
Degree	Micro-scale	The number of connections a node has, *i.e.* the number of neighbours. The degree k of node i is calculated from the adjacency matrix as $k_i = \sum_j a_{ij}$.
Link density	Micro-scale	The number of links l in the network, divided by the maximum number of links that could be present. For a network composed of n nodes, the link density is thus $\frac{l}{n(n-1)} = \frac{1}{n(n-1)} \sum_{i,j} a_{ij}$.
Clustering coefficient	Meso-scale	The clustering coefficient C_i of a node i is the fraction of pairs of its neighbours that are directly connected, and is thus a count of the number of triangles in the network. Such metric can be averaged over all nodes, yielding the network clustering coefficient $C = \frac{1}{n} \sum_i C_i$.
Modularity	Meso-scale	Assesses the presence of a community structure in the network, *i.e.* of groups of nodes densely connected between them [15, 16].
Motifs	Meso-scale	Subgraphs (usually composed of three or four nodes) that appear more frequently than what could be statistically expected [17].
Geodesic distance	Macro-scale	The geodesic distance $d_{i,j}$ between nodes i and j is the minimum number of steps needed to move between them, *i.e.* the length of the shortest path connecting them. The value can be averaged over all possible pairs of nodes, yielding the average shortest path length of the network.
Efficiency	Macro-scale	Mean value of the inverse of the geodesic distance between pairs of nodes, *i.e.* $E = \frac{1}{n(n-1)} \sum_{i,j \neq i} \frac{1}{d_{i,j}}$. It assesses the ease of information flow in the network, and has the advantage of being defined for disconnected networks [18, 19].

[22], and the property can be numerically estimated through the *small-worldness* metric [23, 24], which has been particularly popular in neuroscience, though some important caveats to its applications have recently been pointed out [25].

Until 1998, three classes were thus known: random, regular, and small-world ones. A fourth class of network topologies emerged when it was first noticed that real-world networks are characterised by heterogeneous nodes, some of them (the *hubs*) of high connectivity. A clear example can be found in transportation networks, where few airports create the backbone of the network [26]. Mathematically, these networks are called *scale-free*, as their degree distribution follows a power law, and thus have no characteristic scale [27]. Many models for obtaining scale-free topologies have been proposed [28, 29, 30, 31, 32, 33, 34, 35, 36, 37], the most praised is the Barabási-Albert one [38]. It refers to a *preferential attachment* process, in which new nodes are added to the network, each connected to existing nodes with a probability proportional to their degree.

Structural vs. functional networks

When analysing networks, it is useful to make a distinction between two types: *structural* (also called *physical*) and *functional* networks.

Physical networks are described by explicit interactions; reconstructing them is tantamount to mapping the connections into the corresponding links, as seen in the real system. For instance, in a transportation system nodes can be airports, bus stops, train stations, *etc*, while scheduling tables provide information on the corresponding links. However, some real-world systems lack explicit information about links. One may nevertheless suppose that each element's dynamics is a function of its neighbours. The first step

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

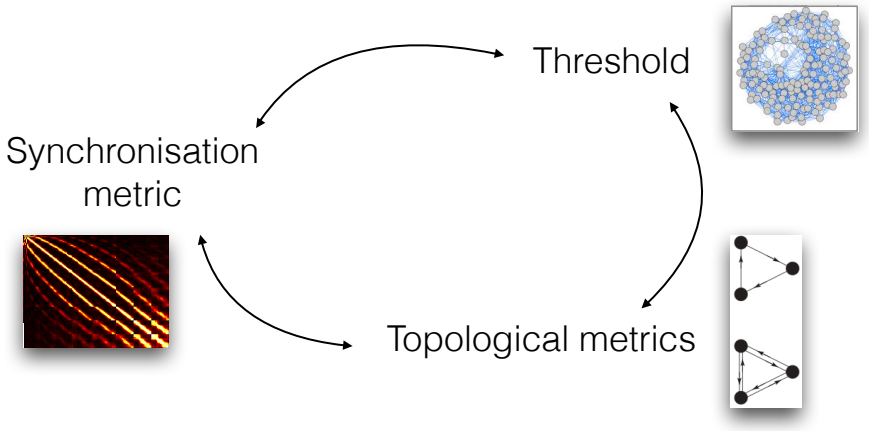


Figure 2: Interactions between different aspects of functional network reconstruction: selection of the synchronisation metric, of the threshold for its binarisation, and of the topological metrics. See the main text for details.

then requires quantifying these functions (hence the name *functional networks*). The resulting network's topological features can finally be extracted as in the previous case.

The study of brain structure and dynamics provides a clear example of this difference. On the one hand, different brain regions are connected by fibres and are organised in a *physical structure*, called the *connectome* [39, 40], which represents the substrate on which brain dynamics takes place. However, on the other hand, one may be interested in understanding how the brain executes a specific task, and thus on how the connectome is used to transfer information between different regions. This can be achieved by reconstructing the functional network associated to the task, by connecting pairs of nodes (brain regions) when some common dynamics (or *synchronisation*) is observed [41, 42].

In the most general case, reconstructing functional networks involves choosing (or tuning) three different aspects of the analysis, namely:

1. *The connectivity metric.* This is the measure assessing the presence of some relationship between the dynamics of two elements of the system.
2. *A threshold.* While not essential, it is customary to eliminate those links that have a low statistical significance, and to consider the binarised version of the network, *i.e.* to discard weights.
3. *The topological metrics.* These metrics describe the structure of the resulting networks and eventually compare different groups of subjects.

In the absence of a set of rules, these three aspects bring in a degree of subjectiveness. For instance, researchers can be tempted to arbitrarily choose a threshold only based on their experience. More importantly, these three steps are strongly interconnected, as depicted in Fig. 2. Each connectivity metric is designed to detect different aspects of dynamics. In the case of functional brain networks, this may include local *vs.* global dynamics, linear and non-linear dependencies, correlations *vs.* causalities, and so forth. Each one of these metrics, in turns, changes the resulting topology, thus requiring different thresholds and topological metrics. Changing the threshold has important implications on the topological metrics that can be assessed. For instance, in dense networks the average geodesic distance always tends to one, while in sparse networks it is difficult to observe complex motifs. Closing the circle, if one wants to observe a given topological metric, it is necessary to choose a suitable synchronisation metric and threshold. For example, motifs containing bi-directional connections are mostly excluded by using a causality metric. These problems will be tackled in Section 4, which is devoted to the use of data mining methods for improving network representations.

While the analysis of physical networks is usually straightforward, it may still require some pre-processing. This pre-processing can include filtering out spurious links (links that are the result of noisy measurements, and that therefore distort the real structure of the network) or reconstructing links in unexplored regions of the system space. These two topics are respectively discussed in Sections 5 and 6.2.

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Recent trends in network theory

Until now we have concisely presented the main elements of complex network theory that are used in the analysis of real systems. Due to the inability of this standard framework to describe scenarios observed in complex real-world scenarios, network theory has been extended within the last years to include concepts such as *temporal* and *multi-layer* networks.

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. Sometimes, edges are active for non-negligible periods of time. This is for instance the case of inpatients proximity patterns within a hospital, where individuals are pairwise connected while they are at the same ward. It is clear that the temporal structure of edge activations can affect the dynamics of the elements interacting through the network, from disease contagion in the patient network to information diffusion over an e-mail network. See [43, 44] for reviews on the topic.

The idea of *multi-layer networks* emerged by observing that connections between the elements constituting real-world networks are seldom of a single type. Let us consider the following three examples, starting with social networks. Social networks as people (or groups of people) have some pattern of contacts or interactions between them [45]. Social interactions seldom develop on a single channel and more than one relationship can bind pairs of people, as was initially proposed by Erving Goffman in 1974, along with his theory of frame analysis [46]. Second, consider a transportation network, as for instance the Air Transportation Network (ATN) [26]. Flights connecting pairs of cities are not homogeneous, as airlines have to be considered. Furthermore, passengers cannot easily connect two flights operated by different airlines, at least connect airlines belonging to different alliances [47, 48]. Finally, biology provides a third example as scientists try to understand the role of specific components in a biological system. The full mapping of the *Caenorhabditis elegans* (or *C. elegans*) neural network is now known, comprising 281 neurons along with more than two thousand connections [49]. Since neurons can be connected by chemical and electrical (ionic) links, this results in two different dynamics, and ultimately does not form a single network.

These three examples explain the efforts for generalising 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 [50, 51, 10]. Multi-layer networks explicitly incorporate multiple channels of connectivity and therefore constitute the natural environment to describe systems interconnected through different categories of connections. A layer represents each channel (relationship, activity, category) and the same node may have different kinds of interactions (different set of neighbours in each layer). For example, in social networks, one can consider several types of relationships: friendship, vicinity, kinship, membership of the same cultural society, partnership or coworker-ship, *etc.*

Software for network analysis

NetworkX [52]. Python library used for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. This allows the creation of networks with different algorithms, evaluation of a large set of standard metrics, and finally display the results in an easily understood way. Freeware. Available at `https://networkx.github.io`.

Cytoscape [53]. Software specialised on the representation of networks, with some additional tools for the integration of biological data. It also provides some basic network analysis capabilities. Freeware. Available at `http://www.cytoscape.org`

Gephi [54]. Interactive visualisation and exploration platform. Freeware. Available at `https://gephi.github.io`

Pajek [55]. Software for representing complex networks, with some basic analysis capabilities. Freeware. Available at `http://mrvar.fdv.uni-lj.si/pajek/`

VisANT [56]. Software for the visual study of metabolic networks and pathways. Freeware. Available at `http://visant.bu.edu`

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

IBM ® i2 Analyst's Notebook. Software for the integration of social data and network analysis. Commercial. Information at `http://www-03.ibm.com/software/products/en/analysts-notebook`

SAS ® Social Network Analysis. Software for the analysis of social networks. Commercial. Information at `http://support.sas.com/software/products/sna/index.html`

2.2. Data Mining

The proliferation, ubiquity and increasing power of computer technology have dramatically enhanced our ability for data collection, storage, and manipulation. This has created a new need for automatic data analysis, classification, and understanding. *Data mining*, with its vision of automatically extracting knowledge from data bases, has attracted growing attention since late 80's. In the current information age, data generated and stored by modern organisations increase in an extraordinary way, and data mining tasks [57] become a necessary and fundamental technology for sustainability and improvement.

The term has also evolved in [57]. A distinction was made between KDD (*Knowledge Discovery in Databases*), which refers to the overall process of discovering useful knowledge from data; and data mining, which refers to a particular step in such process, in which specific algorithms are applied to extract patterns from data. Nowadays, this distinction is almost lost, and "data mining" is used to refer to the overall process performed by combining methodologies and techniques from different fields, such as statistics, databases, machine learning and visualisation. The size of the data increases and becomes challenging and data mining evolves into *Big Data*; a topic that will be further discussed in Section 2.3.

Modern data mining faces two major challenges. The first is being able to share the obtained knowledge. In this sense, the Predictive Model Markup Language (PMML) [58] has become one of the most widely adopted data mining standards used today. The second challenge is defining a methodology to guide the process of discovery. In 1997, an industry group called the *Cross-Industry Standard Process for Data Mining* (CRISP-DM) [59] proposed a methodology for organising the KDD process in six standard steps:

Business (or Problem) Understanding: initial phase that focuses on understanding the project objectives and requirements from a business perspective. This knowledge is then converted into a data mining problem, and a preliminary plan is designed to achieve the project objectives.

Data Understanding: starts with an initial data collection and proceeds with several activities, all aimed at: familiarise with the data; identify data quality problems; discover initial insights about the data; and detect interesting data subsets to form hypotheses about hidden information.

Data preparation: arguably 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 time and resources. This covers all of the activities required to construct the final dataset, such as identifying data that will be fed into the modelling tools from the initial raw data along with data cleaning and preprocessing. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. They include selection and transformation of tables, records, attributes, and data cleaning. Ref. [60] presents several data preparation techniques and algorithms, which can be used for preprocessing data in web mining applications. Additionally, Ref. [61] presents a more general review of techniques, along with a deep discussion on the motivations for data preparation. It is worth noting that the importance of this step is reduced when one is in the Big Data field; a concept that will be further discussed in Section 2.3. Due to the complexity and volume of data, cleaning and preparation may not be feasible and it could lead to an unwanted loss of useful information. In these cases, it is preferable to work with "data in the wild", and sacrifice some accuracy.

Modelling: phase in which data mining algorithms are applied and parameters are calibrated to optimal values. Some of these techniques will be reviewed in detail in the following subsections. Typically, different techniques can be chosen to solve the same modelling problem, each having specific requirements on the format of input data and hypothesis on the patterns to be detected. In these situations, models are optimised and compared; the models reaching a higher performance are passed to the next phase for a final evaluation.

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Evaluation. Once the models have been evaluated from a data mining perspective, it is necessary to review the output patterns obtained considering the business requirements identified in the first phase. Only when all relevant questions have been addressed, can one then move to the deployment of the extracted knowledge.

Deployment. When all of the information about the business problems has been gathered, the information and knowledge then has to be organised and presented.

Data Mining tasks and models

Data mining tasks, or the tasks performed in the modelling phase of the KDD process, can be classified into *predictive* and *descriptive* tasks. Predictive analytics includes all data analysis processes that perform inference in order to extract patterns to be used to make predictions. On the other hand, descriptive tasks group all the processes characterising the general properties of the data. For any of these tasks, a tremendous amount of techniques and algorithms have been described in the literature, the most important of which are reviewed in the next subsections. Delving deeper into such concepts, both families can be defined as follows:

Descriptive Modelling, also called *Exploratory Analysis*, has the main purpose of describing, through patterns, the information encoded in the data set being studied. These techniques have an exploratory nature, in the sense that allows a better characterisation of the existing data, without providing any forecast. They are oriented towards data interpretation, which focuses on understanding the underlying data relations, such as finding correlations, trends, groups, clusters and anomalies. Descriptive models mostly rely on *Clustering* and *Association Rules* techniques and algorithms.

Predictive Modelling: its main goal is to find a model, constructed over the information already labelled in the data set, which can be used in the future to predict information. This model aims at predicting the value of a particular attribute, such as the target or dependent variable, based on the values of other attributes (variables), assuming a set of labelled data (training set) is given. The underlying assumption of the inductive approach is that the data used for training are representative of the whole universe, meaning of all the possible unknown data that may be encountered, and therefore the trained model is able to accurately predict the values of future unseen instances.

It is important to mention that in the case of predictive and descriptive tasks, the methods used to extract knowledge are based on inductive learning, where a model is constructed by generalising a set of training records. In both cases, though, knowledge description comes in different forms. For classification problems, this may be a set of characteristic or discriminant rules, or a decision tree or a neural network with fixed sets of weights. In case of association analyses, it may be a set of associations, or association rules. For cluster analysis, it consists of a set of clusters, each of which has its own description and cluster name (class name).

In the following we will review the main data mining tasks, with a special emphasis will be placed on classification, due to its relevance in problems involving complex networks.

Clustering. Also called *unsupervised learning*, it involves the division of data into groups (clusters) that contain similar records (according to some similarity measures), and those subsequent, separate dissimilar records organised into different clusters. In Ref. [62] clustering is defined as: *partition a given data set in groups, called clusters, so that the points belonging to a cluster are more similar to each other than the rest of the items belonging to other clusters.* In Ref. [63] a taxonomy of clustering techniques is presented, with an overview of its fundamental concepts and methods. Moreover, it describes several successful applications of clustering, such as image segmentation or object and character recognition. Classifying clustering algorithms is not an easy task, as the categories very often overlap. According to the survey found in Ref. [64], one can establish the following classification:

- *Hierarchical-Based.* The hierarchical clustering combines instances of the dataset, to form successive clusters, resulting in a tree form called *dendrogram.* In the lower level of the tree, only one

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

instance is associated to each cluster, and upper levels of the tree are aggregations of the nodes below. Agglomerative and divisive clustering can be distinguished, according to the criteria to group nodes.

- *Partitions-Based.* The clustering methods based on partitions divide the data set into different disjoint subsets. The operation involves assigning points to different clusters, whose number is initially fixed, and iteratively improving the division, until a previously defined heuristic finds the optimal solution. The praised *K-means* [65] algorithm belongs to this class.
- *Density-Based.* In the previously described algorithms, such as the K-means algorithm, the similarity of the points (instances) for cluster assignment is assessed through a specific distance measure. In density-based algorithms, however, clusters are based on density measures. The *DBSCAN* [66] algorithm belongs to this kind of clustering techniques.

Association Rule Mining. This explores the relations between attributes that exist in data, thus detecting attribute-value conditions that occur frequently together. It has been popularised by Agrawal [67], who proposed *Apriori*, the best-known algorithm to mine association rules. Association rules can be used as the basis for decisions about marketing activities, as their best known application conducts analysis of market baskets. In Ref. [68, 69] the authors present two new algorithms for solving basket market analysis. The empirical evaluation presented in the aforementioned paper shows that these algorithms outperform the existing methods for both small and large problems. *Apriori Hybrid* is also shown as a combination of the best features of the two proposed algorithms. A more recent survey of association rule mining can be found in Ref. [70].

Prediction. Also called *supervised learning methods*, its goal is to learn from a set of training data, with the aim of predicting the class of new unlabelled records. More specifically, let $\mathcal{X}$ be the feature space and its possible values, and $\mathcal{Y}$ be the space of possible labels of a target variable. The underlying assumption is that within exists a function $f(\mathcal{X}) \rightarrow \mathcal{Y}$ that assigns a record to a class depending on the values of its describing features. The function f thus aims to predict values of a class attribute from other attributes based on training data. The classification algorithm will try to minimise the distance between the real and the predicted value (*i.e.*, minimise the expected error). Consequently, the classifiers are usually evaluated by assessing their predictive accuracy on a test set. The most frequent measures to validate the models are: (*i*) accuracy, defined as the proportion of correct predictions obtained; (*ii*) precision, the proportion of correct positive forecast; (*iii*) recall, the fraction of relevant instances that have been retrieved; and the (*iv*) F-measure, or $F1$ measure, the harmonic mean of precision and recall.

Depending on the type of variables to predict, we can distinguish between:

- Numeric Value Prediction or regression: which attempts to predict a numeric attribute value by modelling a continuous-valued function.
- Classification: which aims to predict a discrete class label.

Algorithms and models

Many classification and prediction algorithms have been proposed in the last few decades, each bringing different advantages and disadvantages, along with their own requirements on the format of the data. Below are the most successful and well-known techniques.

Naïve Bayes classifiers. This refers to a family of simple probabilistic classifiers based on applying Bayes' theorem with the assumption of independence between the features. Consider an instance, described by a vector of features X, which may belong to several classes C_K (being K the number of classes). The probability of that instance to belong to a class i is then given by:

$$p(C_i|X) = \frac{p(C_i)p(X|C_i)}{p(X)}. \tag{1}$$

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

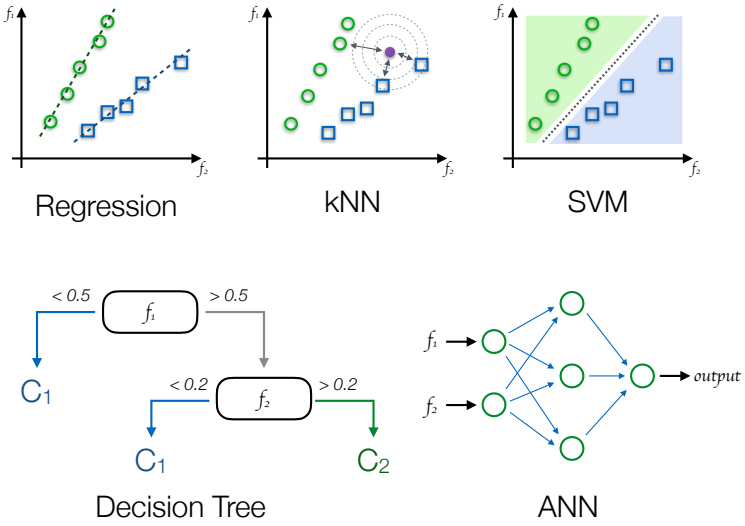


Figure 3: Example of five selected data mining classification algorithms. Green circles and blue squares represent two classes, *e.g.* control subjects and patients. The three algorithms in the top row respectively represent simple linear regression, kNN (in which new instances are classified according to the class of their nearest neighbours), and SVM (which divides the feature space is two non-overlapping regions). In the bottom row, Decision Tree, in which numerical conditions are sequentially tested of the available features; and Artificial Neural Network, which optimises a mathematical function on the input features. Additional graphical examples can be found in Figs. 4 and 9.

In other words, the *posterior* probability of the instance to belong to C_i is proportional to the *prior* probability, to the likelihood, and inversely proportional to the evidence.

Naïve Bayes classifiers are highly scalable, requiring adjustments of parameters that grows linearly with the number of variables in the problem, and thus leading to a linear computation cost. More information can be found in Ref. [71].

Regression. Regression analysis is the most basic and commonly used predictive tool, especially in practical applications [72], and is based on a statistical process for estimating the relationships among variables. As it can be seen in Fig. 3 Top Left, the goal of regression analysis is fitting a linear function through a scatter plot. In the simplest case of univariate regression analysis (*i.e.* with one dependent and one independent variable), this goal can be defined as:

$$h_\theta(x) = \theta_0 + \theta_1(x). \tag{2}$$

The training process thus entails finding a set of θ_i that fit the target population; in other words, this implies minimising the following cost function through, for instance, a gradient descent algorithm or a least squares approach:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h^i_{\theta_x} - y^i)^2. \tag{3}$$

Real-world problems are usually not restricted to one independent variable, thus requiring the use of a multivariate regression. This, nevertheless, leads to the problem of overfitting: when too many features are present, the learned hypothesis may fit the training set very well, but fail to generalise to new examples.

If reducing the number of features is not an option, the solution is to resort to *regularisation*: all features are kept, but the parameter values are reduced. In regularised linear regression, we choose θ

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to minimise:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h^i_{\theta_x} - y^i))^2 + \Lambda\sum_{i=1}^{n}\theta_j^2. \tag{4}$$

Logistic regression. There are many situations in which we need to predict the value of a dependent variable as a function of other independent variables, as in regression, but in which the former is discrete, and specifically binary (*e.g.* the patient improves after a treatment, the credit is paid, the email is spam, *etc.*). The value associated to each observation i is thus either $y_i = 0$ or $y_i = 1$. Mathematically, y_i is a realisation of a random variable Y_i that can take values one and zero with probabilities p and $1 - p$, respectively. Assume that $Pr(Yi_1|X = x) = p(x; \theta)$, for some function p parameterised by θ, and assume that observations are independent of each other. The conditional likelihood function is

$$\prod_i^n PrY = y_i| = x_i = \prod_i^n p(x_i;\theta)^{y_i}(1 - p(x_i;\theta)^{1-y_i}). \tag{5}$$

In this case it is important to see that the mean and variance depend on the underlying probability. As a consequence, any factor affecting the probability will also alter both the mean and the variance of the observations. This suggests that a linear model that allows the predictors to affect the mean but assumes that the variance is constant will not be adequate for the analysis of binary data.

In this kind of models, the constraint, $\pi = p(x_i; \theta)$, tells us that π must be the same whenever x_i is the same, and if π is a continuous function, then similar values of x_i must lead to similar values of π . Assuming p is known (up to parameters), the likelihood is a function of θ, then we can estimate θ by maximising its likelihood.

Summing up, we have a binary output variable Y, and want to model the conditional probability $Pr(Y = 1|X = x)$ as a function of x; any unknown parameters in the function are to be estimated by maximum likelihood. How can we use linear regression to solve this problem? Logistic regression uses a logistic (or *logit*) transformation of $\log p$, by $\log \frac{p}{1-p}$. This can be transformed into a linear function of x without fear of nonsensical result. Formally, the model logistic regression model is:

$$\log \frac{p(x)}{1 - p(x)} = \beta_0 + x\beta \tag{6}$$

The final classification model is then given by the solution of the equation $\beta_0 + x\beta = 0$, which represents the line that separates the two classes. The way to minimise the misclassification error is to predict $Y = 1$ when $p \geq 0.5$ and $Y = 0$ when $p < 0.5$. This is to say that non-negative values of $\beta_0 + x\beta$ predict value 1, and 0 otherwise.

Bayesian networks. Probabilistic graphical model that represents a set of random variables and their conditional dependencies through a directed acyclic graph [73]. Edges represent conditional dependencies, while nodes that are not connected represent variables that are conditionally independent of each other. Each node is associated with a probability function that takes a particular set of values for the node's parent variables as the input, and gives the probability (or probability distribution, if applicable) of the variable represented by the node as the output.

Decision trees. Involves a set of techniques aimed at generating comprehensive 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, while each branch represents a value that the attribute can take - see Fig 3 Bottom Left for a graphical example. Decision trees where the target variable can take on continuous values are called regression trees. Let D be the set of training records that reach a node. The general procedure to build the tree is as follows:

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- If D contains records that belong to the same class, then this is a leaf node.
- If D contains records that belong to more than one class, use an attribute test to split the data into smaller subsets. Apply the procedure recursively on the obtained subset.

Depending on the criteria (information gain, Gini index, *etc.*) chosen to decide the splitting point, different algorithms have been described in the following literature. Gini index is used in CART [74], SLIQ [75], SPRINT [76]; information gain is used in ID3 [77] and in the well known $C.45$ [78]. The main advantage of decision trees is that they are simple to understand and requires little data preparation, being able to handle both numerical and categorical variables. Its ability to perform well with large datasets has made decision trees one of most used techniques, even though they present the drawback of being based on heuristics where locally-optimal decisions are made at each node.

Random forests. When problems are characterised by a large number of variables, each one of them encoding very little information, the best strategy is to grow an ensemble of trees and letting them vote for the most popular class. Following this idea, random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [79]. Each tree in random forest is grown as follows:

1. Sample with replacement a given number of cases from the training set at random. This sample will be the training set for growing the tree.
2. Given M input variables, select randomly at each node $m << M$ variables, and choose the best one to split the node.
3. Grow the tree with no pruning.

Random forests have two important advantages: first, they do not suffer from overfitting, and can thus be use in small data sets. Second, they have been shown to outperform most known algorithms, in terms of accuracy.

