

Analysis of Complex Data by Means of Complex Networks

Massimiliano Zanin, Ernestina Menasalvas, Stefano Boccaletti, Pedro A. Sousa

▶ **To cite this version:**

Massimiliano Zanin, Ernestina Menasalvas, Stefano Boccaletti, Pedro A. Sousa. Analysis of Complex Data by Means of Complex Networks. 5th Doctoral Conference on Computing, Electrical and Industrial Systems (DoCEIS), Apr 2014, Costa de Caparica, Portugal. pp.39-46, 10.1007/978-3-642-54734-8_5 . hal-01274746

HAL Id: hal-01274746

https://inria.hal.science/hal-01274746

Submitted on 16 Feb 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Analysis of Complex Data by Means of Complex Networks

Massimiliano Zanin[1,2,3], Ernestina Menasalvas[2], Stefano Boccaletti[4] and Pedro A. Sousa[1],

[1] Faculdade de Ciências e Tecnologia, Departamento de Engenharia Electrotécnica, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
[2] Center for Biomedical Technology, Universidad Politécnica de Madrid, 28223 Pozuelo de Alarcón, Madrid, Spain
[3] Innaxis Foundation & Research Institute, José Ortega y Gasset 20, 28006, Madrid, Spain
[4] CNR - Institute of Complex Systems, Via Madonna del Piano 10, 50019 Sesto Fiorentino, Florence, Italy
m.zanin@campus.fct.unl.pt; ernestina.menasalvas@upm.es; stefano.boccaletti@fi.isc.cnr.it; pas@holos.pt

Abstract. In the ever-increasing availability of massive data sets describing complex systems, *i.e.* systems composed of a plethora of elements interacting in a non-linear way, *complex networks* have emerged as powerful tools for characterizing these structures of interactions in a mathematical way. In this contribution, we explore how different Data Mining techniques can be adapted to improve such characterization. Specifically, we here describe novel techniques for optimizing network representations of different data sets; automatize the extraction of relevant topological metrics, and using such metrics toward the synthesis of high-level knowledge. The validity and usefulness of such approach is demonstrated through the analysis of medical data sets describing groups of control subjects and patients. Finally, the application of these techniques to other social and technological problems is discussed.

Keywords: Complex systems; complex networks; data mining.

1 Introduction

Networks are all around us: from a social point of view, when we are ourselves, as individuals, the units of a network of social relationships of different kinds [1]; but we are also the result of networks of biochemical reactions, and of electro-chemical interactions between neurons [2]. Furthermore, our world around us is organized in networks, from physical transportation networks [3] up to virtual information webs [4]. While the mathematical formulation of networks as mathematical objects started in 1736 when the Swiss mathematician Leonhard Euler published the solution of the Königsberg bridge problem, graph representations can be found back in 980 AD [5]. Only in recent years, thanks to the increasing capacity of computation centers on the one side, and availability of public data sets on the other, complex network analysis

has witnessed a revolution, which has yielded a vast theoretical and applied body of research. Interested readers may refer to Refs. [6], [7], [8], [9] for further information.

In spite of this evolution, several research questions have still to be tackled, and new ones appear when novel applications of network theory are proposed. Among these, the PhD Thesis "Complex Networks and Data Mining: Toward a new perspective for the understanding of Complex Systems" [10] proposes the use of data mining techniques to solve the following four research questions: *(i)* how to pre-select relevant features for minimizing the cost of network reconstruction, *(ii)* how to design new network reconstruction techniques, *(iii)* how to use network representations to improve data mining tasks, and *(iv)* how to assess and optimize the significance of a network representation.

As will be elaborated in Section 2, the problem posed in the fourth research question is of utmost relevance when the system under analysis is a *Collective Awareness System* (CAS). In this contribution, we review the methodology developed inside [10] to deal with this problem, which allows automatizing the process of obtaining the best network representation of a given system. This guarantees that the highest quantity of information is extracted from the system, thus maximizing the knowledge gained from it.

2 Relationship to Collective Awareness Systems

Under the umbrella of the 'FuturICT' FET Flagship Pilot Project, the European research community has already analyzed the implications and requirements of a collective awareness system, called *Planetary Nervous System* (PNS) [11]. Among the expected benefits of such world-scale sensory system, the PNS would be able to record and mine the digital footsteps created by human activity, as well as to unveil the knowledge hidden in such social big data, thus allowing addressing some fundamental questions about social dynamics. Nevertheless, implementing such system will require overcoming several challenges: some of them of a technical nature and expected to be solved in the next years, as for instance increased computational capabilities, others requiring a change in the data processing paradigm. Specifically, new algorithms for finding patterns in large sets of data will be required, with two specific targets: *i)* handle fragmented, low-level and incomplete data, and *ii)* adapt to the specific characteristics of these data sets, including their networked multi-dimensional nature and semantic richness [11].