Markov Random Field. Markov Random Fields (MRF) are the outcome of a branch of probability introduced in the 80s [80]. Specifically, an MRF is similar to a Bayesian network, in that its aim is to represent dependencies, but extends the latter by introducing concepts such as undirected and cyclic dependencies. This flexibility makes them very appropriate for classification tasks in vision problems, *i.e.* when the focus is on the relationship between a pixel and its neighbours.

Hidden conditional random fields. Hidden conditional random fields (HCRF) are discriminative latent variable models for classification problems in structured domains, where inputs can be represented by a graph of local observations [81]. An HCRF framework learns a set of latent variables conditioned on local features. They have been shown to successfully learn the hidden structure of a given classification problem, provided an appropriate validation of the number of hidden states is provided. HCRFs learn not only the hidden states that discriminate one class (label) from all of the others, but also the structure that is shared among labels. The main limitation of HCRFs is that finding the optimal number of hidden states for a given classification problem is not always intuitive, and consequently learning the correct number of states often involves a trial-and-error process with cross-validation, which can be computationally very expensive.

Artificial Neural Networks(ANN). The structural aspects of biological neural networks inspired this method. ANNs are represented by a set of connected nodes in which each connection has a weight associated with it, and the network learns the classification function adjusting the node weights.

Here we describe the simplest possible neural network, one which comprises a single neuron. Let us imagine that the inputs of this neuron are $x_1, x_2, x_3, \ldots x_n$, plus a constant term 1. The output is defined by:

$$f(W^T, x) = \sum_{1}^{n} W_i x_i + b, \tag{7}$$

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where f is the activation function. If one chooses f to be the sigmoid function

$$f(x) = \frac{1}{1 + \exp(-x)}, \tag{8}$$

then the neuron represents a logistic regression model. A neural network is a set of interconnected neurons, where the neuron's output can represent the input for some other neuron in a different layer.

The simplest kind of neural networks, the single layer perceptron [82], has two important drawbacks: (*i*) perceptron-like methods are binary, and therefore require splitting multi-class problems into multiple binary sub-problems; and (*ii*) 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. The back-propagation algorithm [83], used in conjunction with an optimisation method such as gradient descents, allows the avoidance of those problems. The method calculates the gradient of a loss function with respect to all of the weights in the network. The gradient is fed to the optimisation method, which in turn uses it to update the weights of the nodes in the network, trying to minimise the loss function - see Fig 3 Bottom Right. All of the basics to build neural networks can be found in Refs. [84, 85].

Instance based algorithms, also called *memory-based learning* [86]. Consists of a family of learning algorithms that, instead of performing explicit generalisation, compares new problem instances with instances seen in training, which have been stored in memory. They are called instance-based because they construct hypotheses directly from the training instances themselves. One advantage they provide over other methods is their ability to adapt the model to previously unseen data. This type of algorithms requires less computation time during the training phase compared to other algorithms, as no model should be created beforehand. *kNN* [87] is one of the most well known examples of "lazy-learning algorithms"; it classifies a record using the most frequent class in its k-nearest records, by means of some distance metric - see Fig 3 Top Centre. kNN uses a function that is only locally approximated, and all computation is deferred until classification. The kNN algorithm is among the simplest of all machine-learning algorithms that can be used for classification and regression.

Support Vector Machines(SVM) [88]. Binary linear classifiers that model concepts by creating hyperplanes in a multidimensional space, which can be used for both classification and regression. A good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class, as this minimises the error. The features available in the data set indicate the axes of this space, whose values should always have a numerical form. Records are mapped into this space, and the best linear separation between them is then calculated - see Fig 3 Top Right.

Multiple Kernel Learning [89]. Multiple Kernel Learning (MKL) methods expand the concept of kernel methods to which SVMs belong, by using a predefined set of kernels, and by learning an optimal linear (or non-linear) combination of kernels as part of the algorithm [90]. MKL models are able to select an optimal kernel and the corresponding parameters from a larger set of kernels, reducing the bias introduced in the kernel selection. These different kernels may correspond to using different notions of similarity, but may also respond to the use of information coming from multiple sources (different representations or different feature subsets). By using this technique, it is possible to generate a kernel for multiple data sources, by combining kernels already established for each individual source. For this reason, one of their applications is biomedical data fusion. MKL algorithms have been developed for supervised, semi-supervised, as well as unsupervised learning.

Rule induction. Since regularities hidden in data are frequently expressed in terms of rules, rule induction is one of the tools of machine learning used for classification tasks. Rules are normally expressed in the form: *If* $a_1 = v_1$ *and* $a_2 = v_2, \ldots$ *and* $a_n = v_n$, *then* $b = y$, where a_i and v_i are respectively the condition (independent) variables and their values, and b, y are the decision (dependent) variable and its value. The algorithms aim at creating the smallest rule-set that is consistent with the training

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

data. Refs. [91] and [92] respectively present the LERS and RIPPER induction systems, which have been shown to be competitive with C4.5.

Feature selection

Feature selection methods became popular in the late 90's. They consist of algorithms used for selecting a subset of features, from the original data set, for subsequent analysis; discarding all other features which are expected to be irrelevant for the problem at hand. Data sets in the 90's still had limited sizes and relatively few variables. Feature selection methods were developed to help in data understanding, reducing training times, improving prediction performance, and helping to deal with the problem known as *curse of dimensionality* [93]. Data analytics tasks on domains such as gene and protein expression, chemistry or text classification have made feature selection a need not only for the high number of variables, but also for the shortage of records - *e.g.*, in biophysics, the reduced number of patients usually available. Refs. [94, 95] extensively review methods for feature selection.

Broadly speaking, feature selection methods can be classified in three different families:

- *Filters* select subsets of variables as a pre-processing step, independently of the chosen predictor. In some cases they compete with wrappers as being the most efficient approach. One of the most relevant examples of feature selection filters is the *Recursive Feature Elimination* (RFE), which is based on repeatedly constructing a model (using some suitable classifier) and removing features with low weights (*i.e.* of low relevance). When features are added, instead of being eliminated, the result is a *forward* strategy.

- *Wrappers.* The wrapper methods, popularised by Ref. [96], assess subsets of variables according to their usefulness to a given predictor, using a search algorithm to search through the space of possible features, and evaluating each subset by running a model on it. Only when the number of variables is reduced, all variable combinations can be exhaustively evaluated, thus avoiding the search heuristic. The machine-learning algorithm is taken as a black box, to score subsets of variable according to their predictive power. Wrappers can be computationally expensive and have a risk of overfitting in the model, in which case coarse search strategies may be applied [97].

- *Embedded techniques* implement the same idea of the wrapper methods, but proceed directly to the optimisation of a two-part objective function, with a goodness-of-fit term and a penalty for a large number of variables. These methods perform variable selection in the process of training, and are usually specific to some learning algorithms. Embedded methods that incorporate variable selection as part of the training process may be more efficient in several aspects. First, they make better use of the available data by not needing to split the data into training and validation sets. Additionally, they reach a solution faster by avoiding retraining a predictor from scratch for every variable subset investigated.

Validation

Validation is a broad term, encompassing two different families of metrics:

- *Metrics for Performance Evaluation.* The aim is to evaluate the performance of a model in a reliable way, *i.e.* in a way as general as possible.

- *Methods for Model Comparison.* In this case the problem is how to compare the relative performance among competing models, especially in those cases in which the size of the datasets can make the difference in accuracy not statistically significant. A confidence interval should be established for comparison issues. While such problem will not be tackled in this review, more information can be found in Refs. [98, 99, 100].

Here we focus on the first problem, concerning the evaluation of the predictive capability of the model generated through a data mining task.

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

As previously discussed, the performance of a model is inversely proportional to the number of errors it produces, and can therefore be measured through the *precision* metric (*i.e.* the proportion of correct positive forecasts). However, this is seldom enough to assess the real performance, and an additional distinction should be implemented between the *training* and the *generalisation* errors. The former quantifies the error rate yielded by the model when processing the training set. Clearly, a good classification model should not only fit the training set, but it should also perform correctly against records never seen before. This second part is called the generalisation error, and it is meant to measure the model's performance in a real environment.

Note that the generalisation error can be significantly larger than the training error, and this usually occurs when the training is performed over a small data set, or when this data set does not correctly represent the characteristics of the whole universe. This situation is called *overfitting*, and its probability increases with the complexity of the model (as more training data are required to optimise a complex model). For this, the simplest model is chosen when two models are available and both display the same generalisation error. This is done by applying the *Occam's razor* [101] principle.

The generalisation error of a data mining model is in principle easy to evaluate. For example, one just needs to consider a new data set, *i.e.* one not used in the training phase, and measure the error rate of the model within it. Since the amount of available data may often be limited, it is necessary to resort to the *hold out* method, in which the original dataset is split in training and validation subsets. The former (typically, two thirds of the total available records) is then used to build the model, while the latter (the remaining third) is used to test the model. The global accuracy is defined as the rate of correctly classified instances by the total number of predictions made, while the error rate is defined as the number of errors in the prediction divided by the total number of predictions. The hold out method can be repeated several times, to improve the accuracy of the estimation of the model performance.

A more complete and reliable method is the *cross-validation* [102, 103, 104]. It involves the estimation of the extra-sample error by applying the trained model to several validation sets, and averaging the error rates obtained. In the *k-fold cross-validation*, the original sample is randomly partitioned into k equal sized subsamples. Of these k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k-1$ subsamples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data. The average value of the error obtained in the k executions will be used for estimating the error. In the special case when $K = N$, being N the number of records in the data set, the method is called *Leave-One-Out Cross Validation* (*LOOCV* in short).

Data mining software

Weka [105, 106, 107]. Probably the best known and most used software for scientific data mining, Weka is a collection of machine learning algorithms written in Java. All algorithms can be executed on data files, or called from within any Java program. Freeware. Available at `http://www.cs.waikato.ac.nz/ml/weka/`.

KNIME [108]. Based on the Weka libraries, KNIME is a data analytics platform which allows executing any data mining task (from data access, data transformation, predictive analytics and visualisation) through a graphical workbench. Freeware. Available at `https://www.knime.org`.

R [109]. Free software environment for statistical computing and graphics. Available at `https://www.r-project.org`.

SAS ®. Software suite developed by the SAS Institute for advanced analytics, multivariate analyses, business intelligence, data management, and predictive analytics. Commercial. Information at `https://www.sas.com`.

Beyond these four software packages, the interested reader can refer to `http://www.kdnuggets.com/software/suites.html` for a more comprehensive guide.

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.3. What is "Big Data"?

The challenge to extract patterns from large volume of data was highlighted in the original definition of the Knowledge Discovery in Databases (KDD) process (see Section 2.2) [57]. This definition dates back to 1997, where "large" may have implied just tens or hundreds of megabits of information. Since then, the data targeted for real-world analyses have substantially increased in volume, velocity and variety, increasing the complexity of extracting value. This data explosion is behind the *Big Data* concept.

Commonly found in any new field of research, there is still no generally accepted definition of Big Data and different contexts have given birth to different descriptions and characterisations. The origins of the name seem to date back to 1998, with two specific references: Ref. [110] in computer science and Ref. [111] in statistics/econometrics [1]. For the sake of completeness, we will first review some of the most commonly used definitions and then describe the important goals of Big Data.

One of the first comprehensive definitions of Big Data was provided by McKinsey & Company and states: "Big Data refers to data sets that cannot be acquired, stored, processed and managed by classic database software within a tolerable time" [112]. This definition highlights the importance of the volume of data, which may grow faster than the capacity of storage and computation technologies. In the same year, the International Data Corporation (IDC) stated that "big data technologies describe a new generation of technologies and architectures, designed to economically extract value from large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis" [113]. This can be seen as new elements were introduced, beyond the data size: variety and velocity of data, plus the value of the analyses that can be performed on such data. This was later synthesised by Doug Laney, who in 2013 defined Big Data by a 3 *V*s model: Volume, Velocity and Variety [114]. A different approach was proposed one year later by the National Institute of Standards and Technology (NIST), which focused on the technological aspects of Big Data: "Big Data consists of extensive datasets primarily in the characteristics of volume, variety, velocity, and/or variability that require a scalable architecture for efficient storage, manipulation, and analysis" [115].

It is also worth analysing a recent trend in the study of Big Data, which can be synthesised through the HACE Theorem [116, 117, 118]: "Big Data starts with large-volume, *H*eterogeneous, *A*utonomous sources with distributed and decentralised control, and seeks to explore *C*omplex and *E*volving relationship among data." Two of the characteristics underlying this definition have a clear relation with complex networks:

Autonomous Sources with Distributed and Decentralised Control: Each data source is able to produce and collect information without involving any centralised control. Such distributed nature offers opportunities for understanding patterns of interconnections as complex networks. Examples include the World Wide Web [119, 120], and distributed wireless sensor systems [121, 122]. The challenges are to control and optimise the dynamics of those systems, even in the absence of a central director.

Complex and Evolving Correlations: In standard data sets, *e.g.* such as those used in many biomedical problems, the environmental conditions are kept strictly under control in order to assume the stationarity of the data. In other words, starting records obtained at the beginning of the analysis are as valid as the final ones. This is not necessarily true in a Big Data environment. The real world is dynamic and so are the data acquired from it. Therefore, all relations should be studied taking into account their temporal evolution [43, 44].

Three main features characterise Big Data:

- *Fast Processing*: In order to extract new insights from vast volume of data, as close to real time processing as possible is desired.
- *Prediction*: Accurate prediction is used to apply mathematical and statistical methods to these large quantities of data, in order to infer probabilities, with the aim of extracting useful insights.

[1] While beyond the scope of this review, further information on the historical analysis of terms like "data mining", "data science" and "Big Data" can be found at: `http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/`

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

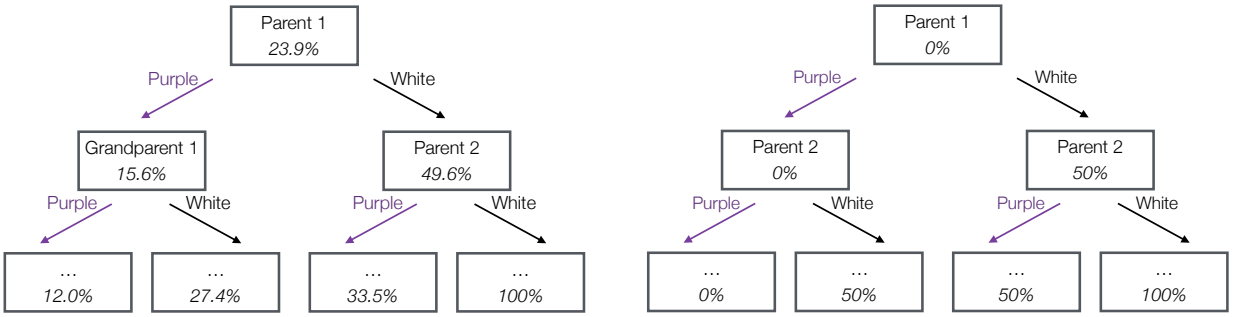


Figure 4: Formulating traits inheritance laws using data mining, using raw data analysis (Left) and by including heuristic about the statistical nature of results (Right). See main text for details.

- *Relation*: The individual pieces of information are not independent and acquire a meaning when combined together.

Despite the variety of definitions and concepts, a common notion is recognised: Big Data is a reality that mankind must both face and exploit in a structured, aggressive and ambitious way, to create value for society, its citizens, and its businesses in all sectors. Data availability and the access to data sources is paramount. Furthermore, there is a broad range of data types and sources, including structured and unstructured data, multi-lingual data sources, data generated from machines and sensors, data-at-rest and-data-in-motion. The overall value is extracted by acquiring those data, combining data from different sources, and providing access to them minimising delays; all while ensuring data integrity and preserving privacy.

While complex network theory and Big Data have had very few contact points, this is expected to change in the near future, as applications combining both concepts are starting to appear - see Section 6.4 for an example.

2.4. The limits of data mining in physics

As previously discussed, there are many real-world problems, especially in the field of biophysics, which can be seen through the lens of data mining. As an example, whenever confronted with situations in which subjects need to be diagnosed into patients and healthy subjects, this boils down to a classification task. However, data mining has not been historically well received by the physics community. Before delving into how the two fields can (and should) be leveraged for one other, we analyse the intrinsic limitations of data mining in terms of its application to physical problems.

One of the distinguishing characteristics of physical science is its effort to achieve higher levels of abstraction. Consider gravity as an example: the aim is not just to explain the trajectory of a pencil falling through the air, but also to explain the dynamics of *any* free-falling object. The concrete nature of the pencil is thus disregarded, or rather, abstracted from the considered dynamics. The same is true in complex networks analysis. The details of each system element are disregarded, and the representation of their interactions is obtained in an abstract form. On the one hand, while data mining provides some level of abstraction, this is based on an inductive (sometimes, an abductive) reasoning over the available data, and may be far from what a physicist would like to achieve. To further illustrate this point we can use a simple biological example. Since the seminal work of Gregor Johann Mendel in 1865, the laws of genetic inheritance are well understood [123]. In his original experiments two types of pea plants were crossed: purple and white flowers. After seeing the characteristics of the offspring, he postulated some simple rules for inheritance traits. Mendel then concluded that each plant has two alleles of the same genes that are randomly inherited from the two parents. This defines the colour of the flowers, with each allele being of two variants (p and w). Purple is the *dominant* form, and flowers will only be white if the pair ww is inherited. Note the abstraction: the plant colour (*i.e.* the phenotype) is explained by something not visible, the genes (*i.e.* the genotype).

If these laws were not known, one may attempt to understand inheritance traits through data mining. As the alleles (*i.e.* the genotype) are not known, the initial data set would contain a list of flower colours (the

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

phenotype), for the target plant, its two parents, and four grandparents. A data mining task could then be executed to predict the colour of the target plant, just by knowing the colour of its predecessors. When this is executed for a data set comprised of 1000 different trials, it results in the Decision Tree model depicted in Fig. 4 Left: each box is a decision node, where one feature (reported within the box) is evaluated, as represented by the arrows. The italic number inside the boxes represents the probability of obtaining a white flower offspring. In other words, if both parents are white then there is a 100% probability of getting a white offspring (right-most sequence of arrows). In terms of precision, the data mining model is not able to exactly forecast the colour of the offspring, and yields an error rate of 18.6% (using a Leave-One-Out cross-validation). Although we know that inheritance traits are random, a preliminary reaction may be to assume that the model is wrong and that it not adequately captures the inheritance principles. In the end this is not completely true, as part of the model is correct; when two parents are white, the offspring will also be white.

To take this approach one step further, we can forecast offspring colour in statistical terms. Here, we introduce *heuristic* into the data mining problem [2]. Instead of trying to predict the colour of the offspring, we may predict the probability of having white offspring. The resulting model is depicted in Fig. 4 Right; note that numbers inside the box now indicate the predicted probability of having a white offspring. The model is still not accurate (for instance, the probability of having white offspring from two purple-coloured parents is not zero, as displayed in the left-most sequence), but it is now able to identify the two most important factors as the colours of each parent.

In this simple example, we observe that it is not sufficient to create a theory of inheritance traits. Data mining is simply unable to provide us with a sufficient level of abstraction from the overall information. Fig. 4 Right does not yield concepts such as the presence of two alleles, or the existence of a dominant trait. Still this exemplifies how data mining can help detect the elements of the analysis that are important, how elements can be combined or simplified, and how data mining can provide insight into a large set of data in an automatic way.

Data mining is an abductive, rather than a deductive process: it simply induces, from the available data, what has happened in the past. If certain important variables are not present, or their description is faulty, then data mining comes to wrong conclusions. Therefore data mining cannot provide insight into what is omitted within the data, and in the case alleles existence, it does not yield higher levels of abstraction. If such data were available, or if all of the data a biologist would use to create an abstract theory were accessible, then such theory could be automatically inferred from the data.

Through this example we can observe a necessary condition. Although data mining creates abstract models from initial data, the data themselves should contain all the elements required to perform an accurate abstraction. Several examples can be found in the Literature in which such abstraction process is performed. For instance, Ref. [124] proposes an algorithm for rediscovering concepts in number theory, given an initial set of concepts and heuristics. Ref. [125] describes a computer program for the discovery of physical laws, which is able to recover the law of ideal gas, using input values observed in experiments, along with number of moles, temperature, pressure and volume. With this the program is able to recover the law of ideal gas. This latter case exemplifies the previously discussed limitation: given all the parameters, it is possible to recover the true law. Nevertheless if some of these elements are missing, no data mining algorithm can suggest, for instance, that the temperature is a key element.

Therefore, this important limitation has to be taken into account by any researcher who wants to use data mining in a physical (or biological) scenario. However, to move from a simple prediction to a general model, another essential ingredient is required: the human ability for creating abstract models. Indeed, good data scientists are not only people who simply applies different algorithms; their *spark* and intelligence should also be included in each step described in Section 2.2, such as business understanding, data understanding and data preparation.

[2] *Heuristic* in data mining refers to any approach to problem solving that employs a practical method not guaranteed to be optimal or perfect, but sufficient for the immediate goals. In other words, heuristic implies the inclusion of *a priori* information inside the problem.

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.5. Hands-on: analysing EEG data

How would a network theory practitioner understand the structure encoded in EEG data? The analysis should, of course, start with a real data set, and in this case we consider a set of EEG recordings for healthy (control subjects) and alcoholic (patients) subjects; the aim will be to assess whether this specific health condition has an effect on the overall brain organisation. The data set is freely accessible at `http://archive.ics.uci.edu/ml/`) [126]. It contains measurements from 19 electrodes placed on subject's scalps which were sampled at 256 Hz (3.9-msec epoch) for 1 second [127], with the electrode positions located at standard sites [128]. Each subject was exposed to a single stimulus, corresponding to pictures of objects chosen from the 1980 Snodgrass and Vanderwart picture set [129]. Of the whole set, we randomly selected 40 recordings of control subjects, and 40 of patients. Our data set is then represented as a $80 \times 19 \times 256$ tensor. An example of four randomly chosen time series is depicted in Fig. 5 Top Left.

To construct a (functional) network for each subject, it is necessary to define the meaning of nodes and links. The solution to the former problem is simple, as EEG electrodes provide a natural set of nodes. A link can then represent whether and the extent to which two nodes are connected, being synchronisation is a typical choice [3]. We here consider two possibilities: a linear Pearson's correlation, and a Granger Causality [130, 131]. As an illustration, in the case of the linear correlation, a pair of nodes i and j would be connected with a strength equal to $w_{i,j} = cov(x_i, x_j)/(\sigma_{x_i}\sigma_{x_j})$, where x_i and x_j are the time series recorded at nodes i and j, and cov is the covariance. For each subject, we thus obtain a *weighted adjacency matrix* $\mathcal{W}$, whose element $w_{i,j}$ indicates the strength of the connection between nodes i and j - see two examples in Fig. 5 Top Right.

Finally, we calculated four common topological metrics from each network after applying an arbitrary threshold; the weighted adjacency matrix is thus converted to an unweighted one $\mathcal{A}$, such that $a_{i,j} = 1$ if $w_{i,j} > \tau$, and $a_{i,j} = 0$ otherwise. Here we consider the following topological metrics: the degree density, transitivity, assortativity and efficiency. The corresponding histograms, comparing control subjects and patients, can be found in Fig. 5 Bottom, for networks created with correlation (top) and Granger Causality (bottom).

Some interesting conclusions can already been drawn. For instance, in the case of the linear correlation, patients seem to have higher degree density, transitivity and efficiency, thus indicating a more strongly connected network. Although less evident, similar results, particularly a higher efficiency for patients, are observed in the case of the Granger Causality.

One may stop here, add some statistical tests to confirm that the differences are indeed significant, and probably have a nice preliminary paper on the effects of alcoholism on brain activity. However, one may also start asking important questions. Is it better to use a linear correlation, or the Granger Causality? In other words, are these two metrics assessing different and significant aspects of brain dynamics? What is the best topological metric to describe the differences between both groups? Is the applied threshold the best possible one or can it be optimised? And finally, what is the discriminative power of this analysis? Could we diagnose the health condition of a new subject?

Answering these questions, or more specifically, understanding how we can answer them by using data mining is the objective of this review. We leave these questions unanswered till the next section where the use of classification algorithms is further discussed.

3. Classification

In this section, we start dealing with one of the most important tasks in data mining: *classification*, *i.e.* the creation of models to predict the class of an unseen instance. As discussed in the introduction, this has many applications in cross-disciplinary fields, as in medicine, with the creation of diagnostic tools. Before presenting examples of classification tasks involving complex networks (Sections 3.3 to 3.5), two additional

[3] The attentive reader would note that some intermediate steps may be required, *e.g.* filtering the data for spurious or erroneous samples, or selecting only some frequency bands. As our aim is not to yield a correct neurological analysis, but instead to exemplify the process to be followed, we have omitted these pre-processing tasks.

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

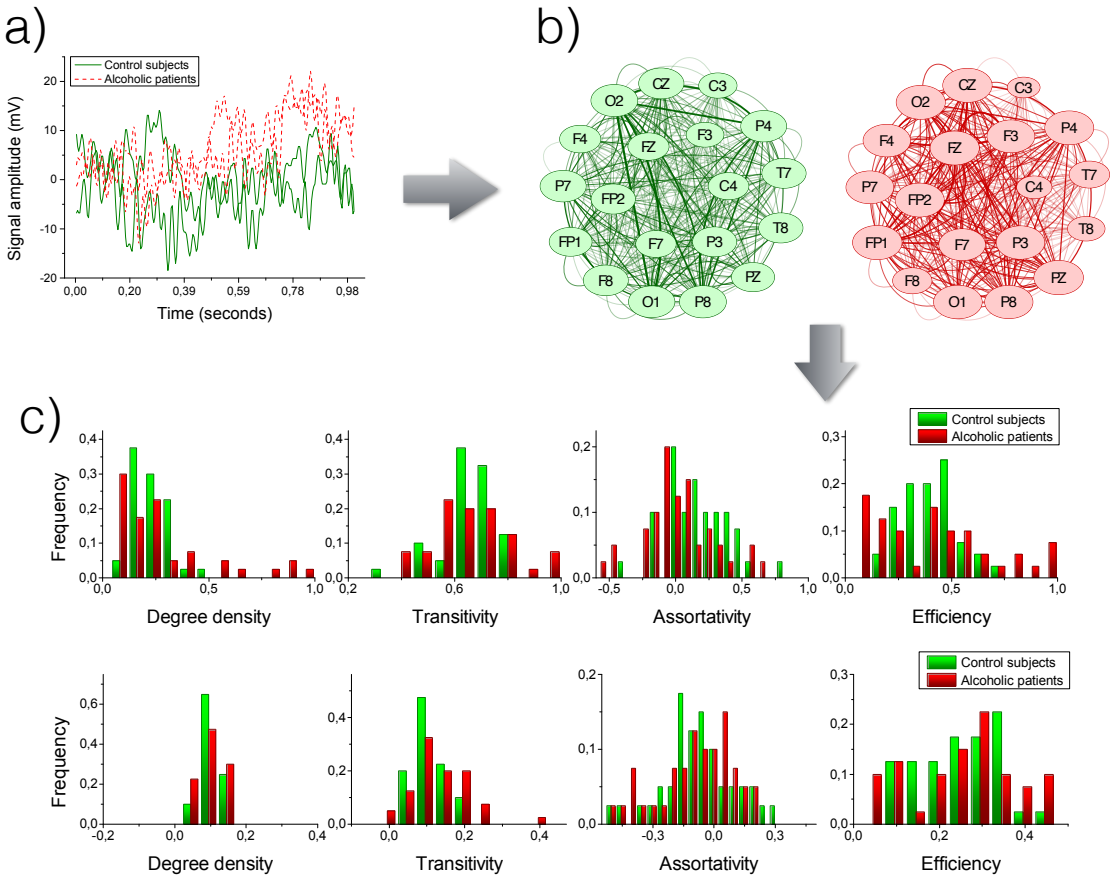


Figure 5: A standard functional network analysis. (Top Left) Example of four time series, two for control subjects and two for patients, representing the brain activity during a cognitive task. (Top Right) Two functional networks (green for the control subject, red for the patient) obtained through linear correlation. The links weight is represented by the width of the line; the size of nodes indicates their centrality. (Bottom) Histograms of four metrics, for networks obtained through correlation (top) and Granger Causality (bottom).

issues should be discussed. First, the relevance of using classification algorithms, against resorting to simple statistical metrics (Section 3.1); second, the advantages associated with using network topological metrics, against the direct analysis of the network adjacency matrix (Section 3.2).

3.1. Statistics vs. classification

Suppose one is studying a given network metric, *e.g.* the clustering coefficient, to see if it indicates relevant differences between populations. Its relevance can be assessed in two different ways. The first relies on the use of some non-parametric statistical test, *e.g.* the Kolmogorov-Smirnov test, to assess if samples corresponding to both groups are drawn from the same distribution. If the hypothesis can be rejected, then the two groups present a statistically significant difference in their network structure. The final result of this analysis would be a p-value, which can then be compared against a significance level threshold (usually denoted by α). The second option involves performing a classification task, in which the subjects within the two groups are classified, using the network metric as the input feature. In this case, the output will be a classification score (or equivalently the classification error). The considered metric would be relevant if the score is sufficiently high (or conversely, the error is sufficiently low).

In principle, one may expect both methods to be equivalent. A low p-value signifies that the values of the considered metric are well separated between both groups, and thus a classification task can be easily

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

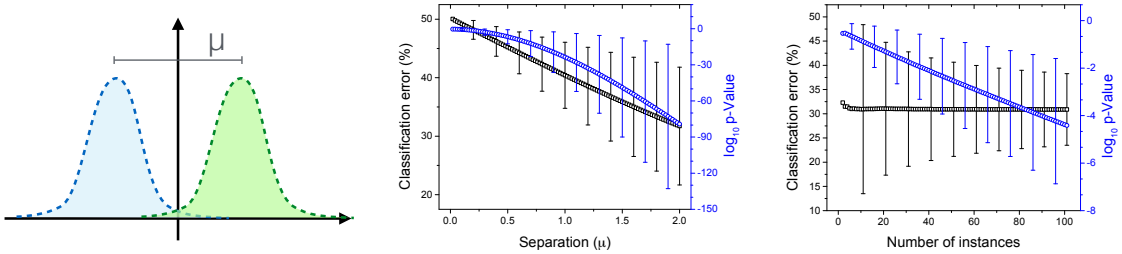


Figure 6: p-value *vs.* classification score. (Left) Data setup, with two normal distributions of $\sigma = 1.0$ and centres separated by a distance μ. (Centre) p-value and classification error as a function of the separation μ. (Right) p-value and classification error as a function of the number of samples drawn from each distribution.

performed. Nevertheless, there are situations in which this is not true. Fig. 6 Left depicts a simple situation, with two probability distributions representing the topological values obtained for the two groups of subjects. For the sake of simplicity, both distributions are normal, with $\sigma_1 = \sigma_2 = 1.0$; their centres are separated according to a parameter μ, such that $\mu_1 = -\frac{1}{2}\mu$ and $\mu_2 = \frac{1}{2}\mu$. When a large number of samples ($n = 1000$) are drawn from both distributions, and the resulting p-values and classification errors are calculated as a function of μ, the obtained results are as depicted in Fig. 6 Centre. While the p-value drops very fast (note the log scale), the error presents a more flat slope. As an example, in the separation of $\mu = 1.0$, the p-value is of the order of 10^{-30}, while the classification error is still around a 40%. This illustrates an important point: when two distributions are different in a statistically significant way, the difference may not be sufficient to enable a successful classification task. While uncommon, the opposite situation can also appear. Fig. 6 Right reports the evolution of both the p-value and the classification error as a function of the number of instances (samples) available in both groups (for $\mu = 1.0$). In the case of very small sets (*i.e.* below 20 subjects), the reduced dimension prevents obtaining a p-value below the desired significance level, even though a classification can be successfully performed. Notice that this situation is quite common in biomedical problems, where the number of subjects available is seldom in the hundreds.

These results suggest that a researcher should err on the side of caution. The p-value provides valuable information about the statistical significance of an analysis, but it is not enough to assess its usefulness. This is especially relevant when the objective is not just theoretical, as when a diagnostic tool is constructed. Conversely, when the available data set is small, the classification score alone may be misleading, and it should be complemented by an analysis of its significance.

3.2. Are network metrics really necessary?

Data mining has previously been defined as the computational process of discovering patterns in sets of data. As such, one may question the necessity and advantage of using a complex network representation to extract structural patterns, as this would be completed by data mining analysis. Suppose we are studying a set of networks, each fully described by an adjacency matrix $\mathcal{A}_i$. The same information can be translated into a vector $\overline{A}_i = \{a_{1,1}, a_{1,2}, \ldots, a_{n,n}\}$, which can then be used as the set of features for training a classification algorithm. If such algorithm were optimal, in that it could detect any class of patterns, it would describe the system as a network and the topological features extracted would be meaningless.

It is simple to disprove this hypothesis. Fig. 7 presents the results of a process in which 20.000 random networks of ten nodes have been generated with a fixed link density ($p = 0.3$). On one hand, two topological metrics (efficiency and clustering coefficient) were extracted from these networks. However, on the other hand, the adjacency matrices have been fed inside an Artificial Neural Network (ANN), trained to recover the obtained metric values. Both outputs, *i.e.* the true topological values and the estimation obtained by the ANN, are fit linearly, and the coefficient of determination R^2 represented in Fig. 7 as a function of the number of nodes and hidden layers in the ANN. The small R^2 obtained in each case (below 0.04) indicates that the data mining model is not able to recover the true topological indicator. This is especially noteworthy in the case of the clustering coefficient, as it is a local metric that can be estimated by just

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

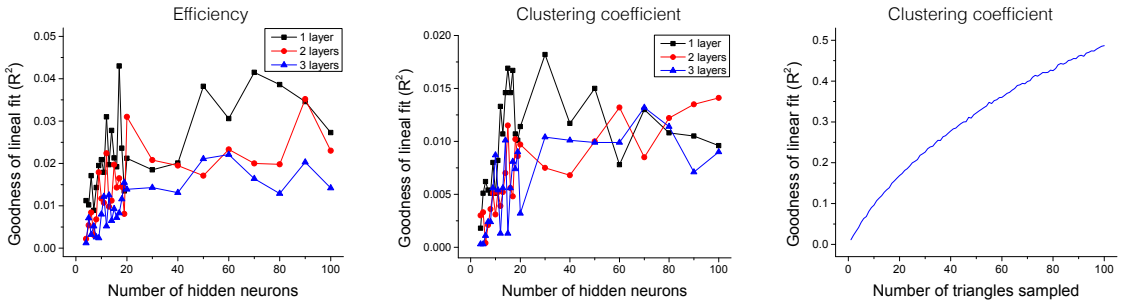


Figure 7: Network metrics *vs.* direct data mining. (Left) Coefficient of determination (R^2) for the linear fit between network efficiency and the prediction of an Artificial Neural Network - see main text for details. (Centre) R^2 for network clustering coefficient. (Right) R^2 for network clustering coefficient, and the approximation obtained sampling triangles of the network.

sampling triplets of nodes and evaluating the fraction of nodes that are connected. The R^2 obtained in this process, as a function of the number of triplets sampled, is shown in Fig. 7 Right.

This simple numerical experiment demonstrates that one single data mining algorithm is in general not sufficient to deal with the structure of a sufficiently complex system. Additional reasons for this can be identified:

Macro- and meso-scale perspectives. Data mining algorithms are optimised to extract relationships between a small set of features, but are not optimised to analyse interactions on a global scale. For instance, calculating the efficiency of a network requires estimating each of the shortest paths within it; a task that cannot be solely accomplished by an ANN.

Statistical perspective. Most data mining algorithms hitherto designed do not handle a statistical approach well. On the contrary, they are optimised to detect patterns between fixed sets of features. Even in the case of local metrics, such as the clustering coefficient, the ANN is not able to statistically sample the data set.

It is then clear that a complex network approach provides an important added value to the study of complex systems, which cannot be obtained by data mining alone. In the following sections, we will review examples, mainly drawn from biomedical and engineering applications, in which data mining and complex networks are used to create classification models.

3.3. Classifying with complex networks: Neuroscience

In biophysics it is often important to be able to *diagnose* a disease. Two or more states, *e.g.* normal (or control) and pathologic, typically need to be distinguished. The problem then consists in creating meaningful network representations, extracting topological metrics, and assessing the differences between the states. Within neuroscience, a natural way to build a network representation is the use of the functional network concept: synchronisation between time series recorded by means of EEG [4], MEG [5] or fMRI [6] can be assessed, and mapped in a functional network.

Historically, most of the classification tasks performed with functional networks and data mining models have been focused on MRI and fMRI. Within this, the most used classification model is by far the Support Vector Machine (SVM). The reasons for this choice are easily identified. SVM is conceptually simple, easy to

[4] *Electroencephalography* (EEG), an electrophysiological monitoring method to record the electrical activity of the brain, both in resting states and during a specific task.

[5] *Magnetoencephalography* (MEG): similar to EEG, but focuses on recording the magnetic signals generated by the brain activity.

[6] *Magnetic Resonance Imaging* (MRI), and its *functional* equivalent (fMRI), is a medical imaging technique that uses magnetic fields and radio waves to form images of the brain.

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

represent, and is simple to extract biological rules from the model structure. In other words, SVM simplifies the translation of results to a biomedical language.

Numerous neurological pathologies have been studied by means of fMRI and SVM. Diagnosing Alzheimer's disease has been extensively discussed in the literature, both with linear [132, 133, 134] and non-linear kernels SVM [135]. The discrepancies between young and older adults, or in general children *vs.* adults, have also been addressed, due to its relevance in understanding how the brain develops and reaches a mature functional status [136, 137, 138]. Additional pathologies include:

- Schizophrenia, a pathology whose main challenges are its complexity, in terms of multiple symptoms, and the lack of characterisation in terms of local abnormalities of brain activity. The fact that this affects the collective, emergent working activity of the brain suggests that a complex network approach can yield relevant results. This topic is discussed in Refs. [139, 140].

- Major Depressive Disorder (MDD) patients, and their discrimination from non-depressed controls, by means of whole-brain tractography [141]. Results indicate that the best network metric is the small-worldness, a measure of the balance between global integration and local specialisation [23, 24].

- Social Anxiety Disorder (SAD), an extreme fear of being scrutinised and judged by others in social or performance situations [142].

- Attention Deficit Hyperactivity Disorder (ADHD), in which symptoms include difficulty staying focused and paying attention, difficulty controlling behaviour, and hyperactivity [143].

- Spinocerebellar ataxia type 7 (SCA7), a genetic disorder characterised by degeneration of the motor and visual systems [144].

- Discrimination between shiverer (C3Fe.SWV Mbp^{shi}/Mbp^{shi}) and background control (C3HeB.FeJ) mice, a model for studying Parkinson's Disease [145].

Kernel learning, the concept underpinning SVM, which evolved to Multiple Kernel Learning, has received increasing attention in neuroscience and has yielded relevant results. As already discussed in Section 2.2, this presents two important advantages. First, it allows the merging of different types of information, for instance coming from different biological sensors. Second, when spatial information is embedded, it detects the regions of the brain which are relevant for the task. Examples of classification tasks using MKL include Ref. [146], where this technique has been applied in two tasks: a classification of patients with schizophrenia *vs.* controls and a classification of wake *vs.* sleep healthy subjects. It has also been used to diagnose amyotrophic lateral sclerosis (ALS) [147], Alzheimer's patients [148, 149], Mild Cognitive Impairment (MCI) [148, 150], post-traumatic stress disorder (PTSD) [151], and image-stimuli data [152].

Most of the examples above share a common feature: fMRI data are analysed by means of a single classification algorithm; specifically SVM (or its evolution MKL). This is an important limitation, as *one algorithm does apply to all scenarios.* Different classification models make various assumptions on the underlying data, which can then translate into important differences in classification scores. The problem of comparing classification algorithms is not new in data mining. For example, the interested reader may refer to Ref. [153] for a comprehensive comparison, and to Ref. [154] for a discussion of the associated methodological problems. Without delving into theoretical details, we stress once again that, to maximise the relevance of obtained results, scientists should try more than one algorithm. Of the works previously described, three papers have performed such comparison. Specifically, Ref. [139] compares Markov Random Field, Gaussian Naive Bayes and SVM; Ref. [137] Naïve Bayes with kernel densities, SVMs with linear and 2nd order normalised polynomial kernels, a radial basis function network, a multi-layer perceptron, and tree-based algorithms (C4.5, functional tree and random forest); Ref. [143] linear SVM, SVM with a RBF-SVM, decision stumps as a base classifier in adaboost, random forests, and C4.5 decision trees. Results are heterogeneous; the Ref. [143] reports that the best classifier is a SVM with radial kernel, while Markov Random Field outperforms SVM in Ref. [139]. These results highlight the importance of

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

comparing different classification algorithms, as even similar data (fMRI) may be better understood by different approaches under different conditions.

Examples of classification tasks on functional networks can also be found in studies using EEG and MEG data. In Ref. [155], a diagnosis in children with partial epilepsy is performed; resting EEG activity away from epileptic seizures is used to create networks, and then classified by means of Decision Trees (DT). The detection of epilepsy in newborn infants is also the topic of Ref. [156]; this study aims at identifying important EEG channels, corresponding to nodes, using SMVs and Gaussian Mixture Models (GMMs) [157]. Ref. [158] uses EEG recordings of children carrying out a face perception task to diagnose autism; the model used is a SVM with polynomial kernels, giving a notable 94.7% accuracy result.

A similar goal is pursued in Ref. [159] but this time using MEG data. The functional networks associated to resting activity of eight autistic subjects reconstructed using Granger Causality is discriminated with an 87.5% accuracy using SVMs. SVM and Granger causality are also used in Ref. [160], to discriminate mild traumatic brain injury (mTBI) patients. Finally, while the classification is not the main topic of the work, SVM are also used in Ref. [161] to discriminate MCI and healthy subjects.

While functional networks obtained from EEG, MEG and fMRI data have yielded important results, Ref. [162] sounds a cautionary note. One of the problems of both EEG and MEG analyses is that the true network, or the network created by all neurons, cannot be recovered due to intrinsic limitations in the available technology. All previous reported studies thus deal with *downsampled* networks, whose number of nodes is defined by the number of sensors in the machine. The authors showed, using a large set of synthetic networks with standard topologies and corresponding down-sampled sub-networks, to then compare the resulting topologies using nearest neighbour classification (NNC), multi-layer perceptron (MLP) and support vector machine (SVM), that the topological measures may substantially deviate from that of an underlying, larger network, when sampled at a few surface sites. This suggests that the use of EEG and MEG recordings introduces an important error in the analysis. The problem of defining, or *sampling*, the nodes of a network will be further discussed in Section 5.1.

Finally, brain connectomics has also been addressed with data mining tools. For instance, Ref. [163] studies the effect of traumatic brain injury, with white matter connections between 165 grey matter brain regions defined using tractography, and structural connectivity matrices calculated from skeletonised diffusion tensor imaging data. Support vector machines trained with graph metrics of white matter connectivity matrices from the microbleed group were able to identify patients with a history of traumatic brain injury with 93.4% accuracy.

3.4. Classifying with complex networks: Brain-Computer Interfaces

While the field of Brain-Computer Interfaces (BCI) also deals with the analysis of the human brain dynamics, its focus is different from the previously reported examples of neuroscience classification problems. Instead of analysing differences between normal and pathological conditions, BCI aims at detecting patterns in brain activity to control external machines, such as computers. Examples of applications include the discrimination between movements involving the right hand and right foot, which, when performed by means of SVM, reaches a 99.1% accuracy [164]; and the discrimination of left and right hand movements, also performed by means of SVM [165], with accuracy rates ranging from 97.77% to 100%.

Beyond SVM, other classification algorithms have been used in BCI problems. An interesting example can be found in Ref. [166], in which a classification of imaginary motor tasks from EEG data was performed by means of *hidden conditional random fields* (HCRFs) [81]. An HCRF generates sequences of labels for sequences of input samples, and thus allows exploiting the temporal structure present in EEG data. In this study, the interest of this methodology goes beyond its improvement of classification accuracy, as it allowed extracting valuable information on phenomena underlying a motor imagery task.

Finally, brain activity associated with the execution of a given task may not be representative of a complete functional network. It may be interesting to reduce their complexity, for instance by extracting the corresponding minimum spanning tree (MST), and by performing the classification on its characteristic parameters. This was proposed in Ref. [167], in the task of discriminating between imagery hand movements (both right and left) and resting state conditions.

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.5. Classifying with complex networks: -omics data

Omics is a term that generically refers to research fields in biology ending in *-omics*: genomics, lipidomics, proteomics, foodomics, transcriptomics and metabolomics [168]. Omics has received increased attention in the last decade due to its possibilities regarding *system biology*. The causes and development of a disease are seldom due to individual elements; *e.g.* single genes, proteins or metabolites. On the contrary, it is usually necessary to characterise their interactions, both within the same omics family, and between elements of a different kind. As such, omics studies naturally merge network representations with data mining analyses.

One of the most important problems is *lethality mining* (also called *dispensability*), which aims at detecting which omics elements are essential for the functions of the cell. Notably, in most organisms knocking out most genes has little or no affect on the cell viability, and only in few cases is lethal. The problem is then detecting such lethal elements, as understanding which biological mechanisms have been affected.

Gene dispensability has been studied by combining network topological features with individual properties related to essentiality, such as flux deviations, centrality, codon frequencies of the sequences, co-regulation and phyletic retention. When this information is combined in a decision tree, it has been possible to identify lethal genes in *Salmonella typhimurium* with an AUC score of 75% - 81% [169]. Expanding from single genes to gene pairs, synthetic lethal (SL) interactions have been studied and can be predicted from the topology of the interaction network. Ref. [170] presents results combining multiple classifiers, namely SVM, neural network, RIPPER (rule-based classifier), random forest, k-nearest neighbours (KNN) and decision trees. Random walks on the interactions networks, combined with a J48 decision tree, also yielded excellent results, which include: a true positive rate of 95% against a false positive rate of 10% for *Saccharomyces cerevisiae*, and a true positive rate of 95% against a false positive rate of 7% for *C. elegans* [171]. A classification task can also be performed to identify "friendly" connections, *i.e.* interactions which are beneficial to the cell. This has been studied in *S. cerevisiae* gene networks through graph diffusion kernels, with a 50% precision and 20% to 50% recall [172] - see Fig. 8 for a graphical representation of the results.

Protein's dispensability has been relatively less studied, due to the challenges associated with modifying the presence of these molecules *in vivo*. Nevertheless, it has been proposed the use of neural networks and SVMs, combined with complex network features (*e.g.* the number of neighbours in the protein-protein interaction network and the gene expression cooperativity network), to predict essential and non-essential genes, for then using trained models to predict associated protein dispensability [173]. SVM classifiers and network properties of protein interaction networks have also been used to predict synthetic sick or lethal (SSL) interactions between proteins, specifically in *S. cerevisiae* [174].

Another relevant problem in omics analysis is the association of functions with specific genes and proteins, such as function annotation. Historically, this task has been performed by using various methods based on sequence or structural similarity [175, 176]; more recently, the topology of protein interaction networks has been proposed as a viable alternative [177, 178]. It has further been proposed that lethal interactions and functional similarity are related problems. For example, kernel machines have been used to predict gene functions starting from lethality and interaction data in yeast [179]; and a logical learning system called eXtended Hybrid Abductive Inductive Learning (XHAIL) for mapping genes of unknown function to enzymes in a known metabolic network [180].

Beyond the use of lethality information, protein function prediction has been performed by means of a large number of complex networks and data mining techniques. Among others, these include: Markov random fields, to infer a protein's functions using protein-protein interaction data and the functional annotations of its interaction protein partners [181]; decision trees and naïve Bayesian networks for predicting protein interactions in yeast [182]; and random forest based methods, combined with bipartite network representations [183]. Related problems are the identification and annotation of proteins that perform their jobs as part of a multi-protein unit, *i.e.* of a *complex*. Methods combining complex networks and data mining have been proposed to tackle this task. To illustrate, the analysis of level-2 neighbours, *i.e.* proteins that do not interact but share interaction partners [184]; and SVM classifiers on yeast protein interaction networks, achieves a coverage of 89% at a false discovery rate of 10% [185].

Omics data and omics interactions have also been used to understand the origin and development of different diseases. The study of endocrine disorders is of special importance. The endocrine system is a

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

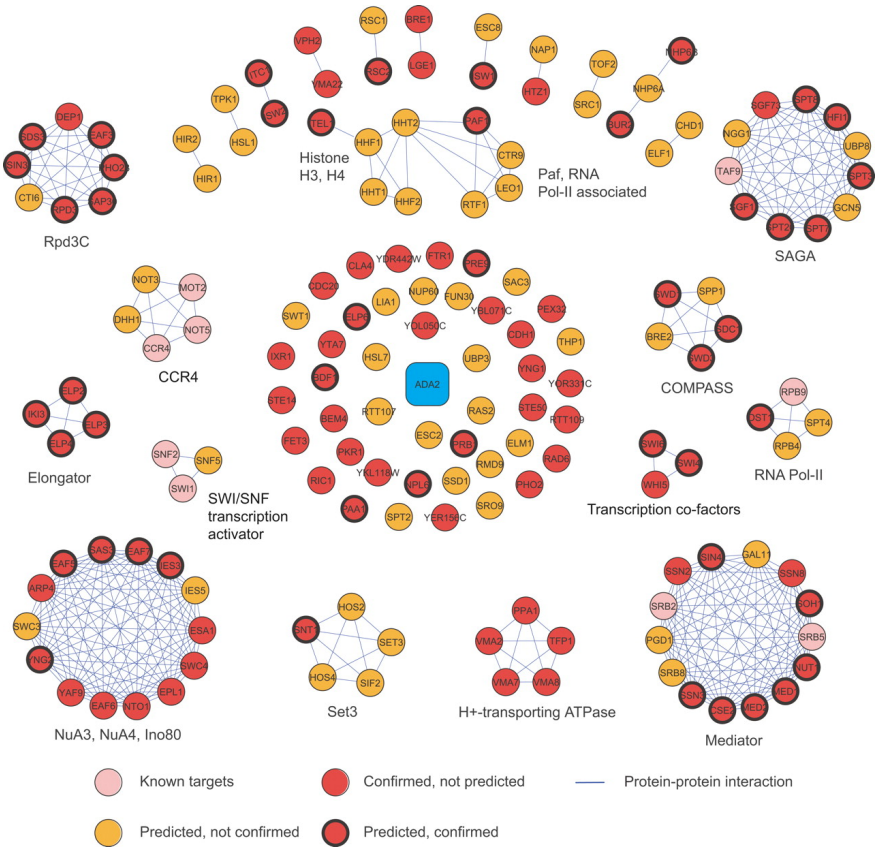


Figure 8: Friendly gene interactions identified by means of data mining techniques, organised according to the involved pathways. Reprinted figure with permission from Ref. [172].

network of glands that produce and release hormones that help control many important body functions, including the body's ability to change calories into energy that powers cells and organs. Endocrine disorders are usually the result of the failure of multiple omics elements, and are therefore well suited to be studied by means of complex networks [186]. Studies include a protein interaction network-based analysis of Juvenile Idiopathic Arthritis (JIA), a common early onset disease encompassing all forms of arthritis, of unknown origin [186]. Additionally, Hepatocellular Carcinoma (HCC) has also been studied through a omics network-based analysis, by identifying clusters of co-expressed proteins between HCC and Metastatic Hepatic Carcinoma (MHCC) [187].

Finally, it is worth citing a few studies proposing different approaches to study relationships between omics networks and diseases. Specifically, text mining has been proposed as a way to predict gene-disease associations, by automatic literature mining based on dependency parsing and support vector machines [188]; the trained system can then be used to classify the new sentences as describing an interaction between a gene pair or not [189]. A new way of creating networks of omics interactions has also been recently proposed, based on the identification of expression deviations from a reference level detected in control subjects, by means of different data mining techniques. These networks, called *parenclitic*, have been applied to problems like the detection of genes responsible for abiotic responses in plants [190], or the detection of key genes and metabolites in different diseases [191, 192, 193, 194, 195].