Within this context, a natural solution has been already identified in the complex network theory, as it provides a rigorous framework for the study of structures created by relationships between the elements of a complex system [12]. Potential applications include the analysis of the dynamics of social systems, *e.g.* the diffusion of opinions [13] or of diseases [14]; the analysis of mobility patterns, by modeling pairs of origin – destination locations as links; or semantic text analysis, for creating structured taxonomies over texts [15].

If one is to apply complex network analysis to data coming from CAS like the PNS, whose size is expected to exceed the capacity of human analysts, principles and methods should be found to ensure a way for an automated knowledge extraction. The

methodology presented in this contribution aims at providing an automatic procedure for obtaining the best network representation of a given system. Traditionally such optimization step has mainly been performed by means of the expert judgment of the researcher: yet, it is unfeasible to manually optimize the network representation of a CAS, due to the quantity of information it encodes. Thus, we expect an important added value when the methodology here presented is positioned between the data gathering, and the human-based network analysis phases of CAS management.

3 Proposed Methodology

As previously introduced, this Section proposes a novel way of optimizing the network representation of a complex system, *e.g.* a CAS, by means of data mining techniques, as developed in [10]. Such methodology will be presented in Section 3.2: before that, Section 3.1 will introduce the reader to the different ways of constructing a network representation starting from a raw data set.

3.1 Network Reconstruction Frameworks

As a first step in the analysis of a complex system, it is necessary to create a network representation of it; this, in turn, requires two steps: map each element of the system into a node of the network, and assess the existence of a relationship between pairs of nodes.

When relationships between the system elements are defined upon a physical support, their identification is a straightforward task, and the researcher only needs to map them into the network representation. For instance, one may consider the air transportation network: when airports are represented by nodes, links are naturally established between pairs of them if at least one direct flight is connecting these two airports [3].

In the absence of such relationships, links can still be built, provided a vector of *observables*, *i.e.* of measurements representing some properties of the system, can be associated to each node. In this case, each link represents the presence of a *functional relationship* between the data corresponding to that pair of nodes, and the resulting networks are called *functional networks*. For instance, if one is to analyze the structure of a stock market, each stock may be represented by a node, with pairs of them connected whenever there is a significant correlation in their price evolution through time [16]. Fig. 1 reports a simple example, in which a network is created by calculating Person's linear correlation between the evolutions of four U.S. stock prices.

It is important to notice that the requirement of having a vector of observable for each node precludes the use of functional representations for systems whose elements are characterized by a single value. Examples include tissues and organic sample analysis, like spectrography; genetic expression levels of individuals, without evolution through time [17]; biomedical analyses, *e.g.* the study of brain oxygen consumption by means of neuroimaging techniques [18]; or social network analyses, when just a snapshot of users characteristics is available [13]. To overcome this

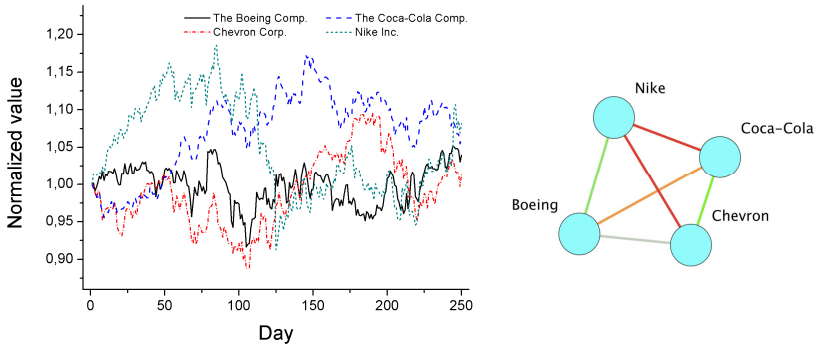


Fig. 1. Example of a *functional network* reconstruction. (Left) Time evolution of four stocks composing the Dow Jones Industrial Average index; each time series corresponds the evolution of prices from 1st January to 31st December 2012, and values are normalized to 1.0 in the first day. (Right) Resulting network, where each node is a stock, and links are weighted according to the Pearson's linear correlation between the corresponding time series; green (red) shades indicate positive (negative) correlations.

limitation, a novel method was recently proposed, which allows treating collections of isolated, possibly heterogeneous, scalars, *e.g.* sets of biomedical tests, as networked systems. The method yields a network where each node represents an observable, and links codify the distance between a pair of observables and a model of their typical relationship within the studied population [19], [20].