3.6. Hands-on: classification

Now that we have a better view of how data mining algorithms can be used to classify complex networks, let us go back to and expand the results presented in Section 2.5. Specifically, we will show how two

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

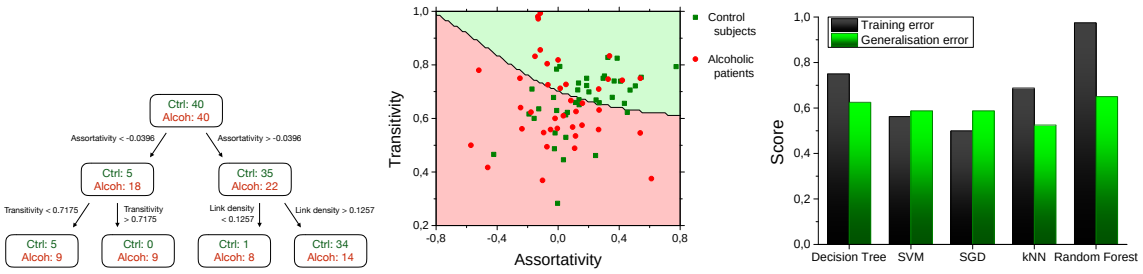


Figure 9: Classification of the data obtained in Section 2.5, for networks obtained with a correlation metric. Left and Centre panels respectively depict the result of applying a Decision Tree and a SVM model. Right panel reports the training and generalisation scores obtained through five different algorithms.

commonly used algorithms, Decision Tree and SVM, would interpret the topological features resulting from those analyses.

First, let us consider a Decision Tree. As discussed in Section 2.2, this class of algorithms constructs a top-down (or reversed) tree structure, in which each node represents an attribute, and branches different values of those attributes. A graphical representation of the resulting classification model is depicted in Fig. 9 Left. Note that the top (or main) node includes all $40 + 40$ subjects. A first selection is subsequently performed, according to the value of the assortativity; for values lower than -0.0396, we have $5 + 18$ subjects (respectively control and alcoholic), and thus more alcoholic than control subjects; on the other hand, assortativity values greater than -0.0396 give rise to the complementary situation. How can this model be used to classify? Quite simple: if we have a subject with an assortativity lower than -0.0396, and a transitivity higher than 0.7175, we can conclude that he/she is alcoholic, as we have not seen an instance of a control subject with such values (second node in the lower row). Of course, this is not always that simple: the last node indicates that when the assortativity is higher than -0.0396 and the link density higher than 0.1257, the subject is *possibly* healthy (although we are not certain).

Let us further consider the case of SVMs. Fig. 9 Centre depicts the plane created by transitivity and assortativity [7], with a black solid line representing the best separation between control subjects and patients (respectively represented by green squares and red circles). Above this line, there are more control subjects than patients, while the opposite occurs below; thus, a new subject could be classified by assessing the position in the plane.

In both cases, we can see that the classification is not perfect, although both algorithms do detect some patterns. Using these 80 subjects, we can calculate the effectiveness of the classification, *i.e.* the training score (*training* due to the fact that the same subjects are used in the training and in the validation). Results are presented in Fig. 9 Right as black bars, both for the two considered algorithms, and for three additional ones (Stochastic Gradient Descent, kNN and Random Forest). The classification task seems quite successful, especially with the latter algorithm, which yields a score close to a 100%. Nevertheless, one should never forget the problem of validation, as discussed in Section 2.2: what would happen if the algorithm was tested with a brand new subject, not present in the training set? To answer this, Fig. 9 Right also reports the *generalisation score* (green bars), calculated through a Leave One Out Cross Validation (LOOCV).

What conclusions can be drawn from this analysis? The differences in topological features that we observed, depicted in Fig. 5, can indeed be used to classify patients, and different algorithms seem to converge towards a score of 60%. This score is not very high, and essentially useless for clinical applications; but has been obtained without optimising any of the parameters (like the threshold for binarising networks). In the next section we will see how data mining can be used to optimise these parameters and ultimately improve this seemingly disappointing result.

[7] Note that the full classification model would require representing a four-dimensional space, quite beyond the possibilities of modern journals!

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4. Improving network significance

As previously introduced in Section 2.1, the analysis of *functional* networks requires tackling three interconnected problems: the selection of a synchronisation metric, of a threshold, and of a set of topological metrics. This Section explores how data mining can help in performing a guided selection, *i.e.* how it can provide a series of rules for choosing the best combination of these three elements given a global objective.

4.1. Selecting the synchronisation metric

The presence of functional relationships between pairs of nodes (*e.g.* of sensors in brain networks) is usually assessed through statistical relations between the corresponding time series. Two problems emerge. First, interpreting the meaning of links, *i.e.* what aspect of the relationship each metric is assessing; and second, electing the most appropriate metric out of the great number of available ones. In neuroscience, this is further complicated by the dynamical nature of brain functional links. Observed activity can be regarded as a sequence of waxing and waning synchronisation and desynchronisation episodes between distant brain regions. Synchronisation facilitates integrative functions, by transiently binding together spatially distributed neural populations, while desynchronisation may allow the brain to flexibly switch from one coherent state to another [196, 42]. In brain functional networks, links are thus unstable objects, whose existence is bounded to specific time frames.

In the last decades, this problem has been tackled by means of different statistical approaches, generally based on assessing the differences between the networks obtained through different metrics. For instance, Directed Transfer Function (DTF), its modification known as direct DTF (dDTF) and the Partial Directed Coherence (PDC) have been compared using EEG recordings [197]; unpartialised and partialised cross-correlations (the second trying to eliminate indirect relationships) [198]; and EEG recordings have been analysed through 9 bivariate and multivariate synchronisation measures, the latter directly ascribing a single value to the synchronisation in a group [199]. In all these cases, differences between networks are studied through p-value analyses, and results indicate that different metrics lead to important differences in the final network structure. An isolated exception is represented by Ref. [200], in which different metrics (non-linear interdependences, phase-synchronizations, Mutual Information, cross-correlation and the coherence function) appeared to detect similar phenomena in EEG recordings.

A step ahead in complexity is provided by those analyses that try to understand how some synchronisation metrics are able to discriminate between different groups of subjects. This has initially been performed through statistical tests. For instance, Ref. [201] characterises the differences between different stages of human sleep, based on scalp EEG and using Synchronisation Likelihood [202] and a non-linear interdependence measure [200]. Differences between sleep stages were assessed through the Variation of Information information criterion [203]. Ref. [204] compared the magnitude square coherence (MSC), cross-correlation (CCOR), normalised Mutual Information (NMI) and normalised joint entropy (NJE) when different stimuli were presented to a group of subjects, ranging in pleasure (valence) and arousal. Finally, Ref. [205] compares four metrics, *i.e.* Pearson product moment correlation, Spearman rank order correlation, Kendall rank order correlation and Mutual Information, in eyes open and closed EEG recordings. It is worth noticing that this last work explicitly studies the ability of each metric to discriminate between conditions, even if no data mining tools were considered.

In recent years, classification models have been included in this task, as for instance in Ref. [206], to discriminate between control and Alzheimer's patients, both in EEG and MEG recordings. Five metrics were considered: approximate entropy, sample entropy, multiscale entropy, auto-Mutual Information and Lempel–Ziv complexity. The ability to discriminate AD patients from control subjects was evaluated using receiver operating characteristic (ROC) curves. The first two metrics resulted to be the most reliable for classification, being significantly lower in the EEG of AD patients. Furthermore, the sample entropy appeared to be less dependent on the signal length and more consistent across different parameters, confirming its usefulness in this classification task. Additionally, it is worth citing Ref. [207], in which a classification was performed between resting states and video watching tasks. Functional connectivities were measured via the Pearson correlation (PeCo), partial correlation (PaCo), Mutual Information (MI), and wavelet transform

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

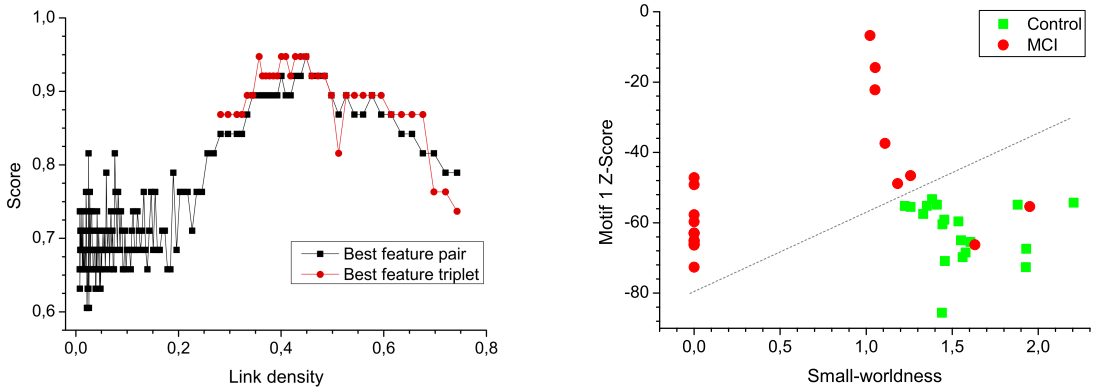


Figure 10: Optimising the network binarization process using classification scores, in a task of differentiating between control subjects and Mild Cognitive Impairment (MCI) patients. (Left) Best classification score as a function of the link density; black (red) points indicate the score obtained using pairs (triplets) of network features. (Right) Best classification of MCI and control subjects. Reprinted figure with permission from Ref. [161].

coherence (WTC). WTC yielded the best classification performance, suggesting that WTC is a preferable functional connectivity metric for functional brain network study.

4.2. Selecting the threshold

Once a synchronisation metric has been selected, a pairwise analysis of the node time series yields a set of weighted cliques, *i.e.* fully connected graphs whose weights correspond to the synchronisation strength detected between the corresponding time series. The direct analysis of such weighted cliques would *prima facie* represent the best solution, as they codify all the available information about the system. Alternatively, one may binarise them, *i.e.* pruning the graphs according to some rule applied to link weights, and assigning unitary weights to surviving links.

While pruning always implies some information loss, it also yields important advantages. Let us consider the creation of brain functional networks - even though similar arguments would hold in other contexts. First of all, brain networks are expected to be naturally sparse, as increasing the connectivity would imply a higher physiological cost. Most topological metrics available in network theory have originally been developed for unweighted graphs, for only subsequently been adapted to weighted ones; thus, in the former case the researcher has at his / her disposal a larger set of (better validated) tools. Furthermore, low synchronisation values may just be the result of statistical fluctuations, such that lowly weighted links may be present just because of noise; in such cases, deleting such noisy information can only improve the understanding of the system. Pruning can also help deleting indirect, second order correlations, which do not represent direct dynamical relationships. Finally, network analysis benefits from their graphical representation, which is meaningful only when depicting sparse structures.

The rules used to binarise networks can roughly be classified into two families: those that are based on statistical arguments, and those which rely on data mining tools. Of the first family, it is worth considering the deletion of all links whose weights are below a given fix threshold; maintains a given fraction of the strongest links; maintain a fixed node degree; or use a statistical test to select those that are significant - see Ref. [208] for a discussion on the pros and cons of each one of them. Data mining provides two additional options, which we are going to review in the remainder of the section: use a classification problem as a criterion for the binarisation, and combine multiple binarisations into a single representation.

The first option does not aim at changing the way networks are binarised, *i.e.* the process is still performed by fixing either a threshold or a link density; it does use data mining to define a criterion for optimising the process. Ref. [161] proposes the use of a classification task to define the best threshold, and specifically to consider the score obtained in the task as a proxy for the significance of the binarised networks. The process works the following way. First, binarised networks are created using different thresholds,

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

covering the whole spectrum (from very dense networks, to sparse ones); second, these networks are used in a classification problem: the threshold corresponding to the highest classification score will be the optimal one, as it allows to extract the most information from the original system. The results of this process are represented in Fig. 10. While this method yields an optimal criterion, it has two important drawbacks: it is computationally costly, as many sets of networks have to be processed; and it yields a single threshold, so that information across different densities cannot be integrated.

This second problem has been tackled in Ref. [209]. Authors propose to split the analysis of a set of networks in different steps. First, as in the method of Ref. [161], multiple thresholds are applied, to generate connectivity networks encoding different levels of details. Afterwards, each set of resulting networks are analysed, and their most relevant features selected by using a recursive feature elimination method based on graph kernel [210, 211] and support vector machine. Lastly, a multikernel SVM is used to fuse all features from multiple thresholded networks, to obtain a global classification model. It is worth stressing the advantage of this method; by using a multiple-kernel learning technique, network properties from multiple thresholded networks are merged together. This allows, in principle, to simultaneously account for different topological metrics at different connectivity levels, *e.g.* motifs in dense networks and efficiency in sparse ones; and to describe the multiscale complexity observed in many real-world systems [212, 213].

4.3. Selecting the topological metrics

Once a set of binary networks has been obtained, the next and final step involves the extraction of some topological metrics, which can ultimately be used *e.g.* to feed a classification task. Most studies in which an automatic selection of topological metrics has been proposed, this was done by using a combination of two techniques: support vector machines (SVM) and Fisher's criterion. The former allows to treat the selection process as a classification task: different sets of features can be tested, and the one yielding the best separation (in terms of classification score between two groups of subjects) is selected. The latter is similar in nature, although it does not explicitly rely on a classification task: if finds a linear combination of features that maximises the distance between the means of the two classes, while minimising the variance within each class.

Both methods have been used to optimise the network analysis of different types of neurological signals. For instance, MEG time series were analysed in Ref. [159] through Granger Causality, to discriminate autistic and control subjects, reaching a 87.5% accuracy. Two strategies were compared: in one case using the adjacency matrix alone as input feature; in the second, adding to the adjacency matrix different topological metrics. Results indicate that, while both strategies reach the same classification score, the latter is more stable with respect to the number of features included - see Fig. 11. Additionally, Ref. [160] dealt with mild traumatic brain injury (mTBI) patients, also through a Granger causality approach; and Ref. [161] focused on the diagnosis of Mild Cognitive Impairment. Structural brain connectivity has been tackled by means of functional magnetic resonance imaging (fMRI), in patients with Alzheimer's disease [133]; whole-brain tractography, to identify graph metrics that best differentiate individuals with Major Depressive Disorder (MDD) from nondepressed controls [141]; and diffusion weighted magnetic resonance imaging (DWI), to classify brain connectivity networks based on sex and kinship [214].

As we end this discussion it is once again useful to recall the work performed in Ref. [209]. By using a multikernel SVM, this paper demonstrates that the selection of topological features and that of thresholds can be carried out simultaneously. The importance of this solution resides in the fact that it tackles all the circle represented in Fig. 2 simultaneously.

4.4. Hands-on: improving significance

Let us now go back to the hands-on exercise we started in Sections 2.5 and 3.6, and see how data mining can further be applied to increase our understanding of the system. We shall focus on two (critical) aspects: the selection of an optimal threshold and of the best synchronisation metric (in this case the choice is between correlation and Granger Causality).

As discussed in Section 4.2, one of the simplest way to optimise the threshold involves extracting networks for different threshold values, and then compare the associated classification score. Fig. 12 Left presents

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

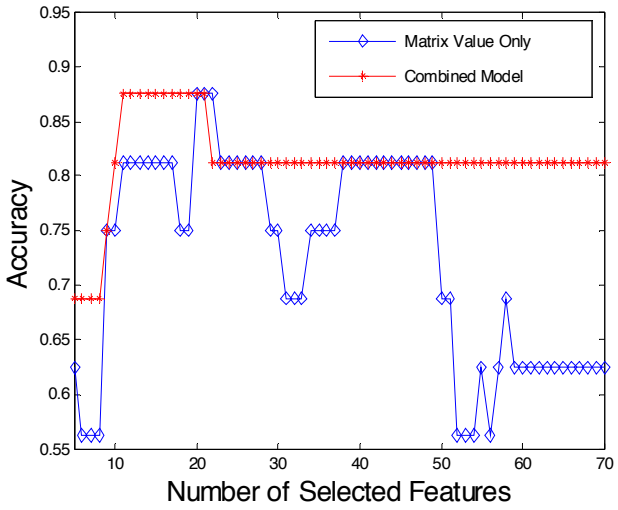


Figure 11: Classification scores obtained in a task of discriminating between autistic and control subjects, as a function of the number of features included in the analysis. The blue line represents the scores obtained by directly using the adjacency matrix as input feature, while the red one corresponds to the use of network topological metrics. Reprinted figure with permission from Ref. [159].

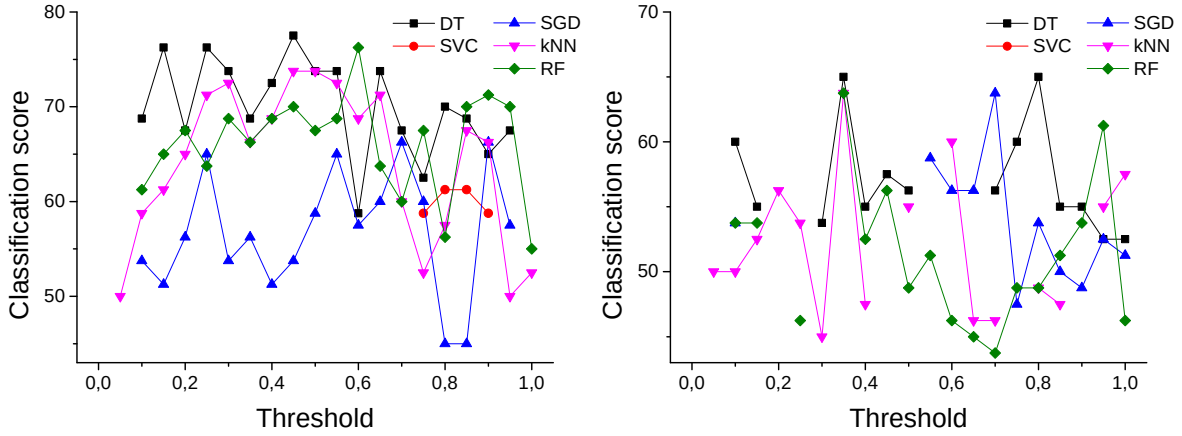


Figure 12: Classification of the data obtained in Section 2.5, as a function of the threshold applied. Left and right panels respectively represent the results for correlation and Granger Causality networks.

the results of this process, for the five classification models here considered (Decision Tree, SVM, Stochastic Gradient Descent, kNN and Random Forest). The maximum of the score is obtained with a Decision Tree, for a threshold of 0.45. Two conclusions can be drawn at this stage: *i*) Decision Tree is the most efficient classification model, for almost any value of the threshold; and *ii*) the best score (77.5%) is much higher than the one previously obtained (63%, threshold of 0.75).

Moving to the results for the Granger Causality networks (Fig. 12 Right), here too we observe a maximum score corresponding to Decision Tree, and for a threshold quite close to the previous one (0.35). On the other hand, it is clear that scores obtained with the Granger Causality are significantly lower than the ones obtained with linear correlation. For some reasons, the former is describing a type of information propagation that is not relevant for the problem at hand (distinguishing control *vs.* alcoholic patients).

In order to increase our understanding of the system, it would be necessary to understand the reasons behind such results. Why does a low threshold, and thus a high link density, better describe the two

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

groups of people? What kind of information transfer is assessed by linear correlation, but not by Granger Causality? Why does Decision Tree systematically outperform other algorithms, which are in principle more advanced? Answering these questions is out of the scope of this review, but will certainly be of interest to all neuroscientists.

5. Feature selection

One of the typical obstacles in the learning phase of any data mining procedure is represented by the high dimensionality of the feature space, even when the model only depends on a reduced number of variables. Indeed, both theoretical and experimental studies indicate that many algorithms scale poorly when a large number of irrelevant features are included, even when many data mining algorithms attempt to automatically discriminate between important features and those that can be eliminated [215, 94]. The goal of feature selection procedures is threefold: reducing the sample size, and thus the computational cost of the analysis; focusing the analysis only on relevant data, thus improving its statistical relevance; and improving 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, *e.g.* gene and protein expressions, and MEG and EEG recordings.

This Section explores how data mining algorithms can be used to reduce the dimensionality of the problem, which, in the context of network analysis, is tantamount to reducing the set of nodes and links. Finally, Section 5.3 will explore the other side of the coin, *i.e.* how complex networks can help performing a feature selection of an arbitrary data set.

5.1. Selecting the nodes of a network

A statistical physics approach to complex networks would suggest that the understanding of a network should improve with the number of nodes included in the analysis, as all nodes are *prima facie* equivalent; thus ideally no node should be discarded. Nevertheless, there exist situations in which performing a selection can be useful. We explore three different applications: neuroscience, omics analysis, and other complex systems.

Neuroscience

Probably the clearest example of the need for node selection is represented by the analysis of fMRI data. The output of an fMRI analysis is a set of images, representing the time evolution of the regional cerebral blood flow (and thus, indirectly, of neural activity). In modern equipment, the number of such regions (also called *voxels*) can be extremely large, of the order of 10^5. Identifying each voxel with a network node supposes two main problems. First, the resulting networks would be extremely large, and thus difficult to analyse: 10^5 nodes imply reconstructing adjacency matrices of 10^{10} elements; representing such networks would require in the order of 10 GB of memory. Second, assessing the synchronisation between each pair of nodes results in numerous statistical tests, which must be appropriately corrected for multiple comparisons in order to avoid a *curse of dimensionality* problem [93].

Two different options have historically been explored to appropriately reduce the number of nodes [216]. The first uses anatomical knowledge to divide the brain into regions of contiguous voxels, which are known to mediate the same cognitive function [217, 218]. Yet, it was shown that selecting regions can yield network properties different from the ones that would be obtained without grouping voxels. Grouping tends to underestimate node connectivity, but overestimates clustering degree; node centralities may also be substantially affected. Although many qualitative network properties are consistent, the voxel-level resolution reveals details that are not visible at the regional-level networks, providing extra information that can be useful in some applications [219]. When such information is not available, or one wants to avoid biases introduced by a parcellation, a data-driven procedure can be used. Voxels are then joined together according to their synchronous activity, by means of techniques like Spatial Independent Component analysis (ICA) [220] or clustering [221, 222].

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Another approach provided by data mining consists in the use of Recursive Feature Elimination (RFE), as described in Section 2.2. This has been used, for instance, to classify functional imaging data using a Support Vector Machine classifier [223]. Forward strategies have also been used - see for instance Ref. [224], in which a Mutual Information is used to assess the redundancy of features [225]. A recent study proposed the use of regularisation trees for fMRI node selection [226], based on constructing trees penalising using a variable similar to the variables selected at previous tree nodes for splitting the current node [227]. This method has the advantage of providing a multi-scale representation of features, which allows selecting brain regions simultaneously at different scales.

Beyond the problems of computational cost and data dimensionality, applying a node selection procedure is expected to increase the reliability and precision of the subsequent classification tasks. While some studies claim an important increase in the classification score [223], the debate is still open regarding the usefulness of a data mining approach against the use of prior knowledge of where to expect signs of a given disease [228]. In general, it seems that node selection is especially useful in cases of small sample sizes.

When reconstructing complex networks of brain activity using other signals, as for instance EEG and MEG, the problem of the large number of channels is less relevant - usually EEG and MEG are composed of between 60 and 250 channels. Some specific applications may nevertheless benefit from a reduction in the number of nodes. This is the case of Brain-Computer Interface (MCI), aimed at allowing severely motor-impaired patients to communicate and interact with a computer [229, 230] - see Section 3.4. If the BCI interface is to be embedded into a portable system, even creating and analysing networks of hundreds of nodes in real time can be a technical challenge, and would, as a minimum, increase the cost of the device. To reduce the complexity of the analysis, Ref. [231] proposed the use of an Evolutionary Algorithm (EA) for selecting relevant subsets of MEG channels, in a classification task aimed at assessing whether the subjects were paying attention to the left or the right. Similarly to the fMRI case, such selection may lead to a distortion in the observed topological features. A full analysis of the effects derived from sampling a subset of nodes (and sensors) in EEG studies can be found in Ref. [232]. It is based in the comparison of the topological metrics of EEG functional networks of 32 normal subjects, when the number of nodes is changed. Larger networks appear to have higher efficiency, higher assortativity and lower modularity, compared to those with same density but different size, highlighting the importance of taking into account the network size when comparing networks across studies.

Omics analysis

The analysis of genomics, proteomics and metabolomics data has yielded promising results in the last years, especially in cancer biology; nevertheless, this has come at the cost of the analysis of high-dimensional data spaces [233] - see Section 3.5. Starting from the hypothesis that the functional state of an organism is determined by its genes expression patterns, functional genetic networks can be created starting from the expression level of each gene and detected pairwise correlations or causalities. The same can also be performed with the other omics data. If no pre-processing step is performed, the result may be a very large network, with tens of thousands of nodes, only a fractions of them being relevant for the study at hand.

Examples of the application of a feature selection algorithm for reducing the number of nodes can be found in Ref. [234], in which a Sequential Backward Selection (SBS) is used: it starts with the full feature set and eliminates the least relevant feature according to the criterion function (top-down approach), repeating this process until a stop condition is satisfied [235]. A similar approach has also been explored in Ref. [192].

Furthermore, it has been proposed that omics data can be analysed in their raw format, *i.e.* starting from spectral information that can be extracted from the analysis of cell or blood samples, and reconstructing a complex network representing the relationships encountered between pairs of spectral frequencies. When compared with the standard binning procedure, in which the full spectrum is divided into a number of non-overlapping windows [236], the use of a Mutual Information criterion substantially improves the usefulness of resulting networks [193].

An alternative approach to feature selection consists in identifying clusters of omics elements that share a similar dynamics; and then map a few cluster-representative elements, or the mean expression level of all elements in a cluster, into a new node. Several algorithms can be used towards this aim, the one most used being k-means and fuzzy c-means [237, 238]. A special note should be devoted to hierarchical clustering

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

techniques: in spite of yielding, in principle, some advantages (like the possibility of choosing the scale of the partition), and their wide use, they also impose artificial restrictions on the results (nodes joined in one cluster cannot be separated and assigned to different clusters in later steps) and perform comparatively poorly [239].

Other complex systems

In recent years, complex networks have found an important field of application in the analysis of climate data, *i.e.* in the identification of patterns in the global climate. This approach is important as it enables a deeper understanding of the complex processes underlying observed phenomena, with important consequences in prediction tasks. Thanks to the high resolution recordings of weather data, and to reanalysis models, it is possible to construct networks with tens of thousands of nodes, pairwise connected when some functional relationship is detected between the corresponding dynamics [240]. While it is technically feasible to analyse the whole network, it may be interesting to cluster regions, *i.e.* join together all nodes with a similar dynamics (even if they are not spatially contiguous), in order to simplify the final results representation. This has been performed by means of several algorithms, *e.g.* k-means [241] and a Shared-Nearest Neighbor approach [242]. It has also been proposed to use the complex network topology towards this clustering task [243, 244] - a topic that will be further discussed in Section 6.1.

When complex networks are applied to the study of transportation systems, as *e.g.* the study of the air transportation network [26], a node selection problem arises. Even if we are studying a specific region, *e.g.* Europe or the USA, the airports (and the flights) in that region are connected to external elements: the problem is thus to select those elements that, even being outside the region of interest, are relevant to understand a given problem. An example of such analysis is presented in Refs. [245, 246], where networks of aircraft are reconstructed to predict the appearance of safety-related events, and relevant aircraft are selected through a wrapper strategy.

5.2. Selecting the links of a network

Once a network has been reconstructed, one may face the problem of selecting a subset of links that are representative of the network structure. There are several reasons for wanting to perform such sampling. First, networks reflect the complexity of the systems they represent, and it is thus easy to obtain intricate interweaved and densely connected structures. Ways of singling out key information by filtering irrelevant links are essential to obtain simpler relevant subgraphs. Second, in the case of functional networks, links may represent both direct and indirect paths: it is then necessary to isolate the link between $X \to Z$ when the true interaction is suspected to be $X \to Y \to Z$, as the former confounds our ability to identify true pairwise interactions. Notice that this is not equivalent to the problem of selecting a suitable threshold, as discussed in Sec. 4.2; here we tackle the problem of extracting a more meaningful representation of an already optimised network, or of a network which has a clear topology - *e.g.* the traffic backbone of a large-scale transportation infrastructure. On the other hand, this is partly related to the problem of link prediction, which, as will be discussed in Sec. 6.2, can be also seen as a problem of filtering non-relevant links. In the last years, several techniques for link selection have been proposed. We review the most important ones, which we organise into two groups depending on their main assumptions: network- *vs.* data mining-based.

Network-based methods

One of the simplest solutions to the link selection problem is provided by Minimum Spanning Trees (MST) [247, 248]. A classical concept of graph theory, a MST is the shortest-length tree subgraph that contains all the nodes of the original network. In simple words, they allow to find the minimum set of links connecting a network, or its *superhighways*. Thanks to their simplicity, they have been largely used in biomedicine [249, 250] and finance [251, 252, 253]. They yet present one big limitation, *i.e.* that are by construction acyclic. This means that all local cycles are destroyed, thus distorting the clustering coefficient and the clustering hierarchies present in the starting network. An evolution of the MST concept was discussed in Ref. [254]. This paper proposed the construction of graphs with different degrees of complexity

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

by linking the most strongly connected nodes, as in MST; the difference is that the resulting graphs should be embeddable on a surface of a given genus, *i.e.* the topologically invariant property of a surface defined as the largest number of nonisotopic simple closed curves that can be drawn on the surface without separating it [255]. Resulting graphs have the same hierarchical tree structure of MSTs, but contain a larger amount of information.

If the network of interest has been created from time series, like in the case of functional networks, one may be interested in identifying those links that are the results of robust statistical properties of the time series, and not of simple fluctuations. This can be performed by calculating surrogate networks [256]: shuffle the data such that they preserve all the statistical quantities, *i.e.* distribution of values and autocorrelation properties, and reconstruct different functional networks. These network properties are only due to the conserved statistical quantities, and therefore represent the spurious links in the original network. A complementary approach was proposed in Ref. [257], involving the local identification of the statistically relevant weight heterogeneities; this allows to filter out the backbone of dominant connections in weighted networks with strong disorder, preserving structural properties and scale hierarchies.

Finally, it is worth citing the method proposed in Ref. [258]. By exploiting the properties of dynamical correlations in networks, this method takes a correlation matrix and uses a matrix transformation to turn it into a highly discriminative silenced matrix, which enhances only the terms associated with direct causal links. Mathematically, let us denote the original observed adjacency matrix by G; the silenced adjacency matrix is then obtained as:

$$S = (G - I + D((G - I)G))G^{-1}, \tag{9}$$

$D(M)$ being the result of setting the off-diagonal terms of M to zero. This transformation thus detects the indirect paths between then nodes of the network and silences them, keeping only the contribution of the direct paths - see Fig. 13 for a graphical representation of the method, and specifically panels *d* - *f* for the silencing part. When tested against the *Escherichia coli* regulatory interaction network, it yields a 50% predictive improvement over traditional correlation measures.

Data mining-based methods

Most of the data mining-based methods for link selection have been developed around two ideas: a recursive feature selection algorithms applied on links; and a metric for deciding the relevance of each link. For the latter, Mutual Information (MI) has demonstrated to be a successful option: while some implementations are reviewed in what follows, the reader may refer to Ref. [259] for a more detailed review and comparison of different algorithms.

One of the first MI-based algorithms to be developed, ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) [260] aims at detecting and eliminating indirect interactions. It is based on the analysis of each triplet of nodes, in order to eliminate the edge with lowest MI (the indirect one) in cases where the difference between the lowest and the second lowest MI is greater than a given threshold. MINET [261] is based on the Maximum Relevance / Minimum Redundancy (MRMR) strategy for feature selections [262]. Starting with a partial set of possible links, new ones are iteratively added, chosen in order to simultaneously maximise their relevance (in terms of their MI with a target external feature) and minimise their redundancy (the MI with the other links on the partial set). SFFS+MCE is a method based on the Sequential Floating Forward Selection feature selection strategy, a wrapper approach that includes features according to some criteria. In the case of SFFS+MCE, the criterion is a conditional entropy that penalises rarely observed links [263]. Finally, MIDER [264] is a method for distinguishing between direct and indirect interactions and to assign directionality; given a set of time-series describing quantitative features of the network nodes, this is accomplished by applying several definitions and normalisation of MI.

Many important link selection methods have been developed with Protein-Protein Interaction (PPIs) networks in mind. PPIs are of special interest for understanding most of the cellular functions of the cell; nevertheless, experimentally detected PPIs have high levels of false positive rate, requiring computational tools for pruning low-confidence networks. Ref. [265] proposes a new geometric approach, called Leave-One-Out Logistic Metric Embedding (LOO-LME), for assessing the reliability of interactions. It first transforms the learning task into an equivalent discriminant form, for then directly dealing with the uncertainty in

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

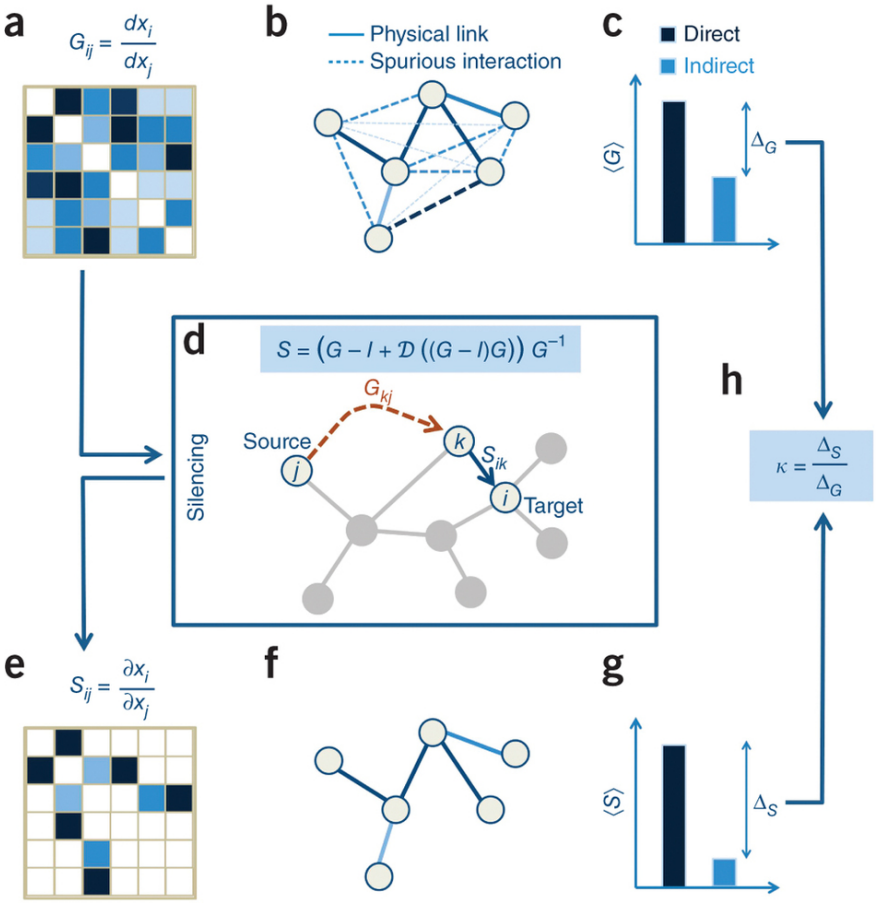


Figure 13: Graphical representation of the method for silencing indirect links in a network. It starts from the observed network G, which accounts for direct as well as indirect correlations (see panel *b*). Silencing is then achieved through the transformation reported in panel *d*, which results in a silenced matrix S (panels *e* and *f*) whose nonzero elements correspond to direct links. Reprinted figure with permission from Ref. [258].

the networks using a leave-one-out-style approach. Experimental results show that LOO-LME substantially outperforms previous methods on PPI assessment problems. On the similar line, Ref. [266] proposes the use of a fast isometric feature mapping (fast-ISOMAP), transforming a PPI network into a low dimensional metric space, thus reinterpreting the problem of assessing protein interactions into the one of measuring similarity between points of its metric space. A link reliability index is then created, based on the similarity between the corresponding nodes (points in the embedding space). Validation results for yeast data demonstrate that the interactions ranked top by this method have high functional homogeneity and localisation coherence.

A rather different approach is presented in Ref. [267], by leveraging on the Boltzmann's concept of entropy maximisation as a way of performing statistical inference with minimal reliance on the form of missing information. The underlying rationale is that each macroscopically observable state of a complex system corresponds to a number of microscopic states. Because the number of ways of realising a given macroscopic state can vary, the most likely state of the system as a whole is the one that corresponds to the largest number of microscopic states. Translated to the problem of filtering network links, entropy maximisation involves the identification of those links that increase the probability of giving rise to the macrostate captured in the observed data. As an illustration, the method is applied to *Saccharomyces cerevisiae* genetic microarray data exhibiting energy metabolic oscillations, for the reconstruction of intracellular communication pathways

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

associated with metabolic oscillations.

5.3. Feature selection through complex networks

In this last subsection, we reverse the approach to the problem: instead of analysing how feature selection techniques can be used in studying complex networks, in this section we sketch how complex networks could be used to improve current feature selection techniques.

One of the most recent trends in feature selection is the so-called Structure Learning [268, 269], which is based on the idea of finding all possible relationships between pairs of variables, and expressing them by mean of a graph. The attentive reader will notice the parallelism between this approach, and the reconstruction of functional brain networks discussed in previous sections. To simplify the analysis, it is usually assumed that data are generated by a Bayesian Network, so that the resulting structure is a directed graphical model. Following this hypothesis, the optimal solution is the Markov blanket of the target node, which is unique in a Bayesian Network. A Markov blanket of a given node is composed of its parents, its children, and its children's other parents; in terms of data relationships, this is equivalent to saying that, in order to understand a given feature, it is necessary to include its direct causes, its direct effects, and the direct causes of its direct effects (also called the "spouses" of the target feature) [270]. It is clear that an essential issue is the correct detection of the relationships between the features; using linear methods, such as SVMs, does not provide reliable results, as linear SVM may give zero weight to important features [271, 272], and it is thus necessary to resort to causality analyses - see for instance Refs. [273, 274].

Once the graph of relationships has been established, complex networks can help identifying the most relevant nodes (*i.e.* features) in an efficient way, without any prior hypotheses on the structure of the network. This can be accomplished in at least two ways. First, by ranking nodes according to their importance, using one of the numerous available centrality metrics [11]; second, by considering that identifying the causal nodes in a network is tantamount to establishing which nodes are controlling its dynamics, thus converting the feature selection into a controllability problem [275, 276]. It is worth noticing two advantages that this approach would yield with respect to the Markov blanket one. First, there is no need to define *a priori* the target node(s), as complex network analysis would provide a global ranking of nodes importance. Second, any complex network analysis takes into account the global structure of the network, thus including structural information that may be missing in a local Markov analysis.

5.4. Hands-on: feature selection

As a final task in this hands-on exercise, we explore how data mining can be used to select a sub-set of significant nodes within the network. While EEG data set is relatively low-dimensional, stemming from only 19 electrodes, further reducing dimensionality may be useful to gain more knowledge on the disease under study.

To do so, we use a *greedy algorithm*. The full network is considered, and nodes are deleted one at the time - thus yielding 19 sets of networks, each one with 18 nodes. A classification score is calculated for each set, and the one yielding the best classification is passed to the following phase. Next, nodes are deleted from this reduced networks, that is, nodes and all associated links are excluded from further analyses; and a new set of scores is calculated - and so forth, until further reduction would make the network too small to be of interest.

Fig. 14 Top Left depicts the results, in terms of classification scores, for the first phase - *i.e.* when only one node is deleted. It can be seen that, in general, the score does not significantly decrease, and even increases when the third node is deleted. This may be surprising, but just confirms what has been discussed throughout this section: some nodes and links may just represent noise, for not being related with the pathology under study, and their deletion only improves the analysis by focusing it to the important elements. Fig. 14 Top Right further depicts the results of the whole process: note how deleting up to ten nodes does not affect our classification capabilities (the score is still above 80%), which in turns illustrates the fact that the nine surviving nodes are the truly important ones to understand the pathology (Fig. 14 Bottom).

In summary, a complex network analysis with arbitrary parameters seems to indicate some interesting differences between control subjects and patients of our EEG data set. Nevertheless, a classification task of

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

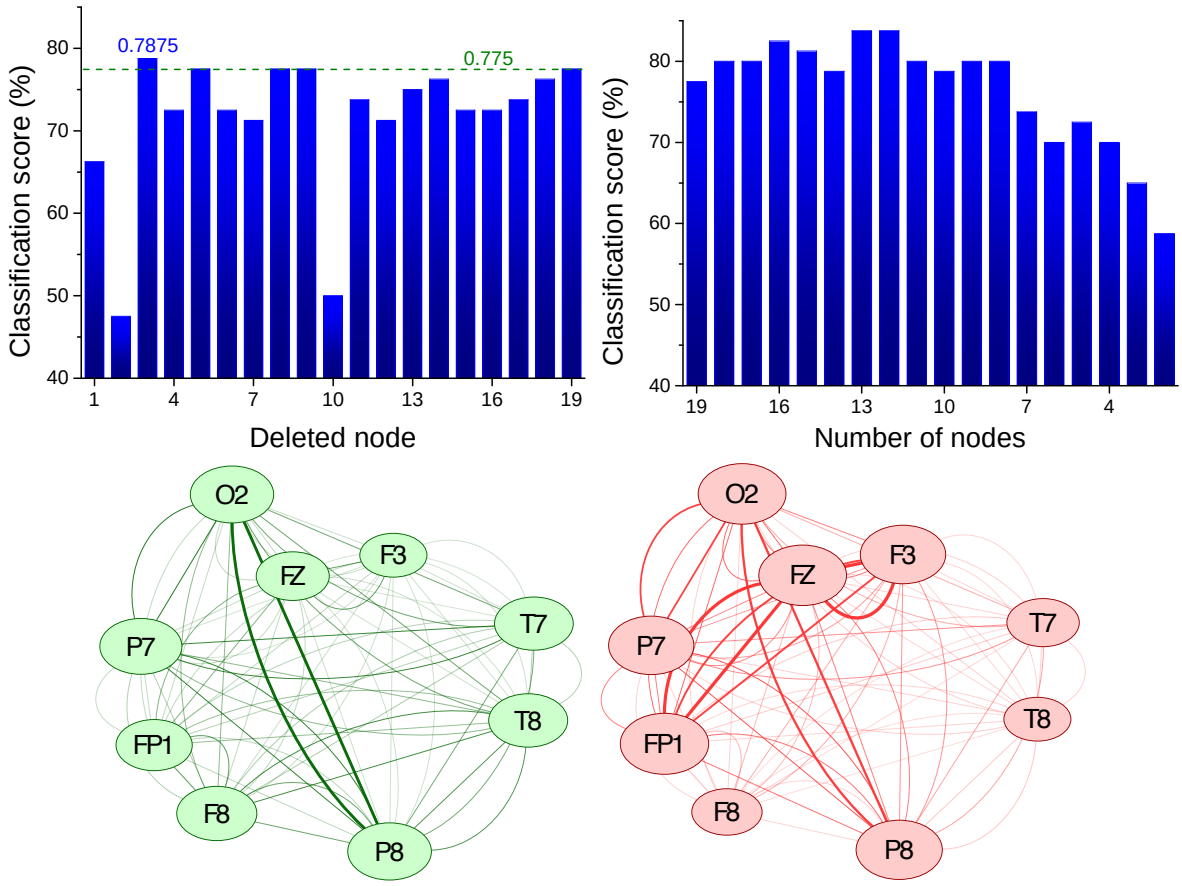


Figure 14: Discarding nodes from the networks. (Top Left) Classification score as a function of the node discarded. The dashed horizontal line represents the best classification score with the complete network (77.5%). (Top Right) Classification score as a function of the number of surviving nodes. (Bottom) The networks of Fig. 5 when only the 9 most important nodes are included.

these topological features yielded a rather low score ($\approx 60\%$). This suggests that these differences were not as important as initially thought, possibly because the analysis (and specifically, its parameters) was not properly tuned. The same data mining techniques were the instrument to increase the significance of the networks, and to point us towards the synchronisation metrics and brain regions most relevant. The upshot is that we end up with higher prognostic capabilities and better understanding of the pathology at hand, when compared with what would be obtained with an unguided complex network analysis.

6. Additional bridges

In the three previous sections we have seen numerous examples of how data mining and complex network theory can be used in a synergistic way. The discussion has been organised around three of the most important tasks in data mining, namely classification, validation and feature selection. Nevertheless, this is clearly not the whole picture. This section presents a selection of four additional problems, in which bridges between the two communities have been successfully created.

6.1. Clustering and community detection

Clustering in data mining and community detection in complex networks analysis are two tasks that present many similarities. Clustering is defined as the task of grouping a set of objects in such a way

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

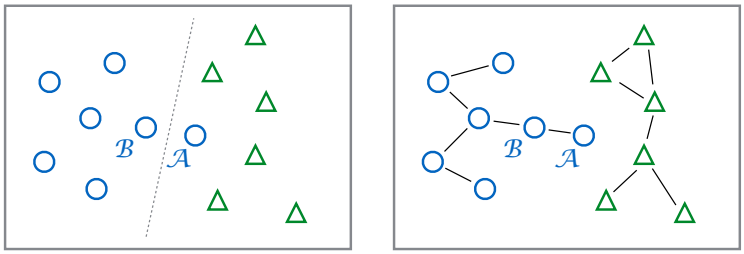


Figure 15: The clustering / community detection problem (Left) Clustering by means of a centroid model; object A is wrongly assign to the right cluster, due to the non-convex nature of such cluster (Right) Result of a community detection process.

that objects in the same group (or cluster) are more similar to each other than to those in other groups [62, 277]. If we take a set of objects, map them into nodes of a network, and pairwise connect nodes whose corresponding objects are similar, clustering is tantamount to finding communities within the network, *i.e.* groups of nodes for which internal connections are much denser than those with the rest of the nodes in the network [278, 16]. The network approach presents some advantages.

Most clustering algorithms exploit local (*i.e.* pairwise) information about similarity. For instance, connectivity models are based on a measure of distance between groups of items; centroid models create clusters based on the distance between objects and a single mean vector; and graph-based models (an intermediate version of both problems) define clusters in which at least a fraction of the nodes are connected by an edge. While the low computational cost represents one of the main advantages of these algorithms, there are some situations in which their detection ability is severely affected. Consider, for instance, the case of an object A which is similar to another object B of its true cluster, but not to the other objects of the same cluster - in other words, B is a mediatory item between A and the cluster itself, see Fig. 15 Left. In this case, a graph-based model fails, as A has not enough connections with its true cluster; a centroid model can also yield wrong results, especially in the case of non-convex clusters [279].

A complex network approach, on the other hand, allows including the full topology of the network into the analysis - provided, of course, that a suitable similarity metrics has been defined, a problem which may be far from trivial. As depicted in Fig. 15 Right, the object A is correctly assigned to the left cluster, although it only shares one link with it, and is closer to the centre of the right cluster (note that, in general, spatial information is eliminated in a network representation). Complex networks also facilitate the detection and characterisation of communities in sets of heterogeneous elements or among which a distance measure cannot easily be assessed [280]; as well as of networks that evolve with time [281, 282]. On the negative side, community detection algorithms can fail in some scenarios, for instance in the identification of very small communities, or when one community is much smaller than the other ones [283, 16].

Overall, a complex network approach to the clustering problem has historically yielded better results than the ones afforded by classical data mining approaches. Beside those reported in Section 5.1 [243, 244], a few examples are worth mentioning. For instance, in music technology a community detection algorithm was used to detect groups of *covers* of the same song, with a result outperforming traditional clustering algorithms [284]. In econophysics [285], clusters of investors trading in a financial market are identified as communities in statistically validated networks [286] - see Fig. 16 for a graphical representation; or stocks are clustered [251] according to the correlation detected between return and volatility time series [287].

It is also worth noticing that clustering (and other data mining) algorithms have recently been used to perform a community detection task. These include fuzzy c-means [288], fuzzy clustering [289], adaptive clustering [290], or genetic algorithms [291].

6.2. Link prediction

Another problem of great interest in complex networks, especially for its biological implications, is the one of *link prediction*. All previous examples covertly assumed that the network is completely known - either because relationships are explicit, or because they could be reconstructed from functional analyses. However,

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

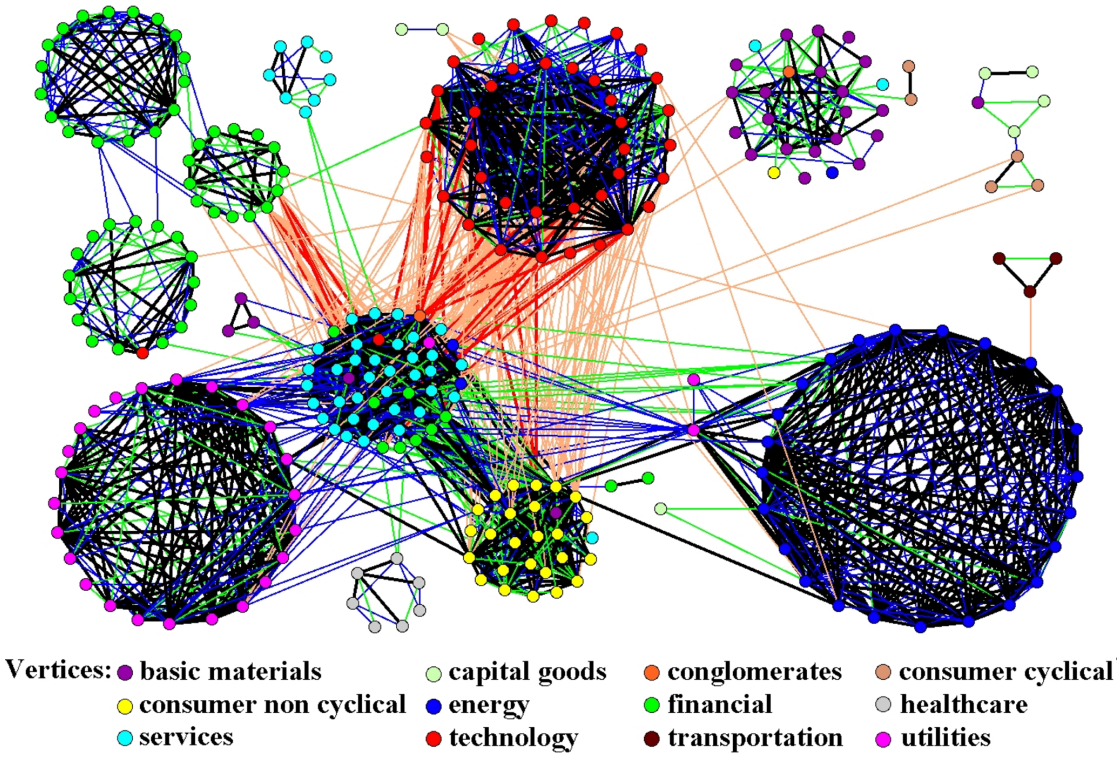


Figure 16: The largest connected component of the Bonferroni network associated with the system of 500 stocks, and its community structure. The nodes represent stocks, and links connecting different stocks correspond to statistically validated relationships. The node color identifies the economic sector of the corresponding stock. Reprinted figure with permission from Ref. [286].

in biological networks such as food webs, protein-protein or metabolic networks, assessing the presence of a link between two elements requires performing costly field and/or laboratory experiments; as a consequence, only a fraction of the total number of links is typically tested, *e.g.* 20% of the molecular interactions in yeast cells [292] and 0.3% in human ones [293] have been described. A solution to this problem involves using some kind of *link prediction* algorithm, *i.e.* an algorithm to estimate the likelihood for a link (yet undiscovered) to be present.