3.2 Network Optimization and Analysis

Whether the network representation is assembled by mapping physical connections, by constructing a *functional* representation, or by using the technique proposed in Refs. [19], [20], two further steps are required: *i)* transform the *fully-connected weighted network* (as the one depicted in Fig. 1) into a *structured unweighted network*, and *ii)* extract a set of metrics describing some topological characteristics. It is worth noticing that both steps are characterized by some level of arbitrariness. While binarizing the network, it is necessary to define a threshold, such that links with a weight lower than this reference value are deleted. Also, among the large group of available topological metrics, the researcher has to choose the one he / she considers being relevant for describing the system under study.

A new methodology has been proposed for addressing these two issues, based on the application of data mining techniques [21]. By starting from an external classification as ground truth, *e.g.* control subjects and patients suffering from some disease, it is possible to use the output of a data mining classification task as a proxy for the relevance of the network representation under study. This yields criteria for an optimal network representation with respect to a given problem.

Following the approach proposed in Ref. [21], instead of applying a single pre-determined threshold τ, such that links whose weight is lower than τ are deleted, a set of thresholds $T = \{\tau_1, \tau_2, \ldots\}$ is applied, covering the whole range of applicable thresholds. Furthermore, a large set of measures M is extracted from each network,

including the most relevant macro-, meso- and micro-scale topological features of a complex network (see Ref. [8] for a review of applicable metrics). At the end of the process, the initial raw data are therefore converted into a large set of measures,

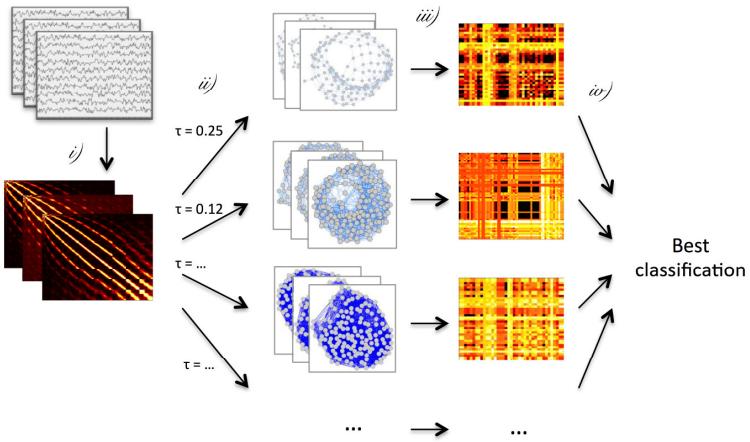


Fig. 2. Optimizing network reconstruction. Following the creation of weighted cliques (step *i*), these are transformed into a set of unweighted adjacency matrices by dint of different thresholds (step *ii*); a set of features is extracted for each network (step *iii*), and this is used as input for a data mining task (step *iv*). Finally, the best classification is used to choose the most relevant threshold and topological metrics. Adapted from Ref. [21].

representing a wide sample of the possible analyses that may be performed from a complex network perspective.

Once the raw data have been transformed into a large set of topological metrics, the problem faced by the researcher is the identification of the optimal subset of metrics for describing the system. Here we propose the use of a data mining classification task for automatizing this process. Specifically, for each threshold τ_i, and for each pair (or triplet) of metrics, subjects are classified; the percentage of subjects correctly classified is then used as a proxy of the relevance of such set of parameters. Indeed, if a good classification is achieved, the considered parameters and network metrics correctly represent the structural differences between the two classes of subjects. Thus, the best classification corresponds to both the best set of metrics and to the corresponding best threshold. Fig. 2 proposes a graphical representation of this process, where information flows from the left (raw data and weighted fully-connected networks) to the right (final classification).