The problem of link prediction has a clear counterpart in the wide family of *recommendation system* algorithms developed in data mining. In its original formulation, a recommender system (also called *information collaborative filtering* system) takes as input information about past purchases (or selections) of goods by a set of users, and suggests the following best buy for each of them. If one considers a bipartite network, where nodes represent both users and goods, and links past interactions between them, a recommendation system is completely equivalent to a link prediction system acting on such network.

In spite of their substantial equivalence, link prediction in complex network theory and recommendation system in data mining have mostly walked independent paths. In the former, network theory has relied more on random walk processes and maximum likelihood methods - see Ref. [294]. In the latter, the most used approaches are based on concepts such as Markov chains and statistical models - see Refs. [295, 296] for a full overview on the topic. Only recently have both communities started joining forces and mixing both approaches [297].

What advantages does the complex network approach offer to data mining in this particular context? As in other cases, like in the clustering problem (see Section 6.1), the answer lies in the global topological perspective of complex networks. For instance, if one were to try to construct a link prediction model for the Internet, a network analysis would be useful to unveil a hierarchical (or *medusa-like*) structure, with a

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

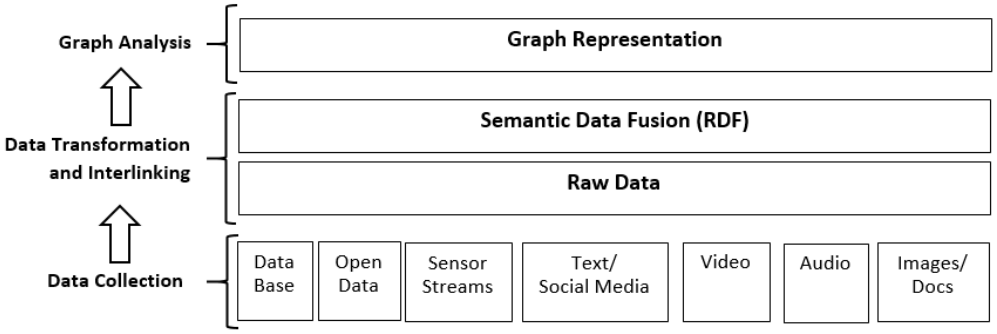


Figure 17: Conceptual representation of a graphical Big Data analysis framework. See main text for details.

strongly connected core, a peer-connected periphery, and a set of isolated nodes [298]. In this case, better performance could be obtained by taking into account the hierarchy of the network, and favouring those links that do not disrupt such a structure [299].

6.3. Evaluating and validating generative models

Suppose that one is analysing the human brain under different conditions, *e.g.* control subjects *vs.* Alzheimer's patients, and that the topologies of the corresponding functional networks have been described. A further question may be asked: can such topologies, and the evolution from the former to the latter, be explained by a set of simple rules? Knowing these rules would allow a better understanding of the disease, possibly shedding light of the underlying causes of the observed phenomena. This can be seen as a "genotype to phenotype problem", that is, the process of identifying the relationships between hidden variables (the genotype) and measured observables (the phenotype) of a system.

Generative models are simple models designed to create synthetic networks with specific topological features. Proposed models include theoretical ones, *i.e.* independent of real-world networks [6, 32, 300, 301]; and models describing brain networks: from cortical connectivity [302, 303, 304, 305], neural ensembles [306, 307], to fMRI [308, 309, 310, 132, 311] and MEG [312] functional networks.

Proposed generative models need to be optimised, that is, one ought to obtain the best set of parameters, such that the networks yielded by the model are topologically equivalent to the real ones according to some criteria. Moreover, even if the model does indeed recover some topological characteristics, this does not guarantee that the generative model itself is representative of what is really happening in the brain, *i.e.* that the model yields new knowledge of the neurological processes involved in the disease. The model must therefore be validated. The attentive reader will notice the similarities between this problem, and the one of improving network significance as tackled in Section 4.

Evaluation and generation of generative models can be achieved by using a classification task in which two conditions are compared [313, 314]. An increase in the knowledge about the conditions implies a higher capability of discriminating between both states; this, in turn, implies that the generative model should increase our capacity of discriminating between both conditions, in order to ensure an increase in knowledge. It has further been demonstrated that, to obtain such an increase in discriminatory capabilities, the relationship between the phenospace (*i.e.* the observed topological metrics) and the genospace (the model parameters) should at least be a non-monotonic function [314].

6.4. Complex networks in Big Data: semantic representations by graphs

To illustrate the possibilities offered by complex networks to simplify the analysis of Big Data, we discuss a standard approach involving the use of non-planar graphs [315]. The approach consists of a methodology for transforming raw data into linked graphical representations of them (see Fig. 17), and is composed of three main steps: Data Collection, Data Transformation and Interlinking and Graph Analysis.

In the initial phase, the user only has access to huge quantities of structured, semi-structured and unstructured data, originating from many independent sources. All these data are heterogeneous and raw,

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

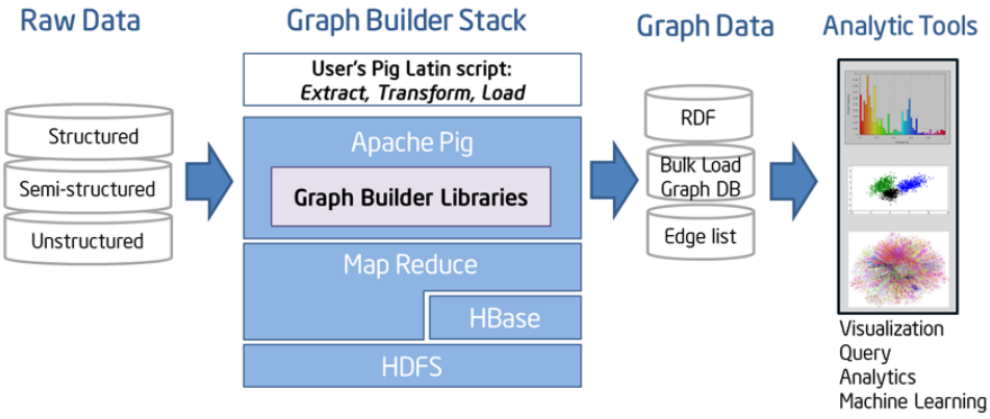


Figure 18: Representation of the software ecosystem supporting a graphical Big Data analysis framework.

and thus require some processing to be understandable. This is achieved in a second phase, termed *Data Transformation and Interlinking phase*, in which the raw data are transformed into *smarted data*. The same data are expressed in a standard format using RDF (*Resource Description Framework*, a standard defined for the web of semantic data [316]). In RDF every statement is expressed by three components (*i.e.* a *triple*): the *subject*, *i.e.* the identifier of the resource; the *predicate*, that indicates the property or attribute of the subject to be described; and the *object*, the value of the predicate related to the subject. Using this semantic data model, it is possible to capture the semantic heterogeneity and the differences in the logical representation of data. Finally, the last phase of the methodology, the *Graph Analysis phase*, deals with the creation of a graphical representation using semantic data fusion. At the end of the process, the original data are mapped into a network structure, on which all network metrics and principles can be applied to discover hidden relationships.

A more concrete example of the framework proposed in Fig. 17 is provided by the Intel Open Source Community [317]. Fig. 18 presents the schema proposed to uncover hidden correlations in large volumes of data, by means of graph analytics. Starting from structured, semi-structured, or unstructured data, the Apache Pig library is used to form RDF triples for graph elements. Afterwards, some graph visualisation tools, such as Gephi (see Section 2.1), are applied to this triples list.

A simple example of an analysis performed following the framework clarifies this issue, as depicted in Fig. 17. The example involves the representation of an anonymised real dataset including subjects with some kind of cardiovascular disease. In particular we consider subjects' BMI (Body Mass Index), following the formulas and the parameters of the Centre for Disease Control and Prevention (CDCP) [318]. Following Fig. 17, the first phase is Data Collection; note that BMI data are only part of the data set, but the only one relevant to our study. In the Data Transformation phase, raw data are transformed to an RDF triple list of the type ("Patient Group" "with BMI Value" "has Weight Status"); according to the previous definition, "Patient Group" is the object, "with BMI Value" the predicate/property and "has Weight Status" the object/value. Two examples of the elements contained in such lists are the following:

<G02 00,68727 Underweight>
<G06 18,51563 Healthy Weight>

From this triple list, and by using the in-memory graph tools like NodeXL [319], one can produce a directed labelled graph as the one shown in Fig. 19. A simple visual inspection makes it clear that most people with cardiovascular disease have some overweight problems - notice the larger degrees of the "Obese" and "Overweight" nodes. Further analyses would require to connect this with other graphs, like for instance with one related to lifestyle, to understand the incidence of overweight in cardiovascular disease.

This simple example shows how RDF, as the basic data representation language for linked information, can be used to control the data variety and the semantic heterogeneity one usually encounters in Big Data,

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

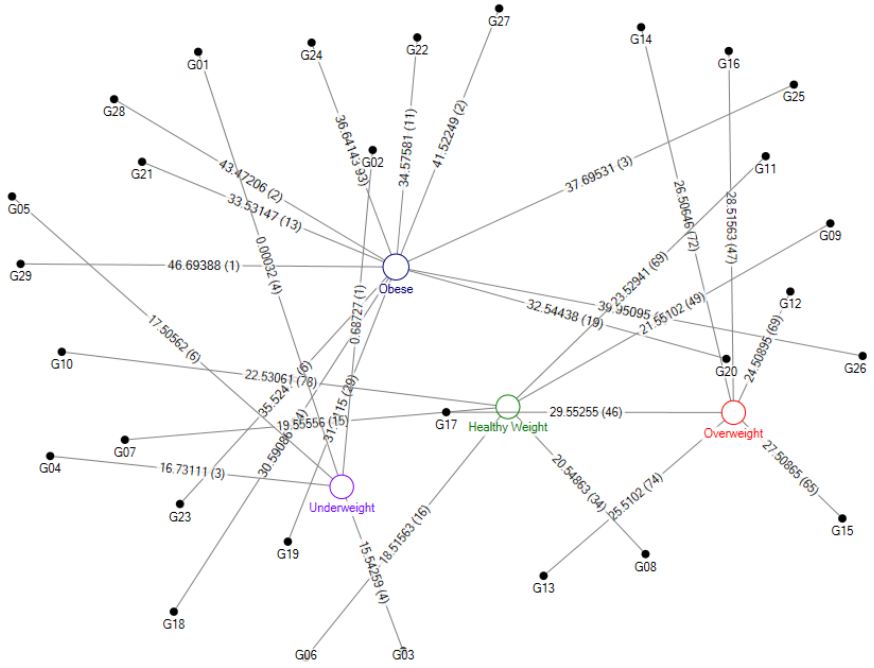


Figure 19: An example of a directed labelled graph of a graphical Big Data analysis framework. See main text for details.

and to produce a graph representation of the available information. Beyond that, there is a clear technical challenge in the system scalability, *i.e.* in the ability of handling and representing very large data sets; nevertheless, this is continuously tackled, and the availability of tools like NodeXL [319] make this problem less and less relevant in most applications.

7. Conclusions

The application of data mining concepts to the study of complex networks is generating great enthusiasm and increasing interest from several scientific communities. Data mining and network theory share the common objective of finding ways to represent complex systems, and ultimately provide a better understanding of their structure and dynamics. At the same time, they provide different view-points. Data mining focuses on the comparative analysis of different instances, to detect similarities and differences between them; complex networks allow representing the internal structure of individual instances in a simple and elegant way. That these two approaches have not yet merged is mainly to put down to the different background of researchers working in statistical physics and artificial intelligence. This problem notwithstanding, in this review we have shown that, when used in synergy these two approaches broaden the scope of complex systems that can be studied in a quantitative way, particularly in interdisciplinary biophysical contexts.

Data mining provides the complex network practitioner with numerous tools for improving the significance of representations. It can for instance be used to identify the important nodes and connections of the system, disregarding less important ones that can be thought to contribute noise. Data mining also allows creating models for discriminating between groups of instances, thus enabling diagnostic and prognostic applications. Additionally, such models can be used to objectively estimate the significance of the reconstructed networks. On the other hand, data scientists can find in network representations a new way of structuring information, providing a wealth of quantitative variables through which systems can be described. When both fields address the same basic goals, as in the case of data clustering and recommendation systems, a cross fertilisation allows coming up with innovative solutions. Finally, we foresee an increasing interest for complex network theory in the raising field of Big Data, in which relationships between data elements are of fundamental importance.

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In spite of all the achievements outlined in this review, many issues still need to be tackled by the scientific community. First, joining data mining and complex network theory is a path often plagued with numerous pitfalls, particularly when some of their respective assumptions and limitations are not properly taken into account. Second, there is a lack of a specific theoretical and numerical foundation for the combined use of concepts from both fields. In this last section, we discuss some of these topics, with the hope of fostering debate in the scientific community.

7.1. Problems to overcome

After reading this review and some of the papers cited within, the attentive reader may notice some recurring issues. The complex network practitioner with insufficient experience with data mining, may be tempted to use it as a black box: results of a data mining task are reported, without a full appraisal of the limitations they present. In what follows, we want to briefly discuss some caveats of merging complex networks and data mining, which we hope will be a guide for future studies.

Method specification. In most studies, data mining tools have by and large been considered as "utilities", used to achieve a specific goal, but not as an end in itself. As a consequence, many papers' abstract omits a description of the classification algorithm or the validation procedure used. This is especially true of biomedical and neuroscience studies, in which the focus is on the biological and clinical aspects of the analysis. It is nevertheless important to understand that data mining techniques have a significant impact in the results. For instance, different classification algorithms make different hypotheses on the patterns to be detected in data; and different validation strategies may highlight limitations in the obtained models. The choice of the algorithm is a fundamental step in the overall methodology, and should be handled as such.

One method does not fit all. Using one single data mining technique in a specific task is bound to return a partial picture, however well-established and widely used the technique. As illustrated in Section 3.3, different classification models may yield important differences in accuracy, due to their underlying hypotheses [139, 137, 143], even for similar data sets, as different problems (pathologies) may require different approximations. It is thus important to always compare different algorithms, and select the one yielding best results - possibly understanding why such differences arise.

Avoiding overfitting. In some cases, the full power of data mining is deployed in different phases of the study: for instance an initial feature selection can be performed through SVM, and the same technique used in a classification task. This may result in an *overfitting* problem: the classification is performed over data that have already been filtered using information about the real class of each instance; in other words, the classification uses data that have been selected using information about the final result. The solution may come from splitting the available data into training, validation and test sets, although a large number of instances is required (a condition not easy to fulfil in biomedical problems). In general, a special attention must be devoted to the problem of model validation - see Section 2.2.

7.2. Open lines of research

Merging data mining and complex network theory is an ongoing process, which the research community has just started dealing with. If such integration is to continue evolving, and especially if concepts such as Big Data are to be included, several technical and theoretical aspects will need to be further developed.

Merging complex networks and data mining. As we have seen throughout this review, complex networks metrics and data mining algorithms have developed and are being applied in an independent fashion. In other words, a standard analysis requires extracting some topological metrics and use them in data mining tasks, see Section 3.6: two steps that have little cross-interactions between them. This is due to the fact that data mining algorithms are not able to extract a pattern in the data that develops in the meso- or macro-scale. An important step ahead would consist in developing models including the network concept in their definition; for instance, a classification model able to internally evaluate topological metrics, like the clustering coefficient, without the need of *a priori* knowledge.

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Optimising network metrics evaluation. As data sets become larger and larger, the evaluation of complex network metrics on them is getting more complicated, especially from a computational point of view. In parallel with the advent of Big Data, the analysis of large social networks, like the one generated by on-line social systems, may be in the future called "Big Complex Networks". If this trend consolidates, complex network metrics will have to be designed with a special attention on their computational cost, in order to ensure their scalability to hundreds of million of nodes.

Software infrastructures for complex network analysis. The challenges in terms of data complexity and size created by the Big Data movement, have forced the IT community to design new software tools for managing and analysing such a large amount of information. In the same vein, important benefits would be obtained from a software infrastructure specifically designed for the analysis of large scale complex networks. While some steps have already been taken in this direction, for instance with the creation of *graph databases* [320], most solutions are limited to the traditional graph concepts, disregarding the complexity of real-world networks.

Acknowledgements

At the end of this review, we would like to acknowledge all scientists with whom we had, have, and surely will have interactions on this topic. Discussions and dialogues have been both stimulating and fruitful, giving birth to many of the ideas included in this work. In this spirit, the authors would like to thank the following colleagues: Emilio Álvarez Pereira (Team&Cloud, Spain), Salvatore Micciche' (University of Palermo, Italy), Fabrizio Lillo (Scuola Normale Superiore di Pisa, Italy), Andrew Cook (University of Westminster, UK), Alexey Zaikin (University College London, UK), José Luis Vallejo (Sngular, Spain), and the ComplexWorld team.

At the same time, we also would like to gratefully acknowledge all the discussions we had (in meetings, conferences, workshops, and personal visits) with other colleagues who are not mentioned in the above list (and, for that, we deeply apologise), which equally inspired our efforts, and opened up our minds in a way that contributed, eventually and substantially, to the realisation of the present survey.

Final note

After editing the present Report, we learned about the publication of other recent and important contributions which relate to subjects and arguments treated in this work. In the following, we provide a list of these additional References that the reader may find useful to consult.

Traxl, D., Boers, N., & Kurths, J. (2016). Deep Graphs-a general framework to represent and analyze heterogeneous complex systems across scales. arXiv preprint arXiv:1604.00971.

Thomas, J. M., Muscoloni, A., Ciucci, S., Bianconi, G., & Cannistraci, C. V. (2016). Machine learning meets network science: dimensionality reduction for fast and efficient embedding of networks in the hyperbolic space. arXiv preprint arXiv:1602.06522.

Armano, G., & Javarone, M. A. (2013). Clustering datasets by complex networks analysis. Complex Adaptive Systems Modeling, 1(1), 1-10.

Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: a survey. Data Mining and Knowledge Discovery, 29(3), 626-688.

Petri, G., Scolamiero, M., Donato, I., & Vaccarino, F. (2013). Topological strata of weighted complex networks. PloS one, 8(6), e66506.

Gauvin, L., Panisson, A., & Cattuto, C. (2014). Detecting the community structure and activity patterns of temporal networks: a non-negative tensor factorization approach. PloS one, 9(1), e86028.

Massara, G. P., Di Matteo, T., & Aste, T. (2015). Network Filtering for Big Data: Triangulated Maximally Filtered Graph. arXiv preprint arXiv:1505.02445.

Barfuss, W., Massara, G. P., Di Matteo, T., & Aste, T. (2016). Parsimonious modeling with Information Filtering Networks. arXiv preprint arXiv:1602.07349.

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

[1] S. A. Huettel, P. B. Mack, G. McCarthy, Perceiving patterns in random series: dynamic processing of sequence in prefrontal cortex, Nature neuroscience 5 (5) (2002) 485–490.

[2] S. Zeki, O. J. Hulme, B. Roulston, M. Atiyah, The encoding of temporally irregular and regular visual patterns in the human brain, PloS one 3 (5) (2008) e2180.

[3] J. Friedman, T. Hastie, R. Tibshirani, The elements of statistical learning, Vol. 1, Springer series in statistics Springer, Berlin, 2001.

[4] V. Vapnik, The nature of statistical learning theory, Springer Science & Business Media, 2013.

[5] V. Mayer-Schönberger, K. Cukier, Big data: A revolution that will transform how we live, work, and think, Houghton Mifflin Harcourt, 2013.

[6] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Reviews of modern physics 74 (1) (2002) 47.

[7] M. E. Newman, The structure and function of complex networks, SIAM review 45 (2) (2003) 167–256.

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

[9] L. d. F. Costa, O. N. Oliveira Jr, G. Travieso, F. A. Rodrigues, P. R. Villas Boas, L. Antiqueira, M. P. Viana, L. E. Correa Rocha, Analyzing and modeling real-world phenomena with complex networks: a survey of applications, Advances in Physics 60 (3) (2011) 329–412.

[10] S. Boccaletti, G. Bianconi, R. Criado, C. I. Del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendina-Nadal, Z. Wang, M. Zanin, The structure and dynamics of multilayer networks, Physics Reports 544 (1) (2014) 1–122.

[11] L. d. F. Costa, F. A. Rodrigues, G. Travieso, P. R. Villas Boas, Characterization of complex networks: A survey of measurements, Advances in Physics 56 (1) (2007) 167–242.

[12] H. Chen, U. Ikeda, M. Shimpo, K. Shimada, Direct effects of statins on cells primarily involved in atherosclerosis., Hypertension Research 23 (2) (2000) 187–192.

[13] M. E. Newman, Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality, Physical review E 64 (1) (2001) 016132.

[14] J. A. Almendral, R. Criado, I. Leyva, J. M. Buldú, I. Sendina-Nadal, Introduction to focus issue: Mesoscales in complex networks, Chaos: An Interdisciplinary Journal of Nonlinear Science 21 (1) (2011) 016101.

[15] M. E. Newman, Modularity and community structure in networks, Proceedings of the National Academy of Sciences 103 (23) (2006) 8577–8582.

[16] S. Fortunato, Community detection in graphs, Physics Reports 486 (3) (2010) 75–174.

[17] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: simple building blocks of complex networks, Science 298 (5594) (2002) 824–827.

[18] V. Latora, M. Marchiori, Efficient behavior of small-world networks, Physical review letters 87 (19) (2001) 198701.

[19] P. Crucitti, V. Latora, M. Marchiori, A. Rapisarda, Efficiency of scale-free networks: error and attack tolerance, Physica A: Statistical Mechanics and its Applications 320 (2003) 622–642.

[20] A. Rényi, P. Erdös, On random graphs, Publicationes Mathematicae 6 (290-297) (1959) 5.

[21] B. Bollobás, Random graphs, Springer, 1998.

[22] D. J. Watts, S. H. Strogatz, Collective dynamics of 'small-world' networks, nature 393 (6684) (1998) 440–442.

[23] M. D. Humphries, K. Gurney, Network 'small-world-ness': a quantitative method for determining canonical network equivalence, PloS one 3 (4) (2008) e0002051.

[24] M. Zanin, On alternative formulations of the small-world metric in complex networks, arXiv preprint arXiv:1505.03689.

[25] D. Papo, M. Zanin, J. H. Martínez, J. M. Buldú, Beware of the small-world neuroscientist!, Frontiers in Human Neuroscience 10.

[26] M. Zanin, F. Lillo, Modelling the air transport with complex networks: A short review, The European Physical Journal Special Topics 215 (1) (2013) 5–21.

[27] A.-L. Barabási, Scale-free networks: a decade and beyond, science 325 (5939) (2009) 412.

[28] A.-L. Barabási, E. Ravasz, T. Vicsek, Deterministic scale-free networks, Physica A: Statistical Mechanics and its Applications 299 (3) (2001) 559–564.

[29] S. Dorogovtsev, J. Mendes, A. Samukhin, Size-dependent degree distribution of a scale-free growing network, Physical Review E 63 (6) (2001) 062101.

[30] P. Holme, B. J. Kim, Growing scale-free networks with tunable clustering, Physical review E 65 (2) (2002) 026107.

[31] G. Caldarelli, A. Capocci, P. De Los Rios, M. A. Munoz, Scale-free networks from varying vertex intrinsic fitness, Physical review letters 89 (25) (2002) 258702.

[32] K. Klemm, V. M. Eguiluz, Highly clustered scale-free networks, Physical Review E 65 (3) (2002) 036123.

[33] K. Klemm, V. M. Eguiluz, Growing scale-free networks with small-world behavior, Physical Review E 65 (5) (2002) 057102.

[34] S. Valverde, R. F. Cancho, R. V. Sole, Scale-free networks from optimal design, EPL (Europhysics Letters) 60 (4) (2002) 512.

[35] S. Jung, S. Kim, B. Kahng, Geometric fractal growth model for scale-free networks, Physical Review E 65 (5) (2002) 056101.

[36] J. Saramäki, K. Kaski, Scale-free networks generated by random walkers, Physica A: Statistical Mechanics and its Applications 341 (2004) 80–86.

[37] M. Catanzaro, M. Boguñá, R. Pastor-Satorras, Generation of uncorrelated random scale-free networks, Physical Review E 71 (2) (2005) 027103.