4 An Application to *Mild Cognitive Impairment*

To demonstrate the validity of the proposed approach, we here consider a set of magneto-encephalographic data (MEG), and identify the features that better differentiate healthy subjects from patients suffering from *Mild Cognitive Impairment*

(MCI). MCI is a disease, considered a prodromal stage of *Alzheimer's*, characterized

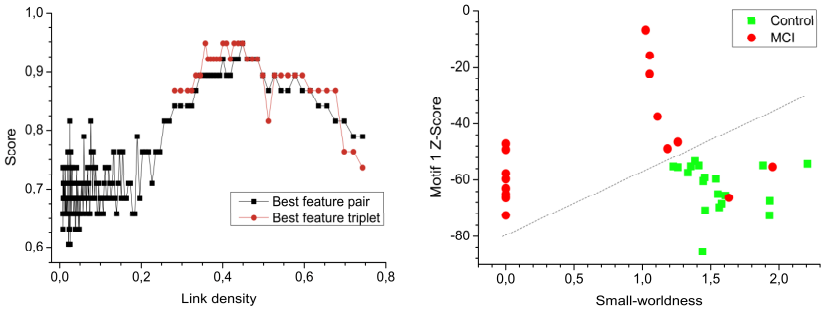


Fig. 3. (Left) Classification score as a function of link density; black (red) points indicate the best classification score obtained using pairs (triplets) of features. (Right) Classification of MCI and healthy subjects; green (red) points represent the position in the space of features of healthy (MCI) patients. Adapted from Ref. [20].

by cognitive impairments beyond those expected based on the age of the patient, but which are not significant enough to interfere with their daily activities. The data set comprises recordings from nineteen patients and nineteen healthy volunteers during a modified Sternberg's letter-probe task, which requires participants to firstly memorize a set of five letters presented on a computer screen, for then pressing a button when a member of the previous set is detected.

Following the methodology proposed in Section 3.2, 178 networks have been created for each subject, corresponding to the number of different thresholds considered. From each one of these networks, 72 different topological metrics have been calculated. A classification task was ultimately performed for each pair and triplet of considered features, using a Support Vector Machine algorithm [22]. Fig. 3 (Left) reports the precision (percentage of correctly classified subjects) corresponding to the most representative pair (triplet, in red) of features, as a function of the link density obtained by applying different thresholds. Classification was also attempted with other algorithms, including Naive Bayes and neural networks [23], producing qualitatively comparable results.

Several relevant conclusions can be derived from Fig. 3. Firstly, the best classification rate (95%) is obtained for sufficiently low threshold values, *i.e.* including a great quantity of links inside the analysis. Specifically, the maximum score corresponds to including about 40% of the links. Remarkably, the functional brain network literature typically considers networks with a 5% link density [24]. The increase in the number of links, as suggested by the proposed methodology, has a major consequence: allowing a better consideration of meso-scale structures, *e.g.* of motifs, that is specific connectivity patterns formed by 3 nodes [17]. Furthermore, results corresponding to low link densities are much more unstable, as demonstrated by the leftmost part of the plot in Fig. 3. Clearly, the addition, or deletion, of a few links has a major effect in the topology, changing the meaning of all metrics calculated on the top of it. Therefore, these results invite to reconsider many studies made in the Literature about functional brain network reconstruction, and validate the

hypothesis that a data mining approach can improve the understanding of complex systems.

5 Conclusion and Discussion

In conclusion, in this contribution we have described and reviewed how the application of data mining techniques can be used to improve and optimize the reconstruction of complex networks, representing for instance Collective Awareness Systems. In turn, the resulting networks can be used to extract knowledge about the topological properties of the corresponding systems, in a way that goes beyond the capacity of classical data mining. Such advantages come at a cost: due to the high number of analyses required, *e.g.* the extraction of several topological metrics for different threshold values, there is an important increase in the computational cost, especially when compared with standard data mining algorithms.

Beyond the proposed biomedical example, such methodology can be applied in any scenario in which a complex network representation is expected to be relevant. Thus, this includes the analysis of any system whose dynamics is defined by the relationships between its elements: from social networks created by interacting individuals, up to technological networks, as communication or transportation systems. Furthermore, such elements may not be homogeneous, or interactions may develop through different channels – what is known as a *multi-layer* (of *multiplex*) network [25], [26]. For all of this, the approach here presented is expected to be of relevance for future applications of complex network techniques to the field of Collective Awareness System.

Acknowledgments. The authors acknowledge the computational resources, facilities and assistance provided by the Centro computazionale di RicErca sui Sistemi COmplessi (CRESCO) of the Italian National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA), and of the Universidad Politécnica de Madrid's CeSViMa (Madrid Supercomputing and Visualization Center).