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[38] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, science 286 (5439) (1999) 509–512.
[39] O. Sporns, G. Tononi, R. Kötter, The human connectome: a structural description of the human brain, PLoS Comput Biol 1 (4) (2005) e42.
[40] O. Sporns, The human connectome: a complex network, Annals of the New York Academy of Sciences 1224 (1) (2011) 109–125.
[41] E. Bullmore, O. Sporns, Complex brain networks: graph theoretical analysis of structural and functional systems, Nature Reviews Neuroscience 10 (3) (2009) 186–198.
[42] D. Papo, M. Zanin, J. M. Buldú, Reconstructing functional brain networks: have we got the basics right?, Frontiers in human neuroscience 8 (2014) 107.
[43] P. Holme, J. Saramäki, Temporal networks, Physics reports 519 (3) (2012) 97–125.
[44] P. Holme, J. Saramäki, Temporal networks, Springer, 2013.
[45] F. Vega-Redondo, Complex social networks, no. 44, Cambridge University Press, 2007.
[46] E. Goffman, Frame analysis: An essay on the organization of experience., Harvard University Press, 1974.
[47] A. Cardillo, J. Gómez-Gardenes, M. Zanin, M. Romance, D. Papo, F. del Pozo, S. Boccaletti, Emergence of network features from multiplexity, Scientific reports 3.
[48] A. Cardillo, M. Zanin, J. Gómez-Gardeñes, M. Romance, A. J. G. del Amo, S. Boccaletti, Modeling the multi-layer nature of the european air transport network: Resilience and passengers re-scheduling under random failures, The European Physical Journal Special Topics 215 (1) (2013) 23–33.
[49] J. White, E. Southgate, J. Thomson, S. Brenner, The structure of the nervous system of the nematode caenorhabditis elegans: the mind of a worm, Phil. Trans. R. Soc. Lond 314 (1986) 1–340.
[50] M. De Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. A. Porter, S. Gómez, A. Arenas, Mathematical formulation of multilayer networks, Physical Review X 3 (4) (2013) 041022.
[51] K.-M. Lee, J. Y. Kim, S. Lee, K.-I. Goh, Multiplex networks, in: Networks of networks: The last frontier of complexity, Springer, 2014, pp. 53–72.
[52] D. A. Schult, P. Swart, Exploring network structure, dynamics, and function using networkx, in: Proceedings of the 7th Python in Science Conferences (SciPy 2008), Vol. 2008, 2008, pp. 11–16.
[53] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome research 13 (11) (2003) 2498–2504.
[54] M. Bastian, S. Heymann, M. Jacomy, Gephi: an open source software for exploring and manipulating networks, ICWSM 8 (2009) 361–362.
[55] V. Batagelj, A. Mrvar, Pajek-program for large network analysis, Connections 21 (2) (1998) 47–57.
[56] Z. Hu, J. Mellor, J. Wu, C. DeLisi, Visant: an online visualization and analysis tool for biological interaction data, BMC bioinformatics 5 (1) (2004) 17.
[57] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, Advances in knowledge discovery and data mining.
[58] R. Pechter, What's pmml and what's new in pmml 4.0?, ACM SIGKDD Explorations Newsletter 11 (1) (2009) 19–25.
[59] R. Wirth, J. Hipp, Crisp-dm: Towards a standard process model for data mining, in: Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Citeseer, 2000, pp. 29–39.
[60] R. Cooley, B. Mobasher, J. Srivastava, Data preparation for mining world wide web browsing patterns, Knowledge and information systems 1 (1) (1999) 5–32.
[61] S. Zhang, C. Zhang, Q. Yang, Data preparation for data mining, Applied Artificial Intelligence 17 (2003) 375–381.
[62] L. Kaufman, P. J. Rousseeuw, Finding groups in data: an introduction to cluster analysis, Vol. 344, John Wiley & Sons, 2009.
[63] A. K. Jain, M. N. Murty, P. J. Flynn, Data clustering: a review, ACM computing surveys (CSUR) 31 (3) (1999) 264–323.
[64] P. Berkhin, A survey of clustering data mining techniques, in: Grouping multidimensional data, Springer, 2006, pp. 25–71.
[65] J. A. Hartigan, M. A. Wong, Algorithm as 136: A k-means clustering algorithm, Applied statistics (1979) 100–108.
[66] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise., in: Kdd, Vol. 96, 1996, pp. 226–231.
[67] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, in: ACM SIGMOD Record, Vol. 22, ACM, 1993, pp. 207–216.
[68] M. J. Zaki, S. Parthasarathy, M. Ogihara, W. Li, New algorithms for fast discovery of association rules, in: KDD, Vol. 97, 1997, pp. 283–286.
[69] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. I. Verkamo, Fast discovery of association rules., Advances in knowledge discovery and data mining 12 (1) (1996) 307–328.
[70] S. Kotsiantis, D. Kanellopoulos, Association rules mining: A recent overview, GESTS International Transactions on Computer Science and Engineering 32 (1) (2006) 71–82.
[71] I. Rish, An empirical study of the naive bayes classifier, in: IJCAI 2001 workshop on empirical methods in artificial intelligence, Vol. 3, IBM New York, 2001, pp. 41–46.
[72] D. A. Freedman, Statistical models and shoe leather, Sociological methodology 21 (2) (1991) 291–313.
[73] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Machine learning 29 (2-3) (1997) 131–163.
[74] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, Classification and regression trees, CRC press, 1984.
[75] M. Mehta, R. Agrawal, J. Rissanen, Sliq: A fast scalable classifier for data mining, in: Advances in Database Technology—EDBT'96, Springer, 1996, pp. 18–32.
[76] J. Shafer, R. Agrawal, M. Mehta, Sprint: A scalable parallel classifier for data mining, in: Proc. 1996 Int. Conf. Very

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Large Data Bases, Citeseer, 1996, pp. 544–555.
[77] J. R. Quinlan, Induction of decision trees, Machine learning 1 (1) (1986) 81–106.
[78] J. R. Quinlan, C4. 5: programs for machine learning, Elsevier, 2014.
[79] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.
[80] R. Kindermann, J. L. Snell, Markov random fields and their applications, American Mathematical Society Providence, 1980.
[81] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, T. Darrell, Hidden conditional random fields, IEEE Transactions on Pattern Analysis & Machine Intelligence (10) (2007) 1848–1852.
[82] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, Psychological review 65 (6) (1958) 386.
[83] P. Werbos, Beyond regression: New tools for prediction and analysis in the behavioral sciences, Ph.D. thesis, Harvard University (1974).
[84] M. T. Hagan, H. B. Demuth, M. H. Beale, Neural network design, Pws Pub. Boston, 1996.
[85] J. M. Zurada, Introduction to artificial neural systems, West St. Paul, 1992.
[86] W. Daelemans, A. Van den Bosch, Memory-based language processing, Cambridge University Press, 2005.
[87] N. S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, The American Statistician 46 (3) (1992) 175–185.
[88] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (3) (1995) 273–297.
[89] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, M. I. Jordan, Learning the kernel matrix with semidefinite programming, The Journal of Machine Learning Research 5 (2004) 27–72.
[90] M. Gönen, E. Alpaydın, Multiple kernel learning algorithms, The Journal of Machine Learning Research 12 (2011) 2211–2268.
[91] J. W. Grzymala-Busse, A new version of the rule induction system lers, Fundamenta Informaticae 31 (1) (1997) 27–39.
[92] W. W. Cohen, Fast effective rule induction, in: In Proceedings of the Twelfth International Conference on Machine Learning, Morgan Kaufmann, 1995, pp. 115–123.
[93] M. Verleysen, D. François, The curse of dimensionality in data mining and time series prediction, in: Computational Intelligence and Bioinspired Systems, Springer, 2005, pp. 758–770.
[94] A. L. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artificial intelligence 97 (1) (1997) 245–271.
[95] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, The Journal of Machine Learning Research 3 (2003) 1157–1182.
[96] R. Kohavi, G. H. John, Wrappers for feature subset selection, Artificial intelligence 97 (1) (1997) 273–324.
[97] J. Reunanen, Overfitting in making comparisons between variable selection methods, The Journal of Machine Learning Research 3 (2003) 1371–1382.
[98] E. Moreno, F. Bertolino, W. Racugno, An intrinsic limiting procedure for model selection and hypotheses testing, Journal of the American Statistical Association 93 (444) (1998) 1451–1460.
[99] P. Lahiri, Model selection, Institute of Mathematical Statistics, 2001.
[100] J. O. Berger, L. R. Pericchi, J. Ghosh, T. Samanta, F. De Santis, J. Berger, L. Pericchi, Objective bayesian methods for model selection: introduction and comparison, Lecture Notes-Monograph Series (2001) 135–207.
[101] C. E. Rasmussen, Z. Ghahramani, Occam's razor, in: In Advances in Neural Information Processing Systems 13, MIT Press, 2001, pp. 294–300.
[102] G. Golub, M. Heath, G. Wahba, Generalized cross-validation as a method for choosing a good ridge parameter, Technometrics 21 (1979) 215–223.
[103] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Vol. 14, 1995, pp. 1137–1145.
[104] B. Efron, R. Tibshirani, Cross-validation and the bootstrap: Estimating the error rate of a prediction rule, Division of Biostatistics, Stanford University, 1995.
[105] G. Holmes, A. Donkin, I. H. Witten, Weka: A machine learning workbench, in: Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on, IEEE, 1994, pp. 357–361.
[106] E. Frank, M. Hall, L. Trigg, G. Holmes, I. H. Witten, Data mining in bioinformatics using weka, Bioinformatics 20 (15) (2004) 2479–2481.
[107] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The weka data mining software: an update, ACM SIGKDD explorations newsletter 11 (1) (2009) 10–18.
[108] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel, Knime: The konstanz information miner, in: Data analysis, machine learning and applications, Springer, 2008, pp. 319–326.
[109] R. Ihaka, R. Gentleman, R: a language for data analysis and graphics, Journal of computational and graphical statistics 5 (3) (1996) 299–314.
[110] S. M. Weiss, N. Indurkhya, Predictive data mining: a practical guide, Morgan Kaufmann, 1998.
[111] F. X. Diebold, 'big data'dynamic factor models for macroeconomic measurement and forecasting, in: Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society,"(edited by M. Dewatripont, LP Hansen and S. Turnovsky), 2003, pp. 115–122.
[112] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. H. Byers, Big data: The next frontier for innovation, competition, and productivity (2011).
[113] J. Gantz, D. Reinsel, Extracting value from chaos, IDC iview (1142) (2011) 9–10.
[114] D. Laney, 3d data management: Controlling data volume, velocity and variety, META Group Research Note 6 (2001)

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

70.
[115] NIST Big Data Public Working Group, Big data interoperability framework: Definitions (2014).
[116] K. S. Vilas, Big data mining, Int J Comput Sci Manag Res 1 (1) (2012) 12–17.
[117] X. Wu, X. Zhu, G.-Q. Wu, W. Ding, Data mining with big data, Knowledge and Data Engineering, IEEE Transactions on 26 (1) (2014) 97–107.
[118] D. S. Tamhane, S. N. Sayyad, Big data analysis using hace theorem, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4.
[119] R. Pastor-Satorras, A. Vázquez, A. Vespignani, Dynamical and correlation properties of the internet, Physical review letters 87 (25) (2001) 258701.
[120] E. Ravasz, A.-L. Barabási, Hierarchical organization in complex networks, Physical Review E 67 (2) (2003) 026112.
[121] G. J. Pottie, W. J. Kaiser, Wireless integrated network sensors, Communications of the ACM 43 (5) (2000) 51–58.
[122] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, Wireless sensor networks: a survey, Computer networks 38 (4) (2002) 393–422.
[123] P. Bowler, The Mendelian revolution: the emergence of hereditarian concepts in modern science and society, A&C Black, 2000.
[124] D. B. Lenat, Automated theory formation in mathematics, in: IJCAI, Vol. 77, 1977, pp. 833–842.
[125] P. Langley, Data-driven discovery of physical laws, Cognitive Science 5 (1) (1981) 31–54.
[126] M. Lichman, UCI machine learning repository (2013).
URL `http://archive.ics.uci.edu/ml`
[127] X. L. Zhang, H. Begleiter, B. Porjesz, W. Wang, A. Litke, Event related potentials during object recognition tasks, Brain Research Bulletin 38 (6) (1995) 531–538.
[128] F. Sharbrough, G. Chatrian, R. Lesser, H. Lüders, M. Nuwer, T. Picton, American electroencephalographic society guidelines for standard electrode position nomenclature, J. clin. Neurophysiol 8 (2) (1991) 200–202.
[129] J. G. Snodgrass, M. Vanderwart, A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity., Journal of experimental psychology: Human learning and memory 6 (2) (1980) 174.
[130] C. W. Granger, Some recent development in a concept of causality, Journal of econometrics 39 (1) (1988) 199–211.
[131] C. W. Granger, Causality, cointegration, and control, Journal of Economic Dynamics and Control 12 (2) (1988) 551–559.
[132] Y. Li, Y. Qin, X. Chen, W. Li, Exploring the functional brain network of alzheimer's disease: based on the computational experiment, PloS one (2013) e73186.
[133] A. Khazaee, A. Ebrahimzadeh, A. Babajani-Feremi, Automatic classification of alzheimer's disease with resting-state fmri and graph theory, in: Biomedical Engineering (ICBME), 2014 21th Iranian Conference on, IEEE, 2014, pp. 252–257.
[134] A. Khazaee, A. Ebrahimzadeh, A. Babajani-Feremi, Identifying patients with alzheimer's disease using resting-state fmri and graph theory, Clinical Neurophysiology 126 (11) (2015) 2132–2141.
[135] Y. Li, Y. Wang, G. Wu, F. Shi, L. Zhou, W. Lin, D. Shen, Alzheimer's Disease Neuroimaging Initiative, Discriminant analysis of longitudinal cortical thickness changes in alzheimer's disease using dynamic and network features, Neurobiology of aging 33 (2) (2012) 427–e15.
[136] N. U. Dosenbach, B. Nardos, A. L. Cohen, D. A. Fair, J. D. Power, J. A. Church, S. M. Nelson, G. S. Wig, A. C. Vogel, C. N. Lessov-Schlaggar, K. A. Barnes, J. W. Dubis, E. Feczko, R. S. Coalson, J. R. Pruett, D. M. Barch, S. E. Petersen, B. L. Schlaggar, Prediction of individual brain maturity using fmri, Science 329 (5997) (2010) 1358–1361.
[137] J. Richiardi, S. Achard, E. Bullmore, D. Van De Ville, Classifying connectivity graphs using graph and vertex attributes, in: Pattern Recognition in NeuroImaging (PRNI), 2011 International Workshop on, IEEE, 2011, pp. 45–48.
[138] T. B. Meier, A. S. Desphande, S. Vergun, V. A. Nair, J. Song, B. B. Biswal, M. E. Meyerand, R. M. Birn, V. Prabhakaran, Support vector machine classification and characterization of age-related reorganization of functional brain networks, Neuroimage 60 (1) (2012) 601–613.
[139] I. Rish, B. Thyreau, B. Thirion, M. Plaze, M.-l. Paillere-martinot, C. Martelli, J.-l. Martinot, J.-B. Poline, G. A. Cecchi, Discriminative network models of schizophrenia, in: Advances in Neural Information Processing Systems, 2009, pp. 252–260.
[140] D. S. Bassett, B. G. Nelson, B. A. Mueller, J. Camchong, K. O. Lim, Altered resting state complexity in schizophrenia, Neuroimage 59 (3) (2012) 2196–2207.
[141] M. D. Sacchet, G. Prasad, L. C. Foland-Ross, P. M. Thompson, I. H. Gotlib, Elucidating brain connectivity networks in major depressive disorder using classification-based scoring, in: Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on, IEEE, 2014, pp. 246–249.
[142] F. Liu, W. Guo, J.-P. Fouche, Y. Wang, W. Wang, J. Ding, L. Zeng, C. Qiu, Q. Gong, W. Zhang, H. Chen, Multivariate classification of social anxiety disorder using whole brain functional connectivity, Brain Structure and Function 220 (1) (2015) 101–115.
[143] J. B. Colby, J. D. Rudie, J. A. Brown, P. K. Douglas, M. S. Cohen, Z. Shehzad, Insights into multimodal imaging classification of adhd, Frontiers in systems neuroscience 6 (2012) 59.
[144] C. R. Hernandez-Castillo, V. Galvez, C. Morgado-Valle, J. Fernandez-Ruiz, Whole-brain connectivity analysis and classification of spinocerebellar ataxia type 7 by functional mri, Cerebellum & Ataxias 1 (1) (2014) 2.
[145] Y. Iturria-Medina, A. P. Fernández, P. V. Hernández, L. G. Pentón, E. J. Canales-Rodríguez, L. Melie-Garcia, A. L. Castellanos, M. O. Ortega, Automated discrimination of brain pathological state attending to complex structural brain network properties: the shiverer mutant mouse case, PloS one (2011) e19071.
[146] T. Fekete, M. Wilf, D. Rubin, S. Edelman, R. Malach, L. R. Mujica-Parodi, Combining classification with fmri-derived complex network measures for potential neurodiagnostics, PloS one (2013) e62867.
[147] T. Fekete, N. Zach, L. R. Mujica-Parodi, M. R. Turner, Multiple kernel learning captures a systems-level functional

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

connectivity biomarker signature in amyotrophic lateral sclerosis, PloS one (2013) e85190.
[148] C.-Y. Wee, P.-T. Yap, D. Shen, Prediction of alzheimer's disease and mild cognitive impairment using cortical morphological patterns, Human brain mapping 34 (12) (2013) 3411–3425.
[149] M. Dyrba, M. Grothe, T. Kirste, S. J. Teipel, Multimodal analysis of functional and structural disconnection in alzheimer's disease using multiple kernel svm, Human brain mapping 36 (6) (2015) 2118–2131.
[150] B. Jie, D. Zhang, W. Gao, Q. Wang, C.-Y. Wee, D. Shen, Integration of network topological and connectivity properties for neuroimaging classification, Biomedical Engineering, IEEE Transactions on 61 (2) (2014) 576–589.
[151] F. Liu, B. Xie, Y. Wang, W. Guo, J.-P. Fouche, Z. Long, W. Wang, H. Chen, M. Li, X. Duan, J. Zhang, M. Qiu, H. Chen, Characterization of post-traumatic stress disorder using resting-state fmri with a multi-level parametric classification approach, Brain topography 28 (2) (2014) 221–237.
[152] D. R. Hardoon, J. Shawe-Taylor, Decomposing the tensor kernel support vector machine for neuroscience data with structured labels, Machine Learning 79 (1-2) (2010) 29–46.
[153] T.-S. Lim, W.-Y. Loh, Y.-S. Shih, A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms, Machine learning 40 (3) (2000) 203–228.
[154] S. L. Salzberg, On comparing classifiers: Pitfalls to avoid and a recommended approach, Data mining and knowledge discovery 1 (3) (1997) 317–328.
[155] E. Van Diessen, W. M. Otte, K. Braun, C. J. Stam, F. E. Jansen, Improved diagnosis in children with partial epilepsy using a multivariable prediction model based on eeg network characteristics, PloS one 8 (4) (2013) e59764.
[156] A. Temko, G. Lightbody, E. M. Thomas, G. B. Boylan, W. Marnane, Instantaneous measure of eeg channel importance for improved patient-adaptive neonatal seizure detection, Biomedical Engineering, IEEE Transactions on 59 (3) (2012) 717–727.
[157] D. Reynolds, Gaussian mixture models, in: Encyclopedia of Biometrics, Springer, 2009, pp. 659–663.
[158] W. Jamal, S. Das, I.-A. Oprescu, K. Maharatna, F. Apicella, F. Sicca, Classification of autism spectrum disorder using supervised learning of brain connectivity measures extracted from synchrostates, Journal of neural engineering 11 (4) (2014) 046019.
[159] L. Pollonini, U. Patidar, N. Situ, R. Rezaie, A. C. Papanicolaou, G. Zouridakis, Functional connectivity networks in the autistic and healthy brain assessed using granger causality, in: Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE, IEEE, 2010, pp. 1730–1733.
[160] G. Zouridakis, U. Patidar, N. Situ, R. Rezaie, E. M. Castillo, H. S. Levin, A. C. Papanicolaou, Functional connectivity changes in mild traumatic brain injury assessed using magnetoencephalography, Journal of Mechanics in Medicine and Biology 12 (02) (2012) 1240006.
[161] M. Zanin, P. Sousa, D. Papo, R. Bajo, J. García-Prieto, F. del Pozo, E. Menasalvas, S. Boccaletti, Optimizing functional network representation of multivariate time series, Scientific reports 2.
[162] L. Antiqueira, F. A. Rodrigues, B. C. van Wijk, L. d. F. Costa, A. Daffertshofer, Estimating complex cortical networks via surface recordings—a critical note, Neuroimage 53 (2) (2010) 439–449.
[163] E. D. Fagerholm, P. J. Hellyer, G. Scott, R. Leech, D. J. Sharp, Disconnection of network hubs and cognitive impairment after traumatic brain injury, Brain 138 (6) (2015) 1696–1709.
[164] T. Li, J. Hong, J. Zhang, Eeg classification based on small-world neural network for brain-computer interface, in: Natural Computation (ICNC), 2010 Sixth International Conference on, Vol. 1, IEEE, 2010, pp. 252–256.
[165] M. Hassan, A. Ali, M. Eladawy, Classification of the imagination of the left and right hand movements using eeg, in: Biomedical Engineering Conference, 2008. CIBEC 2008. Cairo International, IEEE, 2008, pp. 1–5.
[166] J. F. D. Saa, M. Çetin, A latent discriminative model-based approach for classification of imaginary motor tasks from eeg data, Journal of neural engineering 9 (2) (2012) 026020.
[167] M. Demuru, F. Fara, M. Fraschini, Brain network analysis of eeg functional connectivity during imagery hand movements, Journal of integrative neuroscience 12 (04) (2013) 441–447.
[168] J. Lederberg, A. Mccray, The scientist:'ome sweet 'omics–a genealogical treasury of words, The Scientist 17 (7).
[169] K. Plaimas, R. Eils, R. König, Identifying essential genes in bacterial metabolic networks with machine learning methods, BMC systems biology 4 (1) (2010) 56.
[170] G. Pandey, B. Zhang, A. N. Chang, C. L. Myers, J. Zhu, V. Kumar, E. E. Schadt, An integrative multi-network and multi-classifier approach to predict genetic interactions, PLoS Comput Biol 6 (9) (2010) e1000928.
[171] K. C. Chipman, A. K. Singh, Predicting genetic interactions with random walks on biological networks, BMC bioinformatics 10 (1) (2009) 17.
[172] Y. Qi, Y. Suhail, Y.-y. Lin, J. D. Boeke, J. S. Bader, Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions, Genome research 18 (12) (2008) 1991–2004.
[173] Y. Chen, D. Xu, Understanding protein dispensability through machine-learning analysis of high-throughput data, Bioinformatics 21 (5) (2005) 575–581.
[174] S. R. Paladugu, S. Zhao, A. Ray, A. Raval, Mining protein networks for synthetic genetic interactions, Bmc Bioinformatics 9 (1) (2008) 426.
[175] D. W. Huang, B. T. Sherman, R. A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, Nucleic acids research 37 (1) (2009) 1–13.
[176] K. Wang, M. Li, H. Hakonarson, Annovar: functional annotation of genetic variants from high-throughput sequencing data, Nucleic acids research 38 (16) (2010) e164–e164.
[177] S. Zhang, G. Jin, X.-S. Zhang, L. Chen, Discovering functions and revealing mechanisms at molecular level from biological networks, Proteomics 7 (16) (2007) 2856–2869.

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[178] J. Zahiri, J. H. Bozorgmehr, A. Masoudi-Nejad, Computational prediction of protein–protein interaction networks: algorithms and resources, Current genomics 14 (6) (2013) 397.
[179] C. Lippert, Z. Ghahramani, K. M. Borgwardt, Gene function prediction from synthetic lethality networks via ranking on demand, Bioinformatics 26 (7) (2010) 912–918.
[180] O. Ray, C. H. Bryant, Inferring the function of genes from synthetic lethal mutations, in: Complex, Intelligent and Software Intensive Systems, 2008. CISIS 2008. International Conference on, IEEE, 2008, pp. 667–671.
[181] M. Deng, K. Zhang, S. Mehta, T. Chen, F. Sun, Prediction of protein function using protein-protein interaction data, Journal of Computational Biology 10 (6) (2003) 947–960.
[182] L. J. Lu, Y. Xia, H. Yu, A. Rives, H. Lu, F. Schubert, M. Gerstein, Protein interaction prediction by integrating genomic features and protein interaction network analysis, in: Data analysis and visualization in genomics and proteomics, Wiley, 2005, p. 61.
[183] J. Lee, J. Lee, Y. Moreno, Hidden information revealed by optimal community structure from a protein-complex bipartite network improves protein function prediction, PloS one 8 (4) (2013) 60372.
[184] H. N. Chua, K. Ning, W.-K. Sung, H. W. Leong, L. Wong, Using indirect protein–protein interactions for protein complex prediction, Journal of bioinformatics and computational biology 6 (03) (2008) 435–466.
[185] J. Qiu, W. S. Noble, Predicting co-complexed protein pairs from heterogeneous data, PLoS Comput Biol 4 (4) (2008) e1000054.
[186] A. Stevens, C. De Leonibus, D. Hanson, A. Dowsey, A. Whatmore, S. Meyer, R. Donn, P. Chatelain, I. Banerjee, K. Cosgrove, P. Clayton, M. Dunne, Network analysis: a new approach to study endocrine disorders, Journal of molecular endocrinology 52 (1) (2014) R79–R93.
[187] J. Pan, Z. Cong, M. Zhong, Y. Sun, Analysis of hepatocellular carcinoma and metastatic hepatic carcinoma via functional modules in a protein-protein interaction network, Journal of cancer research and therapeutics 10 (7) (2014) 186.
[188] A. Özgür, T. Vu, G. Erkan, D. R. Radev, Identifying gene-disease associations using centrality on a literature mined gene-interaction network, Bioinformatics 24 (13) (2008) i277–i285.
[189] G. Erkan, A. Özgür, D. R. Radev, Semi-supervised classification for extracting protein interaction sentences using dependency parsing., in: EMNLP-CoNLL, Vol. 7, 2007, pp. 228–237.
[190] M. Zanin, J. M. Alcazar, J. V. Carbajosa, M. G. Paez, D. Papo, P. Sousa, E. Menasalvas, S. Boccaletti, Parenclitic networks: uncovering new functions in biological data, Scientific reports 4.
[191] M. Zanin, S. Boccaletti, Complex networks analysis of obstructive nephropathy data, Chaos: An Interdisciplinary Journal of Nonlinear Science 21 (3) (2011) 033103.
[192] M. Zanin, E. Menasalvas Ruiz, P. Sousa, S. Boccaletti, Preprocessing and analyzing genetic data with complex networks: An application to obstructive nephropathy, Networks And Heterogeneous Media 7 (3) (2012) 473–481.
[193] M. Zanin, E. Menasalvas, S. Boccaletti, P. Sousa, Feature selection in the reconstruction of complex network representations of spectral data, PloS one 8 (8).
[194] 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, P. Sousa, Knowledge discovery in spectral data by means of complex networks, Metabolites 3 (1) (2013) 155–167.
[195] A. Karsakov, T. Bartlett, I. Meyerov, A. Zaikin, M. Ivanchenko, Parenclitic network analysis of methylation data for cancer identification, arXiv preprint arXiv:1506.04421.
[196] F. Varela, J.-P. Lachaux, E. Rodriguez, J. Martinerie, The brainweb: phase synchronization and large-scale integration, Nature reviews neuroscience 2 (4) (2001) 229–239.
[197] L. Astolfi, F. Cincotti, D. Mattia, M. G. Marciani, L. A. Baccala, F. de Vico Fallani, S. Salinari, M. Ursino, M. Zavaglia, L. Ding, J. C. Edgar, G. A. Miller, B. He, F. Babiloni, Comparison of different cortical connectivity estimators for high-resolution eeg recordings, Human brain mapping 28 (2) (2007) 143–157.
[198] M. Jalili, M. G. Knyazeva, Constructing brain functional networks from eeg: partial and unpartial correlations, Journal of integrative neuroscience 10 (02) (2011) 213–232.
[199] M. Jalili, E. Barzegaran, M. G. Knyazeva, Synchronization of eeg: Bivariate and multivariate measures, Neural Systems and Rehabilitation Engineering, IEEE Transactions on 22 (2) (2014) 212–221.
[200] R. Q. Quiroga, A. Kraskov, T. Kreuz, P. Grassberger, Performance of different synchronization measures in real data: a case study on electroencephalographic signals, Physical Review E 65 (4) (2002) 041903.
[201] S. I. Dimitriadis, N. A. Laskaris, Y. Del Rio-Portilla, G. C. Koudounis, Characterizing dynamic functional connectivity across sleep stages from eeg, Brain topography 22 (2) (2009) 119–133.
[202] C. Stam, B. Van Dijk, Synchronization likelihood: an unbiased measure of generalized synchronization in multivariate data sets, Physica D: Nonlinear Phenomena 163 (3) (2002) 236–251.
[203] M. Meilă, Comparing clusterings—an information based distance, Journal of multivariate analysis 98 (5) (2007) 873–895.
[204] C. Lithari, M. Klados, C. Papadelis, C. Pappas, M. Albani, P. Bamidis, How does the metric choice affect brain functional connectivity networks?, Biomedical Signal Processing and Control 7 (3) (2012) 228–236.
[205] J. Bonita, L. Ambolode II, B. Rosenberg, C. Cellucci, T. Watanabe, P. Rapp, A. Albano, Time domain measures of inter-channel eeg correlations: a comparison of linear, nonparametric and nonlinear measures, Cognitive neurodynamics 8 (1) (2014) 1–15.
[206] R. Hornero, D. Abásolo, J. Escudero, C. Gómez, Nonlinear analysis of electroencephalogram and magnetoencephalogram recordings in patients with alzheimer's disease, Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 367 (1887) (2009) 317–336.
[207] Y. Zhang, J. Han, X. Hu, L. Guo, T. Liu, Data-driven evaluation of functional connectivity metrics, in: Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on, IEEE, 2013, pp. 532–535.

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[208] B. C. Van Wijk, C. J. Stam, A. Daffertshofer, Comparing brain networks of different size and connectivity density using graph theory, PloS one 5 (10) (2010) e13701.
[209] B. Jie, D. Zhang, C.-Y. Wee, D. Shen, Topological graph kernel on multiple thresholded functional connectivity networks for mild cognitive impairment classification, Human brain mapping 35 (7) (2014) 2876–2897.
[210] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, K. M. Borgwardt, Graph kernels, The Journal of Machine Learning Research 11 (2010) 1201–1242.
[211] N. Shervashidze, P. Schweitzer, E. J. Van Leeuwen, K. Mehlhorn, K. M. Borgwardt, Weisfeiler-lehman graph kernels, The Journal of Machine Learning Research 12 (2011) 2539–2561.
[212] D. Meunier, R. Lambiotte, A. Fornito, K. D. Ersche, E. T. Bullmore, Hierarchical modularity in human brain functional networks, Frontiers in neuroinformatics 3.
[213] Y.-Y. Ahn, J. P. Bagrow, S. Lehmann, Link communities reveal multiscale complexity in networks, Nature 466 (7307) (2010) 761–764.
[214] J. M. Duarte-Carvajalino, N. Jahanshad, C. Lenglet, K. L. McMahon, G. I. de Zubicaray, N. G. Martin, M. J. Wright, P. M. Thompson, G. Sapiro, Hierarchical topological network analysis of anatomical human brain connectivity and differences related to sex and kinship, Neuroimage 59 (4) (2012) 3784–3804.
[215] A. Jain, D. Zongker, Feature selection: Evaluation, application, and small sample performance, Pattern Analysis and Machine Intelligence, IEEE Transactions on 19 (2) (1997) 153–158.
[216] J. Richiardi, S. Achard, H. Bunke, D. Van De Ville, Machine learning with brain graphs: predictive modeling approaches for functional imaging in systems neuroscience, Signal Processing Magazine, IEEE 30 (3) (2013) 58–70.
[217] R. Salvador, J. Suckling, M. R. Coleman, J. D. Pickard, D. Menon, E. Bullmore, Neurophysiological architecture of functional magnetic resonance images of human brain, Cerebral cortex 15 (9) (2005) 1332–1342.
[218] J. Richiardi, H. Eryilmaz, S. Schwartz, P. Vuilleumier, D. Van De Ville, Decoding brain states from fmri connectivity graphs, Neuroimage 56 (2) (2011) 616–626.
[219] J. Tohka, Y. He, A. C. Evans, The impact of sampling density upon cortical network analysis: regions or points, Magnetic resonance imaging 30 (7) (2012) 978–992.
[220] V. Calhoun, T. Adali, G. Pearlson, J. Pekar, A method for making group inferences from functional mri data using independent component analysis, Human brain mapping 14 (3) (2001) 140–151.
[221] D. Cordes, V. Haughton, J. D. Carew, K. Arfanakis, K. Maravilla, Hierarchical clustering to measure connectivity in fmri resting-state data, Magnetic resonance imaging 20 (4) (2002) 305–317.
[222] C.-W. Woo, A. Krishnan, T. D. Wager, Cluster-extent based thresholding in fmri analyses: pitfalls and recommendations, Neuroimage 91 (2014) 412–419.
[223] F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, E. Formisano, Combining multivariate voxel selection and support vector machines for mapping and classification of fmri spatial patterns, Neuroimage 43 (1) (2008) 44–58.
[224] V. Michel, C. Damon, B. Thirion, Mutual information-based feature selection enhances fmri brain activity classification, in: Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on, IEEE, 2008, pp. 592–595.
[225] D. François, F. Rossi, V. Wertz, M. Verleysen, Resampling methods for parameter-free and robust feature selection with mutual information, Neurocomputing 70 (7) (2007) 1276–1288.
[226] R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, E. Eger, F. Bach, B. Thirion, Multiscale mining of fmri data with hierarchical structured sparsity, SIAM Journal on Imaging Sciences 5 (3) (2012) 835–856.
[227] H. Deng, G. Runger, Feature selection via regularized trees, in: Neural Networks (IJCNN), The 2012 International Joint Conference on, IEEE, 2012, pp. 1–8.
[228] C. Chu, A.-L. Hsu, K.-H. Chou, P. Bandettini, C. Lin, Alzheimer's Disease Neuroimaging Initiative, Does feature selection improve classification accuracy? impact of sample size and feature selection on classification using anatomical magnetic resonance images, Neuroimage 60 (1) (2012) 59–70.
[229] J. R. Wolpaw, N. Birbaumer, W. J. Heetderks, D. J. McFarland, P. H. Peckham, G. Schalk, E. Donchin, L. A. Quatrano, C. J. Robinson, T. M. Vaughan, Brain-computer interface technology: a review of the first international meeting, IEEE transactions on rehabilitation engineering 8 (2) (2000) 164–173.
[230] A. Vallabhaneni, T. Wang, B. He, Brain—computer interface, in: Neural engineering, Springer, 2005, pp. 85–121.
[231] R. Santana, C. Bielza, P. Larrañaga, Regularized logistic regression and multiobjective variable selection for classifying meg data, Biological cybernetics 106 (6-7) (2012) 389–405.
[232] A. Joudaki, N. Salehi, M. Jalili, M. G. Knyazeva, Eeg-based functional brain networks: does the network size matter, PloS one 7 (4) (2012) e35673.
[233] R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, Y. Wang, The properties of high-dimensional data spaces: implications for exploring gene and protein expression data, Nature Reviews Cancer 8 (1) (2008) 37–49.
[234] F. M. Lopes, D. C. Martins, J. Barrera, R. M. Cesar, A feature selection technique for inference of graphs from their known topological properties: Revealing scale-free gene regulatory networks, Information Sciences 272 (2014) 1–15.
[235] P. Pudil, J. Novovičová, J. Kittler, Floating search methods in feature selection, Pattern recognition letters 15 (11) (1994) 1119–1125.
[236] J. Griffin, H. Williams, E. Sang, K. Clarke, C. Rae, J. Nicholson, Metabolic profiling of genetic disorders: a multitissue 1 h nuclear magnetic resonance spectroscopic and pattern recognition study into dystrophic tissue, Analytical biochemistry 293 (1) (2001) 16–21.
[237] M. Granzow, D. Berrar, W. Dubitzky, A. Schuster, F. Azuaje, R. Eils, Tumor classification by gene expression profiling: comparison and validation of five clustering methods, ACM Sigbio Newsletter 21 (1) (2001) 16–22.
[238] E. R. Dougherty, J. Barrera, M. Brun, S. Kim, R. M. Cesar, Y. Chen, M. Bittner, J. M. Trent, Inference from clustering

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with application to gene-expression microarrays, Journal of Computational Biology 9 (1) (2002) 105–126.
[239] D. Radke, U. Möller, Quantitative evaluation of established clustering methods for gene expression data, in: Biological and Medical Data Analysis, Springer, 2004, pp. 399–408.
[240] J. F. Donges, Y. Zou, N. Marwan, J. Kurths, Complex networks in climate dynamics, The European Physical Journal Special Topics 174 (1) (2009) 157–179.
[241] R. G. Fovell, M.-Y. C. Fovell, Climate zones of the conterminous united states defined using cluster analysis, Journal of Climate 6 (11) (1993) 2103–2135.
[242] M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, C. Potter, Discovery of climate indices using clustering, in: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2003, pp. 446–455.
[243] K. Steinhaeuser, N. V. Chawla, A. R. Ganguly, Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science, Statistical Analysis and Data Mining: The ASA Data Science Journal 4 (5) (2011) 497–511.
[244] A. A. Tsonis, G. Wang, K. L. Swanson, F. A. Rodrigues, L. da Fontura Costa, Community structure and dynamics in climate networks, Climate dynamics 37 (5-6) (2011) 933–940.
[245] M. Zanin, Synchronization likelihood in aircraft trajectories, in: Proceedings of the Tenth USA/Europe Air Traffic Management Research and Development Seminar, Chicago, USA, 2013.
[246] M. Zanin, Network analysis reveals patterns behind air safety events, Physica A: Statistical Mechanics and its Applications 401 (2014) 201–206.
[247] R. L. Graham, P. Hell, On the history of the minimum spanning tree problem, Annals of the History of Computing 7 (1) (1985) 43–57.
[248] H. N. Gabow, Z. Galil, T. Spencer, R. E. Tarjan, Efficient algorithms for finding minimum spanning trees in undirected and directed graphs, Combinatorica 6 (2) (1986) 109–122.
[249] P. Tewarie, A. Hillebrand, M. Schoonheim, B. Van Dijk, J. Geurts, F. Barkhof, C. Polman, C. Stam, Functional brain network analysis using minimum spanning trees in multiple sclerosis: an meg source-space study, Neuroimage 88 (2014) 308–318.
[250] C. Stam, P. Tewarie, E. Van Dellen, E. Van Straaten, A. Hillebrand, P. Van Mieghem, The trees and the forest: characterization of complex brain networks with minimum spanning trees, International Journal of Psychophysiology 92 (3) (2014) 129–138.
[251] S. Micciché, G. Bonanno, F. Lillo, R. N. Mantegna, Degree stability of a minimum spanning tree of price return and volatility, Physica A: Statistical Mechanics and its Applications 324 (1) (2003) 66–73.
[252] M. Tumminello, C. Coronnello, F. Lillo, S. Micciche, R. N. Mantegna, Spanning trees and bootstrap reliability estimation in correlation-based networks, International Journal of Bifurcation and Chaos 17 (07) (2007) 2319–2329.
[253] G. Bonanno, G. Caldarelli, F. Lillo, S. Micciché, N. Vandewalle, R. N. Mantegna, Networks of equities in financial markets, The European Physical Journal B-Condensed Matter and Complex Systems 38 (2) (2004) 363–371.
[254] M. Tumminello, T. Aste, T. Di Matteo, R. N. Mantegna, A tool for filtering information in complex systems, Proceedings of the National Academy of Sciences of the United States of America 102 (30) (2005) 10421–10426.
[255] T. Aste, T. Di Matteo, S. Hyde, Complex networks on hyperbolic surfaces, Physica A: Statistical Mechanics and its Applications 346 (1) (2005) 20–26.
[256] A. Rheinwalt, N. Marwan, J. Kurths, P. Werner, F.-W. Gerstengarbe, Boundary effects in network measures of spatially embedded networks, EPL (Europhysics Letters) 100 (2) (2012) 28002.
[257] M. Á. Serrano, M. Boguná, A. Vespignani, Extracting the multiscale backbone of complex weighted networks, Proceedings of the national academy of sciences 106 (16) (2009) 6483–6488.
[258] B. Barzel, A.-L. Barabási, Network link prediction by global silencing of indirect correlations, Nature biotechnology 31 (8) (2013) 720–725.
[259] F. M. Lopes, D. Martins, R. M. Cesar, Comparative study of grns inference methods based on feature selection by mutual information, in: Genomic Signal Processing and Statistics, 2009. GENSIPS 2009. IEEE International Workshop on, IEEE, 2009, pp. 1–4.
[260] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, A. Califano, Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, BMC bioinformatics 7 (Suppl 1) (2006) S7.
[261] P. E. Meyer, F. Lafitte, G. Bontempi, minet: Ar/bioconductor package for inferring large transcriptional networks using mutual information, BMC bioinformatics 9 (1) (2008) 461.
[262] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, Pattern Analysis and Machine Intelligence, IEEE Transactions on 27 (8) (2005) 1226–1238.
[263] F. M. Lopes, D. C. Martins, R. M. Cesar, Feature selection environment for genomic applications, BMC bioinformatics 9 (1) (2008) 451.
[264] A. F. Villaverde, J. Ross, F. Morán, J. R. Banga, Mider: network inference with mutual information distance and entropy reduction, PloS one (2014) e96732.
[265] L. Zhu, S. Deng, D.-S. Huang, A two-stage geometric method for detecting unreliable links in protein-protein networks, in: Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on, IEEE, 2014, pp. 31–36.
[266] Y.-K. Lei, Z.-H. You, T. Dong, Y.-X. Jiang, J.-A. Yang, Increasing reliability of protein interactome by fast manifold embedding, Pattern Recognition Letters 34 (4) (2013) 372–379.
[267] T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, N. V. Fedoroff, Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns, Proceedings of the National Academy of Sciences 103 (50) (2006) 19033–19038.

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[268] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, X. D. Koutsoukos, Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation, The Journal of Machine Learning Research 11 (2010) 171–234.
[269] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, X. D. Koutsoukos, Local causal and markov blanket induction for causal discovery and feature selection for classification part ii: Analysis and extensions, The Journal of Machine Learning Research 11 (2010) 235–284.
[270] J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference, Morgan Kaufmann, 2014.
[271] D. Hardin, I. Tsamardinos, C. F. Aliferis, A theoretical characterization of linear svm-based feature selection, in: Proceedings of the twenty-first international conference on Machine learning, ACM, 2004, p. 48.
[272] A. Statnikov, D. Hardin, C. Aliferis, Using svm weight-based methods to identify causally relevant and non-causally relevant variables, sign 1 (2006) 4.
[273] I. Tsamardinos, C. F. Aliferis, Towards principled feature selection: Relevancy, filters and wrappers, in: Proceedings of the ninth international workshop on Artificial Intelligence and Statistics, Morgan Kaufmann Publishers: Key West, FL, USA, 2003.
[274] I. Tsamardinos, C. F. Aliferis, A. Statnikov, Time and sample efficient discovery of markov blankets and direct causal relations, in: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2003, pp. 673–678.
[275] Y.-Y. Liu, J.-J. Slotine, A.-L. Barabási, Controllability of complex networks, Nature 473 (7346) (2011) 167–173.
[276] N. J. Cowan, E. J. Chastain, D. A. Vilhena, J. S. Freudenberg, C. T. Bergstrom, Nodal dynamics, not degree distributions, determine the structural controllability of complex networks, PloS one 7 (6) (2012) e38398.
[277] B. S. Duran, P. L. Odell, Cluster analysis: a survey, Vol. 100, Springer Science & Business Media, 2013.
[278] M. Girvan, M. E. Newman, Community structure in social and biological networks, Proceedings of the national academy of sciences 99 (12) (2002) 7821–7826.
[279] V. Estivill-Castro, Why so many clustering algorithms: a position paper, ACM SIGKDD explorations newsletter 4 (1) (2002) 65–75.
[280] M. Tumminello, S. Micciché, F. Lillo, J. Varho, J. Piilo, R. N. Mantegna, Community characterization of heterogeneous complex systems, Journal of Statistical Mechanics: Theory and Experiment 2011 (01) (2011) P01019.
[281] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, J.-P. Onnela, Community structure in time-dependent, multiscale, and multiplex networks, science 328 (5980) (2010) 876–878.
[282] J. Hopcroft, O. Khan, B. Kulis, B. Selman, Tracking evolving communities in large linked networks, Proceedings of the National Academy of Sciences 101 (1) (2004) 5249–5253.
[283] L. Danon, A. Díaz-Guilera, A. Arenas, The effect of size heterogeneity on community identification in complex networks, Journal of Statistical Mechanics: Theory and Experiment 2006 (11) (2006) P11010.
[284] J. Serrà, M. Zanin, P. Herrera, X. Serra, Characterization and exploitation of community structure in cover song networks, Pattern Recognition Letters 33 (9) (2012) 1032–1041.
[285] M. Tumminello, F. Lillo, J. Piilo, R. N. Mantegna, Identification of clusters of investors from their real trading activity in a financial market, New Journal of Physics 14 (1) (2012) 013041.
[286] M. Tumminello, S. Micciché, F. Lillo, J. Piilo, R. N. Mantegna, Statistically validated networks in bipartite complex systems, PloS one (2011) e17994.
[287] R. N. Mantegna, Hierarchical structure in financial markets, The European Physical Journal B-Condensed Matter and Complex Systems 11 (1) (1999) 193–197.
[288] S. Zhang, R.-S. Wang, X.-S. Zhang, Identification of overlapping community structure in complex networks using fuzzy c-means clustering, Physica A: Statistical Mechanics and its Applications 374 (1) (2007) 483–490.
[289] P. G. Sun, L. Gao, S. S. Han, Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks, Information Sciences 181 (6) (2011) 1060–1071.
[290] Z. Ye, S. Hu, J. Yu, Adaptive clustering algorithm for community detection in complex networks, Physical Review E 78 (4) (2008) 046115.
[291] X. Liu, D. Li, S. Wang, Z. Tao, Effective algorithm for detecting community structure in complex networks based on ga and clustering, in: Computational Science–ICCS 2007, Springer, 2007, pp. 657–664.
[292] H. Yu, P. Braun, M. A. Yıldırım, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.-F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A.-S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A.-L. Barabási, J. Tavernier, D. E. Hill, M. Vidal, High-quality binary protein interaction map of the yeast interactome network, Science 322 (5898) (2008) 104–110.
[293] L. A. N. Amaral, A truer measure of our ignorance, Proceedings of the National Academy of Sciences 105 (19) (2008) 6795–6796.
[294] L. Lü, T. Zhou, Link prediction in complex networks: A survey, Physica A: Statistical Mechanics and its Applications 390 (6) (2011) 1150–1170.
[295] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, Knowledge and Data Engineering, IEEE Transactions on 17 (6) (2005) 734–749.
[296] F. Ricci, L. Rokach, B. Shapira, Introduction to recommender systems handbook, Springer, 2011.
[297] L. Lü, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, T. Zhou, Recommender systems, Physics Reports 519 (1) (2012) 1–49.
[298] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, E. Shir, A model of internet topology using k-shell decomposition, Proceedings of the National Academy of Sciences 104 (27) (2007) 11150–11154.

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[299] A. Clauset, C. Moore, M. E. Newman, Hierarchical structure and the prediction of missing links in networks, Nature 453 (7191) (2008) 98–101.
[300] C. Schmeltzer, J. Soriano, I. M. Sokolov, S. Rüdiger, Percolation of spatially constrained erdős-rényi networks with degree correlations, Physical Review E 89 (1) (2014) 012116.
[301] A. A. Rad, I. Sendiña-Nadal, D. Papo, M. Zanin, J. M. Buldu, F. del Pozo, S. Boccaletti, Topological measure locating the effective crossover between segregation and integration in a modular network, Physical review letters 108 (22) (2012) 228701.
[302] M. Kaiser, C. C. Hilgetag, Modelling the development of cortical systems networks, Neurocomputing 58 (2004) 297–302.
[303] Y. Chen, S. Wang, C. C. Hilgetag, C. Zhou, Trade-off between multiple constraints enables simultaneous formation of modules and hubs in neural systems, PloS one (2013) e1002937.
[304] V. Nicosia, P. E. Vértes, W. R. Schafer, V. Latora, E. T. Bullmore, Phase transition in the economically modeled growth of a cellular nervous system, Proceedings of the National Academy of Sciences 110 (19) (2013) 7880–7885.
[305] F. Klimm, D. S. Bassett, J. M. Carlson, P. J. Mucha, Resolving structural variability in network models and the brain, PLOS Comput Biol 10 (3) (2014) e1003491.
[306] H. F. Kwok, P. Jurica, A. Raffone, C. van Leeuwen, Robust emergence of small-world structure in networks of spiking neurons, Cognitive neurodynamics 1 (1) (2007) 39–51.
[307] C. J. Stam, A. Hillebrand, H. Wang, P. Van Mieghem, Emergence of modular structure in a large-scale brain network with interactions between dynamics and connectivity, Frontiers in computational neuroscience 4.
[308] D. Fraiman, P. Balenzuela, J. Foss, D. R. Chialvo, Ising-like dynamics in large-scale functional brain networks, Physical Review E 79 (6) (2009) 061922.
[309] S. L. Simpson, S. Hayasaka, P. J. Laurienti, Exponential random graph modeling for complex brain networks, PLoS One 6 (5) (2011) e20039.
[310] P. E. Vértes, A. F. Alexander-Bloch, N. Gogtay, J. N. Giedd, J. L. Rapoport, E. T. Bullmore, Simple models of human brain functional networks, Proceedings of the National Academy of Sciences 109 (15) (2012) 5868–5873.
[311] P. E. Vértes, A. Alexander-Bloch, E. T. Bullmore, Generative models of rich clubs in hebbian neuronal networks and large-scale human brain networks, Philosophical Transactions of the Royal Society of London B: Biological Sciences 369 (1653) (2014) 20130531.
[312] J. M. Buldú, R. Bajo, F. Maestú, N. Castellanos, I. Leyva, P. Gil, I. Sendiña-Nadal, J. A. Almendral, A. Nevado, F. Del-Pozo, S. Boccaletti, Reorganization of functional networks in mild cognitive impairment, PLoS One 6 (5) (2011) e19584.
[313] M. Zanin, M. Correia, P. A. Sousa, J. Cruz, Probabilistic constraint programming for parameters optimisation of generative models, in: Progress in Artificial Intelligence, Springer, 2015, pp. 376–387.
[314] M. Zanin, M. Correia, P. A. Sousa, J. Cruz, From phenotype to genotype in complex brain networks, Scientific Reports 6 (2016) 19790.
[315] D. Gupta, A. Sharma, N. Unny, G. Manjunath, Graphical analysis and visualization of big data in business domains, in: Big Data Analytics, Springer, 2014, pp. 53–56.
[316] J. Z. Pan, Resource description framework, in: Handbook on Ontologies, Springer, 2009, pp. 71–90.
[317] 01.org, intel open source technology center, `https://01.org/graphbuilder/`, accessed: 2015-09-30.
[318] Centre for disease control and prevention, `http://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html`, accessed: 2015-09-30.
[319] Nodexl, `http://nodexl.codeplex.com/`, accessed: 2015-09-30.
[320] I. Robinson, J. Webber, E. Eifrem, Graph databases, O'Reilly Media, Inc., 2013.

bioRxiv preprint doi: https://doi.org/10.1101/054064; this version posted May 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

AppendixA. List of acronyms

Acronym	Explanation
ANN	Artificial Neural Network [84, 85], see Section 2.2 for definition.
AUC	Area Under the Curve. In Receiver Operating Characteristic (ROC) curves, AUC measures the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.
BCI	Brain-Computer Interfaces, a computer-based system to acquire, and respond upon, brain signals.
EEG	ElectroEncephaloGram, an electrophysiological monitoring method to record electrical activity of the brain.
fMRI	Functional Magnetic Resonance Imaging, an MRI recorded during a cognitive task.
HCRF	Hidden Conditional Random Fields [81], see Section 2.2 for a definition.
KDD	Knowledge Discovery in Databases, the overall process of discovering useful knowledge from data [57].
KNN	k-Nearest Neighbors algorithm, a classification algorithm based on the class of neighbour training examples. See Section 2.2 for a definition.
LOOCV	Leave One Out Cross Validation, a cross-validation procedure in which one sample at the time is excluded from the training. See Section 2.2 for details.
MEG	MagnetoEncephaloGram, an electrophysiological monitoring method to record the magnetic field generated by the activity of the brain.
MKL	Multiple Kernel Learning [89], see Section 2.2 for a definition.
MRF	Markov Random Field [80], see Section 2.2 for a definition.
MRI	Magnetic Resonance Imaging, a medical imaging technique that uses magnetic fields and radio waves to form images of the brain.
MST	Minimum Spanning Tree, tree that connects all the vertices of a network together with the minimal total weighting for its edges.
NNC	Nearest Neighbour Classifiers, classification method based on copying the class of the closest training examples in the feature space.
PMML	Predictive Model Markup Language, XML-based standard for interchanging predictive models [58].
RDF	Resource Description Framework, general standard for conceptual description or modelling of information [316].
RFE	Recursive Feature Elimination (RFE), a feature selection method based on repeatedly removing features with low relevance. See Section 2.2 for details.
ROC	Receiver Operating Characteristic, graphical plot illustrating the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the true positive rate as a function of the false positive rate, for different thresholds.
SVM	Support Vector Machine [88], see Section 2.2 for a definition.