References

1. Knoke, D., Yang, S.: Social Network Analysis. Sage (2008)
2. Bullmore, E., Sporns, O.: Complex brain networks: graph theoretical analysis of structural and functional systems. Nature Reviews Neuroscience 10, 186--198 (2009)
3. Zanin, M., Lillo, F.: Modelling the air transport with complex networks: A short review. The European Physical Journal Special Topics 215, 5--21 (2013)
4. Albert, R., Jeong, H., Barabási, A. L.: Error and attack tolerance of complex networks. Nature 406, 378--382 (2000)
5. Strano, E., Zanin, M., Estrada, E., Lillo, F.: Spatially embedded socio-technical complex networks. The European Physical Journal Special Topics 215, 1--4 (2013)
6. Albert, R., Barabási, A. L.: Statistical mechanics of complex networks. Reviews of modern physics 74, (2002)

7. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D. U.: Complex networks: Structure and dynamics. Physics reports 424, 175--308 (2006)
8. Costa, L. D. F., Rodrigues, F. A., Travieso, G., Villas Boas, P. R.: Characterization of complex networks: A survey of measurements. Advances in Physics 56, 167--242 (2007)
9. Costa, L. D. F., Oliveira Jr, O. N., Travieso, G., Rodrigues, F. A., Villas Boas, P. R., Antiqueira, L., Viana, M. P., Correa Rocha, L. E.: Analyzing and modeling real-world phenomena with complex networks: a survey of applications. Advances in Physics 60, 329--412 (2011)
10. Zanin, M.: Complex Networks and Data Mining: Toward a new perspective for the understanding of Complex Systems. PhD Thesis (2014)
11. Giannotti, F., Pedreschi, D., Pentland, A., Lukowicz, P., Kossmann, D., Crowley, J., Helbing, D.: A planetary nervous system for social mining and collective awareness. The European Physical Journal Special Topics 214, 49--75 (2012)
12. Havlin, S., Kenett, D. Y., Ben-Jacob, E., Bunde, A., Cohen, R., Hermann, H., Kantelhardt, J. W., Kertész, J., Kirkpatrick, S., Kurths, J., Portugali, J., Solomon, S.: Challenges in network science: Applications to infrastructures, climate, social systems and economics. The European Physical Journal Special Topics 214, 273--293 (2012)
13. Castellano, C., Fortunato, S., Loreto, V.: Statistical physics of social dynamics. Reviews of modern physics 81, (2009)
14. Pastor-Satorras, R., Vespignani, A.: Epidemic dynamics and endemic states in complex networks. Physical Review E 63, 066117 (2001)
15. Navigli, R., Velardi, P., Faralli, S.: A graph-based algorithm for inducing lexical taxonomies from scratch. In: Twenty-Second International Joint Conference on Artificial Intelligence-Volume, pp. 1872—1877. AAAI Press (2011)
16. Mantegna, R. N.: Hierarchical structure in financial markets. The European Physical Journal B-Condensed Matter and Complex Systems 11, 193--197 (1999)
17. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: simple building blocks of complex networks. Science 298, 824--827 (2002)
18. Phelps, M. E., Mazziotta, J. C.: Positron emission tomography: human brain function and biochemistry. Science 228, 799--809 (1985)
19. Zanin, M., Boccaletti, S.: Complex networks analysis of obstructive nephropathy data. Chaos: An Interdisciplinary Journal of Nonlinear Science 21, 033103 (2011)
20. Zanin, M., Alcazar, J. M., Carbajosa, J. V., Sousa, P., Papo, D., Menasalvas, E., Boccaletti, S.: Parenclitic networks' representation of data sets. arXiv:1304.1896 (2013)
21. Zanin, M., Sousa, P., Papo, D., Bajo, R., García-Prieto, J., del Pozo, F., Menasalvas, E., Boccaletti, S.: Optimizing functional network representation of multivariate time series. Scientific reports 2 (2012)
22. Steinwart, I., Christmann, A.: Support vector machines. Springer (2008)
23. Bishop, C. M., Nasrabadi, N. M.: Pattern recognition and machine learning. Springer (2006)
24. Buldú, J. M., Bajo, R., Maestú, F., Castellanos, N., Leyva, I., Gil, P., Sendiña-Nadal, I., Almendral, J. A., Nevado, A., del-Pozo, F., Boccaletti, S.: Reorganization of functional networks in mild cognitive impairment. PLoS One 6, e19584 (2011)
25. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., Porter, M. A.: Multilayer Networks. arXiv:1309.7233 [physics.soc-ph] (2013)
26. Cardillo, A., Gómez-Gardeñes, J., Zanin, M., Romance, M., Papo, D., del Pozo, F., Boccaletti, S.: Emergence of network features from multiplexity. Scientific Reports 3 (2013)